

ABSTRACT

DECHENE, MICHELLE C. Protein Interactions: the Multiple Solvent Crystal Structures of RNase A and Analysis of the RalA and RalBP Complex. (Under the direction of Carla Mattos).

In both structure and function, Ribonuclease A (RNase A) and RalA are two very different proteins. RNase A is an extracellular digestive enzyme that catalyzes the breakdown of 3'-5' phosphodiester linkages in single stranded RNA. RalA is a small monomeric GTPase of the Ras family and is involved in a number of signaling pathways. While the basic fold of RalA is similar to the rest of the Ras family, Ral proteins have a distinct effector binding region and set of effector proteins. RalBP was the first RalA effector identified and it links RalA to receptor mediated-endocytosis and regulation of mitosis.

RNase A is a small kidney shaped protein with a well defined active site cleft running between the two lobes. The active site consists of several pockets, which are responsible for binding nucleotide bases and phosphate moieties of the RNA substrate. This enzyme is well studied. With over 40 years of structural information available, it is an excellent model protein for quantitatively defining the strengths of the Multiple Solvent Crystal Structures (MSCS) Method. MSCS is an experimental method using small organic solvent molecules to map the surface of proteins, and in addition to locating binding sites, provides information about patterns of protein hydration and plasticity. Twenty-two solvent soaked structures were generated revealing 16 organic solvent molecules and 12 sulfate ions clustered in the active site, specifically in the two nucleotide-binding pockets, B₁ and B₂, and in the catalytic pocket, P₁. A comparison of the solvent clusters and the available RNaseA-inhibitor structures revealed that the probe molecules interact with key hot spot residues necessary for

ligand binding. Additionally, conserved water molecules were identified on the surface of RNase A. Outside of the active site, many of these water molecules are involved in stabilizing interactions, or are associated with one of the three helices of RNase A. In the active site, nine well-ordered water molecules, which stabilize the active site, bridge the interaction between the ligand and the active site residues, or are displaced upon ligand binding, were identified. These patterns of hydration are consistent with earlier analyses of RNase A. Finally, RMSD and the hinge angle were used as tools to quantitate the plasticity observed at each residue and overall domain motions relative to one another, respectively. In addition to identifying rigid residues of the active site and those exhibiting more motion, it was found that the trends observed in the MSCS structures correlated well with those observed in other crystal and NMR structures of RNase A.

RalA interacts with effector proteins through its two flexible regions, termed switch I and II, which adopt different conformations in response to the nucleotide binding state. Effector proteins recognize RalA in the GTP-bound “on” state, and bind through these switch regions. Where the Ras Binding Domains (RBD) of Ras effectors all adopt a similar fold and interact with active Ras through an intermolecular β -sheet involving switch I, the recent structures of RalA-effector complex structures of RalA-Sec5 and RalA-Exo84 reveal Ral effector Ral binding domains differ in structure and in the binding mode with RalA. In a third Ral effector, RalBP, the Ral-binding domain is predicted to be α -helical, which is different from the β -sandwich structures of Sec5 and Exo84, suggesting the RalA-RalBP interaction presents a previously unobserved binding mode. Furthermore, structural analysis using circular dichroism revealed that the Ral binding domain of RalBP is intrinsically disordered

and folds upon binding to RalA. This is the first example of a Ras family effector with this behavior. Significant advances have been made towards the crystallizing of the RalA-RalBP complex, resulting in preliminary crystals.

Protein Interactions: the Multiple Solvent Crystal Structures of RNase A and
Analysis of the RalA and RalBP Complex

by
Michelle C. Dechene

A dissertation submitted to the Graduate Faculty of
North Carolina State University
In partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

Functional Genomics

Raleigh, North Carolina

2008

APPROVED BY:

Dr. Carla Mattos
Committee Chair

Dr. Maria Celeste Sagui

Dr. Robert Rose

Dr. Alexander Tropsha

BIOGRAPHY

Michelle Dechene was raised in the suburbs of Chicago, IL. After developing an early interest in biology, she pursued this fascination, as well as an interest in Computer Science, at the University of Dayton, in Dayton, OH. She graduated *cum laude* with a B.S. in Biology and Computer Science. Her undergraduate research project involved protein homology modeling. Michelle then enrolled in the Ph.D. program in Functional Genomics at North Carolina State University, in Raleigh, NC so she could pursue studies combining her two majors. While working in the lab of Dr. Carla Mattos, she optimized the constructs and purification of RalBP used for obtaining crystals of the RalA-RalBP complex and developed computational tools for Multiple Solvent Crystal Structure analysis. In addition to coursework and academic research, Michelle also gained professional experience through two internships. She worked as a software engineering intern developing software for in-house use at CIPHERGEN Biosystems and as a medicinal chemistry intern conducting research in protein crystallography for Celera Genomics.

ACKNOWLEDGMENTS

Thank you to Dr. Carla Mattos for serving as Michelle's advisor and providing amazing opportunities, lab space, guidance, education, ideas, and encouragement.

The National Science Foundation Integrative Graduate Education and Research Traineeship (NSF-IGERT) Fellowship provided key financial support during her first three years and a later semester of graduate school. The IGERT Fellowship provided for tuition and fees, a stipend, and supplemental funds for supplies.

Thank you to Dr. Bob Rose, Dr. Celeste Sagui, and Dr. Alex Tropsha for serving as Michelle's graduate advisory committee and contributing helpful suggestions and guidance.

Data were collected at the Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Supporting institutions may be found at www.ser-cat.org/members.html. Use of the Advanced Photon Source was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. W-31-109-Eng-38.

Thanks also to the past and present members of the Mattos Lab, especially, Dr. Greg Buhrman, Dr. Senthil Kumar, Dr. Nate Nicely, Winnell Newman, and Dr. Paul Swartz. I have learned so much from you and I appreciate your willingness to answer my questions and help me think through problems.

Thank you to Dr. Janet Thornton who welcomed me into her group at the European Bioinformatics Institute for a month. It was this experience and the assistance and discussions from her and Dr. Roman Laskowski, in particular, that helped start my work on the computational analysis of the MSCS data for RNase A.

Finally, thank you to my parents, my family, and my friends. Your support has been invaluable.

TABLE OF CONTENTS

List of Figures	viii
List of Tables	x
Chapter 1: Introduction	1
Protein Interfaces	3
Protein Interaction Hot Spots.....	4
The Role of Water Molecules in Protein Interactions	6
Protein Flexibility Associated with Protein Interactions	8
Intrinsically Disordered Proteins Couple Folding and Binding.....	10
My Work to Study Protein Interactions	12
Ribonuclease A	13
RalA and RalBP	14
The Multiple Solvent Crystal Structures Method	17
References	19
Chapter 2: Multiple Solvent Crystal Structures of Ribonuclease A: A Critical Assessment of the Method.....	27
Abstract.....	27
Introduction.....	27
Materials and Methods.....	31
Crystal Growth, Cross-linking, and Solvent Soaks	31
Data Collection, Processing, Structure Refinement.....	33
Water Renumbering.....	33
Comparison of MSCS with RNase A Structures Solved in Aqueous Solution.....	36
Computational Analysis.....	38
Results.....	40
RNase A MSCS Models	43
Analysis of Plasticity Based on Pairwise RMSD Values Between the Models.....	45
Conserved Water Binding Sites	58
Water Molecules in the Active Site	68
Organic Solvent Binding Sites and Comparison with Inhibitors.....	72
Solvents Bound in the B ₁ Subsite.....	74
Solvents Bound in the P ₁ Subsite	81
Solvents Bound in the B ₂ Subsite.....	84
Additional Active Site Solvent	86
Solvents in Crystal Contacts	86
Other Surface Solvents	87
Discussion.....	89
Protein Plasticity	89
Hydration	91
Binding of Organic Solvents.....	94

Conclusions.....	96
Acknowledgements.....	97
References.....	97
Chapter 3: Multiple Solvent Crystal Structures of Ribonuclease A: A Comparison of the P3 ₂ 21 and C2 Spacegroups	102
Abstract.....	102
Introduction	103
Materials and Methods.....	106
Crystal Growth, Cross-linking, and Solvent Soaks	106
Data Collection, Processing, Structure Refinement.....	107
Water Renumbering.....	107
Comparison of Structures and Preparation of Files	110
Computational Analysis.....	111
Results.....	113
RNAse A Models.....	113
Protein Plasticity	114
Molecular Hinge	122
Conserved Water Binding Sites.....	128
Active Site Waters	129
Waters Outside the Active Site.....	136
Water in Crystal Contacts.....	138
Water Molecules Not Conserved in the P3 ₂ 21 MSCS Structures	138
Organic Solvent Binding Sites.....	140
Solvents Bound in the Active Site	147
Solvents Bound in Crystal Contacts	151
Other Surface Solvents	152
Discussion	153
Protein Plasticity.....	153
Hydration	157
Binding of Organic Solvents.....	159
Conclusions.....	161
Acknowledgements.....	161
References.....	162
Chapter 4: The Interaction of RalA and RalBP	165
Abstract.....	165
Introduction.....	166
Materials and Methods.....	171
Expression of GST-RalBP (all constructs).....	171
Batch Affinity Purification of GST-RalBP(397-518x) and Thrombin Cleavage.....	172
Original Purification of RalBP(397-518x).....	173
Expression and Purification of RalA	174
RalA-RalBP(all constructs) Complex Formation and Purification	176

Optimized Purification of GST-RalBP(all constructs) and Solution Thrombin Cleavage	176
DNA Sequencing	177
Mutagenesis	178
Prediction of Naturally Disordered Regions.....	179
Circular Dichroism Spectroscopy	179
Engineering Truncated Forms of the Double Serine Mutant of RalBP	181
Concentrated Thrombin Cleavage and Purification of RalBP(391-444).....	182
Expression and Purification of RalA(11-178)	184
Crystallization Trials.....	186
Results.....	187
Initial Purification of RalBP(397-518x)	187
RalA-RalBP(397-518x) Complex Formation	189
RalBP(397-518x) Oligomerizes Through its Two Cysteine Residues	191
Sequencing and Double Serine Mutant.....	193
RalBP(397-518x) Folds Upon Binding to RalA.....	196
RalBP(403-499) Does Not Bind to RalA.....	202
Six New Truncated Constructs of RalBP Designed	204
RalBP(397-518) Binds RalA	206
RalBP(391-444) Binds RalA	208
Crystal Screening.....	211
RalA(11-178) Binds RalBP(391-444) and Crystal Screening.....	213
Discussion.....	215
Advances Toward the Crystal Structure of the RalA-RalBP Complex	215
RalA-Effector Interactions.....	219
RalA Hot Spot Identification Through Structural Analysis.....	220
Conformational Change at the Ral-Effector Interface	221
Summary	223
Acknowledgements.....	224
References.....	224
Concluding Remarks.....	231
Appendices.....	233
Appendix A: Perl Scripts	234
Appendix B: Downloaded PDB Files.....	300

LIST OF FIGURES

Chapter 2: Multiple Solvent Crystal Structures of Ribonuclease A: A Critical Assessment of the Method	27
Figure 1: Structure of RNase A	41-42
Figure 2: RMSD calculations: High, Low, and Average RMSD per residue..	49-54
Figure 3: Solvation of RNase A.....	59
Figure 4: Plasticity and conserved water binding sites of RNase A.....	65
Figure 5: Conserved water molecules in the active site of RNase A.....	70
Figure 6: Organic solvent binding sites	73
Figure 7: Organic solvent and inhibitor binding in the active site.....	80
Chapter 3: Multiple Solvent Crystal Structures of Ribonuclease A: A Comparison of the P3 ₂ 21 and C2 Spacegroups	
Figure 1: RMSD calculations: High, Low, and Average RMSD per residue..	116-118
Figure 2: Hinge motions of RNase A	123
Figure 3: RNase A hinge calculations: Thr 45 N – Phe 120 N distance vs hinge Angle	128
Figure 4: Conserved water molecules in the active site.....	133
Figure 5: Organic solvent binding sites	146
Figure 6: Organic solvent binding sites across all MSCS structures.....	146
Figure 7: Sulfate ion binding in the P ₁ pocket of the active site	148
Figure 8: Organic solvent and inhibitor binding in the active site.....	150
Chapter 4: The interaction of RalA and RalBP	
Figure 1: SDS-PAGE gel of concentrated RalBP(397-518x).....	188
Figure 2: SDS-PAGE gel showing the co-elution of RalA and RalBP(397-518x) in size-exclusion chromatography	189
Figure 3: SDS-PAGE gel showing the anomalous behaviour of RalBP397-518x) during the co-elution of RalA and RalBP(397-518x) in size-exclusion chromatography	190
Figure 4: SDS-PAGE gel showing RalBP(397-518x) after the second incubation with glutathione-agarose.....	191
Figure 5: Native and SDS-PAGE gels showing the oligomerization of RalBP(397-518x).....	192
Figure 6: Translated DNA sequencing results of GST-RalBP(397-518x)	195
Figure 7: Translated DNA sequencing results of the GST-RalBP(397-518x) double mutant, GST-RalBP(397-518x;C411S,C451S)	195
Figure 8: The double serine mutant, RalBP(397-518x;C411S,C451S), binds RalA	197
Figure 9: PONDR predicts RalBP(397-518x) to be mostly disordered.....	198
Figure 10: Circular Dichroism shows RalBP(397-518x) folds upon binding to RalA.....	200-201

Figure 11: Native and SDS-PAGE gels showing RalBP(403-499) does not bind to RalA.....	203
Figure 12: Sequences of RalBP constructs	206
Figure 13: RalBP(397-518) binds RalA	207
Figure 14: SDS-PAGE gel of purified RalBP(391-444).....	209
Figure 15: RalBP(391-444) binds RalA	210
Figure 16: Select examples of High-Throughput Screening results	212
Figure 17: RalBP(391-444) binds RalA(11-178)	214
Figure 18: Diffraction pattern from the RalA(11-178)-RalBP(391-444) complex	215

LIST OF TABLES

Chapter 2: Multiple Solvent Crystal Structures of Ribonuclease A: A Critical Assessment of the Method	27
Table 1: Data Collection and Refinement Statistics for Apo-RNase A.....	34
Table 2: Percentage of residues in each domain that fall at or below the baseline of the backbone average RMSD calculations for each set of structures.....	47
Table 3: Conserved Water Molecules identified by SEWS for the MSCS structures of RNase A	61-62
Table 4: Water Molecules with less than 80% conservation in the MSCS Structures	63
Table 5: Bound Organic Solvents and the interactions made with RNase A ..	75-77
Chapter 3: Multiple Solvent Crystal Structures of Ribonuclease A: A Comparison of the P3 ₂ 21 and C2 Spacegroups	102
Table 1: Data Collection and Refinement Statistics for RNase A with a Sulfate Ion in the Active Site	108-109
Table 2: Hinge Angle Calculations.....	125-127
Table 3: Conserved Water Molecules identified by SEWS for the MSCS structures of RNase A	130-132
Table 4: Water Molecules with less than 80% conservation in the MSCS structures	139
Table 5: Bound Organic Solvents and the interactions made with RNase A ..	141-145
Chapter 4: The Interaction of RalA and RalBP	165
Table 1: Parameters used with the Jasco J-600 for Circular Dichroism	180
Table 2: Primers for RalBP truncation.....	182
Table 3: Nomenclature of expressed and purified GST-RalBP constructs.....	183
Table 4: Crystallization Trials	185-186
Appendix B: Downloaded PDB Files	300
Table 1: Downloaded PDB Files	301-310

CHAPTER 1: Introduction

Proteins have the ability to bind with molecules in the cell, such as other proteins, nucleic acids, polysaccharides, and additional ligands. It is through these interactions that proteins perform many of their biological roles; malfunctions in this activity can lead to disease.

Thus, significant effort has been made to understand the fundamental features that facilitate molecular complexes involving proteins. One approach to determining characteristics of interaction interfaces has been done through database analysis of the protein complex structures available in the Protein Data Bank (PDB). With approximately 44,500 crystal structures compared to about 7,500 NMR structures, database analysis of the PDB is biased in favor of protein interactions for which there are crystal structures. Additionally, structural database analysis is biased towards protein interactions for which there are available structures, which excludes most intrinsically disordered proteins.

In this dissertation, I use the X-ray crystallography based Multiple Solvent Crystals Structures (MSCS) method to predict the binding sites and analyze the surface of Ribonuclease A (RNase A) in order to elucidate the interaction features between this enzyme and its substrate. As there are 40 years of structural biology studying this protein, the project is focused on quantitatively defining the strengths of MSCS as a tool for obtaining binding site properties relevant to aqueous solution using organic solvents as probes. The latter portion of this dissertation is focused on a less studied protein, RalBP, and its complex with the GTPase RalA. I use various methods, including liquid chromatography, polyacrylimide gel electrophoresis, and circular dichroism to study the protein-protein

interaction of RalA and RalBP, while working toward protein crystals of the complex. RNase A is a small extracellular enzyme that binds and cleaves RNA in its well-defined active site cleft. While it is essentially a “lock-and-key” type enzyme, RNase A exhibits domain breathing motions in response to ligand binding. RalA is a GTPase belonging to the Ras family. It is an intracellular signaling protein that interacts with effector proteins through its flexible switch regions, which adopt different conformations based on its nucleotide binding state. Presumably, effector molecules bind RalA when an appropriate switch conformation is present. RalBP is multidomain protein that interacts with a number of proteins, in addition to being an effector of RalA. Little is known about the structure of RalBP, aside from being predicted to be highly α -helical, or how RalBP binds to RalA, except that it is via the predicted minimum Ral-binding domain.

This treatise is composed of five chapters. Chapter 1 includes an overview of the features of protein interactions which appear later in this dissertation. This includes hot spots, water molecules, and protein flexibility and plasticity observed in protein interactions, and is followed by a brief description of intrinsically disordered proteins. Additionally, RNase A, the interaction of RalA and RalBP, and the MSCS method are briefly introduced. Chapters 2 and 3 contain the MSCS of RNase A. Two crystal forms were obtained for this protein and Chapter 2 presents the results from the crystals grown with symmetry of the C2 space group containing the apo-enzyme. These structures contain two RNase A molecules in the asymmetric unit. Chapter 3 compares the results from the second crystal form with symmetry of the P3₂21 space group, to those described in Chapter 2. The P3₂21 crystals

have only one RNase A molecule in the asymmetric unit, and each protein has a sulfate ion bound in the active site. Chapter 4 details the work with the complex of RalA and RalBP, including the discovery that the Ral-binding domain of RalBP is a natively unfolded protein in solution and folds upon binding to RalA, and the growth of an initial crystal of the protein complex. Brief concluding remarks follow Chapter 4. The Perl scripts written for the analysis of MSCS data of RNase A are included in Appendix A, and references for and descriptions of the RNase A structures downloaded from the Protein Data Bank (PDB) are presented in Appendix B.

Protein Interfaces

As protein interactions are critical to biological function, a substantial amount of work has gone into studying them. A logical place to start is the protein interface where these interactions occur, i.e., the regions of proteins designed to interact with other proteins and substrates. An analysis of available homodimeric, heterodimeric, enzyme-inhibitor, and antibody-protein complexes was performed to characterize the features of these interfaces and to determine if there was a predictive element distinguishing the protein interface from the rest of the surface (Jones and Thornton, 1996). While a number of characteristics were examined, it was determined that there are no definitive parameters for these characteristics and no single characteristic differentiates binding sites (Jones and Thornton, 1996; Jones and Thornton, 1997a). Protein-protein surfaces tend to be somewhat planar and accessible, and enzymes that bind smaller ligands tend to have a surface cleft that binds partner molecules, but the other properties differ based on the nature of the complex (Jones and Thornton,

1997a). In addition to surface shape and accessibility, properties of protein interactions and the predictability of interfaces were examined, including electrostatic and shape complementarity between interacting surfaces, the likelihood of any given residue of being found at an interface, hydrophobicity, conformational changes upon complex formation, solvation potential of surface patches, and the protrusion of a residue from the surface of the protein (Jones and Thornton, 1996; Jones and Thornton, 1997a). Ensuing studies on protein interaction interfaces have proven the complexity of protein interactions and highlighted differences observed in different complexes.

This introduction will concentrate on the features of protein interactions that are examined in this dissertation: protein interaction hot spots, the role of water molecules in protein interactions, and protein flexibility associated with protein interactions. As much of what is known about protein interfaces is derived from the available three-dimensional structures, the topic of intrinsically disordered proteins and binding is presented following the topic of protein flexibility.

Protein Interaction Hot Spots

Many groups have studied the interface of protein interactions and identified so-called hot spots, but in each instance this term indicates something slightly different. In alanine scanning experiments, a hot spot indicates a single residue that can contribute a large portion of the binding free energy to the interface (Clackson and Wells, 1995). A later computational study identified hot spots as “adhesive” areas on the protein surface where

displacement of water during binding would have a bond stabilizing effect (Fernández and Scott, 2003). In studies probing the surface of proteins with small organic molecules, hot spots describe areas where the probe molecules cluster (Landon et al., 2007; Mattos et al., 2006; Sheu et al., 2005). While the approaches and descriptions are different, the overall idea is the same, hot spots are areas or residues on the surface of the protein, which fall in a binding site and are important for the specificity of the interaction.

Hot spots confer specificity to the binding site. In a study examining hot spots determined by alanine scanning, it was found that the hot spot of one protein packs against the hot spot of its binding partner in a heterodimer interface (Bogan and Thorn, 1998). These hot spots are enriched in tryptophan, arginine, and tyrosine residues, which are presumably preferred at hot spots because of their ability to make multiple types of favorable interactions (e.g. hydrogen bonding, hydrophobic, and aromatic) that can be recognized through counterpart hot spots (Bogan and Thorn, 1998). Furthermore, it has been shown that conserved polar residues correlate with interface hot spots (Ma et al., 2003). This is consistent with the finding that interface residues are generally more highly conserved than other surface residues (Yao, et al., 2003). However, the conservation of residues at interfaces compared to the rest of the surface was determined using available structures in the PDB, and it has been suggested that residue conservation may not be as discriminating as these studies proposed (Caffrey et al., 2004), and, in some instances, may not be useful at all (Bradford and Westhead, 2003).

As hot spots contribute to the specificity of binding, it is not surprising that there are numerous methods used to locate them and the functional groups that interact with these residues. There are several computational methods, including GRID (Goodford, 1985; Wade and Goodford, 1993), MCSS (Caflisch et al., 1993; Miranker and Karplus, 1991), and CS-Map (Kortvelyesi et al., 2003; Silberstein et al, 2003), which use small molecular probes to identify favorable interactions sites. MSCS (a description follows in a later section) is an experimental crystallographic method that uses small organic solvent molecules to probe the surface of proteins (Allen et al., 1996; Mattos et al., 2006; Ringe and Mattos, 1999). SAR by NMR uses small organic molecules or compounds with different functional groups to bind to subsites of the binding surface and then these molecules are optimized and linked together to produce high affinity ligands (Hajduk et al, 1997; Shuker et al, 1996). Using small organic compounds to probe the binding surface and then using them to build larger, more potent molecules is the same principle behind fragment based lead discovery, which may use NMR or crystallography (Carr et al, 2005; Hartshorn et al., 2005; Rees et al., 2004).

The Role of Water Molecules in Protein Interactions

Considering that water is the solvent of life and that proteins have evolved in this ubiquitous solvent, it is not surprising that some water molecules have been specifically incorporated into protein structure and interactions. Our understanding of these water molecules comes from X-ray crystallography, NMR, and computational studies (Karplus and Faerman, 1994).

The traditional view of the role of water in protein interactions is one of departure: protein-protein or -ligand binding is entropically favorable because of the release of ordered water into the bulk solvent. Indeed, it has been found that hot spots tend to be located at the center of interfaces while being surrounded by “energetically unimportant” residues with the function of protecting the hot spot residue from the bulk solvent during protein interactions (Bogan and Thorn, 1998). Furthermore, database analysis revealed that a majority of specific protein interaction interfaces were “dry” with a ring of water molecules around the interface (Rodier et al., 2005). Additionally, binding sites have been found to contain certain easily displaceable bound water molecules, which make specific interactions with the polar groups. These interactions are reproduced by polar functional groups of the binding partner (Ringe, 1995). It is not surprising that the solvation potential of surface patches and dehydrons, or defectively packed backbone hydrogen bonds where water exclusion would have a bond stabilizing effect, have been used as computational predictors of protein interaction interfaces (Jones and Thornton, 1997a; Jones and Thornton, 1997b; Fernández and Scott, 2003).

While removal of water molecules is important, it is only half of the story, and the tightly bound ordered water molecules that stick around during complex formation have drawn considerable interest. It has been proposed that water-mediated contacts in protein-protein interactions are specific and add another mechanism for recognition (Papoian et al., 2003). On average, protein-protein interfaces contain 22 water molecules with 11 water-mediated hydrogen bonds, or about one interface water per 100 Å² (Janin, 1999; Lo Conte et al., 1999). These water molecules contribute to the packing of the interface atoms and to the shape and

charge complementarity of the interacting surfaces (Lo Conte et al., 1999). The role of water in complex formation is not exclusive to protein-protein interactions. In a study analyzing 392 high resolution crystal structures of protein-ligand structures, it was found that over 85% of these complexes have one or more bridging water molecules present at the interface (Lu et al. 2007). Also, it is estimated that 15% of all protein-DNA hydrogen bonds are water mediated (Luscombe et al, 2001). These water-mediated interactions are as important as direct hydrogen bonds when it comes to stability and specificity (Janin, 1999). Water molecules add information to the protein interface, and can be considered as part of the binding site, particularly when they make multiple hydrogen bonds with the protein (Mattos and Ringe, 2001; Teyra and Pisabarro, 2007). All of this illustrates that water is a key player in protein interactions, and it is necessary to consider the role water molecules play at interfaces.

Protein Flexibility Associated with Protein Interactions

When working with protein structures, it is easy to think of them as rigid molecules as described by the lock-and-key representation of protein interactions. However, proteins are flexible molecules and can adapt their structure to enhance binding to partner molecules as is portrayed in the induced fit model (Koshland, 1960). Conformational change upon complex formation is characteristic of protein interfaces, although there are many variations on this theme. While some proteins may exhibit no change upon complex formation, side chain movements, main chain segment movements, and entire domain movements have been observed as a result of binding (Jones and Thornton, 1996, Najmanovich et al., 2000).

Domain movements are particularly relevant to enzyme complexes and can result in actions that enhance binding and catalysis, such as optimally orienting the substrate or squeezing out excess water from the active site. (Hammes, 2002). In a database analysis on cases where structures of the free and complexed forms of proteins were available, it has been found that the size of the interaction interface is related to the conformational changes occurring upon association. “Large” interfaces are those that bury 2000-4660 Å² upon complex formation. Binding at these interfaces is observed to include large conformational changes, such as disorder-to-order transitions, large movements in loops switching to different conformations, and changes in relative positions of domains in multidomain proteins (Lo Conte et al., 1999). Binding at interfaces designated as “standard size” (a total buried area of 1600 (±400) Å²) involves small conformational changes, such as shifts in surface loops, small movements of short portions of the polypeptide chain, and adjustments in the surface side chains (Lo Conte et al., 1999). Additionally, another database study examining side chain flexibility of available apo- and holo- forms of protein structures revealed that only a small number of interface side chains undergo changes, with about 85% of the cases examined showing changes in three residues or less (Najmanovich et al., 2000).

Plasticity of interface residues contributes to specificity. While changes in conformation seem to be important for optimizing the complex, residues that adopt a more fixed arrangement tend to lend specificity to the interaction. This is illustrated by idea of “anchor” and “latch” side chains in protein-protein interactions (Rajamani et al., 2004). Anchor side chains belong to functionally important residues and in the apo-protein are found in

conformations similar to those in the bound complex. Once the anchor docks into a structurally constrained binding groove of the other protein, an induced fit process occurs where the flexible latch side chains adjust forming the final high affinity complex. Latch side chains are found at the periphery of the pocket in conformations allowing for the clamping that leads to the high affinity complex (Rajamani et al., 2004). This idea is confirmed by molecular dynamics simulations, which found that even in the absence of a binding partner, central interface residues tend to be less mobile than other amino acids on the surface of the protein. Where the side chains at the periphery of the pocket demonstrate more motion (Smith et al., 2005). A second contribution to specificity through plasticity is through adjustments in the size of the binding site. For example, the mutants of α -Lytic protease have a broader specificity than the wild type protein resulting from the increased plasticity of the active site. This allows for adjustments in the size of the binding pocket, thus accommodating both large and small substrates (Bone et al., 1989).

Intrinsically Disordered Proteins Couple Folding and Binding

Intrinsically disordered (or natively unfolded or unstructured) proteins have gained interest in recent years. These proteins are common in eukaryotes and less so in prokaryotes, and it is thought that possibly 30% of eukaryotic proteins are fully or partially disordered (Dyson and Wright, 2002; Fink, 2005). The high proportion of these proteins in genomes suggests that unstructured proteins play an important role in the cell and, in fact, there are numerous examples with a broad range of functions, including signal transduction, cell cycle control, transcriptional and translational regulation, protein phosphorylation, small molecule storage,

and regulation of multiprotein complex self-assembly (Wright and Dyson, 1999; Dyson and Wright, 2005, Uversky, 2002a). With the growing numbers of known disordered proteins and increasing interest in them, efforts have been made to develop methods to computationally predict regions of disorder. IUPred predicts disordered regions by estimating the total pair-wise inter-residue interaction energy from an amino acid sequence (Dosztányi et al., 2005a; Dosztányi et al, 2005b). This approach is different from PONDR VL3H and DISOPRED2, which are predictors trained on structurally determined regions of disorder (Obradovic, 2003; Ward, 2004). These methods have had some success and an experiment using PONDR showed that compared to eukaryotic proteins in the Swiss-Prot database and well-ordered protein segments from the PDB, proteins involved in signaling and those associated with cancer had much higher percentages of sequences predicted to have long consecutive disordered regions (Iakoucheva et al., 2002).

Disordered proteins lack a stable, well-defined structure and exist in a range of states from molten globule to random coil (Uversky, 2002b). Often, these proteins fold upon binding to their biological targets, involving anywhere from a few residues to an entire domain (Wright and Dyson, 1999; Dyson and Wright, 2002; Dyson and Wright, 2005). Lack of structure and the disorder to order transition during binding of specific targets are believed to give disordered proteins the advantages of being able to bind several different targets, and of having precise control over the thermodynamics of the binding process (Wright and Dyson, 1999; Dyson and Wright, 2002; Uversky, 2002a, Fink, 2005). It is thought that elements of secondary structure serve as targets for recognition initiating the coupled folding and binding

process. Illustrating this, a database analysis of intrinsically disordered proteins with known three-dimensional structures revealed that preformed elements of structure could serve as initial contact points, followed by folding of the unstructured protein onto the template of the interface (Fuxreiter et al., 2004). Alternatively, a molecular dynamics study presented that the folding of the disordered p27 protein was not caused by its own structural preferences, but instead by the requirements to form a specific molecular interface. When bound in complex to cyclin A, p27 adopts a β -hairpin and β -strand conformation, as opposed to its natural folding preference for an α -helix (Verkhivker et al., 2003). This selection among conformations adopted by a disordered protein is potentially what allows for their ability to bind several different targets. As disordered proteins pose a major challenge for crystallography and NMR studies, these proteins are under-represented in structural databases, and much remains unknown about how these proteins function and interact.

My Work to Study Protein Interactions

The object of this dissertation is to study protein interactions by examining three different proteins: Ribonuclease A, RalA, and RalBP. Ribonuclease A binds and catalyzes the hydrolysis of RNA in a well-formed cleft, and is an example of a protein-ligand interaction. RalA, on the other hand provides a good system to study protein-protein interactions, as it binds to numerous effector proteins dependent upon conformational changes in its switch regions. RalBP is an effector for RalA, and little is known about its three dimensional structure or how it interacts with RalA.

Ribonuclease A

Ribonuclease A (RNase A) has been studied for almost a hundred years, beginning with the report that the surviving active agent from boiled aqueous pig pancreatic extract could break down yeast nucleic acid into nucleotides (Jones, 1920). The crystallization of this protein in 1939 began the modern work with RNase A (Kunitz, 1939). As evidenced by the fact that RNase A can survive being boiled, RNase A is a highly stable protein. This coupled with its ease of collection, as it is excreted from the pancreas, purification, and crystallization, resulted in Armour, Inc preparing over a kilogram of highly pure crystalline protein which it then sold to the biochemical community for a small fee (Richards and Wyckoff, 1971). The ready availability of RNase A kicked off a flurry of research on this protein.

RNase A is an endoribonuclease that catalyzes the breakdown of 3'5'-phosphodiester linkages in single stranded RNA at the 3' side of pyrimidine nucleotides. Cleavage of RNA occurs in the two subsequent transphosphorylation and hydrolysis reactions where His12 and His19 are involved in general acid-base catalysis. Lys41 stabilizes the 2'3'-cyclicphosphodiester intermediate. This mechanism has been well studied and has been reviewed (Raines, 1998).

The structure of Bovine Pancreatic Ribonuclease A has been studied for over 40 years. Fankuchen conducted the first x-ray diffraction study in 1941 (Fankuchen, 1941). This led to further crystallographic investigation of RNase A, and in 1967, it was the fourth protein to have its three-dimensional structure determined by X-ray crystallography (Kartha et al. 1967;

Wyckoff et al. 1967). RNase A is 124 amino acids long and contains eight cysteines, all of which are involved in disulfide bonds (Smyth et al., 1963). The overall shape of the protein resembles a kidney with the active site in the cleft between the two lobes. The structure is dominated by a β -sheet and there are three short α -helices. The two lobes have been shown to exhibit breathing or hinge motions where the hinge closes slightly upon substrate binding (Radha Kishan et al, 1995; Sadasivan et al, 1998; Merlino et al, 2002; Vitagliano et al, 2002; Beach et al, 2005). The active site is divided into several subsites, used to describe substrate binding. Phosphate moieties bind in subsites P₀, P₁, and P₂, where P₁ is the site where cleavage occurs. Nucleotide bases bind in subsites B₁ and B₂, where B₁ is selects for pyrimidines (Raines, 1998). Structural studies of RNase A have continued over the years and there are over 100 structures of this protein deposited in the PDB, with about half containing an inhibitor bound in the active site.

RalA and RalBP

Compared to RNase A, RalA and RalBP are newcomers to the world of protein research. RalA was identified a little over 20 years ago in a search of a cDNA library to find Ras family genes (Chardin and Tavitian, 1986). Since then, work has been done to reveal the structure of this protein and some of the ways it functions in the cell. (For a summary of roles and binding partners of RalA, see Chapter 4). In many ways RalA is similar to Ras. It is a small GTPase, which hydrolyzes GTP to GDP and in so doing, an accompanying conformational change occurs in the so-called switch regions. The mechanism of the conformational change in the switch regions has been described as a loaded spring (Vetter

and Wittinghofer, 2001). In the GTP-bound form, the backbone amino groups of Thr46 and Gly71 form hydrogen bonds with γ -phosphate oxygen atoms. With the release of the γ -phosphate after GTP hydrolysis, the two switch regions can relax into their GDP conformation (Vetter and Wittinghofer, 2001). It is through these switch regions that effectors are bound, recognizing the “on” GTP-bound state.

Ras and RalA are over 50% homologous (Chardin and Tavitian, 1986), and share the same overall structural fold (Nicely et al., 2004). RalA is 206 amino acids long and the structure is composed of a six-stranded β -sheet, five α -helices, and ten connecting loops (Nicely et al., 2004). Residues 40-50 of loop 2 comprise switch I, and residues 70-83 of loop 4 and part of helix 2 comprise switch II (Nicely et al., 2004). RalA has a unique switch I sequence, YEPTKAD, and a unique set of effectors. In addition to a unique switch I region, the structure of RalA revealed a clustering of conserved residues on the surface of the protein, which identifies two potential sites for protein-protein interaction. These sites are adjacent to or modified by the switches (Nicely et al., 2004). When compared to the structures of Ras and Rap, another Ras family member, there are structural differences in these sites, particularly in shape and charge distribution, suggesting these sites contribute to the specificity of binding between each of these GTPases and their respective binding partners (Nicely et al., 2004). The first structure of RalA-effector complex was that of RalA-Sec5 (Fukai et al., 2003). Even though the structure of Sec5 is different from the standard ubiquitin-like fold of the Ras-binding domain, it formed an intermolecular β -sheet through switch I, a binding motif common to Ras-effector complexes (Fukai et al., 2003; Mott et al.,

2003). A second RalA-effector complex was determined of RalA-Exo84 (Jin et al., 2005). The Exo84 Ral-binding domain adopts a different fold from those of Ras effectors and Sec5, and unlike the RalA-Sec5 complex, Exo84 interacts with RalA through both switch regions (Jin et al., 2005).

Almost 10 years after the discovery of the RalA protein, the first effector of Ral was identified by three groups concurrently (Cantor et al., 1995; Jullien-Flores et al., 1995; Park and Weinberg, 1995). Numerous interacting proteins and functions have been identified for RalBP (for a summary, see Chapter 4), but there is currently no structure available for RalBP. At 655 amino acids long, RalBP is a large multidomain protein. It was originally identified as having a central GAP domain, a basic α -helix in the amino-terminal portion, and an acidic α -helix and coiled-coil in the carboxy-terminal region (Cantor et al., 1995; Jullien-Flores et al., 1995). The Ral-binding domain was predicted to overlap with the acidic α -helix and part of the coiled coil, either from residue 391-444 (Park and Weinberg, 1995) or from residue 403-499 (Jullien-Flores et al., 1995). Since then, two ATP-binding sites have been identified, one in the amino terminal region, and one in the carboxy terminal region (Awasthi et al., 2001). With the structure of RalBP predicted as being highly helical, particularly in the Ral-binding domain, the structure of the complex is likely to present a previously unobserved binding mode between a GTPase and its effector.

The Multiple Solvent Crystal Structures Method

The Multiple Solvent Crystal Structures (MSCS) Method uses small organic molecules to experimentally probe the surface of a crystalline protein. This method began with the idea that any organic inhibitor molecule can be dissected into functional groups, which can be mimicked by small organic molecules (Ringe, 1995). To perform an MSCS experiment, a protein crystal is crosslinked, if necessary to prevent dissolution, and transferred into a solution of a high concentration of organic solvent. In order to allow the organic solvent probes to compete with water, instead of other solvent molecules, only one solvent is used at a time (Allen et al., 1996). By the end of the process, the solvent has been allowed to fill the interstices between the protein molecules (Ringe, 1995). However, early studies revealed that organic solvent probes are not observed covering the surface of the protein (Allen et al., 1996). Instead, they displace water molecules in a few well-defined sites, which correlated to the active or binding sites on the protein surface (Ringe and Mattos, 1999).

MSCS of thermolysin in increasing concentrations of isopropanol, from 2% to 100%, revealed that the binding of organic solvents was additive. While the isopropanol molecules were found to occupy all four of the main subsites of the active site, 75% of the subsites were occupied only at high concentrations of solvent (English et al., 1999). These results and those of the MSCS of thermolysin crystals soaked in acetone, acetonitrile, and phenol were similar, with only minor changes observed in the conformation of the protein, and probe molecules clustering in the main specificity pocket of the active site, in positions consistent with known protein-ligand complexes (English et al., 1999; English et al, 2001).

Additionally, when compared to computational results from MCSS (Caflisch et al., 1993; Miranker and Karplus, 1991) and GRID (Goodford, 1985; Wade and Goodford, 1993), it was found that MSCS identified the same interactions, however, far fewer binding sites were observed experimentally by MSCS (English et al. 2001). At that time, neither GRID nor MCSS allowed for the flexibility of the protein surface or took into account competing water molecules (Ringe, 1995). A more recent computational method, CS-Map (Computational Solvent Mapping of Proteins) takes desolvation into account and when compared to GRID and MCSS produces similar results, but with far fewer predicted hot spots (Kortvelyesi, 2003). The results of CS-Map were similar to those observed with MSCS and not only was the consensus site of solvent molecule probes always located in the active site, but the probes in this location formed hydrogen bonds and non bonded interactions with residues in the same way a bound ligand would interact (Kortvelyesi et al., 2003; Silberstein et al, 2003).

Elastase was the first protein extensively mapped by MSCS (Allen et al., 1996; Mattos et al., 2006). Similar to the results with thermolysin, the organic solvent probe molecules cluster in the pockets of the active site, and their positions correspond with known inhibitors (Mattos et al., 2006). In addition to probing hot spots, the elastase study revealed areas of plasticity in the protein, particularly side chains that make subtle adjustments in response to the bound organic solvent probes (Mattos et al, 2006). Finally, the ensemble of structures allowed for the nearly complete first hydration shell to be visualized with over 400 unique water binding sites. When superimposed with an inhibitor, the pattern of crystallographic water molecules in the active site could trace the location of the inhibitor binding sites (Mattos et al., 2006).

MSCS is a fast and flexible method for mapping the surface of crystalline macromolecules generating high-resolution structures of 2Å or better (Ringe and Mattos, 1999). The strength of this method lies in the collective analysis of the ensemble of protein structures that are produced. In addition to providing information about organic solvent binding sites and interactions with hot spot residues, the multiple structures provide a more complete picture of the solvation and plasticity patterns on the surface of the protein than any single crystal structure could. Unlike the computational methods, MSCS easily allows for the observation of the effects of protein plasticity and competing water molecules. While this method is powerful and generates a lot of data, it is limited to proteins with crystals that diffract to high resolution and that can withstand the transfer into organic solvents.

References

- Allen, K.N., Bellamacina, C.R., Ding, X., Jeffery, C.J., Mattos, C., Petsko, G.A., and Ringe, D. (1996). An experimental approach to mapping the binding surfaces of crystalline proteins. *Journal of Physical Chemistry* *100*, 2605-2611.
- Awasthi, S., Cheng, J.Z., Singhal, S.S., Pandya, U., Sharma, R., Singh, S.V., Zimniak, P., and Awasthi, Y.C. (2001). Functional reassembly of ATP-dependent xenobiotic transport by the N- and C-terminal domains of RLIP76 and identification of ATP binding sequences. *Biochemistry* *40*, 4159-4168.
- Beach, H., Cole, R., Gill, M.L., and Loria J.P. (2005). Conservation of μ s-ms enzyme motions in the apo- and substrate-mimicked state. *Journal of the American Chemical Society*. *127*, 9167-9176.
- Bogan, A.A., and Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology* *280*, 1-9.

Bone, R., Silen, J.L., and Agard D.A. (1989). Structural plasticity broadens the specificity of an engineered protease. *Nature* 339, 191-195.

Bradford, J.R. and Westhead, D.R. (2003) Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem. *Protein Science* 12, 2099-2103.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science* 13,190-202.

Cafilisch, A., Miranker, A., and Karplus, M. (1993). Multiple Copy Simultaneous Search and construction of ligands in binding sites: Application to inhibitors of HIV-1 Aspartic Proteinase. *Journal of Medicinal Chemistry* 36, 2142-2167.

Carr, R.A.E., Congreve, M., Murray, C.W., Rees, D.C. (2005). Fragment-based lead discovery: leads by design. *Drug Discovery Today* 10, 987-992.

Cantor, S.B., Urano, T., and Feig, L.A. (1995). Identification and characterization of Ral-binding protein 1, a potential downstream target of Ral GTPases. *Mol Cell Biol* 15, 4578-4584.

Chardin, P., and Tavitian, A. (1986). The *ral* gene: a new *ras* related gene isolated by the use of a synthetic probe. *EMBO Journal* 5, 2203-2208.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science* 267, 383-386.

Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. (2005a). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433-3434.

Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. (2005b). The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *Journal of Molecular Biology* 347, 827-839.

Dyson, H.J., and Wright, P.E. (2002). Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology* 12, 54-60.

Dyson, H.J. and Wright, P.E. (2005). Intrinsically Unstructured Proteins and Their Functions. *Nature Reviews* 6, 197-208.

English, A.C., Done, S.H., Caves, L.S.D., Groom, C.R., and Hubbard, R.E. (1999). Locating interaction sites on proteins: The crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins: Structure, Function, and Genetics* 37, 628-640.

English, A.C., Groom, C.R., and Hubbard, R.E. (2001). Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Engineering* 14, 47-59.

Fankuchen, I. (1941). An x-ray and crystallographic study of ribonuclease. *Journal of General Physiology* 24, 315-316.

Fernández, A., and Scott, R. (2003). Dehydron: A Structurally Encoded Signal for Protein Interaction. *Biophysical Journal*. 85, 1914-1928.

Fink, A.L. (2005). Natively unfolded proteins. *Current Opinion in Structural Biology* 15, 35-41.

Fukai, S., Matern, H.T., Jagath, J.R., Scheller, R.H., and Brunger, A.T. (2003). Structural basis of the interaction between RalA and Sec5, a subunit of the sec6/8 complex. *EMBO J* 22, 3267-3278.

Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. (2004). Preformed Structural Elements Feature in Partner Recognition by Intrinsically Unstructured Proteins. *Journal of Molecular Biology*. 338, 1015-1026.

Goodford, P.J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* 28, 849-857.

Hajduk, P.J., Sheppard, G., Nettesheim, D.G., Olejniczak, E.T., Shuker, S.B., Meadows, R.P., Steinman, D.H., Carrera, G.M., Jr., Marcotte, P.A., Severin, J., Walter, K., Smith, H., Gubbins, E., Simmer, R., Holzman, T.F., Morgan, D.W., Davidsen, S.K., Summers, J.B., and Fesik, S.W. (1997). Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR. *Journal of the American Chemical Society* 119, 5818-5827.

Hammes, G.G. (2002). Multiple conformational changes in enzyme catalysis. *Biochemistry* 41, 8221-8228.

Hartshorn, M.J., Murray, C.W., Cleasby, A., Frederickson, M., Tickle, I.J., and Jhoti, H. (2005). Fragment-Based Lead Discovery Using X-ray Crystallography. *Journal of Medicinal Chemistry* 48, 403-413.

- Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradović, Z., and Dunker, A.K. (2002). Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *Journal of Molecular Biology* 323, 573-584.
- Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure* 7, R277-279.
- Jin, R., Junutula, J.R., Matern, H.T., Ervin, K.E., Scheller, R.H., and Brunger, A.T. (2005). Exo84 and Sec5 are competitive regulatory Sec6/8 effectors to the RalA GTPase. *EMBO Journal* 24, 2064-2074.
- Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences USA* 93, 13-20.
- Jones, S., and Thornton, J.M. (1997a). Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology* 272, 121-132.
- Jones, S., and Thornton, J.M. (1997b). Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology* 272, 133-143.
- Jones, W. (1920). The Action of Boiled Pancreas Extract on Yeast Nucleic Acid. *American Journal of Physiology* 52, 203-207.
- Jullien-Flores, V., Dorseuil, O., Romero, F., Letourneur, F., Saragosti, S., Berger, R., Tavitian, A., Gacon, G., and Camonis, J.H. (1995). Bridging Ral GTPase to Rho pathways. RLIP76, a Ral effector with CDC42/Rac GTPase-activating protein activity. *J Biol Chem* 270, 22473-22477.
- Karplus, P.A, and Faerman, C. (1994). Ordered water in macromolecular structure. *Current Opinion in Structural Biology* 4, 770-776.
- Kartha, G., Bello, J., and Harker, D. (1967). Tertiary structure of ribonuclease. *Nature* 213, 862-865.
- Kortvelyesi, T., Dennis, S., Silberstein, M., Brown, L., 3rd, and Vajda, S. (2003). Algorithms for Computational Solvent Mapping of Proteins. *Proteins: Structure, Function, and Genetics* 51, 340-351.
- Koshland, D.E. Jr. (1960). The active site and enzyme action. *Advances in enzymology and related subjects of biochemistry* 22, 45-97.
- Kunitz, M. (1939). Isolation from Beef Pancreas of a Crystalline Protein Possessing Ribonuclease Activity. *Science* 90, 112-113.

- Landon, M.R., Lancia, D.R., Jr., Yu, J., Thiel, S.C., and Vajda, S. (2007). Identification of Hot Spots within Druggable Binding Regions by Computational Solvent Mapping of Proteins. *Journal of Medicinal Chemistry* 50, 1231-1240.
- Lo Conte, L., Chothia, C., Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology* 285, 2177-2198.
- Lu, Y., Wang, R., Yang, C.-Y., and Wang, S. (2007). Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling* 47, 668-675.
- Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research* 29, 2860-2874.
- Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences USA* 100,5772-5777.
- Mattos, C., and Ringe, D. (2001). *International Tables for Crystallography*, edited by M.G. Rossmann and E. Arnold pp. 623-647. Dordrecht: Kluwer Academic Publishers.
- Mattos, C., Bellamacina, C.R., Peisach, E., Pereira, A., Vitkup, D., Petsko, G.A., and Ringe, D. (2006). Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. *Journal of Molecular Biology* 357, 1471-1482.
- Merlino, A., Vitagliano, L., Ceruso, M.A., Di Nola, A., and Mazzarella L. (2002). Global and local motions in ribonuclease A: A molecular dynamics study. *Biopolymers* 65, 274-283.
- Miranker, A., and Karplus, M. (1991). Functionality maps of binding sites: A Multiple Copy Simultaneous Search Method. *Proteins: Structure, Function, and Genetics* 11, 29-34.
- Najmanovich, R., Kuttner, J., Sobolev, V., and Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins: Structure, Function, and Genetics* 39, 261-268.
- Nicely, N.I., Kosak, J., de Serrano, V., and Mattos, C. (2004). Crystal Structures of Ral-GppNHp and Ral-GDP Reveal Two Binding Sites that Are Also Present in Ras and Rap. *Structure* 12, 2025-2036.
- Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., and Dunker, A.K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins* 53, 566-572.

- Papouian, G.A., Ulander, J., and Wolynes, P.G. (2003). Role of Water Mediated Interactions in Protein-Protein Recognition Landscapes. *Journal of the American Chemical Society* *125*, 9170-9178.
- Park, S.H., and Weinberg, R.A. (1995). A putative effector of Ral has homology to Rho/Rac GTPase activating proteins. *Oncogene* *11*, 2349-2355.
- Powers, R.A., and Shoichet, B.K. (2002). Structure-based approach for binding site identification on AmpC beta-lactamase. *Journal of Medicinal Chemistry* *45*, 3222-3234.
- Radha Kishan, K.V., Chandra, N.R., Sudarsanakumar, C., Suguna, K., and Vijayan, M. (1995). Water-dependent domain motion and flexibility in ribonuclease A and the invariant features in its hydration shell. An x-ray study of two low-humidity crystal forms of the enzyme. *Acta Crystallographica Section D* *51*, 703-710.
- Rajamani, D., Thiel, S., Vajda, S., and Camacho C.J. (2004). Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences* *101*, 11287-11292.
- Raines, R.T. (1998). Ribonuclease A. *Chemical Reviews* *98*, 1045-1065.
- Rees, D.C., Congreve, M., Murray, C.W., and Carr, R. (2004). Fragment-based lead discovery. *Nature Reviews Drug Discovery* *3*, 660-672.
- Richards, F.M., and Wyckoff, H.W. (1971). *The Enzymes*, edited by P. Boyer pp. 647-806. New York: Academic Press.
- Ringe, D. (1995). What makes a binding site a binding site? *Current Opinion in Structural Biology* *5*, 825-829.
- Ringe, D., and Mattos, C. (1999). Analysis of the binding surfaces of proteins. *Medicinal Research Reviews* *19*, 321-331.
- Rodier, F., Bahadur, R.P., Chakrabarti, P., and Janin, J. (2005). Hydration of protein-protein interfaces. *Proteins* *60*, 36-45.
- Sadasivan, C., Nagendra, H.G., and Vijayan, M. (1998). Plasticity, Hydration and Accessibility in Ribonuclease A. The Structure of a New Crystal Form and its Low-Humidity Variant. *Acta Crystallographica D54*, 1343-1352.
- Sheu, S.-H., Kaya, T., Waxman, D.J., and Vajda, S. (2005). Exploring the Binding Site Structure of the PPAR γ Ligand-Binding Domain by Computational Solvent Mapping. *Biochemistry* *44*, 1193-1209.

Shuker, S.B., Hajduk, P.J., Meadows, R.P., and Fesik, S.W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274, 1531-1534.

Silberstein, M., Dennis, S., Brown, L., 3rd, Kortvelyesi, T., Clodfelter, K., and Vajda, S. (2003). Identification of Substrate Binding Sites in Enzymes by Computational Solvent Mapping. *Journal of Molecular Biology* 332, 1095-1113.

Smith, G.R., Sternberg, M.J.E., and Bates, P.A. (2005). The Relationship between the Flexibility of Proteins and their Conformational States on Forming Protein-Protein Complexes with an Application to Protein-Protein Docking. *Journal of Molecular Biology* 347, 1077-1101.

Smyth, D.G., Stein, W.H., and Moore, S. (1963). The sequence of amino acid residues in bovine pancreatic ribonuclease: revisions and confirmations. *Journal of Biological Chemistry* 238, 227-234.

Teyra, J. and Pisabarro, M.T. (2007). Characterization of Interfacial Solvent in Protein Complexes and Contribution of Wet Spots to the Interface Description. *Proteins: Structure, Function, and Bioinformatics* 67, 1087-1095.

Uversky, V.N. (2002a). What does it mean to be natively unfolded? *European Journal of Biochemistry* 269, 2-12.

Uversky, V.N. (2002b). Natively unfolded proteins: A point where biology waits for physics. *Protein Science* 11, 739-756.

Verkhivker, G.M., Bouzida, D., Gehlhaar, D.K., Rejto, P.A., Freer, S.T., and Rose, P.W. (2003). Stimulating disorder-order transitions in molecular recognition of unstructured proteins: Where folding meets binding. *Proceedings of the National Academy of Sciences* 100, 5148-5153.

Vetter, I.R., and Wittinghofer, A. (2001). The guanine nucleotide-binding switch in three dimensions. *Science* 294, 1299-1304.

Vitagliano, L., Merlino, A., Zagari, A., and Mazzarella, L. (2002). Reversible substrated-induced domain motions in Ribonuclease A. *Proteins: Structure, Function, and Genetics*. 46, 97-104

Wade, R.C., and Goodford, P.J. (1993). Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *Journal of Medicinal Chemistry* 36, 148-156.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology* 337, 635–645.

Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology* 293, 321-331.

Wyckoff, H.W., Hardman, K.D., Allewell, N.M., Inagami, T., Johnson, L.N., and Richards, F.M. (1967). The structure of ribonuclease-s at 3.5 Å resolution. *Journal of Biological Chemistry* 242, 3984-3988.

Yao, H., Kristensen, D.M., Mihalek, I., Sowa, M.E., Shaw, C., Kimmel, M., Kaviraki, L., and Lichtarge, O. (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *Journal of Molecular Biology* 326, 255-261.

CHAPTER 2: Multiple Solvent Crystal Structures of Ribonuclease A: A Critical Assessment of the Method

Abstract

The Multiple Solvent Crystal Structures (MSCS) uses organic solvents to map the surfaces of proteins. In addition to identifying binding sites, this method allows for a more thorough examination of protein plasticity and hydration than could be achieved by a single crystal structure. The crystal structures of Bovine Pancreatic Ribonuclease A (RNase A) soaked in a variety of solvents were solved (50% dioxane, 50% dimethylformamide, 70% dimethylsulfoxide, 70% 1,6-hexanediol, 70% 2-propanol, 50% R,S,R-bisfuran alcohol, 70% t-butanol, 50% trifluoroethanol, or 1.0 M trimethylamine-N-oxide). As RNase A is a well-studied protein, it serves as an excellent model for the development of quantitative tools for MSCS data analysis. The MSCS structures of RNase A reveal patterns of plasticity and conserved water binding sites consistent with crystal and NMR structures available in the Protein Data Bank (PDB). Additionally, comparison of the MSCS structures with inhibitor-bound structures of RNase A from the PDB reveals that the organic solvent molecules bound in the active site pick out key interactions made by bound inhibitor molecules and highlight ligand binding hot spots in the active site of RNase A.

Introduction

The Multiple Solvent Crystal Structures (MSCS) method was first published in 1996 as an experimental method for locating and characterizing protein binding sites for potential use in ligand design (Mattos and Ringe, 1996). The crystal structures of Subtilisin (Fitzpatrick et

al., 1994) and Elastase (Allen et al., 1996) in neat acetonitrile had unexpectedly revealed that the protein was far from solvated by its organic solvent host, but rather retained a substantial first hydration shell with only a few molecules of the solvent bound primarily in the active site. This observation led to the idea that by soaking a protein in several distinct organic solvents and superimposing the resulting structures, the bound organic solvent molecules would reveal the shape and chemical complementarity in the active site that could then be used to develop functional groups in a larger ligand. The idea of linking fragments to produce more potent ligands has its origins in earlier pioneering work (Jencks, 1981) and was at the time also being explored by NMR, not with organic solvents, but with small solutes in aqueous solution (Shuker et al., 1996). The approach of screening large numbers of small solute molecules has contributed with some degree of success to pharmaceutical research, combining both X-ray crystallography and NMR spectroscopy in fragment-based lead discovery and drug design (Rees et al., 2004).

The MSCS method is currently being developed as a powerful tool to experimentally locate and determine fundamental properties of protein binding sites (Mattos, 2002; Mattos et al., 2006). It has also played a role in the development of computational solvent mapping (CS-Map), an *in silico* counterpart to MSCS used to locate hot spots for protein-ligand interactions (Dennis et al., 2002; Sheu et al., 2005a; Silberstein et al., 2003; Silberstein et al., 2006). Results from other computational methods, such as Multiple Copy Simultaneous Search (MCSS) that aim to place chemical functional groups in protein active sites, have been compared to those of experimental MSCS data sets (English et al., 2001; Miranker and

Karplus, 1991; Stultz and Karplus, 2000). Thus, MSCS has played an important role in the synergy that exists between experimental and computational strategies that aim to predict and understand protein-binding sites.

MSCS has been shown to correctly locate the known substrate binding sites for both Elastase (Ringe and Mattos, 1999) and Thermolysin (English et al., 1999; English et al., 2001).

Locating binding sites in proteins for which structures exist resulting from the various ongoing genomics projects has become one of the great challenges facing this new era of structural biology. This problem has preoccupied many structure prediction and protein bioinformatics groups, resulting in a variety of computational methods for binding site prediction and characterization, with various degrees of success (Glaser et al., 2006; Jones and Thornton, 2004; Laurie and Jackson, 2006; Sheu et al., 2005b). MSCS is unique however, not only because it is an experimental method, but also because in addition to locating binding sites with the organic solvent probes it provides a picture of hydration and plasticity on the entire surface of the protein (Mattos et al., 2006; Mattos and Ringe, 2001). To date, however, the patterns of clustering of organic solvent molecules, plasticity and hydration have been analyzed only qualitatively by MSCS. In the present article we use the very well studied protein Bovine Pancreatic Ribonuclease A (RNase A) in a detailed MSCS analysis that can be compared with information accumulated over 40 years of structural biology research on this protein in order to establish more quantitatively the extent to which physical and chemical properties of protein surfaces are captured by MSCS. What aspects of ligand binding does MSCS predict? Do the trends and range in plasticity obtained from

MSCS reflect those obtained from the over 100 RNase A structures in the Protein Data Bank (PDB) that have been solved in aqueous solution from various crystalline environments and bound to countless inhibitors? Even more interesting, do the plasticity trends reflect the qualitative features obtained by solution NMR? Are water molecules conserved in the MSCS set the same as those observed to be conserved in the large number of structures solved in aqueous solution? In short, the goal of this study is to provide an assessment of how well the MSCS set reflects the properties of RNase A in aqueous solution, delineating the strengths and limitations inherent in the method. This is an important step in defining how the MSCS method can be used to understand the details of surfaces in general and binding sites in particular.

In addition to providing an excellent model for our current work on MSCS, RNase A has received significant attention as a model for members of the Ribonuclease family to which it belongs, many of which have been shown to possess potent physiological activities (D'Alessio, 1993). Examples of therapeutically important RNase A homologues include angiogenin, eosinophil-derived neurotoxin, and eosinophil cationic protein (Leonidas et al., 2003; Leonidas et al., 2006; Leonidas et al., 1999). These studies add to a wealth of structural information available in the PDB of RNase A bound to inhibitors.

RNase A catalyzes the breakdown of 3'5'-phosphodiester linkages in single stranded RNA at the 3' side of pyrimidine nucleotides. It uses an acid-base mechanism in which His 12 and His 119 play a central role. A review of the structure, function and catalytic mechanism of

RNAse A provides an excellent summary for decades of work on this enzyme (Raines, 1998). The plasticity and hydration features obtained by MSCS are compared in the present work with those published in a set of eight structures which encompasses two different crystal forms (Sadasivan et al., 1998a), and a partially overlapping set of nine structures representing four crystal forms (Zegers et al., 1994). None of the crystal forms included in these two crystallographic studies of RNAse A include the form which we used for the MSCS work (Vitagliano et al., 2000). In addition the plasticity captured by MSCS is compared to that observed in the NMR structure consisting of 32 models deposited in the PDB (Santoro et al., 1993). For the assessment of how well the organic solvent probes reveal functional group binding sites for RNAse A we use 58 inhibitor-bound structures taken from the PDB. The published RNAse A structures in aqueous solution provide a diverse and robust set for assessment and validation of the MSCS results in terms of plasticity, hydration and identification of hot spots on protein surfaces.

Materials and Methods

Crystal Growth, Cross-linking, and Solvent Soaks

A procedure similar to that described previously (Vitagliano et al., 2000) was used to remove the sulfate from the RNAse A active site. Dialysis of the protein into buffers of increasing pH results in deprotonation of the two catalytic His residues, releasing the negatively charged sulfate molecule. Bovine Pancreatic Ribonuclease A (type XII A) was purchased from Sigma and dissolved in de-ionized water for a final concentration of 30 mg/mL. Protein was then dialyzed against a 0.15 M Tris-HCl buffer at pH 7.2 at 4° C for 24 hours. The dialysis was

repeated with a 0.15M Tris-HCl buffer at pH 8.2, and subsequently with a 0.15 M Tris-HCl buffer at pH 9.0, each for 24 hours. RNase A was then transferred to fresh buffer at pH 9.0 and the pH was adjusted to 5.5 over 8 hours by adding small aliquots of 1 M HCl. In the final step, RNase A protein was dialyzed over night into a 10mM sodium citrate buffer pH 5.0 and concentrated to 12 mg/mL. As in previous reports (Leonidas et al., 1997; Vitagliano et al., 2000), RNase A crystals were grown using hanging drop vapor diffusion at 18 °C with 10 μ L drops containing half protein solution and half reservoir solution over a 500 μ L reservoir. The reservoir solution contained 20 mM sodium citrate buffer at pH 5.0 and PEG4000 levels of 30% w/v, 32% w/v, or 35% w/v. All crystallization drops resulted in high diffraction quality crystals after 5 months.

RNase A crystals were manually transferred with a cryo-loop to a 10 μ L drop of 0.08% glutaraldehyde (8.26 % w/v in distilled water, pH 4.25, Electron Microscopy Sciences) in stabilization buffer (0.15M HEPES, pH 7.5, 32% w/v PEG 4000) over a 300 μ L reservoir (the same glutaraldehyde-stabilization buffer as the drop) and the cross-linking reaction was allowed to proceed at room temperature for 30 minutes. Cross-linked crystals were then transferred with a cryo-loop to new drops containing stabilization buffer and an organic solvent and allowed to soak for 1 to 2 hours at room temperature. Soaked crystals were then collected, cryo-protected by dunking in stabilization buffer containing 20% glycerol, and flash frozen in liquid nitrogen.

Data Collection, Processing, Structure Refinement

Diffraction data were collected at 100K at the SER-CAT ID-22 beamline at APS (Argonne, IL) using 1.0 Å wavelength radiation and a Mar300 CCD detector at a crystal to detector distance of 100 mm. The data were processed and scaled using HKL2000 (Otwinowski and Minor, 1997). A published model of RNase A (PDB code 1JVT) with the water molecules removed was used to calculate the initial electron density maps for all models of RNase A. The models were refined with CNS (Brunger et al., 1998). The program Coot (Emsley and Cowtan, 2004) was used to manually rebuild the models and to identify water and organic solvent molecules using the Fo-Fc electron density maps contoured at 3σ and the 2Fo-Fc electron density difference maps contoured at 1σ . Data collection, refinement statistics, and PDB codes are shown in Table 1. The RNase A crystals have symmetry of space group C2 and have two protein molecules in the asymmetric unit. These molecules are designated as chain A and chain B in the coordinates deposited in the PDB and will be referred to as protein chain or molecule A and B.

Water Renumbering

The water molecules in the MSCS models were renumbered for consistency. Due to the presence of two molecules in the asymmetric unit, each of the ten crystallographic coordinate files was divided into two files: one for each RNase A molecule and the associated water and organic solvent molecules within 5Å of the protein. The resulting monomers were then all superimposed using least squares superposition of the entire protein main chain atoms. Molecule A from the cross-linked RNase A structure in aqueous solution was used as the

Table 1. Data Collection and Refinement Statistics for Apo-RNase A.

Solvent	Crosslinked	Dioxane	Dimethyl- formamide	Dimethyl- sulfoxide	1,6- Hexanediol	Isopropanol	R,S,R- Bisfuranol	t-Butanol	Trifluoro- ethanol	Trimethyl- amine N- oxide
% volume		50%	50%	70%	70%	70%	50%	70%	50%	1M
Space Group	C2	C2	C2	C2	C2	C2	C2	C2	C2	C2
Unit Cell	a = 100.68 Å b = 32.69 Å c = 72.59 Å α = 90.00° β = 90.99° γ = 90.00°	a = 100.74 Å b = 32.82 Å c = 72.69 Å α = 90.00° β = 90.72° γ = 90.00°	a = 100.68 Å b = 32.79 Å c = 72.61 Å α = 90.00° β = 90.44° γ = 90.00°	a = 100.97 Å b = 32.64 Å c = 72.91 Å α = 90.00° β = 90.66° γ = 90.00°	a = 100.60 Å b = 32.71 Å c = 73.41 Å α = 90.00° β = 90.60° γ = 90.00°	a = 99.87 Å b = 32.77 Å c = 72.09 Å α = 90.00° β = 90.73° γ = 90.00°	a = 100.63 Å b = 32.96 Å c = 72.70 Å α = 90.00° β = 90.64° γ = 90.00°	a = 100.83 Å b = 32.71 Å c = 72.68 Å α = 90.00° β = 90.64° γ = 90.00°	a = 100.56 Å b = 32.83 Å c = 72.52 Å α = 90.00° β = 90.59° γ = 90.00°	a = 100.04 Å b = 32.77 Å c = 72.81 Å α = 90.00° β = 90.40° γ = 90.00°
Temperature of data collection (K)	100	100	100	100	100	100	100	100	100	100
Resolution (Å)	1.65	1.95	1.84	1.76	2.00	2.02	1.76	1.68	1.93	1.68
# of Reflections	23665	15668	20712	22635	15129	16257	23668	26786	17788	26241
Redundancy	2.6 (1.2)	3.4 (3.1)	3.4 (2.9)	3.2 (2.9)	2.7 (1.7)	2.6 (2.1)	3.3 (3.0)	2.7 (1.9)	3.4 (3.4)	2.9 (2.4)
R sym (%)	10.2 (59.5)	6.8 (40.5)	4.1 (20.2)	6.9 (41.1)	12.3 (77.8)	7.4 (37.2)	6.0 (45.6)	13.1 (26.4)	7.6 (25.2)	5.9 (32.1)
completeness (%)	81.7 (33.8)	88.5 (93.9)	88.8 (95.5)	96.9 (94.0)	91.3 (65.9)	51.5 (80.7)	97.1 (96.2)	97.2 (93.4)	98.7 (98.0)	97.1 (92.4)
average I/σ	9.2 (7.0)	29.3 (3.8)	58.9 (7.4)	42.1 (4.1)	11.4 (1.6)	19.9 (2.5)	38.0 (3.1)	28.7 (4.14)	31.2 (6.9)	53.9 (4.2)
Rwork/Rfree (%)	21.3/24.8	20.1/24.0	19.2/22.4	20.9/23.6	23.4/29.3	19.4/24.3	21.3/24.8	20.7/23.0	19.0/22.0	22.6/25.8
rmsd from ideal geometry for bond lengths (Å)	0.005	0.005	0.005	0.005	0.006	0.005	0.005	0.005	0.005	0.01
rmsd from ideal geometry for bond angles (°)	1.2	1.2	1.2	1.3	1.3	1.2	1.2	1.3	1.2	1.5
# protein atoms	1892	1864	1848	1896	1892	1879	1889	1857	1868	1896
# water molecules	263	183	242	197	167	169	233	320	222	208
# organic solvent molecules	0	5	1	10	4	4	6	1	6	2
orientation of His 119 in molecule A	A & B	A & B	A & B	A	A	A	A	A & B	A	A
orientation of His 119 in molecule B	A & B	A	A	A	A	A	A	A & B	A	A
PDB ID	3EUX	3EUY	3EUZ	3EV0	3EV1	3EV2	3EV6	3EV3	3EV4	3EV5

reference structure. For the analysis of surface hydration it is important for water molecules at a particular site in any of the MSCS structures to have the same residue number in all the coordinate files. For this purpose, a unique set of water positions was compiled from all of the MSCS structures to create a consensus list with assigned numbers, which was then used as the template to renumber the water molecules in each structure. This consensus list was generated by first including all water molecules located within 5 Å of molecule A of the cross-linked structure in aqueous solution as unique water binding positions. Next, all water molecules located within 5 Å of molecule B of the same structure were compared with the consensus list of unique water binding sites. Water molecules from this structure that were found within 1.4 Å of a water position on the consensus list were not added; however, the water position from the consensus list was averaged with the position of the comparison water molecule resulting in an average water position in the consensus list. Water molecules associated with chain B that were not located within 1.4 Å of a water molecule on the consensus list were added to the list as unique water binding sites. This process was repeated until the consensus list was populated with average unique water binding positions from molecules A and B of all 10 MSCS structures. For the final renumbering, water molecules associated with molecules A and B from the MSCS structures were each assigned the number of the closest average position in the consensus list. These coordinates were used for the analysis presented in this paper.

For final PDB deposition molecules A and B in the asymmetric unit of each of the structures were merged back into a single coordinate file using the least squares superposition in Coot.

Water molecules associated with protein chain A were included in chain C and water molecules associated with protein chain B were included in chain D. A number of water molecules had been duplicated because they were located at the interface between protein chains A and B, within 5 Å of both molecules in the asymmetric unit. These water molecules were removed from chain D and included only in chain C. The final coordinates deposited in the PDB are those resulting from one last round of refinement with CNS after merging of the files with the renumbered water molecules.

Comparison of MSCS with RNase A Structures Solved in Aqueous Solution

The SAS tool (Milburn et al., 1998) was used via the European Bioinformatics Institute website to probe the Protein Data Bank (PDB) (Berman et al., 2000) for models of RNase A with 100% sequence identity to the published model with PDB code 1JVT. Five sets of RNase A models were downloaded from the PDB for comparison with the RNase A MSCS structures. All represent structures solved in aqueous solution and serve to assess whether results from the MSCS in organic solvents correspond to those observed from structures solved in an aqueous environment. (Details about downloaded files can be found in Appendix B.) These files were grouped into the following sets for analysis:

- 1) RNase A in complex with inhibitor molecules solved from crystals with symmetry of the C2 space group (referred to as the C2 inhibitor set throughout this chapter). The structures in this set were obtained from crystals isomorphous with the crystals used to obtain the MSCS set. It consists of 18 models with the following PDB codes:

1AFK, 1AFL, 1EOS, 1JN4, 1JVU, 1O0F, 1O0H, 1O0M, 1O0N, 1O0O, 1QHC, 1W4O, 1W4P, 1W4Q, 1WBU, 1Z6D, 1Z6S, and 2G8R.

- 2) RNase A in complex with inhibitor molecules solved from crystals with symmetry of the following six space groups $C2$, $P2_1$, $P2_12_12_1$, $P3_121$, $P3_221$, $P4_12_12$ (referred to as the inhibitor set throughout this chapter). This set of 37 structures includes the 18 models in the $C2$ inhibitor set and have the following PDB codes: 1AFK, 1AFL, 1EOS, 1EOW, 1JN4, 1JVU, 1O0F, 1O0H, 1O0M, 1O0N, 1O0O, 1QHC, 1RAR, 1RAS, 1RBJ, 1RCA, 1RCN, 1RNC, 1RND, 1RNM, 1RNN, 1ROB, 1RPF, 1RPG, 1RSM, 1RTA, 1RUV, 1U1B, 1W4O, 1W4P, 1W4Q, 1WBU, 1Z6D, 1Z6S, 2G8R, 8RSA, and 9RSA.
- 3) Models included in the Sadasivan study (Sadasivan et al., 1998a) consist of eight RNase A structures, all of which were obtained from crystals with symmetry of space group $P2_1$, with one exception (PDB code 1RPH), for which crystals were obtained with $P3_221$ symmetry. The PDB codes for structures in this set are: 7RSA, 5RSA, 3RN3, 1RHB, 1RHA, 1RPH, 1XPT, and 1XPS. This study presents a detailed analysis of plasticity in RNase A.
- 4) Models included in the Zegers study (Zegers et al., 1994) consist of nine RNase A structures from crystals with symmetry of the four space groups $P2_1$, $P2_12_12_1$, $P3_121$, and $P3_221$. The PDB codes for this set are: 1RPG, 1ROB, 7RSA, 1RPH, 1RPF, 4SRN, 1RSM, 8RSA, and 2RNS. The Zegers set has two structures in common with the Sadasivan set: 1RPH and 7RSA. An extensive crystallographic water analysis was done in this study.

- 5) Solution NMR structures of RNase A (Santoro et al., 1993) contain a total of 32 models and were deposited in the PDB with code 2AAS. This file was separated into 32 individual coordinate files – one for each model.

The models taken from the PDB and the MSCS models were superimposed using least squares backbone superposition of the entire protein chain using the program Coot. Molecule A from the cross-linked RNase A in aqueous solution was used as the reference structure. In the cases where there were two protein molecules in the asymmetric unit, the PDB files were divided into two as described for the MSCS set: one for each protein monomer and the associated water, inhibitor, and organic solvent molecules within 5 Å of the protein. Each model was then superimposed with the rest of the ensemble.

Computational Analysis

Quantitative structural analysis was performed on the superimposed models of the structures downloaded from the PDB and on the MSCS structures to explore the parameters of (i) plasticity, (ii) conserved water molecules and (iii) superposition between organic solvents and inhibitor functional groups. A series of calculations and corresponding analysis was done independently on the MSCS and on the five sets of structures described in the previous section: the C2-inhibitor set, the all-inhibitor set, the Sadasivan set, the Zegers set, and the NMR structures.

(i) Plasticity: A script was written in Perl to calculate the RMSD per residue, between each pair of structures. The RMSD was calculated as

$$\text{RMSD}^2 = \{\sum_{i=1 \text{ to } n} [(x_{2,i} - x_{1,i})^2 + (y_{2,i} - y_{1,i})^2 + (z_{2,i} - z_{1,i})^2]\}/n$$

where i is the atom type in each residue and n is the number of atoms in each residue. This calculation was performed for all atoms in the protein and also for just the backbone atoms (N, CA, C, and O). The highest RMSD, lowest RMSD, and average of all calculated pairwise RMSD values were then plotted for each residue. This was done for the MSCS set as well as for the five sets of structures in aqueous solution outlined in the previous section.

(ii) Conserved water molecules: The Structurally Equivalent Water System (SEWS) program (Bottoms et al., 2006) was used to identify conserved water positions among the sets of structures. Water molecules were identified as belonging to the same cluster if their positions were within 1.4 Å of a calculated peak of atom density. Conserved water positions were identified as clusters of water molecules having at least one common interaction with the protein within a distance cut-off of 3.4 Å. A water molecule was considered conserved when present in at least 80% of the structures in the set. The NMR structures were not included in this analysis because these models contain no water molecules.

(iii) Organic solvents and inhibitor functional groups: A script was written in Perl to calculate the distances between the atom positions in the organic solvents and corresponding atom positions in the inhibitors. The distance was calculated as

$$\text{distance}^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2$$

Using a cut-off of 1.0 Å or less with atoms of the same type (e.g. carbon and carbon), areas of overlap were identified between organic solvents in the MSCS structures and inhibitor atoms in the inhibitor-bound structures.

Results

RNAse A is composed of 124 amino acid residues with eight cysteines, all of which are involved in disulfide bonds shown to be critical for structural stability (Klink et al., 2000). The structure is dominated by a β -sheet and there are three short α -helices. The β -sheet folds into two lobes (domains A and B) with the active site cleft situated between them (Figure 1a). Domain A is defined as residues 1-13, 49-79, and 105-124, and Domain B consists of residues 16-46, and 82-101. Residues 14-15, 47-48, 80-81, and 102-104 form the hinge connecting the two lobes (Kishan et al., 1995; Sadasivan et al., 1998a). As an enzyme, RNAse A is an endoribonuclease that functions in the degradation of RNA (EC 3.1.27.5). It catalyzes the breakdown of 3'5'-phosphodiester linkages in single stranded RNA at the 3' side of pyrimidine nucleotides. Cleavage of the P-O5' bond occurs by acid-base catalysis in two steps: a transphosphorylation reaction leading to a 2'3'-cyclicphosphodiester intermediate and the leaving group on the 5' end, followed by a nucleophilic attack by water to regenerate the nucleotide. As shown in Figure 1b, the active site has well defined subsites for the nucleotide bases on either side of the scissile bond and a less well-defined site for a second base on the 3' end (B₁, B₂, and B₃ respectively); phosphate moieties bind in the P₀, P₁, and P₂ subsites, where P₁ is the site of cleavage (Raines, 1998). The phosphate group containing the scissile bond in the P₁ subsite is braced on one side by the catalytic base (His

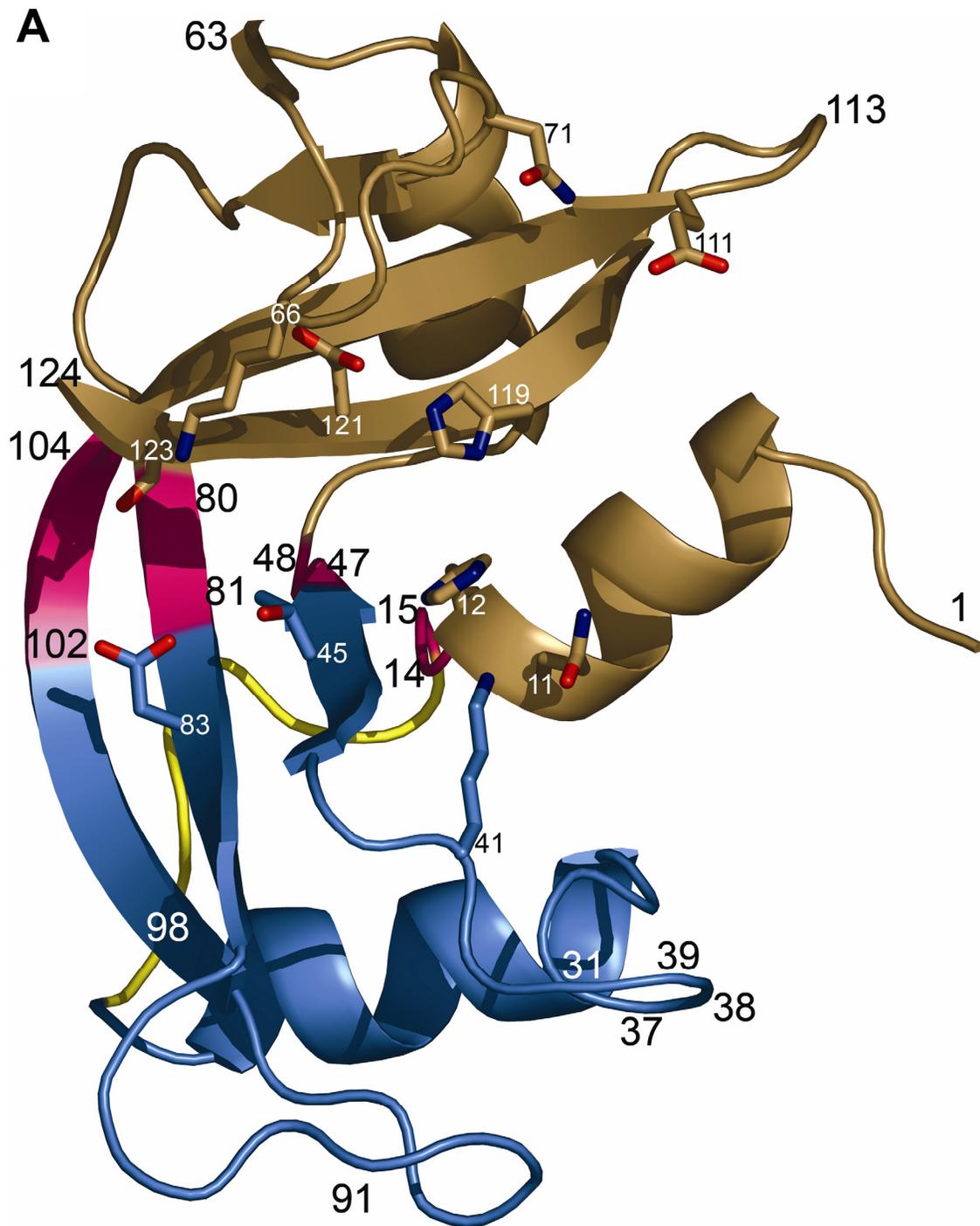


Figure 1. Structure of RNase A.

A. Ribbon diagram of RNase A with Domain A colored in sand, Domain B colored in blue, and the Hinge colored fuchsia. The disordered loop region 16-22 falls in Domain B and is colored yellow. The side chains identified in part A are represented as sticks and the atoms are colored by atom type with oxygen colored red and nitrogen colored blue. Select residue numbers are labeled.

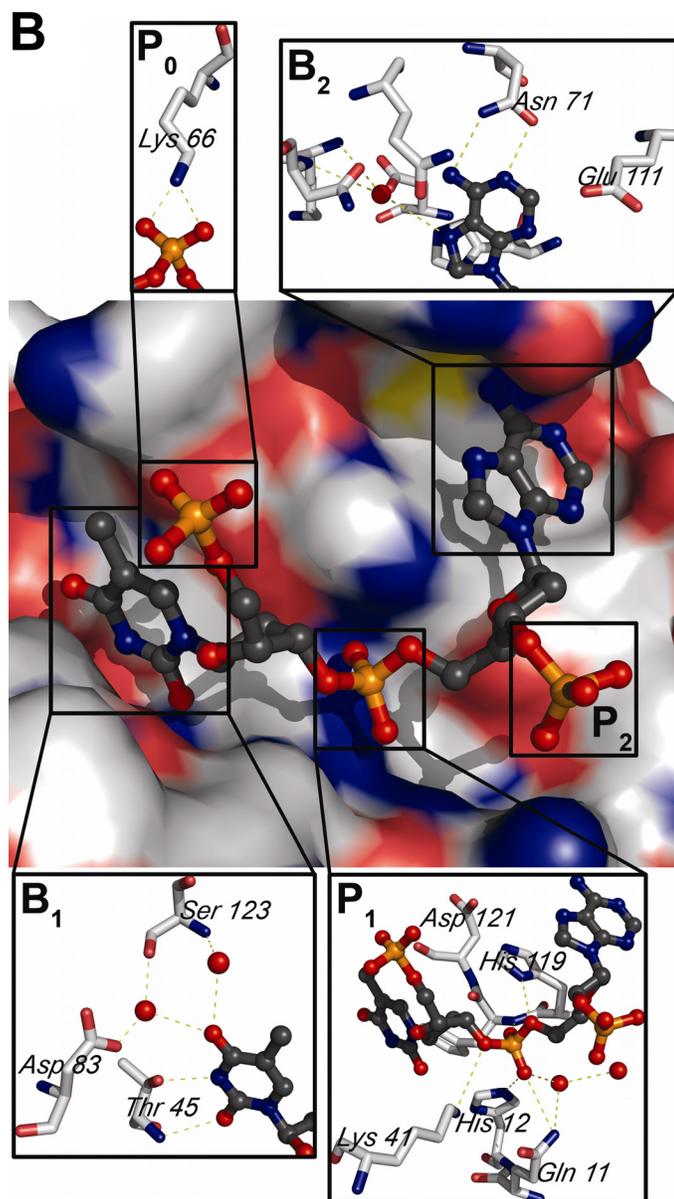


Figure 1. Structure of RNase A (continued).

B. Active Site of RNase A: Surface of RNase A with ball-and-stick representation of dT 2037, dA 2038, and the phosphate from dA 2039 from the deposited coordinates of PDB ID 1RCN (Fontecilla-Camps et al., 1994). Representative conserved water molecules are included in the expanded B₁, P₁, and B₂ pockets. P₀ consists of Lys 66, which interacts with a phosphate moiety of the bound substrate. The pyrimidine bound in B₁ hydrogen bonds with Thr 45 N and O γ 1. Conserved water molecules bridge additional interactions between the pyrimidine and Ser 123 and Asp 83. The catalytic P₁ pocket binds a phosphoryl group through interactions with His 119, His 12, Gln 111, and Lys K41. Asp 121 stabilizes His 119 into its catalytic conformation, commonly referred to as A. Conformation B of His 119 is related to conformation A by rotations of approximately 180° about the C α -C β and C β -C γ bond. The purine bound in B₂ pocket hydrogen bonds with Asn 71, stacks on top of the imidazole of His 119, and interacts with Glu 111. A conserved water molecule bridges additional interactions with the backbone nitrogens of Lys 66 and Asn 67. Figures 1, and 3-7 were prepared using Pymol (DeLano).

12) and on the other by the catalytic acid (His 119). Residues Gln 11, Lys 41 and Asp 121 also play a role in catalysis (Wlodawer et al., 1983). Figure 1 shows the overall arrangement of the active site of RNase A and depicts key interactions between active site residues and bound ligands.

RNase A MSCS Models

RNase A crystals grown in aqueous solution, with symmetry of space group C2, were cross-linked with glutaraldehyde and transferred to one of a variety of solutions containing high concentrations of organic solvents. This resulted in ten structures, each with two molecules in the asymmetric unit. Table 1 shows the data collection, refinement statistics, and PDB codes for the MSCS models. In addition to cross-linked RNase A in aqueous solution (XLINK), the crystal structures were solved in the following conditions: 50% dioxane (DIO), 50% dimethylformamide (DMF), 70% dimethylsulfoxide (DMS), 70% 1,6-hexanediol (HEZ), 70% isopropanol (IPA), 50% R,S,R-bisfuranol (RSF), 70% t-butanol (TBU), 50% trifluoroethanol (ETF), and 1M trimethylamine N-oxide (TMO).

Electron density was poor or missing for the side chains of Lys 1, Lys 31, Lys 37, Asp 38, Arg 39, Lys 91, Lys 98 in both molecules of the asymmetric unit in all ten structures (Figure 1a). These residues are on the surface of RNase A either in loops or at the very ends of secondary structural elements. The side chains for these residues were modeled as common rotamers, using care to avoid steric hindrance with neighboring residues. The one exception

is Lys 1 in the DMF model, which was removed from molecule A due to lack of electron density for the entire residue. Additionally, the region containing residues 16-22, which corresponds to the loop between the first two helices, has poor density, and residues were removed from each structure when there was no density to support their placement. Residues were removed from the models as follows: XLINK and HEZ, 19 and 20 from molecule A; DMS and TMO, 21 from molecule B; DIO, 18-20 from molecule A, 18-21 from molecule B; DMF, 17-20 from molecule A, 18-21 from molecule B; IPA, 17-20 from molecule A; RSF, 17 and 21 from molecule A; TBU, 19-21 from molecule A, 17-21 from molecule B; and ETF, 18-21 from molecule A, 18 and 21 from molecule B.

As observed previously, His 119 is found in one of two conformations, A or B (Borkakoti et al., 1982; Zegers et al., 1994). In conformation A, His 119 interacts with Asp 121 in what is considered the active form of the enzyme. In this conformation, the aromatic ring of the histidine stacks with the purine base of the substrate in the B₂ pocket, while one of its nitrogen atoms donates a proton to the leaving group in the first part of the cleavage reaction. The MSCS models presented here have His 119 in conformation A alone, or show clear electron density for both conformations A and B, depending on the solvent conditions (Table 1). His 119 is found in only the A conformation in the DMS, HEZ, IPA, RSF, ETF, and TMO models. In the XLINK and TBU structures, His 119 adopts both conformations in molecules A and B in the asymmetric unit. For the DIO and DMF structures, the conformation of His 119 is different in each molecule of the asymmetric unit: both

conformations are evident in molecule A, where only conformation A is found in molecule B.

Analysis of Plasticity Based on Pairwise RMSD Values Between the Models

Each MSCS model was divided into two files, which separated the two protein molecules found in the asymmetric unit and their associated water and organic solvent molecules, and resulted in 20 crystallographically independent models of RNase A in ten different solvent environments. All the protein molecules from the MSCS models were then superimposed as described in the Materials and Methods. Previous studies have shown by visual inspection that the comparison of the MSCS models of Elastase reveal subtle changes in side chain conformation due to different solvent environments (Mattos et al., 2006). In RNase A, significant plasticity has been previously observed and analyzed in detail in the Sadasivan study (Sadasivan et al., 1998a), which uses a set of structures solved in aqueous solution (see Materials and Methods section above). In the present study the pairwise RMSD per residue is calculated for all of the structures in the MSCS set as well as for structures in each of the five sets taken from the PDB for comparison. The goal is to quantify plasticity in each set and to assess the correspondence between plasticity results determined from MSCS and those obtained from an aqueous environment. Figure 2a shows a plot of the average pairwise RMSD for main chain atoms in the MSCS set for each of the 124 residues in RNase A. The highest and lowest pairwise RMSD values are also included for each residue to indicate the range of plasticity per residue within the set. In Figures 2b-2f the plots in Figure 2a are superimposed on corresponding plots calculated from the five comparison data sets.

The MSCS set shows peaks and valleys in the RMSD plots that correspond primarily to the loop regions and ordered secondary structures (α -helices and β -strands) respectively. The areas of high average RMSD's in the MSCS set are areas described in the Sadasivan study as having high plasticity and are more prominent in domain B than in A (Sadasivan et al., 1998a). The hinge region has low average backbone RMSD values ranging from 0.12 to 0.29 Å, with the exception of His 48, which has an average backbone RMSD value of 0.57 Å. His 48 lies adjacent to the loop region 16-22, which is disordered in the MSCS structures. This probably destabilizes His 48, allowing for more variation in structure. With the exception of Lys 1 (average RMSD of 1.79 Å), which is highly disordered in virtually all known structures of RNase A, all of the highest RMSD values fall within Domain B. Excluding Lys 1, the maximum average RMSD in Domain A is found for Asn 113 at 0.63 Å. Nine residues in Domain B have average RMSD values higher than that of Asn 113 (Thr 17, 1.17 Å; Ala 19, 0.90 Å; Ala 20, 0.82 Å; Ser 21, 1.6 Å; Lys 37, 0.68 Å; Asp38, 0.86 Å; Gly 88, 0.96 Å; Ser 89, 1.0 Å; and Ser 90, 0.72 Å). Most of these residues are in loop regions with high B-factors in the models (Figure 1a). Overall, the hinge region contains the highest concentration of residues with a low average backbone RMSD, followed by Domain A and Domain B, respectively (Table 2).

In general, the active site residues shown in Figure 1b have relatively low average backbone RMSD values. In the P₁ pocket, Gln 11, His 12, and His 119 all have values at or below 0.2 Å RMSD (0.17 Å, 0.14 Å, and 0.20 Å, respectively). Asp 121, which stabilizes the catalytic

Table 2. Percentage of residues in each domain that fall at or below the baseline of the backbone average RMSD calculations for each set of structures. Domain A consists of residues 1-13, 49-79, and 105-124; domain B consists of residues 16-46, and 82-101; and the hinge consists of residues 14-15, 47-48, 80-81, and 102-104.

	Baseline RMSD	<=Baseline in Domain A	<=Baseline in Domain B	<=Baseline in Hinge
MSCS	0.21Å	48.4%	33.3%	66.7%
Inhibitor	0.35Å	59.4%	31.4%	77.8%
C2 Inhibitor	0.35Å	62.5%	43.1%	77.8%
Sadasivan	0.35Å	68.8%	29.4%	100.0%
Zegers	0.35Å	54.7%	33.3%	77.8%
NMR	0.60Å	59.4%	21.6%	88.9%

conformation of His 119, has a RMSD value of 0.21Å. Lys 41 has the highest value of the residues of the P₁ pocket with an average RMSD of 0.31 Å. In the B₁ pocket, both Thr 45 and Asp 83 have low values of 0.14 Å and 0.19 Å, respectively, while Ser 123 has a higher value of 0.28 Å. Both Asn 71 and Glu 111 of the B₂ pocket have average RMSD values of 0.30 Å and 0.32 Å, respectively, and Lys 66 of P₀ has the highest average RMSD of the active site residues at 0.47 Å. The average backbone RMSDs across the MSCS structures highlights the residues of the active site that remain rigid upon ligand binding and those that adjust to improve the interaction. In general, residues in the P₀ and B₂ pockets further removed from the catalytic center show greater plasticity (or potential for adjustment) than residues in the P₁ and B₁ sites immediately adjacent to the scissile bond.

The trends in plasticity observed in the MSCS set are representative of the trends revealed in all six plots in Figure 2. In all cases, it is clear that domain B shows a significantly higher degree of plasticity as previously observed (Sadasivan et al., 1998a) and overall, the areas of greater plasticity are the same in all six data sets. The details change in going from one plot

to another in Figure 2 and interesting differences between the various sets of structures are reflected in the magnitude of the average RMSD per residue and in the range given by the low and high RMSD values. The set most closely related to the MSCS set in our study is that of the inhibitor-bound structures solved from crystals with symmetry of the C2 space group (Figure 2b). In these two sets the crystallographic environment is the same, with similar unit cell parameters and two molecules in the asymmetric unit. Any difference in the RMSD plots between the MSCS set and the C2 set is likely due to the solvent environment and this is the comparison most relevant in assessing whether MSCS can reflect the changes in structure that are represented in an aqueous environment. Figure 2b shows the average, high and low RMSDs per residue for the MSCS set in black lines and those for the C2 inhibitor set in gray. The average and low RMSDs for the two sets superimpose very well, with no significant areas of deviation. The high RMSDs follow similar trends, although with somewhat greater variability between the two sets. Residue 66, an active site Lys found in the P₀ pocket, is an exception to the overall similar trend in the high RMSD values, with that of the C2 inhibitor set (gray dotted line in Figure 2b) showing a significantly higher value than in the MSCS set (2.9 Å and 1.7 Å respectively). This is due to a single inhibitor structure (PDB code 1O0F) where in molecule B of the asymmetric unit residues 65-67 were built with an altered backbone conformation. Excluding this one molecule from the set results in much closer agreement between the high RMSD values for that region in the two data sets (compare the black and gray lines with squares in Figure 2b). Figure 2c shows a comparison between the MSCS set and the all-inclusive inhibitor set, which contains the effects of six different crystal environments. Due to the fact that molecule B of the 1O0F model dominates the high RMSD

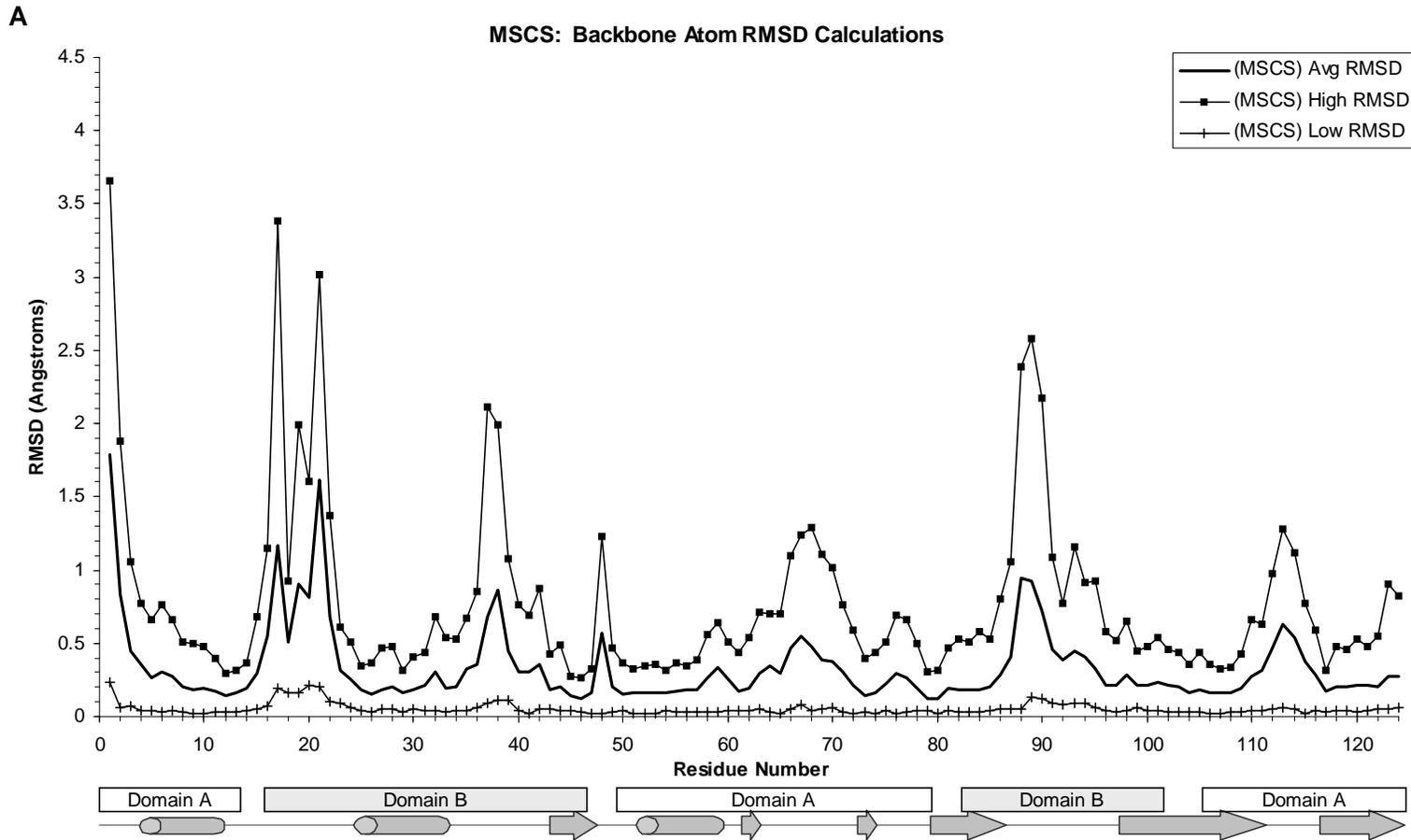


Figure 2. RMSD calculations: High, Low, and Average RMSD per residue. Secondary structural elements are depicted below the x-axis with cylinders and arrows illustrating helices and strands, respectively; and the boxes indicate residues belonging to Domain A and Domain B. A. Backbone RMSD calculations for 20 MSCS structures of RNase A.

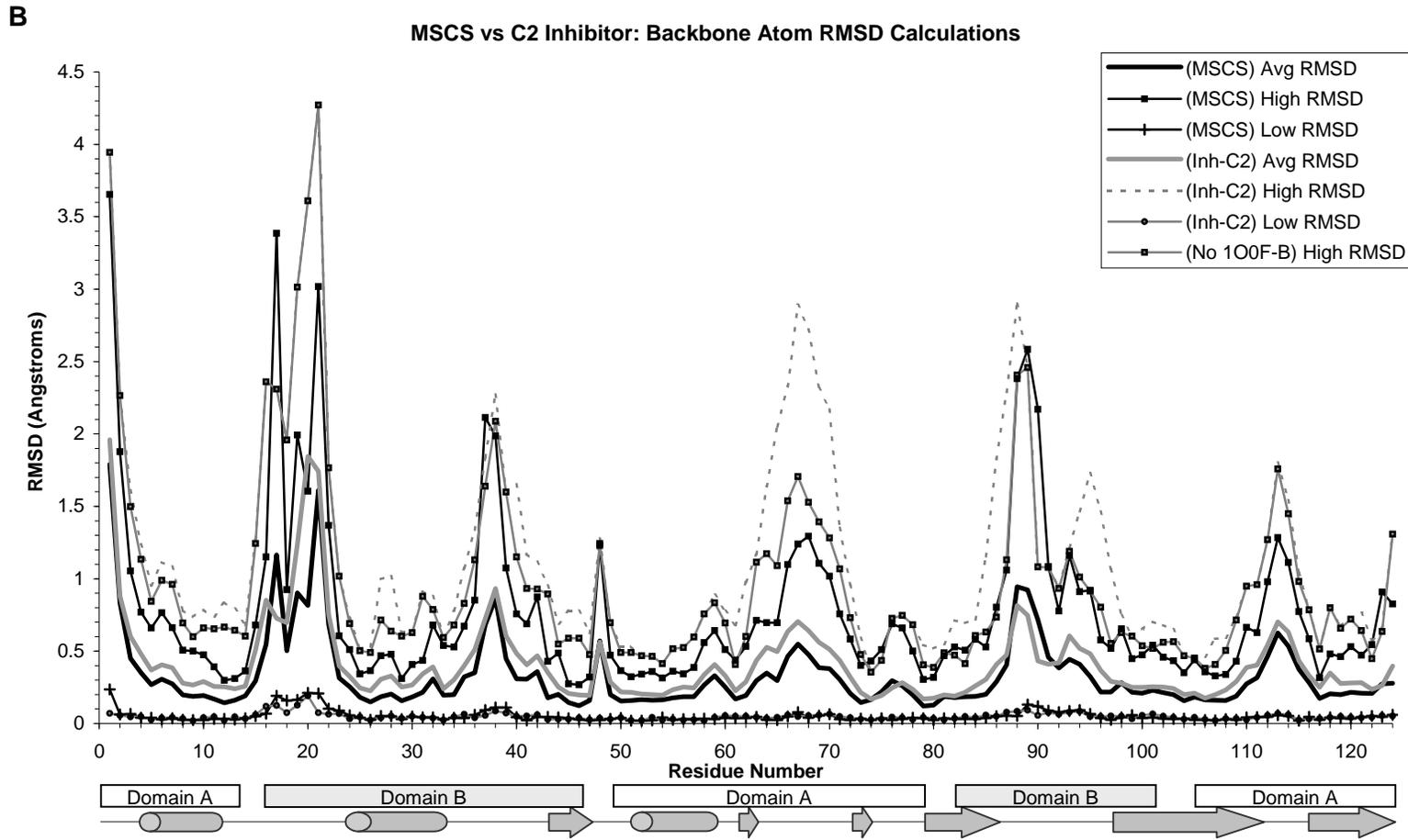


Figure 2. RMSD Calculations (continued).

B. Backbone RMSD calculations for inhibitor-bound structures in the C2 space group and MSCS structures of RNase A

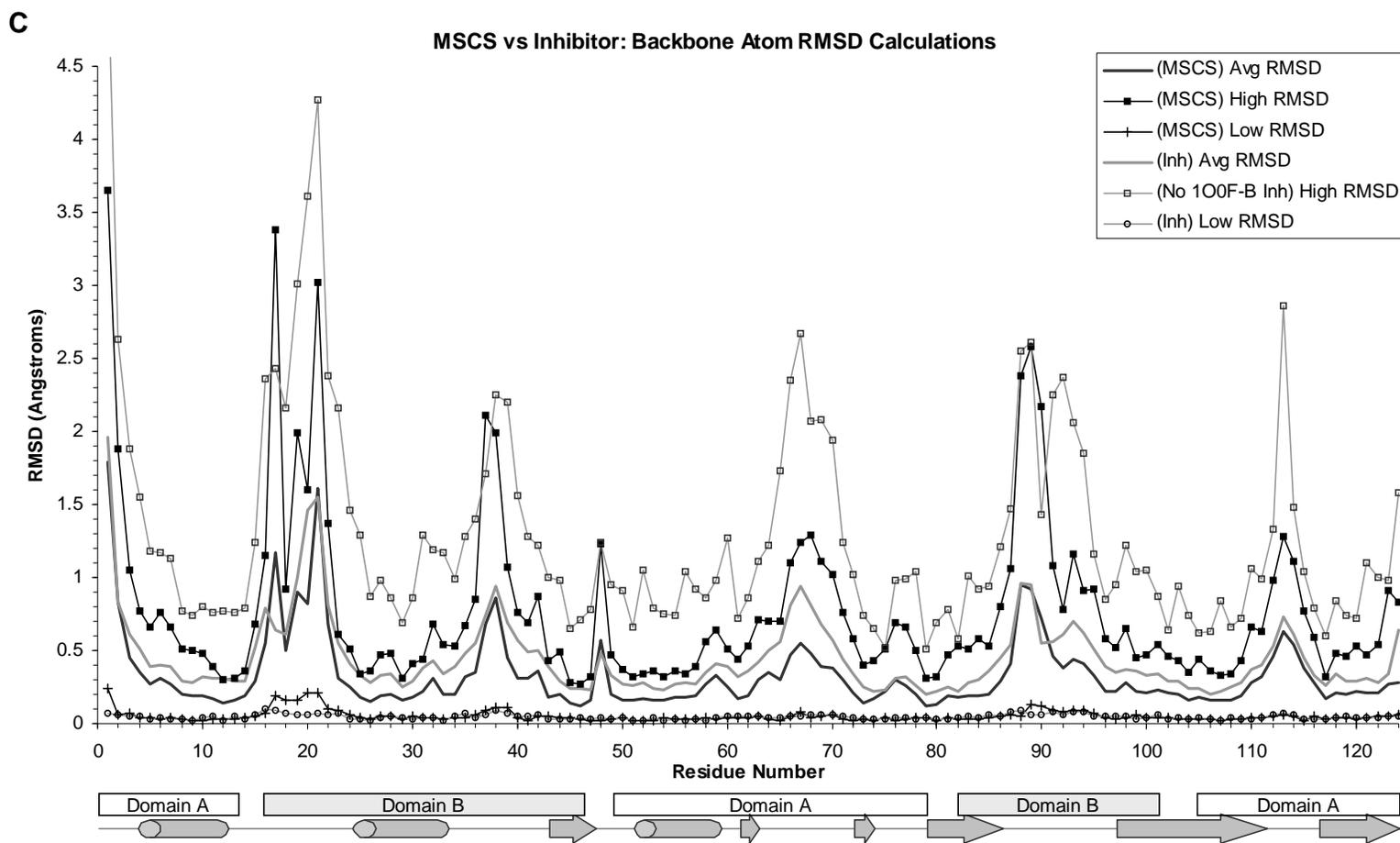


Figure 2. RMSD Calculations (continued)

C. Backbone RMSD calculations for inhibitor-bound and MSCS structures of RNase A. The High RMSD for residue 1 of the inhibitor-bound structures is 5.23\AA .

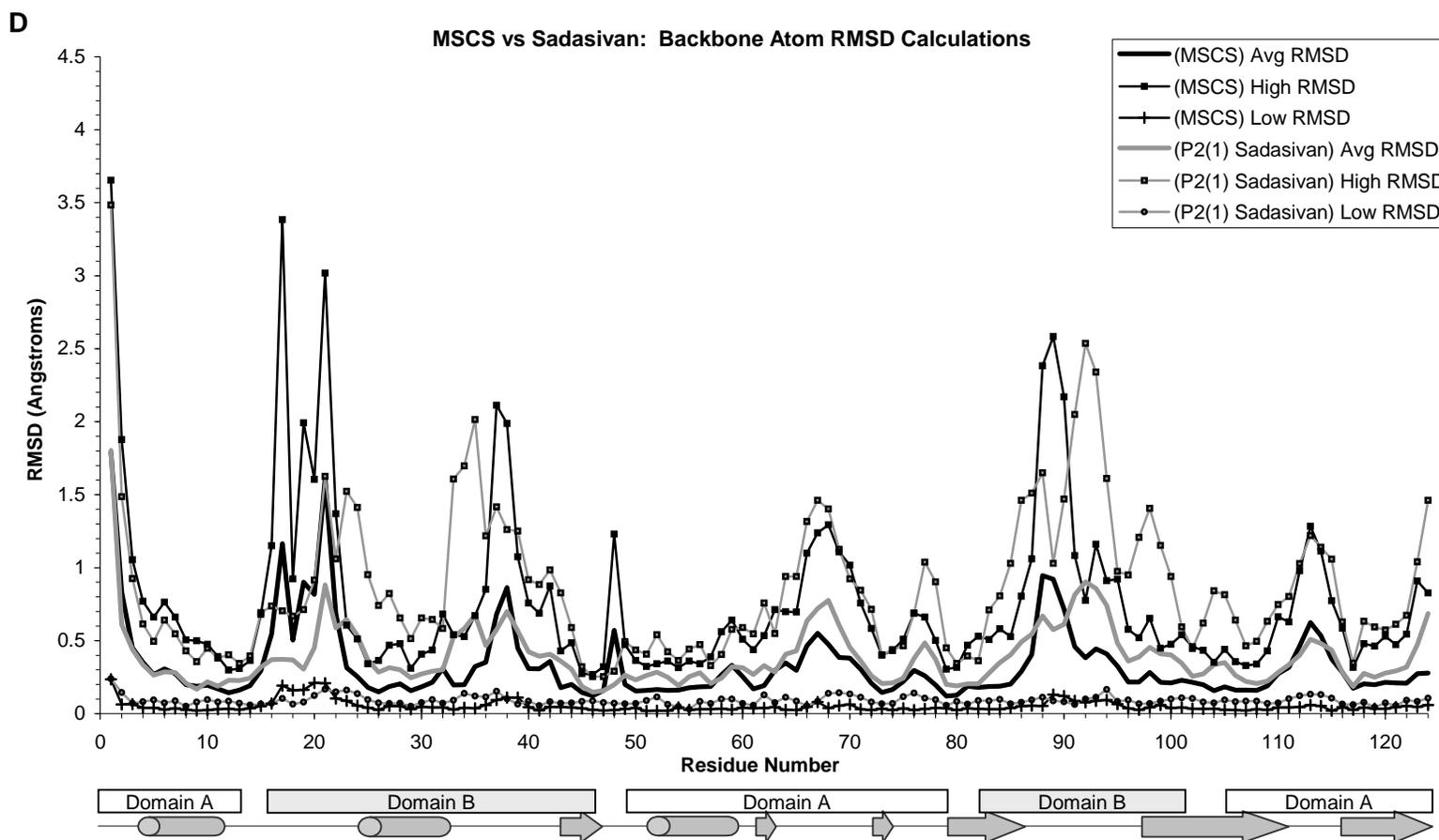


Figure 2. RMSD Calculations (continued)

D. Backbone RMSD calculations for the Sadasivan set and MSCS structures of RNase A. In the Sadasivan set, 1RPH model is excluded and only the structures in the P2₁ space group are compared.

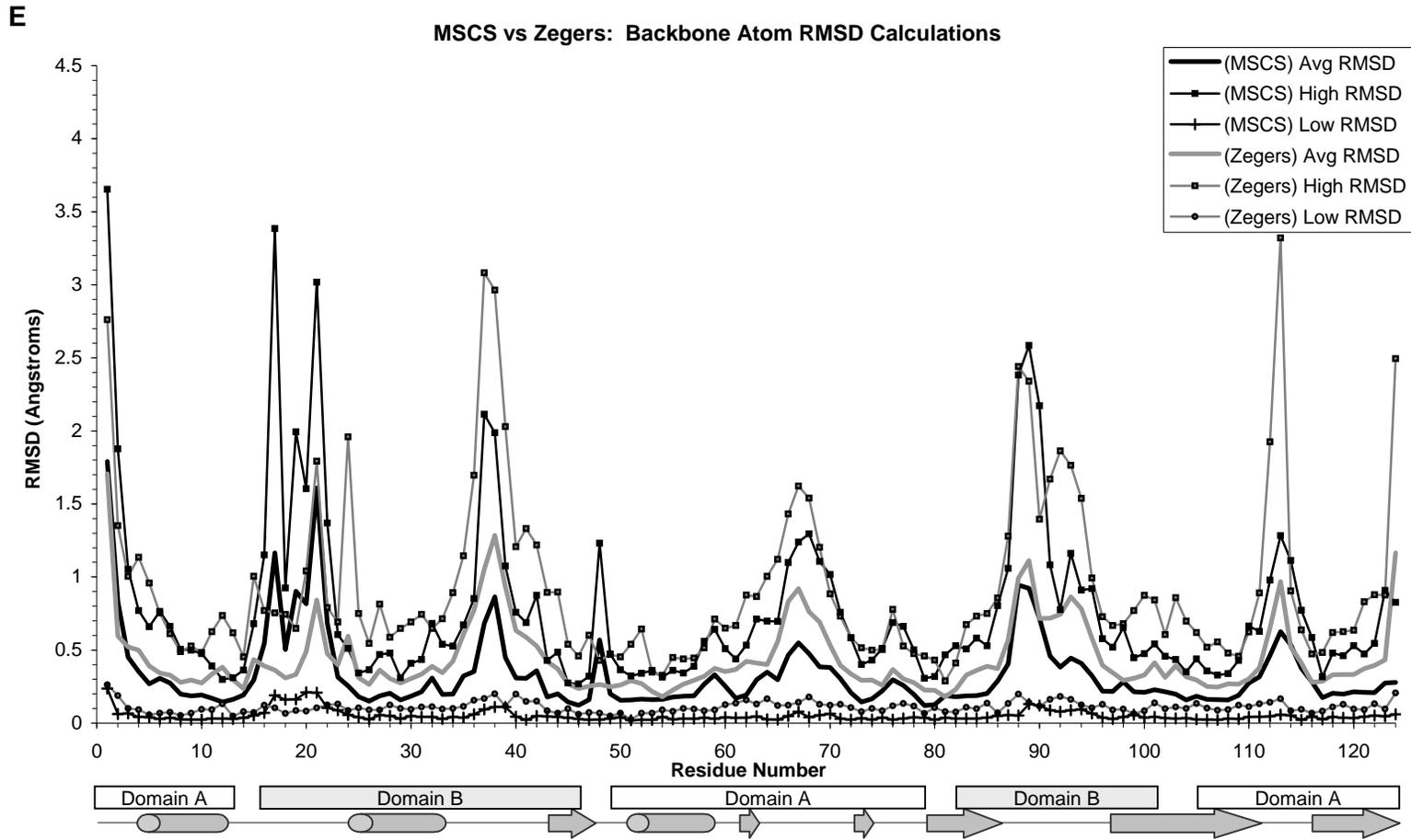


Figure 2. RMSD Calculations (continued)
 E. Backbone RMSD calculations for the Zegers set and MSCS structures of RNase A.

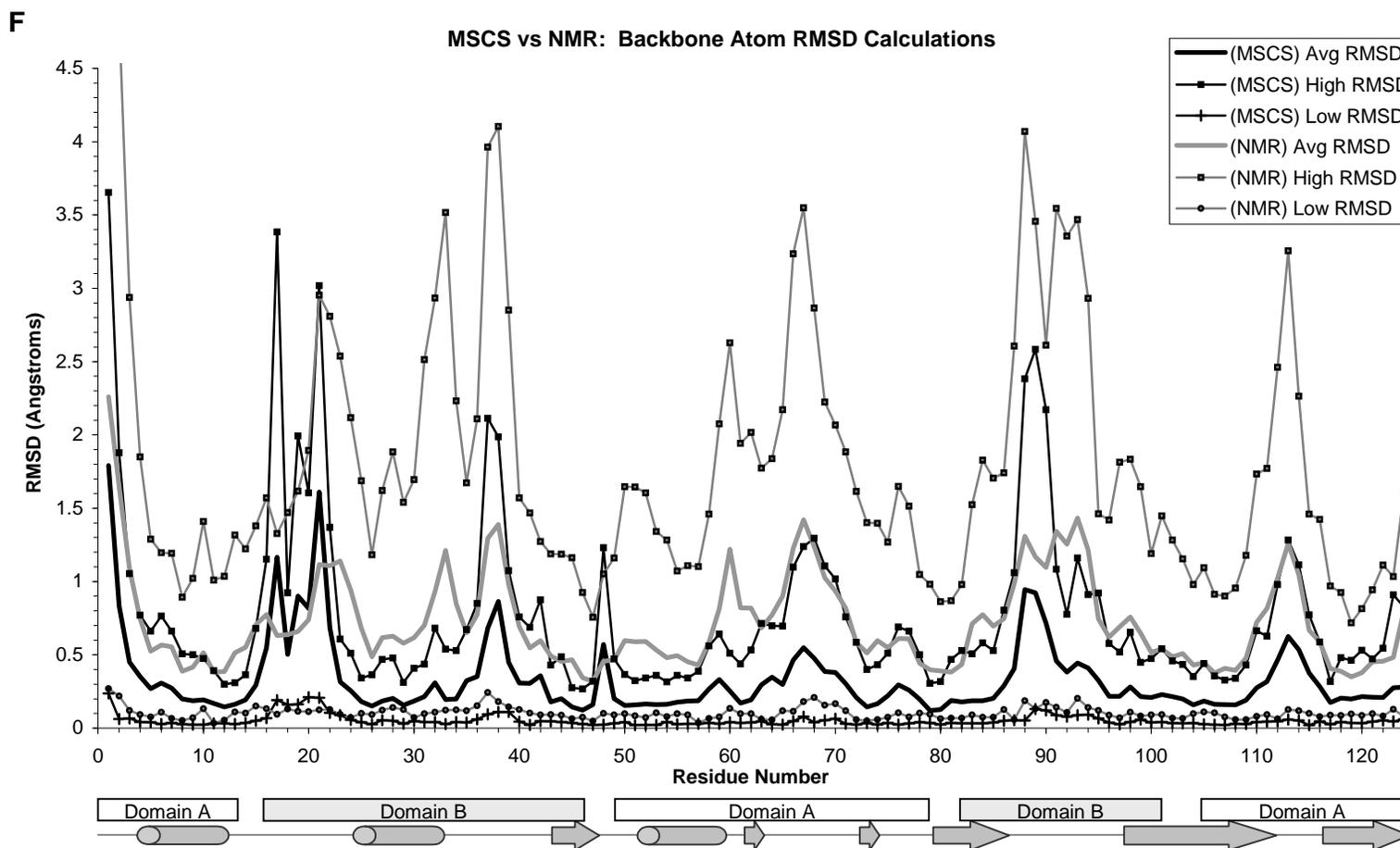


Figure 2. RMSD Calculations (continued)

F. Backbone RMSD calculations for 32 NMR models and MSCS structures of RNase A. The High RMSD values for residues 1 and 2 of the inhibitor-bound structures are 7.29Å and 4.85Å, respectively.

values, it was removed from the RMSD calculations in Figure 2c, but its inclusion results in a peak at residues 65-70 which is identical to that shown by the dotted line in figure 2b (not shown). In the case of the all-inclusive inhibitor set the trends in average RMSDs are again very similar to those observed for the MSCS sets, except around residues 66 and 93, where the inhibitor set shows higher RMSD averages. Unlike the case for the C2 inhibitor set, however, the high RMSDs for the all-inhibitor set is overall higher than for the MSCS set. This is a reflection of the fact that, in this set, RNase A structures solved in six distinct crystallographic space groups are superimposed and compared to each other. The high RMSDs reflect comparisons between the most different structures derived from crystals having distinct symmetry constraints.

The Sadasivan set provides an opportunity to directly compare the plasticity of structures solved from crystals with symmetries of two distinct space groups. The MSCS models were from crystals with symmetry of the C2 space group and all but one of the structures (PDB code 1RPH) in the Sadasivan set were derived from crystals with P2₁ symmetry. Figure 2d shows the RMSD plots of the MSCS set superimposed on those for the Sadasivan set with the 1RPH model removed. While the average RMSD values per residue are very similar for the two sets, there are six regions where they differ significantly. The first three include residues for which the average RMSDs in the MSCS set are larger than those in the Sadasivan set. Residues 16-20 are in a region of disorder in structures derived from crystals with symmetry of the C2 space group (MSCS set) while in the P2₁ form this is a region of crystal contact (Sadasivan set). His 48 lies beneath residues 16-22 and is stabilized in the Sadasivan

set. Residues 88-90 are also near crystal contacts in the Sadavisan set, but not in the MSCS set. There are another three regions for which the reverse is true: residues 32-36, 91-94 and 123-124 are stabilized by crystal contacts in the MSCS set, but not in the Sadavisan set, resulting in higher average RMSD values for these residues in the latter set. These trends are observed much more dramatically when considering the high RMSD values in both sets, but the overall differences are similar due to the distinct areas of crystal contacts found in the C2 and P2₁ crystal forms. The Sadasivan study (Sadasivan et al., 1998b) identified residues 14-15, 25-30, 46-48, 50, 52-61, 63, 72-75, 80-84, 97-100, 102-112, and 116-120 as areas of low plasticity based on the analysis of ten RNase A molecules in their study. All of these residues, with the exception of 48 and 88-90 which are stabilized by crystal contacts in the P2₁ but not in the C2 form, have the lowest average RMSDs in the MSCS set as well. Note that residues 16-20 are not included in the Sadavisan invariant set and this is due to increased disorder for these residues in the structure with PDB code 1RPH, solved from crystals with symmetry P3₂21, as well as in several other models in this set.

The Zegers study (Zegers et al., 1994) focused primarily on analysis of conserved water molecules, but they did note that between 1RPH and 1RPF (both of the P3₂21 space group), the differences were found primarily in the regions of 19-23 and 87-91. Additionally, among the three structures of 1RPH, 1RPF, and 1RPG (from crystals with P2₁ symmetry), the differences were observed in the regions of 35-42, 65-71, and 88-96. When the average RMSDs obtained from the Zegers set are compared to those of the MSCS structures, seven regions of differences are highlighted (Figure 2e). Three of these regions, 16-22, 48, 93-94

have the same explanation for their differences as between the MSCS structures and the Sadasivan set of structures, that is, they are due to differences in crystal contacts. Residues 24, 113 and 124 are stabilized by crystal contacts in the MSCS structures and show lower average RMSDs than in the Zegers set. In the Zegers structures the crystal packing varies for these residues, allowing for a greater range of conformations to be adopted across the set as reflected in the higher RMSD values. The final region, residues 11-13, adopts a greater range of conformation in the Zegers set. This is due the 8RSA structures, in which His 12 is covalently modified, resulting in a shift in residues 11-13 relative to the rest of the models.

The 32 NMR models from PDB ID 2AAS (Santoro et al., 1993) were considered individually and the RMSD values were calculated and plotted as was done with the MSCS models.

Model 16 and model 32 are identical; therefore, to prevent these two models from biasing the low RMSD value, it was not allowed to adopt the value of zero for any residue. When the RMSD plot of the NMR models was compared to the plot for the MSCS models, the overall trends were again the same (Figure 2f). The NMR models exhibited a larger range between the low and high RMSD values for each residue, and the average RMSD value per residue was consistently above that of the MSCS set. This is expected given the lack of crystal environment constraints in solution. Interestingly, the high RMSDs from the MSCS set are comparable to the average RMSD values in the NMR set. Areas that differ in trends between these two sets are those that have been previously noted to be in crystal contacts in the MSCS set: residues 16-22 and 48 have notoriously high RMSDs in the MSCS set due to poor electron density for model building in the C2 crystal form. Residues 33-34, 58-63 and 91-94

are in crystal contacts in the MSCS set but showed high plasticity in other crystal forms. It is therefore not surprising that these regions show higher RMSDs in the NMR set. Overall, the plasticity described by RMSD calculations across the MSCS structures is qualitatively similar to plasticity seen in NMR structures, or across a set of crystal structures encompassing a wider range of solvent and crystal environments. Differences among the sets of structures tend to be in the magnitudes of the values, reflecting the various crystallographic environment or the intrinsic differences expected when comparing X-ray crystallography or NMR methods of data collection. There appears to be no effect uniquely attributed to the organic solvent environments on the overall magnitude or trends in plasticity within the MSCS set.

Conserved Water Binding Sites

With Multiple Solvent Crystal Structures a protein is observed in a diverse set of environments where the first hydration shell is the primary mediator between the protein and the bulk solvent (Mattos, 2002). Thus, as the protein structure makes small adjustments to its environment, so do the accompanying water molecules, with the result that the MSCS models collectively sample far more hydration sites on the protein surface than are seen in any one structure (Figure 3). Within this collective set of crystallographic water molecules there is a small number found in the same position in all of the structures (Mattos et al., 2006). The question is whether these conserved water positions are representative of conserved crystallographic water in aqueous environments. The four RNase A sets of crystal structures downloaded from the PDB as described in the Methods serve as the test sets in this

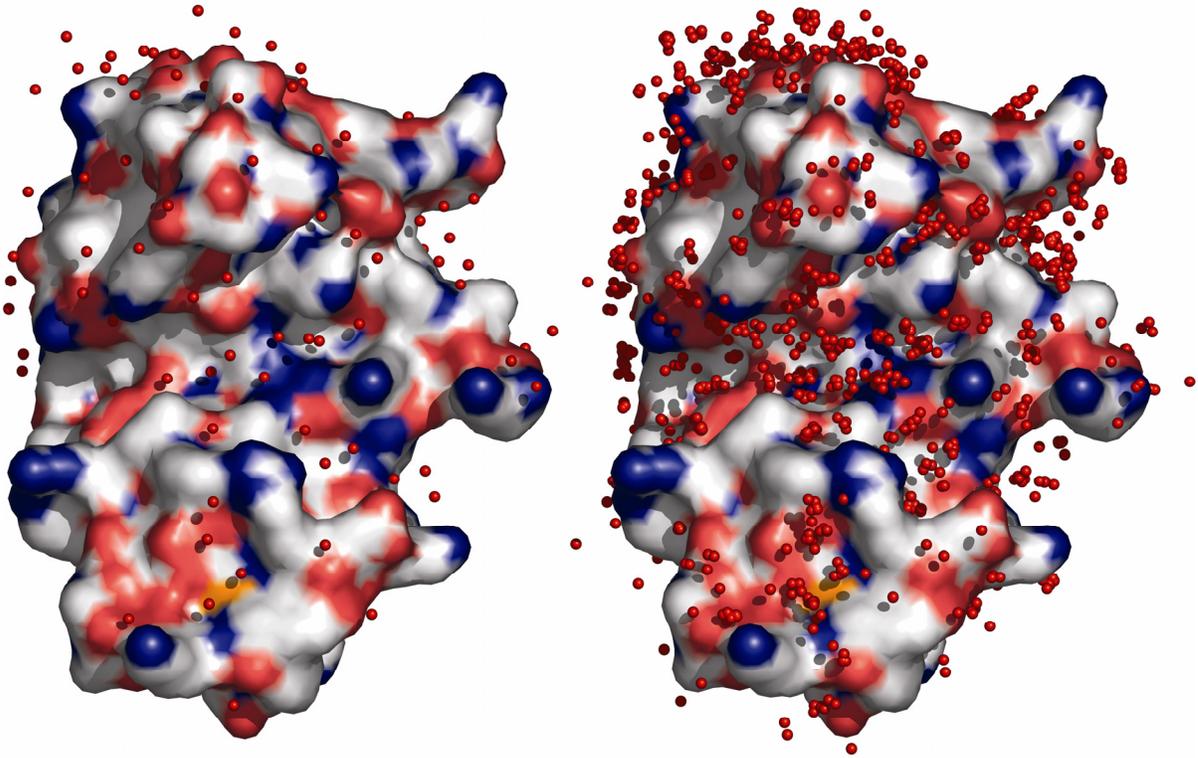


Figure 3. Solvation of RNase A. RNase A with all water molecules from a single MSCS model is shown on the left and an RNase A molecule with water molecules superimposed from all MSCS models is shown on the right.

study to determine the relevance of conserved water positions derived from MSCS of RNase A. As with the analysis of plasticity, these four sets allow comparison within the same crystallographic environment (C2 inhibitor set), within multiple crystallographic environments (all inhibitor set), and with two studies that had previously determined conserved water molecules within smaller sets of structures (Sadasivan and Zegers sets). The latter two sets also provide a check for consistency between different methods used in the determination of conserved crystallographic water molecules.

The conserved water binding sites in the present study were identified with the SEWS program using a previously published procedure based on the densities of water molecules at particular sites within a set of structures (Bottoms et al., 2006). Considering that many of the structures involved have two molecules in the asymmetric unit, there are 20 models in the MSCS set with a total of 2077 water molecules, 36 in the C2 inhibitor set with 5141 water molecules, 56 in the all-inhibitor set with 6948 water molecules (models with PDB codes 1RBJ and 1RTA were excluded because they do not contain water molecules), and 10 each in the Sadasivan and Zegers sets with 1353 and 1190 water molecules, respectively.

A water-binding site was considered conserved within a set when occupied in at least 80% of the structures. This avoided bias from one or a few structures where there may not be a water molecule at a conserved position for reasons such as bound organic solvents or crystal contacts. Table 3 shows the 31 conserved water binding sites within the MSCS set, the percent occupancy of each of those water positions in the other four sets, as well as the interactions with protein atoms. Of these, 14 are conserved across all of the five sets of structures based on the SEWS analysis as described above (see Table 3 for the identity of these water molecules). There are 22 conserved water positions in the C2-inhibitor set of structures obtained from crystals isomorphous to those used for MSCS. Interestingly, the majority of the nine water molecules found to be conserved in MSCS but not in the C2-inhibitor set are found in crystal contacts in both sets, but at 65-75% in the latter, not quite making the cutoff of 80%. Furthermore, of the 22 conserved water position in the C2-inhibitor set 19 are identified by MSCS. The three remaining water molecules are found in

Table 3. Conserved Water Molecules identified by SEWS for the MSCS structures of RNase A.

Conserved Water Position	MSCS Conservatn	Inhibitor-Bound Conservatn	C2 Inhibitor-Bound Conservatn	Sadasivan Conservatn	Zegers Conservatn	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)
2	100%	61%	94%	<50%	<50%	Tyr 76 OH	2.74						
4	100%	100%	100%	100%	100%	Ala 5 O	2.92	Pro 117 O	2.71				
5	80%	80%	82%	80%	70%	Ser 15 N	3.01						
6	100%	96%	97%	100%	100%	Ser 50 N	2.88	Glu 49 OE1	2.80	Asp 53 OD2	2.77		
7	90%	82%	94%	90%	80%	Ala 52 O	2.67						
10	100%	98%	97%	100%	100%	Ser 77 N	3.01	Tyr 76 N (95%)	3.22	Gln 60 OE1	2.81		
13	90%	89%	94%	80%	70%	Asn 67 N	2.97	Lys 66 N (75%)	3.26	Asp 121 OD1	2.72		
15	100%	93%	100%	100%	100%	Glu 9 OE2 (95%)	2.92	Gln 55 OE1/NE2 (3)	2.85				
16	90%	89%	100%	100%	100%	Gln 11 OE1 (9) / NE2 (9)	2.90						
17	100%	95%	100%	90%	90%	Pro 114 O	2.73						
18	85%	89%	94%	100%	80%	Ala 6 O	2.80						
19	100%	<50%	<50%	50%	<50%	Phe 120 N	2.99	His 12 NE2	2.82				
20	85%	89%	100%	90%	90%	Ala 4 O	2.77	Val 118 O	2.63				
21	95%	96%	100%	100%	90%	Ser 23 O	2.86	Asn 27 N	3.00	Tyr 97 O	3.08	Thr 99 OG1	2.94
23	100%	55%	69%	70%	<50%	Cys 110 O	2.82						
26	100%	75%	89%	90%	60%	Cys 58 O	2.73						
32	85%	59%	75%	50%	50%	Ser 50 OG	2.78						

Table 3 continued.

Conserved Water Position	MSCS Conservatn	Inhibitor-Bound Conservatn	C2 Inhibitor-Bound Conservatn	Sadasivan Conservatn	Zegers Conservatn	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)
34	85%	52%	<50%	70%	70%	Asn 71 ND2	2.96	Glu 111 OE2 (55%)	3.03				
39	95%	93%	97%	100%	90%	Glu 111 O	2.88						
57	80%	75%	75%	90%	90%	Arg 33 NE	2.97						
63	80%	77%	75%	100%	90%	Ser 23 O	2.85	Thr 99 N	2.84				
64	95%	98%	100%	100%	100%	Asp 83 O	3.09	Lys 98 O	2.84	Thr 100 OG1	2.80		
69	100%	89%	94%	100%	100%	Ala 5 N	2.94	Ala 4 N (70%)	3.27				
78	90%	<50%	50%	<50%	<50%	Asn 62 O	2.93						
82	100%	98%	100%	100%	100%	Asp 53 O	2.73	Gln 60 NE2	2.92				
90	80%	50%	72%	<50%	<50%	Thr 78 N	3.05						
129	80%	<50%	<50%	<50%	<50%	Thr 78 OG1	3.00	Asn 103 OD1	2.82	Thr 78 O (70%)	3.21		
158	85%	59%	69%	<50%	<50%	Gln 74 OE1	2.65						
243	80%	61%	72%	50%	<50%	Ala 4 N	3.03						
318	80%	80%	83%	70%	100%	Asn 62 N	2.93						
423	85%	73%	86%	80%	60%	Gln 60 NE2	3.03						

Table 4. Water Molecules with less than 80% conservation in the MSCS structures.

Conserved Water Position	MSCS Conservatn	Inhibitor-Bound Conservatn	C2 Inhibitor-Bound Conservatn	Sadasivan Conservatn	Zegers Conservatn	Interaction	Interaction	Interaction
B₁	Pocket:							
22	70%	<50%	<50%	80%	50%	Thr 45 N		
60	75%	77%	89%	50%	70%	Ser 123 N		
153	45%	80%	83%	80%	90%	Asp 83 OD1	Thr 45 OG1	Ser 123 OG
Missing	from	MSCS:						
12	70%	55%	56%	60%	80%	Val 43 O		
43	55%	55%	58%	100%	50%	Thr 36 O	Pro 93 O	
44	65%	82%	89%	70%	70%	Glu 2 OE2		
56	<50%	73%	78%	80%	80%	Lys 31 O	Thr 36 OG1	
110	50%	77%	69%	80%	90%	Asn 27 O	Cys 95 O	
115	65%	59%	64%	80%	70%	Arg 10 NH1		
143	<50%	<50%	53%	80%	50%	Glu 86 OE2		
163	50%	68%	64%	100%	80%	Thr 3 N		
215	40%	<50%	<50%	70%	100%	Gln 101 N		
229	<50%	55%	56%	90%	70%	Gln 55 O		
282	35%	61%	56%	100%	80%	Ala 52 N		
291	<50%	<50%	<50%	100%	70%	Asp 14 OD2		
361	<50%	<50%	<50%	80%	50%	Leu 51 N		
368	<50%	<50%	<50%	<50%	50%	Ser 18 N		
443	<50%	<50%	<50%	50%	80%	Asn 103 OD1		

Table 4, which lists water-binding sites that were conserved in one or more of the comparison sets, but not in the MSCS set. Two of the three water molecules are in the B₁ pocket where they are often displaced by organic solvents in the MSCS set (60 and 153) and

one is at the interface between molecules A and B (44). In the MSCS set Wat 44 is 100% conserved in molecule B, but it is only found associated with molecule A in three of the ten structures. There are 18 conserved water positions in the all-inhibitor set, 31 in the Sadasivan set and 24 in the Zegers set. The majority of the conserved water binding sites in all five sets of structures are associated with domain A of RNase A, with 87%, 86%, 83%, 61%, and 58% of the water molecules associated with domain A in the MSCS, C2-inhibitor, all-inhibitor, Sadasivan, and Zegers sets, respectively. Interestingly, but not surprisingly, the conserved water sites are in the areas of least plasticity in the structures (Figure 4).

The most thorough crystallographic water analysis on RNase A to date was done on the Zegers set of structures, where 17 water molecules were found to be 100% conserved in all members of the set, except when modifications precluded the binding of water (Zegers et al., 1994). The less strict 80% conservation criterion applied in the present study explains the additional seven conserved water binding sites obtained with SEWS. All but four of the 17 published conserved water molecules were found to be associated with one of the three α -helices in RNase A, often linking it to a nearby β -strand (Zegers et al., 1994). There are seven conserved water positions that link the N-terminal α -helix containing the active site residues Gln 11 and His 12 to the C-terminal β -strand with residues His 119 and Phe 120 (Figure 1b). These water molecules were suggested to effectively stabilize the active site (Zegers et al., 1994). These seven water molecules were identified by MSCS and have the following numbers in Table 3: 20 (P_1 pocket), 16 (P_1 pocket), 4, 15, 17, 39, 69. Two water molecules, 21 and 63, link the beginning of helix 2 with a nearby β -strand and are both

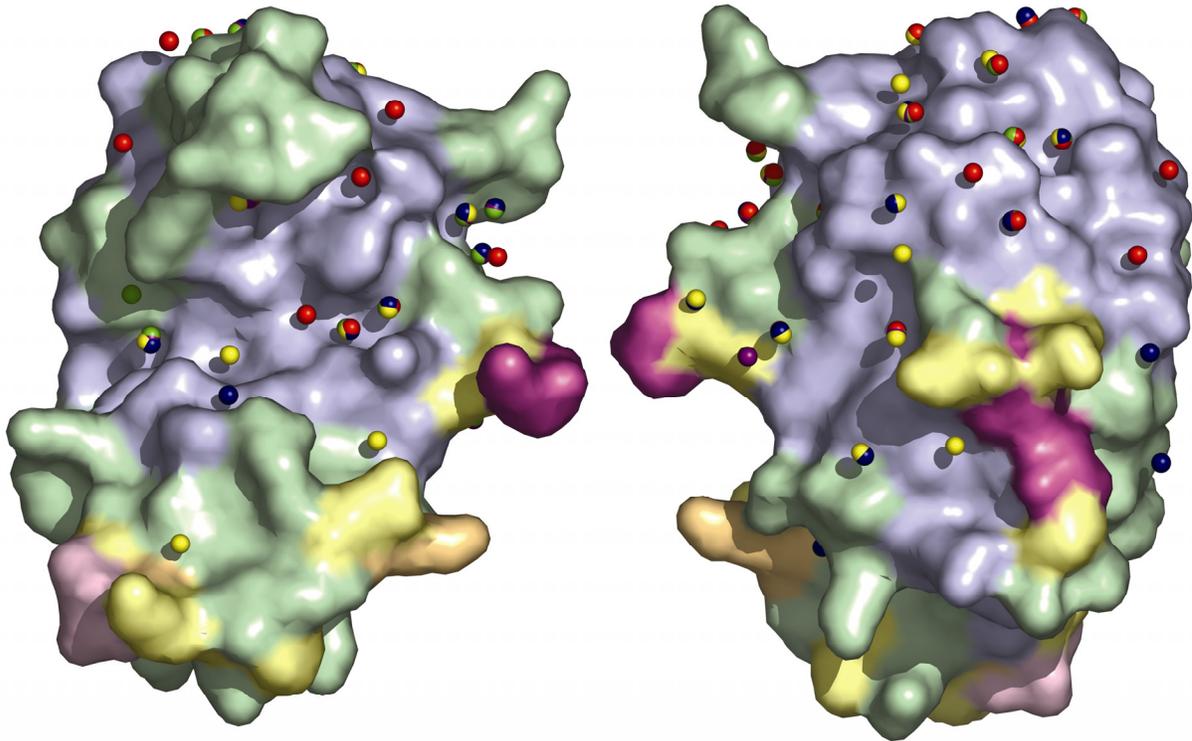


Figure 4. Plasticity and conserved water binding sites of RNase A. Front view (looking down on the active site) of RNase A molecule is shown on the left and the reverse view is shown on the right. The amino acids of RNase A are colored according to their calculated average backbone RMSD in the MSCS set of structures using the following coloring scheme: 0.0-0.2Å, light blue; 0.2-0.4Å, pale green; 0.4-0.6Å, pale yellow; 0.6-0.8Å, light orange; 0.8-1.0Å, light pink; 1.0+Å, light magenta. Conserved water binding positions found in each of the sets of structures are superimposed on the RNase A molecule and colored as follows: MSCS, red; Inhibitor, purple; C2 Inhibitor, green; Sadasivan, yellow; Zegers, dark blue. The RNase A molecule is oriented with domain A at the top and domain B at the bottom. A majority of conserved water binding sites is associated with domain A, which generally demonstrates less motion than domain B.

identified as conserved by MSCS. Four additional conserved water molecules, 6, 10, 82 and 282, help maintain a distorted structure for α -helix 3 and bridge it to a β -strand. Water molecules 6, 10 and 82 are found to be conserved by MSCS, but 282 is only 40% conserved (Table 4). It is located in an area of crystal contacts in the MSCS structures and water molecules in this position are most commonly associated with a symmetry-related molecule

(103 in chain C and 155 in chain D). Taking this into consideration, this water molecule is 80% conserved in the MSCS set. The remaining four water-binding sites published in the Zegers study (Zegers et al., 1994) are on the protein surface. Two are identified by MSCS as conserved (64 and 318) and the remaining two are not (215 and 368). Wat 215 (Table 4) is 100% conserved in the Zegers set, 70% conserved in the Sadasivan set, and less than 50% conserved in the MSCS and both the C2- and all-inhibitor sets. The protein residue, Gln 101, with which it interacts in molecule B in the C2 crystal form packs against Ser 16 and the O γ group of this residue takes the place of Wat 215 in the MSCS and C2-inhibitor sets. This is not the case for molecule A, where Wat 215 is present in 80% of the MSCS structures, resulting in an overall conservation level of 40% when molecules A and B are taken together. Wat 318 is less than 50% conserved in all five sets of structures based on the SEWS analysis criteria. It is located in a position where it sometimes interacts with the backbone nitrogen of Ser 18. While it is found in this position in 80% of the structures in the Zegers set, it only interacts with the backbone nitrogen of Ser 18 or with the O γ group of the same residue in 30% and 20% of the structures, respectively. It was found to be conserved by the Zegers study (Zegers et al., 1994) because they did not use interaction with a common protein residue as one of their criteria for conserved water-binding sites. In the MSCS structures, this is in an area of disorder where water molecules were often not modeled. Since similar disorder can be observed in structures of the C2-inhibitor set for residues 16-22, it is unlikely that in the MSCS structures the disorder is due to the effects of organic solvents.

A conserved water analysis was also published for the Sadasivan set (Sadasivan et al., 1998a). In this study a water site was taken as conserved if a bound water molecule made at least one interaction with the protein in common in all structures belonging to the set and the distance between equivalent water molecules was 1.8 Å or less after superposition. By these criteria there are 14 invariant water binding sites identified by the Sadasivan study (Sadasivan et al., 1998b), ten of which are found among the 17 invariant water-binding sites published in the Zegers study (Zegers et al., 1994). The variation in conserved water molecules between the Zegers and Sadasivan studies have been previously discussed (Sadasivan et al., 1998a) and are in large part due to the differences in selection criteria and crystallographic environments between structures in the two studies.

Of the 14 conserved water molecules published in the Sadasivan study ten are also identified as conserved by MSCS and have the following numbers in Table 3: 4, 6, 10, 18, 21, 39, 63, 64, 69, and 82. Water molecules 43, 163 and 282 appear on Table 4 because they are not conserved in the MSCS set but are identified as conserved elsewhere, including in the Sadasivan study. Wat 282 has already been discussed above because it was also found to be conserved in the Zegers study. Wat 43 is only found to be conserved in the Sadasivan set because there it often is involved in a H-bonding network across a crystal contact. In all four other sets, Wat 43 is around 50-60% conserved. Wat 163 is 50% conserved in the overall MSCS set. However, it is 100% conserved in molecule B of the asymmetric unit in all MSCS structures, but completely absent in molecule A due to a crystal contact that takes its place. In the C2-inhibitor set Wat 163 sometimes appears at the crystal contact associated with

molecule A and thus it is 67% conserved in this isomorphous set of structures. Finally, Wat 291 is associated with residues 14, 16, and 17 in the Sadasivan set. It was added seldomly to the MSCS set of structures due to disorder in the 16-22 region as previously discussed and is therefore not included in Table 3.

Water Molecules in the Active Site

An analysis of water conservation in the active site is complicated by the fact that in many of the structures, including those in the MSCS set, there are bound molecules associated with this area. Nevertheless, there are nine water-binding sites that appear frequently in all five sets of structures. Water molecules 22, 60, 153 are in the B₁ pocket (Figure 5a), 16, 19, 20 are in the P₁ pocket (Figure 5b), and 13,14 and 34 are in the B₂ pocket (Figure 5c). As shown in Table 3, water molecules 13, 16, 19, 20 and 34 are better than 80% conserved in the MSCS set, with 16 and 20 in the P₁ pocket highly conserved in all five sets, while Wat 13 is within the cutoff in all but the Zegers set, where it is 70% conserved. Wat 19 in P₁ and Wat 34 in B₂ are often displaced by inhibitors, lowering the conservation level in all four comparison sets. The three water molecules in the B₁ pocket, 22, 60 and 153 are shown in Table 4 to be conserved in at least one of the four comparison sets and when the two molecules of the asymmetric unit of the MSCS structures are considered independently, all of these water molecules are at least 80% conserved in either molecule A or molecule B, but not both. Wat 14 is present in 75% of the MSCS models but has no common interaction with

protein atoms and thus is not picked up as conserved by the SEWS criteria used here. It is displaced in all inhibitor-bound structures.

The water molecules bound in these nine water-binding sites can be categorized as structural, bridging, or displaced. Structural water molecules are rarely displaced and form multiple hydrogen bonds with polar groups of the active site and other well-ordered water molecules. When an inhibitor molecule is bound, bridging water molecules bridge the interaction between residues lining the active site and the inhibitor. Finally, inhibitors bound in the active site replace hydrogen bonds formed by displaced water molecules.

Water molecules 16 and 20 of the P₁ pocket can be classified as structural. These water molecules participate in a hydrogen-bonding network involving polar groups that line the pocket, and water molecule 19, when it is present (Figure 5b). This network adds to the stability of the active site by connecting helix 1 to β -strand 6, which belong to the two lobes on either side of the active site (Zegers et al., 1994). Wat 16 forms hydrogen bonds to Gln 11 N ϵ 2 (or O ϵ 1, depending on the orientation of the side chain) and sometimes to the backbone oxygen of Val 118. In most cases when an inhibitor is bound in the P₁ pocket, water molecules present in conserved site 16 interact with the phosphate moiety. It is displaced by inhibitors bound in the 1RAR and 1RAS structures, and by a formate molecule bound in the 1RNN structure. Wat 16 is missing in the active site of molecule A of 9RSA, but there is no inhibitor binding in the P₁ pocket. Both molecules of 1U1B have an inhibitor phosphate group bound in the P₁ pocket of the active site, but there is no Wat 16 to interact with it,

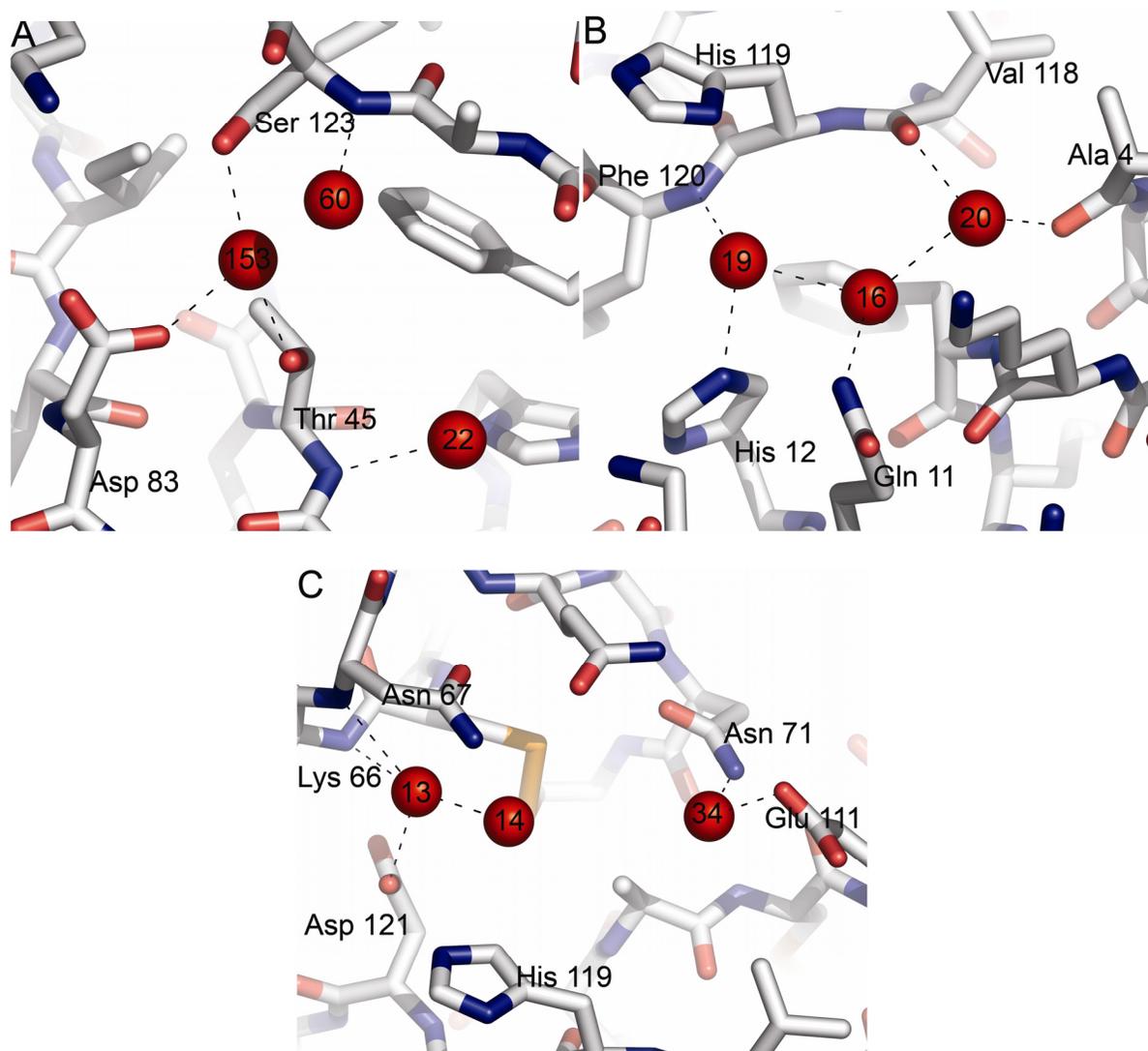


Figure 5. Conserved water molecules in the active site of RNase A. Conserved water binding positions are shown in red and the numbers correspond to the numbering of MSCS water molecules. The protein is drawn as sticks, colored by atom type, and the dashed lines indicate hydrogen bonds. A. The B₁ pocket B. The P₁ pocket. C. The B₂ pocket

contrary to the previous observation that this water always interacts with P₁ phosphate or sulfate group of inhibitors in RNase A-inhibitor complexes (Zegers et al., 1994). Wat 20 hydrogen bonds to the backbone oxygen atoms of Val 118 and Ala 4. Binding of inhibitors

generally does not disrupt Wat 20, however this water molecule is not present in the 1RCN, 1RSM, 9RSA, and 1U1B models.

Water molecules 60 and 153 of the B₁ pocket, and Wat 13 of the B₂ pocket bridge the interactions between the residues lining the active site and a bound inhibitor. Wat 153 interacts with Asp 83 O δ 1, Thr 45 O γ 1, and Ser 123 O γ (Figure 5a). In inhibitor bound structures, it bridges a pyrimidine moiety to the active site residues Asp 83 and Ser 123. In the MSCS structures, this water molecule is not conserved because it is only found associated with molecule B, whereas in molecule A, Asp 83 is not within H-bonding interaction. Water molecule 60 interacts with Ser 123 N and bridges a bound pyrimidine to this side chain.

Organic solvents displace this water molecule in two of the MSCS structures. Water molecules 153 and 60 are both identified as changing their roles as hydrogen bond donors or acceptors depending on whether a uracil or cytosine is bound in the B₁ pocket (Gilliland et al., 1994). Wat 13 forms hydrogen bonds with the backbone nitrogen atoms of Lys 66 and Asn 67, with the Asp 121 O δ 1 atom and with water molecule 14 (Figure 5c). When inhibitors bound in the B₂ pocket displace Wat 14, Wat 13 bridges the interaction between the inhibitor and Lys 66, Asn 67, and Asp 121.

Upon inhibitor binding, water molecules 22 of the B₁ pocket, 19 of the P₁ pocket, and 14 and 24 of the B₂ pocket are displaced. Wat 22 interacts with Thr 45 N and is displaced by every inhibitor and a number of organic solvent molecules bound in the B₁ pocket, where functional groups can form a similar interaction with Thr 45 N (Figure 5a). Wat 19 hydrogen

bonds with Nε2 of His 12 and the backbone nitrogen of Phe 120. It is invariably displaced when a phosphate binds in the catalytic site (Figure 5b). Wat 34 forms hydrogen bonds with Asn 71 Nδ2 and, when Glu 111 is oriented to be within hydrogen bonding distance, either with the Oε1 or Oε2 group of Glu 111 (Figure 5c). While it does not interact with the protein, Wat 14 is frequently observed in the MSCS models and forms a hydrogen bond with Wat 13. Both Wat 34 and Wat 14 are displaced when inhibitors bind in the B₂ pocket. These inhibitors then form some of the same hydrogen bonds made by water molecules found in these positions, particularly with Asn 71 Nδ2 and with conserved Wat 13.

Organic Solvent Binding Sites and Comparison with Inhibitors

Each of the superimposed MSCS structures was analyzed considering the positions of the superimposed downloaded inhibitors. A Perl script was used to identify atom types in common between the solvents and inhibitors (e.g. two carbons) that had positions within 1Å of each other. These atoms were identified as “overlapping.” Additional atoms that overlap (within 1Å), but do not have the same atom type were identified by visual inspection.

Thirty-nine organic solvent molecules bind at the surface of RNase A (Figure 6), occupying 23 unique binding sites, which are designated as 901-923. Five of the solvents are found at the interface between molecules A and B in the asymmetric unit, and were duplicated when the two protein molecules from the MSCS structures were separated into individual files.

An “A” or “B” is added to the solvent designation to distinguish with which protein molecule

the solvent associates. Fourteen of the organic solvents are found to bind in the active site, clustering into four unique binding sites. Two unique binding sites are located in the B₁ subsite, one in the P₁ subsite, and one in the B₂ subsite.

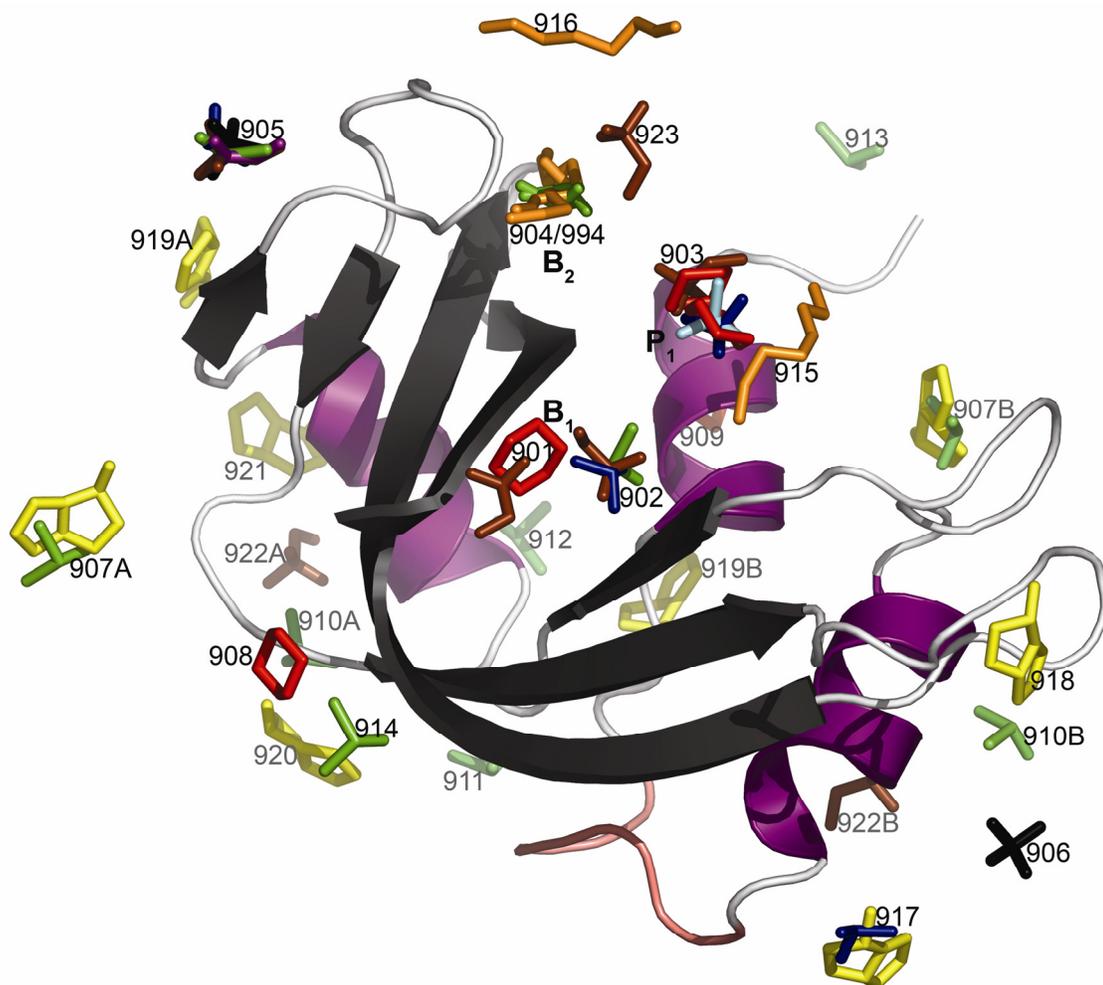


Figure 6. Organic solvent binding sites. Ribbon diagram of RNase A showing the binding sites for organic solvent molecules. The β -strands are shown in dark gray and the α -helices are shown in purple. The disordered loop region 16-22 is colored salmon. The organic solvent molecules are colored as follows: DIO, red; DMF, purple; DMS, green; HEZ, orange; IPA, blue; RSF, yellow; TBU, cyan; ETF, brown; and TMO, black. Organic solvent binding sites and RNase active site pockets are labeled.

Solvents Bound in the B₁ Subsite

Solvents in the B₁ pocket group into two clusters, 901 and 902 (Figure 7A). Cluster 901 contains a dioxane (DIO) and a trifluoroethanol (ETF) molecule and interacts primarily with the backbone nitrogen and O_γ of Ser 123. A dimethylsulfoxide (DMS), isopropanol (IPA), and a second trifluoroethanol (ETF) form cluster 902 and the main sites of interaction are the backbone nitrogen and O_{γ1} of Thr 45.

The current orientation of DIO901 allows the O2 atom of the dioxane to form a hydrogen bond with the backbone nitrogen of Ser 123. This replaces the hydrogen bond made by Wat 60. Water molecules found in this position bridge the interaction between the backbone nitrogen of Ser 123 and the pyrimidine base. An alternative orientation of DIO901 in which the O1 and O2 atoms occupied the current positions of the C2 and C4 atoms of the dioxane would allow the oxygen at the current C2 position to hydrogen bond with the backbone oxygen of Asp 121 and the oxygen at the current C4 position to hydrogen bond with Wat 153. The water molecule present in this position would bridge the interaction between the oxygen of the dioxane and the O_γ atom of Ser 123, the O_{δ1} atom of Asp 83, and the O_{γ1} atom of Thr 45. Additionally, if DIO901 is oriented in this way, the oxygen in the C4 position would be in a similar position to and mimicking the interactions made by the oxygen or nitrogen atom bound to the C4 atom of uracil or cytosine, respectively. DIO901 is probably sampling all of these interactions.

Table 5. Bound Organic Solvents and the interactions made with RNase A

Organic Solvent	Solvent Atom	Protein Atom	Distance												
DIO901	O2	Ser 123 N	3.15	O2	Ala 122 CA	3.11	O2	Ala 122 CB	3.24	C1	Ala 122 CB	3.55	C2	Asp 121 O	3.14
	C4	HOH 153 D	2.97	C3	Thr 45 OG1	3.67									
ETF901	O	Ser 123 OG	3.14	F1	Ser 123 OG	3.78	F1	Ser 123 O	3.26	F3	Ser 123 OG	3.33	F3	Ser 123 N	2.75
	F3	Phe 120 CE1	3.55	F1	Ser 123 N	2.99									
DMS902	O	Thr 45 N	2.77	O	Asn 44 OD1	3.54	C1	Thr 45 OG1	3.10	C1	Val 43 CG1	3.59	S	Phe 120 CD1	3.36
IPA902	O	Val 43 CG1	3.37	O	Asn 44 CA	3.31	O	Asn 44 C	3.04	O	Asn 44 O	3.40	CB2	Phe 120 CD1	3.57
	CB1	Thr 45 N	3.36	CB1	Thr 45 OG1	3.55	CB1	His 12 CE1	3.63	O	Thr 45 OG1	2.55			
ETF902	C2	Phe 120 CD1	3.41	F2	Phe 120 CD1	3.24	F2	Thr 45 OG1	3.18	F2	Thr 45 CB	3.34	F2	Thr 45 N	3.13
	F2	His 12 CE1	3.76	F1	Thr 45 OG1	2.76	F1	Val 43 CG1	3.55	F3	Val 43 CG1	3.90	F3	Asn 44 CA	3.54
	F3	Val 43 O	3.34												
DIO903	O2	His 119 ND1 (B conf)	2.54	C4	His 119 ND1 (B conf)	3.50	C4	Gln 11 OE1	3.88	C3	His 119 CE1 (B conf)	3.96	C3	Lys 7 NZ	3.95
	C1	His 119 ND1 (A conf)	2.90	C1	His 119 CE1 (A conf)	3.63	O2	HOH 19 C	2.72	C4	HOH 238 C	3.44	O2	HOH 16 C	3.64
DIO993	O2	Gln 11 OE1	2.91	O2	Gln 11 NE2	3.35	C3	HOH 19 D	3.23	C3	His 119 ND1 (A conf)	3.37	C1	Lys 41 CE	3.67
ETF903	O	Lys 7 NZ	3.45	F1	HOH 16 C	2.95	F2	HOH 19 C	2.72	F2	His 119 ND1 (A conf)	2.84	F3	His 119 ND1 (A conf)	2.87
	F3	His 119 CB	3.32	F3	Val 118 O	3.65									
IPA903	O	His 119 ND1	3.16	O	His 119 CE1 (A conf)	3.29	O	HOH 19 D	2.96	CA	Gln 11 OE1	3.64	CB2	Lys 41 CE	3.50
	CB2	Lys 41 NZ	3.83												
TBU903	O	Gln 11 NE2	3.49	O	Lys 7 NZ	3.52	O	His 119 ND1 (Conf B)	3.96	O	HOH 19 D	3.31	C2	His 119 CE1 (Conf A)	3.55
	C2	His 119 ND1 (Conf A)	3.53	C1	Gln 11 NE2	3.80	C1	Lys 41 CE	4.03						
DMS904	O	Asn 67 ND2	3.38	S	His 119 CG	3.57	S	His 119 CD2	3.37	C1	Gln 69 OE1	3.70			
DMS994	O	His 119 NE2	3.45	O	His 119 CE1 (A conf)	3.31	S	His 119 CB	3.91	S	Asn 67 ND2	3.83	S	Gln 69 OE1	3.94
	C2	Cys 65 SG	3.52	C2	Cys 65 CB	3.71	C2	Ala 109 CB	3.68	C1	His 119 CB	3.49	C1	Ala 109 CB	3.76

ETF901 is the second solvent in the 901 cluster that binds in the B₁ pocket of the active site of RNase A. F1 of ETF901 displaces Wat 60 and interacts with the backbone nitrogen and oxygen of Ser 123. Waters at this position are also displaced by a sulfate that binds in this pocket in the 1RNM structure, and by the phosphate of the inhibitor in the 1Z6D(B) structure. The sulfate O2 of the 1RNM structure overlaps the F1 of ETF901, and forms a hydrogen bond with the backbone nitrogen of Ser 123. The O3P of the inhibitor from the 1Z6D structure overlaps F1 and forms hydrogen bonds with the backbone nitrogen and O_γ of Ser 123. Additionally ETF901 serves as a partial shape mimic of the 1Z6D inhibitor as O1P and P of the 1Z6D inhibitor overlap F3 and C1 of ETF901, respectively. O of ETF901 forms a hydrogen bond with O_γ of Ser 123.

DMS902 binds in the B₁ pocket as part of a second cluster, 902, and displaces Wat 22, which is part of the water network in the active site of RNase A. The O of DMS902 forms a hydrogen bond with the backbone nitrogen of Thr 45 and if it occupied the current position of C1, it would form a hydrogen bond with O_{γ1} of Thr 45. The O2 and N3 groups of pyrimidine rings that bind in the B₁ pocket also form both of these interactions with Thr 45 (1WBU, 1W4Q, 1W4P, 1W4O, 1U1B, 1RUV, 1RPG, 1RPF, 1ROB, 1RNN, 1RNM, 1RCN, 1QHC, 1O0N, 1O0M, 1JVU, 1JN4, 1EOS). In addition to identifying interactions made by the pyrimidines made in this pocket, DMS902 serves as a shape mimic of part of the pyrimidine ring, with the S of DMS902 superimposing with the C2 group of the pyrimidine. The purine base in 1EOW, 1RBJ, 1RCA, 1RNC, 1RND, and 1Z6D is oriented in such a way that the O6 and N7 groups reproduce the interactions made by the O2 and N3 groups of the

pyrimidines, as does the carboxyl group (O12 and O13) of the inhibitor of 2G8R. A chloride ion from the 1RAR structures overlaps with the position of the O in DMS902, Wat 22, and the O2 group of the pyrimidines.

IPA902 is the second solvent molecule of the 902 cluster that binds in the B₁ pocket of the active site. As with DMS902, it displaces any water molecules located in position 22. The O group forms a hydrogen bond with the O γ 1 of Thr 45. The shape of IPA902 mimics some of the shape of the purines bound in the B₁ pocket. The O overlaps with the N7 group, the CA overlaps with the C5 group, and the CB1 overlaps with the C6 group. IPA902 does not mimic pyrimidines as well, but the CB1 and CB2 overlap with the C2 and C4, respectively, in pyrimidines.

The final solvent molecule in cluster 902 is ETF902, and as the rest of the solvents in the cluster, it displaces Wat 22. F1 and F2 both interact with O γ 1 of Thr 45 and F3 interacts with Val 43 O (both F and O are electronegative and this interaction is unlikely to be favorable). The C1 position of ETF902 picks out the position of C2 in pyrimidines and C5 in purines. The C2 of ETF902 picks out the C4 position in purines. Additionally, F2 of ETF902 overlaps the position of a chloride ion (2026) from 1RAR.

Collectively, the organic solvents bound in the B₁ subsite pick out key interactions made by the inhibitor molecules bound in this pocket. Hot spots for ligand binding in the B₁ pocket include the backbone nitrogen and O γ 1 of Thr 45, and the backbone nitrogen and O γ of Ser

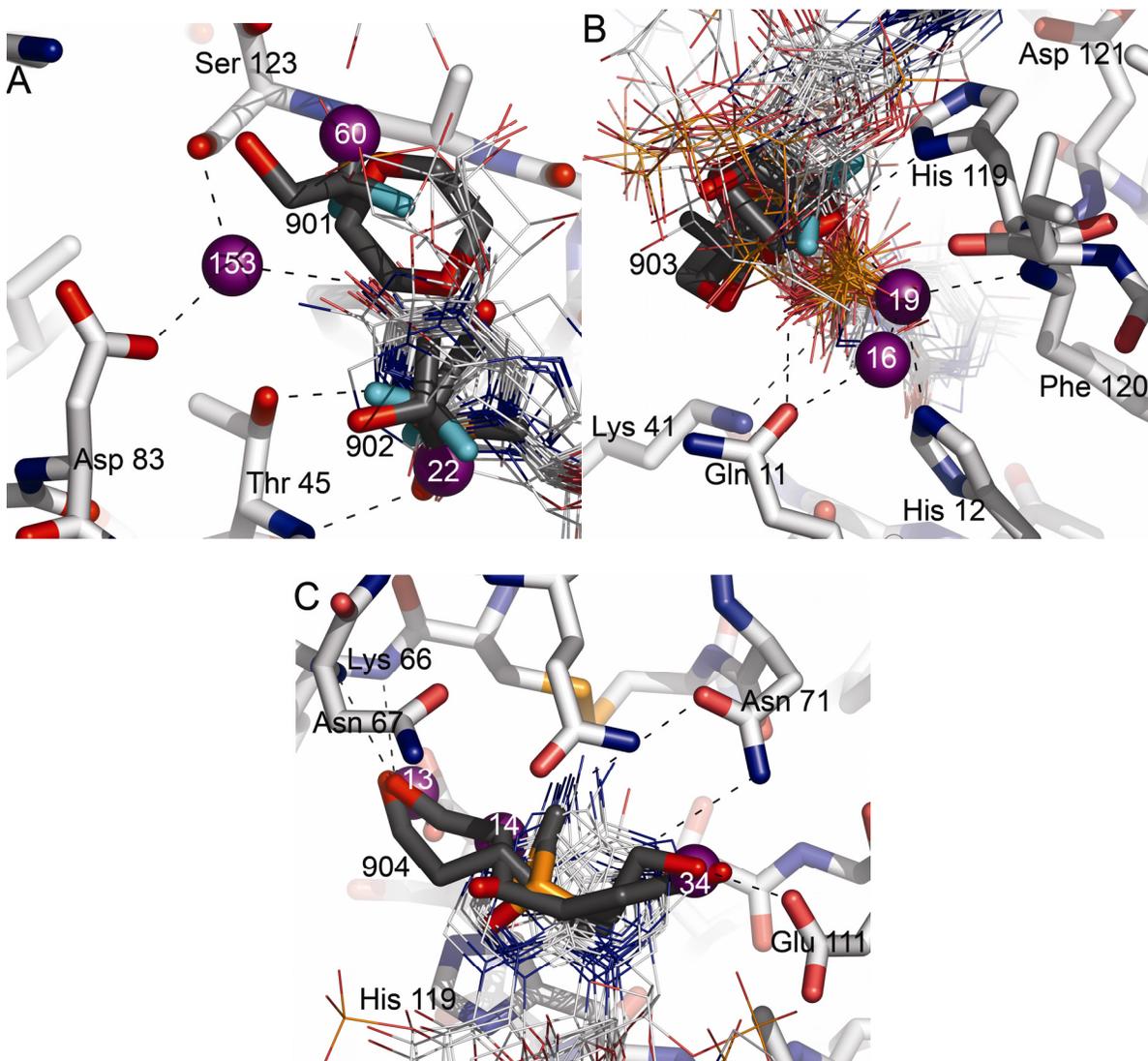


Figure 7. Organic solvent and inhibitor binding in the active site. RNase A is drawn with sticks with carbon atoms colored light gray, solvent molecules are drawn with sticks and carbon atoms colored dark gray. Representative conserved water molecules are colored purple. Inhibitor molecules are drawn as lines and colored with carbons as light gray. A black dashed line designates interactions between the inhibitors and water molecules or protein. A. The B₁ pocket B. The P₁ pocket C. The B₂ pocket

123, either directly or by a bridging water. Additionally, the bridging interaction made by Wat 153, Thr 45 O_γ1, Asp 83 O_δ1, and a bound ligand appears as an important interaction.

Solvents Bound in the P₁ Subsite

One cluster of organic solvents, 903, binds in the P₁ pocket of the active site of RNase A (Figure 7B). None of the solvents displace Wat 19, which bridges His12 and His 119, as the inhibitors bound in the P₁ subsite do, instead, they interact with waters bound in this position. Cluster 903 contains two dioxane molecules, DIO903 and DIO993, which are associated with molecules B and A, respectively, and are found in two different orientations in the P₁ pocket. In addition to the two dioxanes, these clusters contain ETF903, IPA903, and TBU903. These solvents make a number of interactions with residues of this pocket, such as His 119 N δ 1 and Gln 11 N ϵ 2.

ETF903 is part of the 903 cluster found in the P₁ pocket of the active site of RNase A. The F1 and F2 groups interact with Wat 16 and Wat 19, respectively. These water molecules form part of a hydrogen-bonding network in the active site and link the two sides of the cleft. The F2 group also interacts with the N δ 1 group of His 119. The O of ETF903 overlaps with the O1P group of the 1O0F(B) ligand on the phosphate which hydrogen bonds with the N ζ group of Lys 7 of the P₂ pocket. (If there were a phosphate/sulfate bound in the P₁ pocket pulling the two lobes closer together, the interaction between the O of ETF903 and Lys 7 N ζ would be much more distinct, in contrast to the current distance of 3.45Å.) Additionally, C1 of ETF903 picks out the 5' carbon of the ribose following the phosphate bound in the P₁ pocket.

The second solvent of cluster 903 is DIO903, and it hydrogen bonds with the Wat 19, which then hydrogen bonds with N ϵ 2 of His 12 and the backbone nitrogen of Phe 120. This solvent molecule picks out the location of inhibitor binding, but there is not much atom overlap because the inhibitors at this point in the active site do not superimpose well as a result of differences among the inhibitors (e.g. a single phosphate or double phosphate groups in between R1 and R2). An oxygen in the C1 position of DIO903 would be a good mimic of a phosphate-bound oxygen interacting with N δ 1 of His 119 in the A conformation. The O2 atom of DIO903 forms a hydrogen bond with the N δ 1 group of His 119 in conformation B. This is the same interaction made by a phosphate oxygen of an inhibitor when a phosphate is bound in the P₁ pocket and His 119 is found in the B conformation (e.g. 1O0O(A), 1RPF). Additionally, if O1 or O2 were located in the current position of C1, it would form a hydrogen bond of the N δ 1 group of His 119 when in conformation A, as would be observed when a phosphate group is bound in the P₁ pocket and His 119 adopts conformation A.

The second dioxane of the 903 cluster in the P₁ pocket is DIO993. The O2 of DIO993 forms a hydrogen bond with both the O ϵ 1 and the N ϵ 2 of Gln 11 in the P₁ pocket. A similar interaction is seen when a phosphate is bound in the P₁ pocket: a phosphate oxygen hydrogen bonds with the N ϵ 2 of Gln 11. Additionally, if either O1 or O2 were positioned in the current location of C3 in DIO993, there would be hydrogen bonds formed with the N δ 1 of His 119 in the A conformation and with the Wat 19. Wat 19 would then bridge the interaction between DIO993 and His 12. A phosphate bound in the P₁ pocket displaces Wat

19 and would form a hydrogen bond with Nε2 of His 12, while another oxygen of this group would interact with His 119 as illustrated with DIO993.

IPA903 is the fourth solvent molecule part of cluster 903. Through a hydrogen bond with Wat 19, the O of IPA903 interacts with Nε2 of His 12 and the backbone nitrogen of His 119. The O group of IPA903 overlaps with an oxygen of the phosphate of inhibitors bound in the P₁ pocket. This oxygen of the phosphate will form a hydrogen bond with the Nδ1 of His 119 when the distance and angle are appropriate. (The O of IPA903 suggests this interaction, but the angle is not optimal.) Another oxygen of the phosphate overlaps with Wat 19 and hydrogen bonds with Nε2 of His 12 and the backbone nitrogen of His 119.

The final solvent in cluster 903 is TBU903. As seen with the other solvents bound in the P₁ pocket, it does not displace any conserved water molecules from the active site, but instead forms a hydrogen bond with Wat 19. This water molecule forms additional hydrogen bonds with Wat 16, the backbone nitrogen of Phe 120, and the Nε2 group of His 12. His 119 is found in both the A and B conformations in the TBU MSCS structures. The O of TBU903 overlaps with the O3A of the inhibitor from 1QHC (molecule B). The O3A group forms a hydrogen bond with Lys 7 Nζ, which the O of TBU903 is positioned to do (though a little long in this structure at 3.52Å). Additionally, the phosphate displaces Wat 19 and reproduces the interactions made by a water molecule in this position by the oxygen groups on the phosphate.

The organic solvents in cluster 903 identify important groups in the P₁ site residues (and a P₂ residue with Lys 7) that are important for inhibitor or substrate binding. They mimic the interactions made by some of the phosphate oxygens with His 119 Nδ1 and Gln 11 Nε2. However, the solvents couldn't displace conserved Wat 19, which makes a number of hydrogen bonds with the surrounding protein (the backbone nitrogen of Phe 120, and Nε2 of His 12) and bridges the two lobes of RNase A. Its importance is highlighted by the comparison of the solvent binding and the inhibitor binding. This water is observed to be displaced by a phosphate or sulfate oxygen, highlighting the specificity of the P₁ pocket. Instead of displacing the water molecule from this critical position, the organic solvents interact with it.

Solvents Bound in the B₂ Subsite

Two 1,6-hexanediol molecules and two dimethylsulfoxide molecules bind in the B₂ subsite of the RNase A active site (Figure 7C). Because these molecules bind in approximately the same orientation as the other member of its pair, they are grouped in a single cluster, 904, but to distinguish the members of the pair, solvent molecules associated with protein molecule B are designated 994. As seen with the previous two subsites, solvents in cluster 904 interact with a number of residues in the B₂ pocket, such as Asn 67 Nδ1, Asn 71 Nδ2, and stacking with His 119.

DMS904 and DMS994 both bind in the B₂ pocket of the active site in nearly the same position and in orientations that are rotated 180° about an axis similar to the S-O bond. Both

these solvent molecules stack on top of His 119, which is only found in the A conformation in the DMS MSCS structures, and displace Wat 14. The O group of DMS904 forms a hydrogen bond with the N δ 1 of Asn 67, and interaction also made by the sulfate group of 5-(1-sulfonaphthyl)-acetylamino-ethylamine in the 1RAS and 1RAR inhibitor-bound structures. C1 and C2 interact with hydrophobic areas of the B₂ pocket, such as Cys 65 CB and Ala 109 CB. Additionally, in DMS904, C1 and C2 overlap with nine and thirteen carbon atoms in inhibitor molecules, respectively, and for DMS994, there are eight and nine overlaps of inhibitor carbon atoms by C1 and C2.

HEZ904 and HEZ994 both bind in the B₂ pocket of the active site of molecule A and B, respectively. These two molecules bind in overlapping positions, make similar interactions with residues in the B₂ pocket, and displace water molecules from three conserved binding sites. The O6 group of the HEZ molecules displaces Wat 13 and replaces the interactions these water molecules would make by forming hydrogen bonds with the backbone nitrogens of Asn 67 and Lys 66, and with the O δ 1 group of Asp 121. Additionally, the HEZ molecules displace Wat 14, which hydrogen bonds with Wat 13, and are displaced by inhibitors binding in the B₂ pocket. Inhibitor molecules hydrogen bond with Wat 13, and are not found to displace water molecules from this position. The importance of Wat 13 is highlighted by its displacement by HEZ molecules, which replace the hydrogen bonds formed by Wat 13. At the opposite end of the HEZ molecules, the O1 group displaces Wat 34 and replaces the interactions by hydrogen bonding with the N δ 2 group of Asn71 and the O ϵ 2 group of Glu 111. His 119 is found in only the A conformation in the 1,6-hexanediol MSCS structures,

and the HEZ molecules stack on top of this side chain as is observed by the inhibitors bound in this pocket.

The four molecules bound in the B₂ subsite collectively pick out key interactions made by inhibitor molecules bound in this pocket. Hot spots for ligand binding in the B₂ pocket include stacking on the His 119 imidazole ring, an interaction with Asn 71 Nδ2, and less commonly, an interaction with Glu 111 Oε2. The 904/994 solvents miss the hydrogen bond made between Asn 71 Oδ1 and the inhibitors. Both HEZ904 and HEZ994 displace and replace the interactions of Wat 13, which bridges the interaction between the inhibitors and the backbone nitrogens of Lys 66 and Asn 67, and the Oδ1 group of Asp 121, thus highlighting the role this water plays in ligand binding.

Additional Active Site Solvent

ETF923 is located between the P₁ and B₂ pocket of the active site, however, it does not displace any conserved waters or overlap any inhibitors. F2 of ETF923 interacts with Nδ2 of Asn 67. Asn 67 is part of the loop that covers the B₂ pocket and has higher flexibility because of this; however, the side chain adopts much the same conformation for the majority of structures (only four structures do not see Asn 67 in this similar conformation).

Solvents in Crystal Contacts

There are 17 binding sites of organic solvents on the surface of RNase A that do not fall in the active site of the protein, but are located in areas of crystal contact. Most do not displace

conserved water molecules or overlap with inhibitors, however DMS907B and RSF907B overlap a citrate molecule, and RSF919B displaces Wat 5. The rest of the solvents include the final large cluster, containing DMF905, DMS905, IPA905, ETF905, and TMO905, and a number of smaller clusters or single solvents including: DMS907A, RSF907A, DMS910 (A and B), DMS911, DMS913, HEZ915, HEZ916, IPA917, RSF917, RSF918, RSF919A, RSF920, RSF921, and ETF922 (A and B).

Other Surface Solvents

There are five organic solvent molecules that bind the surface of RNase A outside of the active site that do not bind in areas of crystal contact. These solvents tend to pick out areas of the RNase surface that exhibit low plasticity, and sometimes pick out hinge residues. None of these solvents have overlaps with bound inhibitor molecules or displace conserved waters. These solvents are found isolated, rather than in clusters on the protein surface, and are often near a crystal contact.

DIO908 is located outside of the active site and hydrogen bonds with the terminal backbone oxygen of Val 124 and the backbone oxygen of Asn 103 through the O1 atom and O2 atom, respectively. Considering all the MSCS and the inhibitor-bound structures, this small, shallow pocket on the surface has low plasticity and is part of the hinge region. There is some variation in the terminal COOH and in the side chain of Lys 104, in fact, Lys 104 adopts a conformation that is less sampled among all the structures. In this conformation, the N ζ atom could form a hydrogen bond with DIO908 when either O1 or O2 occupies the

current location of the C3 atom. DIO 908 does not occupy a binding site of either an inhibitor or a conserved water molecule.

DIO909 is located outside the active site alongside helix 1 in an area of low plasticity. There are few interactions made by this solvent molecule, but there is a potential hydrogen bond (distance is good but the angle is weird) with the O ϵ 2 of Glu 9. Additionally, it neither displaces water molecules bound in conserved binding sites, nor overlaps binding sites observed with inhibitors.

DMS912 is located outside the active site and is not located in an area of crystal contact. The O group forms a hydrogen bond with the N ϵ 2 group of Gln 55. If the O group was positioned in the current location of C2, it would hydrogen bond with Wat 282, which was not found to be conserved in the MSCS structures, but was previously identified as conserved (Sadasivan et al., 1998b; Zegers et al., 1994). This water molecule then hydrogen bonds with the backbone nitrogen of Ala 52. DMS912 is found in an area of low plasticity (52-53, 55) along helix 3.

DMS914 binds RNase A approximately 2Å away from DIO908 in a hinge region with low plasticity. The O group forms a hydrogen bond with the backbone nitrogen of Asn 103. If the O group occupies the current C1 position, it could potentially form a hydrogen bond with the backbone oxygen of Asn 103.

TMO906 interacts with Gln 28 on the face of RNase A on the opposite side of the active site in domain B. It overlaps with no inhibitors, and displaces no waters. Gln 28 is defined as a rigid residue (Sadasivan et al., 1998b), but the side chains of this and other local residues (Asn 24, Lys 31) are flexible.

Discussion

The MSCS method is a powerful experimental tool that can be used to ascertain the fundamental properties of protein binding sites. In addition to providing information about binding sites, the analysis of a set of superimposed structures of a protein in high concentrations of organic solvents illustrates patterns of hydration and plasticity on the entire surface of the protein. This study brings the MSCS analysis of these properties a quantitative aspect. Bovine pancreatic RNase A is a good model for the development and assessment of these methods because it is a well studied protein and has hundreds of structures available in the Protein Data Bank. All of this previous work allows for comparison with the MSCS results of the plasticity, hydration, and organic solvent binding on the surface of RNase A.

Protein Plasticity

Previous studies of RNase A have identified a number of areas of protein plasticity. First, there is the overall domain motion observed in relation to an inhibitor binding in the active site or a sulfate or phosphate ion binding in the P₁ pocket. Second, the loops in between elements of secondary structure exhibit a greater range of motion than residues involved in the α -helices or the β -sheet. This effect increases as the distance from the hinge increases.

Third, individual residues exhibit greater plasticity than the majority of the protein, owing either to their solvent accessibility or functionality in the active site.

By soaking crystals in a variety of organic solvents and comparing the structures, a greater range of conformations is observed in the protein than would be seen in a single crystal structure. This is easily seen visually once the structures are superimposed. As seen in earlier work on RNase A, the residues exhibiting increased plasticity compared to the rest of the protein tended to fall in loop regions, and in those residues with greater solvent accessibility. Additionally, certain active site residues displayed increased plasticity. To effectively compare the plasticity observed in the MSCS structures to that of previous work, RMSD calculations were introduced. For each pair of structures in a set, the RMSD was calculated by residue. The lowest value, the highest value, and the average of all values for each residue were used to illustrate quantitatively the plasticity each residue demonstrates.

To evaluate the MSCS method and the RMSD calculations for quantitating plasticity, three previously studied sets of structures were used for comparison. These structures were analyzed using different methods. The structure of RNase A determined by NMR contained 32 models, which were compared using the RMS differences for the backbone Φ and Ψ angles (Santoro et al., 1993). The Sadasivan study (Sadasivan et al., 1998b) examined ten molecules using a computational method calculating the RMS of each residue using the C_{α} atom. Out of the ten molecules examined in the Zegers study (Zegers et al., 1994), only the three newly solved structures were visually inspected for structural differences. When each

of these sets of structures was analyzed using the RMSD calculations as used for the MSCS structures, the results reflected the observations made in the previous studies. Additionally, when these results were compared to the MSCS results, the trends, if not the magnitude, were the same. Where there were differences in the trends, it was a matter of differences in crystal contacts. This illustrates that soaking crystals in organic solvents causes a sampling of similar conformations as observed in structures from crystals in aqueous solution and plasticity follows the qualitative trends observed by NMR.

As a further test of the method, two additional sets of structures were included. Fifty-eight inhibitor-bound molecules of RNase A and a subset of 36 molecules belonging to the C2 space group were analyzed with the same method. The comparison with the MSCS structures again presented similar trends, further illustrating that MSCS in organic solvents exhibits what is observed in aqueous solution in crystals with the differences explained by crystal contacts. More importantly, qualitative trends are trends seen in solution NMR structures.

Hydration

Compared to the number of structures of RNase A that are available, there are relatively few analyses of the hydration of RNase A involving more than a few structures. Both the Zegers and the Sadasivan studies identified a number of conserved water molecules associated with their respective ten molecules of RNase A (Sadasivan et al., 1998b; Zegers et al., 1994). Fourteen water molecules were identified as invariant in the Sadasivan study, which is

composed of structures primarily from the $P2_1$ space group. Waters were considered to be invariant when there is at least one interaction of 3.6 Å or less with the protein in common for every structure, and after superposition, all homologous water molecules are no further than 1.8 Å apart (Kishan et al., 1995; Sadasivan et al., 1998b). The Zegers study, with only two structures in common with the Sadasivan study and which is composed of structures from four space groups, identified 17 waters as conserved. Using the FIXWAT program (Lisgarten et al., 1993), water molecules were identified as conserved when they were located within 5 Å of the protein and waters from all the structures clustered in a sphere with a radius of 0.5 Å (Zegers et al., 1994). Many of these waters are involved in stabilizing the tertiary structure of RNase A and ten of these waters are identified by both studies.

To determine which water molecules in the MSCS structures were conserved, the SEWS program (Bottoms et al., 2006) was used with a 3.4 Å interaction distance between a water molecule and the protein, and a sphere of radius 1.4 Å. Water molecules were considered conserved if they fell within the same sphere and had at least one common interaction with the protein in 80% of the structures. These parameters were varied to find those that were optimal and it was found that an interaction distance of less than 3.4 Å missed water molecules, where a greater value didn't significantly improve the performance of the program. Altering the sphere radius did not change the results drastically because water molecules needed to have at least one common interaction with the protein to be considered conserved; however, a radius of 1.4 Å prevented two water molecules from being grouped in the same sphere and limited the number of missed water molecules. SEWS was used to

identify conserved water molecules from the two sets of ten molecules used in the previous studies. For the Sadasivan set of structures, all previously identified conserved water molecules were found to be 100% conserved, with 31 total water molecules at least 80% conserved. With 24 water molecules identified at 80% conservation or greater, SEWS finds all the waters previously identified by the Zegers study (Zegers et al., 1994), except for Wat 144. Wat 144 was identified as only 50% conserved. Unlike the results with the Sadasivan structures, not all previously identified waters from the Zegers set were found to be 100% conserved. This is probably due to the fact that waters in the Zegers study were considered conserved if they were missing from covalently modified structures near the modified region, and to the requirement in our analysis for at least one common interaction with the protein. Additionally, SEWS analysis was run on the 56 inhibitor-bound structures (two of the original set did not contain water molecules and were removed) and on the subset of 36 structures in the C2 space group. Eighteen water molecules were identified as at least 80% conserved in the inhibitor-bound set and 22 in the C2 inhibitor-bound subset.

The sets of conserved water molecules were compared with those identified from the MSCS structures. We expected the sets of structures with the same space group, the MSCS structures and the inhibitor-bound structures in the C2 space group, to produce the same conserved waters, particularly outside of the active site where they wouldn't be displaced by bound inhibitor molecules. Indeed, the conserved waters identified from the C2 subset of the inhibitor-bound structures overlap closely with those identified in the MSCS structures. Only three water molecules from the C2 subset were missed in the MSCS structures.

Interestingly, the MSCS conserved waters only miss two of the waters identified in the full inhibitor-bound set of structures, however, the C2 subset is more similar with the MSCS waters, picking out three waters missing in the full inhibitor-bound set. The conserved waters from the Sadasivan set of ten structures that are primarily of one space group, which is different to that of the MSCS structures, had the greatest difference compared to the conserved waters identified in the MSCS structures. Twelve of the waters identified from this set did not have a corresponding conserved water in the MSCS structures. Finally, the conserved waters identified from the Zegers set of ten structures had eight waters that did not have a corresponding MSCS water. The differences in conserved waters in these two sets of structures can be explained by differences in crystal contacts.

These results illustrate that even though the protein crystals are soaked with organic solvents, MSCS reveals the same conserved waters as would be seen if the crystal were in aqueous solution. Interestingly, the conserved waters from the MSCS structures are the most similar to those identified in the large set of inhibitor-bound structures, both of the same space group as MSCS and of all space groups available.

Binding of Organic Solvents

As seen with computational solvent mapping, where most probe molecules are found in a number of conformations (Dennis et al., 2002; Silberstein et al., 2003), many of the solvent molecules (e.g. DIO901, DMS 902) could be modeled in more than one conformation, allowing the solvents to sample a number of interactions in a given binding site. This

rotational freedom makes organic solvent molecules more favorable probes for protein binding sites than a larger ligand, which must form a greater number of interactions with the binding site. Additionally, the clustering of organic solvents highlights important hot spots on protein surfaces. The organic solvents cluster in the three most distinct subsites of the active site of RNase A: B₁ and B₂, which differentially bind a nucleotide base, and P₁, which binds a phosphate moiety and is the site of catalysis. Each of these subsites contains a hot spot for organic solvent binding and these hot spots can be defined by the key interactions they highlight. The hot spot in the B₁ subsite is characterized by hydrogen bonding with the backbone nitrogen and the O_{γ1} of Thr 45, hydrogen bonding with the backbone nitrogen and the O_γ of Ser 123, either directly or through a bridging water molecule (Wat 60 in MSCS structures), and a bridging interaction with Wat 153 connecting to Ser 123 O_γ, Asp 83 O_{δ1}, and Thr 45 O_{γ1}. Interactions with Wat 16, and hydrogen bonding with His 119 N_{δ1}, Gln 11 N_{ε2}, and His 12 N_{ε2} characterize the P₁ subsite hot spot. Additionally, if Wat 19 cannot be removed by a phosphate moiety, hydrogen bonding with this water is another characteristic of this hot spot. Finally, the interactions describing the hot spot in the B₂ subsite are stacking with the imidazole ring of His 119, hydrogen bonding with Asn 67 N_{δ1} and Asn 71 N_{δ1}, and hydrogen bonding with the backbone nitrogens of Asn 67 and Lys 66 either directly or through a bridging interaction with Wat 13.

The interactions delineating these hot spots are further highlighted when the MSCS structures are compared with the inhibitor-bound structures of RNase A. When inhibitors are bound in each of these hot spots, the same interactions made by the organic solvents are made by the

inhibitors, often with functional group overlap of 1Å or less. The comparison of the MSCS and the inhibitor-bound structures also highlights that inhibitor design is not a simple process of connecting the dots between the organic solvents. However, this does not decrease the strength of the information gathered from the organic solvents clustering in hot spots. The organic solvents in MSCS identify the positions of functional groups and key interactions that are present over and over again in the inhibitor-bound structures of RNase A. Any inhibitor designed should strive to reproduce all the key interactions defining the hot spots in the three subsites of the RNase A active site.

Conclusions

With MSCS, a number of crystal structures are produced and when analyzed collectively, the result is a clearer, more complete picture of the protein than could be illustrated with a single structure alone. While the crystals are soaked in organic solvents, the results presented here for RNase A show that MSCS produces plasticity, hydration, and active site interaction information relevant to structures in aqueous solution. The plasticity observed in these structures reflects the trends repeatedly observed in both aqueous crystal and solution NMR structures. Conserved water molecules in aqueous structure correspond to conserved water molecules in MSCS structures. And, organic solvents cluster in hot spots defined by interactions observed repeatedly in the inhibitor-bound structures. The information discussed here could potentially be used to improve inhibitors of RNase A, and when the MSCS method is applied to a lesser-studied protein, important interactions in the active site and properties of the protein as a whole would be unveiled.

Acknowledgements

We thank Senthil Kumar for his efforts to reproduce crystals. We appreciate all the discussions and assistance, which helped start the computational analysis, provided by Janet Thornton and Roman Laskowski. Christopher Bottoms generously provided the SEWS program and assistance in using the software. Data were collected at the Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Use of the Advanced Photon Source was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. W-31-109-Eng-38. We are grateful to the staff at the SER-CAT beamline for their guidance during data collection. This research was supported by a grant from the National Science Foundation under grant number 0237297.

References

- Allen, K.N., Bellamacina, C.R., Ding, X., Jeffery, C.J., Mattos, C., Petsko, G.A., and Ringe, D. (1996). An experimental approach to mapping the binding surfaces of crystalline proteins. *Journal of Physical Chemistry* *100*, 2605-2611.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic acids research* *28*, 235-242.
- Borkakoti, N., Moss, D.A., and Palmer, R.A. (1982). Ribonuclease A: Least squares refinement of structure at 1.45 Å resolution. *Acta Crystallographica B* *38*, 2210-2217.
- Bottoms, C.A., White, T.A., and Tanner, J.J. (2006). Exploring structurally conserved solvent sites in protein families. *Proteins* *64*, 404-421.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., *et al.* (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* *54*, 905-921.

D'Alessio, G. (1993). New and cryptic biological messages from RNases. *Trends in cell biology* *3*, 106-109.

DeLano, W.L. The PyMOL Molecular Graphics System (Palo Alto, CA, USA, DeLano Scientific). <http://www.pymol.org>

Dennis, S., Kortvelyesi, T., and Vajda, S. (2002). Computational mapping identifies the binding sites of organic solvents on proteins. *Proc Natl Acad Sci U S A* *99*, 4290-4295.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* *60*, 2126-2132.

English, A.C., Done, S.H., Caves, L.S., Groom, C.R., and Hubbard, R.E. (1999). Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins* *37*, 628-640.

English, A.C., Groom, C.R., and Hubbard, R.E. (2001). Experimental and computational mapping of the binding surface of a crystalline protein. *Protein engineering* *14*, 47-59.

Fitzpatrick, P.A., Ringe, D., and Klibanov, A.M. (1994). X-ray crystal structure of cross-linked subtilisin Carlsberg in water vs. acetonitrile. *Biochemical and biophysical research communications* *198*, 675-681.

Fontecilla-Camps, J.C., de Llorens, R., le Du, M.H., and Cuchillo, C.M. (1994). Crystal structure of ribonuclease A.d(ApTpApApG) complex. Direct evidence for extended substrate recognition. *J Biol Chem* *269*, 21526-21531.

Gilliland, G.L., Dill, J., Pechik, I., Svensson, L.A., and Sjolín, L. (1994). The active site of bovine pancreatic ribonuclease: an example of solvent modulated specificity. *Protein and Peptide Letters* *1*, 60-65.

Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A., and Thornton, J.M. (2006). A method for localizing ligand binding pockets in protein structures. *Proteins* *62*, 479-488.

Jencks, W.P. (1981). On the attribution and additivity of binding energies. *Proc Natl Acad Sci U S A* *78*, 4046-4050.

Jones, S., and Thornton, J.M. (2004). Searching for functional sites in protein structures. *Current opinion in chemical biology* 8, 3-7.

Kishan, R.V., Chandra, N.R., Sudarsanakumar, C., Suguna, K., and Vijayan, M. (1995). Water-dependent domain motion and flexibility in ribonuclease A and the invariant features in its hydration shell. An X-ray study of two low-humidity crystal forms of the enzyme. *Acta Crystallogr D Biol Crystallogr* 51, 703-710.

Klink, T.A., Woycechowsky, K.J., Taylor, K.M., and Raines, R.T. (2000). Contribution of disulfide bonds to the conformational stability and catalytic activity of ribonuclease A. *European journal of biochemistry / FEBS* 267, 566-572.

Laurie, A.T., and Jackson, R.M. (2006). Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Current protein & peptide science* 7, 395-406.

Leonidas, D.D., Chavali, G.B., Oikonomakos, N.G., Chrysina, E.D., Kosmopoulou, M.N., Vlassi, M., Frankling, C., and Acharya, K.R. (2003). High-resolution crystal structures of ribonuclease A complexed with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors. *Protein Sci* 12, 2559-2574.

Leonidas, D.D., Maiti, T.K., Samanta, A., Dasgupta, S., Pathak, T., Zographos, S.E., and Oikonomakos, N.G. (2006). The binding of 3'-N-piperidine-4-carboxyl-3'-deoxy-ara-uridine to ribonuclease A in the crystal. *Bioorganic & medicinal chemistry* 14, 6055-6064.

Leonidas, D.D., Shapiro, R., Irons, L.I., Russo, N., and Acharya, K.R. (1997). Crystal structures of ribonuclease A complexes with 5'-diphosphoadenosine 3'-phosphate and 5'-diphosphoadenosine 2'-phosphate at 1.7 Å resolution. *Biochemistry* 36, 5578-5588.

Leonidas, D.D., Shapiro, R., Irons, L.I., Russo, N., and Acharya, K.R. (1999). Toward rational design of ribonuclease inhibitors: high-resolution crystal structure of a ribonuclease A complex with a potent 3',5'-pyrophosphate-linked dinucleotide inhibitor. *Biochemistry* 38, 10287-10297.

Lisgarten, J.N., Gupta, V., Maes, D., Wyns, L., Zegers, I., Palmer, R.A., Dealwis, C.G., Aguilar, C.F., and Hemmings, A.M. (1993). Structure of the crystalline complex of cytidylic acid (2'-CMP) with ribonuclease at 1.6 Å resolution. Conservation of solvent sites in RNase-A high-resolution structures. *Acta Crystallogr D Biol Crystallogr* 49, 541-547.

Mattos, C. (2002). Protein-water interactions in a dynamic world. *Trends in biochemical sciences* 27, 203-208.

- Mattos, C., Bellamacina, C.R., Peisach, E., Pereira, A., Vitkup, D., Petsko, G.A., and Ringe, D. (2006). Multiple solvent crystal structures: probing binding sites, plasticity and hydration. *Journal of molecular biology* 357, 1471-1482.
- Mattos, C., and Ringe, D. (1996). Locating and characterizing binding sites on proteins. *Nature biotechnology* 14, 595-599.
- Mattos, C., and Ringe, D. (2001). Solvent Structure. *International Tables of Crystallography* (Rossman, M G, Arnold, E Eds) *Kluwer Academic Publishers, Vol F*, 623-640.
- Milburn, D., Laskowski, R.A., and Thornton, J.M. (1998). Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Engineering, Design and Selection* 11, 855-859.
- Miranker, A., and Karplus, M. (1991). Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* 11, 29-34.
- Otwinowski, Z., and Minor, W. (1997). Processing of x-ray diffraction data collected in oscillation mode. *Methods in enzymology* 276, 307-326.
- Raines, R.T. (1998). Ribonuclease A. *Chem Rev* 98, 1045-1066.
- Rees, D.C., Congreve, M., Murray, C.W., and Carr, R. (2004). Fragment-based lead discovery. *Nature reviews* 3, 660-672.
- Ringe, D., and Mattos, C. (1999). Analysis of the binding surfaces of proteins. *Medicinal research reviews* 19, 321-331.
- Sadasivan, C., Nagendra, H.G., and Vijayan, M. (1998a). Plasticity, Hydration and Accessibility in Ribonuclease A. The Structure of a New Crystal Form and its Low-Humidity Variant. *Acta Crystallographica D* 54, 1343-1352.
- Santoro, J., Gonzalez, C., Bruix, M., Neira, J.L., Nieto, J.L., Herranz, J., and Rico, M. (1993). High-resolution three-dimensional structure of ribonuclease A in solution by nuclear magnetic resonance spectroscopy. *Journal of molecular biology* 229, 722-734.
- Sheu, S.H., Kaya, T., Waxman, D.J., and Vajda, S. (2005a). Exploring the binding site structure of the PPAR gamma ligand-binding domain by computational solvent mapping. *Biochemistry* 44, 1193-1209.
- Sheu, S.H., Lancia, D.R., Jr., Clodfelter, K.H., Landon, M.R., and Vajda, S. (2005b). PRECISE: a Database of Predicted and Consensus Interaction Sites in Enzymes. *Nucleic acids research* 33, D206-211.

Shuker, S.B., Hajduk, P.J., Meadows, R.P., and Fesik, S.W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science* (New York, NY) *274*, 1531-1534.

Silberstein, M., Dennis, S., Brown, L., Kortvelyesi, T., Clodfelter, K., and Vajda, S. (2003). Identification of substrate binding sites in enzymes by computational solvent mapping. *Journal of molecular biology* *332*, 1095-1113.

Silberstein, M., Landon, M.R., Wang, Y.E., Perl, A., and Vajda, S. (2006). Computational methods for functional site identification suggest a substrate access channel in transaldolase. *Genome informatics* *17*, 13-22.

Stultz, C.M., and Karplus, M. (2000). Dynamic ligand design and combinatorial optimization: designing inhibitors to endothiapepsin. *Proteins* *40*, 258-289.

Vitagliano, L., Merlino, A., Zagari, A., and Mazzeola, L. (2000). Productive and nonproductive binding to ribonuclease A: X-ray structure of two complexes with uridylyl(2',5')guanosine. *Protein Sci* *9*, 1217-1225.

Wlodawer, A., Miller, M., and Sjolín, L. (1983). Active site of RNase: neutron diffraction study of a complex with uridine vanadate, a transition-state analog. *Proc Natl Acad Sci U S A* *80*, 3628-3631.

Zegers, I., Maes, D., Dao-Thi, M.H., Poortmans, F., Palmer, R., and Wyns, L. (1994). The structures of RNase A complexed with 3'-CMP and d(CpA): active site conformation and conserved water molecules. *Protein Sci* *3*, 2322-2339.

CHAPTER 3: Multiple Solvent Crystal Structures of Ribonuclease A: A Comparison of the P3₂21 and C2 Spacegroups

Abstract

The Multiple Solvent Crystal Structures (MSCS) method uses organic solvents to map the surfaces of proteins. In addition to identifying binding sites on the surfaces of proteins, this method allows for a more thorough examination of protein plasticity and hydration than could be achieved by a single crystal structure. Generally, this method is applied to crystals of a protein grown in a single space group, but for Ribonuclease A (RNase A), solvent mapping results were obtained for crystals in two different space groups, P3₂21 and C2, presenting a unique opportunity for comparison. The crystal structures in the P3₂21 space group of RNase A soaked in cyclopentanol, cyclopentanone, cyclohexanol, cyclohexanone, dioxane, dimethylformamide, dimethylsulfoxide, 1,6-hexanediol, 2-propanol, S,R,S-bisfuran alcohol, and trifluoroethanol were compared to the solvent soaked crystal structures in the C2 space group, which are presented in Chapter 2. In addition to the solvent molecule probes, each P3₂21 MSCS structure has a sulfate anion bound in the catalytic site of RNase, which simulates the phosphate moiety that normally binds at the location occupied by the sulfate. This ion serves as an additional molecular probe of RNase A and provides a mimic of the inhibitor-bound state of RNase with which to compare to the apo- state found in the C2 MSCS structures and examine the range of molecular hinge opening observed in the MSCS crystals of RNase A. Quantitative analysis of the two sets of crystal structures reveals similar plasticity trends that are modulated by crystal contacts; conserved water binding positions common to both sets of MSCS structures are found both in the active site and

bridging elements of secondary structure. And, a comparison of the bound molecular probes in the active site adds to the observation of hot spots for substrate binding discussed in Chapter 2.

Introduction

Over the past decade, the Multiple Solvent Crystal Structures (MSCS) method has been developed as a powerful experimental method to locate and characterize binding sites of proteins (Allen et al., 1996; Mattos, 2002; Mattos et al., 2006; Mattos and Ringe, 1996). Binding sites are identified by the clustering of organic solvents on the surface of the protein, as is observed with the complementary *in silico* technique, computational solvent mapping (CS-Map; (Dennis et al., 2002; Sheu et al., 2005; Silberstein et al., 2003; Silberstein et al., 2006). In addition to locating binding sites, the MSCS method produces a set of structures, in which protein plasticity and hydration can be examined. Until recently, MSCS data, including the binding of organic solvents, protein plasticity, and hydration, were examined qualitatively. In the last chapter, the first quantitative analysis of MSCS structures was presented using Ribonuclease A (RNase A) as a model.

RNase A is a kidney shaped protein with two lobes, domain A and domain B, connected by a nine residue (14-15, 47-48, 80-81, and 102-104) hinge region. The structure is dominated by an anti-parallel β -sheet and also has three short α -helices. The β -sheet spans the two lobes of RNase A, and the portion contained in domain A is designated β_1 (residues 61-63, 71-75, 105-111, and 116-124) and the portion contained in domain B is designated β_2 (residues 42-

46, 82-87, and 96-101). Together, with the hinge region, β_1 and β_2 form a characteristic V-shaped motif (Vitagliano et al., 2002). Single-stranded RNA binds in the well-defined subsites of the active site cleft that lies in between β_1 and β_2 . The B_1 pocket binds pyrimidine nucleotides, which interact directly with Thr 45 and indirectly via bridging water molecules with Asp 83 and Ser 123. The 3' phosphate, which contains the scissile bond, binds in the P_1 pocket where it interacts with Lys 41, Gln 11, and the catalytic residues His 12 and His 119. The subsequent nucleotide base binds in the B_2 subsite. Figure 1 of Chapter 2 illustrates the overall arrangement of the active site subsites and the key interactions made between active site residues and substrates.

It was first suggested over a decade ago that enzyme flexibility was required for the catalytic function of RNase A (Rasmussen et al., 1992) and since then work has been done to examine the flexibility of the protein. The two β -sheets participate in breathing or hinge motions, which have been observed in molecular dynamics simulations in addition to NMR and crystal structures (Beach et al., 2005; Kishan et al., 1995; Merlino et al., 2002; Sadasivan et al., 1998; Vitagliano et al., 2002), and this motion has been shown to be vital to substrate binding and release (Vitagliano et al., 2002). Additionally, it was suggested that RNase A uses this hinge in catalysis (Merlino et al., 2002). A substrate binds in the active site of RNase A and a conformational change occurs in which β_1 and β_2 , and consequently domain A and B, move closer and reduce the size of the active groove, creating a hydrophobic environment (Beach et al., 2005; Hammes, 2002). In the P_1 subsite, the transphosphorylation reaction proceeds followed by hydrolysis, resulting in the cleaved P-O^{5'} bond at the 3' side of

the pyrimidine nucleotide bound in the B₁ subsite (Raines, 1998). The product is released, which is likely facilitated by the relaxation of the hinge (Beach et al., 2005).

Initial work on the MSCS of RNase A produced crystals in the P₃₂₁ space group, which had a sulfate anion bound in the P₁ subsite of the active site cleft. Further analysis of these structures revealed that only two organic solvents were bound in the active site. Whether this was a result of the sulfate ion in the active site or the fact that the high salt content of the crystallization conditions prevented high concentrations of organic solvents from being used in the solvent soaks is unknown, but both of these reasons prompted efforts for new crystals to be grown in different conditions with no bound sulfate. This resulted in crystals of RNase A in the C₂ space group with an empty active site and successful soaks in high concentrations of organic solvents. The two sets of MSCS structures of RNase A present a unique opportunity to compare the MSCS of the same protein in two space groups, and examine the plasticity, hydration, and organic solvent binding patterns. In addition to the 20 MSCS structures in the C₂ space group, this work adds 12 MSCS structures in the P₃₂₁ space group, and compares these models to 17 structures in the P₃₂₁ space group downloaded from the PDB. Also, this work expands upon the analysis of the plasticity of RNase A by comparing the range of hinge conformations observed in the MSCS structures to all of downloaded crystal and NMR structures used in the examination of the C₂ and P₃₂₁ MSCS structures.

Materials and Methods

Crystal Growth, Cross-linking, and Solvent Soaks

RNAse A (type X1IA) from Sigma was dissolved into 20mM sodium phosphate pH 6, 20mM sodium acetate buffer with a final protein concentration of 35 mg/mL. As done previously (Zegers et al., 1994), crystals were grown at 18 °C and room temperature using the vapor diffusion method with 10 μ L drops containing half protein solution and half reservoir solution over a 500 μ L reservoir. The reservoir solution consisted of 35% ammonium sulfate and 1.5M sodium chloride. Crystals grew overnight.

RNAse A crystals were manually transferred with a cryo-loop to a 10 μ L drop of 0.04% glutaraldehyde (8.26 % w/v in distilled water, pH 4.25, Electron Microscopy Sciences) in stabilization buffer (0.15M HEPES, pH 7.2, 50% v/v PEG 400) over a 300 μ L reservoir (the same glutaraldehyde-stabilization buffer as the drop) and the cross-linking reaction was allowed to proceed at room temperature for 1 hour. Cross-linked crystals were then transferred using a cryo-loop to new drops containing stabilization buffer and an organic solvent (25% cyclopentanol, 40% cyclopentanone, 40% cyclohexanol, 40% cyclohexanone, 40% dioxane, 40% dimethylformamide, 25% dimethylsulfoxide, 50% 1,6-hexanediol, 40% 2-propanol, 50% S,R,S-bisfuran alcohol, or 40% trifluoroethanol) and allowed to soak for 1 hour at room temperature. Soaked crystals were then collected, cryo-protected by dunking in stabilization buffer containing 20% glycerol, and flash frozen in liquid nitrogen.

Data Collection, Processing, Structure Refinement

Diffraction data were collected at 100K at the SER-CAT ID-22 beamline at APS (Argonne, IL) using 1.0 Å wavelength radiation and a Mar300 CCD detector at a crystal to detector distance of 100 mm. The data were processed and scaled using HKL2000 (Otwinowski and Minor, 1997). Published models of RNase A (PDB code 1JVT, and 1RPH for the apo-RNase A and sulfate-bound RNase A, respectively) with the water molecules removed were used to calculate the initial electron density maps for all models of RNase A. The models were refined using CNS (Brunger et al., 1998). The program Coot (Emsley and Cowtan, 2004) was used to manually rebuild the models and to identify water and organic solvent molecules using the Fo-Fc electron density maps contoured at 3σ and the 2Fo-Fc electron density difference maps contoured at 1σ . Data collection and refinement statistics are shown in Table 1.

Water Renumbering

For consistency, the MSCS water molecules were renumbered. First, the structures were superimposed using least squares superposition of the entire protein chain in the program Coot and the cross-linked model as the reference structure. To renumber the water molecules, a consensus list of all the unique water positions in the MSCS structures was compiled. To continue using the same water numbering used in the MSCS structures in the C2 space group, the consensus list from those structures was used as a starting point. Next, all water molecules located within 5Å of the cross-linked molecule were compared with the consensus list of unique water binding positions. Water molecules from this structure that

Table 1 continued.

Solvent	Isopropanol	S,R,S-Bisfuranol	Trifluoroethanol
% volume	40%	50%	40%
Space Group	P3(2)21	P3(2)21	P3(2)21
Unit Cell	a = 63.48 Å b = 63.48 Å c = 64.23 Å $\alpha = 90.00^\circ$ $\beta = 90.00^\circ$ $\gamma = 120.00^\circ$	a = 63.77 Å b = 63.77 Å c = 63.87 Å $\alpha = 90.00^\circ$ $\beta = 90.00^\circ$ $\gamma = 120.00^\circ$	a = 63.58 Å b = 63.58 Å c = 63.90 Å $\alpha = 90.00^\circ$ $\beta = 90.00^\circ$ $\gamma = 120.00^\circ$
Temperature of data collection (K)	100	100	100
Resolution (Å)	1.78	1.90	2.04
# of Reflections	14778	12507	9877
Redundancy	10.8 (10.0)	10.2 (9.4)	10.3 (8.1)
R sym (%)	7.6 (78.7)	9.0 (0.0)	12.5 (57.9)
completeness (%)	100.0 (100.0)	100.0 (100.0)	100.0 (99.9)
average I/σ	43.1 (2.9)	33.4 (1.95)	21.7 (3.5)
Rwork/Rfree (%)	20.22/21.55	21.67/25.13	19.72/23.67
rmsd from ideal geometry for bond lengths (Å)	0.005	0.01	0.005
rmsd from ideal geometry for bond angles (°)	1.3	1.6	1.3
# protein atoms	951	951	951
# water molecules	87	85	87
# sulfate molecules	1	2	1
# organic solvent molecules	3	9	2
orientation of His 119	A	A & B	A

were found within 1.4Å of a water position on the consensus list were not added; however, the water position from the consensus list was averaged with the position of the comparison water molecule resulting in an average water position in the consensus list. Comparison water molecules that were not located within 1.4Å of a water molecule on the consensus list were added to the list as a unique water binding position. This process was repeated until the consensus list was populated with average unique water binding positions from all the MSCS structures. Water molecules from the MSCS structures were then assigned the number of the closest average position in the consensus list.

Comparison of Structures and Preparation of Files

RNAse A models in the P₃₂21 space group were downloaded from the Protein Data Bank (Berman et al., 2000). The downloaded structures had the following PDB codes: 1FS3, 1RAR, 1RAS, 1RNM, 1RNN, 1RNO, 1RNQ, 1RNW, 1RNX, 1RNY, 1RNZ, 1RPF, 1RPH, 2BLP, 2BLZ, 2G4X, and 4SRN. (Further details about the downloaded files can be found in Appendix B.) Using the cross-linked model as the reference structure, the downloaded models and the MSCS models were superimposed using least squares backbone superposition of the entire protein chain in the program Coot.

The MSCS structures in the C2 space group were also used as a comparison set of structures. The details of the crystallization, refinement, and preparation of these structures can be found in Chapter 2. These structures were superimposed with the P₃₂21 MSCS and downloaded structures.

Computational Analysis

Computational analysis was performed on the superimposed files of the C2 and P3₂21 MSCS and the P3₂21 downloaded. The analysis was run independently on the all three of these sets.

To identify conserved water positions, the SEWS (Structurally Equivalent Water System) program was used (Bottoms et al., 2006). Water molecules were identified as belonging to the same cluster if their positions fell within 1.4 Å of a calculated peak of atom density. Conserved water positions were identified as a cluster of water molecules that had at least one interaction with the protein in common. The interaction distance cutoff was 3.4 Å. The percent conservation of each conserved water position is the percentage of the structures with a water molecule located in a cluster and making the same interactions with the protein molecule. For each set of structures, a conservation level of 80% or higher across the set of structures was necessary for a water molecule to be considered conserved.

A Perl script was written to calculate the RMSD for each residue, between each pair of proteins. The RMSD was calculated as

$$\text{RMSD}^2 = (\{\sum_{i=1 \text{ to } n} [(x_{2,i} - x_{1,i})^2 + (y_{2,i} - y_{1,i})^2 + (z_{2,i} - z_{1,i})^2]\}/n)$$

where *i* is the atom type in each residue and *n* is the number of atoms in each residue. This calculation was performed for all atoms in the protein and also for just the backbone atoms (C, CA, O, and N). For each residue, the highest RMSD, lowest RMSD, and average of all calculated RMSD values were then plotted. For each set of structures, a baseline value was estimated using the plot of the average RMSD value for the backbone atom RMSD

calculations. A horizontal trendline was drawn across the low values of average RMSD plot and the y-intercept was used as the baseline value.

As done in previous work, the hinge angle of RNase A was calculated (Kishan et al., 1995; Sadasivan et al., 1998) for all MSCS and downloaded structures. A Perl script was written to estimate the centers of mass of domain A (residues 1-13, 49-79, and 105-124), domain B (residues 16-46 and 82-101), and the hinge (residues 14, 15, 47, 48, 80, 81, 102-104) using the C α atoms and then calculate the hinge angle. The vector defining the center of mass was calculated as

$$R = M^{-1} \sum_n (m_n r_n)$$

where r_n is the vector of the position of the nth C α , m_n is the mass of the atom at corresponding to r_n , which was set to 1 for all C α s, and M is the total mass of all C α . Since the mass of C α is the same for each vector, M was equal to n, resulting in R being the average position of all the C α s in a given domain. The hinge angle is defined as the angle between the two vectors formed by the center of mass of the hinge with the centers of mass of domain A and B. The hinge angle is calculated as

$$\text{Hinge Angle} = \text{acos} \{ (a_x b_x + a_y b_y + a_z b_z) / [(a_x^2 + a_y^2 + a_z^2)^{1/2} (b_x^2 + b_y^2 + b_z^2)^{1/2}] \}$$

where the Hinge Angle is in radians, a_x , a_y , and a_z are the x, y, and z components of the vector connecting the centers of mass of the hinge and domain A, and b_x , b_y , and b_z are the x, y, and z components of the vector connecting the centers of mass of the hinge and domain B. An additional Perl script was written to calculate the distance between the backbone nitrogen of Thr 45 and the backbone nitrogen of Phe 120, as done previously (Vitagliano et al., 2002).

Because the analysis of the hinge angle was not done in previous work of the MSCS of RNase A, the hinge angle and the Thr 45 N – Phe 120 N distance were also calculated on the following sets of structures: Zegers, Sadasivan, NMR, downloaded inhibitor structures, which includes the subset of downloaded C2 inhibitor structures. These sets of structures are listed and described in more detail in Chapter 2 and also are listed in the table legend for Table 2 in this chapter.

Results

RNase A Models

As described in the Materials and Methods, RNase A crystals grown in aqueous solution were cross-linked with glutaraldehyde and transferred to one of a variety of solutions containing high concentrations of organic solvents, resulting in 12 structures (Table 1). In addition to cross-linked (XLINK) RNase A in aqueous solution, the crystal structures were solved in the following conditions: 25% cyclopentanol (C5L), 40% cyclopentanone (C5N), 40% cyclohexanol (CXL), 40% cyclohexanone (CYH), 40% dioxane (DIO), 40% dimethylformamide (DMF), 25% dimethylsulfoxide (DMS), 50% 1,6-hexanediol (HEZ), 40% 2-propanol (IPA), 50% S,R,S-bisfuranol (SRF), and 40% trifluoroethanol (ETF).

Because of the high salt concentrations in the mother liquor, it was not possible to obtain data for higher concentrations of organic solvent soaks. All of the models are in the P3₂21 space group and have one protein molecule in the asymmetric unit.

In each of the structures, density was poor or missing for the side chains of Lys 1, Lys 37, Asp 38, Arg 39, Lys 66, Lys 91, and Lys 104. When it was impossible to discern a side chain position, the side chain was modeled as a common rotamer. Lys 104 packs against the same residue in a symmetry-related molecule of RNase A, preventing it being modeled into a common rotamer because of steric collisions. In addition to these seven residues, electron density for three additional side chains was incomplete in certain structures: Lys 7 in the CYH structure, Lys 31 in the DIO, DMF, and CYH structures, and Asn 94 in the ETF structure.

All of these structures of RNase A have a sulfate ion bound in the active site between the two catalytic histidines, 12 and 119. The bound sulfate ion interacts with the backbone nitrogen of Phe 120, N ϵ 2 of His 12, and N δ 1 of His 119, similar to the interactions made by a phosphate moiety of a substrate molecule. His 119, which has been previously described as adopting one of two conformations, A or B (Borkakoti et al., 1982; Zegers et al., 1994), was found in the catalytically active conformation A in all the structures of RNase A except for the XLINK model, in which it adopts conformation B, and the SRF structure, in which His 119 is found in both conformations (Table 1). An additional sulfate ion is found near the backbone nitrogen and O γ of Ser 23 in the XLINK, CYH, and SRF structures.

Protein Plasticity

The protein molecules from each of the 12 MSCS models were superimposed as described in the Materials and Methods, and using a Perl script, the residue RMSD of each pair of

structures was calculated. This was done in order to quantify the range of motion observed for each amino acid in response to the different solvent environments, as has been observed with elastase in earlier work (Mattos et al., 2006). The highest RMSD value (resulting from the pair of structures that are the most divergent at residue x), the lowest RMSD value (resulting from the pair of structures that are the most similar at residue x), and the average of all the pair-wise RMSD values for residue x are plotted in Figure 1a.

As observed with the MSCS structures of RNase A in the C2 spacegroup, the plot of the backbone RMSD calculations highlights clusters of peaks where the backbone plasticity is higher than the rest of the molecule of RNase A. These peaks correspond with loop regions of RNase A, particularly: the loop connecting α -helix 1 and α -helix 2 consisting of residues 13-24, the loop containing residues 33-42 connecting α -helix 2 to β -strand 1, residues 64-71 connecting β -strand 2 to β -strand 3, the loop between β -strand 4 and β -strand 5 consisting of residues 87-96 with Gly 88 having the highest calculated RMSD values across this set of models, and the loop connecting β -strand 5 and β -strand 6 consisting of residues 112-115. After Gly 88, Lys 1 has the second highest calculated RMSD values for this set of models, and this residue falls in the loop before α -helix 1. The three α -helices and six β -strands of RNase A have lower RMSD trends across the MSCS models. In addition to corresponding to secondary structural elements, a number of the highest peaks of calculated RMSDs correspond to areas of disorder and high B-factors in the models: Lys 1, Lys 37, Asp 38, Arg 39, Lys 66, Lys 91, and Asn 94.

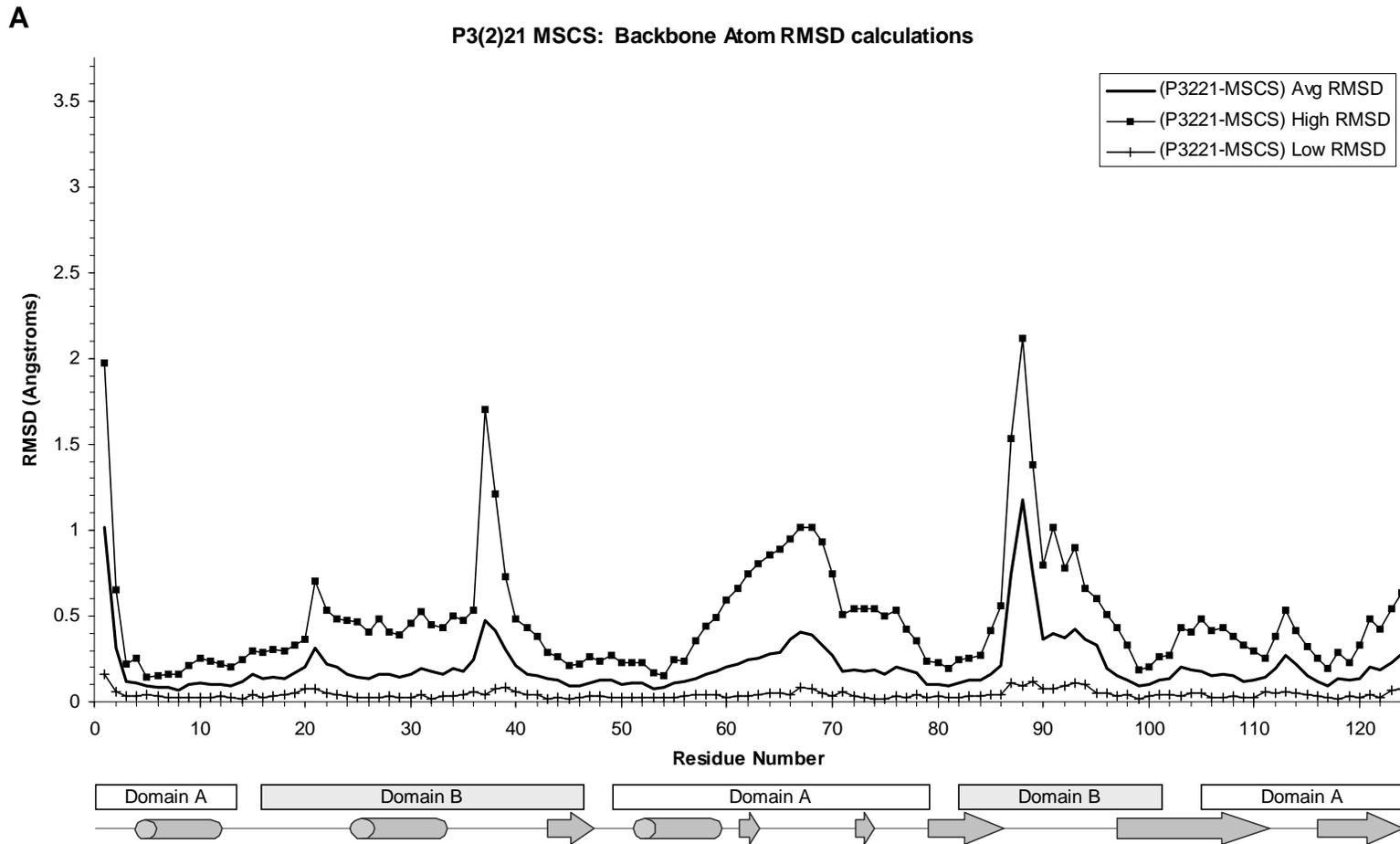


Figure 1. RMSD calculations: High, Low, and Average RMSD per residue. Secondary structural elements are depicted below the x-axis with cylinders and arrows illustrating helices and strands, respectively; and the boxes indicate residues belonging to Domain A and Domain B.

A. Backbone RMSD calculations for 12 MSCS structures ($P3_221$) of RNase A with a bound sulfate ion in the active site.

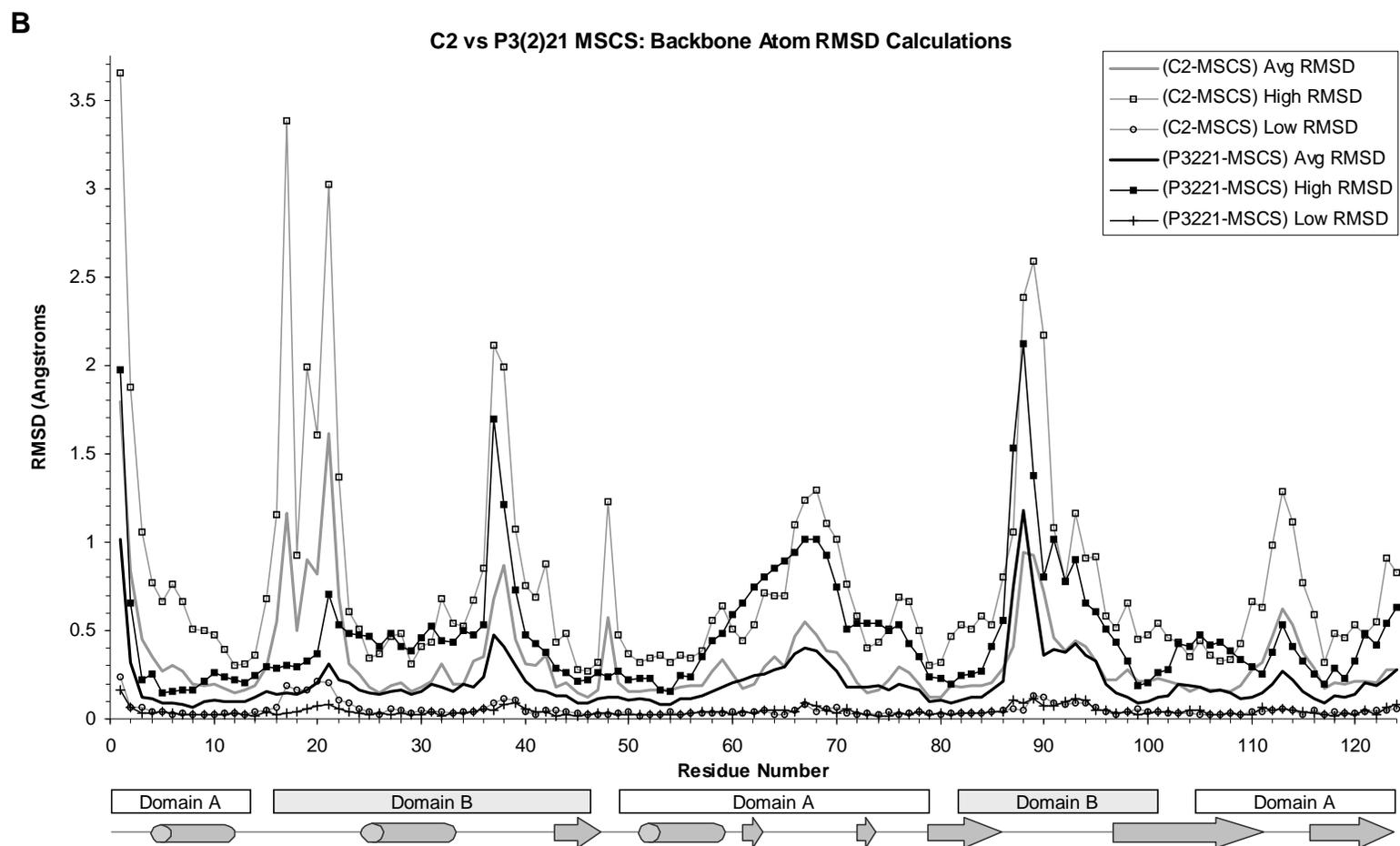


Figure 1 continued.

B. Backbone RMSD calculations for 12 P₃2₁ MSCS structures (sulfate ion in the active site) and 20 MSCS structures in the C₂ space group.

C

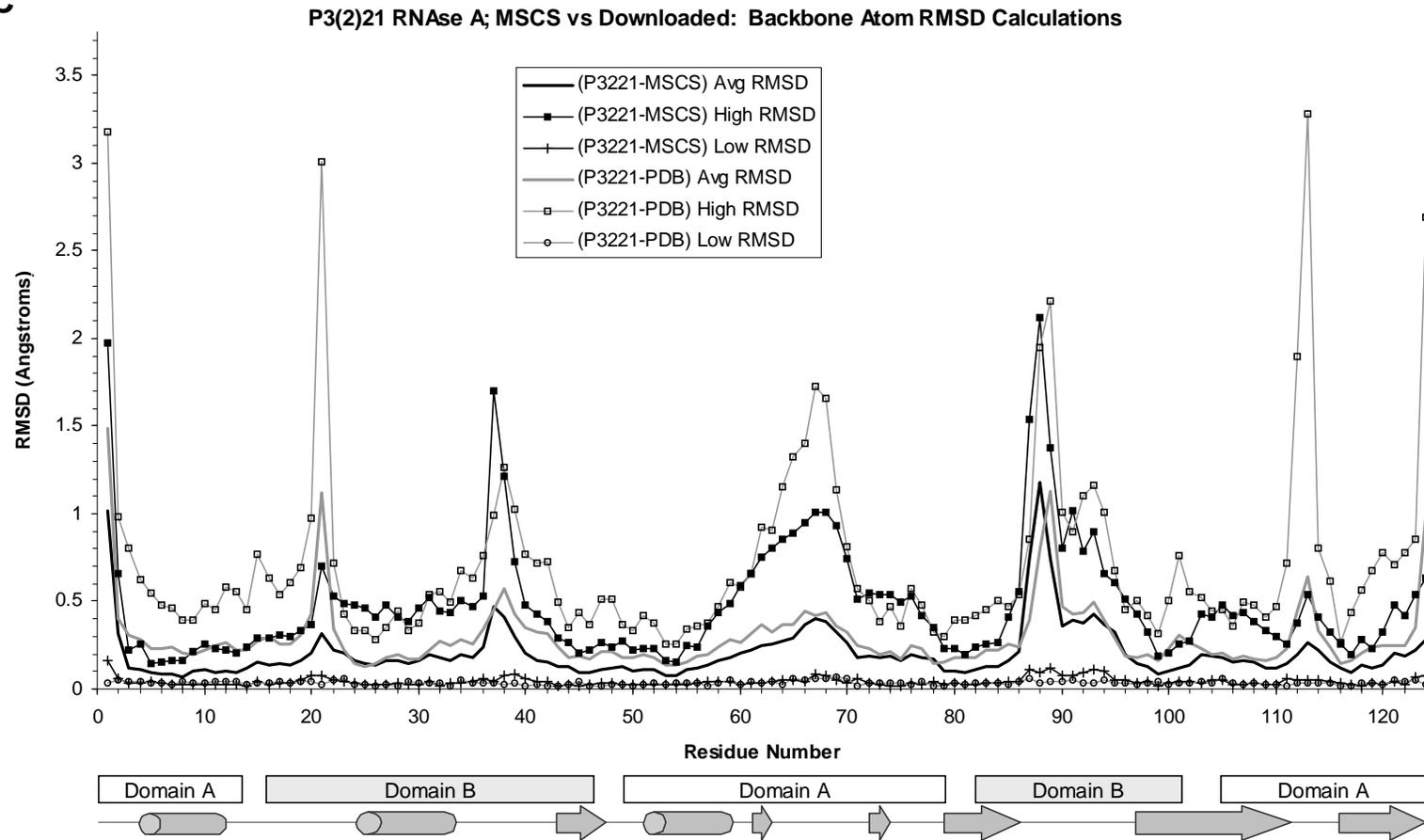


Figure 1 continued.

C. Backbone RMSD calculations for 12 MSCS structures with a bound sulfate ion in the active site (P₃₂₁ space group) and 17 downloaded structures in the P₃₂₁ space group.

In addition to differences among elements of secondary structure, differences in RMSD among domains of RNase A are also observed. First, the hinge region has low average backbone RMSD values ranging from 0.09 to 0.20 Å. Second, Domain A has average backbone RMSD values ranging from 0.07 to 0.40 Å with the exception of Lys 1, which has an average backbone RMSD of 1.01 Å. All of the highest RMSD values fall in Domain B, with the exception of Lys 1. The average backbone RMSD values range from 0.09 to 0.47 Å in Domain B, except for Thr 87 (0.74 Å), Gly 88 (1.18 Å), and Ser 89 (0.74 Å), and three additional residues have higher RMSD values than the maximum of Domain A, excluding Lys 1: Lys 37, 0.47 Å; Asp 38, 0.41; and Pro 93, 0.43 Å. Using the plot of the average RMSD values for the backbone atoms, an approximate baseline was drawn with a value of 0.18 Å. 74 residues (3-19, 24-30, 32-33, 41-58, 71, 73, 75, 78-85, 97-102, 105-111, and 115-120) were found with average RMSD values falling at or below this baseline. The majority of all three regions of RNase A are comprised of these residues with 77.8%, 60.9%, and 54.9% of the residues in the hinge region, domain A, and domain B having average RMSD values below the 0.18 Å cutoff, respectively. Domain B has the highest plasticity of the three regions, where the hinge is the most rigid, both of which is consistent with observations in other studies (Kishan et al., 1995; Merlino et al., 2002; Sadasivan et al., 1998).

In the P₃₂₁ MSCS structures of RNase A, most of the active site residues have low average backbone RMSD values. In the P₁ pocket, all of the residues have an average RMSD below the 0.18 Å cutoff (Gln 11, 0.10 Å; His 12, 0.10 Å; Lys 41, 0.16 Å; His 119, 0.12 Å), except for Asp 121, which has an average backbone RMSD of 0.21 Å, a value that falls just outside

the baseline cutoff. In the C2 MSCS structures, Lys 41 had a value 0.31 Å, well above the baseline cutoff, but there were no sulfate ions bound in the active site to tether Lys 41 as is observed in the P₃₂₂₁ MSCS structures. This change illustrates the observation that Lys 41 adjusts to improve binding with inhibitors (Vitagliano et al., 2000). Similar to what was observed in the C2 MSCS structures, Thr 45 and Asp 83 have low average RMSD values of 0.09 Å and 0.12 Å, respectively, where Ser 123 had a higher value of 0.23 Å. The B₂ pocket has both Asn 71 and Glu 111 falling below the baseline with RMSD values of 0.18 Å and 0.15 Å, respectively. This is different from what is observed in the C2 MSCS structures, where both of these residues exhibit higher average backbone RMSD values of 0.30 Å and 0.32 Å, respectively. Both of these residues belong to loops, which are distal from the hinge region. Generally, these loops exhibit the greatest movement across superimposed structures of RNase A, because of the hinge action of the protein. In the case of the C2 MSCS structures, the two molecules in the asymmetric unit have different ranges of hinge motion, resulting in a larger calculated RMSD value for residues in the distal loops.

The highest average RMSD value for an active site residue belongs to Lys 66 (0.37 Å) of the P₀ pocket. The average backbone RMSD across the MSCS structures highlights the residues of the active site that remain rigid versus those that have higher plasticity and adjust upon ligand binding. In the case of the P₁ subsite, the sulfate ion is a mimic of a bound inhibitor and the rigidity of the residues are in response to this. The differences between the C2 and P₃₂₂₁ MSCS structures in this subsite highlight the residues, particularly Lys 41, that adjust in response to substrate binding.

These data were compared against the trends seen in the RMSD plots for two other sets of structures. First, the C2 MSCS structures were used for comparison. The baseline used for the C2 structures was 0.21 Å, and 54 residues (8-14, 25-30, 33-34, 43-47, 49-57, 61-62, 72-74, 78-85, 100, 103-109, 117-119, and 122) were found with average RMSD values falling at or below this baseline. As observed with the P3₂21 structures, the hinge region of the C2 structures is the most rigid with 66.7% of the hinge residues having an RMSD below the cutoff, followed by domain A and B, which have 48.4% and 33.3% of their residues falling below 0.21 Å. The plots of the calculated backbone RMSD values are similar for these two sets of structures (Figure 1b), but there are some differences in the trends. These five differences can be explained by the differences in crystal packing between the two space groups. Residues 4-8 in the P3₂21 structures are stabilized by crystal contact, resulting in lower RMSD values compared to those of the C2 structures. In the C2 structures, the loop encompassing residues 16-20 is disordered and this is reflected by the high RMSD values, where this region is stabilized in the P3₂21 structures and has RMSD near or below the baseline for this set of structures. This affects His 48, which lies beneath the 16-20 loop, and results in the high RMSD values in the C2 structures and the below baseline values in the P3₂21 structures. Pro 42, and residues 97-102 are stabilized by crystal contacts in the P3₂21 structures, but not the C2 structures, which is exhibited by the elevated values observed in the C2 plots. While the overall trends are similar, the P3₂21 structures generally exhibit lower RMSD values, which may be a result of fewer structures used in this set (12 in P3₂21 vs. 20 in C2), and the different crystal packing environments for the two RNase A molecules in the C2 crystals. Second, the set of 17 P3₂21 structures downloaded from the PDB were

used for comparison. The plots of the P3₂21 MSCS and P3₂21 downloaded structures (Figure 1c) exhibit similar trends. The main difference is in magnitude, which may be a result of more downloaded structures (17 downloaded vs. 12 MSCS) or the fact that the downloaded structures are quite diverse: this collection contains apo structures, and structures with a sulfate ion, chloride ion, formic acid, or inhibitor bound in the P₁ pocket.

Molecular Hinge

In addition to individual side chain fluctuations, RNase A exhibits molecular hinge bending motions. The two β -sheets, and consequently the two domains A and B, move in concert producing a more open or more closed conformation. This fluctuation is observed in the MSCS structures, as illustrated by two representative models in Figure 2. Because the hinge angle was not investigated with the previous MSCS work, all sets of comparison structures were used analysis. The structure sets included are the P3₂21 MSCS structures, the C2 MSCS structures, the Sadasivan structures, the Zegers structures, the 2AAS NMR models, the downloaded inhibitor structures, which includes the C2 inhibitor subset, and the P3₂21 downloaded structures. A listing of these structures can be found in the caption of Table 2 and in Appendix B.

To measure the fluctuations of the hinge observed in RNase A, two metrics were used. The first is the measure of the distance between the backbone nitrogens of Thr 45 and Phe 120, as used by the Vitagliano study (Vitagliano et al., 2000). Both of these residues are located in the P₁ subsite of the active site cleft, interact with substrates, and are located on different β -

sheets: Thr 45 in β_2 (domain B) and Phe 120 in β_1 (domain A). The second measure is that of the hinge angle, which is described as the angle between the two vectors connecting the center of mass of the hinge with the center of mass of domain A, and the center of mass of the hinge with the center of mass of domain B. The center of mass for each domain is

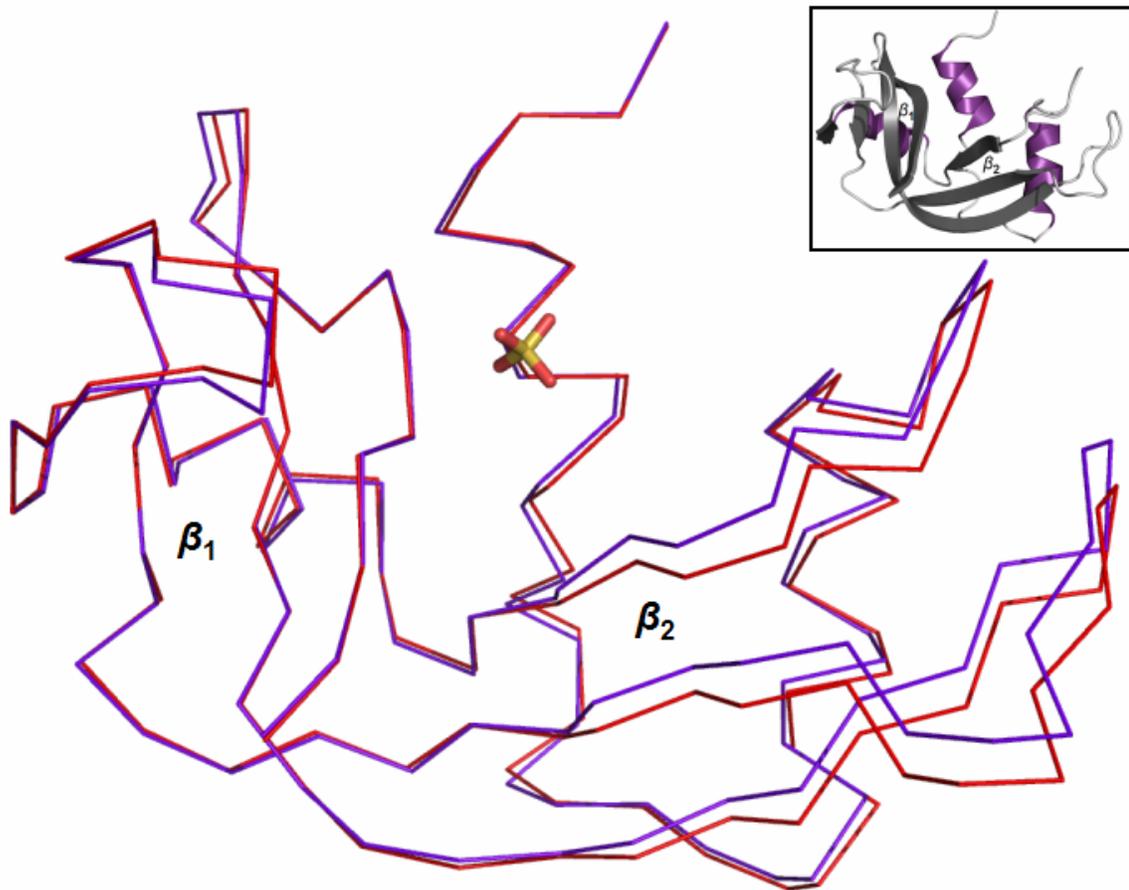


Figure 2. Hinge motions of RNase A. $\text{C}\alpha$ ribbon drawing of C2 HEZ molecule B (red) and P3₂21 XLINK (blue) with a representative sulfate ion bound in the P₁ pocket (MSCS numbering 903). The β -sheet of domain A is labeled β_1 , and that of domain B is labeled β_2 . Structures are superimposed on β_1 , specifically, residues 61-63, 71-75, 105-111, 116-124, as described previously (Vitagliano et al., 2002). Of the MSCS structures, P3₂21 XLINK has the shortest distance between Thr 45 N and Phe 120 N at 8.06 Å and C2 HEZ molecule B has the greatest at 8.81 Å. Inset: Cartoon rendering of RNase A with helices colored purple, and β -strands colored dark gray. The orientation is approximately that of the larger figure and β_1 and β_2 are labeled. Figures 2, and 4-8 were prepared using Pymol (DeLano).

calculated using the positions of the C α atoms, as was done previously (Kishan et al., 1995; Sadasivan et al., 1998). The results of these calculations done for each RNase A molecule in all the sets of structures are presented in Table 2. There do not appear to be any clear patterns, but there are a few general trends. First, in the structures in the C2 space group, molecule A tends to adopt a more closed conformation than molecule B, which is consistent with what was observed previously (Vitagliano et al., 2002). Also, P3₂21 MSCS structures, all of which have a sulfate ion bound in the P₁ pocket of the active site adopt a more closed conformation, similar to that of molecule A in the C2 structures. This does not appear to be an effect of the P3₂21 space group, because a number of the downloaded structures have hinge angle values reflecting a more open conformation. Finally, the solution NMR structures, which have no ligands bound in the active site, exhibit the highest Thr 45 N – Phe 120 N distance and hinge angle values of all the sets of structures, which given the lack of crystal packing constraints, is not surprising.

To visualize these trends more easily, the Thr 45 N – Phe 120 N distance (Å) was plotted against the hinge angle (degrees) for each molecule of RNase A. Each point was color-coded to identify which values were from the NMR, downloaded crystal, or MSCS crystal structures (Figure 3a). From this plot, it is clear that the NMR structures have the largest hinge angles of all the structures and very few crystal structures reach this range of values. On the other hand, while the Thr 45 N – Phe 120 N distances of the NMR structures are at the larger end of the spectrum, a number of MSCS and downloaded structures have distances

Table 2. Hinge Angle Calculations. Results of the hinge angle calculation (degrees) and the calculation of the distance between Thr 45 N and Phe 120 N (Angstroms) for downloaded and MSCS structures. The calculations are described in the Materials and Methods and were performed on each RNase A molecule in the asymmetric unit. Downloaded structures in the P₃₂₁ space group: 1FS3, 1RAR, 1RAS, 1RNM, 1RNN, 1RNO, 1RNQ, 1RNW, 1RNX, 1RNY, 1RNZ, 1RPF, 1RPH, 2BLP, 2BLZ, 2G4X, 4SRN; Downloaded inhibitor structures: 1AFK, 1AFL, 1EOS, 1EOW, 1JN4, 1JVU, 1O0F, 1O0H, 1O0M, 1O0N, 1O0O, 1QHC, 1RAR, 1RAS, 1RBJ, 1RCA, 1RCN, 1RNC, 1RND, 1RNM, 1RNN, 1ROB, 1RPF, 1RPG, 1RSM, 1RTA, 1RUV, 1U1B, 1W4O, 1W4P, 1W4Q, 1WBU, 1Z6D, 1Z6S, 2G8R, 8RSA, 9RSA; Downloaded inhibitor structures in the C₂ space group: 1AFK, 1AFL, 1EOS, 1JN4, 1JVU, 1O0F, 1O0H, 1O0M, 1O0N, 1O0O, 1QHC, 1W4O, 1W4P, 1W4Q, 1WBU, 1Z6D, 1Z6S, 2G8R; Downloaded structures included in the Zegers study (Zegers et al., 1994): 1RPG, 1ROB, 7RSA, 1RPH, 1RPF, 4SRN, 1RSM, 8RSA, 2RNS; Downloaded structures included in the Sadasivan study (Sadasivan et al., 1998): 7RSA, 5RSA, 3RN3, 1RHB, 1RHA, 1RPH, 1XPT, and 1XPS; Downloaded NMR structures: 2AAS Models 1-32; MSCS models in the P₃₂₁ space group: SRF P₃(2)₂₁, C5L P₃(2)₂₁, C5N P₃(2)₂₁, CXL P₃(2)₂₁, CYH P₃(2)₂₁, DMF P₃(2)₂₁, DMS P₃(2)₂₁, DIO P₃(2)₂₁, ETF P₃(2)₂₁, HEZ P₃(2)₂₁, IPA P₃(2)₂₁, XLINK P₃(2)₂₁; MSCS models in the C₂ space group: RSF C₂, DMF C₂, DMS C₂, DIO C₂, ETF C₂, HEZ C₂, IPA C₂, TBU C₂, TMO C₂, XLINK C₂. Further details about the downloaded structures can be found in Appendix B.

Structure	Thr45N-Phe120N Distance	Hinge Angle
1AFK molecule A	8.28	94.60
1AFK molecule B	8.59	95.57
1AFL molecule A	8.37	94.66
1AFL molecule B	8.54	95.60
1EOS molecule A	8.23	94.02
1EOS molecule B	8.70	95.91
1EOW	8.19	93.90
1FS3	8.34	94.60
1JN4 molecule A	8.07	93.23
1JN4 molecule B	8.58	95.41
1JVU molecule A	8.12	93.83
1JVU molecule B	8.70	96.02
1O0F molecule A	7.98	93.78
1O0F molecule B	8.72	97.64
1O0H molecule A	7.96	93.49
1O0H molecule B	8.59	95.78
1O0M molecule A	7.86	93.03
1O0M molecule B	8.51	95.62
1O0N molecule A	8.14	94.00
1O0N molecule B	8.44	95.02
1O0O molecule A	7.97	93.42
1O0O molecule B	8.59	95.69
1QHC molecule A	8.31	94.53
1QHC molecule B	8.58	95.57
1RAR	8.32	96.41
1RAS	8.26	96.15

Structure	Thr45N-Phe120N Distance	Hinge Angle
1RBJ	8.13	93.42
1RCA	8.21	93.76
1RCN	7.98	93.67
1RHA	8.02	93.15
1RHB	8.29	94.71
1RNC	8.08	94.34
1RND	8.17	94.19
1RNM	8.13	93.40
1RNN	8.17	93.48
1RNO	8.25	94.85
1RNQ	8.25	93.75
1RNW	8.27	94.82
1RNX	8.33	94.66
1RNY	8.30	94.84
1RNZ	8.31	94.65
1ROB	8.05	94.11
1RPF	8.17	93.77
1RPG	8.26	94.08
1RPH	8.38	94.98
1RSM	8.75	95.65
1RTA	8.48	94.00
1RUV	8.22	94.42
1U1B molecule A	8.30	94.11
1U1B molecule B	8.31	93.00
1W4O molecule A	8.02	93.60
1W4O molecule B	8.55	95.02

Table 2 continued.

Structure	Thr45N-Phe120N Distance	Hinge Angle
1W4P molecule A	8.11	93.72
1W4P molecule B	8.53	95.33
1W4Q molecule A	7.97	93.03
1W4Q molecule B	8.26	94.25
1WBU molecule A	8.15	93.97
1WBU molecule B	8.39	95.55
1XPS molecule A	8.20	94.73
1XPS molecule B	8.82	95.04
1XPT molecule A	8.37	94.75
1XPT molecule B	8.56	94.96
1Z6D molecule A	7.98	93.10
1Z6D molecule B	8.49	95.55
1Z6S molecule A	8.17	94.45
1Z6S molecule B	8.67	96.21
2AAS Model 1	8.63	97.58
2AAS Model 2	8.41	98.51
2AAS Model 3	8.56	100.52
2AAS Model 4	8.50	100.74
2AAS Model 5	8.46	97.24
2AAS Model 6	8.68	99.58
2AAS Model 7	8.71	98.89
2AAS Model 8	8.56	100.42
2AAS Model 9	8.73	98.69
2AAS Model 10	8.94	99.48
2AAS Model 11	8.67	98.88
2AAS Model 12	8.95	99.28
2AAS Model 13	8.79	99.68
2AAS Model 14	8.74	101.88
2AAS Model 15	8.66	98.03
2AAS Model 16 & 32	8.80	100.39
2AAS Model 17	8.73	98.42
2AAS Model 18	8.71	100.14
2AAS Model 19	8.60	98.61
2AAS Model 20	8.51	98.95
2AAS Model 21	8.60	96.77
2AAS Model 22	8.59	98.58
2AAS Model 23	8.6	97.74
2AAS Model 24	8.54	100.60
2AAS Model 25	8.69	96.85
2AAS Model 26	8.92	98.09

Structure	Thr45N-Phe120N Distance	Hinge Angle
2AAS Model 27	8.82	97.46
2AAS Model 28	8.73	96.64
2AAS Model 29	8.65	96.67
2AAS Model 30	8.79	99.63
2AAS Model 31	8.67	98.49
2BLP	8.29	95.23
2BLZ	8.33	95.20
2G4X	8.33	94.90
2G8R molecule A	8.33	94.52
2G8R molecule B	8.61	96.51
2RNS	8.56	91.07
3RN3	8.37	94.65
4SRN	8.40	95.90
5RSA	8.43	94.34
7RSA	8.55	94.76
8RSA molecule A	8.46	95.97
8RSA molecule B	8.46	96.39
9RSA molecule A	8.75	95.64
9RSA molecule B	8.65	96.10
RSF C2 molecule A	8.42	93.43
RSF C2 molecule B	8.65	95.83
SRF P3(2)21	8.27	94.09
C5L P3(2)21	8.15	93.42
C5N P3(2)21	8.19	94.02
CXL P3(2)21	8.20	93.07
CYH P3(2)21	8.20	93.84
DMF C2 molecule A	8.25	93.31
DMF C2 molecule B	8.64	93.72
DMF P3(2)21	8.09	92.28
DMS C2 molecule A	8.43	94.81
DMS C2 molecule B	8.68	94.61
DMS P3(2)21	8.13	93.04
DIO C2 molecule A	8.36	93.11
DIO C2 molecule B	8.73	93.90
DIO P3(2)21	8.21	94.10
ETF C2 molecule A	8.37	92.23
ETF C2 molecule B	8.69	94.91
ETF P3(2)21	8.13	93.53
HEZ C2 molecule A	8.73	93.96
HEZ C2 molecule B	8.81	95.70

Table 2 continued.

Structure	Thr45N-Phe120N Distance	Hinge Angle
HEZ P3(2)21	8.29	94.54
IPA C2 molecule A	8.22	92.48
IPA C2 molecule B	8.63	95.78
IPA P3(2)21	8.14	95.08
TBU C2 molecule A	8.26	92.76
TBU C2 molecule B	8.58	93.56
TMO C2 molecule A	8.33	94.15
TMO C2 molecule B	8.70	95.65
XLINK C2 molecule A	8.34	93.80
XLINK C2 molecule B	8.55	95.74
XLINK P3(2)21	8.06	93.61

in this range. When considering the crystal structures in this plot, the MSCS structures sample well the values observed in the downloaded crystal structures, but the MSCS structures tend to have lower overall hinge angles. A second plot was created to compare the MSCS and downloaded structures of the P3₂21 and C2 space groups (Figure 3b). The C2 structures have the same trends as observed in the overall crystal comparison. The C2 MSCS structures samples the C2 values fairly well, but the hinge angles in the MSCS structures are lower, overall. The P3₂21 structures show a different trend. The P3₂21 MSCS structures adopt a more closed conformation and have lower distances and hinge angles than the downloaded structures in the same space group. This is probably because all of the P3₂21 MSCS structures have a sulfate ion in the active site, where the downloaded structures have a wide variety of conditions of the active site (apo, sulfate, inhibitor molecule, chloride ion, formic acid).

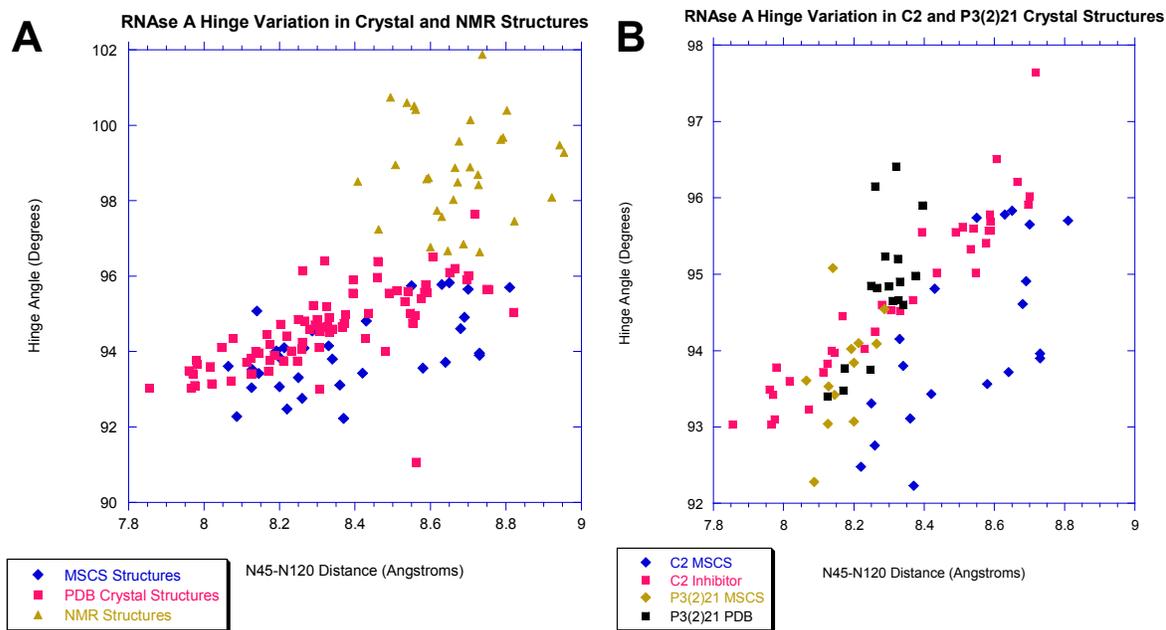


Figure 3. RNase A hinge calculations: Thr 45 N – Phe 120 N distance vs hinge angle. Calculations are described in Materials and Methods and the values are listed in Table 2.

A. Hinge variation of RNase A observed in crystal and NMR structures. A listing of individual structures included in this plot is included in Table 2. MSCS structures includes both those in the C2 and the P₃21 space groups, NMR structures include all models from PDB ID 2AAS, and PDB crystal structures include all downloaded crystal structures, which are listed as parts of the following groups: downloaded inhibitor structures (which includes the C2 space group), Zegers study (Zegers et al., 1994), Sadasivan study (Sadasivan et al., 1998), downloaded structures in the P₃21 space group. B. Hinge variation of RNase A observed in crystal structures of the C2 and P₃21 space groups. A listing of structures included in this plot is included in Table 2. C2 MSCS includes MSCS structures in the C2 space group, C2 inhibitor includes downloaded inhibitor structures in the C2 space group, P₃21 MSCS includes MSCS structures in the P₃21 space group, P₃21 PDB includes downloaded structures in the P₃21 space group.

Conserved Water Binding Sites

The SEWS (Structurally Equivalent Water System) program, which was developed by and uses the method described previously (Bottoms et al., 2006), was used to computationally identify conserved water binding sites for three sets of structures: the 12 P₃21 MSCS structures, the 20 C2 MSCS structures, and the 17 P₃21 structures downloaded from the PDB. Forty-two conserved water binding sites were identified for the MSCS structures at

80% conservation or greater (Table 3), meaning that these water molecules were present and made the same interactions with the protein in at least 10 of the 12 P₃₂21 MSCS structures. Using the same 80% conservation cutoff for the other two sets of structures, 31 conserved water binding sites were found in the C2 MSCS structures (Chapter 2, Table 2), and 29 conserved sites were found in the P₃₂21 downloaded structures. Of the 42 conserved waters identified in the P₃₂21 MSCS structures, 22 were 100% conserved, eight were 92% conserved, and 12 were 83% conserved. Furthermore, the majority of conserved water binding sites are found interacting with domain A for each of the set of structures, with 71%, 69%, and 87% of the waters from the P₃₂21 MSCS structures, the P₃₂21 downloaded structures, and the C2 MSCS structures, respectively, associating with domain A of RNase A.

Active Site Waters

There are four 80% conserved water binding sites in the active site of RNase A from the P₃₂21 MSCS structures. Wat 22 binds in the B₁ pocket, Wat 16 and Wat 20 bind in the P₁ pocket, and Wat 34 binds in the B₂ pocket. An additional two waters are identified as conserved in the C2 MSCS structures, but not in the P₃₂21 structures: Wat 19 in the P₁ subsite and Wat 13 in the B₂ subsite.

Wat 22 binds in the B₁ subsite and is 100% conserved in the P₃₂21 MSCS structures (Figure 4a). Wat 22 is within hydrogen bonding distance of the backbone nitrogen of Thr 45, but its

Table 3. Conserved Water Molecules identified by SEWS for the MSCS structures of RNase A. Water molecules that are not 80+% conserved in the P3₂₁ MSCS structures (2, 7, 13, 18, 19, 23, 78, 90, 129, 158, 243, and 423) do not have values listed for Average Distance (Å).

Conserved Water Position	SO4 MSCS Conservatn	MSCS Conservatn	P3221 Conservatn	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)
2	<50%	100%	<50%	Tyr 76 OH							
4	100%	100%	100%	Ala 5 O	2.89	Pro 117 O	2.74				
5	83%	80%	88%	Ser 15 N	2.85						
6	100%	100%	100%	Ser 50 N	2.86	Glu 49 OE1	2.82	Asp 53 OD2	2.66		
7	50%	90%	94%	Ala 52 O							
10	100%	100%	100%	Ser 77 N	2.88	Tyr 76 N	3.17	Gln 60 OE1	2.78		
13	67%	90%	71%	Asn 67 N		Lys 66 N		Asp 121 OD1			
15	100%	100%	94%	Glu 9 OE2	2.75	Gln 55 NE2/OE1 (1)	2.79				
16	100%	90%	76%	Gln 11 NE2	2.89						
17	92%	100%	94%	Pro 114 O	2.82						
18	25%	85%	59%	Ala 6 O							
19	<50%	100%	<50%	Phe 120 N		His 12 NE2					
20	100%	85%	100%	Ala 4 O	2.73	Val 118 O	2.60				
21	100%	95%	100%	Ser 23 O	2.74	Asn 27 N	3.08	Tyr 97 O	2.89	Thr 99 OG1	2.88
22	100%	70%	<50%	Thr 45 N	2.99						
23	<50%	100%	53%	Cys 110 O							
26	100%	100%	94%	Cys 58 O	2.86						
32	100%	85%	94%	Ser 50 OG	2.70	Asp 53 OD1 (75%)	3.19				
34	83%	85%	88%	Glu 111 OE2	2.61	Asn 71 ND2 (75%)	2.91	Gln 69 NE2 / OE1 (75%)	3.02		
39	92%	95%	94%	Glu 111 O	3.02						
47	100%	<50%	88%	Thr 99 O	2.62	Lys 98 NZ	2.73				
56	100%	<50%	76%	Lys 31 O	2.78	Thr 36 N	3.17	Thr 36 OG1	2.74		

Table 3 continued.

Conserved Water Position	SO4 MSCS Conservatn	MSCS Conservatn	P3221 Conservatn	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)
57	83%	80%	76%	Ser 32 OG	2.71	Arg 33 NE	2.85				
63	83%	80%	100%	Ser 23 O	3.02	Thr 99 N	2.80				
64	100%	95%	94%	Asp 83 O	3.00	Lys 98 O	2.83	Thr 100 OG1	2.71		
69	83%	100%	100%	Ala 5 N	2.94						
71	83%	<50%	65%	Asn 34 ND2/OD1	2.76						
78	<50%	90%	<50%	Asn 62 O							
82	100%	100%	100%	Asp 53 O	2.68	Gln 60 NE2	2.93				
90	<50%	80%	<50%	Thr 78 N							
110	100%	50%	94%	Asn 27 O	2.92						
123	83%	70%	82%	Gln 60 O	2.75						
129	<50%	80%	65%	Thr 78 OG1		Asn 103 OD1		Thr 78 O (70%)			
149	92%	50%	88%	Asp 53 OD2	3.05	Glu 49 OE1 (75%)	3.27				
158	<50%	85%	<50%	Gln 74 OE1							
163	100%	50%	100%	Thr 3 N	3.05						
166	100%	<50%	<50%	His 48 O	2.95						
220	83%	<50%	59%	Glu 49 OE1	2.82						
229	92%	<50%	100%	Gln 55 O	3.12						
243	<50%	80%	<50%	Ala 4 N							
282	100%	35%	94%	Ala 52 N	2.98						
286	92%	<50%	88%	Glu 9 OE1	2.63						
318	92%	80%	94%	Gln 60 O	3.17	Asn 62 N	2.94				
344	83%	<50%	76%	Gln 55 OE1/NE2	2.83						

Table 3 continued.

Conserved Water Position	SO4 MSCS Conservatn	MSCS Conservatn	P3221 Conservatn	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)	Interaction	Average Distance (Å)
361	100%	<50%	76%	Leu 51 N	2.96						
368	100%	<50%	76%	Ser 18 N	2.90	Ser 18 OG (92%)	3.15				
423	58%	85%	<50%	Gln 60 NE2							
435	83%	<50%	<50%	Ser 21 OG	2.65	Ser 22 OG	2.93	Gln 28 NE2/OE1	2.86		
437	92%	60%	82%	Tyr 115 N	2.96						
463	92%	<50%	65%	Thr 3 OG1	2.76						
480	100%	<50%	<50%	Thr 70 N	3.11						
483	100%	<50%	65%	Gly 112 N	2.98						
486	83%	<50%	<50%	Ser 32 O	2.71						
509	83%	<50%	<50%	Ser 32 OG	3.06						

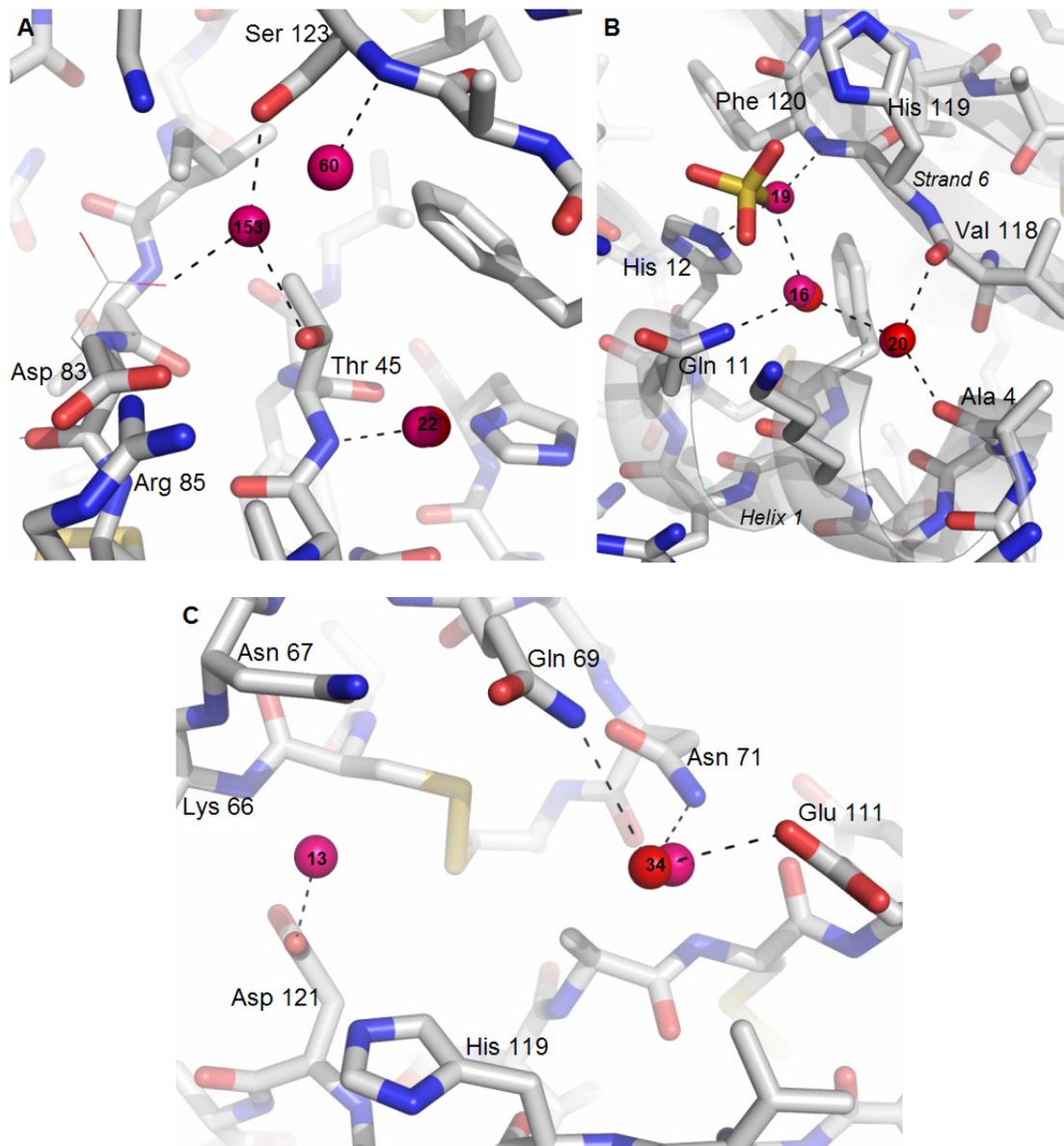


Figure 4. Conserved water molecules in the active site. RNase A is drawn with sticks with carbon atoms colored light gray, a representative sulfate ion is drawn as sticks and colored by atom type (sulfur, orange; oxygen, red), and representative conserved water molecules from the P3₂21 MSCS structures are colored red. Comparison representative conserved water molecules from the C2 models are colored hot pink. A black dashed line designates interactions between water molecules and other water molecules, protein, or ion atoms. An alternate conformation of Asp 83 is represented as lines using the same color scheme for RNase A. A. Conserved water positions in the B₁ pocket. B. Conserved water positions in the P₁ pocket. C. Conserved water positions in the B₂ pocket.

positioning is not optimal for hydrogen bond formation. This water is displaced by a number of organic solvents (binding site 902) in the C2 MSCS structures, and inhibitors, chloride ions, and formic acid in the P3₂21 downloaded structures, resulting in this water not being at least 80% conserved in either of those two sets of structures. An earlier study (Gilliland et al., 1994) discussed two other water molecules that change roles as hydrogen bond donor or acceptor depending on if a cytosine or uracil is bound in the B₁ pocket. These water molecules correspond to MSCS Wat 60 and Wat 153, which are not found to be over 80% conserved in any of the three sets of structures presented here. Wat 60 is the most conserved of the two in the MSCS structures at 75%. It interacts with the backbone nitrogen of Ser 123, and is displaced by organic solvents occupying binding site 901 in the MSCS structures. In the three instances where a pyrimidine inhibitor is bound in the B₁ pocket, a water molecule in this position is present, except in the one case where a sulfate ion (1RNM) is bound in this position and an oxygen atom of this ion displaces the water molecule. In the P3₂21 MSCS structures, Asp 83 adopts a conformation turned away from Ser 123 and toward Arg 85. This positioning of Asp 83 does not favor the binding of Wat 153, and this water molecule is only found in three of the P3₂21 MSCS structures. While a similar conformation of Asp 83 is observed in the P3₂21 downloaded structures, Wat 153 appears more often in those structures, and is always present when a pyrimidine inhibitor is bound in the B₁ pocket.

Wat 16 and Wat 20 bind in or near the P₁ pocket and are both 100% conserved in the P3₂21 MSCS structures (Figure 4b). Similarly, these water molecules are 100% conserved in the P3₂21 downloaded structures, except in the case of Wat 16, when it is missing only when a

bound inhibitor or a formic acid molecule displaces it. Wat 16 and Wat 20 are considered as part of a group of waters that bridge different elements of secondary structure. Through hydrogen bonds with the backbone oxygens of Ala 4 and Val 118, Wat 20 connects α -helix 1 to β -strand 6. Wat 20 also hydrogen bonds with Wat 16, which also links helix 1 through a hydrogen bond with Gln 11 N ϵ 1. These interactions contribute to the stability of the active site. Wat 16 generally interacts with sulfate ions or the phosphate groups of inhibitors bound in the P₁ subsite. In the case of the C2 MSCS structures, Wat 16 forms a hydrogen bond with conserved Wat 19. This water is never present in the P3₂21 MSCS structures because it is displaced by the always present sulfate ion. When not displaced by inhibitors, formic acid, sulfate or chloride ions, this water molecule is found in the P3₂21 downloaded structures.

Wat 34 binds in the B₂ pocket and is 83% conserved in the P3₂21 MSCS structures (Figure 4c). This water hydrogen bonds with Glu 111 O ϵ 2, Asn 71 N δ 2, and Gln 69 N ϵ 2 or O ϵ 1, depending on the orientation of this side chain. This water is displaced when inhibitors bind in the B₂ pocket and some of the interactions are replaced, particularly the hydrogen bond with Asn 71 N δ 2. At the opposite end of the B₂ pocket, Wat 13 is found to be conserved in the C2 MSCS structures, but less than 80% conserved in the P3₂21 structures. This is probably a result of crystal packing. Residues 66 and 67 belong to a loop, which packs up against a symmetry molecule. This packing results in a slight reorientation of these residues and their backbone nitrogens are adjusted slightly away from the orientation observed in the C2 MSCS structures. While Wat 13 in the P3₂21 structures is located a similar distance from these nitrogens as is observed in the C2 structures, the orientations of the nitrogens makes

hydrogen bond formation less favorable. In effect, these water molecules are probably forming a single hydrogen bond with RNase A through Asp 121 O δ 1, causing Wat 13 to be less fixed and less conserved in the P3₂21 structures.

Waters Outside the Active Site

Fifteen of the water binding sites outside the active site are at least 80% conserved in both MSCS sets and in the P3₂21 downloaded structures. Eleven of these waters (4, 6, 10, 15, 17, 21, 39, 63, 64, 69, and 82) bridge areas of secondary structure and two (64, 318) are associated with one of the helices. These conserved water molecules are discussed in detail in Chapter 2. Water molecules 5, 26, and 32 are located in crystal contacts in the C2 MSCS structures. Wat 5 is also bound in an area of crystal contact in the P3₂21 structures and forms a hydrogen bond with the backbone nitrogen of Ser 15. Associating with helix 3, Wat 26 forms a hydrogen bond with a free backbone oxygen at the C-terminal end, and Wat 32 with Ser 50 O γ and Asp 53 O δ 1 at the N-terminal end of the helix.

A number of conserved water molecules in the P3₂21 MSCS structures are associated with one of the helices of RNase A. Four conserved water molecules, 56, 57, 110, and 435, interact with helix 2. At the N-terminal end of this helix, Wat 435 interacts with the O γ s of Ser 21 and Ser 22. Wat 110 interacts with this helix through hydrogen bonds with the backbone oxygen of Asn 27. At the C-terminal end of helix 2, Wat 56 forms a hydrogen bond with the backbone oxygen of Lys 31, and Wat 57 interacts with Ser 32 O γ and Arg 33 N ϵ . Water molecules 56, 110, and 435 are not conserved in the C2 MSCS structures because

they fall in regions of disorder (Lys 37, Lys 31, and 16-22, respectively) where water molecules were not modeled. Water molecules 229 and 344 are both associated with helix 3. Wat 229 forms a hydrogen bond with the backbone oxygen of Gln 55, where Wat 344 hydrogen bonds with O ϵ 1 (or N ϵ 2 depending on the orientation of the end of the side chain) of that same residue.

Three additional water molecules, 123, 166, and 368, are at least 80% conserved in the P3₂21 MSCS structures, but are not associated with the helices. Wat 166 interacts with the C-terminal end of β -strand 1 through a hydrogen bond with the backbone oxygen of His 48. This water is not conserved in the C2 MSCS structures, and there are two possible explanations for this. First, the backbone oxygen of His 48 points in a different direction from that observed in the P3₂21 structures, which disrupts the hydrogen bonding with water molecules found in this position. Second, His 48 lies directly beneath the disordered region 16-22 and waters would not have been modeled around this region. Wat 123 and Wat 368 associate with loops. Water 123 forms a hydrogen bond with the backbone oxygen of Gln 60 in the loop connecting helix 3 and β -strand 2. Wat 368 hydrogen bonds with the backbone nitrogen of Ser 18 in the loop between the first two helices, and corresponds to a previously identified conserved water molecule (Zegers et al., 1994); Wat 144). Because of its association with a residue that falls in the disordered 16-22 region in the C2 MSCS structures, this water is not found to be conserved in the C2 MSCS structures.

Water in Crystal Contacts

A number of conserved water binding positions are located in areas of crystal contacts. Water molecules 2, 18, 23, 78, 90, 158, and 423 are stabilized by crystal contacts in the C2 structures and are not identified as being 80% conserved in the P3₂21 structures. Wat 243, which is conserved in the C2 MSCS structures, is often displaced by a sulfate ion or an organic solvent (organic solvent binding site 929) caught in a crystal contact in the P3₂21 structures preventing this water from being identified as conserved in either the P3₂21 MSCS or downloaded structures. Water molecules 47, 71, 149, 163, 220, 286, 361, 486, and 509 are stabilized by crystal contacts in the P3₂21 structures and are not conserved in the C2 MSCS structures. Additionally, Wat 486 and Wat 509 are not conserved in the P3₂21 downloaded structures. Crystal contacts in the C2 MSCS structures result in water molecules 282, 437, 463, 480, and 483 either being associated with a symmetry molecule or missing from these structures.

Water Molecules Not Conserved in P3₂21 MSCS Structures

A couple of waters are associated with elements of secondary structure, but not conserved in the P3₂21 MSCS structures. Wat 7 associates with helix 3 through a hydrogen bond with the backbone oxygen of Ala 52, but is only 50% conserved in the P3₂21 MSCS structures. This water is displaced by a three organic solvents, a DIO and SRF bound in binding position 926 and symmetry related solvent C5L927. The loss of this water molecule in these three structures alone would be enough to prevent it from being identified as 80% conserved. Wat 129 interacts with the N-terminal end of β -strand 4 through a hydrogen bond with Asn 103

Table 4. Water molecules with less than 80% conservation in the MSCS structures.

Conserved Water Position	SO4 MSCS Conservatn	MSCS Conservatn	P3221 Conservatn	Interaction	Interaction	Interaction
B₁	Pocket:					
60	75%	75%	<50%	Ser 123 N		
153	25%	45%	76%	Asp 83 OD1	Thr 45 OG1	Ser 123 OG
Missing	from	MSCS				
131	<50%	55%	82%	Cys 40 O		
291	58%	10%	100%	Asp 14 OD1/OD2		
443	<50%	<50%	82%	Gln 101 NE2	Asn 103 ND2/OD1	
537	<50%	<50%	100%	Ser 15 O	Thr 17 O	

O δ 1. While this water is conserved in the C2 MSCS structures, the side chain of Asn 103 commonly adopts a conformation that clashes with this water binding in both the P3₂21 MSCS and downloaded structures, preventing it from binding in this position.

Five waters are found to be conserved in the P3₂21 downloaded structures, but not in the MSCS structures (Table 4). The first is Wat 7, which is the only water molecule that is conserved in the C2 MSCS structures (Table 3), and is displaced in three of the P3₂21 MSCS structures by organic solvents. Wat 291 and Wat 537 form hydrogen bonds with Asp 14 O δ 1, and the backbone oxygen of Ser 15, respectively. Both of these water molecules are located in the disordered region of the C2 MSCS structures and would not have been modeled. There is a loose cluster of water molecules for both of these binding positions in both the P3₂21 downloaded and MSCS structures, but in the MSCS structures, water

molecules in these positions are not more than 67% conserved. Wat 131 forms a hydrogen bond with the backbone oxygen of Cys 40, which is located in the loop connecting α -helix 2 to β -strand 1. This water falls in the vicinity of Arg 39, which in the P3₂21 MSCS structures is disordered and this water would not have been modeled. Finally, Wat 443 forms hydrogen bonds with the side chains of residues Gln 101 and Asn 103, which are located in β -strand 5 in or adjacent to the hinge region. In the C2 and P3₂21 MSCS structures, Gln 101 and Asn 103 (respectively) occupy another conformation in a majority of the structures, which destabilizes the water bound in this position.

Organic Solvent Binding Sites

In the 12 MSCS structures in the P3₂21 space group, there are 15 sulfate ions bound in two locations on the surface of RNase A. A sulfate ion occupies the P₁ pocket of the active site of all 12 models, and each has been numbered 903 in these models to correspond with the numbering of the organic solvents bound in this pocket from the MSCS structures in the C2 space group. An additional sulfate ion is modeled into the XLINK, CYH, and SRF structures. This sulfate ion interacts with the backbone nitrogen and O γ of Ser 23 and is bound in an area of crystal contact. This sulfate ion is numbered 929 to correspond with the numbering of organic solvents modeled in this location in other P3₂21 MSCS structures.

In addition to the sulfate ions, 40 organic solvents bind at the surface of RNase A (Table 5 and Figure 5), occupying 21 unique binding sites with 16 of these binding sites being unique from those observed in the MSCS structures in the C2 space group (Figure 6). Solvents

Table 5. Bound Organic Solvents and the interactions made with RNase A. All distances less than 4 Å between an organic solvent atom and an atom of RNase A, a water molecule, or a nearby organic solvent are listed. Atoms belonging to a symmetry-related molecule are designated by (sym).

Organic Solvent	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance
BIS901	O1	Ser 123 OG	3.26	O1	Val 124 CG2 (sym)	3.80	O2	Asp 83 OD1	2.47	O2	Arg 85 NH1	3.30	O3	Asp 121 O	3.57
	C3	Lys 66 CE (sym)	3.50	C4	Thr 45 OG1	2.96	C4	Val 43 CG1	3.38	C5	Ser 123 OG	3.07	C5	Ala 122 CB (sym)	3.79
	C5	Ser 123 O (sym)	2.89	C5	BIS 908 C5 (sym)	2.88	C6	Ala 122 CA	3.49	C6	BIS 908 C6 (sym)	3.18			
SO4 903	O1	Gln 11 NE2	2.83	O1	HOH 16	3.30	O1	HOH 544	3.12	O2	His 12 NE2	3.63	O2	Lys 41 NZ	3.01
	O2	HOH 552	2.77	O3	His 12 NE2	2.73	O3	Phe 120 N	2.99	O3	HOH 16	2.85	O4	His 119 ND1 (conf A)	2.42
	O4	His 119 ND1 (conf A)	3.33												
HEZ904	O1	Asn 67 N	2.94	O1	Lys 66 N	3.33	O1	Asp 121 OD1	2.99	O1	Asn 67 CB	3.52	C1	His 119 NE2	3.57
	C1	Asn 67 CG	3.65	C2	Asn 67 CG	3.39	C2	Asn 67 OD1	3.10	C2	Asn 67 ND2	3.58	C3	His 119 CD2	3.62
	C3	Gln 69 OE1	3.84	C4	His 119 CG	3.46	C4	Ala 109 CB	3.62	C6	Gln 69 NE2	3.43	O6	Glu 111 OE2	2.59
	O6	Asn 71 ND2	2.87												
CPL905	OAA	Asn 62 OD1	3.09	OAA	Thr 70 O	3.86	OAA	Thr 70 CB	3.54	OAA	Thr 70 CG2	3.82	CAF	Asn 62 OD1	3.25
	CAD	Gly 88 O (sym)	3.18	CAB	Ser 90 O (sym)	3.59	CAB	Ser 90 CB (sym)	3.90	CAC	Ala 64 CB	3.94	CAC	Thr 70 CG2	3.64
	CAE	Val 63 C	3.61												
CPN905	O1	Asn 62 OD1	2.98	O1	Thr 70 O	3.39	O1	Thr 70 CB	3.53	C3	Thr 70 CG2	3.53	C3	Ala 64 CB	3.93
	C2	Glu 86 CB (sym)	3.55	C1	Ser 90 CB (sym)	3.78	C1	Thr 87 O (sym)	3.80	C1	Gly 88 O (sym)	3.64	C5	Gly 88 O (sym)	3.43
CXN905	C2	Asn 62 OD1	3.46	C2	Gly 88 O (sym)	3.76	C3	Val 63 O	3.92	C4	Ala 64 N	3.45	C4	Glu 86 C (sym)	3.78
	C5	Ala 64 CB	3.86	C5	Thr 70 CG2	3.87	C6	Ser 90 O (sym)	3.66						
DMS905	O	Asn 62 OD1	2.87	O	Thr 70 O	3.55	C1	Ala 64 CB	3.56	C1	Thr 70 CG2	3.72	C2	Ser 90 O (sym)	3.94
	C2	Thr 87 O	3.89	S	Thr 87 O (sym)	3.43	S	Thr 87 C	3.76						

Table 5 continued.

Organic Solvent	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance
DMS924	O	Ala 4 N	3.03	O	Thr 3 CB	3.19	O	Thr 3 CA	3.28	O	HOH 69	2.77	C1	Ala 4 CB	3.92
	C2	Ser 23 N (sym)	3.40	S	Ser 23 OG (sym)	3.42									
DOX924	O1	Ser 23 N (sym)	3.35	C2	Ser 23 OG (sym)	3.11	C2	HOH 69	3.27	C1	Ala 4 CB	3.76	O2	Ala 4 CB	3.27
	O2	Ala 4 N	2.91	C4	DOX 2071 C4	3.92	C3	DOX 2071 O2	3.54	C3	DOX 2071 C4	3.52			
DOX925	O1	Thr 70 CG2 (sym)	3.34	O1	Ala 64 CB (sym)	3.76	C2	Val 63 C (sym)	3.49	C1	Asn 62 OD1	3.21	O2	Gly 88 N	2.93
	O2	Gly 88 O	3.45	C4	Gly 88 O	3.33	C4	Ser 90 O	3.98	C3	Glu 86 CB	3.80			
IOH925	O	Glu 86 O	3.52	O	Gly 88 N	3.13	O	Val 63 C (sym)	3.79	O	Asn 62 O (sym)	3.96	CA	Thr 70 CB (sym)	3.92
	CB1	Asn 62 OD1 (sym)	3.49	CB2	Thr 70 CG2 (sym)	3.64									
BIS926	O2	Gln 55 OE1	3.86	O3	HOH 229	2.71	C3	Gln 55 CB	3.83	C5	Ala 52 CB	3.83	C5	BIS 902 O1 (sym)	2.83
	C6	Ser 16 CB	3.94	C6	Ala 52 O	3.57									
DOX926	O1	Gln 55 OE1	3.51	O1	Gln 55 CB	3.88	C1	Ala 56 CB	3.74	C1	Ala 52 O	3.37	C1	Ser 16 OG (sym)	3.59
CPL927	OAA	Ser 16 OG	3.37	OAA	HOH 291	3.40	OAA	Ala 52 O (sym)	3.77	CAF	Ser 16 O	3.96	CAD	HOH 370	3.14
	CAB	Ala 52 CB (sym)	3.74	CAE	Ala 52 O (sym)	3.27	CAE	Ala 56 CB	3.98						
BIS928	O1	Lys 31 NZ	3.47	O2	Lys 31 CD	3.13	O3	Asn 24 ND2	2.60	C3	Asn 27 CB	3.78	C3	Gln 28 CG	3.99
CPL928	OAA	Asn 24 ND2	3.64	CAF	Asn 27 C	3.96	CAD	Asn 27 O	3.58	CAB	Lys 31 CD	3.68	CAE	Asn 27 CB	3.54
CPN928	O1	Asn 24 ND2	3.39	C1	Lys 31 CD	3.54	C1	Asn 27 O	3.75	C3	Asn 27 CB	3.76			
IOH928	O	Asn 27 O	3.74	CB1	Asn 27 CB	3.65	CB1	Asn 24 ND2	3.65	CB2	Lys 31 CD	3.56			
CPN929	O1	Ser 23 N	3.23	C4	Ser 23 CB	3.56	C5	Ser 23 OG	3.44	C5	HOH 157 (sym)	3.24	C1	HOH 69 (sym)	2.98
	C1	Ala 4 CB (sym)	3.53	C2	Ala 4 N (sym)	3.52	C3	CPX 903 C5	3.79						

Table 5 continued.

Organic Solvent	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance	Solvent Atom	Protein Atom	Distance
DMF936	O	Met 79 CE (sym)	3.52	O	HOH 82 (sym)	2.71	O	HOH 10 (sym)	2.80	N	Ser 32 CB	3.81	N	Met 79 CG (sym)	3.98
	C	Ser 32 O	3.13	C	Ser 77 CB (sym)	3.69	C1	Ser 32 O	3.12	C1	Thr 78 O (sym)	3.23	C1	Glu 49 OE2 (sym)	3.56
	C2	Ser 32 OG	3.88	C2	Asp 53 OD2 (sym)	3.14	C2	Glu 49 OE1 (sym)	3.45						
CXL937	O	Asn 62 O	3.79	O	HOH 318	3.39	O	Leu 51 CD2 (sym)	3.48	C2	Leu 51 CD2 (sym)	3.74	C2	Gln 55 NE2 (sym)	2.92
	C2	Lys 61 NZ	2.72	C3	Gln 55 NE2 (sym)	2.76	C5	Ser 89 OG (sym)	2.78	C6	Gly 88 N (sym)	3.00			
CXN937	O1	Asn 62 N	2.88	C1	Asn 62 CB	3.76	C1	Leu 51 CD2 (sym)	3.56	C3	Gln 55 OE1 (sym)	3.38	C4	Gln 55 OE1 (sym)	3.42
	C4	Gly 88 CA (sym)	3.75	C5	Lys 61 NZ	3.15	C5	Gln 55 NE2 (sym)	3.55	C6	Lys 61 CD	3.63	C6	Thr 87 OG1 (sym)	3.44
DMF937	O	Leu 51 CD2 (sym)	3.38	O	HOH 344 (sym)	2.73	C	Lys 61 NZ	2.82	C	Leu 51 CD2 (sym)	3.70	C	Gln 55 NE2 (sym)	3.12
	C1	Gln 55 NE2 (sym)	3.54	C1	Ser 89 OG (sym)	3.66	C2	Ser 89 CB (sym)	3.41	C2	Thr 87 OG1 (sym)	3.39	C2	Ser 89 N (sym)	2.96
	C2	Gly 88 N (sym)	3.32												
CXL938	O	Thr 99 N	2.86	O	HOH 463 (sym)	2.71	C1	Ser 23 O	3.00	C2	Tyr 97 O	3.26	C5	Ser 23 OG	3.25
	C5	CXL 2266 C2 (sym)	3.37	C6	Ser 23 CB	3.65	C6	Ala 5 CB (sym)	3.95						
DMF938	O	Thr 99 N	2.77	O	HOH 463 (sym)	2.71	N	Ser 23 CB	3.68	C	Tyr 97 O	3.38	C	Ser 23 O	3.01
	C2	Ala 5 CB (sym)	3.70												
CXL939	C1	Tyr 115 CZ	3.64	C3	Tyr 115 CD2	3.63	C5	Tyr 115 CG	3.81						

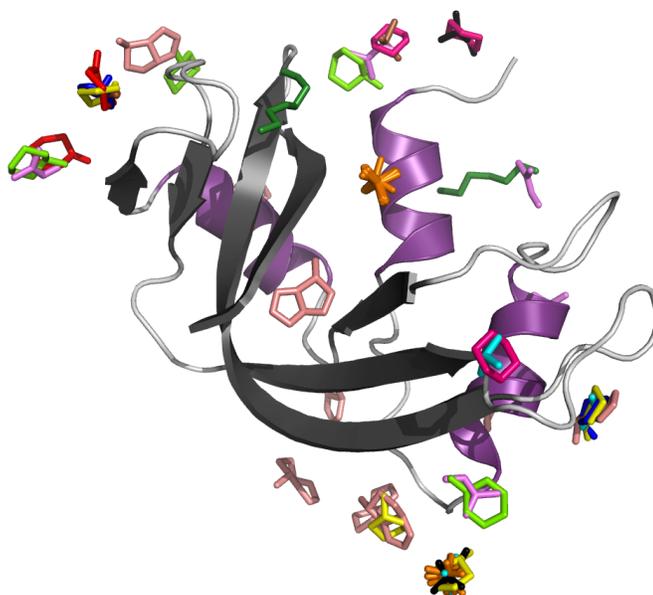


Figure 5. Organic solvent binding sites. Ribbon diagram of RNase A showing the binding sites for organic solvent molecules. The β -strands are shown in dark gray and the α -helices are shown in purple. Sulfate ions are colored orange and the organic solvent molecules are colored as follows: CYH, red; CXL, light green; C5N, yellow; C5L, blue; DIO, hot pink; DMF, violet; DMS, brown; HEZ, forest green; IPA, cyan; SRF, salmon; and ETF, black.

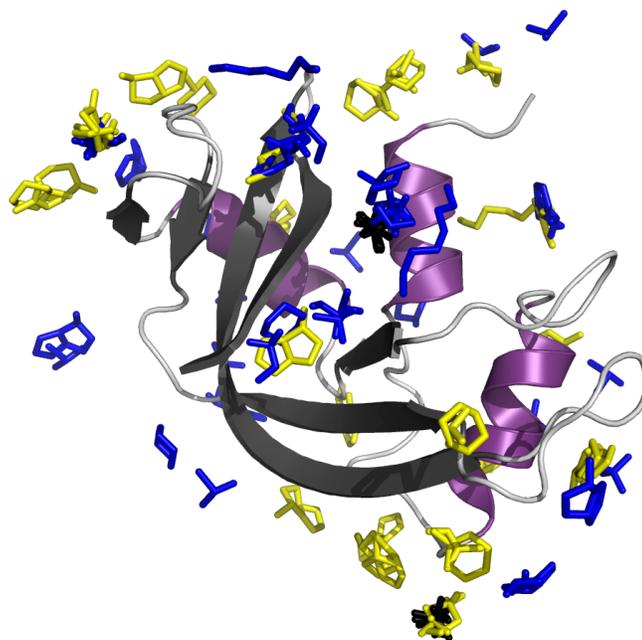


Figure 6. Organic solvent binding sites across all MSCS structures. Ribbon diagram of RNase A showing the binding sites for organic solvent molecules observed in both the $P3_221$ and $C2$ space groups. The β -strands are shown in dark gray and the α -helices are shown in purple. The organic solvent molecules from the models in the $C2$ space group are colored blue and those from the $P3_221$ space group are colored yellow. The sulfate ions from the $P3_221$ models are colored black.

occupying binding sites observed in the C2 space group structures are numbered 901-923 and the binding sites unique to the P3₂21 space group are designated at 924-939. Two of the organic solvents bind in the active site of RNase A, one in the B₁ pocket and one in B₂, and eight solvents occupy additional binding sites observed in the MSCS structures in the C2 space group.

Solvents Bound in the Active Site

Unlike the MSCS structures in the C2 space group, there are very few organic solvents bound in the active site of RNase A, and there is an ever-present sulfate ion in the P₁ pocket, bound between the catalytic histidines, His 12 and His 119 (Figure 7a). Sulfate ions bound in the P₁ pocket have been numbered 903 to be consistent with the numbering of the organic solvents that bind in this pocket in the C2 MSCS structures. The oxygen atoms of this ion form hydrogen bonds with the backbone nitrogen of Phe 120, Lys 41 N ζ , Gln 11 N ϵ 2, and His 119 N δ 1. The O3 atom of the sulfate ion is within hydrogen-bonding distance of His 12 N ϵ 2 (2.73 Å in the XLINK structure), however the orientation is not optimal for this hydrogen bond. Additionally the O3 atom of the sulfate ion forms a hydrogen bond with Wat 16, which has been previously observed to interact with the P₁ phosphate or sulfate groups of inhibitors in RNase A–inhibitor complexes (Zegers et al., 1994). When compared to the organic solvents bound in the P₁ pocket from the C2 MSCS models (Figure 7b), it is observed that sulfate ion binds deeper in the P₁ pocket than the organic solvents, and it forms better interactions with the active site residues, Gln 11, His 12, Lys 41, and His 119. Additionally, the sulfate ion bound in this position displaces Wat 19, which is found to be

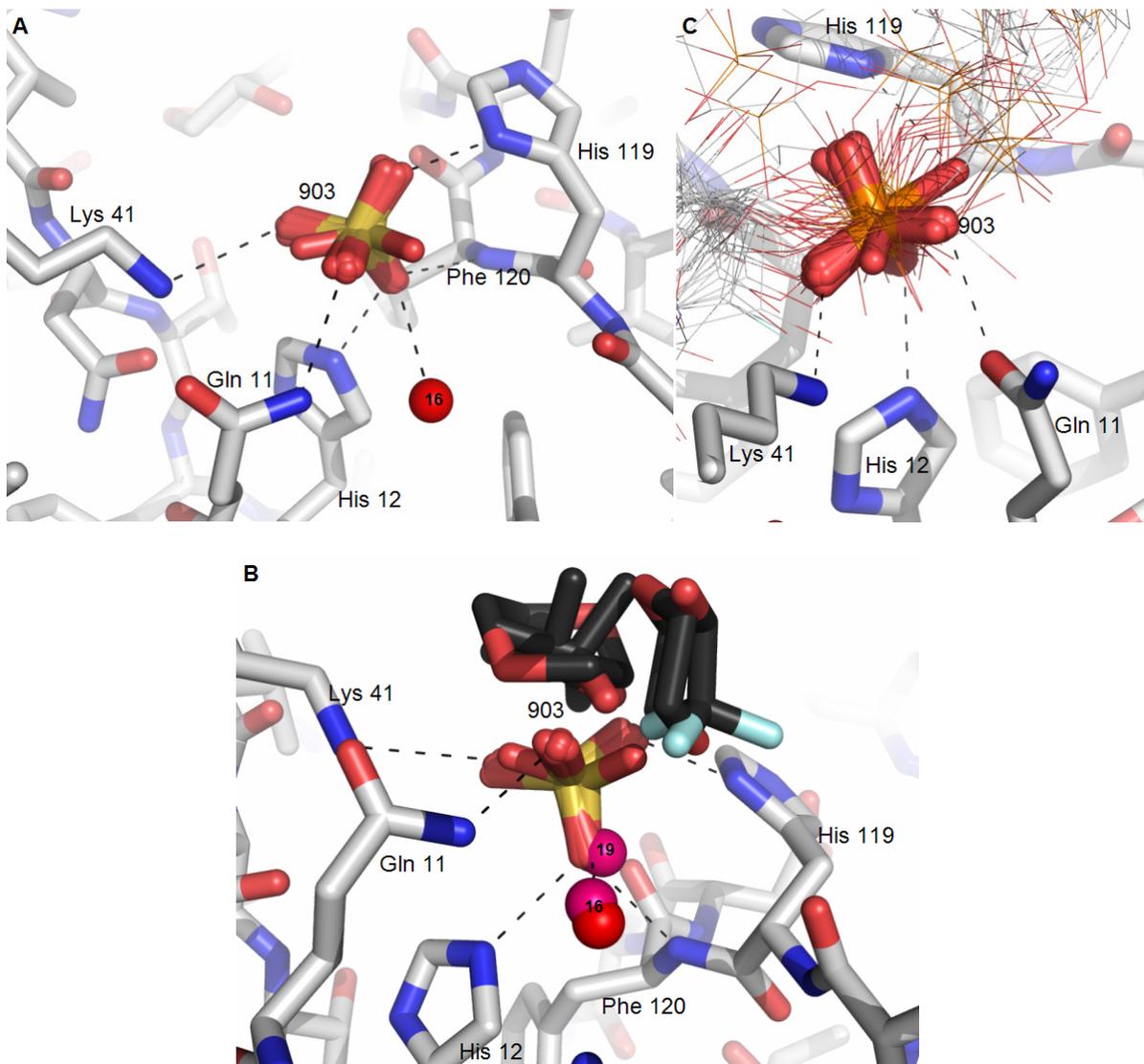


Figure 7. Sulfate ion binding in the P_1 pocket of the active site. RNase A is drawn with sticks with carbon atoms colored light gray, all of the superimposed sulfate ions are drawn as sticks and colored by atom type (sulfur, orange; oxygen, red), and representative conserved water molecules are colored red. Comparison organic solvent molecules from the C2 MSCS structures are drawn with sticks and carbon atoms colored dark gray, representative conserved water molecules from these models are colored hot pink. Inhibitor molecules are drawn as lines and phosphorus atoms are colored orange. A black dashed line designates interactions between the sulfate ions and water molecules or protein. A. Sulfate ions bound in the P_1 pocket in the P3₂21 MSCS models of RNase A. B. Organic solvents and sulfate ions bound in the P_1 pocket of RNase A. C. Sulfate ions and inhibitors bound in the P_1 pocket of RNase A

conserved in the C2 MSCS models, but displaced by inhibitors binding in the P₁ pocket. This sulfate ion binds in the P₁ pocket similarly to the phosphate moiety of a substrate or inhibitor molecule (Figure 7c). The sulfur atom overlaps well with the sulfur or phosphorus atoms of inhibitors and the oxygen atoms make interactions with the residues of the P₁ pocket, similar to the oxygen atoms of sulfate or phosphate moieties of larger bound molecules.

The first organic solvent bound in the active site is SRF901, which binds in the B₁ pocket of the active site of RNase A (Figure 8a). SRF901 displaces Wat 60, as seen with organic solvents, DIO901 and ETF901, bound in this position in the C2 MSCS structures. Unlike the solvents observed in the C2 MSCS structures, the O2 atom of SRF901 forms hydrogen bonds with Asp 83 O δ 1 and Arg 85 N η 1. These interactions are most likely the result of crystal contacts. In the P3₂21 structures, the Asp 83, Arg 85, and SRF901 are all located in an area of crystal contacts, and the conformations of Asp 83 and Arg 85 observed in these structures are favored because of these contacts. Asp 83 points away from Ser 123 and toward Arg 85, allowing the O δ 1 atom to form a hydrogen bond with SRF901. Additionally, in the P3₂21 structures, Arg 85 adopts a conformation oriented closer to Asp 83 and the B₁ pocket than is observed in the C2 structures. In the SRF structure, this positions the N η 1 of Arg 85 for hydrogen bonding with SRF901. In the C2 structures, neither the solvents nor Asp 83 are located in crystal contacts, however Arg 85 in molecule B is. While this restricts the motion of Arg 85, it does not adopt a conformation as close to the B₁ pocket as is observed in the P3₂21 structures. With no interactions in common with other solvents bound in the P₁

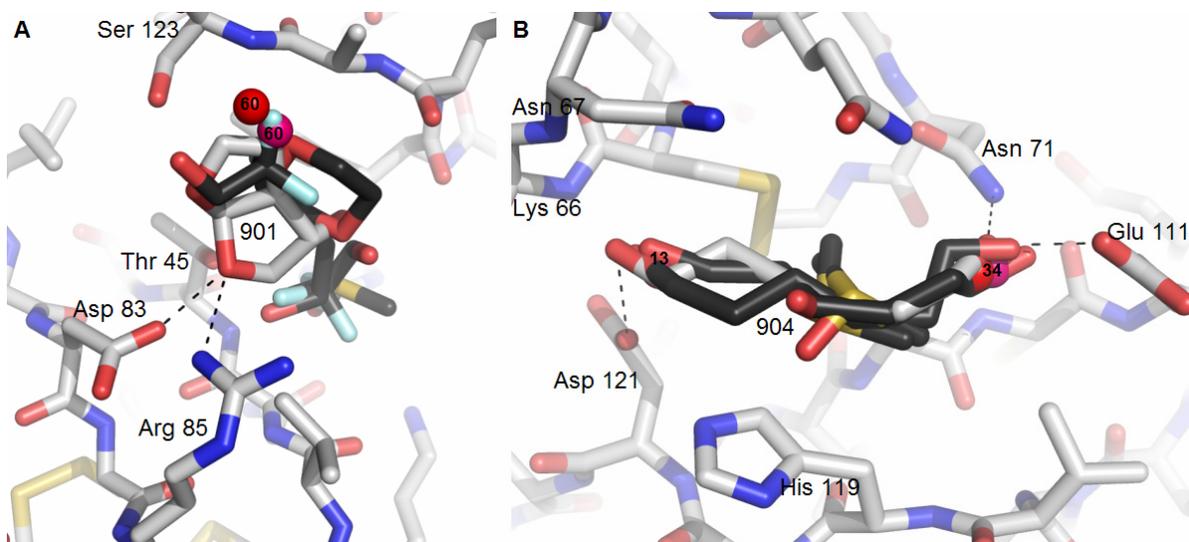


Figure 8. Organic solvent and inhibitor binding in the active site. RNase A and organic solvents from the P3₂21 space group are drawn with sticks with carbon atoms colored light gray, and representative conserved water molecules are colored red. Comparison organic solvent molecules from the C2 MSCS models are drawn with sticks and carbon atoms colored dark gray, and representative conserved water molecules from these models are colored hot pink. A black dashed line designates interactions between the solvents and water molecules or protein. A. The B₁ pocket B. The B₂ pocket

subsite, it seems likely that the binding of SRF901 is more a result of crystal contact than interactions with active site residues.

The second organic solvent bound in the active site is HEZ904, which binds in the B₂ pocket of RNase A (Figure 8b). HEZ904 binds in a similar fashion to HEZ904 and HEZ994 of the C2 MSCS structures. As observed in the C2 MSCS structures, the O6 atom (O1 in the C2 structures) of HEZ904 displaces Wat 34 and forms hydrogen bonds with Asn 71 Nδ2 and Glu 111 Oε2. The O1 atom (O6 in the C2 structures) of HEZ904 displaces Wat 13, which is only 67% conserved in the P3₂21 structures, and reproduces the interactions made by this water molecule by forming hydrogen bonds with Oδ1 of Asp 121. As seen with the HEZ

molecules of the C2 structures, HEZ904 stacks on top of His 119, similar to the inhibitor molecules binding in the B₂ pocket of the RNase A active site.

Solvents Bound in Crystal Contacts

There are 13 binding sites of organic solvents on the surface of RNase A that do not fall in the active site of the protein and are located in areas of crystal contact. Of these, three binding sites coincide with sites of organic solvent binding in the C2 MSCS structures and include eight solvent molecules: C5L905, C5N905, CYH905, DMS905, DMF907, HEZ907, DIO913, and ETF913. The remaining ten sites in crystal contacts include seven clusters and three single solvent binding sites consisting of CXL924, DMF924, DMS924, DIO924, DIO925, IPA925, DIO926, SRF926, C5L927, C5N929, ETF929, IPA929, SRF930, C5N930, SRF931, DMF936, CXL937, CYH937, DMF937, CXL938, DMF938, and the three sulfate ions 929 from XLINK, SRF, and CYH. Many do not displace conserved water molecules or overlap with previously observed bound molecules, but there are some exceptions. HEZ907 displaces Wat 18 and Wat 286, and DMF907 displaces Wat 71 and overlaps with chloride ions observed in the 1RNY, 1RAR, 1RNZ, and 1RNX models in the P3₂21 space group. Both of these molecules overlap with a citrate molecule as observed in the C2 MSCS models. Binding sites 924 and 929 are symmetry related and both overlap with sulfate ions from P3₂21 models: 1RNM (929), 1RNO (924), 1RNW (924), and G4X (924). While solvents in binding site 929 do not displace water molecules, CXL924 and DMF924 displace Wat 69, CXL924 displaces Wat 39, and DMF924, DMS924, and DIO924 displace Wat 243. Finally solvents in four binding sites displace four water molecules: SRF926 and DIO926 displace

Wat 7, DMF936 displaces Wat 486, CYH937 displaces Wat 318, and CXL938 and DMF938 displace Wat 63.

Other Surface Solvents

There are nine organic solvents that are found in six binding sites that are outside of the active site and not involved in crystal contacts. None of these solvents overlap with previously observed binding molecules nor do they displace any conserved water molecules. Most of these binding sites occur in small shallow pockets on the surface of the protein. In two such pockets on the surface of RNase A, SRF932 and SRF933 each bind near the loop region encompassing residues 18-21, which is partly disordered in the C2 MSCS structures. Both of these solvents interact with Ser 18 with the O2 atom of SRF932 forming a hydrogen bond with the backbone oxygen of Ser 18, and the O1 atom of SRF933 forming a hydrogen bond with the O γ . Within 8Å of both of these SRF molecules, SRF935 binds in another shallow pocket, where it potentially forms a hydrogen bond between its O2 atom and the backbone oxygen of Thr 17. A cluster of organic solvents, including SRF928, C5L928, C5N928, and IPA928, binds in a small shallow pocket right at the edge of a small deeper pocket, which is large enough to accommodate Wat 110. This pocket is fairly rigid across the P3₂21 structures and there is the potential for the solvent molecules to form a hydrogen bond with Asn 24 N δ 2. Four Å away from the 905 cluster, SRF934 occupies a fifth shallow pocket located outside the B₂ pocket and forms a hydrogen bond between its O1 atom and Tyr 115 O η . Finally, approximately 3Å away from SRF934, CXL939 binds on the surface of RNase A. Unlike the previous five binding sites, CXL939 does not bind in a shallow

pocket, instead the ring of the CXL stacks on top of the ring for Tyr 115, and the oxygen atoms of both of these rings orient in approximately the same direction.

Discussion

The initial challenges in the MSCS of RNase A provided a unique opportunity to compare the solvent mapping results of a protein from crystals in two different space groups.

Additionally, the sulfate anion in the P₁ subsite of the active site cleft in the P3₂21 structures provided a small molecule mimic of a substrate bound in the active site, and served as an additional small molecular probe. This mimic offered the contrasting bound state to the apo-RNase A in the C2 space group with which to examine global plasticity observable in MSCS structures with the hinge motions of RNase A.

Protein Plasticity

Aside from the differences caused by crystal contacts in the different space groups, the trends in the plasticity of RNase A were the same for the MSCS structures as for the compared structures. The previous observations finding the regions of high plasticity primarily in loops, the rigid regions tending to coincide with areas of secondary structure and the hinge region, and domain B having higher plasticity than domain A (Kishan et al., 1995; Merlino et al., 2002; Sadasivan et al., 1998), are also observed in the MSCS structures.

The fluctuations of the hinge have been observed with molecular dynamics, and crystal and NMR structures (Beach et al., 2005; Kishan et al., 1995; Merlino et al., 2002; Sadasivan et

al., 1998; Vitagliano et al., 2002), and a handful of similar metrics have been used to measure the fluctuations of the hinge. The two used in this study was the measure of the hinge angle (Kishan et al., 1995; Sadasivan et al., 1998) and the measure of the distance between the Thr 45 N and Phe 120 N (Vitagliano et al., 2002). An alternative method to calculating the hinge angle is to superimpose two structures on a single domain, and then find the degree of rotation to superimpose on the second domain. This method produces results similar to the difference between the hinge angles for two structures (Vitagliano et al., 1998). The drawback of using rotation to examine the hinge angle is that there must be a reference structure, where calculating the hinge angle considers each molecule independently of one another. The hinge angle calculation used in this study considered the C α positions for the whole protein in determining the center of mass for each domain as was done in a pair of previous studies (Kishan et al., 1995; Sadasivan et al., 1998). Another group used the hinge angle as a metric, but only considered the C α atoms of the residues involved in β_1 , β_2 , and seven residues of the hinge (Vitagliano et al., 1998; Vitagliano et al., 2002). The β -sheet hinge angle was applied to six molecules of RNase A (Vitagliano et al., 2002), and when compared to the hinge angle calculated with the whole domain C α s, the results were within 2.2° for five of the six structures and the last structure saw a difference of 2.9°. The results are similar in a crystal environment, but this is most likely not the case in solution structures. In contrast to the global approach of the hinge angle calculation, the distance between Thr 45 N and Phe 120 N is localized to the active site. While these residues lie on opposite β -sheets, they are both located in the active site and are involved in ligand binding.

The differences in these two metrics become more apparent when the crystal and NMR structures are compared (Figure 3a). While the NMR structures have Thr 45 N – Phe 120 N distances in the larger half of the range, there is significant overlap of crystal structures and NMR structures found in this range, and the MSCS structures sample most of the range of this distance observed in both the crystal and NMR structures. The higher distance values observed in the NMR structures is not surprising because they reflect a more open conformation of the hinge, which is expected in the absence of sulfate anion or substrate, as is the case in the NMR structures. The distribution of the hinge angles is different. Again, the NMR structures are found to have values in the upper half of the range, but there is very little overlap between the solution and the crystal structures. Presumably, this is a result of the hinge angle calculation accounting for the residues of entire domains as opposed to only the β -sheets. In this respect, the hinge angle reflects the hinge movements over the whole molecule, as opposed to just the active site cleft. As the NMR structures do not experience the damping effects of crystal contacts, the greater molecular movements illustrated here are not surprising. However, it would be interesting to see how the hinge angle calculated using only the β - sheet compares.

When the crystal structures of the C2 and P3₂21 space groups are compared (Figure 3b), it is found that the P3₂21 space group samples a tighter range of hinge conformations. This is probably a result of the P3₂21 structures having only one protein molecule in the asymmetric unit, where the C2 structures have two, and one molecule has a more closed hinge conformation and the other has a more open one (Vitagliano et al., 2002). The P3₂21 MSCS

structures adopt a more closed conformation than most of the P3₂21 downloaded structures. This is not entirely surprising because all of the P3₂21 MSCS structures have a sulfate anion in the active site, which causes RNase A to adopt a more closed conformation, and the downloaded structures have a variety of states, with only a few of them having an inhibitor molecule or a sulfate anion bound in the active site. The C2 MSCS structures sample the hinge conformations of the C2 downloaded structures comparably well, which is interesting because at least one molecule in each of the downloaded structures has an inhibitor molecule bound. This could be explained by the earlier observation that the apo form of RNase A samples the inhibitor bound conformation and vice versa (Beach et al., 2005).

From the results of a molecular dynamics study, it was proposed that the mixed $\alpha/3_{10}$ helix 3 functioned as a mechanical hinge for RNase A (Merlino et al., 2002). This helix, consisting of residues 50-60, is linked directly to β_1 through the 58-110 disulfide bond and to β_2 through the 47-50 loop. In these motions, it was observed that helix 3 bends around the midpoint (residue 56-57), which coincides with the change from an α to a 3_{10} helix (Merlino et al., 2002). Knowing that variation of the hinge angle is observed across the MSCS structures of RNase A, it is interesting to highlight that while residues 50-58 (50-57 in the C2 MSCS structures) for helix 3 display baseline average backbone RMSD values, the later residues in this helix do not, which is consistent with the observations from the molecular dynamics study.

The superimposed MSCS structures highlight areas of local and global protein plasticity that are observed in NMR and molecular dynamics studies, and across the set of crystal structures of RNase A. Additionally, the MSCS structures in two different space groups show similar trends and the differences are the result of crystal contacts.

Hydration

The SEWS (Structurally Equivalent Water System) program (Bottoms et al., 2006) was used to identify the conserved water binding positions in the MSCS and the downloaded P3₂21 structures. Water molecules must make at least one common interaction with the protein with a distance of 3.4 Å or less and occupy the same position in at least 80% of the structures to be identified as conserved. To be considered as occupying the same position, water molecules from the superimposed structures must fall within the same 1.4 Å radius sphere. The SEWS program identified 42 water molecules to be at least 80% conserved in the P3₂21 MSCS structures, 29 in the P3₂21 downloaded structures, and 31 in the C2 MSCS structures (Table 3). The conserved water molecules common among all three sets of structures tended to be found in the active site, or on the surface of the protein and involved in bridging elements of secondary structure or fulfilling hydrogen-bonding requirements of surface residues. The differences between the MSCS structures in the two space groups tended to be a result of crystal contacts. Finally, the differences between the P3₂21 MSCS and P3₂21 downloaded structures were usually the result of the water molecules being located near disordered side chains and therefore not modeled.

The conserved water molecules located in the active site have been of particular interest to those studying RNase A. For example, Wat 22, which is conserved in the P₃₂₁ MSCS structures, and displaced by organic solvents (numbered 902) binding in the B₁ subsite, has been observed to be displaced by inhibitors binding in this pocket, and when the inhibitor is soaked out of the crystal, this water is observed again (Vitagliano et al., 2000, 2002). The other two water molecules of interest in the B₁ subsite are not identified as conserved in the MSCS structures. Wat 60 just misses the that qualification because it is only 75% conserved in the MSCS structures and Wat 153 is less than 50% conserved in the MSCS structures, where the Asp 83 side chain often adopts a conformation that can not form hydrogen bonds with this water. The interesting thing about these two waters is that they are commonly observed to bridge the interaction of nucleotide bases in the B₁ subsite and the residues in this pocket, and have been identified as changing their roles of hydrogen bond donor or acceptor depending on presence of a cytidine or uracil in the B₁ subsite (Gilliland et al., 1994). A group working on designing specific inhibitors for RNase A has suggested that by exploiting the position of Wat 153 by adding a functional group that would reach into this position and make the same interactions with the protein, it might be possible to design a more potent and specific inhibitor (Leonidas et al., 2003). In the P₁ subsite, Wat 16 and Wat 20 bridge helix 1 and strand 6 (part of β_1), adding stability to the active site. Wat 19 is not observed in the P₃₂₁ MSCS structures because it displaced by an oxygen of the sulfate anion, which replaces the hydrogen bond interactions made by Wat 19. This is also observed in crystal structures where an inhibitor molecule is soaked in and the water is displaced by the phosphate moiety of the inhibitor. In the structures where the inhibitor has been retro-

soaked out, Wat 19 is present (Vitagliano et al., 2002). In the C2 MSCS structures this water molecule is not displaced by organic solvents, instead the solvent molecules are observed to interact with this water molecule.

While there are differences, which are mostly the result of crystal contacts, the C2 and P3₂21 MSCS structures pick out similar patterns of conserved hydration on the surface of RNase A. As observed in previous studies (Kishan et al., 1995; Sadasivan et al., 1998; Zegers et al., 1994) and in sets of comparison structures (Chapter 2), most of these conserved water molecules play structural roles or are bound in the active site. More importantly, many of the active site waters identified as conserved in the MSCS structures highlight direct interactions substrate molecules make with RNase A.

Binding of Organic Solvents

A combination of problematic solvent soaks and a sulfate anion bound in the P₁ pocket of the active site made clustering of organic solvent molecules in the active site spotty. While this is unfortunate, it presented a new small molecular probe in the active site.

The analysis of the organic solvents bound in C2 MSCS structures, revealed three binding sites or hot spots in the active site of RNase A (Chapter 2). Comparing the molecules that bind in the P3₂21 structures mostly confirms these hot spots, with the exception of the hot spot in the B₁ pocket. The hot spot in the B₁ subsite is characterized by interactions with the backbone nitrogen and the O_γ1 of Thr 45, hydrogen bonding with the backbone nitrogen and

O γ of Ser 123, either directly or through a bridging water, and finally, interactions with Ser 123 O γ , Asp 83 O δ 1, and Thr 45 O γ 1 bridged by Wat 153. While bound in the B₁ subsite, SRF901 makes none of these interactions, however, it is likely that this binding is more the result of crystal contacts as opposed to the interactions in the subsite.

The sulfate ion is bound universally in the P₁ subsite of the P₃₂21 MSCS structures and superimposes in nearly identical positions and orientations. It forms the hydrogen bonds defining the hotspot in the P₁ subsite: with Wat 16, His 119 N δ 1, Gln 11 N ϵ 2, and His 12 N ϵ 2. Additionally, the sulfate anion superimposed well with inhibitors bound in this pocket (Figure 7c) and formed the same interactions with the protein as observed with the inhibitor molecules.

HEZ904 binds in the B₂ subsite, similar to HEZ904 and HEZ994 of the C2 MSCS structures. The interactions made by this molecule highlight a couple of the interactions defining the hot spot in this pocket: hydrogen bonding with Asn 71 N δ 1 and stacking with the imidazole ring of His 119. A curiosity in this pocket develops when it is noted that organic solvent binding in this pocket is not obscured by crystal contacts or the sulfate ion, as evidenced by the binding of HEZ904 in the P₃₂21 structures. However, there is no binding of any DMS molecules in this pocket as is observed in both RNase molecules in the C2 MSCS structures. This can be explained by the DMS soaks of the P₃₂21 crystals only reached 25% DMSO where the C2 crystals were soaked in 70% DMSO. The only two DMS molecules bound in the P₃₂21 structures were located in crystal contacts. It could be speculated that if the

concentration of this solvent could have been increased, a DMS molecule would bind in the B₂ pocket at a higher solvent concentration.

Conclusions

The comparison of the MSCS structures in the C2 and P3₂21 space groups revealed that space group does play a role in the results found in each set for plasticity, hydration, and binding sites of the organic solvents, but the role it plays is expected. First, crystal packing has a damping influence on overall molecular fluctuations, but molecular motions are observable in the MSCS structures. Second, crystal contacts have a stabilizing influence on some loops or residues, and these vary by space group. Finally, some binding sites of surface water molecules or organic solvents are stabilized by crystal contacts, and this varies by space group. These are not surprising or unexpected effects, but this highlights the necessity of considering the role of crystal contacts in structural analysis of MSCS.

Acknowledgements

Many thanks to Crystal Cholewa for growing, cross-linking, and soaking crystals of P3₂21 RNase A and the members of the spring 2005 BCH 590M course (Crystal Cholewa, Kelly Daughtry, Alison Fraser, Stephanie Holsenbeck, Amit Kumar, Chris Miller, Andrea Moon, and Jad Walters for their work in the initial stages of refinement for a number of RNase A structures with a sulfate ion in the active site. We appreciate all the discussions and assistance, which helped start the computational analysis, provided by Janet Thornton and Roman Lakowski. Christopher Bottoms generously provided the SEWS program and

assistance in learning the software. Data were collected at the Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Supporting institutions may be found at www.ser-cat.org/members.html. Use of the Advanced Photon Source was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. W-31-109-Eng-38. We are grateful to the staff at the SER-CAT beamline for their guidance during data collection. This research was supported by a grant from the National Science Foundation under grant number 0237297.

References

- Allen, K.N., Bellamacina, C.R., Ding, X., Jeffery, C.J., Mattos, C., Petsko, G.A., and Ringe, D. (1996). An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J Phys Chem* *100*, 2605-2611.
- Beach, H., Cole, R., Gill, M.L., and Loria, J.P. (2005). Conservation of μ s-ms enzyme motions in the apo- and substrate-mimicked state. *J Am Chem Soc* *127*, 9167-9176.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res* *28*, 235-242.
- Borkakoti, N., Moss, D.A., and Palmer, R.A. (1982). Ribonuclease A: Least squares refinement of structure at 1.45 Å resolution. *Acta Crystallogr B* *38*, 2210-2217.
- Bottoms, C.A., White, T.A., and Tanner, J.J. (2006). Exploring structurally conserved solvent sites in protein families. *Proteins* *64*, 404-421.
- Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., *et al.* (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* *54*, 905-921.
- DeLano, W.L. The PyMOL Molecular Graphics System (Palo Alto, CA, USA, DeLano Scientific). <http://www.pymol.org>

- Dennis, S., Kortvelyesi, T., and Vajda, S. (2002). Computational mapping identifies the binding sites of organic solvents on proteins. *Proc Natl Acad Sci U S A* *99*, 4290-4295.
- Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* *60*, 2126-2132.
- Gilliland, G.L., Dill, J., Pechik, I., Svensson, L.A., and Sjolín, L. (1994). The active site of bovine pancreatic ribonuclease: an example of solvent modulated specificity. *Protein and Peptide Letters* *1*, 60-65.
- Hammes, G.G. (2002). Multiple conformational changes in enzyme catalysis. *Biochemistry* *41*, 8221-8228.
- Kishan, R.V., Chandra, N.R., Sudarsanakumar, C., Suguna, K., and Vijayan, M. (1995). Water-dependent domain motion and flexibility in ribonuclease A and the invariant features in its hydration shell. An X-ray study of two low-humidity crystal forms of the enzyme. *Acta Crystallogr D Biol Crystallogr* *51*, 703-710.
- Leonidas, D.D., Chavali, G.B., Oikonomakos, N.G., Chrysina, E.D., Kosmopoulou, M.N., Vlassi, M., Frankling, C., and Acharya, K.R. (2003). High-resolution crystal structures of ribonuclease A complexed with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors. *Protein Sci* *12*, 2559-2574.
- Mattos, C. (2002). Protein-water interactions in a dynamic world. *Trends Biochem Sci* *27*, 203-208.
- Mattos, C., Bellamacina, C.R., Peisach, E., Pereira, A., Vitkup, D., Petsko, G.A., and Ringe, D. (2006). Multiple solvent crystal structures: probing binding sites, plasticity and hydration. *J Mol Biol* *357*, 1471-1482.
- Mattos, C., and Ringe, D. (1996). Locating and characterizing binding sites on proteins. *Nat Biotechnol* *14*, 595-599.
- Merlino, A., Vitagliano, L., Ceruso, M.A., Di Nola, A., and Mazzarella, L. (2002). Global and local motions in ribonuclease A: a molecular dynamics study. *Biopolymers* *65*, 274-283.
- Otwinowski, Z., and Minor, W. (1997). Processing of x-ray diffraction data collected in oscillation mode. *Methods in enzymology* *276*, 307-326.
- Raines, R.T. (1998). Ribonuclease A. *Chem Rev* *98*, 1045-1066.
- Rasmussen, B.F., Stock, A.M., Ringe, D., and Petsko, G.A. (1992). Crystalline ribonuclease A loses function below the dynamical transition at 220 K. *Nature* *357*, 423-424.

- Sadasivan, C., Nagendra, H.G., and Vijayan, M. (1998). Plasticity, hydration and accessibility in ribonuclease A. The structure of a new crystal form and its low-humidity variant. *Acta Crystallogr D Biol Crystallogr* *54*, 1343-1352.
- Sheu, S.H., Kaya, T., Waxman, D.J., and Vajda, S. (2005). Exploring the binding site structure of the PPAR gamma ligand-binding domain by computational solvent mapping. *Biochemistry* *44*, 1193-1209.
- Silberstein, M., Dennis, S., Brown, L., Kortvelyesi, T., Clodfelter, K., and Vajda, S. (2003). Identification of substrate binding sites in enzymes by computational solvent mapping. *J Mol Biol* *332*, 1095-1113.
- Silberstein, M., Landon, M.R., Wang, Y.E., Perl, A., and Vajda, S. (2006). Computational methods for functional site identification suggest a substrate access channel in transaldolase. *Genome Inform* *17*, 13-22.
- Vitagliano, L., Adinolfi, S., Riccio, A., Sica, F., Zagari, A., and Mazzarella, L. (1998). Binding of a substrate analog to a domain swapping protein: X-ray structure of the complex of bovine seminal ribonuclease with uridylyl(2',5')adenosine. *Protein Sci* *7*, 1691-1699.
- Vitagliano, L., Merlino, A., Zagari, A., and Mazzarella, L. (2000). Productive and nonproductive binding to ribonuclease A: X-ray structure of two complexes with uridylyl(2',5')guanosine. *Protein Sci* *9*, 1217-1225.
- Vitagliano, L., Merlino, A., Zagari, A., and Mazzarella, L. (2002). Reversible substrate-induced domain motions in ribonuclease A. *Proteins* *46*, 97-104.
- Zegers, I., Maes, D., Dao-Thi, M.H., Poortmans, F., Palmer, R., and Wyns, L. (1994). The structures of RNase A complexed with 3'-CMP and d(CpA): active site conformation and conserved water molecules. *Protein Sci* *3*, 2322-2339.

CHAPTER 4: The Interaction of RalA and RalBP

Abstract

Ral proteins are members of the Ras family of GTPases with unique switch I region and distinct effectors from the rest of the family. Interest in Ral has been growing as a mediator of Ras functioning, as an important player in cancer, and as a potential target for chemotherapy. Through its effectors, Ral plays a role in numerous cellular processes, such as exocytosis, filopodia formation, and transcriptional regulation. The first effector identified for Ral is the Ral Binding Protein (RalBP) and it links Ral to receptor-mediated endocytosis and regulation of mitosis. RalBP has been identified as being important in oncogenic signaling of Ral and also as a chemotherapy target in its own right. Compared to Ras, not much is known about the interaction of Ral and its effectors. About five years ago, the first structure of a Ral-effector complex demonstrated an intermolecular β -sheet, which is commonly observed in all Ras-Ras binding domain (RBD) complexes. Predictions that RalBP has high α -helical content suggest the Ral-RalBP interaction might be novel from those observed in Ras-RBD structures, making the structure of the Ral-RalBP complex of particular interest. This chapter discusses progress towards obtaining the crystal structure of the Ral-RalBP complex. While the specific structure has not been obtained, we have learned that (1) RalBP folds upon binding to RalA, (2) the Ral binding domain of RalBP(403-499) does not bind to RalA where that of RalBP(391-444) does, and (3) using RalA(11-178) and RalBP(391-444) produced preliminary crystals. These results should materially advance the goal of obtaining the crystal structure of the Ral-RalBP complex.

Introduction

RalA and RalB are members of a branch of the Ras family of GTPases and are >50% homologous to Ras (Chardin and Tavitian, 1986). As with other GTPases, Ral functions as a molecular switch, becoming active when bound to GTP through the aid of a Guanine-nucleotide Exchange Factor (GEF), such as Ral-GDS (Albright et al., 1993). Ral inactivation occurs when a GTPase Activating Protein (Ral-GAP) catalyzes the hydrolysis of GTP to GDP (Emkey et al., 1991). The primary differences between RalA and RalB, which are 85% identical, are in the C-terminal region (Feig, 2003). Both proteins are ubiquitously expressed and geranylgeranylated (Bhullar and Yang, 1998; Jilkina and Bhullar, 1996; Kinsella et al., 1991; Wildey et al., 1993). However, the subcellular location of these two proteins is different, partly as a result of the variation in the C-terminus (Shipitsin and Feig, 2004). Both RalA and RalB have been found associated with the membrane fractions of cells, but only RalB has been located in the cytosol, and Ral A is commonly associated with vesicles (Bhullar et al., 1990; Jilkina and Bhullar, 1996). Geranylgeranylation and proper localization have been shown to be critical for Ral function and the recruitment of effectors to the membrane (Bodemann and White, 2008; Matsubara et al., 1997; Shipitsin and Feig, 2004).

Effector proteins preferentially bind the active, GTP-bound form of GTPases. As the primary structural differences between the GDP- and GTP-bound forms lie in the flexible regions known as switch I and switch II, it follows that effector binding occurs through the switches (Vetter and Wittinghofer, 2001). Ral proteins have a unique switch I sequence, YEPTKAD, compared to the switch I sequence, YDPTIED, of most Ras family members

(Bos, 1998; Feig, 2003). It has been proposed that secondary binding sites could convey further specificity to Ras family GTPases (Nicely et al., 2004), and also has been shown that binding sites distal to the switch I effector region affect affinity of effector binding in Ral proteins (Shipitsin and Feig, 2004). Along with a unique switch I, Ral proteins have a distinct set of effectors through which Ral plays a role in a number of cellular processes. Ral interacts with the exocyst through subunits Sec5 and Exo84 and is critically involved with secretion, vesicle trafficking, and basolateral membrane trafficking in polarized cells (Brymora et al., 2001; Chen et al., 2007; Lopez et al., 2008; Moskalenko et al., 2002). Through its interaction with filamin, Ral induces filopodia formation (Ohta et al., 1999). Recently, Ral has been shown to interact with ZONAB, a transcriptional repressor, through which Ral influences transcription (Frankel et al., 2005). Finally, RalBP links Ral to receptor-mediated endocytosis and regulation of mitosis (Ikeda et al., 1998; Nakashima et al., 1999; Quaroni and Paul, 1999).

RalBP was the first effector found to bind Ral proteins and was identified concurrently in mouse, rat, and human and called RIP1 (Park and Weinberg, 1995), RalBP1 (Cantor et al., 1995), and RLIP76 (Jullien-Flores et al., 1995), respectively. This effector was predicted to be largely α -helical and found to contain a GAP domain for CDC42/Rac1. The N- and C-terminal portions of the protein, which flank the GAP and Ral-binding domains, have been found to have numerous functions. RalBP binds partners such as POB1, Repl1, and the AP2 complex and through these partners, RalBP is involved in endocytosis with activated Ral (Ikeda et al., 1998; Jullien-Flores et al., 2000; Matsuzaki et al., 2002; Nakashima et al., 1999;

Xu et al., 2001; Yamaguchi et al., 1997). RalBP was also found to be 98.4% identical to cytoctrin, which associates with the centrosome and regulates the assembly of the mitotic spindle (Quaroni and Paul, 1999). Additionally, it was proposed that RalBP has a role in turning off endocytosis during mitosis (Rosse et al., 2003). RalBP was found to also function as a non-ABC, ATP-dependent transporter of glutathione conjugates and doxorubicin and has been suggested to contribute to multidrug resistance in cancer (Awasthi et al., 2000; Awasthi et al., 2003a; Sharma et al., 2002).

In addition to the elucidation of functional roles in healthy cells for Ral and RalBP proteins, evidence for important contributions to tumorigenesis is growing and both Ral and RalBP have been identified as potential chemotherapy targets (Awasthi et al., 2007; Chien and White, 2003; Lim et al., 2005; Nadkar et al., 2006; Panner et al., 2006; Singhal et al., 2007; Smith et al., 2007). Over a decade ago, Ral was found to belong to a distinct downstream pathway from Ras through Ral-GDS and to promote proliferation and complement transformation of other Ras effectors (Reuther and Der, 2000; Urano et al., 1996; Wolthuis et al., 1997). Since then, the knowledge of the contributions of Ral in tumorigenesis has grown. It has been found that Ral activity in the tumor directly correlates to the metastatic potential of tumor cells (Tchevkina et al., 2005). Also, Ral has been shown to have a significant role in controlling cell proliferations in estrogen-independent breast cancer cells (Yu and Feig, 2002). The two isoforms of Ral have different, though collaborative, roles to maintain tumorigenicity. RalB is required for cell survival and prevents transformed cells from initiating apoptosis where RalA is required for anchorage-independent proliferation (Chien

and White, 2003). Additionally, RalA has been linked to the translational machinery involved in tumor suppression (Panner et al., 2006). The subcellular localization of these proteins contributes to the ability of RalA and RalB to affect transformation, with RalA activation being a critical step of transformation and tumorigenesis in human cells (Lim et al., 2005; Lim et al., 2006). Oncogenic signaling of Ral requires the binding of RalBP and exocyst effectors (Lim et al., 2005). RalBP is upregulated in cancer and, through its function as a transporter, has a key role in cancer cell survival (Awasthi et al., 2000; Awasthi et al., 2003a; Sharma et al., 2002). Additionally, inhibition or depletion of RalBP in cancer cells prevents its transport function and leads to apoptosis and remission of cancer, while sparing nonmalignant cells (Awasthi et al., 2003b; Awasthi et al., 2003c; Nadkar et al., 2006). Recently, additional regulatory steps have been identified in Ral function and cell transformation. The kinase Aurora A has been shown to phosphorylate Ral and potentially contribute to anchorage-independent growth (Wu et al., 2005). Where the phosphatase PP2A A β dephosphorylates Ral and abolishes its transforming function (Sablina et al., 2007). Also, Aurora A, Ral, and RalBP have been shown to be overexpressed in bladder cancer (Smith et al., 2007), adding further evidence that the interaction of Ral and RalBP plays an important role in cancer.

The Ras binding domain (RBD) of Ras effectors, including Ral-GDS (Huang et al., 1997), adopt a ubiquitin-like fold. The interaction between switch I of Ras and the RBD forms an intermolecular β -sheet, and the differences in each complex of Ras and various effectors can be observed in the interaction of the side chains (Huang et al., 1998). Less is known about

the structures of Ral-effector complexes. The Ral binding domain of Sec5 consists of an IPT, or β -sandwich, domain, and that of Exo84 adopts a pleckstrin homology domain fold, both of which are novel for Ras effectors (Jin et al., 2005; Mott et al., 2003). While the structure of Sec5 is distinct from that of RBDs, the Ral-Sec5 interaction forms an intermolecular β -sheet through switch I, similar to the complexes of Ras and RBDs (Fukai et al., 2003). In contrast, Exo84 interacts with RalA through both switch regions, presenting a new binding mode for Ral-effector complexes (Jin et al., 2005). With RalBP predicted to be largely α -helical (Cantor et al., 1995; Jullien-Flores et al., 1995; Park and Weinberg, 1995), the question is, will the interaction between Ral and RalBP be similar to what is observed with Ras-effector complexes? The high helical content suggests that the answer is no. However, mutation of switch I region residues Lys 47 and Ala 48, which are unique to Ral proteins, to their counterpart amino acids in Ras (Ile and Glu, respectively) prevents the formation of the Ral-RalBP complex. The reverse mutations in the Ras switch I region allow for the formation of the Ras-RalBP complex (Bauer et al., 1999). This suggests that despite the helical content, the interaction of Ral and RalBP may be similar to that of Ras and RBDs. A further question involves the conformation of the switch regions of Ral. Two different conformations were observed in the switch I region when active Ral is bound to Sec5 compared to Ral-GppNHp alone (Nicely et al., 2004). In addition, switch II in the RalA-Exo84 complex adopts a unique conformation from those previously observed (Jin et al., 2005). Does RalBP bind a switch I or switch II conformation of Ral that has been previously observed or another discrete conformation adopted by active Ral?

Significant progress was made to obtain the crystal structure of the RalA-RalBP complex. First, RalBP is intrinsically disordered and folds upon binding to RalA. This is the first example of a Ras family effector with this behavior. Second, two cysteine-to-serine mutations alleviate the oligomerization problem observed in the Ral binding domain of RalBP. These mutations did not disrupt the formation of the complex between RalA and RalBP. Third, the double mutant of RalBP was truncated to the published Ral binding domain of residues 403-499 (Jullien-Flores et al., 1995) to improve chances for crystallization. However this construct of RalBP did not bind to RalA. Other constructs were designed and the construct corresponding to an alternative published Ral binding domain of residues 391-444 (Park and Weinberg, 1995) did bind RalA. A shorter construct of RalA, residues 11-178, then was used when none of the prior crystallization attempts were successful. Using RalBP(391-444) and RalA(11-178), preliminary protein crystals were obtained. These results provide an excellent starting point for obtaining crystals of the RalA(11-178)-RalBP(391-444) complex.

Materials and Methods

Expression of GST-RalBP (all constructs)

E. coli BL21 Rosetta cells (Novagen) containing the pGEX-2T GST fusion vector (GE Healthcare) with RalBP(all constructs) were grown in LB broth at 37°C with shaking until an OD₆₀₀ of approximately 0.6-0.8 was reached. Protein expression was induced by adding 0.5 mM isopropyl- β -D-thiogalactopyranoside (IPTG). The temperature was maintained at 37°C

and the cells were allowed to grow for 4 hours before harvesting by centrifugation. Cell pellets were stored at -80°C .

Batch Affinity Purification of GST-RalBP(397-518x) and Thrombin Cleavage

Frozen cell pellets were thawed on ice and resuspended in 50 mL Lysis Buffer (50 mM HEPES pH 7.5, 5 mM MgCl_2 , 200 mM NaCl, 1 mM DTT or 5 mM TCEP HCl, and 5% glycerol) with protease inhibitors (5 mM benzamidine, 1mM pefabloc, 2 $\mu\text{g}/\text{mL}$ antipain, 1 $\mu\text{g}/\text{mL}$ leupeptin, 1 $\mu\text{g}/\text{mL}$ pepstatin A). After sonication and centrifugation, the supernatant of the cell lysate was mixed with a 20 mL 1:1 suspension of glutathione-agarose beads equilibrated with Lysis Buffer and incubated with rocking at 4°C for 2 hours. Unbound protein was removed with centrifugation and subsequent washes with Lysis B buffer, until the absorbance at 280nm of the supernatant was < 0.1 .

Glutathione-agarose beads with bound GST-RalBP(397-518x) were transferred into Thrombin Buffer (50 mM HEPES pH 8.0, 5 mM MgCl_2 , 150 mM NaCl, 2.5 mM CaCl_2 , 1mM DTT or 5mM TCEP HCl, and 5% glycerol). Thrombin cleavage of GST-RalBP(397-518x) was performed with the protein bound to the glutathione-agarose beads by adding 10 units of thrombin for every mL 1:1 suspension of glutathione-agarose beads in Thrombin Buffer. The cleavage reaction was allowed to proceed for 2 hours at room temperature with rocking. Cleavage product was collected by centrifugation and three subsequent washes with Thrombin Buffer. Thrombin was removed from the collected supernatant by adding 100 μL p-aminobenzamidine-agarose for every 50 units of thrombin and incubating for 30 minutes at

room temperature with rocking. The p-aminobenzamidine-agarose was removed by filtration. Pefabloc was added to 1 mM to inactivate any remaining thrombin. (The original purification protocol continues from this point with FPLC purification described in “Original Purification of RalBP(397-518x)”.) 2mL of fresh 1:1 suspension of glutathione-agarose beads equilibrated with Thrombin Buffer was added to the RalBP(397-518x) solution and allowed to incubate overnight at 4°C with rocking. The glutathione-agarose beads were removed by filtration to withdraw any cleaved GST or fusion protein that had unbound from the beads during the washes.

Original Purification of RalBP(397-518x)

After the thrombin cleavage of GST-RalBP(397-518x), RalBP(397-518x) was dialyzed overnight into RalBP QFF Buffer A (20 mM HEPES pH 8.0, 5 mM MgCl₂, 20 mM NaCl, 1 mM DTT or 5 mM TCEP HCl, 5% glycerol). RalBP(391-518x) was applied to a HiPrep 16/10 Q Sepharose FF column (GE Healthcare) at 1 mL/min and eluted over a 300 mL gradient of 0-60% RalBP QFF Buffer B in RalBP QFF Buffer A. RalBP QFF Buffer B differs from RalBP QFF Buffer A only in that it contains 800mM NaCl. Fractions containing the protein were pooled, concentrated to less than 1 mL, and applied at 1 mL/min to a HiPrep 26/60 Sephacryl S-100 HR column (GE Healthcare) equilibrated in RalBP Gel Filtration Buffer (20 mM HEPES pH 8.0, 5 mM MgCl₂, 150 mM NaCl, and 1 mM DTT or 5 mM TCEP HCl). As described in the results, the NaCl content of the RalBP Gel Filtration Buffer was eventually changed to 300 mM. Protein was eluted with 200 mL RalBP Gel Filtration Buffer. Fractions containing the protein were pooled, and were applied to a HiTrap Q

Sepharose HP column (GE Healthcare) at 1 mL/min and eluted over a 250 mL gradient of 0-50% RalBP QHP Buffer B in RalBP QHP Buffer A. RalBP QHP Buffers A and B are identical to RalBP QFF buffers A and B, except the QHP buffers do not include glycerol.

Expression and Purification of RalA

As described previously (Nicely et al., 2004), *E. coli* BL21 Rosetta cells (Novagen) containing the pET21a(+) vector (Novagen) with the C-terminal truncated of simian RalA (residues 1-178) were grown in LB broth at 37°C with shaking until an OD₆₀₀ of approximately 0.6-0.8 was reached. Protein expression was induced by adding 0.15 mM isopropyl- β -D-thiogalactopyranoside (IPTG). The temperature was reduced to 32°C and cells were allowed to grow for 5 hours before harvesting by centrifugation. Cell pellets were stored at -80°C. Frozen cell pellets were thawed on ice and resuspended in 50 mL QFF Buffer A (20mM HEPES pH 7.6, 5 mM MgCl₂, 50 mM NaCl, 1 mM DTT, 10 μ M GDP, and 5% glycerol) with protease inhibitors (5 mM benzamidine, 1mM pefabloc, 2 μ g/mL antipain, 1 μ g/mL leupeptin, 1 μ g/mL pepstatin A). After sonication and centrifugation, the supernatant of the cell lysate was applied to a HiPrep 16/10 Q Sepharose FF column (GE Healthcare) at 5 mL/min and eluted over a 200 mL gradient of 0-40% QFF Buffer B in QFF Buffer A. QFF Buffer B differs from QFF Buffer A only in that it contains 1M NaCl. Fractions containing the protein were pooled, concentrated to less than 1 mL, and applied at 1.3 mL/min to a HiPrep 26/60 Sephacryl S-100 HR column (GE Healthcare) equilibrated in Gel Filtration Buffer (20 mM HEPES pH 7.6, 5 mM MgCl₂, 150 mM NaCl, 1 mM DTT, and 10 μ M GDP). Protein was eluted with 200 mL Gel Filtration Buffer. Fractions containing

the protein were pooled, and were applied to a HiTrap Q Sepharose HP column (GE Healthcare) at 1 mL/min and eluted over a 110 mL gradient of 0-11% QHP Buffer B in QHP Buffer A. QHP Buffers A and B are identical to QFF buffers A and B, except the QHP buffers do not include glycerol.

Similar to the method published for Ras (John et al., 1990), RalA-GppNHp was obtained by nucleotide exchange. RalA was concentrated to less than 1 mL and applied to a PD-10 Sephadex G-25 column (GE Healthcare) equilibrated in Reaction Buffer (32 mM Tris-HCl pH 8.0, 200 mM $(\text{NH}_4)_2\text{SO}_4$, 10 mM DTT, 0.1% n-octyl glucopyranoside) to remove the excess Mg^{2+} and GDP. The protein was eluted using 5 mL Reaction Buffer. The protein containing fractions were pooled and a 4x molar excess of GMPPNP and alkaline phosphatase-agarose beads (approximately 100 units alkaline phosphatase per 20 mg protein) were added. The mixture was incubated at 37°C with end-over-end rotation for 45 minutes. After incubation, MgCl_2 was added to 20 mM and the mixture was centrifuged at 14000 rpm for 3 minutes to remove the alkaline phosphatase-agarose beads. The supernatant was applied to a PD-10 Sephadex G-25 column (GE Healthcare) equilibrated in Happy Lite Buffer (10 mM HEPES pH 7.5, 10 mM NaCl, 5mM MgCl_2 , 1 mM DTE, 1 μM GMPPNP). Protein containing fractions were pooled and if not used immediately, flash frozen and stored at -80°C . Unless specified otherwise, “RalA” is used to denote RalA-GMPPNP in this chapter.

RalA-RalBP(all constructs) Complex Formation and Purification

With concentrated volumes of <5 mL for both RalA and RalBP(all constructs), the two proteins were mixed in a 1:1 ratio and incubated on ice or at 4°C for 1 hour. To remove uncomplexed protein, the protein complex was concentrated to <1 mL and applied at 1.3 mL/min to a HiPrep 26/60 Sephacryl S-100 HR column (GE Healthcare) equilibrated in Complex Gel Filtration Buffer (20 mM HEPES pH 8.0, 5 mM MgCl₂, 150 mM NaCl, and 1 mM DTT or 5 mM TCEP HCl). Protein was eluted with 200 mL Complex Gel Filtration Buffer. Fractions containing the protein were pooled and concentrated and exchanged into Complex Happy Buffer (10 mM HEPES pH 7.5, 5 mM MgCl₂, 1 μM GMPPNP, and 1 mM DTE or 5 mM TCEP).

Optimized Purification of GST-RalBP(all constructs) and Solution Thrombin Cleavage

Frozen cell pellets were thawed on ice and resuspended with 50 mL GST Binding Buffer (PBS pH 7.3: 140 mM NaCl, 2.7 mM KCl, 10 mM NaH₂PO₄*H₂O, 1.8 mM KH₂PO₄; 10 mM DTT) with protease inhibitors (5 mM benzamidine, 1mM pefabloc, 2 μg/mL antipain, 1 μg/mL leupeptin, 1 μg/mL pepstatin A). After sonication and centrifugation, the supernatant of the cell lysate was applied to two HiTrap Glutathione Sepharose 4 FF columns (GE Healthcare) in series (2-5 mL columns for 10 mL total volume) at 1 mL/min in GST Binding Buffer. To ensure maximum GST-RalBP retention, a quarter of the total volume of the cell lysate was applied at a time and then eluted. Each elution was performed at 4 mL/min with 20-30 mL of Elution Buffer (50 mM Tris-HCl pH 8.0, 10 mM reduced glutathione, 10 mM DTT). Protein containing fractions were pooled and CaCl₂ was added to 2.5 mM. For every

milliliter of solution, 90 units of thrombin were added and cleavage reaction was allowed to proceed for 45 minutes at room temperature. Thrombin was removed by adding 100 μ L p-aminobenzamidine-agarose for every 50 units of thrombin and incubating for 30 minutes at room temperature with rocking. The p-aminobenzamidine-agarose was removed by filtration. Pefabloc was added to 1 mM to inactivate any remaining thrombin. Protein was applied to a HiTrap Q Sepharose HP column (GE Healthcare) at 1 mL/min and eluted over a 2-step gradient: 15 mL 0-15% RalBP QHP Buffer B (20 mM HEPES pH 8.0, 1 M NaCl, 1 mM DTT, 5 mM TCEP HCl) in QHP Buffer A (20 mM HEPES pH 8.0, 20 mM NaCl, 1 mM DTT, 5 mM TCEP HCl), 150 mL 15-45% RalBP QHP Buffer B in QHP Buffer A. RalBP(397-518x) containing fractions were pooled and 4mL of fresh 1:1 suspension of glutathione-agarose beads equilibrated with RalBP QHP Buffer A was added to the solution and allowed to incubate overnight at 4°C with rocking. The glutathione-agarose beads were removed by filtration to withdraw any cleaved GST or fusion protein that had not been separated by the HiTrap Q Sepharose HP column. In the constructs where serine residues replace the cyteines, only 1 mM DTT is used and no TCEP HCl is included. If the protein was being used for crystallization trials instead of complex formation, it was transferred into RalBP Happy Buffer (10 mM HEPES pH 7.6, 10 mM NaCl, and 1 mM DTE, or 5 mM TCEP).

DNA Sequencing

E. coli BL21 Rosetta (Novagen) or DH5 α (Invitrogen) containing the pGEX-2T GST fusion vector (GE Healthcare) with RalBP(all constructs) were grown in 50 mL LB at 37°C with

shaking overnight and cells were harvested with centrifugation. Plasmid DNA was purified using the QIAprep Spin Miniprep Kit (QIAGEN) and a microcentrifuge. 2-3 µg DNA in a volume of 20 µL was sent to MWG-Biotech for DNA sequencing using special-ordered forward (5'-GGA CCC AAT GTG CCT GGA TGC G-3') and reverse (5'-AAG TGC CAC CTG ACG TCT-3') sequencing primers.

Mutagenesis

Met472 was restored from Thr using the QuikChange II Site-Directed Mutagenesis Kit (Stratagene) and special-ordered forward (5'-GGA GGA TGT TTC CAA AGA AGA AAT GAA CGA AAA CGA GGA GGT C-3') and reverse (5'-GAC CTC CTC GTT TTC GTT CAT TTC TTC TTT GGA AAC ATC CTC C-3') site-directed mutagenesis primers. Four reactions were prepared with varying amounts of double-stranded plasmid DNA (5 ng, 10 ng, 20 ng, and 50 ng).

Cys451 was mutated to Ser using the QuikChange II Site-Directed Mutagenesis Kit (Stratagene) and special-ordered forward (5'-GGG AAG CTA AAA GAC AAG AGT CTG AGA CCA AGA TTG CAC AGG-3') and reverse (5'-CCT GTG CAA TCT TGG TCT CAG ACT CTT GTC TTT TAG CTT CCC-3') site-directed mutagenesis primers. Four reactions were prepared with varying amounts of double-stranded plasmid DNA (5 ng, 10 ng, 20 ng, and 50 ng). Cys 411 was mutated to Ser in the same manner using special-ordered forward (5'-GGA GAC AGG AGT TTC TTT TGA ACT CTT TAC ATC GAG ATC TGC AGG

GCG-3') and reverse (5'-CGC CCT GCA GAT CTC GAT GTA AAG AGT TCA AAA GAA ACT CCT GTC TCC-3') site-directed mutagenesis primers.

After the mutagenesis had been verified by sequencing, the T472M corrected DNA, C451S single mutant and C411S,C451S double mutant DNA were each transformed into both OneShot Top10 Competent Cells (Invitrogen) and BL21(DE3) cells (Novagen). The C411S,C451S double mutant with the correction at 472 is termed RalBP(397-518x;C411S, C451S).

Prediction of Naturally Disordered Regions

The protein sequence of RalBP(397-518x) was used as input for the Predictor of Naturally Disordered Regions (PONDR; <http://www.pondr.com>; Romero et al., 2002) using the VL-XT predictor (Li et al., 1999; Romero et al., 1997; Romero et al., 2001). Predictors were trained on sequences of previously determined ordered and disordered regions. According to the PONDR tutorial, the VL-XT predictor is trained with long regions of 40 or more residues characterized by NMR and crystallography, and crystallographically characterized C-terminal and N-terminal short regions of five or more residues.

Circular Dichroism Spectroscopy

Both RalA and RalBP(397-518x) were purified normally, except that phosphate buffer (pH 8) was used in place of HEPES in the RalBP QHP A and B buffers, and the Complex Gel Filtration Buffer. 20 μ M samples of RalA, RalBP(397-518x), and the complex were

Table 1. Parameters used with the Jasco J-600 for Circular Dichroism.

	Near UV	Far UV
Data Mode	CD	CD
Band Width	1.0 nm	1.0 nm
Slit Width	Auto	Auto
Sensitivity	20 mdeg	100 mdeg
Time Constant	4.0 sec	4.0 sec
Start Wavelength	320 nm	250 nm
End Wavelength	250 nm	207 nm
Step Resolution	1.0 nm	1.0 nm
Scan Speed	20 nm/min	10 nm/min
# of scans	10	10
Alternate	off	off

prepared in CD Sample Buffer (10 mM phosphate buffer pH 8, 5 mM Na₂SO₄, 5 mM MgSO₄, 10 mM DTT), with each sample having a volume of 1 mL. An additional milliliter of CD Sample Buffer was used for the blank.

Near-UV and Far-UV spectra were collected on a Jasco J-600 for the three samples and the blank. The parameters used in data acquisition are listed in Table 1. To prevent damage to the instrument, Far-UV data could not be collected at wavelengths shorter than 207 nm for our samples.

Data was plotted after first subtracting the blank readings at each wavelength for each sample. Theoretical signals were calculated by subtracting the values from one signal from those of another, for example: subtracting values for the RalBP(397-518x) signal from the

values of the RalA-RalBP(397-518x) complex signal to produce the “Complex-RalBP” signal in Figure 10b.

Engineering Truncated Forms of the Double Serine Mutant of RalBP

All truncated forms of RalBP used DNA from the GST-RalBP(397-518x;C411S,C451S) construct as the template. The GST-RalBP(403-499) construct was engineered using the Blade and Edge primers (Table 2) and PCR amplification using the NovaTaq PCR Kit (Novagen). These primers amplified the RalBP sequence from residues 403-499 with a triple stop codon at the 3' end and 5' BamHI and 3' EcoRI restriction sites. Amplified insert and fresh pGEX-2T vector (GE Healthcare) were digested with BamHI (New England Biolabs) and EcoRI (New England Biolabs). Ligation was performed using the Novagen DNA Ligation Kit. Ligated vector was transformed into XL1-Blue (Stratagene) and BL21(DE3) cells (Novagen). DNA engineering was confirmed by sequencing miniprep DNA.

All additional constructs were made with the same methodology as GST-RalBP(403-499), but with different primers. The GST-RalBP(397-518) construct was engineered using the Jude and Eustace primers (Table 2). These primers amplified the RalBP sequence from residues 397-518 with a triple stop codon at the 3' end and 5' BamHI and 3' EcoRI restriction sites. The GST-RalBP(391-444) construct was engineered using the Brighton and Edinburgh primers (Table 2). These primers amplified the RalBP sequence from residues 391-444 with a triple stop codon at the 3' end and 5' BamHI and 3' EcoRI restriction sites. Additionally, the Brighton primer inserts codons for residues 391-396 in between the 5'

Table 2. Primers for RalBP truncation. Note: The Brighton primer engineers in six additional amino acid residues at the N-terminus: Residues 391-396, LPETQA.

Name	Description	Sequence
Blade	BamHI-Residue 403	5'-CGC GGA TCC AGG AGA CAG GAG TTT CTT TTG AAC TCT TTA CAT CG-3'
Brighton	BamHI-Residue 391	5'-CGC GGA TCC CTG CCA GAG ACC CAA GCA GGC ATC AAG GAA GAA ATC AGG-3'
Jude	BamHI-Residue 397	5'-CGC GGA TCC GGC ATC AAG GAA GAA ATC AGG AGA CAG G-3'
Edinburgh	Residue 444-Stop-Stop-Stop-EcoRI	5'-CCG GAA TTC CGG TCA TCA TCA CCT CAG CTT TCT CTT GAG GGC-3'
Edge	Residue 499-Stop-Stop-Stop-EcoRI	5'-CCG GAA TTC CGG TCA TCA TCA CAT GGC CAG GAG C-3'
Eustace	Residue 518-Stop-Stop-Stop-EcoRI	5'-CCG GAA TTC CGG TCA TCA TCA TCG GAG GCG GTC AAT C-3'

BamHI site and the codon for Gly 397. All constructs of GST-RalBP that were expressed and purified are listed in Table 3.

Concentrated Thrombin Cleavage and Purification of RalBP(391-444)

GST-RalBP(391-444) was resuspended and purified with affinity chromatography identically to the other constructs of the fusion protein. Protein containing fractions were pooled and concentrated to less than 1 mL. Ten units of thrombin per milligram of protein were added and the mixture was incubated at 37°C for 2 minutes with end-over-end rotation.

Benzamidine was added to 150mM and protein was diluted with RalBP QHP Buffer A (20 mM HEPES pH 8.0, 20 mM NaCl, and 1 mM DTT) to 40 mL. Protein was applied to a HiTrap Q Sepharose HP column (GE Healthcare) immediately preceding a 26/10 HiLoad SP Sepharose HP column (GE Healthcare). Once all the protein was loaded and unbound HiTrap Q Sepharose HP column (GE Healthcare) immediately preceding a 26/10 HiLoad SP

Table 3. Nomenclature of expressed and purified GST-RalBP constructs. Two values are given for the pI of each protein. The first is for the GST-RalBP fusion protein, and the second is for the RalBP protein after thrombin cleavage.

Name	pI	Description
GST-RalBP(397-518x)	5.32/4.87	Original GST-RalBP fusion provided by Larry Feig
GST-RalBP(397-518x;C411S,C451S)	5.32/4.87	GST-RalBP fusion provided by Larry Feig with residue 472 reverted to methionine as observed in the wild-type sequence and with a double serine mutation at residue 411 and 451
GST-RalBP(403-499)	5.48/4.93	GST-RalBP fusion including the double serine mutation at residues 411 and 451, but truncated to include only residues 403-499 (the published (Cantor et al 1995) Ral minimum binding domain)
GST-RalBP(397-518)	5.45/4.99	GST-RalBP fusion with residue 472 reverted to methionine as observed in the wild-type sequence and with a double serine mutation at residue 411 and 451; a triple stop codon is included after the codon for residue 518 thus eliminating the non-RalBP residues at the C-terminus
GST-RalBP(391-444)	7.05/9.81	GST-RalBP fusion including the serine mutation at residue 411, truncated to include only residues 391-444 (the published (Park and Weinberg 1995) Ral minimum binding domain); Residues 391-396 were not present in any previous construct and needed to be engineered into the sequence.

Sepharose HP column (GE Healthcare). Once all the protein was loaded and unbound protein had washed off the columns, the HiTrap Q Sepharose HP column was removed and RalBP(391-444) was eluted at 4 mL/min from the 26/10 HiLoad SP Sepharose HP column over a 200 mL gradient of 0-50% RalBP QHP Buffer B in RalBP QHP Buffer A. RalBP QHP Buffer B differs from Buffer A only in that it contains 1M NaCl. Protein containing fractions were pooled and concentrated. If the protein was being used for crystallization trials instead of complex formation, it was transferred into RalBP Happy Buffer (10 mM HEPES pH 7.6, 10 mM NaCl, and 1 mM DTE).

Expression and Purification of RalA(11-178)

E. coli BL21 (DE3) cells (Novagen) containing the pET21a(+) vector (Novagen) with the N-terminal and C-terminal truncated of simian RalA (residues 11-178) were grown in LB broth at 37°C with shaking until an OD₆₀₀ of approximately 0.6-0.8 was reached. Protein expression was induced by adding 0.15 mM isopropyl- β -D-thiogalactopyranoside (IPTG). The temperature was reduced to 32°C and cells were allowed to grow for 5 hours before harvesting by centrifugation. Cell pellets were stored at -80°C. Frozen cell pellets were thawed on ice and resuspended in 50 mL QFF Buffer A (20mM HEPES pH 7.6, 5 mM MgCl₂, 50 mM NaCl, 1 mM DTT, 10 μ M GDP, and 5% glycerol) with protease inhibitors (5 mM benzamidine, 1mM pefabloc, 2 μ g/mL antipain, 1 μ g/mL leupeptin, 1 μ g/mL pepstatin A). After sonication and centrifugation, the supernatant of the cell lysate was applied to a HiPrep 16/10 Q Sepharose FF column (GE Healthcare) at 5 mL/min and eluted over a 200 mL gradient of 0-40% QFF Buffer B in QFF Buffer A. QFF Buffer B differs from QFF Buffer A only in that it contains 1M NaCl. Fractions containing the protein were pooled, concentrated to less than 1 mL, and applied at 1.3 mL/min to a HiPrep 26/60 Sephacryl S-100 HR column (GE Healthcare) equilibrated in Gel Filtration Buffer (20 mM HEPES pH 7.6, 5 mM MgCl₂, 150 mM NaCl, 1 mM DTT, and 10 μ M GDP). Protein was eluted with 200 mL Gel Filtration Buffer. Fractions containing the protein were pooled, and were applied to a HiTrap Q Sepharose HP column (GE Healthcare) at 1 mL/min and eluted over a 150 mL gradient of 0-15% QHP Buffer B in QHP Buffer A. QHP Buffers A and B are identical to QFF buffers A and B, except the QHP buffers do not include glycerol. RalA(11-178)-GppNHp was obtained by nucleotide exchange using the same methodology as RalA(1-

Table 4. Crystallization Trials.

Protein	Protein Conc.	Temp	Well Additives	Screens	TCEP in Buffer?
RalBP(397-518x)	10 mg/mL	18°C		Hampton Peg Ion	no
	14 mg/mL	18°C		Hampton Crystal Screen	no
	15 mg/mL	18°C		Hampton Peg Ion, Hampton Crystal Screen, Hampton Crystal Screen (Sitting Drop), Hampton Crystal Screen 2, Nextal Classics, Nextal Peg, Nextal AmSO ₄ , Emerald Wizard I	yes
	18 mg/mL	18°C	10 mM TCEP	Hampton Peg Ion, Hampton Crystal Screen Lite	yes
GST-RalBP(397-518x)	18 mg/mL	18°C		Nextal Classics, Nextal Peg, Nextal AmSO ₄ , Peg Ion, Crystal Screen, Crystal Screen 2, Wizard I	yes
RalA-RalBP(397-518x) Complex	4 mg/mL	18°C		Hampton Crystal Screen 2, Hampton Crystal Screen Lite	no
	10-15 mg/mL	18°C		Nextal AmSO ₄	yes
	13 mg/mL	18°C		Hampton Peg Ion, Hampton Crystal Screen, Hampton Crystal Screen 2	no
	15 mg/mL	18°C		Hampton Peg Ion Screen	no
	20 mg/mL	18°C	10 mM TCEP	Hampton Peg Ion, Hampton Crystal Screen, Hampton Crystal Screen 2, Hampton Crystal Screen Lite	yes
	20 mg/mL	room	10 mM TCEP	Hampton Peg Ion, Hampton Crystal Screen, Hampton Crystal Screen 2, Hampton Crystal Screen Lite	yes
	20 mg/mL	4°C		Hampton Crystal Screen Lite	yes
	25 mg/mL	18°C		Hampton Crystal Screen 2, Hampton Crystal Screen Lite	no
	40 mg/mL	18°C		Hampton Crystal Screen Lite, Emerald Wizard I	yes
	40 mg/mL	4°C		Hampton Peg Ion, Emerald Wizard I	yes

Table 4 continued.

Protein	Protein Conc.	Temp	Well Additives	Screens	TCEP in Buffer?
RalA-RalBP(397-518x;C411S,C451S) Complex	12 mg/mL	18°C		Hampton Peg Ion, Hampton Crystal Screen, Hampton Crystal Screen 2	no
	30 mg/mL	18°C		Hampton Peg Ion, Hampton Crystal Screen 2, Hampton Crystal Screen Lite, Emerald Wizard I	no
RalA-RalBP(403-499) Complex	10 mg/mL	18°C		Hampton Peg Ion	no
RalA-RalBP(397-518) Complex	20 mg/mL	18°C		Emerald Wizard I, Emerald Wizard II	no
	17 mg/mL	18°C		Hampton Peg Ion, Emerald Wizard I, Emerald Wizard II, Emerald Wizard III	no
RalA-RalBP(391-444)	10 mg/mL	18°C		Hampton Peg Ion, Hampton Crystal Screen, Hampton Crystal Screen 2, Nextal ProComplex	no
RalA(11-178)-RalBP(391-444)	18 mg/mL	18°C		Hampton Peg Ion, Hampton Crystal Screen, Hampton Crystal Screen 2	no

178). Protein containing fractions were pooled and if not used immediately, flash frozen and stored at -80°C . Complex formation and purification using RalA(11-178) was performed in the same manner as RalA(1-178). Unless specified otherwise, “RalA(11-178)” is used to denote RalA(11-178)-GMPPNP in this chapter. Residue 11 in this construct is mutated from a serine to a methionine.

Crystallization Trials

Sparse matrix crystal screens were set up using hanging drop vapor diffusion and commercially manufactured screens. Each experiment used a well volume of 500 μL of precipitant solution and a drop size of 2 μL protein solution plus 2 μL precipitant solution.

Some of the crystallization experiments included an additional 10 mM TCEP in the precipitant solution to prevent the oligomerization of RalBP(397-518x). The screens used were: Hampton Peg Ion, Hampton Crystal Screen, Hampton Crystal Screen Lite, Hampton Crystal Screen 2, Nextal Peg-Ion, Nextal Classics, Nextal AmSO₄, Nextal Pro-Complex, Emerald Wizard I, Emerald Wizard II, and Emerald Wizard III. A summary of the crystallization trials is given in Table 4.

Results

Initial Purification of RalBP(397-518x)

Initial purification of the GST-RalBP(397-518x) protein posed some unexpected challenges. After batch affinity purification of GST-RalBP(397-518x), thrombin cleavage, and subsequent anion exchange and size-exclusion chromatography, RalBP(397-518x) was found to be mostly pure, but appeared to co-elute with a number of high molecular weight contaminants. This was surprising, because these contaminants should have been removed with size-exclusion chromatography. Presuming this was a result of protein-protein interactions, subsequent size-exclusion chromatography was run with added detergent (0.1% n-octyl- β -D-glucopyranoside) and increased salt concentration (320 mM NaCl) to attempt to disrupt these interactions. When this did not improve purification, size-exclusion chromatography was run using a sample of RalBP(397-518x) and protein size standards (myoglobin, cytochrome c, alcohol dehydrogenase, bovine serum albumin, and carbonic anhydrate) and increased salt concentration (300 mM NaCl). Compared to the protein standards, RalBP(397-518x) eluted at the appropriate time for its molecular weight, leading

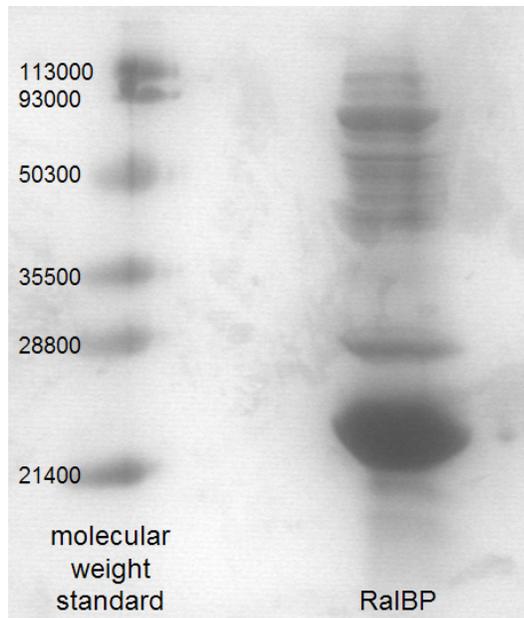


Figure 1. SDS-PAGE gel of concentrated RalBP(397-518x). The first lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards with their respective molecular weights in Daltons listed at the left. The second lane contains concentrated RalBP(397-518x) after being purified with batch affinity purification, and anion exchange and size exclusion chromatography.

to the continued use of 300 mM NaCl in the buffer for size exclusion chromatography. The use of a second anion exchange step (HiTrap Q Sepharose HP column, GE Healthcare) was introduced to remove contaminants still present after size exclusion chromatography. Upon concentration, it was clear that while RalBP(397-518x) had been subjected to affinity batchpurification, anion exchange, and size-exclusion chromatography, there were still impurities in the protein sample (Figure 1).

RalA-RalBP(397-518x) Complex Formation

As the RalA-RalBP complex was the ultimate goal, work began to obtain the complex with the hope that complex formation would aid in purification. As the RalA expression and purification procedures had been developed previously (Nicely et al., 2004), it did not take long to have enough pure RalA to test for complex formation. The complex was allowed to form as described in the Materials and Methods and then purified using size-exclusion chromatography to remove any uncomplexed protein or other contaminants. RalA and RalBP(397-518x) co-eluted in a 1:1 ratio, indicating complex formation (Figure 2).

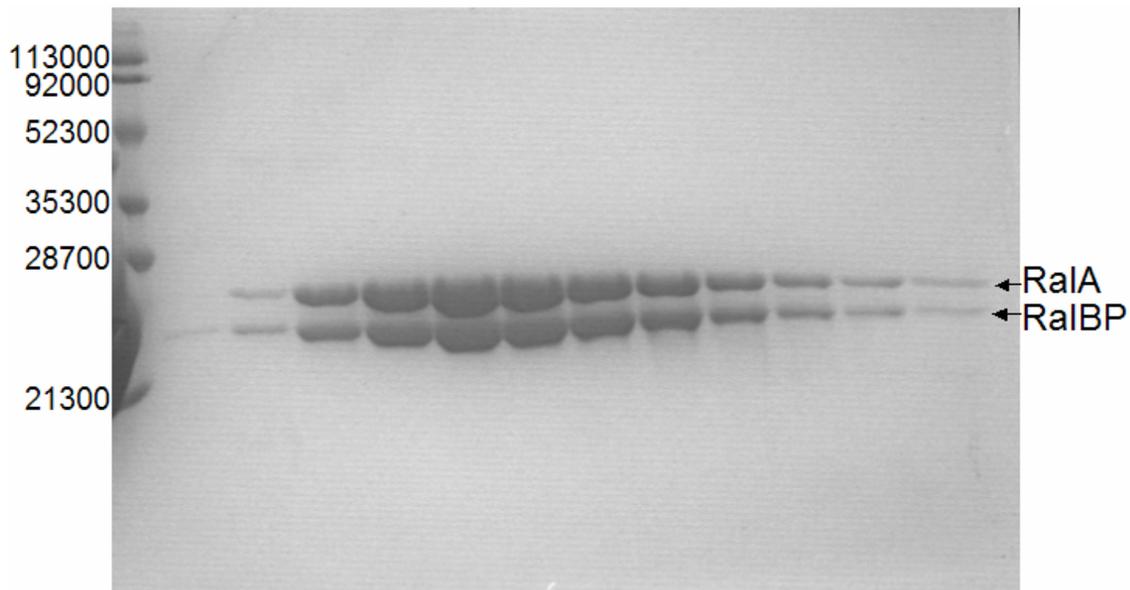


Figure 2. SDS-PAGE gel showing the co-elution of RalA and RalBP(397-518x) in size-exclusion chromatography. The first lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards with their respective molecular weights in Daltons listed at the left. Fractions from the RalA-RalBP(397-518x) complex peak are run in subsequent lanes. RalA and RalBP(397-518x) are labeled at the right and appear to elute in a 1:1 ratio indicating co-elution and complex formation.

Unfortunately, concentration of the complex revealed that the protein was not free from impurities. Upon further investigation, it was confirmed that the RalBP(397-518x) sample was the source of these impurities and that RalBP(397-518x) behaved unusually. During size-exclusion chromatography, the RalA-RalBP(397-518x) complex eluted in a 1:1 ratio, with excess RalA eluting after the complex, as was expected because RalA has a lower molecular weight than that of the complex. The unusual observation was that RalBP(391-597x), which has a lower molecular weight than RalA, eluted first, where the largest molecules should be eluting (Figure 3).

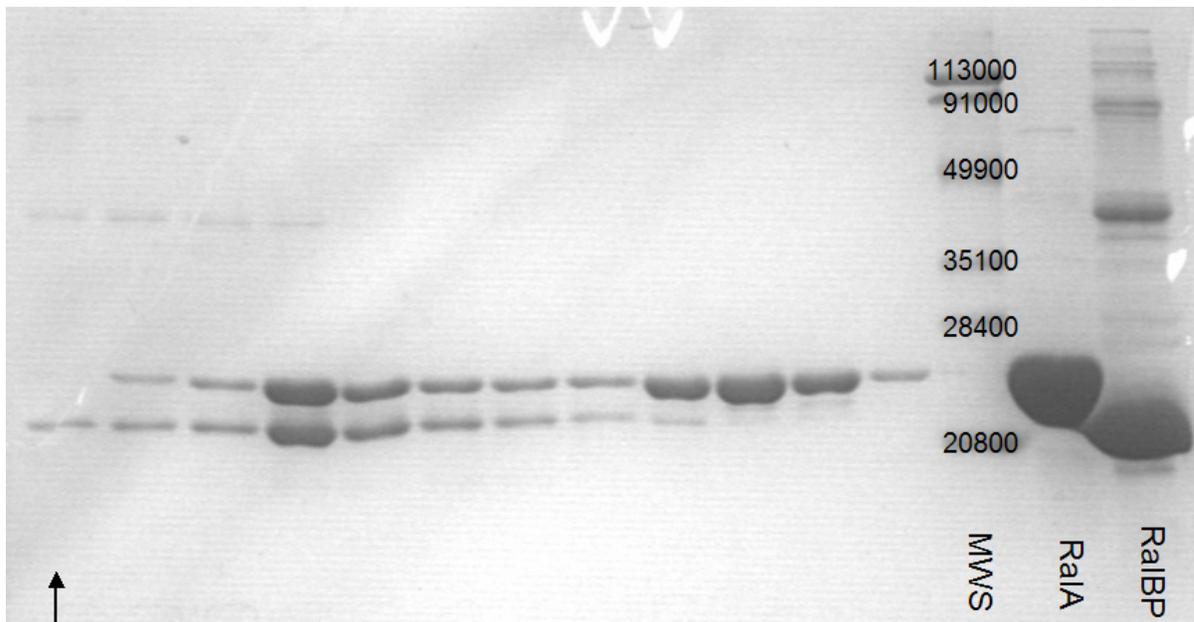


Figure 3. SDS-PAGE gel showing the anomalous behaviour of RalBP397-518x) during the co-elution of RalA and RalBP(397-518x) in size-exclusion chromatography. The MWS lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards with their respective molecular weights in Daltons listed over each standard band. Concentrated samples of RalA and RalBP(397-518x) are run in the two subsequent lanes. The lanes previous to MWS contain fractions from size exclusion chromatography of the RalA-RalBP(397-518x) complex. RalA and RalBP(397-518x) appear to elute in a 1:1 ratio with excess RalA eluting after the complex. The arrow marks the surprising result: unbound RalBP(397-518x) elutes before the complex. Also of note are the faint molecular weight bands in the first four lanes, indicating impurities.

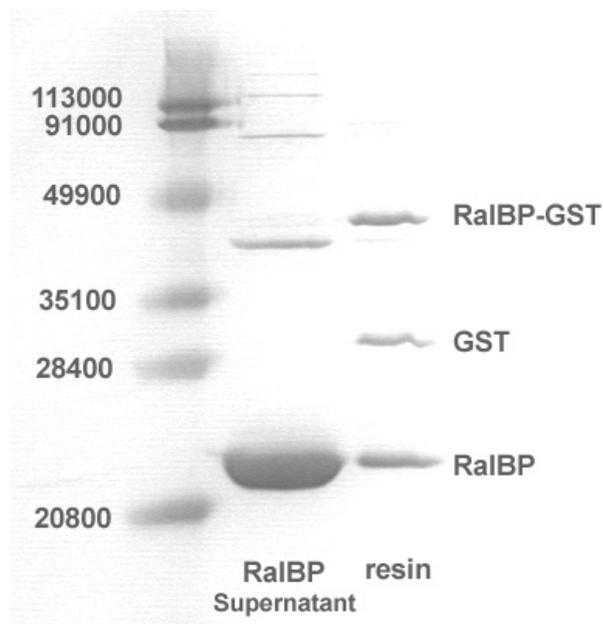


Figure 4. SDS-PAGE gel showing RalBP(397-518x) after the second incubation with glutathione-agarose. The first lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards with their respective molecular weights in Daltons listed at the left. The subsequent lanes show the supernatant containing RalBP(397-518x) and the proteins bound to the glutathione-agarose beads (resin). The proteins are identified at the right.

RalBP(397-518x) Oligomerizes Through its Two Cysteine Residues

Because the complex was still impure, and size-exclusion chromatography after complex formation did not remove the impurities, the focus shifted back to purification of RalBP(397-518x) alone. A number of different strategies were used, including using a new size-exclusion column which had a different molecular weight range for separation, using detergent during cell lysis to prevent RalBP(397-518x) from associating with bacterial proteins, using hydrophobic interaction chromatography, and using a decreased pH (7) for ion exchange to alter the elution profile. None of these strategies removed the impurities. However, introducing a second incubation with glutathione-agarose, as is described in the

Materials and Methods, produced the purest form of RalBP observed thus far (Figure 4), even though impurities were still evident. A native PAGE was run to test for the possibility of using an acrylamide column to separate RalBP(392-518x) from the remaining impurities and a ladder pattern was observed. This suggested that the high molecular weight impurities were actually RalBP(397-518x) oligomers of increasing size, formed through the two cysteine residues present in this construct. To test this, both a native and an SDS-PAGE

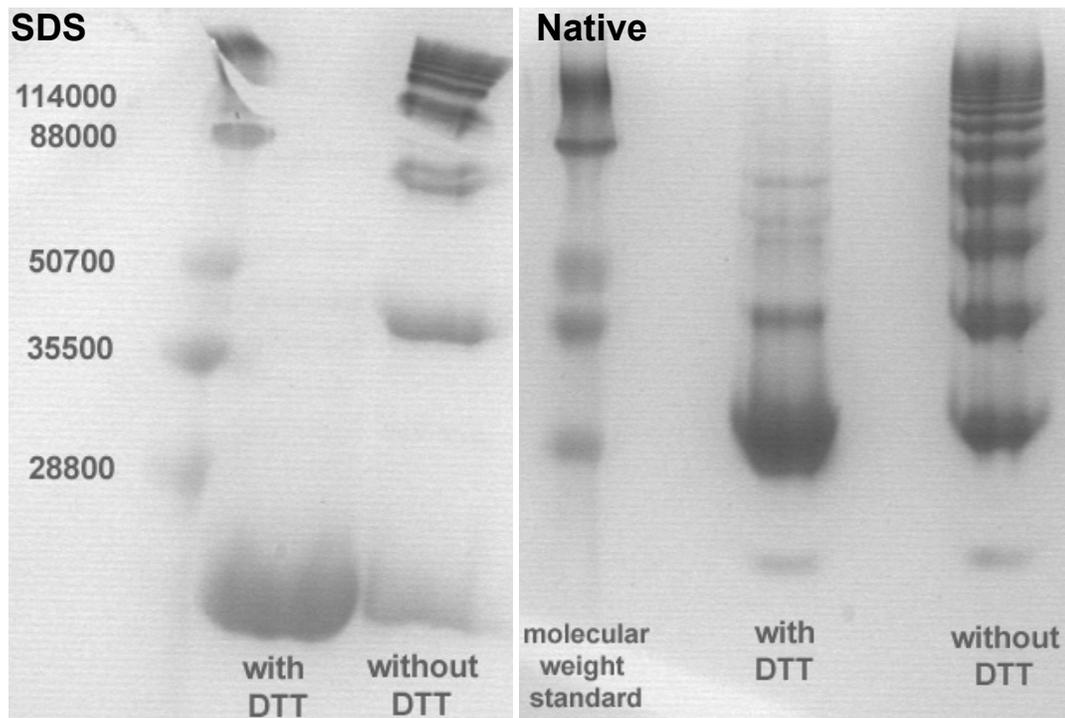


Figure 5. Native and SDS-PAGE gels showing the oligomerization of RalBP(397-518x). The SDS-PAGE gel is located on the left and the Native PAGE gel is located on the right. In both gels, the first lane contains the Bio-Rad Rad Prestained SDS-PAGE Low Range Standards (their respective molecular weights in Daltons listed at the left of the SDS-PAGE gel). In the Native-PAGE, this standard was included for orientation purposes. The middle lane contains the RalBP(397-518x) sample with 10 mM DTT added (with DTT) and the right lane contains no additional DTT (1 mM was contained in the buffer; without DTT). In both the native and SDS-PAGE gels, the disruption of the ladder pattern is observed with the additional DTT.

were run on samples of RalBP(397-518x) with an additional 10 mM DTT compared to the RalBP(397-518x) in the Thrombin Buffer, which contains 1 mM DTT (Figure 5). The additional DTT disrupted the disulfide bonds between the cysteine residues, causing the ladder pattern and the high-molecular weight “contaminants” to compress down into a single band of RalBP(397-518x) monomer. Additionally, a gel slice of one of the higher molecular weight bands was tested using mass spectrometry, which confirmed the identity of the protein as RalBP. This discovery of the oligomerization of RalBP(397-518x) explains the unusual early elution it displays in the size-exclusion chromatography of the complex. While the 300 mM NaCl buffer used for RalBP(397-518x) in size exclusion chromatography appears to disrupt these oligomers, the 150 mM NaCl buffer used for the complex did not.

To prevent this oligomerization, 5 mM TCEP HCl was added to all the buffers used with RalBP(397-518x). This aided in purification of both RalBP(397-518x) and the complex. Crystallization trials were set as described in the Materials and Methods for both RalBP(397-518x) alone and in complex with RalA. This process did not lead to the production of workable protein crystals.

Sequencing and Double Serine Mutant

To remove the need to use TCEP HCl, it was decided to engineer a GST-RalBP construct with serine residues replacing the cysteines. Initial DNA sequencing results revealed that there was an unexpected mutation at residue 472 in the GST-RalBP(397-518x) construct, and additional C-terminal residues that are not found in the wild-type RalBP protein sequence

(Figure 6). The extraneous C-terminal residues (EFIVTD) are a result of relying on the stop codon present in the pGEX-2T vector instead of engineering in a stop codon at the end of the RalBP sequence. Before the double serine mutant was engineered, the mutation at residue 472 was corrected, replacing the threonine with a methionine as is observed in the human wild-type sequence. Once the correction was confirmed by sequencing, the double serine mutant was engineered starting with C451S and then adding C411S, creating the GST-RalBP(397-518x;C411S,C451S) construct. Sequencing verified each of these mutations was present (Figure 7).

GST-RalBP(397-518x;C411S,C451S) was expressed and purified as normal with similar results to the GST-RalBP(397-518x) protein. Complex formation was tested to determine if the serine mutations disrupted the binding of RalBP(397-518x;C411S,C451S) with RalA. Results from size exclusion chromatography (Figure 8a) show RalA and RalBP(397-518x;C411S,C451S) elute together as observed when the cysteine residues are present, indicating complex formation. To further test this observation, native-PAGE was performed on a concentrated sample of RalA, a concentrated sample of RalBP(397-518x;C411S,C451S), and a concentrated sample of the two proteins mixed together but before any purification steps were taken (Figure 8b). The three samples run discretely in the gel. If the mixture of the two proteins did not interact, it would be expected that two bands running similarly to those of the individual components would be observed. What was observed, however, was a single prevalent band, which appears distinct from the other two samples, indicating complex formation. Finally, a Western Blot was performed on a

```

*          300          *          320          *          340          *          360          *          380
Minimum_Bi : ----- : -
Full_Lengt : RFEEACGRITTEKRVQEFQRLKELPECNYLLISWLIVHMDHVIAKELETMKNIQNISIVLSPTVQISNRVLYVFFTHVQELFGNVVVKQVMKPL : 380
Theoretica : ----- : -
Sequenced_ : ----- : -

*          400          *          420          *          440          *          460          *
Minimum_Bi : -----RRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENEN : 73
Full_Lengt : RWSNMATMPTLPETCAGSGIKEETRRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENEN : 475
Theoretica : -----GSGIKEETRRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENEN : 81
Sequenced_ : -----GSGIKEETRRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENEN : 81
                gikeeiRRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENEN

*          480          *          500          *          520          *          540          *          560          *
Minimum_Bi : -----EEVINILLAQENEILTEQEELLAM----- : 97
Full_Lengt : EEVINILLAQENEILTEQEELLAMEQFLRRQIASEKEEIDRLREFIVTDIAEIQRQHGHRSETEEYSESESESEDEEELQIILEDLQRQNEELETKN : 570
Theoretica : -----EEVINILLAQENEILTEQEELLAMEQFLRRQIASEKEEIDRLREFIVTD----- : 130
Sequenced_ : -----EEVINILLAQENEILTEQEELLAMEQFLRRQIASEKEEIDRLREFIVTD----- : 130
                EEVINILLAQENEILTEQEELLAMeqflrrqiasekeei rlr i

*          580          *          600          *          620          *          640          *
Minimum_Bi : ----- : -
Full_Lengt : NHLNCAIHEEREAIIELRVQLRLQLMQRAKAEQCAQEDEEPEWGGAVQPPRDGVLEPKAAKEQPKAGKEFAKPSRDRKETS : 655
Theoretica : ----- : -
Sequenced_ : ----- : -

```

Figure 6. Translated DNA sequencing results of GST-RalBP(397-518x). “Minimum_Bi” is a published Ral minimum binding domain of RalBP (Jullien-Flores et al., 1995), “Full_Lengt” is a portion of the published full-length sequence of RalBP (Jullien-Flores et al., 1995), “Theoretica” is the sequence we had on record, and “Sequenced_” is the translated result from DNA sequencing of the GST-RalBP(397-518x) construct. Including the full-length sequence in the alignment produced the correct numbering for the amino acid residues across the top. The sequences were aligned with ClustalX (Thompson et al., 1994; Thompson et al., 1997) and Genedoc (Nicholas et al., 1997) was used to remove the gaps in the flanking regions and produce the figure. Black indicates residues conserved across all four sequences, dark gray indicates residues conserved across three sequences, and light gray indicates residues conserved across two sequences. The additional glycine and serine residues at the N-terminal end of “Theoretica” and “Sequenced_” belong to the linker sequence joining GST to RalBP and are remnants from thrombin cleavage.

```

*          20          *          40          *          60          *          80
Minimum_Bi : -----RRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENENEVIN : 78
Sequenced_ : GSGIKEETRRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENENEVIN : 86
Dbl_Mutant : GSGIKEETRRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENENEVIN : 86
                gsgikeeiRRQEFLLNCLHRDLGGGKIDLSKEERLWEVQRILTALKRKLREAKRQECETKIAQETIASLSKEDVSKEEENENEVIN

*          100          *          120          *
Minimum_Bi : -----ILLAQENEILTEQEELLAM----- : 97
Sequenced_ : -----ILLAQENEILTEQEELLAMEQFLRRQIASEKEEIDRLREFIVTD : 130
Dbl_Mutant : -----ILLAQENEILTEQEELLAMEQFLRRQIASEKEEIDRLREFIVTD : 130
                ILLAQENEILTEQEELLAMeqflrrqiasekeeiidrlrefivtd

```

Figure 7. Translated DNA sequencing results of the GST-RalBP(397-518x) double mutant, GST-RalBP(397-518x;C411S,C451S). “Minimum_Bi” is a published Ral minimum binding domain of RalBP (Jullien-Flores et al., 1995), “Sequenced_” is the translated result from DNA sequencing of the GST-RalBP(397-518x) construct, and “Dbl_Mutant” is the translated result from DNA sequencing of the GST-RalBP fusion protein with the corrected Thr 472, and the double mutation (C411S, C451S), also known as the GST-RalBP(397-518x;C411S, C451S) construct. According to the numbering along the top of this figure, C411S is numbered 17, C451S is numbered 57, and Thr 472 is numbered 78. The sequences were aligned with ClustalX (Thompson et al., 1994; Thompson et al., 1997) and Genedoc (Nicholas et al., 1997) was used to remove the gaps in the flanking regions and produce the figure. Black indicates residues conserved across all three sequences, and gray indicates residues conserved across two sequences. In the “Dbl_Mutant” sequence, note the two serine mutations and the reversion back to methionine.

similarly run native-PAGE (Figure 8c) with anti-RalA (N-19) antibody (Santa Cruz Biotechnology). As was expected, this antibody did not detect protein in the molecular weight standard or the RalBP(397-518x;C411S,C451S) sample. In the RalA and mixed RalA-RalBP(397-518x;C411S,C451S) samples, the antibody indicates a discrete shift in RalA, which is further evidence of the formation of the RalA-RalBP(397-518x;C411S,C451S) complex.

RalBP(397-518x) Folds Upon Binding to RalA

The initial sequencing results highlighted the difference between RalBP(397-518x) and the published Ral minimum binding domain of RalBP (Jullien-Flores et al., 1995) as can be observed in Figures 6 and 7. It was thought that a shorter construct of RalBP would improve the chances of crystallization, but before embarking on engineering the shorter constructs, the protein sequence of RalBP(397-518x) was used as input with the Predictor Of Naturally Disordered Regions (PONDR; Romero et al., 2002) to computationally determine if there were any flanking regions of disorder which would best be excluded from the new construct. The VL-XT predictor (Li et al., 1999; Romero et al., 1997; Romero et al., 2001), which was trained on sequences from NMR or crystallographically determined regions of 40 or more residues, and crystallographically determined amino- and carboxyl-terminal regions of five or more residues, was used. To our surprise, the majority of the RalBP(397-518x) protein was predicted to be disordered (Figure 9). With values greater than or equal to 0.5 being considered as indicative of protein disorder, only residues 8-15 and 124-130 of the RalBP(397-518x) construct (see “Sequenced_” in Figure 7 for numbering) had values below

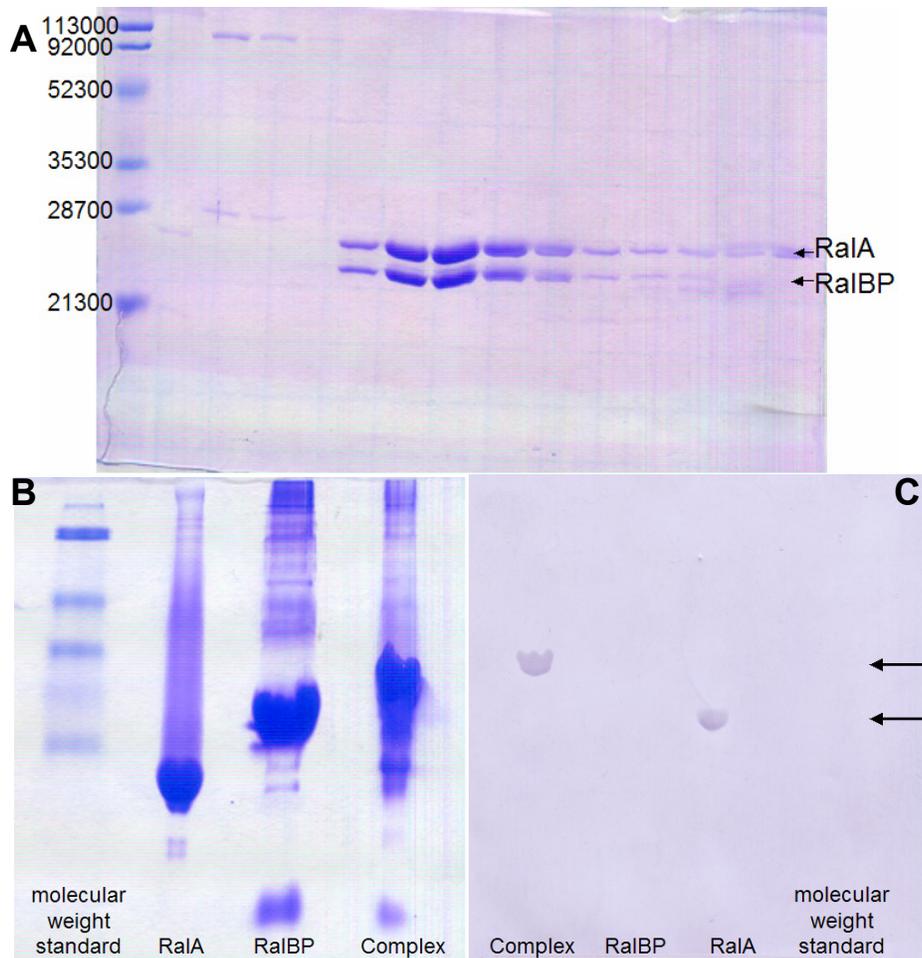


Figure 8. The double serine mutant, RalBP(397-518x;C411S,C451S), binds RalA.

A. (top) SDS-PAGE gel showing the co-elution of RalA and RalBP(397-518x;C411S,C451S) during size exclusion chromatography. The first lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards with their respective molecular weights in Daltons listed at the left. Eluted fractions are run in subsequent lanes with high molecular weight contaminants eluting first, then the RalA-RalBP(397-518x;C411S,C451S) complex, and finally excess RalA. RalA and RalBP(397-518x;C411S,C451S) are labeled at the right and appear to elute in a 1:1 ratio indicating co-elution and complex formation. B. (bottom left) Native-PAGE gel of the RalA-RalBP(397-518x;C411S,C451S) complex. The first lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards, which were included for orientation purposes. The subsequent lanes contain a concentrated sample of RalA, a concentrated sample of RalBP(397-518x;C411S,C451S), and the RalA-RalBP(397-518x;C411S,C451S) complex. Note that in the rightmost lane, there is only one dominant band and it has shifted from those observed with either RalA or RalBP(397-518x;C411S,C451S), indicating that the complex formation. C. (bottom right) Western Blot of a Native-PAGE gel of the RalA-RalBP(397-518x;C411S,C451S) complex. This blot was performed on a similarly run Native-PAGE gel as in Figure 8b, only the lanes are reversed. The rightmost lane contains the Bio-Rad Kaleidoscope Precision Plus Protein Standards, which were included for orientation purposes, and as an additional negative control. The preceding lanes contain concentrated samples of the RalA-RalBP(397-518x;C411S,C451S) complex, RalBP(397-518x;C411S,C451S), and RalA. The anti-RalA (N-19) antibody (Santa Cruz Biotechnology) detects the presence of RalA in the RalA and complex samples as indicated by the arrow at the right. The discrete shift in RalA indicates complex formation.

0.5, and residues 125-130 were remnants of the pGEX-2T vector and are not contained in the wild-type RalBP protein sequence.

To test this surprising information, we began circular dichroism studies. 20 μ M samples were prepared of RalA, RalBP(397-518x), and the RalA-RalBP(397-518x) complex in a phosphate buffer lacking Cl⁻. The change in buffer contents was made because earlier attempts with HEPES buffer and the Cl⁻ from NaCl and MgCl₂ prevented data acquisition at shorter wavelengths. Near-UV (250-350 nm) and Far-UV (190-250 nm) spectra were collected for

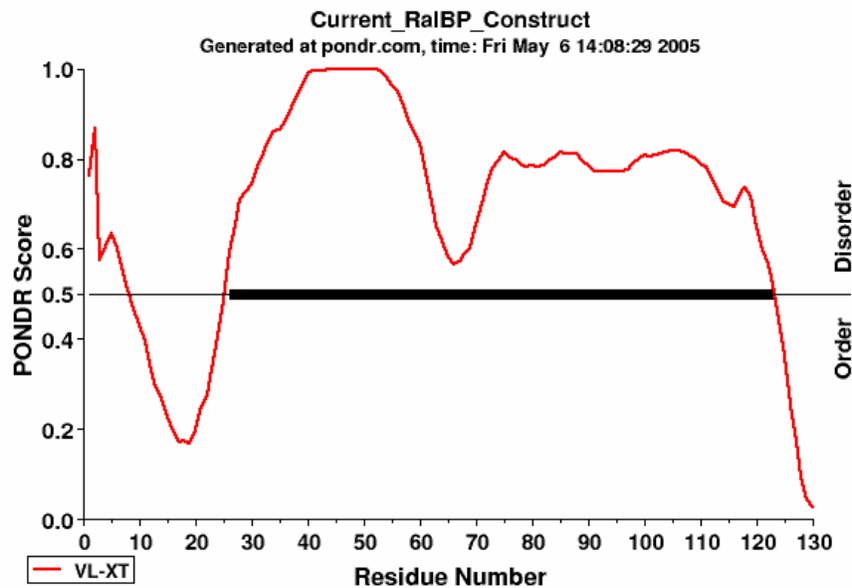


Figure 9. PONDNR predicts RalBP(397-518x) to be mostly disordered. The protein sequence of RalBP(397-518x) was used as input for PONDNR(Romero et al., 2002) using the VL-XT predictor (Romero et al., 1997; Li et al., 1999; Romero et al., 2001). Values of 1 and 0 correspond to ideal predictions of disorder and order, respectively, and values greater than or equal to 0.5 indicate protein disorder. Residue numbering corresponds to that of “Sequenced_” in Figure 7.

the three samples and a blank, although it was only possible to acquire data for wavelengths longer than 207 nm, possibly because of the high concentration (10 mM) of DTT present in the samples. The spectra of the three samples with blank subtracted are plotted in Figure 10a. The Near-UV (250-350 nm) spectra provide information about the overall tertiary structure, and proteins with secondary but no defined tertiary structure will have a signal of nearly zero. The Far-UV (190-250 nm) is useful in determining secondary structure, as α -helices, β -sheets, and random coils all exhibit distinct spectra (Kelly and Price, 2000). RalA and the Complex have non-zero Near-UV spectra, indicating a defined tertiary structure for both of these samples. On the other hand, RalBP(397-518x) has a nearly zero Near-UV spectrum, indicating that RalBP(397-518x) is behaving like a molten globule, and confirming the PONDR results. The Far-UV spectrum for RalBP(397-518x) has a minimum near 222 nm suggesting an α -helical character, which is consistent with secondary structure prediction (Cantor et al., 1995; Jullien-Flores et al., 1995). The α/β structure of RalA is reflected in its Far-UV spectrum, and the Far-UV spectrum for the Complex shows an increase in secondary structure compared to RalA and RalBP(397-518x).

Further comparisons of the spectra were made. First, the RalBP(397-518x) signal was subtracted from the Complex signal, resulting in the theoretical spectrum named “Complex-RalBP” (Figure 10b). If there were no conformational changes in either protein upon formation of the complex, this theoretical spectrum should equal that of RalBP alone. With both the Near-UV and the Far-UV, this is not the case, and higher intensity of this signal compared to the RalA signal indicates that the conformational change increases the presence

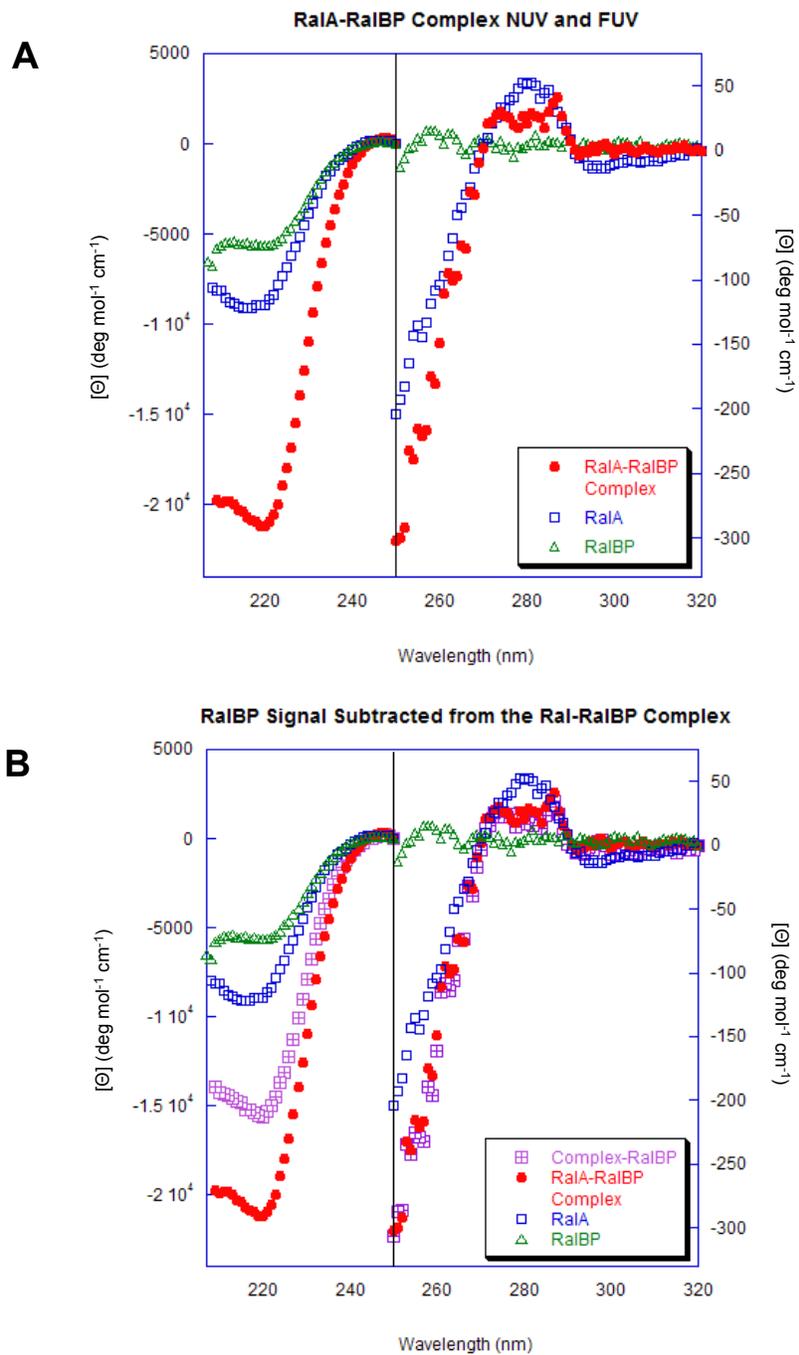


Figure 10. Circular Dichroism shows RalBP(397-518x) folds upon binding to RalA. A. (top) CD spectra from RalA alone, RalBP(397-518x) alone, and the RalA-RalBP(397-518x) Complex. The Near UV (320-250 nm) signal for RalBP(397-518x) shows very little tertiary structure and the Far UV (250-210 nm) for RalBP(397-518x) shows the least amount of secondary structure of the three samples B. (bottom) CD spectra from RalA alone, RalBP(397-518x) alone, the RalA-RalBP(397-518x) Complex, and the theoretical signal of RalBP(397-518x) subtracted from the Complex signal. This theoretical signal should equal the RalA signal if there is no conformational change in RalBP(397-518x) or RalA upon complex formation.

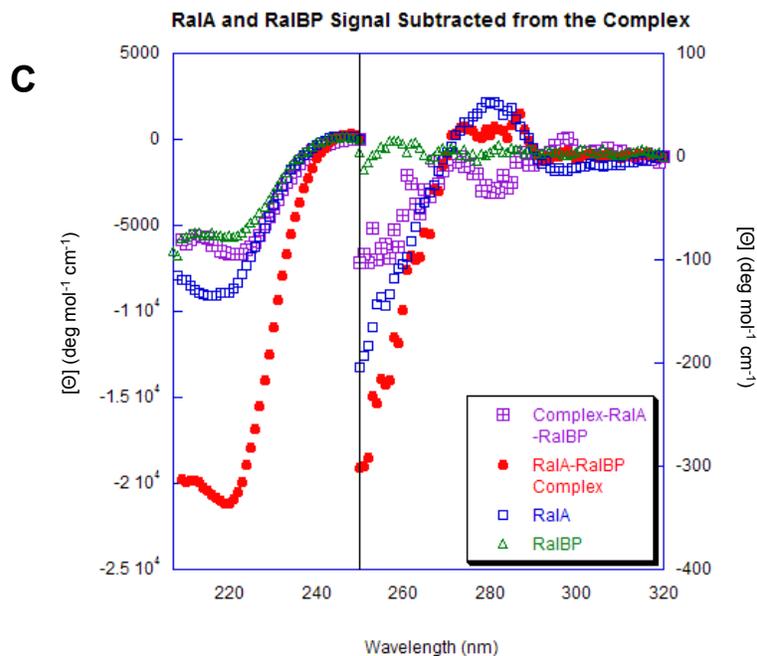


Figure 10 continued. C. CD spectra from RalA alone, RalBP(397-518x) alone, the RalA-RalBP(397-518x) Complex, and the theoretical signal of RalA and RalBP(397-518x) subtracted from the Complex signal. This theoretical signal should equal nearly zero if there is no conformational change in RalBP(397-518x) or RalA upon complex formation.

of defined secondary and tertiary structure. An additional theoretical spectrum was calculated by further subtracting the RalA signal from “Complex-RalBP”, and this new theoretical spectrum is named “Complex-RalA-RalBP” (Figure 10c). If there is no conformational change in either protein upon binding, this theoretical spectrum should equal zero, which is not the case. Given that RalA is a mostly ordered protein, it was concluded that this difference is primarily the result of RalBP(397-518x) folding upon binding to RalA.

RalBP(403-499) Does Not Bind to RalA

Based on the conclusion that RalBP(397-518x) is mostly disordered in solution, it was decided that the best strategy for crystallization was to focus on the RalA-RalBP complex. To improve the chances of complex crystallization, extraneous residues outside of the published minimum Ral binding domain (Cantor et al., 1995) were removed and the RalBP(403-499) construct was engineered. This construct contains the double serine mutation to prevent oligomerization.

When GST-RalBP(403-499) was expressed and purified, it was discovered that a second protein was being overexpressed, and it co-eluted with GST-RalBP(403-499) during affinity chromatography. Because of this, the purification protocol had to be adjusted to include an anion exchange step at pH 6.2 before the affinity chromatography. Once purified from this contaminant, the thrombin cleavage and further purification proceeded as normal.

Unfortunately, complex formation did not. The complex reaction was allowed to proceed as normal and then size exclusion chromatography was used to remove any contaminants. The SDS-PAGE gel (Figure 11a) from the resulting elution profile raised concerns that RalBP(403-499) did not bind RalA. While the elution of RalA and RalBP(403-499) overlaps, it does not appear as though the two proteins are eluting together in the expected 1:1 ratio. Furthermore, it appears that RalBP(403-499), the smaller of the two proteins, elutes slightly earlier than RalA, which is unexpected.

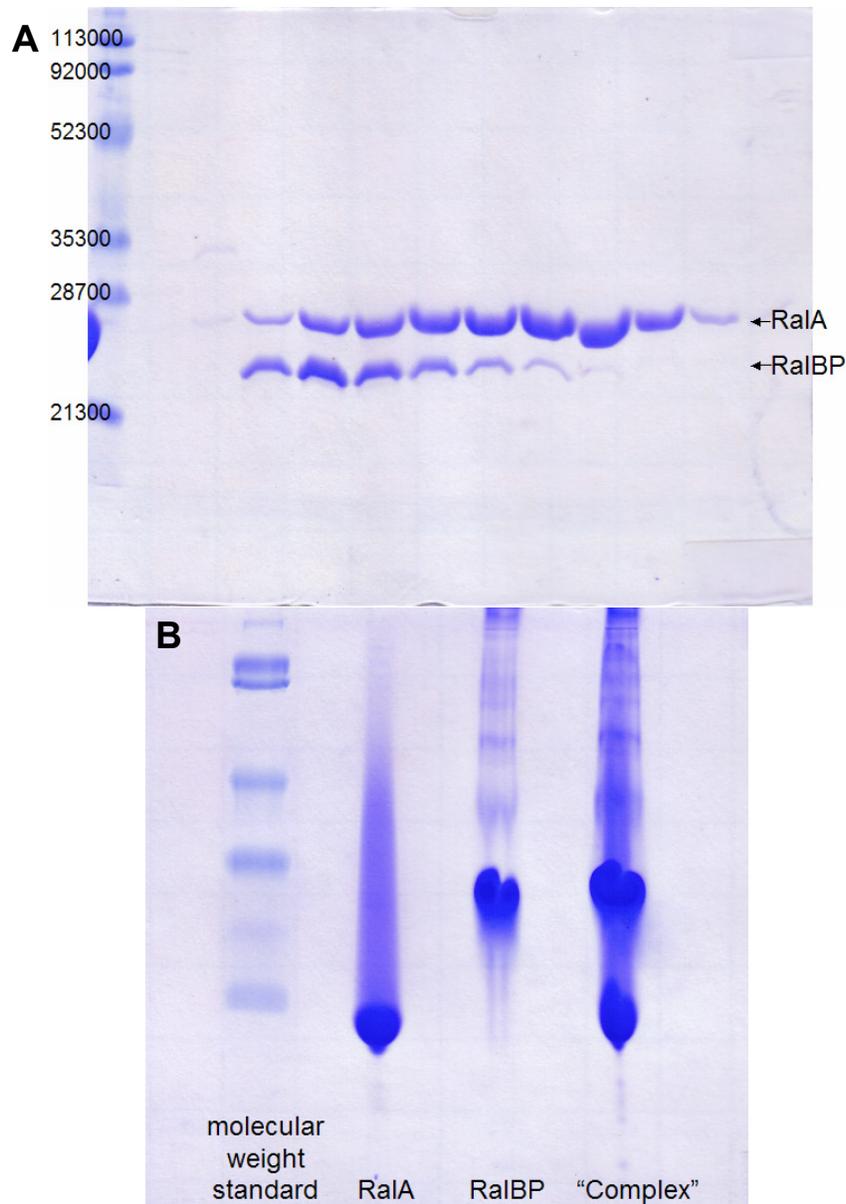


Figure 11. Native and SDS-PAGE gels showing RalBP(403-499) does not bind to RalA. A. (top) SDS-PAGE gel showing RalA and RalBP(403-499) do not co-elute during size exclusion chromatography. The first lane contains the BioBio-Rad Prestained SDS-PAGE Low Range Standards with their respective molecular weights in Daltons listed at the left. Eluted fractions are run in subsequent lanes with high molecular weight contaminants eluting first. The elution of RalBP(403-499) and RalA overlaps, but it is not clear that the two proteins co-elute in a 1:1 ratio, which calls complex formation into question. RalA and RalBP(403-499) are labeled at the right. B. (bottom) Native-PAGE gel confirming the RalA-RalBP(403-499) complex does not form. The first lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards, which were included for orientation purposes. The subsequent lanes contain a concentrated sample of RalA, a concentrated sample of RalBP(403-499), and the result of the RalA-RalBP(403-499) complex formation. RalBP(403-499) was added in 2:1 molar excess to RalA, and if these two proteins bound in complex, the lower band corresponding to RalA should not appear in the “Complex” lane. The presence of the two bands in the rightmost lane indicates that the RalA-RalBP(403-499) complex does not form.

To test for the formation of the RalA-RalBP(403-499) complex, the two proteins were allowed to incubate as normal, but with RalBP(403-499) in 2:1 molar excess of RalA. A native PAGE (Figure 11b) was run using samples of RalA, RalBP(403-499), and the mixture of the two proteins. As there are twice the molecules of RalBP(403-499) present in the mixture, it was expected that the band corresponding to RalA would not be present if the complex of these two proteins was forming. As is observed in Figure 11b, this is not the case and the lower protein band, corresponding to RalA, is present in the “Complex” sample indicating that the RalA-RalBP(403-499) complex does not form.

The reason for the anomalous elution of RalBP(403-499) in the size exclusion chromatography step after the mixing of Ral and RalBP(403-499) is unclear. It is possible that the truncation of RalBP disrupted structural elements of the protein introducing structural disorder and causing the protein to migrate through the matrix differently than a folded protein of a similar molecular weight.

Six new truncated constructs of RalBP designed

Because RalBP(403-499) did not bind to RalA, new truncated constructs of RalBP were designed. As RalBP(397-518x;C411S,C451S) binds RalA, it was decided to continue to use the DNA from this construct as the starting point for sequence engineering and to maintain the serine mutations in the new constructs of RalBP. The first sequence designed was identical to RalBP(397-518x;C411S,C451S), only it has a triple stop codon immediately

following the codon for Arg 518, resulting in a protein that does not contain C-terminal artifacts of the vector. This construct is named RalBP(397-518).

With the failure of RalBP(403-499) to bind RalA, the question now was what the truncated sequence should be. Three concurrent studies identified RalBP with different methods and giving it different names (Cantor et al., 1995; Jullien-Flores et al., 1995; Park and Weinberg, 1995). Two of these studies identified specific residues for the minimum Ral binding domain, and one of these protein constructs, RalBP(403-499) (Jullien-Flores et al., 1995), failed to bind RalA. The next truncated construct corresponds to the minimum Ral binding domain of the second study (Park and Weinberg, 1995), RalBP(391-444). Residues 391-396 of RalBP were not present in any of the previous constructs and engineering them into the sequence of RalBP(391-444) was required.

In addition to RalBP(397-518) and RalBP(391-444), four additional constructs of RalBP were designed. To design these constructs, combinations of endpoints were used. Residues 391 and 397 were used as start points and residues 444, 499, and 518 were used as end points. All of these protein sequences along with the two published minimum Ral binding domains are shown in the sequence alignment of Figure 12. Of these six new constructs, only RalBP(397-518) and RalBP(391-444) were successfully cloned into the pGEX-2T vector.

RalBP(397-518) Binds RalA

RalBP(397-518) was expressed and purified as normal with results similar to RalBP(397-518x). Size exclusion chromatography was run to test for the formation of the complex, and the results (Figure 13a) show RalA and RalBP(397-518) elute together as was observed with the RalBP(397-518x;C411S,C451S) construct. This is not surprising as the only difference between these two constructs is the removal of the six C-terminal residues (Figure 12), which are remnants of the vector and should not contribute to RalA-RalBP binding. For further verification, a native-PAGE was performed on concentrated samples of RalA, RalBP(397-518), and the RalA-RalBP(397-518) complex before size exclusion chromatography (Figure 13b). When comparing the RalA and the Complex sample, a shift of RalA is observed,

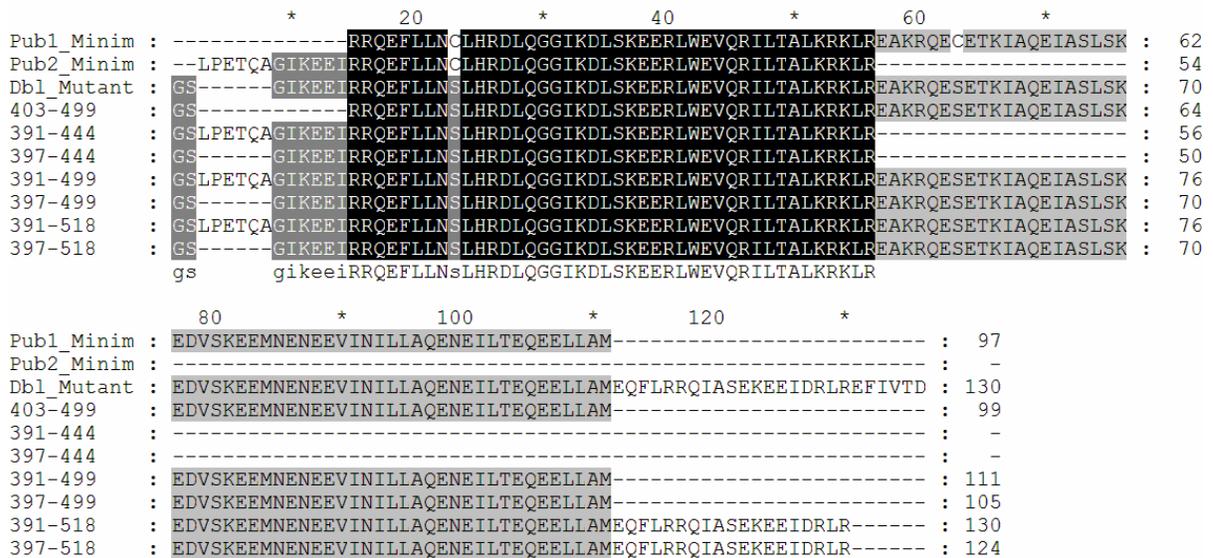


Figure 12. Sequences of RalBP constructs. “Pub1_Minim” is the Ral minimum binding domain of RalBP published by Jullien-Flores *et al* (1995), “Pub2_Minim” is the Ral minimum binding domain of RalBP published by Park and Weinberg (1995), “Dbl_Mutant” is RalBP(397-518x;C411S,C451S). “403-499”, “391-444”, “397-444”, “391-499”, “397-499”, “391-518”, and “397-518” are the truncated RalBP constructs with the range of residues designated by each label.

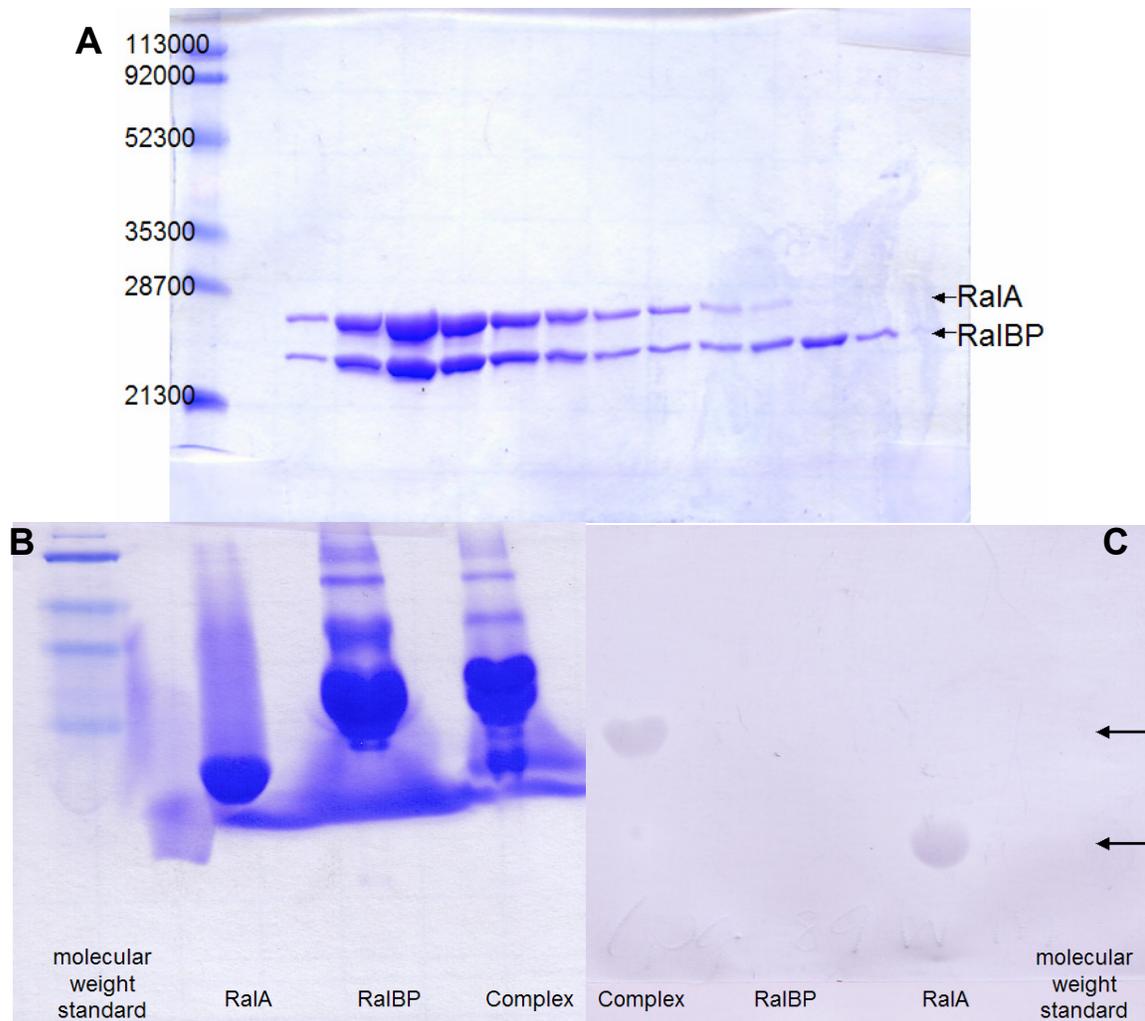


Figure 13. RalBP(397-518) binds RalA.

A. (top) SDS-PAGE gel showing the co-elution of RalA and RalBP(397-518) during size exclusion chromatography. The first lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards with their respective molecular weights in Daltons listed at the left. Eluted fractions are run in subsequent lanes with high molecular weight contaminants eluting first, then the RalA-RalBP(397-518) complex, and finally excess RalBP(397-518). RalA and RalBP(397-518) are labeled at the right and appear to elute in a 1:1 ratio indicating co-elution and complex formation. B. (bottom left) Native-PAGE gel of the RalA-RalBP(397-518) complex. The first lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards, which were included for orientation purposes. The subsequent lanes contain a concentrated sample of RalA, a concentrated sample of RalBP(397-518), and the RalA-RalBP(397-518) complex. Note that in the rightmost lane, there is only one dominant band and it has shifted from those observed with either RalA or RalBP(397-518), indicating that the complex formation. The shift is especially apparent for RalA. C. (bottom right) Western Blot of a Native-PAGE gel of the RalA-RalBP(397-518) complex. This blot was performed on a similarly run Native-PAGE gel as in Figure 13b, only the lanes are reversed. The rightmost lane contains the Bio-Rad Prestained SDS-PAGE Low Range Standards, which were included for orientation purposes, and as an additional negative control. The preceding lanes contain concentrated samples of the RalA-RalBP(397-518) complex, RalBP(397-518), and RalA. The anti-RalA (N-19) antibody (Santa Cruz Biotechnology) detects the presence of RalA in the RalA and complex samples as indicated by the arrows at the right. The discrete shift in RalA indicates complex formation.

indicating complex formation. Finally, a Western Blot was performed on a similarly run native-PAGE (Figure 13c) with anti-RalA (N-19) antibody (Santa Cruz Biotechnology). As was expected, RalA was not detected in the molecular weight standard or the RalBP(397-518) sample. In the RalA and the mixed RalA-RalBP(397-518) samples, the antibody indicates a discrete shift in RalA, which further confirms the formation of the RalA-RalBP(397-518) complex.

RalBP(391-444) binds RalA

RalBP(391-444) was expressed as normal, but the purification introduced new challenges. First, it had been noticed that the large amounts of thrombin being used for cleavage of the fusion protein were introducing impurities into the protein sample. By decreasing the reaction volume and increasing the temperature, it was possible to reduce the amount of thrombin necessary for cleavage, and thus reduce the contaminants introduced with it. The second challenge was that the RalBP(391-444) protein had a significantly different pI than the other constructs of RalBP (Table 3). This did not affect the affinity chromatography or the thrombin cleavage reaction, but RalBP(391-444) could not be purified using anion exchange chromatography. The most successful strategy for purifying RalBP(391-444) was to apply the unbound protein from an anion exchange column directly onto a cation exchange column, and then elute the protein from the cation exchange column with a gradient (Figure 14).

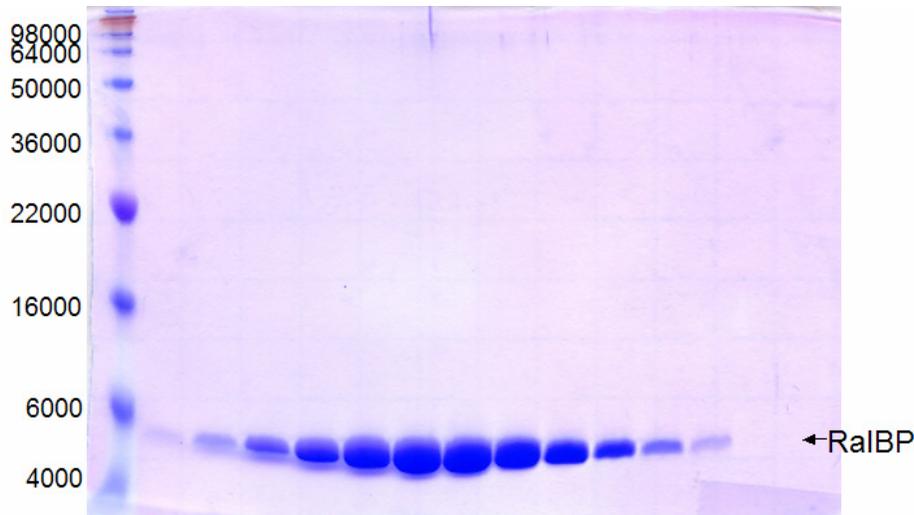


Figure 14. SDS-PAGE gel of purified RalBP(391-444). The leftmost lane contains the Invitrogen SeeBlue Plus2 Pre-Stained Standards with their respective molecular weights in Daltons listed at the left. The subsequent lanes show eluted fractions of RalBP(391-444), which is indicated at the right, from cation exchange chromatography (26/10 HiLoad SP Sepharose HP column; GE Healthcare).

Once purification of RalBP(391-444) was successful, the question was if it bound RalA.

RalBP(391-444) was mixed with RalA and allowed to incubate on ice for about an hour to allow the complex to form. Because of the small size of RalBP(391-444), 6.5 kD compared to 20 kD of RalA, the proteins were mixed in a 2:1 RalBP(391-444):RalA molar ratio to allow for better separation during size exclusion chromatography. When size is the basis for separation, a protein complex with a molecular weight of 26.5 kD can be separated from a protein with a molecular weight of 6.5 kD more distinctly than from a protein with a molecular weight of 20 kD. This molar ratio resulted in no detectable excess RalA and the RalA-RalBP(391-444) complex eluting discretely from RalBP(391-444) as is observed in the size-exclusion chromatogram (Figure 15a). Fractions corresponding to the two peaks observed in the chromatogram were run on an SDS-PAGE and RalA and RalBP(391-444)

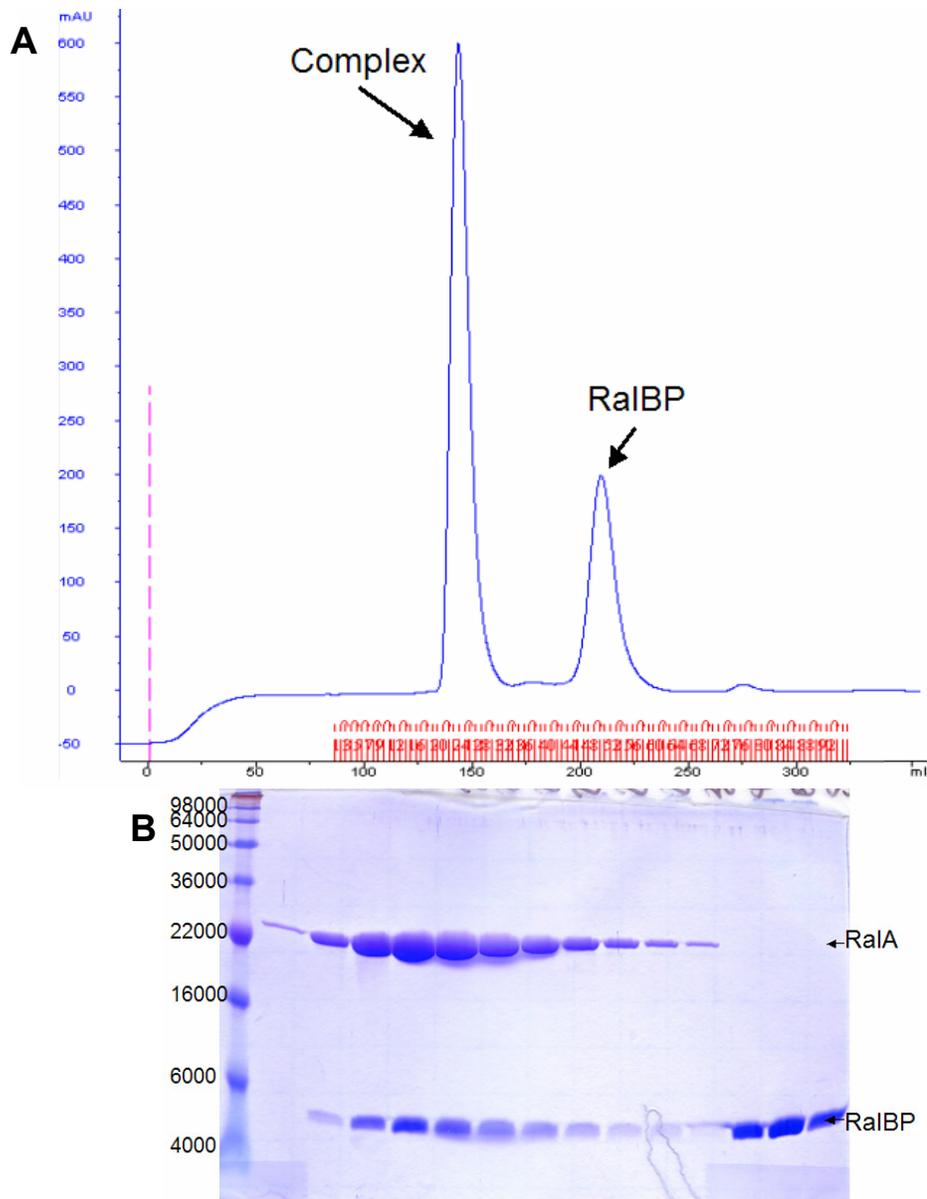


Figure 15. RalBP(391-444) binds RalA.

A. (top) Typical size-exclusion chromatogram of RalA-RalBP(391-444) complex. RalA and RalBP(391-444) are mixed together and incubated on ice before being applied to an equilibrated size exclusion column, where the RalA-RalBP(391-444) complex elutes first, followed by any excess unbound protein. Because of the difference in molecular weight of the complex and RalBP(391-444), including excess RalBP(391-444) ensures that there is no excess RalA and improves separation between the complex and any unbound protein. B. (bottom) SDS-PAGE gel showing the co-elution of RalA and RalBP(391-444) during size exclusion chromatography. The leftmost lane contains the Invitrogen SeeBlue Plus2 Pre-Stained Standards with their respective molecular weights in Daltons listed at the left. The subsequent lanes are samples of fractions corresponding to the peaks in Figure 15a. Samples from the first peak show RalA and RalBP(391-444) eluting together, where the last three lanes, which were taken from the second peak, have RalBP(391-444) eluting alone. The co-elution of RalA and RalBP(391-444) and the distinct separation of the complex and RalBP(391-444) alone indicate complex formation.

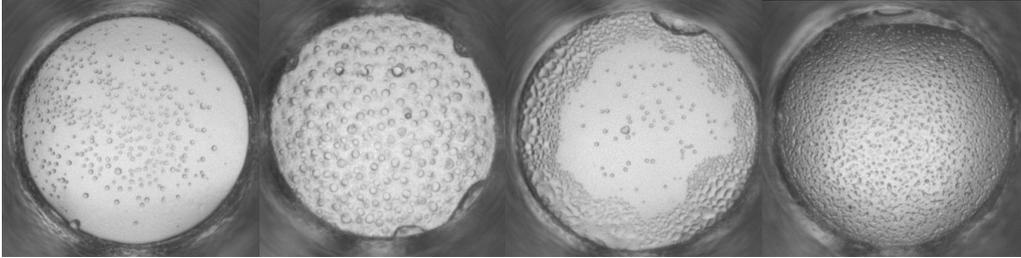
elute together in the first peak, and excess RalBP(391-444) elutes in the second peak (Figure 15b). The elution of RalBP(391-444) at two clearly separated points in size-exclusion chromatography indicates the formation of the RalA-RalBP(391-444) complex.

Crystal Screening

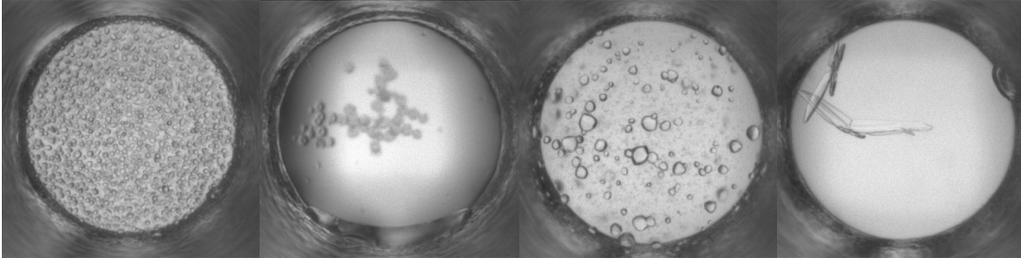
Numerous crystallization trials were set up with commercially manufactured screens when a sufficient amount of pure protein was available. A summary of the crystallization experiments is given in Table 4. None of the crystallization trials produced protein crystals. In addition to the in-house screening efforts, samples of RalBP(391-444), the RalA-RalBP(391-444) complex, and the RalA-RalBP(397-518) complex were sent to the high-throughput screening facility at the Hauptman-Woodward Medical Research Institute. For each sample, 400 μ L of protein with a concentration of 10 mg/mL was used for 1536 crystallization experiments. Pictures were taken at time points of 0, 1 day, 1 week, 2 weeks, 3 weeks, and 4 weeks. Some examples of promising results for RalBP(391-444), the RalA-RalBP(391-444) complex, and the RalA-RalBP(397-518) complex are shown in Figure 16a, 16b, and 16d, respectively. Many of the promising hits were from conditions with phosphate buffers, leading to the suspicion that they were Magnesium Phosphate crystals. To avoid this, an additional sample was sent to the high-throughput screening facility: RalA-RalBP(391-444) in a buffer containing 0.5 MgCl₂ and 5 mM NaCl in place of the normal 5 mM MgCl₂. Select examples of promising hits for this sample are shown in Figure 16c.

Promising hits from the various samples were reproduced in-house and none of the conditions produced protein crystals.

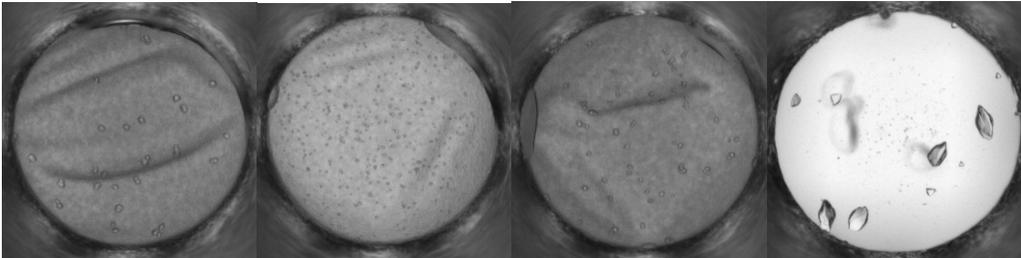
A.



B.



C.



D.

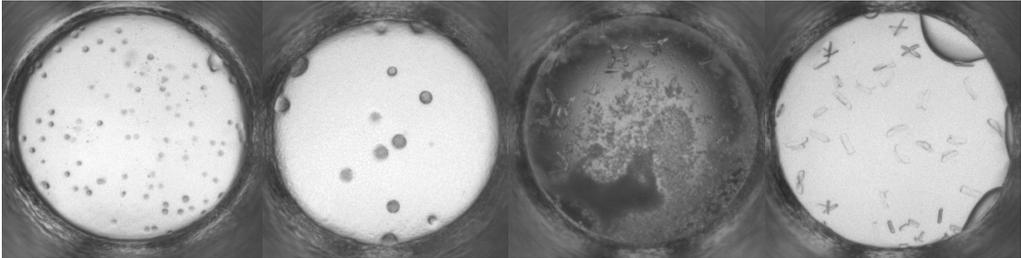


Figure 16. Select examples of High-Throughput Screening results. A. High-Throughput Screening Results from RalBP(391-444). B. High-Throughput Screening Results from RalA-RalBP(391-444) Complex. C. High-Throughput Screening Results from RalA-RalBP(391-444) Complex in a buffer with low $MgCl_2$. D. High-Throughput Screening Results from RalA-RalBP(397-518) Complex

RalA(11-178) Binds RalBP(391-444) and Crystal Screening

Because the crystallization trials using RalA(1-178) were unsuccessful, the N-terminal truncated form, RalA(11-178) was used to form the complex. This construct of RalA is similar to that in which crystals were obtained for the complexes of RalA-Sec5, RalA-Exo84, and RalA-C3bot (Fukai et al., 2003; Jin et al., 2005; Pautsch et al., 2005). RalA(11-178) was expressed and purified similarly to RalA(1-178). Size exclusion chromatography of the mixture of RalA(11-178) and RalBP(391-444) revealed a single peak, which eluted similarly to the RalA(1-178)-RalBP(391-444) complex (Figure 17a). Fractions corresponding to the peak observed in the chromatogram were run on an SDS-PAGE and RalA(11-178) and RalBP(391-444) elute together (Figure 17b). These results indicate the formation of the RalA(11-178)-RalBP(391-444) complex.

Crystallization trials were set up with this complex and a summary of these experiments is given in Table 4. One protein crystal was found with Hampton Crystal Screen 2 condition number 34 (0.05 M cadmium sulfate hydrate, 0.1 M HEPES pH 7.5, 1 M sodium acetate trihydrate). This single crystal was flash-frozen without cryoprotection and sent to the Advanced Photon Source for data collection. The diffraction pattern from this crystal revealed it to be a weakly diffracting protein crystal (Figure 18). Attempts to reproduce and improve these crystals by varying protein concentration, precipitant concentration, and pH, are ongoing.

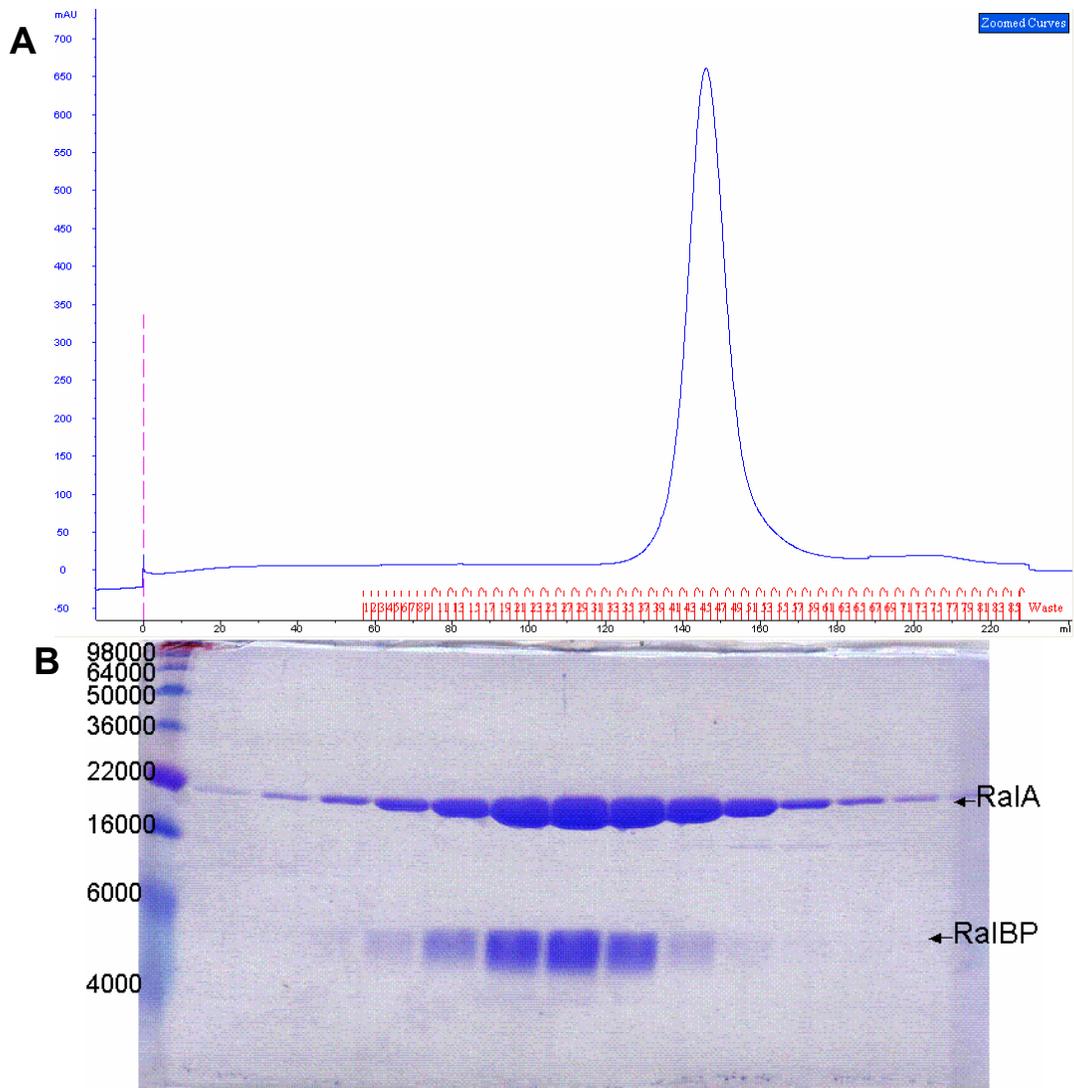


Figure 17. RalBP(391-444) binds RalA(11-178).

A. Typical size-exclusion chromatogram of RalA(11-178)-RalBP(391-444) complex. RalA(11-178) and RalBP(391-444) are mixed together and incubated on ice before being applied to an equilibrated size exclusion column, where the RalA(11-178)-RalBP(391-444) complex elutes first, followed by any excess unbound protein. In this chromatogram, the complex elutes as a single peak with no excess unbound protein.

B. SDS-PAGE gel showing the co-elution of RalA(11-178) and RalBP(391-444) during size exclusion chromatography. The leftmost lane contains the Invitrogen SeeBlue Plus2 Pre-Stained Standards with their respective molecular weights in Daltons listed at the left. The subsequent lanes are samples of fractions corresponding to the peak in Figure 17a. Samples show RalA(11-178) and RalBP(391-444) eluting together. The co-elution of RalA(11-178) and RalBP(391-444) indicate complex formation.

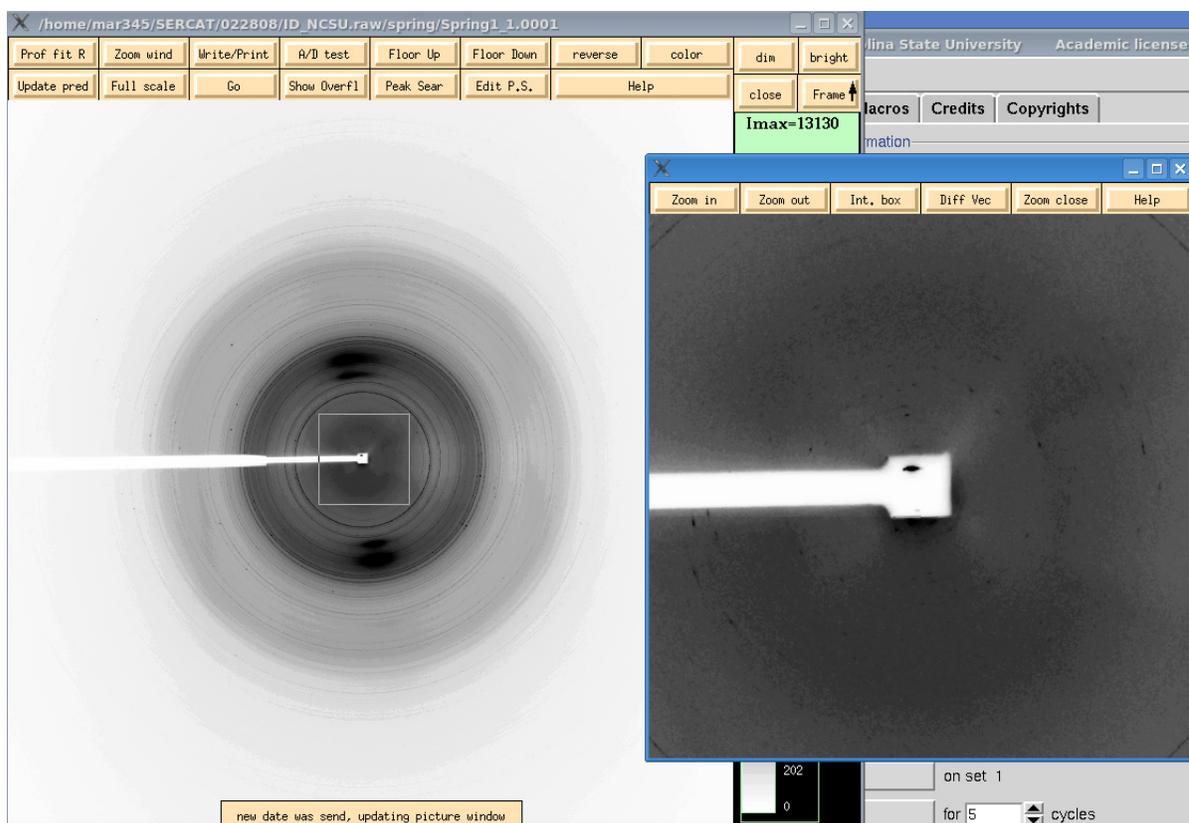


Figure 18. Diffraction pattern from the RalA(11-178)-RalBP(391-444) complex. Data were collected from a crystal grown in Hampton Crystal Screen 2 condition number 34 (0.05 M cadmium sulfate hydrate, 0.1 M HEPES pH 7.5, 1 M sodium acetate trihydrate) and flash-frozen without cryoprotection. The blue box at the right contains a zoomed-in view of the center of the diffraction pattern at the left (delineated by the white square in the left diffraction pattern). The faint reflections near the beamstop are indicative of a weakly diffracting protein crystal.

Discussion

Advances Toward the Crystal Structure of the RalA-RalBP Complex

From the beginning, work with RalBP has revealed a protein that has behaved in unexpected ways. Initial characterization of the full-length protein found the macromolecule to migrate in a manner demonstrative of a protein with a significantly larger molecular weight than the one predicted for RalBP (Cantor et al., 1995; Park and Weinberg, 1995). Later work

observed this behavior together with protein migration indicative of a variety of lower molecular weights (Awasthi et al., 2002a; Awasthi et al., 2002b). Somewhat consistent with these findings, the smaller construct of RalBP used in this work also migrated in a manner inconsistent with the expected molecular weight of the protein construct (Figure 3). While no evidence had been presented to explain the observed behavior of RalBP, it was suggested that perhaps the observed larger and smaller molecular weights were the result of dimerization through the predicted coiled-coil at the C-terminus, or proteolytic processing, respectively (Awasthi et al., 2002a; Awasthi et al., 2002b; Jullien-Flores et al., 1995). The shorter constructs of RalBP in this work excluded most of the coiled-coil region (residues 440-610) and do not demonstrate the numerous smaller molecular weights as observed in previous work. The oligomerization observed in the Ral binding domain of RalBP was due to the two cysteines of this protein, as shown by the ability of DTT to disrupt the oligomers (Figure 5). An additional finding, which contributes to the unexpected behavior of RalBP, is the evidence for RalBP disorder and folding upon binding to RalA (Figure 10), which presents the first example of this behavior for a Ras family effector. The disordered behavior of RalBP may also enhance the oligomerization effect by allowing the two cysteine residues greater accessibility than would be found in the folded conformation of the protein. The intrinsic disorder of RalBP could make it more susceptible to proteolysis, which potentially explains the lower molecular weight variants of RalBP observed in other studies. Additionally, RalBP has been found to have a variety of functions and to interact with a number of different proteins; the folding of RalBP upon binding to a partner protein may be

important for these diverse functions and interactions. However, more work is necessary to verify these possibilities and to improve the samples for complete CD spectra.

Cysteine to serine mutations alleviated the oligomerization problem in the Ral binding domain of RalBP. Additionally, this double mutation of cysteines 411 and 451 did not appear to affect binding of RalA and RalBP as the double mutant of RalBP coeluted with RalA during size exclusion chromatography and the complex displayed a shift in Native PAGE, which is indicative of complex formation (Figure 8). The additional mutation made at residue 472 from a threonine to a methionine made the amino acid sequence consistent to wild-type RalBP in humans. A threonine residue is observed at position 472 in the rat RalBP. As with the cysteine to serine mutations, the threonine to methionine mutation did not appear to affect the formation of the complex between RalBP and RalA. This residue change was engineered before the mutation of the cysteine residues, so the results shown in Figure 8 also support formation of the complex with a methionine as residue 472.

Additionally, the change from threonine to methionine is irrelevant for the shorter RalBP(391-444) protein as it does not include residue 472.

While the three mutations of residues 411, 451, and 472 did not inhibit complex formation, the truncation to the published Ral binding domain of RalBP residues 403-499 (Jullien-Flores et al., 1995) did not form the complex with RalA. This was an unexpected result; however, the study identifying the Ral binding domain of RalBP used deletion studies and didn't specifically test residues 403-499 alone (Jullien-Flores et al., 1995). This region of RalBP is

predicted to be highly α -helical (Cantor et al., 1995; Jullien-Flores et al., 1995) and the truncation of RalBP to residues 403-499 potentially could have caused disruptions to the secondary structure of this region, which could interfere with binding. Also, it is possible that segments of RalBP outside of this region bind to RalA at secondary binding sites and improve the interaction, which could account for the previously observed complex formation in contrast to the results presented here. A second published Ral binding domain of RalBP spanned from residues 391-444 (Park and Weinberg, 1995), raising the suspicion that the 403 endpoint in the RalBP(403-499) was problematic. With this in mind, a number of RalBP constructs were designed using the N-terminal endpoints of residues 391 and 397, and the C-terminal endpoints of residues 444, 499, and 518 (Figure 12). Residues 397 and 518 were the limits of our starting RalBP construct and the other three endpoints were consistent with the Ral binding domain of RalBP as published in the literature (Jullien-Flores et al., 1995; Park and Weinberg, 1995). The RalBP construct consistent with the Park and Weinberg published Ral binding domain, RalBP(391-444), successfully bound to RalA (Figure 15). Interestingly, of all the RalBP constructs tested, the only one that did not bind RalA was RalBP(403-499).

As the ultimate goal of this work was to obtain a crystal structure of the RalA-RalBP complex, numerous crystallization trials were attempted with the variety of RalBP constructs. No diffraction-quality protein crystals were obtained until very recently when a truncated RalA(11-178) was used in conjunction with RalBP(391-444). While these preliminary crystals have not yet been reproduced, they provide a starting point for protein crystallization

experiments. Additionally, using the shorter RalA(11-178) construct is consistent with other successful RalA complex structures (Fukai et al., 2003; Pautsch et al., 2005) and uses a logical N-terminal truncation as the first 10-11 residues in RalA are observed to be disordered in crystal structures (Nicely et al., 2004). While more work is necessary to achieve a crystal structure, the use of RalA(11-178) and RalBP(391-444) is very promising for success.

RalA-Effector Interactions

RalA is over 50% homologous to Ras (Chardin and Tavittian, 1986) and shares the same overall structural fold (Nicely et al., 2004). However, a unique effector recognition sequence in switch I and a distinct set of effector proteins differentiate RalA from the rest of the Ras family. Two structures have been solved of RalA in complex with the Ral-binding domain of effector molecules: RalA-Sec5 (Fukai et al., 2003) and RalA-Exo84 (Jin et al., 2005). While the canonical Ras-binding domain adopts an Ubiquitin-like fold, neither Sec5 nor Exo84 do. Sec5 has an immunoglobulin-like β -sandwich structure and interacts with RalA primarily through the switch I region, forming an intermolecular β -sheet (Fukai et al., 2003). While the structure of Sec5 is different from Ras-binding domains, the interaction it makes with RalA is the same as is observed with Ras or Rap and Ras-binding domains (Fukai et al., 2003). Unlike Sec5, Exo84 adopts a pleckstrin homology domain fold and interacts with RalA through both switch regions (Jin et al., 2005). To further diversify the RalA-effector interactions, RalBP is predicted to be highly α -helical, particularly in the Ral-binding domain (Cantor et al., 1995; Jullien-Flores et al., 1995; Park and Weinberg, 1995), suggesting that

the structure of this complex will present a previously unobserved binding mode, not only for RalA, but also for Ras GTPases.

RalA hot spot identification through structural analysis

Analysis of the RalA-Sec5 structure revealed RalA residues that contribute to specificity of the interaction. Lys47 has been shown in mutational analysis to prevent binding of Ras binding domains (RBDs; Bauer et al., 1999). Superposition of the RalA-Sec5 structure onto a Ras-RBD structure revealed that the Lys47 side chain would sterically hinder a RBD from binding RalA. Sec5 avoids this clash by a kink between β strands 1 and 2 (Fukai et al., 2003). Additionally, Glu38 and Tyr36, which are conserved in Ral but not Ras, make important interactions with Sec5. When these residues were mutated in Ral, the interaction with Sec5 was disrupted (Fukai et al., 2003).

The RalA-Exo84 structure revealed a different Ral-effector binding mode, and also different RalA residues important for the specificity of the interaction. Lys47 and Ala48, belonging to switch I, and Ile78 and Asn81, belonging to switch II, serve as specificity determinants for the interaction (Jin et al., 2005). These residues are four of the so-called Ral tree determinant residues because they are conserved in Ral, but not other Ras GTPases. Ala48 and Ile78 stabilize the Tyr82 side chain, allowing it to make a water-mediated hydrogen bond and hydrophobic interactions with Exo84 (Jin et al., 2005). Asn81 forms two hydrogen bonds with Exo84 and contributes to the surface complementarity between RalA and Exo84 (Jin et

al., 2005). Finally, positively charged Lys47, which sterically prevents binding of RBDs, contributes to binding through a favorable charge interaction (Jin et al., 2005).

Adjacent to the switch regions are two pockets lined with the Ral tree determinant residues. These pockets have been identified as potential novel binding sites of RalA (Nicely et al., 2004). Considering the important contributions made by conserved RalA residues in both the RalA-Sec5 and RalA-Exo84 complexes, these seem likely sites for interaction. Additionally, comparison of these pockets with Ras and Rap reveals that they vary in size and charge distribution, suggesting these sites may contribute to binding specificity (Nicely et al., 2004).

The structural analyses of the Ral complexes with Sec5 and Exo84 have revealed a number of hot spot residues for protein binding. It is possible RalBP will interact through these residues, or even bind in one of the predicted binding sites. So far, mutational studies have shown that Ala48 is important for RalBP binding (Bauer et al., 1999). A structure of the complex will reveal how.

Conformational Change at the Ral-Effector Interface

Conformational change plays a central role in GTPase-effector interactions. GTPases cycle between an active GTP-bound state and an inactive GDP-bound state. A conversion in nucleotide binding is accompanied by a nucleotide-dependent conformational change in the switch regions (Vetter and Wittinghofer, 2001). As effector proteins recognize the active GTP-bound state of GTPases through the switch regions, it stands to reason that this state

must facilitate the selection of switch conformations complementary to effectors.

Comparison of the GDP-bound RalA structure to those of the RalA-GTP complexes with Sec5 and Exo84 reveals switch conformations that sterically prevent complex formation. For the RalA-Sec5 complex, the GDP switch I conformation causes Tyr43 to flip to a position pointing toward Asp49. This causes the side chain of Asp49 to be exposed to the solvent in a position that would clash with Sec5 (Fukai et al., 2003). For the RalA-Exo84 complex, both the switch regions of RalA-GDP would collide with Exo84. The positions of residues Tyr75, Ile78, Asn81, and Tyr82 of switch II and Asp49 of switch I would clash with Exo84 binding (Jin et al., 2005).

In the GTP-bound RalA structures further differences are observed in the switch regions. In the RalA-Sec5 complex, switch II does not interact with Sec5 and is disordered (Fukai et al., 2003). The switch I conformation is different from the one observed in the RalA-GppNHp structure (Nicely et al., 2004). In fact, the switch I conformation in the RalA-GppNHp structure would clash with Sec5 binding (Nicely et al., 2004). This alternate conformation of switch I suggests the “conformational selection” model of binding (Kumar et al., 2000) and may be preferred by RalBP or another effector. Comparison of the RalA-Sec5 complex with the RalA-Exo84 complex reveals a similar conformation of switch I (Jin et al., 2005). Additionally, where switch II is disordered in the complex with Sec5, it adopts a unique conformation with Glu73 and Tyr75 interacting directly with Exo84 (Jin et al., 2005).

The RalA-RalBP complex is likely to present a previously unobserved binding mode between a GTPase and its effector. Unlike the β -sandwich structures of Exo84 and Sec5, RalBP is predicted to have a high α -helix content. The Ral binding domain also overlaps with the putative coiled-coil region (Cantor et al., 1995; Jullien-Flores et al., 1995; Park and Weinberg, 1995). Additionally, my work has shown evidence that RalBP is disordered in solution and folds upon binding to RalA. Presumably, an α -helix of RalBP serves as a target for recognition for RalA-GTP and then RalBP folds onto the template of RalA. This is consistent with the process of binding for intrinsically disordered proteins proposed previously (Fuxreiter et al., 2004). The loss of binding caused by the truncation of RalBP to residues 403-499 might be related to this process. It is possible that the truncation, most likely the amino-terminal trimming to residue 403, cut in the middle of an α -helix. This cut may have disrupted this structural element and caused a partial or complete unwinding of the helix. The loss of this secondary structure could have prevented recognition and complex formation.

Summary

Much progress has been made towards the crystal structure of the RalA-RalBP complex, but more work is necessary to obtain the structure and to answer the questions about the specificity of this interaction. The work presented in this chapter shows the significant advances made to overcome the problems encountered when working with RalBP, particularly those of oligomerization and intrinsic disorder. While the disorder of RalBP presents challenges to crystallization, it also is an exciting result as it is the first example of a

Ras family effector that folds upon binding. Furthermore, identification of the RalA(11-178) and RalBP(391-444) constructs that result in a protein crystal puts this project in a position primed for success.

Acknowledgements

Larry Feig kindly donated the clones for full-length simian RalA and for the original GST-fusion with rat RalBP. Nate Nicely provided the purification protocol for RalA and valuable discussions about the project. Many thanks to Brett Feeney, Sara Milam, and Clay Clark for their help with CD data collection and analysis. Mike Goshe generously provided mass spectrometry analysis for RalBP. Winnell Newman was always available for discussions about cloning and helped with later cloning experiments, and Gerald Guanga offered invaluable suggestions for protein purification.

References

- Albright, C.F., Giddings, B.W., Liu, J., Vito, M., and Weinberg, R.A. (1993). Characterization of a guanine nucleotide dissociation stimulator for a ras-related GTPase. *EMBO J* 12, 339-347.
- Awasthi, S., Cheng, J., Singhal, S.S., Saini, M.K., Pandya, U., Pikula, S., Bandorowicz-Pikula, J., Singh, S.V., Zimniak, P., and Awasthi, Y.C. (2000). Novel function of human RLIP76: ATP-dependent transport of glutathione conjugates and doxorubicin. *Biochemistry* 39, 9327-9334.
- Awasthi, S., Sharma, R., Singhal, S.S., Zimniak, P., and Awasthi, Y.C. (2002a). RLIP76, a novel transporter catalyzing ATP-dependent efflux of xenobiotics. *Drug Metab Dispos* 30, 1300-1310.
- Awasthi, S., Sharma, R., Yang, Y., Singhal, S.S., Pikula, S., Bandorowicz-Pikula, J., Singh, S.V., Zimniak, P., and Awasthi, Y.C. (2002b). Transport functions and physiological

significance of 76 kDa Ral-binding GTPase activating protein (RLIP76). *Acta Biochim Pol* 49, 855-867.

Awasthi, S., Singhal, S.S., Sharma, R., Zimniak, P., and Awasthi, Y.C. (2003a). Transport of glutathione conjugates and chemotherapeutic drugs by RLIP76 (RALBP1): a novel link between G-protein and tyrosine kinase signaling and drug resistance. *Int J Cancer* 106, 635-646.

Awasthi, S., Singhal, S.S., Singhal, J., Cheng, J., Zimniak, P., and Awasthi, Y.C. (2003b). Role of RLIP76 in lung cancer doxorubicin resistance: II. Doxorubicin transport in lung cancer by RLIP76. *Int J Oncol* 22, 713-720.

Awasthi, S., Singhal, S.S., Singhal, J., Yang, Y., Zimniak, P., and Awasthi, Y.C. (2003c). Role of RLIP76 in lung cancer doxorubicin resistance: III. Anti-RLIP76 antibodies trigger apoptosis in lung cancer cells and synergistically increase doxorubicin cytotoxicity. *Int J Oncol* 22, 721-732.

Awasthi, Y.C., Sharma, R., Yadav, S., Dwivedi, S., Sharma, A., and Awasthi, S. (2007). The non-ABC drug transporter RLIP76 (RALBP-1) plays a major role in the mechanisms of drug resistance. *Curr Drug Metab* 8, 315-323.

Bauer, B., Mirey, G., Vetter, I.R., Garcia-Ranea, J.A., Valencia, A., Wittinghofer, A., Camonis, J.H., and Cool, R.H. (1999). Effector recognition by the small GTP-binding proteins Ras and Ral. *J Biol Chem* 274, 17763-17770.

Bhullar, R.P., Chardin, P., and Haslam, R.J. (1990). Identification of multiple ral gene products in human platelets that account for some but not all of the platelet Gn-proteins. *FEBS Lett* 260, 48-52.

Bhullar, R.P., and Yang, S. (1998). Immunodetection of ralA and ralB GTP-binding proteins in various rat tissues and platelets. *Mol Cell Biochem* 179, 49-55.

Bodemann, B.O., and White, M.A. (2008). Ral GTPases and cancer: linchpin support of the tumorigenic platform. *Nat Rev Cancer* 8, 133-140.

Bos, J.L. (1998). All in the family? New insights and questions regarding interconnectivity of Ras, Rap1 and Ral. *EMBO J* 17, 6776-6782.

Brymora, A., Valova, V.A., Larsen, M.R., Roufogalis, B.D., and Robinson, P.J. (2001). The brain exocyst complex interacts with RalA in a GTP-dependent manner: identification of a novel mammalian Sec3 gene and a second Sec15 gene. *J Biol Chem* 276, 29792-29797.

Cantor, S.B., Urano, T., and Feig, L.A. (1995). Identification and characterization of Ral-binding protein 1, a potential downstream target of Ral GTPases. *Mol Cell Biol* *15*, 4578-4584.

Chardin, P., and Tavittian, A. (1986). The ral gene: a new ras related gene isolated by the use of a synthetic probe. *EMBO J* *5*, 2203-2208.

Chen, X.W., Leto, D., Chiang, S.H., Wang, Q., and Saltiel, A.R. (2007). Activation of RalA is required for insulin-stimulated Glut4 trafficking to the plasma membrane via the exocyst and the motor protein Myo1c. *Dev Cell* *13*, 391-404.

Chien, Y., and White, M.A. (2003). RAL GTPases are linchpin modulators of human tumour-cell proliferation and survival. *EMBO Rep* *4*, 800-806.

Emkey, R., Freedman, S., and Feig, L.A. (1991). Characterization of a GTPase-activating protein for the Ras-related Ral protein. *J Biol Chem* *266*, 9703-9706.

Feig, L.A. (2003). Ral-GTPases: approaching their 15 minutes of fame. *Trends Cell Biol* *13*, 419-425.

Frankel, P., Aronheim, A., Kavanagh, E., Balda, M.S., Matter, K., Bunney, T.D., and Marshall, C.J. (2005). RalA interacts with ZONAB in a cell density-dependent manner and regulates its transcriptional activity. *EMBO J* *24*, 54-62.

Fukai, S., Matern, H.T., Jagath, J.R., Scheller, R.H., and Brunger, A.T. (2003). Structural basis of the interaction between RalA and Sec5, a subunit of the sec6/8 complex. *EMBO J* *22*, 3267-3278.

Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. (2004). Preformed Structural Elements Feature in Partner Recognition by Intrinsically Unstructured Proteins. *Journal of Molecular Biology*. *338*, 1015-1026.

Huang, L., Hofer, F., Martin, G.S., and Kim, S.H. (1998). Structural basis for the interaction of Ras with RalGDS. *Nat Struct Biol* *5*, 422-426.

Huang, L., Weng, X., Hofer, F., Martin, G.S., and Kim, S.H. (1997). Three-dimensional structure of the Ras-interacting domain of RalGDS. *Nat Struct Biol* *4*, 609-615.

Ikeda, M., Ishida, O., Hinoi, T., Kishida, S., and Kikuchi, A. (1998). Identification and characterization of a novel protein interacting with Ral-binding protein 1, a putative effector protein of Ral. *J Biol Chem* *273*, 814-821.

Jilkina, O., and Bhullar, R.P. (1996). Generation of antibodies specific for the RalA and RalB GTP-binding proteins and determination of their concentration and distribution in human platelets. *Biochim Biophys Acta* 1314, 157-166.

Jin, R., Junutula, J.R., Matern, H.T., Ervin, K.E., Scheller, R.H., and Brunger, A.T. (2005). Exo84 and Sec5 are competitive regulatory Sec6/8 effectors to the RalA GTPase. *EMBO Journal* 24, 2064-2074.

John, J., Sohmen, R., Feuerstein, J., Linke, R., Wittinghofer, A., and Goody, R.S. (1990). Kinetics of interaction of nucleotides with nucleotide-free H-ras p21. *Biochemistry* 29, 6058-6065.

Jullien-Flores, V., Dorseuil, O., Romero, F., Letourneur, F., Saragosti, S., Berger, R., Tavitian, A., Gacon, G., and Camonis, J.H. (1995). Bridging Ral GTPase to Rho pathways. RLIP76, a Ral effector with CDC42/Rac GTPase-activating protein activity. *J Biol Chem* 270, 22473-22477.

Jullien-Flores, V., Mahe, Y., Mirey, G., Leprince, C., Meunier-Bisceuil, B., Sorkin, A., and Camonis, J.H. (2000). RLIP76, an effector of the GTPase Ral, interacts with the AP2 complex: involvement of the Ral pathway in receptor endocytosis. *J Cell Sci* 113 (Pt 16), 2837-2844.

Kelly, S.M., and Price, N.C. (2000). The use of circular dichroism in the investigation of protein structure and function. *Curr Protein Pept Sci* 1, 349-384.

Kinsella, B.T., Erdman, R.A., and Maltese, W.A. (1991). Carboxyl-terminal isoprenylation of ras-related GTP-binding proteins encoded by rac1, rac2, and ralA. *J Biol Chem* 266, 9786-9794.

Kumar, S., Ma, B., Tsai, C.-J., Sinha, N., and Nussinov, R. (2000). Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Science* 9, 10-19.

Li, X., Romero, P., Rani, M., Dunker, A.K., and Obradovic, Z. (1999). Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform Ser Workshop Genome Inform* 10, 30-40.

Lim, K.H., Baines, A.T., Fiordalisi, J.J., Shipitsin, M., Feig, L.A., Cox, A.D., Der, C.J., and Counter, C.M. (2005). Activation of RalA is critical for Ras-induced tumorigenesis of human cells. *Cancer Cell* 7, 533-545.

Lim, K.H., O'Hayer, K., Adam, S.J., Kendall, S.D., Campbell, P.M., Der, C.J., and Counter, C.M. (2006). Divergent roles for RalA and RalB in malignant growth of human pancreatic carcinoma cells. *Curr Biol* 16, 2385-2394.

Lopez, J.A., Kwan, E.P., Xie, L., He, Y., James, D.E., and Gaisano, H.Y. (2008). The RalA GTPase is a central regulator of insulin exocytosis from pancreatic islet beta-cells. *J Biol Chem*.

Matsubara, K., Hinoi, T., Koyama, S., and Kikuchi, A. (1997). The post-translational modifications of Ral and Rac1 are important for the action of Ral-binding protein 1, a putative effector protein of Ral. *FEBS Lett* 410, 169-174.

Matsuzaki, T., Hanai, S., Kishi, H., Liu, Z., Bao, Y., Kikuchi, A., Tsuchida, K., and Sugino, H. (2002). Regulation of endocytosis of activin type II receptors by a novel PDZ protein through Ral/Ral-binding protein 1-dependent pathway. *J Biol Chem* 277, 19008-19018.

Moskalenko, S., Henry, D.O., Rosse, C., Mirey, G., Camonis, J.H., and White, M.A. (2002). The exocyst is a Ral effector complex. *Nat Cell Biol* 4, 66-72.

Mott, H.R., Nietlispach, D., Hopkins, L.J., Mirey, G., Camonis, J.H., and Owen, D. (2003). Structure of the GTPase-binding domain of Sec5 and elucidation of its Ral binding site. *J Biol Chem* 278, 17053-17059.

Nadkar, A., Pungaliya, C., Drake, K., Zajac, E., Singhal, S.S., and Awasthi, S. (2006). Therapeutic resistance in lung cancer. *Expert Opin Drug Metab Toxicol* 2, 753-777.

Nakashima, S., Morinaka, K., Koyama, S., Ikeda, M., Kishida, M., Okawa, K., Iwamatsu, A., Kishida, S., and Kikuchi, A. (1999). Small G protein Ral and its downstream molecules regulate endocytosis of EGF and insulin receptors. *EMBO J* 18, 3629-3642.

Nicely, N.I., Kosak, J., de Serrano, V., and Mattos, C. (2004). Crystal structures of Ral-GppNHp and Ral-GDP reveal two binding sites that are also present in Ras and Rap. *Structure* 12, 2025-2036.

Nicholas KB, Nicholas HB Jr, Deerfield DW II. 1997. "GeneDoc: Analysis and Visualization of Genetic Variation." *EMBNEW.NEWS* 4:14.<http://www.psc.edu/biomed/genedoc>

Ohta, Y., Suzuki, N., Nakamura, S., Hartwig, J.H., and Stossel, T.P. (1999). The small GTPase RalA targets filamin to induce filopodia. *Proc Natl Acad Sci U S A* 96, 2122-2128.

Panner, A., Nakamura, J.L., Parsa, A.T., Rodriguez-Viciana, P., Berger, M.S., Stokoe, D., and Pieper, R.O. (2006). mTOR-independent translational control of the extrinsic cell death pathway by RalA. *Mol Cell Biol* 26, 7345-7357.

Park, S.H., and Weinberg, R.A. (1995). A putative effector of Ral has homology to Rho/Rac GTPase activating proteins. *Oncogene* 11, 2349-2355.

Pautsch, A., Vogelsgesang, M., Trankle, J., Herrmann, C., and Aktories, K. (2005). Crystal structure of the C3bot-RalA complex reveals a novel type of action of a bacterial exoenzyme. *EMBO J* 24, 3670-3680.

Quaroni, A., and Paul, E.C. (1999). Cytocentrin is a Ral-binding protein involved in the assembly and function of the mitotic apparatus. *J Cell Sci* 112 (Pt 5), 707-718.

Reuther, G.W., and Der, C.J. (2000). The Ras branch of small GTPases: Ras family members don't fall far from the tree. *Curr Opin Cell Biol* 12, 157-165.

Romero, P., Obradovic, Z., and Dunker, A.K. (1997). Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Informatics* 8, 110-124.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Sequence complexity of disordered protein. *Proteins* 42, 38-48.

Romero P, Dunker AK, Li X, Obradovic Z. (2002) "PONDR®: Predictor Of Naturally Disordered Regions." Molecular Kinetics, Indianapolis, IN, USA. <http://www.pondr.com>

Rosse, C., L'Hoste, S., Offner, N., Picard, A., and Camonis, J. (2003). RLIP, an effector of the Ral GTPases, is a platform for Cdk1 to phosphorylate epsin during the switch off of endocytosis in mitosis. *J Biol Chem* 278, 30597-30604.

Sablina, A.A., Chen, W., Arroyo, J.D., Corral, L., Hector, M., Bulmer, S.E., DeCaprio, J.A., and Hahn, W.C. (2007). The tumor suppressor PP2A A β regulates the RalA GTPase. *Cell* 129, 969-982.

Sharma, R., Sharma, A., Yang, Y., Awasthi, S., Singhal, S.S., Zimniak, P., and Awasthi, Y.C. (2002). Functional reconstitution of Ral-binding GTPase activating protein, RLIP76, in proteoliposomes catalyzing ATP-dependent transport of glutathione conjugate of 4-hydroxynonenal. *Acta Biochim Pol* 49, 693-701.

Shipitsin, M., and Feig, L.A. (2004). RalA but not RalB enhances polarized delivery of membrane proteins to the basolateral surface of epithelial cells. *Mol Cell Biol* 24, 5746-5756.

Singhal, S.S., Singhal, J., Yadav, S., Dwivedi, S., Boor, P.J., Awasthi, Y.C., and Awasthi, S. (2007). Regression of lung and colon cancer xenografts by depleting or inhibiting RLIP76 (Ral-binding protein 1). *Cancer Res* 67, 4382-4389.

Smith, S.C., Oxford, G., Baras, A.S., Owens, C., Havaleshko, D., Brautigan, D.L., Safo, M.K., and Theodorescu, D. (2007). Expression of ral GTPases, their effectors, and activators in human bladder cancer. *Clin Cancer Res* 13, 3803-3813.

- Tchevkina, E., Agapova, L., Dyakova, N., Martinjuk, A., Komelkov, A., and Tatosyan, A. (2005). The small G-protein RalA stimulates metastasis of transformed cells. *Oncogene* *24*, 329-335.
- Thompson JD, Higgins DG, Gibson TJ. 1994. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice." *Nucleic Acids Research*. *22*:4673-4680.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. "The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." *Nucleic Acids Research*. *24*:4876-4882.
- Urano, T., Emkey, R., and Feig, L.A. (1996). Ral-GTPases mediate a distinct downstream signaling pathway from Ras that facilitates cellular transformation. *EMBO J* *15*, 810-816.
- Vetter, I.R., and Wittinghofer, A. (2001). The guanine nucleotide-binding switch in three dimensions. *Science* *294*, 1299-1304.
- Wildey, G.M., Viggesswarapu, M., Rim, S., and Denker, J.K. (1993). Isolation of cDNA clones and tissue expression of rat ral A and ral B GTP-binding proteins. *Biochem Biophys Res Commun* *194*, 552-559.
- Wolthuis, R.M., de Rooter, N.D., Cool, R.H., and Bos, J.L. (1997). Stimulation of gene induction and cell growth by the Ras effector Rlf. *EMBO J* *16*, 6748-6761.
- Wu, J.C., Chen, T.Y., Yu, C.T., Tsai, S.J., Hsu, J.M., Tang, M.J., Chou, C.K., Lin, W.J., Yuan, C.J., and Huang, C.Y. (2005). Identification of V23RalA-Ser194 as a critical mediator for Aurora-A-induced cellular motility and transformation by small pool expression screening. *J Biol Chem* *280*, 9013-9022.
- Xu, J., Zhou, Z., Zeng, L., Huang, Y., Zhao, W., Cheng, C., Xu, M., Xie, Y., and Mao, Y. (2001). Cloning, expression and characterization of a novel human REPS1 gene. *Biochim Biophys Acta* *1522*, 118-121.
- Yamaguchi, A., Urano, T., Goi, T., and Feig, L.A. (1997). An Eps homology (EH) domain protein that binds to the Ral-GTPase target, RalBP1. *J Biol Chem* *272*, 31230-31234.
- Yu, Y., and Feig, L.A. (2002). Involvement of R-Ras and Ral GTPases in estrogen-independent proliferation of breast cancer cells. *Oncogene* *21*, 7557-7568.

CONCLUDING REMARKS

Protein interactions have been the subject of intense research, as they are important to understanding protein function and the development of disease. The work in this dissertation has examined protein interactions in two different ways. First, the well-studied RNase A was used as a model to assess the strengths of the Multiple Solvent Crystal Structure Method as a tool for obtaining binding site properties. This work identified detailed information about the surface of the protein, highlighting trends observed in the available structures of RNase A in the Protein Data Bank. Second, the interaction of RalA and its effector protein RalBP was examined, revealing the flexibility of RalBP as a protein, and proposing a novel binding mode of a RalA-effector complex.

The MSCS method is a fast, flexible way to study binding sites experimentally when high diffracting crystals are available. As is illustrated with the model protein RNase A, MSCS provides detailed information about interaction hot spots, solvation, and plasticity of the protein. The clustering of the organic solvents identifies important interactions that are made consistently by inhibitor molecules bound in the active site. Additionally, a number of well-ordered water molecules are found in the same position in a majority of structures. These waters participate in different functional roles, such as stabilizing the active site, bridging the interaction between active site residues and inhibitors, and leaving upon inhibitor binding thus allowing for direct interactions between the functional groups of an inhibitor or probe molecule and active site residues. Finally, analysis of which active site residues remain invariant in response to the different solvent environments of MSCS compared to those with

more flexibility, highlights the residues important for initial recognition and those that adjust to improve binding to a substrate. All of this information helps in understanding the binding interaction and potentially could aid ligand design.

In spite of being very similar to Ras, RalA is proving to be a unique member of the family. Ras and its associated effectors interact in a conserved way, forming an intermolecular β -sheet through switch I of Ras and the ubiquitin-like fold common to the Ras binding domain. The two structures of RalA-effector complexes reveal that neither of these effector proteins adopt a ubiquitin-like fold. While Sec5 forms an intermolecular β -sheet similar to that seen in Ras-RBD structures, Exo84 interacts in a distinct manner through both switch I and switch II of RalA. To further diversify RalA effectors, RalBP is predicted to be highly α -helical and the structure of the complex, for which we have obtained preliminary crystals, is likely to present a previously unobserved binding mode between a GTPase and an effector. Additionally, it was found in this work that RalBP is intrinsically disordered in solution and folds upon binding to RalA, representing the first example of this behavior in a Ras GTPase effector.

APPENDICES

APPENDIX A: Perl Scripts

Included here are the Perl scripts written for the analysis of RNase A structures. Use of these programs with other structures may require code modification.

AllAtomSolventLigandSuper.pl

This program identifies atom overlaps found between atoms in organic solvents and atoms in non-organic solvents, which are generally bound inhibitor molecules for RNase A. Atoms from lines with a HETATM tag that overlap the position of another HETATM with a distance of less than 1.0Å are included in the list of results. Lines with an ATOM tag are ignored (this generally includes protein and water molecules). Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the structures must be superimposed. This program currently only recognizes the following solvents: ACN, DOX, DMF, DMS, HEZ, IOH, BIR, TBU, ETF, and TMA. For additional solvent recognition, the code will have to be edited. The final results will be written in a file called

“AllAtomSolvenLigandSuper.log”.

```
#!/usr/bin/perl
# *****
#
# AllAtomSolventLigandSuper.pl
# Author: Michelle Dechene
#
# Perl script to identify solvent superpositions with ligands
# in all PDB files in the current directory. Structures must be
# superimposed.
#
```

```

# Solvents must be identified by HETATM instead of ATOM
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ###   ##.###   ##.###   ##.###   #.##   ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "AllAtomSolvenLigandSuper.log";
select RESULTS;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for HETATMs
            if ($Line =~ m/^HETATM/){
                @Atom = split /\s+/, $Line;
                # SolventID is ResidueName + Chain ID + Residue #
                $SolventID = $Atom[3] . $Atom[4] . $Atom[5];
                $SolventHash{$SolventID}{@Atom[2]} = [@Atom];
            }
        }
        close PDBFILE;
    }
}

for $ResID1 ( sort keys %SolventHash ) {
    for $AtomName1 ( sort keys %{ $SolventHash{$ResID1} } ) {
        for $ResID2 ( sort keys %SolventHash ) {
            for $AtomName2 ( sort keys %{ $SolventHash{$ResID2} } ) {
                # Don't compare atoms from the same residue
                if ($ResID1 ne $ResID2){
                    @Atom1 = @{ $SolventHash{$ResID1}{$AtomName1} };
                    @Atom2 = @{ $SolventHash{$ResID2}{$AtomName2} };

                    # Calculate the distance between 2 atoms
                    $X1 = @Atom1[6];
                    $Y1 = @Atom1[7];
                    $Z1 = @Atom1[8];

                    $X2 = @Atom2[6];
                    $Y2 = @Atom2[7];
                    $Z2 = @Atom2[8];

                    $XDiff = $X2 - $X1;
                    $YDiff = $Y2 - $Y1;
                    $ZDiff = $Z2 - $Z1;
                    $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
                    $Distance = sqrt $SumOfSquares;

                    # If the atom centers are closer than 1.0 Angstroms, they are stored with their
                    # distances

```

```

if ($Distance <= 1.0) {
  if (
    ( ( ($ResID1 =~ m/^ACN/) || ($ResID1 =~ m/^DOX/) || ($ResID1 =~ m/^DMF/)
      || ($ResID1 =~ m/^DMS/) || ($ResID1 =~ m/^HEZ/) || ($ResID1 =~ m/^IOH/)
      || ($ResID1 =~ m/^BIR/) || ($ResID1 =~ m/^TBU/) || ($ResID1 =~ m/^ETF/)
      || ($ResID1 =~ m/^TMA/ ) )
      &&
      !( ($ResID2 =~ m/^ACN/) || ($ResID2 =~ m/DOX/) || ($ResID2 =~ m/^DMF/)
        || ($ResID2 =~ m/^DMS/) || ($ResID2 =~ m/^HEZ/) || ($ResID2 =~ m/^IOH/)
        || ($ResID2 =~ m/^BIR/) || ($ResID2 =~ m/^TBU/) || ($ResID2 =~ m/^ETF/)
        || ($ResID2 =~ m/^TMA/ ) ) )
    ||
    ( !( ($ResID1 =~ m/^ACN/) || ($ResID1 =~ m/^DOX/) || ($ResID1 =~ m/^DMF/)
      || ($ResID1 =~ m/^DMS/) || ($ResID1 =~ m/^HEZ/) || ($ResID1 =~ m/^IOH/)
      || ($ResID1 =~ m/^BIR/) || ($ResID1 =~ m/^TBU/) || ($ResID1 =~ m/^ETF/)
      || ($ResID1 =~ m/^TMA/ ) )
      &&
      ( ($ResID2 =~ m/^ACN/) || ($ResID2 =~ m/DOX/) || ($ResID2 =~ m/^DMF/)
        || ($ResID2 =~ m/^DMS/) || ($ResID2 =~ m/^HEZ/) || ($ResID2 =~ m/^IOH/)
        || ($ResID2 =~ m/^BIR/) || ($ResID2 =~ m/^TBU/) || ($ResID2 =~ m/^ETF/)
        || ($ResID2 =~ m/^TMA/ ) ) )
    ) {

    # AtomKeys store the keys to %SolventHash so that data can be accessed
    $Atom1Key = $ResID1 . ":" . $AtomName1;
    $Atom2Key = $ResID2 . ":" . $AtomName2;
    if ( ($ResID2 =~ m/^ACN/) || ($ResID2 =~ m/^DOX/) || ($ResID2 =~ m/^DMF/)
      || ($ResID2 =~ m/^DMS/) || ($ResID2 =~ m/^HEZ/) || ($ResID2 =~ m/^IOH/)
      || ($ResID2 =~ m/^BIR/) || ($ResID2 =~ m/^TBU/) || ($ResID2 =~ m/^ETF/)
      || ($ResID2 =~ m/^TMA/ ) ) {
      $TempKey = $Atom1Key;
      $Atom1Key = $Atom2Key;
      $Atom2Key = $TempKey;
    }
    $AtomOverlap{$Atom1Key}{$Atom2Key} = $Distance;
  }
}
}
}
}
}
}

for $AtomID1 ( sort keys %AtomOverlap ) {
  for $AtomID2 ( sort keys %{ $AtomOverlap{$AtomID1} } ) {
    printf "Distance between $AtomID1 and $AtomID2: %2.3f\n" ,
    $AtomOverlap{$AtomID1}{$AtomID2};
  }
}

close RESULTS;

```

AllAtomSolventSuperPos.pl

This program identifies atom overlaps found between atoms in organic solvents. Atoms from lines with at HETATM tag that overlap the position of another HETATM with a distance of less than 1.0Å are included on the list of results. Lines with an ATOM tag are ignored (this

generally includes protein and water molecules). Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the structures must be superimposed. This program currently only recognizes the following solvents: ACN, DOX, DMF, DMS, HEZ, IOH, BIR, TBU, ETF, and TMA. For additional solvent recognition, the code will have to be edited. The final results will be written in a file called “SolventSuperPos1.log”.

```
#!/usr/bin/perl

# *****
#
# AllAtomSolventSuperPos.pl
# Author: Michelle Dechene
#
# Perl script to identify solvent superpositions (with other solvents)
# in all PDB files in the current directory. Structures must be
# superimposed.
#
# Solvents must be identified by HETATM instead of ATOM
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ###   ##.###   ##.###   ##.###   #.##   ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "SolventSuperPos1.log";
select RESULTS;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
```

```

# Search for HETATMs
if ($Line =~ m/^HETATM/){
  @Atom = split /\s+/, $Line;
  # SolventID is ResidueName + Chain ID + Residue #
  $SolventID = $Atom[3] . $Atom[4] . $Atom[5];
  $SolventHash{$SolventID}{@Atom[2]} = [@Atom];
}
}
close PDBFILE;
}

for $ResID1 ( sort keys %SolventHash ) {
  for $AtomName1 ( sort keys %{ $SolventHash{$ResID1} } ) {
    for $ResID2 ( sort keys %SolventHash ) {
      for $AtomName2 ( sort keys %{ $SolventHash{$ResID2} } ) {
        # Don't compare atoms from the same residue
        if ($ResID1 ne $ResID2){
          @Atom1 = @{ $SolventHash{$ResID1}{$AtomName1} };
          @Atom2 = @{ $SolventHash{$ResID2}{$AtomName2} };

          # Calculate the distance between 2 atoms
          $X1 = @Atom1[6];
          $Y1 = @Atom1[7];
          $Z1 = @Atom1[8];

          $X2 = @Atom2[6];
          $Y2 = @Atom2[7];
          $Z2 = @Atom2[8];

          $XDiff = $X2 - $X1;
          $YDiff = $Y2 - $Y1;
          $ZDiff = $Z2 - $Z1;
          $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
          $Distance = sqrt $SumOfSquares;

          # If the atom centers are closer than 1.0 Angstroms, they are stored with their
          # distances
          if ($Distance <= 1.0) {
            if ( ($ResID1 =~ m/^ACN/) || ($ResID1 =~ m/^DOX/) || ($ResID1 =~ m/^DMF/)
              || ($ResID1 =~ m/^DMS/) || ($ResID1 =~ m/^HEZ/) || ($ResID1 =~ m/^IOH/)
              || ($ResID1 =~ m/^BIR/) || ($ResID1 =~ m/^TBU/) || ($ResID1 =~ m/^ETF/)
              || ($ResID1 =~ m/^TMA/)
              || ($ResID2 =~ m/^ACN/) || ($ResID2 =~ m/^DOX/) || ($ResID2 =~ m/^DMF/)
              || ($ResID2 =~ m/^DMS/) || ($ResID2 =~ m/^HEZ/) || ($ResID2 =~ m/^IOH/)
              || ($ResID2 =~ m/^BIR/) || ($ResID2 =~ m/^TBU/) || ($ResID2 =~ m/^ETF/)
              || ($ResID2 =~ m/^TMA/) ){
              # AtomKeys store the keys to %SolventHash so that data can be accessed
              $Atom1Key = $ResID1 . ":" . $AtomName1;
              $Atom2Key = $ResID2 . ":" . $AtomName2;
              @Sorted = sort ($Atom1Key, $Atom2Key);
              $AtomOverlap{$Sorted[0]}{$Sorted[1]} = $Distance;
            }
          }
        }
      }
    }
  }
}

for $AtomID1 ( sort keys %AtomOverlap ) {
  for $AtomID2 ( sort keys %{ $AtomOverlap{$AtomID1} } ) {
    printf "Distance between $AtomID1 and $AtomID2: %2.3f\n" ,
    $AtomOverlap{$AtomID1}{$AtomID2};
  }
}
}

```

```
close RESULTS;
```

AtomDist.pl

This program calculates the distance between the backbone nitrogens of Thr 45 and Phe 120 for each protein molecule in a file. Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations. Results will be written to a file called "AtomDist.log".

```
#!/usr/bin/perl

# *****
#
# Author: Michelle Dechene
#
# Perl script to calculate distances between two user-defined atoms
# in user defined PDB files. Currently is hardcoded for N45 and N120
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ###   ##.###   ##.###   ##.###   #.#   ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

chomp(@PDBFileNames = <STDIN>);
open RESULTS, ">", "AtomDist.log";
select RESULTS;
foreach $PDBFileName (@PDBFileNames) {
    print "\nPDB File: $PDBFileName\n";
    open PDBFILE, "<", $PDBFileName;
    # Read in PDB file into array
    chomp(@FileLines = <PDBFILE>);
    foreach $Line (@FileLines) {
        if ($Line =~ m/ATOM.....N.....45/){
            @N45 = split /\s+/, $Line;
        }
        if ($Line =~ m/ATOM.....N.....120/){
            @N120 = split /\s+/, $Line;
        }
        if (@N45 && @N120){
            # Calculate the distance between 2 points

```

```

    $X1 = @N45[6];
    $Y1 = @N45[7];
    $Z1 = @N45[8];

    $X2 = @N120[6];
    $Y2 = @N120[7];
    $Z2 = @N120[8];

    $XDiff = $X2 - $X1;
    $YDiff = $Y2 - $Y1;
    $ZDiff = $Z2 - $Z1;
    $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
    $Distance = sqrt $SumOfSquares;

    # Print the results
    print "\nChain @N45[4]\n";
    print "@N45[3] @N45[5] Atom @N45[2] X = $X1, Y = $Y1, Z = $Z1\n";
    print "@N120[3] @N120[5] Atom @N120[2] X = $X2, Y = $Y2, Z = $Z2\n";
    printf "Atom Distance: %2.3f\n", $Distance;
    # Clear the arrays holding atom information
    @N45 = ();
    @N120 = ();
}
}
print "\n-----\n";
close PDBFILE;
}
close RESULTS;

```

AtomDist1.pl

This program calculates the distance between the backbone nitrogens of Thr 45 and Phe 120 for each protein molecule in a file. Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

The user must manually input all the PDB filenames to be used for the calculations. An input file containing one filename per line may be created and specified when running the script with a command similar to “perl AtomDist1.pl <inputfile”’. Results will be written to a file called “AtomDist1.log”’.

```

#!/usr/bin/perl
# *****

```

```

#
# AtomDist1.pl
# Author: Michelle Dechene
#
# Perl script to calculate distances between two user-defined atoms
# in user defined PDB files. Currently is hardcoded for N45 and N120
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ###   ##.###   ##.###   ##.###   #.##   ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

# user input for filenames
print "Enter PDB filenames: ";
chomp(@PDBFileNames = <STDIN>);

# open output file
open RESULTS, ">", "AtomDist1.log";
select RESULTS;

foreach $PDBFileName (@PDBFileNames) {
    print "\nPDB File: $PDBFileName\n";
    open PDBFILE, "<", $PDBFileName;

    # Read in PDB file into array
    chomp(@FileLines = <PDBFILE>);

    foreach $Line (@FileLines) {
        # Search for specified atoms
        if ($Line =~ m/ATOM.....N.....45/){
            @N45 = split /\s+/, $Line;
        }
        if ($Line =~ m/ATOM.....N.....120/){
            @N120 = split /\s+/, $Line;
        }
        if (@N45 && @N120){
            # Calculate the distance between 2 points
            $X1 = @N45[6];
            $Y1 = @N45[7];
            $Z1 = @N45[8];

            $X2 = @N120[6];
            $Y2 = @N120[7];
            $Z2 = @N120[8];

            $XDiff = $X2 - $X1;
            $YDiff = $Y2 - $Y1;
            $ZDiff = $Z2 - $Z1;
            $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
            $Distance = sqrt $SumOfSquares;

            # Print the results
            print "\nChain @N45[4]\n";
            print "@N45[3] @N45[5] Atom @N45[2] X = $X1, Y = $Y1, Z = $Z1\n";
            print "@N120[3] @N120[5] Atom @N120[2] X = $X2, Y = $Y2, Z = $Z2\n";
            printf "Atom Distance (Angstroms): %2.3f\n", $Distance;

            # Clear the arrays holding atom information
            @N45 = ();
            @N120 = ();
        }
    }
}

```

```

    print "\n-----\n";
    close PDBFILE;
}
close RESULTS;

```

AtomDist2.pl

This program calculates the distance between two user-defined atoms for each protein molecule in a file. Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

The user must manually input the atom name and residue number of the two atoms between which the distance will be calculated and all the PDB filenames to be used for the calculations. An input file containing the atom name and residue number of the first atom on the first two lines, and the atom name and residue number of the second atom on the next two lines, followed by one filename per line may be created and specified when running the script with a command similar to “perl AtomDist2.pl <inputfile”.

Results will be written to a file called “AtomDist2.log”.

```

#!/usr/bin/perl

# *****
#
# AtomDist2.pl
# Author: Michelle Dechene
#
# Perl script to calculate distances between two user-defined atoms
# in user defined PDB files.
#
# There is no error checking for Atom types, etc
#
# This program works best with two atoms in the same chain.
#
# Assumes the following PDB format
# ATOM   ###  xxx xxx x  ###      ##.###  ##.###  ##.###  #.##  ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number

```

```

#
# *****

# user input for atom type, residue number
print "Atom 1 Atom type: ";
chomp($AtomType1 = <STDIN>);
print "Atom 1 Residue number: ";
chomp($ResidueNum1 = <STDIN>);
print "Atom 2 Atom type: ";
chomp($AtomType2 = <STDIN>);
print "Atom 2 Residue number: ";
chomp($ResidueNum2 = <STDIN>);

# user input for filenames
print "Enter PDB filenames: ";
chomp(@PDBFileNames = <STDIN>);

# open output file and echo user input
open RESULTS, ">", "AtomDist2.log";
select RESULTS;
print "Atom 1 = $AtomType1 for residue $ResidueNum1 \n";
print "Atom 2 = $AtomType2 for residue $ResidueNum2 \n";

foreach $PDBFileName (@PDBFileNames) {
    print "\nPDB File: $PDBFileName\n";
    open PDBFILE, "<", $PDBFileName;

    # Read in PDB file into array
    chomp(@FileLines = <PDBFILE>);

    foreach $Line (@FileLines) {
        # Search for specified atoms
        if ($Line =~ m/\b$AtomType1\b .* \b$ResidueNum1\s/){
            @Atom1 = split /\s+/, $Line;
        }
        if ($Line =~ m/\b$AtomType2\b .* \b$ResidueNum2\s/){
            @Atom2 = split /\s+/, $Line;
        }
        if (@Atom1 && @Atom2){
            # Calculate the distance between 2 atoms
            $X1 = @Atom1[6];
            $Y1 = @Atom1[7];
            $Z1 = @Atom1[8];

            $X2 = @Atom2[6];
            $Y2 = @Atom2[7];
            $Z2 = @Atom2[8];

            $XDiff = $X2 - $X1;
            $YDiff = $Y2 - $Y1;
            $ZDiff = $Z2 - $Z1;
            $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
            $Distance = sqrt $SumOfSquares;

            # Print the results
            print "\nChain @Atom1[4]\n";
            print "@Atom1[3] @Atom1[5] Atom @Atom1[2] X = $X1, Y = $Y1, Z = $Z1\n";
            print "@Atom2[3] @Atom2[5] Atom @Atom2[2] X = $X2, Y = $Y2, Z = $Z2\n";
            printf "Atom Distance (Angstroms): %2.3f\n", $Distance;

            # Clear the arrays holding atom information
            @Atom1 = ();
            @Atom2 = ();
        }
    }
}
}

```

```

        print "\n-----\n";
        close PDBFILE;
    }
close RESULTS;

```

AtomDist3.pl

This program calculates the distance between the backbone nitrogens of Thr 45 and Phe 120 for each protein molecule in a file. Coordinates from the file are read into an array based on the line formatting of the PDB file. Because of this, the presence or absence of a chain name in the PDB file makes no difference to the functionality of the script.

All PDB files in the current folder will be used for the calculations. Results will be written to a file called "AtomDist.log".

```

#!/usr/bin/perl

# *****
#
# Author: Michelle Dechene
#
# Perl script to calculate distances between N45 and N120 in all PDB
# files in the current folder.
#
# There is no error checking for Atom types, etc
#
# *****
# Get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# Open output file and echo user input
open RESULTS, ">", "AtomDist.log";
select RESULTS;

# Read in protein atom coordinates, specifically: CA
foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File input: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);
        foreach $Line (@FileLines) {
            if ($Line =~ m/ATOM.....N.....45/){
                @N45[0] = substr($Line, 13, 3); #Atom Name
            }
        }
    }
}

```

```

@N45[1] = substr($Line, 17, 3); #Res Name
@N45[2] = substr($Line, 21, 1); #Chain
@N45[3] = substr($Line, 23, 3); #Res Number
@N45[4] = substr($Line, 30, 8); #X
@N45[5] = substr($Line, 38, 8); #Y
@N45[6] = substr($Line, 46, 8); #Z
}
if ($Line =~ m/ATOM.....N.....120/){
  @N120[0] = substr($Line, 13, 3); #Atom Name
  @N120[1] = substr($Line, 17, 3); #Res Name
  @N120[2] = substr($Line, 21, 1); #Chain
  @N120[3] = substr($Line, 23, 3); #Res Number
  @N120[4] = substr($Line, 30, 8); #X
  @N120[5] = substr($Line, 38, 8); #Y
  @N120[6] = substr($Line, 46, 8); #Z
}
}
if (@N45 && @N120){
  # Calculate the distance between 2 points
  $X1 = @N45[4];
  $Y1 = @N45[5];
  $Z1 = @N45[6];

  $X2 = @N120[4];
  $Y2 = @N120[5];
  $Z2 = @N120[6];

  $XDiff = $X2 - $X1;
  $YDiff = $Y2 - $Y1;
  $ZDiff = $Z2 - $Z1;
  $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
  $Distance = sqrt $SumOfSquares;

  # Print the results
  print "\nChain @N45[2]\n";
  print "@N45[1] @N45[3] Atom @N45[0] X = $X1, Y = $Y1, Z = $Z1\n";
  print "@N120[1] @N120[3] Atom @N120[0] X = $X2, Y = $Y2, Z = $Z2\n";
  printf "Atom Distance: %2.3f\n", $Distance;
  # Clear the arrays holding atom information
  @N45 = ();
  @N120 = ();
}
}
print "\n-----\n";
close PDBFILE;
}
close RESULTS;

```

AtomList.pl

This program identifies the atom names in a set of PDB files that do not begin with an N, C, O, S, P, G, V, or F. This was used to aid in the writing of other scripts for analysis. Data from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein, however

this script does not use any data found in a line after the chain, so the presence or absence of a chain should not make a difference of the functionality of this program.

All PDB files in the current folder will be examined.

```
#!/usr/bin/perl

# *****
#
# AtomList.pl
# Author: Michelle Dechene
#
# Perl script to identify atom types without a leading N, C, O, S, or P
# in the name
#
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ###   ##.###   ##.###   ##.###   #.##   ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

opendir MYDIR, ".";
@contents = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "AtomList.log";
select RESULTS;

foreach $PDBFileName (@contents) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Input atom information if a HETATM
            if ($Line =~ m/^HETATM/){
                @Atom = split /\s+/, $Line;
                if ( !($Atom[2] =~ m/^C/) && !($Atom[2] =~ m/^O/) && !($Atom[2] =~ m/^N/)
                    && !($Atom[2] =~ m/^S/) && !($Atom[2] =~ m/^P/) && !($Atom[2] =~ m/^G/)
                    && !($Atom[2] =~ m/^V/) && !($Atom[2] =~ m/^F/) ){
                    print "$Atom[2]\n";
                }
            }
        }
        close PDBFILE;
    }
}
}
```

BFactor.pl

This program calculates the average B-factor of each amino acid across a collection of structures. Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the script assumes there is one protein molecule per file. Solvent and ligand atoms must be designated as HETATM, or they will be included in the calculations corresponding to their respective residue number. The final results of these calculations are written in a tab formatted file called "BFactor.log" and each intermediate calculation documented in a file named "Debug.log" for monitoring and error-checking purposes.

```
#!/usr/bin/perl
# *****
#
# BFactor.pl
# Author: Michelle Dechene
#
# Perl script to calculate the average B-factor of each amino acid
# across a collection of structures (of the same superimposed protein)
#
# Solvents must be identified by HETATM instead of ATOM
#
# Assumes only one protein chain in each PDB file
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM    ###  xxx xxx x  ###      ##.###  ##.###  ##.###  #.#  ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "BFactor.log";
```

```

open LOG, ">", "Debug.log";
select LOG;

# read in atom coordinates, for just protein atoms...
foreach $PDBFileName (@files) {
  if ($PDBFileName =~ m/pdb$/){
    print "\nPDB File input: $PDBFileName\n";
    open PDBFILE, "<", $PDBFileName;

    # Read in PDB file into array
    chomp(@FileLines = <PDBFILE>);

    foreach $Line (@FileLines) {
      # Search for HETATMs
      if ($Line =~ m/^ATOM/){
        @Atom = split /\s+/, $Line;
        if ($Atom[3] ne "HOH") {
          # $AtomHash { Residue Number } { PDB File } { Atom Type } = [ line from PDB
          # file ]
          $AtomHash{$Atom[5]}{$PDBFileName}{$Atom[2]} = [@Atom];
        }
      }
    }
    close PDBFILE;
  }
}

$atomcounter = 0;
$ResidueSumOfSquares = 0;
$highRMSD = -1;
$lowRMSD = -1;
$residuecounter = 0;
$RMSDSum = 0;

# calculate RMSD across each residue for each pair of structures
for $ResNum ( sort keys %AtomHash ) {
  for $PDBFile ( sort keys %{ $AtomHash{$ResNum} } ) {
    for $AtomType (sort keys %{ $AtomHash{$ResNum}{$PDBFile} } ) {
      @Atom1 = @{$AtomHash{$ResNum}{$PDBFile}{$AtomType} };

      # get B-factor
      $BFactor = @Atom1[10];

      # put B-factor and a count of values in an array
      $ResidueBFactor[$ResNum][0] = $ResidueBFactor[$ResNum][0] + $BFactor;
      $ResidueBFactor[$ResNum][1] = $ResidueBFactor[$ResNum][1] + 1;

      print "Res: $ResNum, BFactor Sum: $ResidueBFactor[$ResNum][0] Count:
        $ResidueBFactor[$ResNum][1]\n";
    }
  }
}

for $ResNum ( sort keys %AtomHash ) {
  if ($ResidueBFactor[$ResNum][1] != 0){
    $AverageBFactor[$ResNum] = $ResidueBFactor[$ResNum][0] / $ResidueBFactor[$ResNum][1];
  }
}
close LOG;

select RESULTS;
# output an excel formatted file with residue number and B-Factor
$i = 1;
print "Residue\tB-Factor\n";

```

```

while ($i <= $#AverageBFactor){
    printf "%i\t%2.6f\n", $AverageBFactor[$i];
    $i += 1;
}

# close files
close RESULTS;

```

HETATM.pl

This program edits PDB files and changes ATOM to HETATM for all atoms that do not belong to a protein or water molecule. Changed files are saved in a new directory called “Hetatm_PDB” and filenames have a prefix of “Edit_”.

All PDB files in the current folder will be modified, if necessary, and saved in the new directory. This program assumes that the residue name for water molecules is called “HOH” in the PDB file.

```

#!/usr/bin/perl

# *****
#
# HETATM.pl
# Author: Michelle Dechene
#
# Perl script to change non-water and non-protein atoms to HETATM
#
# This script works on all PDB files in the current directory.
#
# Old files are left alone and new files (with HETATM changes) are
# written in a new directory called Hetatm_PDB
#
# *****

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

mkdir "Hetatm_PDB", 0755;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        open PDBFILE, "<", $PDBFileName;
        open PDBWRITE, ">", "Hetatm_PDB/Edit_" . $PDBFileName;
        select PDBWRITE;

        # Read in PDB file into array

```

```

chomp(@FileLines = <PDBFILE>);

foreach $Line (@FileLines) {
    $_=$Line;
    # Locate lines of non-protein and non-water atoms
    if (($Line =~ m/^ATOM/) && (substr($Line,17,3) ne "ALA") && (substr($Line,17,3) ne
        "CYS") && (substr($Line,17,3) ne "ASP") && (substr($Line,17,3) ne
        "GLU") && (substr($Line,17,3) ne "PHE") && (substr($Line,17,3) ne
        "GLY") && (substr($Line,17,3) ne "HIS") && (substr($Line,17,3) ne
        "ILE") && (substr($Line,17,3) ne "LYS") && (substr($Line,17,3) ne
        "LEU") && (substr($Line,17,3) ne "MET") && (substr($Line,17,3) ne
        "ASN") && (substr($Line,17,3) ne "PRO") && (substr($Line,17,3) ne
        "GLN") && (substr($Line,17,3) ne "ARG") && (substr($Line,17,3) ne
        "SER") && (substr($Line,17,3) ne "THR") && (substr($Line,17,3) ne
        "VAL") && (substr($Line,17,3) ne "TRP") && (substr($Line,17,3) ne
        "TYR") && (substr($Line,17,3) ne "HOH")) {
        s/ATOM /HETATM/;
    }
    print "$_\n";
}
close PDBFILE;
close PDBWRITE;
}
}

```

HingeAngle.pl

This program finds the centers of mass for Domain A (residues 1-13, 49-79, 105-124), Domain B (residues 16-46, 82-101), and the Hinge region (residues 14, 15, 47, 48, 80, 81, 102-104) of RNase A, using the positions of the C α atoms and weighting them with the mass of the entire amino acid residue. The hinge angle is defined as the angle between the two vectors connecting the hinge center of mass with the domain A center of mass, and the hinge center and the domain B center.

All PDB files in the current folder will be used for the calculations and the script assumes one protein molecule per file. Results are written in a file called "HingeAngles.log".

```

#!/usr/bin/perl
# *****
#
# Author: Michelle Dechene
# HingeAngle.pl
#

```

```

# Perl script to calculate the hinge angle between Domain A and
# Domain B in RNase A using weighed CA atoms.
#
# There is no error checking for Atom types, etc
#
# Assumes a single protein molecule per file
#
# *****

# Get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# Open output file and echo user input
open RESULTS, ">", "HingeAngles.log";
select RESULTS;

# Read in protein atom coordinates, specifically: CA
foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File input: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Initialize variables for center of mass calculation
        $XDomainA = 0;
        $YDomainA = 0;
        $ZDomainA = 0;
        $MassDomainA = 0;
        $XDomainB = 0;
        $YDomainB = 0;
        $ZDomainB = 0;
        $MassDomainB = 0;
        $XHinge = 0;
        $YHinge = 0;
        $ZHinge = 0;
        $MassHinge = 0;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for CAs
            if (($Line =~ m/^ATOM/) && (index($Line, "CA") == 13)){
                # Assign residue mass
                $ResN = substr($Line, 17, 3);
                if ($ResN =~ m/^GLY/){ $ResMass = 57.05; }
                elsif ($ResN =~ m/^ALA/){ $ResMass = 71.08; }
                elsif ($ResN =~ m/^SER/){ $ResMass = 87.08; }
                elsif ($ResN =~ m/^PRO/){ $ResMass = 97.12; }
                elsif ($ResN =~ m/^VAL/){ $ResMass = 99.13; }
                elsif ($ResN =~ m/^THR/){ $ResMass = 101.10; }
                elsif ($ResN =~ m/^CYS/){ $ResMass = 103.14; }
                elsif (($ResN =~ m/^LEU/) || ($ResN =~ m/^ILE/)){ $ResMass = 113.16; }
                elsif ($ResN =~ m/^ASN/){ $ResMass = 114.10; }
                elsif ($ResN =~ m/^ASP/){ $ResMass = 115.09; }
                elsif ($ResN =~ m/^GLN/){ $ResMass = 128.13; }
                elsif ($ResN =~ m/^LYS/){ $ResMass = 128.17; }
                elsif ($ResN =~ m/^GLU/){ $ResMass = 129.11; }
                elsif ($ResN =~ m/^MET/){ $ResMass = 131.20; }
                elsif ($ResN =~ m/^HIS/){ $ResMass = 137.14; }
                elsif ($ResN =~ m/^PHE/){ $ResMass = 147.18; }
                elsif ($ResN =~ m/^ARG/){ $ResMass = 156.19; }
                elsif ($ResN =~ m/^TYR/){ $ResMass = 163.18; }
                elsif ($ResN =~ m/^TRP/){ $ResMass = 186.21; }
            }
        }
    }
}

```

```

# Get coordinates and weight them
$WeightedX = substr($Line, 30, 8) * $ResMass;
$WeightedY = substr($Line, 38, 8) * $ResMass;
$WeightedZ = substr($Line, 46, 8) * $ResMass;

$ResNum = substr($Line, 23, 3);
# Domain A Summation
if ((($ResNum >= 1) && ($ResNum <= 13)) || (($ResNum >= 49) && ($ResNum <= 79)) ||
    (($ResNum >= 105) && ($ResNum <= 124))){
    $XDomainA += $WeightedX;
    $YDomainA += $WeightedY;
    $ZDomainA += $WeightedZ;
    $MassDomainA += $ResMass;
}

# Domain B Summation
elseif ((($ResNum >= 16) && ($ResNum <= 46)) || (($ResNum >= 82) && ($ResNum <=
    101))){
    $XDomainB += $WeightedX;
    $YDomainB += $WeightedY;
    $ZDomainB += $WeightedZ;
    $MassDomainB += $ResMass;
}

# Hinge Summation
elseif (($ResNum == 14) || ($ResNum == 15) || ($ResNum == 47) || ($ResNum == 48) ||
    ($ResNum == 80) || ($ResNum == 81) || (($ResNum >= 102) && ($ResNum <= 104))){
    $XHinge += $WeightedX;
    $YHinge += $WeightedY;
    $ZHinge += $WeightedZ;
    $MassHinge += $ResMass;
}
}
}
close PDBFILE;

# Find the center of mass for Domain A
$XCenterA = $XDomainA / $MassDomainA;
$YCenterA = $YDomainA / $MassDomainA;
$ZCenterA = $ZDomainA / $MassDomainA;
print "Domain A Center: $XCenterA, $YCenterA, $ZCenterA\n";

# Find the center of mass for Domain B
$XCenterB = $XDomainB / $MassDomainB;
$YCenterB = $YDomainB / $MassDomainB;
$ZCenterB = $ZDomainB / $MassDomainB;
print "Domain B Center: $XCenterB, $YCenterB, $ZCenterB\n";

# Find the center of mass for Hinge
$XCenterHinge = $XHinge / $MassHinge;
$YCenterHinge = $YHinge / $MassHinge;
$ZCenterHinge = $ZHinge / $MassHinge;
print "Hinge Center: $XCenterHinge, $YCenterHinge, $ZCenterHinge\n";

# Calculate the Hinge Angle from the three centers of mass
# Calculate the values of the vectors
$XVectorA = $XCenterA - $XCenterHinge;
$YVectorA = $YCenterA - $YCenterHinge;
$ZVectorA = $ZCenterA - $ZCenterHinge;

$XVectorB = $XCenterB - $XCenterHinge;
$YVectorB = $YCenterB - $YCenterHinge;
$ZVectorB = $ZCenterB - $ZCenterHinge;

# dot product between vector A and vector B
$DotProduct = (($XVectorA * $XVectorB) + ($YVectorA * $YVectorB) + ($ZVectorA *

```

```

    $ZVectorB));

# take the norm (or length) of vector A and vector B
$SumofSquaresVectorA = $XVectorA**2 + $YVectorA**2 + $ZVectorA**2;
$SumofSquaresVectorB = $XVectorB**2 + $YVectorB**2 + $ZVectorB**2;

$NormVectorA = sqrt $SumofSquaresVectorA;
$NormVectorB = sqrt $SumofSquaresVectorB;

# cos(theta) = ( A dot B ) / ((norm A) (normB))
$CosTheta = $DotProduct / ($NormVectorA * $NormVectorB);

# find theta
use Math::Trig;

$Theta = acos($CosTheta);
$Thetadegrees = rad2deg($Theta);

print "Theta = $Thetadegrees\n";
}
}

```

HingeAngleBB.pl

This program finds the centers of mass for Domain A (residues 1-13, 49-79, 105-124), Domain B (residues 16-46, 82-101), and the Hinge region (residues 14, 15, 47, 48, 80, 81, 102-104) of RNase A, using the positions of the C α atoms. The hinge angle is defined as the angle between the two vectors connecting the hinge center of mass with the domain A center of mass, and the hinge center and the domain B center.

All PDB files in the current folder will be used for the calculations and the script assumes one protein molecule per file. Results are written in a file called BBHingeAngles.log

```

#!/usr/bin/perl

# *****
#
# Author: Michelle Dechene
# HingeAngle.pl
#
# Perl script to calculate the hinge angle between Domain A and
# Domain B in RNase A using CA atoms.
#
# By leaving the weights out of the calculations, essentially only
# the backbone is considered and the side chains are ignored.
#

```

```

# There is no error checking for Atom types, etc
#
# Assumes a single protein molecule per file
#
# *****

# Get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# Open output file and echo user input
open RESULTS, ">", "BBHingeAngles.log";
select RESULTS;

# Read in protein atom coordinates, specifically: CA
foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File input: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Initialize variables for center of mass calculation
        $XDomainA = 0;
        $YDomainA = 0;
        $ZDomainA = 0;
        $MassDomainA = 0;
        $XDomainB = 0;
        $YDomainB = 0;
        $ZDomainB = 0;
        $MassDomainB = 0;
        $XHinge = 0;
        $YHinge = 0;
        $ZHinge = 0;
        $MassHinge = 0;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for CAs
            if (($Line =~ m/^ATOM/) && (index($Line, "CA") == 13)){

                # Get coordinates, no weights
                $WeightedX = substr($Line, 30, 8);
                $WeightedY = substr($Line, 38, 8);
                $WeightedZ = substr($Line, 46, 8);

                $ResNum = substr($Line, 23, 3);
                # Domain A Summation, MassDomainA is simply a number of residues
                if (((($ResNum >= 1) && ($ResNum <= 13)) || (($ResNum >= 49) && ($ResNum <= 79)) ||
                    (($ResNum >= 105) && ($ResNum <= 124)))){
                    $XDomainA += $WeightedX;
                    $YDomainA += $WeightedY;
                    $ZDomainA += $WeightedZ;
                    $MassDomainA ++;
                }

                # Domain B Summation, MassDomainB is simply a number of residues
                elsif (((($ResNum >= 16) && ($ResNum <= 46)) || (($ResNum >= 82) && ($ResNum <=
                    101)))){
                    $XDomainB += $WeightedX;
                    $YDomainB += $WeightedY;
                    $ZDomainB += $WeightedZ;
                    $MassDomainB ++;
                }
            }
        }
    }
}

```

```

# Hinge Summation, MassHinge is simply a number of residues
elseif (($ResNum == 14) || ($ResNum == 15) || ($ResNum == 47) || ($ResNum == 48) ||
($ResNum == 80) || ($ResNum == 81) || (($ResNum >= 102) && ($ResNum <= 104))){
    $XHinge += $WeightedX;
    $YHinge += $WeightedY;
    $ZHinge += $WeightedZ;
    $MassHinge ++;
}
}
}
close PDBFILE;

# Find the center of mass for Domain A
$XCenterA = $XDomainA / $MassDomainA;
$YCenterA = $YDomainA / $MassDomainA;
$ZCenterA = $ZDomainA / $MassDomainA;
print "Domain A Center: $XCenterA, $YCenterA, $ZCenterA\n";

# Find the center of mass for Domain B
$XCenterB = $XDomainB / $MassDomainB;
$YCenterB = $YDomainB / $MassDomainB;
$ZCenterB = $ZDomainB / $MassDomainB;
print "Domain B Center: $XCenterB, $YCenterB, $ZCenterB\n";

# Find the center of mass for Hinge
$XCenterHinge = $XHinge / $MassHinge;
$YCenterHinge = $YHinge / $MassHinge;
$ZCenterHinge = $ZHinge / $MassHinge;
print "Hinge Center: $XCenterHinge, $YCenterHinge, $ZCenterHinge\n";

# Calculate the Hinge Angle from the three centers of mass
# Calculate the values of the vectors
$XVectorA = $XCenterA - $XCenterHinge;
$YVectorA = $YCenterA - $YCenterHinge;
$ZVectorA = $ZCenterA - $ZCenterHinge;

$XVectorB = $XCenterB - $XCenterHinge;
$YVectorB = $YCenterB - $YCenterHinge;
$ZVectorB = $ZCenterB - $ZCenterHinge;

# dot product between vector A and vector B
$DotProduct = (($XVectorA * $XVectorB) + ($YVectorA * $YVectorB) + ($ZVectorA *
$ZVectorB));

# take the norm (or length) of vector A and vector B
$SumofSquaresVectorA = $XVectorA**2 + $YVectorA**2 + $ZVectorA**2;
$SumofSquaresVectorB = $XVectorB**2 + $YVectorB**2 + $ZVectorB**2;

$NormVectorA = sqrt $SumofSquaresVectorA;
$NormVectorB = sqrt $SumofSquaresVectorB;

# cos(theta) = ( A dot B ) / ((norm A) (normB))
$CosTheta = $DotProduct / ($NormVectorA * $NormVectorB);

# find theta
use Math::Trig;

$Theta = acos($CosTheta);
print "Theta Radians: $Theta\n";
$Thetadegrees = rad2deg($Theta);

print "Theta = $Thetadegrees\n";
}
}
}

```

LSQMAN.pl

This program opens LSQMAN (of the DEJAVU software package from the Uppsala Software Factory) to superimpose all PDB files in the current folder. LSQMAN is run and the commands described in the comment box of the code are echoed to produce a superposition based on all atoms of residues 1-124 of chain A in both files. This was written specifically for RNase A, so the commands will mostly likely need to be modified for use with other structures.

All PDB files in the current directory will be superimposed on the reference structure defined in the code, in this case, a file named “xlinkA.pdb”. The code naming the reference structure will need to be edited as necessary. To capture a log of each superposition made by LSQMAN (DEJAVU), type “perl LSQMAN.pl > LSQ.log” at the prompt (as described in the comment box of the code). Superimposed structures are saved in a new directory named “sup”. The reference structure is not superimposed nor is it saved in the “sup” directory.

```
#!/usr/bin/perl

# *****
#
# LSQMAN.pl
# Author: Michelle Dechene
#
# Perl script to open lsqman and superimpose all pdb files in the
# current folder
#
# Superimposed files are written in a new folder: sup
#
# lsqman path: /usr/local/dejavu/lx_lsqman
# Commands called in lsqman:
# re m1 <Reference Structure Filename>
# re m2 <Moving Structure Filename>
# at ex
# ex m1
# a1-124
# m2
# a1-124
# apply m1 m2
# wr m2 sup/<Moving Structure Filename>
# quit
```

```

#
# To run, type at prompt: perl LSQMAN.pl > LSQ.log
# where LSQ.log can be named anything and is the logfile storing
# all the lsqman results that would normally be displayed on the
# screen
#
# *****

# The reference structure: All structures will be superimposed onto $BaseStruct
$BaseStruct = 'xlinkA.pdb';

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

mkdir "sup", 0755;

foreach $PDBFileName (@files) {
  if ($PDBFileName =~ m/pdb$/){
    if ($PDBFileName ne $BaseStruct){
      system "echo '\nre m1 $BaseStruct\nre m2 $PDBFileName\nat ex\nex m1\nal-
124\nm2\nal-124\napply m1 m2\nwr m2 sup/$PDBFileName\nquit' |
/usr/local/dejavu/lx_lsqman";
    }
  }
}

```

LSQMAN_DomainA.pl

This program opens LSQMAN (of the DEJAVU software package from the Uppsala Software Factory) to superimpose all PDB files in the current folder. LSQMAN is run and the commands described in the comment box of the code are echoed to produce a superposition based on all atoms of residues belonging to domain A of RNase A (1-13, 49-79, and 105-124) of chain A in both files. This was written specifically for RNase A, so the commands will mostly likely need to be modified for use with other structures.

All PDB files in the current directory will be superimposed on the reference structure defined in the code, in this case, a file named “xlinkA.pdb”. The code naming the reference structure will need to be edited as necessary. To capture a log of each superposition made by LSQMAN (DEJAVU), type “perl LSQMAN_DomainA.pl > LSQ-A.log” at the prompt (as

described in the comment box of the code). Superimposed structures are saved in a new directory named “supA”. The reference structure is not superimposed nor is it saved in the “supA” directory.

```
#!/usr/bin/perl

# *****
#
# LSQMAN_DomainA.pl
# Author: Michelle Dechene
#
# Perl script to open lsqman and superimpose all pdb files in the
# current folder
#
# Superimposes to Domain A of RNase A, defined as 1-13, 49-79, 105-124
#
# Superimposed files are written in a new folder: supA
#
# lsqman path: /usr/local/dejavu/lx_lsqman
# Commands called in lsqman:
# re m1 <Reference Structure Filename>
# re m2 <Moving Structure Filename>
# at ex
# ex m1
# "a1-13 a49-79 a105-124"
# m2
# "a1-13 a49-79 a105-124"
# apply m1 m2
# wr m2 supA/<Moving Structure Filename>
# quit
#
# To run, type at prompt: perl LSQMAN_DomainA.pl > LSQ-A.log
# where LSQ-A.log can be named anything and is the logfile storing
# all the lsqman results that would normally be displayed on the
# screen
#
# *****

# The reference structure: All structures will be superimposed onto $BaseStruct
$BaseStruct = 'xlinkA.pdb';

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

mkdir "supA", 0755;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        if ($PDBFileName ne $BaseStruct){
            system "echo '\nre m1 $BaseStruct\nre m2 $PDBFileName\nat ex\nex m1\n\"a1-13 a49-79
a105-124\"\nm2\n\"a1-13 a49-79 a105-124\"\napply m1 m2\nwr m2
supA/$PDBFileName\nquit' | /usr/local/dejavu/lx_lsqman";
        }
    }
}
}
```

LSQMAN_DomainB.pl

This program opens LSQMAN (of the DEJAVU software package from the Uppsala Software Factory) to superimpose all PDB files in the current folder. LSQMAN is run and the commands described in the comment box of the code are echoed to produce a superposition based on all atoms of residues belonging to domain B of RNase A (16-46, 82-101) of chain A in both files. This was written specifically for RNase A, so the commands will mostly likely need to be modified for use with other structures.

All PDB files in the current directory will be superimposed on the reference structure defined in the code, in this case, a file named “xlinkA.pdb”. The code naming the reference structure will need to be edited as necessary. To capture a log of each superposition made by LSQMAN (DEJAVU), type “perl LSQMAN_DomainB.pl > LSQ-B.log” at the prompt (as described in the comment box of the code). Superimposed structures are saved in a new directory named “supB”. The reference structure is not superimposed nor is it saved in the “supB” directory.

```
#!/usr/bin/perl
# *****
#
# LSQMAN_DomainB.pl
# Author: Michelle Dechene
#
# Perl script to open lsqman and superimpose all pdb files in the
# current folder
#
# Superimposes to Domain B of RNase A, defined as 16-46, and 82-101
#
# Superimposed files are written in a new folder: supB
#
# lsqman path: /usr/local/dejavu/lx_lsqman
# Commands called in lsqman:
# re m1 <Reference Structure Filename>
# re m2 <Moving Structure Filename>
# at ex
# ex m1
# "a16-46 a82-101"
```

```

# m2
# "a16-46 a82-101"
# apply m1 m2
# wr m2 supB/<Moving Structure Filename>
# quit
#
# To run, type at prompt: perl LSQMAN_DomainB.pl > LSQ-B.log
# where LSQ-B.log can be named anything and is the logfile storing
# all the lsqman results that would normally be displayed on the
# screen
#
# *****

# The reference structure: All structures will be superimposed onto $BaseStruct
$BaseStruct = 'xlinkA.pdb';

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

mkdir "supB", 0755;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        if ($PDBFileName ne $BaseStruct){
            system "echo '\nre m1 $BaseStruct\nre m2 $PDBFileName\nat ex\nex m1\n\"a16-46 a82-
101\"\nm2\n\"a16-46 a82-101\"\napply m1 m2\nwr m2 supB/$PDBFileName\nquit' |
                /usr/local/dejavu/lx_lsqman";
        }
    }
}

```

MeanBBRMSD.pl

This program calculates the RMSD of each amino acid across a collection of superimposed structures in comparison with the mean backbone for the set of structures. For the set of structures, the mean backbone is calculated using only the backbone atoms (C, CA, O, N) of each amino acid. Then, backbone atoms of each structure in the set are compared to the mean backbone and the RMSD is calculated for each residue. The high RMSD value, the low RMSD value, and the average RMSD value for each amino acid are then calculated and included as the final result. Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the script assumes there is one protein molecule per file. Solvent and ligand atoms must be designated as HETATM, or they will be included in the calculations corresponding to their respective residue number. The final results of these calculations are written in a tab formatted file called “MeanBBRMSD.log” and each intermediate calculation is documented in a file named “MeanBBRMSD_Details.log” for monitoring and error-checking purposes.

```
#!/usr/bin/perl

# *****
#
# Author: Michelle Dechene
# MeanBBRMSD.pl
#
# Perl script to calculate the mean BB of a set of structures and
# then the RMSD of each amino acid (BB atoms) in each structure of
# the set in comparison with the mean BB.
#
# Solvents must be identified by HETATM instead of ATOM
#
# Assumes residues are numbered 1-999
#
# There is no error checking for Atom types, etc
#
# Assumes a single protein molecule per file
#
# Assumes the following PDB format
# ATOM    ###  xxx xxx x  ###      ##.###  ##.###  ##.###  #.##  ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

# Get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# Open output file and echo user input
open RESULTS, ">", "MeanBBRMSD.log";
open LOG, ">", "MeanBBRMSD_Details.log";
select LOG;

# Initialize $CoordinateSum and Hash, assumes residues are numbered 1-999
$CoordinateSum[0] = 0; # X coordinate sum
$CoordinateSum[1] = 0; # Y coordinate sum
$CoordinateSum[2] = 0; # Z coordinate sum
$CoordinateSum[3] = 0; # Number of Structures
$i = 1;
while ($i <= 999){
    $CoordinateSumHash{$i}{"CA"} = [@CoordinateSum]; # CA coordinates
    $CoordinateSumHash{$i}{"C"} = [@CoordinateSum]; # C coordinates
}
```

```

$CoordinateSumHash{$i}{"O"} = [@CoordinateSum]; # O coordinates
$CoordinateSumHash{$i}{"N"} = [@CoordinateSum]; # N coordinates
$i ++;
}

# Initialize Residue Count
$ResidueCount = 0;

# read in atom coordinates, for just protein atoms...
foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/) {
        print "\nPDB File input: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for HETATMs
            if ($Line =~ m/^ATOM/){
                @Atom = split /\s+/, $Line;

                if (($Atom[3] eq "ALA") || ($Atom[3] eq "CYS") || ($Atom[3] eq "ASP") ||
                    ($Atom[3] eq "GLU") || ($Atom[3] eq "PHE") || ($Atom[3] eq "GLY") ||
                    ($Atom[3] eq "HIS") || ($Atom[3] eq "ILE") || ($Atom[3] eq "LYS") ||
                    ($Atom[3] eq "LEU") || ($Atom[3] eq "MET") || ($Atom[3] eq "ASN") ||
                    ($Atom[3] eq "PRO") || ($Atom[3] eq "GLN") || ($Atom[3] eq "ARG") ||
                    ($Atom[3] eq "SER") || ($Atom[3] eq "THR") || ($Atom[3] eq "VAL") ||
                    ($Atom[3] eq "TRP") || ($Atom[3] eq "TYR")) {

                    # $AtomHash { Residue Number }{ PDB File }{ Atom Type }=[ line from PDB file ]
                    $AtomHash{$Atom[5]}{$PDBFileName}{$Atom[2]} = [@Atom];

                    # Add X,Y,Z coordinates to find mean backbone structure
                    if (($Atom[2] eq "N") || ($Atom[2] eq "CA") || ($Atom[2] eq "C") ||
                        ($Atom[2] eq "O")) {
                        @CoordinateSum = @{ $CoordinateSumHash{$Atom[5]}{$Atom[2]} };
                        $CoordinateSum[0] += $Atom[6]; # X coordinate sum
                        $CoordinateSum[1] += $Atom[7]; # Y coordinate sum
                        $CoordinateSum[2] += $Atom[8]; # Z coordinate sum
                        $CoordinateSum[3] ++; # Number of Structures
                        $CoordinateSumHash{$Atom[5]}{$Atom[2]} = [@CoordinateSum];
                    }

                    if ($Atom[5] > $ResidueCount){
                        $ResidueCount = $Atom[5];
                    }
                }
            }
        }
        close PDBFILE;
    }
}

for $ResNum ( sort keys %CoordinateSumHash ) {
    for $AtomType ( sort keys %{ $CoordinateSumHash{$ResNum} } ) {
        if ($ResNum <= $ResidueCount){
            print "Sum: $ResNum, $AtomType: @{ $CoordinateSumHash{$ResNum}{$AtomType} }\n";
        }
    }
}

# Find average backbone coordinates for each residue
for $ResNum ( sort keys %CoordinateSumHash ) {
    for $AtomType ( sort keys %{ $CoordinateSumHash{$ResNum} } ) {
        if ($ResNum <= $ResidueCount){

```

```

@CoordinateSum = @{ $CoordinateSumHash{$ResNum}{$AtomType} };
if ($CoordinateSum[3] != 0) {
    $CoordinateAvg[0] = $CoordinateSum[0]/$CoordinateSum[3]; # X coordinate average
    $CoordinateAvg[1] = $CoordinateSum[1]/$CoordinateSum[3]; # Y coordinate average
    $CoordinateAvg[2] = $CoordinateSum[2]/$CoordinateSum[3]; # Z coordinate average
}
else {
    $CoordinateAvg[0] = $CoordinateSum[0];
    $CoordinateAvg[1] = $CoordinateSum[1];
    $CoordinateAvg[2] = $CoordinateSum[2];
}
$CoordinateAvg[1], Avg Z = $CoordinateAvg[2]\n";
$CoordinateAvgHash{$ResNum}{$AtomType} = [@CoordinateAvg];
}
}
}

for $ResNum ( sort keys %CoordinateAvgHash ) {
    for $AtomType ( sort keys %{ $CoordinateAvgHash{$ResNum} } ) {
        print "Avg: $ResNum, $AtomType: @{ $CoordinateAvgHash{$ResNum}{$AtomType} }\n";
    }
}

$atomcounter = 0;
$ResidueSumOfSquares = 0;
$highRMSD = -1;
$lowRMSD = -1;
$residuecounter = 0;
$RMSDSum = 0;

# calculate RMSD across each residue for each pair of structures
for $ResNum ( sort keys %AtomHash ) {
    for $PDBFile ( sort keys %{ $AtomHash{$ResNum} } ) {
        for $AtomType (sort keys %{ $AtomHash{$ResNum}{$PDBFile} } ) {
            if (( $AtomType eq "N" ) || ( $AtomType eq "CA" ) || ( $AtomType eq "C" ) ||
                ( $AtomType eq "O" )) {
                @Atom1 = @{ $AtomHash{$ResNum}{$PDBFile}{$AtomType} };
                @AvgAtom2 = @{ $CoordinateAvgHash{$ResNum}{$AtomType} };

                # get x, y, z for atom in first residue
                $X1 = @Atom1[6];
                $Y1 = @Atom1[7];
                $Z1 = @Atom1[8];

                # get x, y, z for atom in second residue (mean residue)
                $X2 = @AvgAtom2[0];
                $Y2 = @AvgAtom2[1];
                $Z2 = @AvgAtom2[2];

                $XDiff = $X2 - $X1;
                $YDiff = $Y2 - $Y1;
                $ZDiff = $Z2 - $Z1;
                $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
                $atomcounter += 1;
                $ResidueSumOfSquares += $SumOfSquares;
            }
        }
    }
}

# calculate RMSD
if ($atomcounter != 0) {
    $Mean = $ResidueSumOfSquares / $atomcounter;
    $RMSD = sqrt $Mean;
    print "Residue: $ResNum, PDB1: $PDBFile\n";
    print "RMSD: $RMSD\n";

    # add to sum, increment residue counter
}

```

```

$RMSDSum += $RMSD;
$residuecounter += 1;

# save high value
if (($highRMSD == -1) || ($RMSD > $highRMSD)) {
    $highRMSD = $RMSD;
}

# save low value
if (($lowRMSD == -1) || ($RMSD < $lowRMSD)) {
    $lowRMSD = $RMSD;
}

$ResidueSumOfSquares = 0;
$SumOfSquares = 0;
$atomcounter = 0;
}
}

if ($residuecounter != 0){
# calculate average of all the RMSD's for the residue
$AverageRMSD = $RMSDSum / $residuecounter;

# store high, low, avg, and res# (in an array for sorting purposes)
$ResidueRMSD[$ResNum][0] = $AverageRMSD;
$ResidueRMSD[$ResNum][1] = $highRMSD;
$ResidueRMSD[$ResNum][2] = $lowRMSD;

# reset high, low, sum, counter
$highRMSD = -1;
$lowRMSD = -1;
$residuecounter = 0;
$RMSDSum = 0;
}
}
close LOG;

select RESULTS;
# output an excel formatted file with residue number, high, low, and average RMSD values
$i = 1;
print "Residue\tAvg RMSD\tHigh RMSD\tLow RMSD\n";
while ($i <= $#ResidueRMSD){
    printf "%i\t%2.6f\t%2.6f\t%2.6f\n", $ResidueRMSD[$i][0], $ResidueRMSD[$i][1],
$ResidueRMSD[$i][2];
    $i += 1;
}

# close files
close RESULTS;

```

MeanBBRMSDMedian.pl

This program calculates the RMSD of each amino acid across a collection of superimposed structures in comparison with the mean backbone for the set of structures. For the set of structures, the mean backbone is calculated using only the backbone atoms (C, CA, O, N) of

each amino acid. Then, backbone atoms of each structure in the set are compared to the mean backbone and the RMSD is calculated for each residue. The high RMSD value, the low RMSD value, the average RMSD value, and the median of all RMSD values for each amino acid are then calculated and included as the final result. Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the script assumes there is one protein molecule per file. Solvent and ligand atoms must be designated as HETATM, or they will be included in the calculations corresponding to their respective residue number. The final results of these calculations are written in a tab formatted file called “MeanBBRMSDMedian.log” and each intermediate calculation is documented in a file named “MeanBBRMSDMedian_Details.log” for monitoring and error-checking purposes.

```
#!/usr/bin/perl

# *****
#
# Author: Michelle Dechene
# MeanBBRMSDMedian.pl
#
# Perl script to calculate the mean BB of a set of structures and
# then the RMSD of each amino acid (BB atoms) in each structure of
# the set in comparison with the mean BB.
#
# Solvents must be identified by HETATM instead of ATOM
#
# Assumes residues are numbered 1-999
#
# There is no error checking for Atom types, etc
#
# Assumes a single protein molecule per file
#
# Assumes the following PDB format
# ATOM    ###  xxx xxx x  ###      ##.###  ##.###  ##.###  #.##  ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
```

```

# *****

# Get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# Open output file and echo user input
open RESULTS, ">", "MeanBBRMSDMedian.log";
open LOG, ">", "MeanBBRMSDMedian_Details.log";
select LOG;

# Initialize $CoordinateSum and Hash, assumes residues are numbered 1-999
$CoordinateSum[0] = 0; # X coordinate sum
$CoordinateSum[1] = 0; # Y coordinate sum
$CoordinateSum[2] = 0; # Z coordinate sum
$CoordinateSum[3] = 0; # Number of Structures

$i = 1;
while ($i <= 999){
    $CoordinateSumHash{$i}{"CA"} = [@CoordinateSum]; # CA coordinates
    $CoordinateSumHash{$i}{"C"} = [@CoordinateSum]; # C coordinates
    $CoordinateSumHash{$i}{"O"} = [@CoordinateSum]; # O coordinates
    $CoordinateSumHash{$i}{"N"} = [@CoordinateSum]; # N coordinates
    $i ++;
}

# Initialize Residue Count
$ResidueCount = 0;
@RMSDList = undef;

# read in atom coordinates, for just protein atoms...
foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File input: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for HETATMs
            if ($Line =~ m/^ATOM/){
                @Atom = split /\s+/, $Line;

                if (($Atom[3] eq "ALA") || ($Atom[3] eq "CYS") || ($Atom[3] eq "ASP") ||
                    ($Atom[3] eq "GLU") || ($Atom[3] eq "PHE") || ($Atom[3] eq "GLY") ||
                    ($Atom[3] eq "HIS") || ($Atom[3] eq "ILE") || ($Atom[3] eq "LYS") ||
                    ($Atom[3] eq "LEU") || ($Atom[3] eq "MET") || ($Atom[3] eq "ASN") ||
                    ($Atom[3] eq "PRO") || ($Atom[3] eq "GLN") || ($Atom[3] eq "ARG") ||
                    ($Atom[3] eq "SER") || ($Atom[3] eq "THR") || ($Atom[3] eq "VAL") ||
                    ($Atom[3] eq "TRP") || ($Atom[3] eq "TYR")) {

                    # $AtomHash { Residue Number } { PDB File } { Atom Type }=[ line from PDB file ]
                    $AtomHash{$Atom[5]}{$PDBFileName}{$Atom[2]} = [@Atom];

                    # Add X,Y,Z coordinates to find mean backbone structure
                    if (($Atom[2] eq "N") || ($Atom[2] eq "CA") || ($Atom[2] eq "C")
                        || ($Atom[2] eq "O")) {
                        @CoordinateSum = @{$CoordinateSumHash{$Atom[5]}{$Atom[2]}};
                        $CoordinateSum[0] += $Atom[6]; # X coordinate sum
                        $CoordinateSum[1] += $Atom[7]; # Y coordinate sum
                        $CoordinateSum[2] += $Atom[8]; # Z coordinate sum
                        $CoordinateSum[3] ++; # Number of Structures
                        $CoordinateSumHash{$Atom[5]}{$Atom[2]} = [@CoordinateSum];
                    }
                }
            }
        }
    }
}

```

```

        if ($Atom[5] > $ResidueCount){
            $ResidueCount = $Atom[5];
        }
    }
}
close PDBFILE;
}
}

# Find average backbone coordinates for each residue
for $ResNum ( sort keys %CoordinateSumHash ) {
    for $AtomType ( sort keys %{ $CoordinateSumHash{$ResNum} } ) {
        if ($ResNum <= $ResidueCount){
            @CoordinateSum = @{ $CoordinateSumHash{$ResNum}{$AtomType} };
            if ($CoordinateSum[3] != 0){
                $CoordinateAvg[0] = $CoordinateSum[0]/$CoordinateSum[3]; # X coordinate average
                $CoordinateAvg[1] = $CoordinateSum[1]/$CoordinateSum[3]; # Y coordinate average
                $CoordinateAvg[2] = $CoordinateSum[2]/$CoordinateSum[3]; # Z coordinate average
            }
            else {
                $CoordinateAvg[0] = $CoordinateSum[0];
                $CoordinateAvg[1] = $CoordinateSum[1];
                $CoordinateAvg[2] = $CoordinateSum[2];
            }
            $CoordinateAvg[1], Avg Z = $CoordinateAvg[2]\n";
            $CoordinateAvgHash{$ResNum}{$AtomType} = [@CoordinateAvg];
        }
    }
}

$atomcounter = 0;
$ResidueSumOfSquares = 0;
$highRMSD = -1;
$lowRMSD = -1;
$residuecounter = 0;
$RMSDSum = 0;
#$ResNumcounter = 0;

# calculate RMSD across each residue for each pair of structures
for $ResNum ( sort keys %AtomHash ) {
    for $PDBFile ( sort keys %{ $AtomHash{$ResNum} } ) {
        for $AtomType (sort keys %{ $AtomHash{$ResNum}{$PDBFile} } ) {
            if (($AtomType eq "N") || ($AtomType eq "CA") || ($AtomType eq "C") ||
                ($AtomType eq "O")) {
                @Atom1 = @{ $AtomHash{$ResNum}{$PDBFile}{$AtomType} };
                @AvgAtom2 = @{ $CoordinateAvgHash{$ResNum}{$AtomType} };

                # get x, y, z for atom in first residue
                $X1 = @Atom1[6];
                $Y1 = @Atom1[7];
                $Z1 = @Atom1[8];

                # get x, y, z for atom in second residue (mean residue)
                $X2 = @AvgAtom2[0];
                $Y2 = @AvgAtom2[1];
                $Z2 = @AvgAtom2[2];

                $XDiff = $X2 - $X1;
                $YDiff = $Y2 - $Y1;
                $ZDiff = $Z2 - $Z1;
                $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;

                $atomcounter += 1;
                $ResidueSumOfSquares += $SumOfSquares;
            }
        }
    }
}

```

```

    }
}

# calculate RMSD
if ($atomcounter != 0) {
    $Mean = $ResidueSumOfSquares / $atomcounter;
    $RMSD = sqrt $Mean;
    print "Residue: $ResNum, PDB1: $PDBFile\n";
    print "RMSD: $RMSD\n";

    # add to sum, increment residue counter
    $RMSDSum += $RMSD;
    $residuecounter += 1;

    # build RMSD array; for whatever reason, array element with index 0 is left as
    # undefined
    push(@RMSDList, $RMSD);

    # save high value
    if (($highRMSD == -1) || ($RMSD > $highRMSD)) {
        $highRMSD = $RMSD;
    }

    # save low value
    if (($lowRMSD == -1) || ($RMSD < $lowRMSD)) {
        $lowRMSD = $RMSD;
    }

    $ResidueSumOfSquares = 0;
    $SumOfSquares = 0;
    $atomcounter = 0;
}
}

# Sort the array and find the median
@SortedRMSD = sort(@RMSDList);
$RMSDelements = @SortedRMSD;

print "\n\nSorted RMSD Array: @SortedRMSD, Elements: $RMSDelements\n\n";
# if statements are backwards because the zeroth element is undefined
if ( ($RMSDelements % 2) == 0) {
    $median = $SortedRMSD[($RMSDelements / 2)];
}
else {
    $median = ($SortedRMSD[int($RMSDelements / 2)] + $SortedRMSD[int($RMSDelements / 2)
+ 1]) / 2;
}

if ($residuecounter != 0){
    # calculate average of all the RMSD's for the residue
    $AverageRMSD = $RMSDSum / $residuecounter;

    # store high, low, avg, and res# (in an array for sorting purposes)
    $ResidueRMSD[$ResNum][0] = $AverageRMSD;
    $ResidueRMSD[$ResNum][1] = $highRMSD;
    $ResidueRMSD[$ResNum][2] = $lowRMSD;
    $ResidueRMSD[$ResNum][3] = $median;

    # reset high, low, sum, counter
    $highRMSD = -1;
    $lowRMSD = -1;
    $residuecounter = 0;
    $RMSDSum = 0;
    @RMSDList = undef;
}
}
}

```

```

close LOG;

select RESULTS;
# output an excel formatted file with residue number, high, low, and average RMSD values
$i = 1;
print "Residue\tAvg RMSD\tHigh RMSD\tLow RMSD\tMedian RMSD\n";
while ($i <= $#ResidueRMSD){
    printf "$i\t%.2f\t%.2f\t%.2f\t%.2f\n", $ResidueRMSD[$i][0], $ResidueRMSD[$i][1],
$ResidueRMSD[$i][2], $ResidueRMSD[$i][3];
    $i += 1;
}

# close files
close RESULTS;

```

MeanBBRMSDSubstr.pl

This script probably does not run properly as it was scrapped in the interests of time. It was started to do the same thing as MeanBBRMSD.pl, but without the need for defined protein chains (using substr instead of splitting on spaces).

This program calculates the RMSD of each amino acid across a collection of superimposed structures in comparison with the mean backbone for the set of structures. For the set of structures, the mean backbone is calculated using only the backbone atoms (C, CA, O, N) of each amino acid. Then, backbone atoms of each structure in the set are compared to the mean backbone and the RMSD is calculated for each residue. The high RMSD value, the low RMSD value, and the average RMSD value for each amino acid are then calculated and included as the final result. Coordinates from the file are read into an array from each line filling the array with substrings of each line. The script should work independently of a defined or undefined protein chain.

All PDB files in the current folder will be used for the calculations and the script assumes there is one protein molecule per file. Solvent and ligand atoms must be designated as HETATM, or they will be included in the calculations corresponding to their respective residue number. The final results of these calculations are written in a tab formatted file called “MeanBBRMSD.log” and each intermediate calculation is documented in a file named “MeanBBRMSD_Details.log” for monitoring and error-checking purposes.

```
#!/usr/bin/perl

# *****
#
# Author: Michelle Dechene
# MeanBBRMSDSubstr.pl
#
# Perl script to calculate the mean BB of a set of structures and
# then the RMSD of each amino acid (BB atoms) in each structure of
# the set in comparison with the mean BB.
#
# Solvents must be identified by HETATM instead of ATOM
#
# Assumes residues are numbered 1-999
#
# There is no error checking for Atom types, etc
#
# Assumes a single protein molecule per file
#
# *****

# Get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# Open output file and echo user input
open RESULTS, ">", "MeanBBRMSD.log";
open LOG, ">", "MeanBBRMSD_Details.log";
select LOG;

# Initialize $CoordinateSum and Hash, assumes residues are numbered 1-999
$CoordinateSum[0] = 0; # X coordinate sum
$CoordinateSum[1] = 0; # Y coordinate sum
$CoordinateSum[2] = 0; # Z coordinate sum
$CoordinateSum[3] = 0; # Number of Structures
$i = 1;
while ($i <= 999){
    $CoordinateSumHash{$i}{"CA"} = [@CoordinateSum]; # CA coordinates
    $CoordinateSumHash{$i}{"C"} = [@CoordinateSum]; # C coordinates
    $CoordinateSumHash{$i}{"O"} = [@CoordinateSum]; # O coordinates
    $CoordinateSumHash{$i}{"N"} = [@CoordinateSum]; # N coordinates
    $i ++;
}

# Initialize Residue Count
$ResidueCount = 0;
```

```

# read in atom coordinates, for just protein atoms...
foreach $PDBFileName (@files) {
  if ($PDBFileName =~ m/pdb$/){
    print "\nPDB File input: $PDBFileName\n";
    open PDBFILE, "<", $PDBFileName;

    # Read in PDB file into array
    chomp(@FileLines = <PDBFILE>);

    foreach $Line (@FileLines) {
      # Search for HETATMs
      if ($Line =~ m/^ATOM/){
        # Get residue type
        $ResType = substr($Line, 17, 3);
        if (($ResType eq "ALA") || ($ResType eq "CYS") || ($ResType eq "ASP") ||
            ($ResType eq "GLU") || ($ResType eq "PHE") || ($ResType eq "GLY") ||
            ($ResType eq "HIS") || ($ResType eq "ILE") || ($ResType eq "LYS") ||
            ($ResType eq "LEU") || ($ResType eq "MET") || ($ResType eq "ASN") ||
            ($ResType eq "PRO") || ($ResType eq "GLN") || ($ResType eq "ARG") ||
            ($ResType eq "SER") || ($ResType eq "THR") || ($ResType eq "VAL") ||
            ($ResType eq "TRP") || ($ResType eq "TYR")) {

          # Assumes residue are numbered 1-999
          $Atom[0] = substr($Line, 23, 3);
          # $Atom[1] = Atom Type
          $Atom[1] = substr($Line, 13, 3);
          # $Atom[2] = X coordinate
          $Atom[2] = substr($Line, 30, 8);
          # $Atom[3] = Y coordinate
          $Atom[3] = substr($Line, 38, 8);
          # $Atom[4] = Z coordinate
          $Atom[4] = substr($Line, 46, 8);

          # $AtomHash { Residue Number } { PDB File } { Atom Type } = [ Atom Array:
          # Residue Number, Atom Type, & X,Y,Z coordinates]
          $AtomHash{$Atom[0]}{$PDBFileName}{$Atom[1]} = [@Atom];

          # Add X,Y,Z coordinates to find mean backbone structure
          if (($Atom[1] eq "N ") || ($Atom[1] eq "CA ") || ($Atom[1] eq "C ") ||
              ($Atom[1] eq "O ")) {
            @CoordinateSum = @($CoordinateSumHash{$Atom[0]}{$Atom[1]});
            $CoordinateSum[0] += $Atom[2]; # X coordinate sum
            $CoordinateSum[1] += $Atom[3]; # Y coordinate sum
            $CoordinateSum[2] += $Atom[4]; # Z coordinate sum
            $CoordinateSum[3] ++; # Number of Structures
            $CoordinateSumHash{$Atom[0]}{$Atom[1]} = [@CoordinateSum];
          }

          if ($Atom[0] > $ResidueCount){
            $ResidueCount = $Atom[0];
          }
        }
      }
    }
    close PDBFILE;
  }
}

for $ResNum ( sort keys %CoordinateSumHash ) {
  for $AtomType ( sort keys %{ $CoordinateSumHash{$ResNum} } ) {
    print "Sum: $ResNum, $AtomType: @($CoordinateSumHash{$ResNum}{$AtomType})\n";
  }
}

# Find average backbone coordinates for each residue
for $ResNum ( sort keys %CoordinateSumHash ) {

```

```

for $AtomType ( sort keys %{ $CoordinateSumHash{$ResNum} } ) {
    if ($ResNum <= $ResidueCount){
        @CoordinateSum = @{ $CoordinateSumHash{$ResNum}{$AtomType} };

        if ($CoordinateSum[3] != 0){
            $CoordinateAvg[0] = $CoordinateSum[0]/$CoordinateSum[3]; # X coordinate average
            $CoordinateAvg[1] = $CoordinateSum[1]/$CoordinateSum[3]; # Y coordinate average
            $CoordinateAvg[2] = $CoordinateSum[2]/$CoordinateSum[3]; # Z coordinate average
        }
        else {
            $CoordinateAvg[0] = $CoordinateSum[0];
            $CoordinateAvg[1] = $CoordinateSum[1];
            $CoordinateAvg[2] = $CoordinateSum[2];
        }
        $CoordinateAvgHash{$ResNum}{$AtomType} = [@CoordinateAvg];
    }
}

for $ResNum ( sort keys %CoordinateAvgHash ) {
    for $AtomType ( sort keys %{ $CoordinateAvgHash{$ResNum} } ) {
        print "Avg: $ResNum, $AtomType: @{ $CoordinateAvgHash{$ResNum}{$AtomType} }\n";
    }
}

```

NoTer.pl

This program edits PDB files and removes TER tags. Changed files are saved in a new directory called “NoTER_PDB”.

All PDB files in the current folder will be modified, if necessary, and saved in the new directory.

```

#!/usr/bin/perl

# *****
#
# NoTER.pl
# Author: Michelle Dechene
#
# Perl script to remove TER from PDB files
#
# This script works on all PDB files in the current directory.
#
# Old files are left alone and new split files are written in a new
# directory called NoTer_PDB
#
# *****

# get all file names from the current directory
opendir MYDIR, ".";

```

```

@files = readdir MYDIR;
closedir MYDIR;

mkdir "NoTER_PDB", 0755;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        open PDBFILE, "<", $PDBFileName;

        open PDBNEW, ">", "NoTER_PDB/" . $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        select PDBNEW;
        foreach $Line (@FileLines) {
            # Locate lines of non-protein and non-water atoms
            if ($Line =~ m/^TER/){
                # Do Not Print
            }
            else{
                print "$Line\n";
            }
        }

        close PDBNEW;
        close PDBFILE;
    }
}

```

RMSD4NMRAAll.pl

This program is identical to RMSD4ProteinsAll.pl, except that it does not save a value of zero as the low RMSD value.

This program calculates the RMSD of each amino acid across a collection of superimposed structures. For each pair of structures, the RMSD is calculated per residue using all atoms available for that amino acid. The high RMSD value, the low RMSD value, and the average RMSD value for each residue are then calculated and included as the final result.

Coordinates from the file are read into an array from each line splitting the values by spaces.

The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the script assumes there is one protein molecule per file. Solvent and ligand atoms must be designated as HETATM, or they will be included in the calculations corresponding to their respective residue number. The final results of these calculations are written in a tab formatted file called “RMSD4NMRAll.log” and each intermediate calculation is documented in a file named “RMSD_NMRAll_Details.log” for monitoring and error-checking purposes.

```
#!/usr/bin/perl

# *****
#
# RMSD4ProteinsALL.pl
# Author: Michelle Dechene
#
# Perl script to calculate the RMSD of each amino acid across a
# collection of structures (of the same superimposed protein)
#
# Solvents must be identified by HETATM instead of ATOM
#
# Assumes only one protein chain in each PDB file
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ###   ##.###   ##.###   ##.###   #.##   ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# 0 will not be stored as a low RMSD value
#
# *****

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "RMSD4NMRAll.log";
open LOG, ">", "RMSD_NMRAll_Details.log";
select LOG;

# read in atom coordinates, for just protein atoms...
foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File input: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for HETATMs
            if ($Line =~ m/^ATOM/){
                @Atom = split /\s+/, $Line;
            }
        }
    }
}
```

```

        if ($Atom[3] ne "HOH") {
            # $AtomHash { Residue Number } { PDB File } { Atom Type } = [ line from PDB
            # file ]
            $AtomHash{$Atom[5]}{$PDBFileName}{$Atom[2]} = [@Atom];
        }
    }
}
close PDBFILE;
}
}

$atomcounter = 0;
$ResidueSumOfSquares = 0;
$highRMSD = -1;
$lowRMSD = -1;
$residuecounter = 0;
$RMSDSum = 0;
#$ResNumcounter = 0;

# calculate RMSD across each residue for each pair of structures
for $ResNum ( sort keys %AtomHash ) {
    for $PDBFile1 ( sort keys %{ $AtomHash{$ResNum} } ) {
        for $PDBFile2 ( sort keys %{ $AtomHash{$ResNum} } ) {
            for $AtomType1 ( sort keys %{ $AtomHash{$ResNum}{$PDBFile1} } ) {
                for $AtomType2 ( sort keys %{ $AtomHash{$ResNum}{$PDBFile2} } ) {
                    if (($PDBFile1 ne $PDBFile2) && ($AtomType1 eq $AtomType2)) {

                        @Atom1 = @{ $AtomHash{$ResNum}{$PDBFile1}{$AtomType1} };
                        @Atom2 = @{ $AtomHash{$ResNum}{$PDBFile2}{$AtomType2} };

                        # get x, y, z for atom in first residue
                        $X1 = @Atom1[6];
                        $Y1 = @Atom1[7];
                        $Z1 = @Atom1[8];

                        # get x, y, z for atom in second residue
                        $X2 = @Atom2[6];
                        $Y2 = @Atom2[7];
                        $Z2 = @Atom2[8];

                        $XDiff = $X2 - $X1;
                        $YDiff = $Y2 - $Y1;
                        $ZDiff = $Z2 - $Z1;
                        $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
                        $atomcounter += 1;
                    }
                }
            }
            $ResidueSumOfSquares += $SumOfSquares;
        }
    }
}

# calculate RMSD
if ($atomcounter != 0) {
    $Mean = $ResidueSumOfSquares / $atomcounter;
    $RMSD = sqrt $Mean;
    print "Residue: $ResNum, PDB1: $PDBFile1, PDB2: $PDBFile2\n";
    print "RMSD: $RMSD\n";

    # add to sum, increment residue counter
    $RMSDSum += $RMSD;
    $residuecounter += 1;

    # save high value
    if (($highRMSD == -1) || ($RMSD > $highRMSD)) {
        $highRMSD = $RMSD;
    }

    # save low value
}

```

```

        if (($lowRMSD == -1) || ($RMSD < $lowRMSD)) && ($RMSD != 0) {
            $lowRMSD = $RMSD;
        }

        $ResidueSumOfSquares = 0;
        $SumOfSquares = 0;
        $atomcounter = 0;
    }
}

if ($residuecounter != 0){
    # calculate average of all the RMSD's for the residue
    $AverageRMSD = $RMSDSum / $residuecounter;
    # store high, low, avg, and res# (in an array for sorting purposes)
    $ResidueRMSD[$ResNum][0] = $AverageRMSD;
    $ResidueRMSD[$ResNum][1] = $highRMSD;
    $ResidueRMSD[$ResNum][2] = $lowRMSD;

    # reset high, low, sum, counter
    $highRMSD = -1;
    $lowRMSD = -1;
    $residuecounter = 0;
    $RMSDSum = 0;
}
}
close LOG;

select RESULTS;
# output an excel formatted file with residue number, high, low, and average RMSD values
$i = 1;
print "Residue\tAvg RMSD\tHigh RMSD\tLow RMSD\n";
while ($i <= $#ResidueRMSD){
    printf "$i\t%.2f\t%.2f\t%.2f\n", $ResidueRMSD[$i][0], $ResidueRMSD[$i][1],
        $ResidueRMSD[$i][2];
    $i += 1;
}

# close files
close RESULTS;

```

RMSD4NMRBB.pl

This program is identical to RMSD4ProteinsBB.pl, except that it does not save a value of zero as the low RMSD value.

This program calculates the RMSD of each amino acid across a collection of superimposed structures. For each pair of structures, the RMSD is calculated per residue using only the backbone atoms (C, CA, O, N) for that amino acid. The high RMSD value, the low RMSD

value, and the average RMSD value for each residue are then calculated and included as the final result. Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the script assumes there is one protein molecule per file. Solvent and ligand atoms must be designated as HETATM, or they will be included in the calculations corresponding to their respective residue number. The final results of these calculations are written in a tab formatted file called "RMSD4NMRBB.log" and each intermediate calculation is documented in a file named "RMSD_NMRBB_Details.log" for monitoring and error-checking purposes.

```
#!/usr/bin/perl

# *****
#
# RMSD4ProteinsBB.pl
# Author: Michelle Dechene
#
# Perl script to calculate the RMSD of each amino acid across a
# collection of structures (of the same superimposed protein)
#
# Solvents must be identified by HETATM instead of ATOM
#
# Assumes only one protein chain in each PDB file
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###  xxx xxx x  ###      ##.###  ##.###  ##.###  #.##  ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# 0 will not be stored as a low RMSD value.
#
# *****

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "RMSD4NMRBB.log";
open LOG, ">", "RMSD_NMRBB_Details.log";
```

```

select LOG;

# read in atom coordinates, for just protein atoms...
foreach $PDBFileName (@files) {
  if ($PDBFileName =~ m/pdb$/){
    print "\nPDB File input: $PDBFileName\n";
    open PDBFILE, "<", $PDBFileName;

    # Read in PDB file into array
    chomp(@FileLines = <PDBFILE>);

    foreach $Line (@FileLines) {
      # Search for HETATMs
      if ($Line =~ m/^ATOM/){
        @Atom = split /\s+/, $Line;
        if ($Atom[3] ne "HOH") {
          # $AtomHash { Residue Number } { PDB File } { Atom Type } = [ line from PDB
          # file ]
          $AtomHash{$Atom[5]}{$PDBFileName}{$Atom[2]} = [@Atom];
        }
      }
    }
    close PDBFILE;
  }
}

$atomcounter = 0;
$ResidueSumOfSquares = 0;
$highRMSD = -1;
$lowRMSD = -1;
$residuecounter = 0;
$RMSDSum = 0;
#$ResNumcounter = 0;

# calculate RMSD across each residue for each pair of structures
for $ResNum ( sort keys %AtomHash ) {
  for $PDBFile1 ( sort keys %{ $AtomHash{$ResNum} } ) {
    for $PDBFile2 ( sort keys %{ $AtomHash{$ResNum} } ) {
      for $AtomType1 (sort keys %{ $AtomHash{$ResNum}{$PDBFile1} } ) {
        for $AtomType2 (sort keys %{ $AtomHash{$ResNum}{$PDBFile2} } ) {
          if (($PDBFile1 ne $PDBFile2) && ($AtomType1 eq $AtomType2)) {
            if (($AtomType1 eq "N" || ($AtomType1 eq "CA" || ($AtomType1 eq "C" ||
              ($AtomType1 eq "O")) {

              @Atom1 = @{ $AtomHash{$ResNum}{$PDBFile1}{$AtomType1} };
              @Atom2 = @{ $AtomHash{$ResNum}{$PDBFile2}{$AtomType2} };

              # get x, y, z for atom in first residue
              $X1 = @Atom1[6];
              $Y1 = @Atom1[7];
              $Z1 = @Atom1[8];

              # get x, y, z for atom in second residue
              $X2 = @Atom2[6];
              $Y2 = @Atom2[7];
              $Z2 = @Atom2[8];

              $XDiff = $X2 - $X1;
              $YDiff = $Y2 - $Y1;
              $ZDiff = $Z2 - $Z1;
              $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
              $atomcounter += 1;
              $ResidueSumOfSquares += $SumOfSquares;
            }
          }
        }
      }
    }
  }
}

```

```

    }
    # calculate RMSD
    if ($atomcounter != 0) {
        $Mean = $ResidueSumOfSquares / $atomcounter;
        $RMSD = sqrt $Mean;
        print "Residue: $ResNum, PDB1: $PDBFile1, PDB2: $PDBFile2\n";
        print "RMSD: $RMSD\n";

        # add to sum, increment residue counter
        $RMSDSum += $RMSD;
        $residuecounter += 1;

        # save high value
        if (($highRMSD == -1) || ($RMSD > $highRMSD)) {
            $highRMSD = $RMSD;
        }

        # save low value
        if (($lowRMSD == -1) || ($RMSD < $lowRMSD) && ($RMSD != 0)) {
            $lowRMSD = $RMSD;
        }

        $ResidueSumOfSquares = 0;
        $SumOfSquares = 0;
        $atomcounter = 0;
    }
}

if ($residuecounter != 0) {
    # calculate average of all the RMSD's for the residue
    $AverageRMSD = $RMSDSum / $residuecounter;

    # store high, low, avg, and res# (in an array for sorting purposes)
    $ResidueRMSD[$ResNum][0] = $AverageRMSD;
    $ResidueRMSD[$ResNum][1] = $highRMSD;
    $ResidueRMSD[$ResNum][2] = $lowRMSD;

    # reset high, low, sum, counter
    $highRMSD = -1;
    $lowRMSD = -1;
    $residuecounter = 0;
    $RMSDSum = 0;
}
}
close LOG;

select RESULTS;
# output an excel formatted file with residue number, high, low, and average RMSD values
$i = 1;
print "Residue\tAvg RMSD\tHigh RMSD\tLow RMSD\n";
while ($i <= $#ResidueRMSD) {
    printf "$i\t%.2f\t%.2f\t%.2f\n", $ResidueRMSD[$i][0], $ResidueRMSD[$i][1],
        $ResidueRMSD[$i][2];
    $i += 1;
}

# close files
close RESULTS;

```

RMSD4ProteinsAll.pl

This program calculates the RMSD of each amino acid across a collection of superimposed structures. For each pair of structures, the RMSD is calculated per residue using all atoms available for that amino acid. The high RMSD value, the low RMSD value, and the average RMSD value for each residue are then calculated and included as the final result.

Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the script assumes there is one protein molecule per file. Solvent and ligand atoms must be designated as HETATM, or they will be included in the calculations corresponding to their respective residue number. The final results of these calculations are written in a tab formatted file called "RMSD4Proteins.log" and each intermediate calculation is documented in a file named "RMSD_All_Details.log" for monitoring and error-checking purposes.

```
#!/usr/bin/perl

# *****
#
# RMSD4ProteinsALL.pl
# Author: Michelle Dechene
#
# Perl script to calculate the RMSD of each amino acid across a
# collection of structures (of the same superimposed protein)
#
# Solvents must be identified by HETATM instead of ATOM
#
# Assumes only one protein chain in each PDB file
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###  xxx xxx x  ###      ##.###  ##.###  ##.###  #.# #.#
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

# get all file names from the current directory
```

```

opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "RMSD4Proteins.log";
open LOG, ">", "RMSD_All_Details.log";
select LOG;

# read in atom coordinates, for just protein atoms...
foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File input: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for HETATMs
            if ($Line =~ m/^ATOM/){
                @Atom = split /\s+/, $Line;
                if ($Atom[3] ne "HOH") {
                    # $AtomHash { Residue Number } { PDB File } { Atom Type } = [ line from PDB
                    # file ]
                    $AtomHash{$Atom[5]}{$PDBFileName}{$Atom[2]} = [@Atom];
                }
            }
        }
        close PDBFILE;
    }
}

$atomcounter = 0;
$ResidueSumOfSquares = 0;
$highRMSD = -1;
$lowRMSD = -1;
$residuecounter = 0;
$RMSDSum = 0;
#$ResNumcounter = 0;

# calculate RMSD across each residue for each pair of structures
for $ResNum ( sort keys %AtomHash ) {
    for $PDBFile1 ( sort keys %{ $AtomHash{$ResNum} } ) {
        for $PDBFile2 ( sort keys %{ $AtomHash{$ResNum} } ) {
            for $AtomType1 (sort keys %{ $AtomHash{$ResNum}{$PDBFile1} } ) {
                for $AtomType2 (sort keys %{ $AtomHash{$ResNum}{$PDBFile2} } ) {
                    if (($PDBFile1 ne $PDBFile2) && ($AtomType1 eq $AtomType2)) {

                        @Atom1 = @{ $AtomHash{$ResNum}{$PDBFile1}{$AtomType1} };
                        @Atom2 = @{ $AtomHash{$ResNum}{$PDBFile2}{$AtomType2} };

                        # get x, y, z for atom in first residue
                        $X1 = @Atom1[6];
                        $Y1 = @Atom1[7];
                        $Z1 = @Atom1[8];

                        # get x, y, z for atom in second residue
                        $X2 = @Atom2[6];
                        $Y2 = @Atom2[7];
                        $Z2 = @Atom2[8];

                        $XDiff = $X2 - $X1;
                        $YDiff = $Y2 - $Y1;
                        $ZDiff = $Z2 - $Z1;
                        $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
                    }
                }
            }
        }
    }
}

```

```

        $atomcounter += 1;
    }
    $ResidueSumOfSquares += $SumOfSquares;
}
# calculate RMSD
if ($atomcounter != 0) {
    $Mean = $ResidueSumOfSquares / $atomcounter;
    $RMSD = sqrt $Mean;
    print "Residue: $ResNum, PDB1: $PDBFile1, PDB2: $PDBFile2\n";
    print "RMSD: $RMSD\n";

    # add to sum, increment residue counter
    $RMSDSum += $RMSD;
    $residuecounter += 1;

    # save high value
    if (($highRMSD == -1) || ($RMSD > $highRMSD)) {
        $highRMSD = $RMSD;
    }

    # save low value
    if (($lowRMSD == -1) || ($RMSD < $lowRMSD)) {
        $lowRMSD = $RMSD;
    }

    $ResidueSumOfSquares = 0;
    $SumOfSquares = 0;
    $atomcounter = 0;
}
}
}

if ($residuecounter != 0) {
    # calculate average of all the RMSD's for the residue
    $AverageRMSD = $RMSDSum / $residuecounter;

    # store high, low, avg, and res# (in an array for sorting purposes)
    $ResidueRMSD[$ResNum][0] = $AverageRMSD;
    $ResidueRMSD[$ResNum][1] = $highRMSD;
    $ResidueRMSD[$ResNum][2] = $lowRMSD;

    # reset high, low, sum, counter
    $highRMSD = -1;
    $lowRMSD = -1;
    $residuecounter = 0;
    $RMSDSum = 0;
}
}
close LOG;

select RESULTS;
# output an excel formatted file with residue number, high, low, and average RMSD values
$i = 1;
print "Residue\tAvg RMSD\tHigh RMSD\tLow RMSD\n";
while ($i <= $#ResidueRMSD) {
    printf "$i\t%.2f\t%.2f\t%.2f\n", $ResidueRMSD[$i][0], $ResidueRMSD[$i][1],
        $ResidueRMSD[$i][2];
    $i += 1;
}

# close files
close RESULTS;

```

RMSD4ProteinsBB.pl

This program calculates the RMSD of each amino acid across a collection of superimposed structures. For each pair of structures, the RMSD is calculated per residue using only the backbone atoms (C, CA, O, N) for that amino acid. The high RMSD value, the low RMSD value, and the average RMSD value for each residue are then calculated and included as the final result. Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the script assumes there is one protein molecule per file. Solvent and ligand atoms must be designated as HETATM, or they will be included in the calculations corresponding to their respective residue number. The final results of these calculations are written in a tab formatted file called "RMSD4ProteinsBB.log" and each intermediate calculation is documented in a file named "RMSD_BB_Details.log" for monitoring and error-checking purposes.

```
#!/usr/bin/perl
# *****
#
# RMSD4ProteinsBB.pl
# Author: Michelle Dechene
#
# Perl script to calculate the RMSD of each amino acid across a
# collection of structures (of the same superimposed protein)
#
# Solvents must be identified by HETATM instead of ATOM
#
# Assumes only one protein chain in each PDB file
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ##   ##   ##   #.##   ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
```

```

# *****

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "RMSD4ProteinsBB.log";
open LOG, ">", "RMSD_BB_Details.log";
select LOG;

# read in atom coordinates, for just protein atoms...
foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File input: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for HETATMs
            if ($Line =~ m/^ATOM/){
                @Atom = split /\s+/, $Line;
                if ($Atom[3] ne "HOH") {
                    # $AtomHash { Residue Number } { PDB File } { Atom Type } = [ line from PDB
                    # file ]
                    $AtomHash{$Atom[5]}{$PDBFileName}{$Atom[2]} = [@Atom];
                }
            }
        }
        close PDBFILE;
    }
}

$atomcounter = 0;
$ResidueSumOfSquares = 0;
$highRMSD = -1;
$lowRMSD = -1;
$residuecounter = 0;
$RMSDSum = 0;
#$ResNumcounter = 0;

# calculate RMSD across each residue for each pair of structures
for $ResNum ( sort keys %AtomHash ) {
    for $PDBFile1 ( sort keys %{ $AtomHash{$ResNum} } ) {
        for $PDBFile2 ( sort keys %{ $AtomHash{$ResNum} } ) {
            for $AtomType1 (sort keys %{ $AtomHash{$ResNum}{$PDBFile1} } ) {
                for $AtomType2 (sort keys %{ $AtomHash{$ResNum}{$PDBFile2} } ) {
                    if (($PDBFile1 ne $PDBFile2) && ($AtomType1 eq $AtomType2)) {
                        if (($AtomType1 eq "N") || ($AtomType1 eq "CA") || ($AtomType1 eq "C") ||
                            ($AtomType1 eq "O")) {

                            @Atom1 = @{ $AtomHash{$ResNum}{$PDBFile1}{$AtomType1} };
                            @Atom2 = @{ $AtomHash{$ResNum}{$PDBFile2}{$AtomType2} };

                            # get x, y, z for atom in first residue
                            $X1 = @Atom1[6];
                            $Y1 = @Atom1[7];
                            $Z1 = @Atom1[8];

                            # get x, y, z for atom in second residue
                            $X2 = @Atom2[6];
                            $Y2 = @Atom2[7];
                            $Z2 = @Atom2[8];

```

```

        $XDiff = $X2 - $X1;
        $YDiff = $Y2 - $Y1;
        $ZDiff = $Z2 - $Z1;
        $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
        $atomcounter += 1;
        $ResidueSumOfSquares += $SumOfSquares;
    }
}
}
}
# calculate RMSD
if ($atomcounter != 0) {
    $Mean = $ResidueSumOfSquares / $atomcounter;
    $RMSD = sqrt $Mean;
    print "Residue: $ResNum, PDB1: $PDBFile1, PDB2: $PDBFile2\n";
    print "RMSD: $RMSD\n";

    # add to sum, increment residue counter
    $RMSDSum += $RMSD;
    $residuecounter += 1;

    # save high value
    if (($highRMSD == -1) || ($RMSD > $highRMSD)) {
        $highRMSD = $RMSD;
    }

    # save low value
    if (($lowRMSD == -1) || ($RMSD < $lowRMSD)) {
        $lowRMSD = $RMSD;
    }

    $ResidueSumOfSquares = 0;
    $SumOfSquares = 0;
    $atomcounter = 0;
}
}
}

if ($residuecounter != 0){
    # calculate average of all the RMSD's for the residue
    $AverageRMSD = $RMSDSum / $residuecounter;

    # store high, low, avg, and res# (in an array for sorting purposes)
    $ResidueRMSD[$ResNum][0] = $AverageRMSD;
    $ResidueRMSD[$ResNum][1] = $highRMSD;
    $ResidueRMSD[$ResNum][2] = $lowRMSD;

    # reset high, low, sum, counter
    $highRMSD = -1;
    $lowRMSD = -1;
    $residuecounter = 0;
    $RMSDSum = 0;
}
}
close LOG;

select RESULTS;
# output an excel formatted file with residue number, high, low, and average RMSD values
$i = 1;
print "Residue\tAvg RMSD\tHigh RMSD\tLow RMSD\n";
while ($i <= $#ResidueRMSD){
    printf "$i\t%.2f\t%.2f\t%.2f\n", $ResidueRMSD[$i][0], $ResidueRMSD[$i][1],
        $ResidueRMSD[$i][2];
    $i += 1;
}
}

```

```
# close files
close RESULTS;
```

SameAtomSolventLigandSuper.pl

This program identifies atom overlaps found between atoms in organic solvents and atoms in non-organic solvents, which are generally bound inhibitor molecules for RNase A. Atoms from lines with a HETATM tag that overlap the position of another HETATM with a distance of less than 1.0Å and are of a similar atom type (i.e. carbon and carbon, or oxygen and oxygen) are included on the list of results. Lines with an ATOM tag are ignored (this generally includes protein and water molecules). Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the structures must be superimposed. This program currently only recognizes the following solvents: ACN, DOX, DMF, DMS, HEZ, IOH, BIR, TBU, ETF, and TMA. For additional solvent recognition, the code will have to be edited. The final results will be written in a file called “SameAtomSolventLigandSuper.log”.

```
#!/usr/bin/perl
# *****
#
# SameAtomSolventLigandSuper.pl
# Author: Michelle Dechene
#
# Perl script to identify solvent superpositions with ligands
# in all PDB files in the current directory. Structures must be
# superimposed.
#
# Solvents must be identified by HETATM instead of ATOM
#
```

```

# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ###   ##.###   ##.###   ##.###   #.# #.#
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "SameAtomSolventLigandSuper.log";
select RESULTS;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for HETATMs
            if ($Line =~ m/^HETATM/){
                @Atom = split /\s+/, $Line;
                # SolventID is ResidueName + Chain ID + Residue #
                $SolventID = $Atom[3] . $Atom[4] . $Atom[5];
                $SolventHash{$SolventID}{@Atom[2]} = [@Atom];
            }
        }
        close PDBFILE;
    }
}

for $ResID1 ( sort keys %SolventHash ) {
    for $AtomName1 ( sort keys %{ $SolventHash{$ResID1} } ) {
        for $ResID2 ( sort keys %SolventHash ) {
            for $AtomName2 ( sort keys %{ $SolventHash{$ResID2} } ) {
                # Don't compare atoms from the same residue
                if ($ResID1 ne $ResID2){
                    @Atom1 = @{ $SolventHash{$ResID1}{$AtomName1} };
                    @Atom2 = @{ $SolventHash{$ResID2}{$AtomName2} };

                    # Calculate the distance between 2 atoms
                    $X1 = @Atom1[6];
                    $Y1 = @Atom1[7];
                    $Z1 = @Atom1[8];

                    $X2 = @Atom2[6];
                    $Y2 = @Atom2[7];
                    $Z2 = @Atom2[8];

                    $XDiff = $X2 - $X1;
                    $YDiff = $Y2 - $Y1;
                    $ZDiff = $Z2 - $Z1;
                    $SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
                    $Distance = sqrt $SumOfSquares;

                    # If the atom centers are closer than 1.0 Angstroms, they are stored with their
                    # distances
                    if ($Distance <= 1.0) {
                        if (

```

```

( ( ($ResID1 =~ m/^ACN/) || ($ResID1 =~ m/^DOX/) || ($ResID1 =~ m/^DMF/)
  || ($ResID1 =~ m/^DMS/) || ($ResID1 =~ m/^HEZ/) || ($ResID1 =~ m/^IOH/)
  || ($ResID1 =~ m/^BIR/) || ($ResID1 =~ m/^TBU/) || ($ResID1 =~ m/^ETF/)
  || ($ResID1 =~ m/^TMA/) )
  &&
  ! ( ($ResID2 =~ m/^ACN/) || ($ResID2 =~ m/DOX/) || ($ResID2 =~ m/^DMF/)
    || ($ResID2 =~ m/^DMS/) || ($ResID2 =~ m/^HEZ/) || ($ResID2 =~ m/^IOH/)
    || ($ResID2 =~ m/^BIR/) || ($ResID2 =~ m/^TBU/) || ($ResID2 =~ m/^ETF/)
    || ($ResID2 =~ m/^TMA/) ) )
||
( ! ( ($ResID1 =~ m/^ACN/) || ($ResID1 =~ m/^DOX/) || ($ResID1 =~ m/^DMF/)
  || ($ResID1 =~ m/^DMS/) || ($ResID1 =~ m/^HEZ/) || ($ResID1 =~ m/^IOH/)
  || ($ResID1 =~ m/^BIR/) || ($ResID1 =~ m/^TBU/) || ($ResID1 =~ m/^ETF/)
  || ($ResID1 =~ m/^TMA/) )
  &&
  ( ($ResID2 =~ m/^ACN/) || ($ResID2 =~ m/DOX/) || ($ResID2 =~ m/^DMF/)
    || ($ResID2 =~ m/^DMS/) || ($ResID2 =~ m/^HEZ/) || ($ResID2 =~ m/^IOH/)
    || ($ResID2 =~ m/^BIR/) || ($ResID2 =~ m/^TBU/) || ($ResID2 =~ m/^ETF/)
    || ($ResID2 =~ m/^TMA/) ) )
) {

# Assign Atom to Atom 1
if ($AtomName1 =~ m/^C/){
  $Name1 = "C";
}
elseif ($AtomName1 =~ m/^N/){
  $Name1 = "N";
}
elseif ($AtomName1 =~ m/^O/){
  $Name1 = "O";
}
elseif ($AtomName1 =~ m/^S/){
  $Name1 = "S";
}
elseif ($AtomName1 =~ m/^P/){
  $Name1 = "P";
}
elseif ($AtomName1 =~ m/^F/){
  $Name1 = "F";
}
elseif ($AtomName1 =~ m/^V/){
  $Name1 = "V";
}
}
elseif ($AtomName1 =~ m/^G/){
  if ($AtomName1 =~ m/^.C/){
    $Name1 = "C";
  }
  elseif ($AtomName1 =~ m/^.N/){
    $Name1 = "N";
  }
  elseif ($AtomName1 =~ m/^.O/){
    $Name1 = "O";
  }
  elseif ($AtomName1 =~ m/^.S/){
    $Name1 = "S";
  }
  elseif ($AtomName1 =~ m/^.P/){
    $Name1 = "P";
  }
  elseif ($AtomName1 =~ m/^.F/){
    $Name1 = "F";
  }
  elseif ($AtomName1 =~ m/^.V/){
    $Name1 = "V";
  }
}
}

```

```

else {
    $Name1 = "-1";
}

# Assign Atom to Atom 2
if ($AtomName2 =~ m/^C/){
    $Name2 = "C";
}
elseif ($AtomName2 =~ m/^N/){
    $Name2 = "N";
}
elseif ($AtomName2 =~ m/^O/){
    $Name2 = "O";
}
elseif ($AtomName2 =~ m/^S/){
    $Name2 = "S";
}
elseif ($AtomName2 =~ m/^P/){
    $Name2 = "P";
}
elseif ($AtomName2 =~ m/^F/){
    $Name2 = "F";
}
elseif ($AtomName2 =~ m/^V/){
    $Name2 = "V";
}
}
elseif ($AtomName2 =~ m/^G/){
    if ($AtomName2 =~ m/^.C/){
        $Name2 = "C";
    }
    elseif ($AtomName2 =~ m/^.N/){
        $Name2 = "N";
    }
    elseif ($AtomName2 =~ m/^.O/){
        $Name2 = "O";
    }
    elseif ($AtomName2 =~ m/^.S/){
        $Name2 = "S";
    }
    elseif ($AtomName2 =~ m/^.P/){
        $Name2 = "P";
    }
    elseif ($AtomName2 =~ m/^.F/){
        $Name2 = "F";
    }
    elseif ($AtomName2 =~ m/^.V/){
        $Name2 = "V";
    }
}
}
else {
    $Name2 = "-2";
}

if ($Name1 eq $Name2){
    # AtomKeys store the keys to %SolventHash so that data can be accessed
    $Atom1Key = $ResID1 . ":" . $AtomName1;
    $Atom2Key = $ResID2 . ":" . $AtomName2;
    if ( ($ResID2 =~ m/^ACN/) || ($ResID2 =~ m/^DOX/) || ($ResID2 =~ m/^DMF/)
        || ($ResID2 =~ m/^DMS/) || ($ResID2 =~ m/^HEZ/) || ($ResID2 =~ m/^IOH/)
        || ($ResID2 =~ m/^BIR/) || ($ResID2 =~ m/^TBU/) || ($ResID2 =~ m/^ETF/)
        || ($ResID2 =~ m/^TMA/) ) {
        $TempKey = $Atom1Key;
        $Atom1Key = $Atom2Key;
        $Atom2Key = $TempKey;
    }
    $AtomOverlap{$Atom1Key}{$Atom2Key} = $Distance;
}
}

```

```

    }
  }
}

for $AtomID1 ( sort keys %AtomOverlap ) {
  for $AtomID2 ( sort keys %{ $AtomOverlap{$AtomID1} } ) {
    printf "Distance between $AtomID1 and $AtomID2: %2.3f\n" ,
$AtomOverlap{$AtomID1}{$AtomID2};
  }
}

close RESULTS;

```

SameAtomSolventSuperPos.pl

This program identifies atom overlaps found between atoms in organic solvents. Atoms from lines with a HETATM tag that overlap the position of another HETATM with a distance of less than 1.0Å and are of a similar atom type (i.e. carbon and carbon, or oxygen and oxygen) are included on the list of results. Lines with an ATOM tag are ignored (this generally includes protein and water molecules). Coordinates from the file are read into an array from each line splitting the values by spaces. The script uses array numbering assuming that there will be a defined chain for the protein.

All PDB files in the current folder will be used for the calculations and the structures must be superimposed. This program currently only recognizes the following solvents: ACN, DOX, DMF, DMS, HEZ, IOH, BIR, TBU, ETF, and TMA. For additional solvent recognition, the code will have to be edited. The final results will be written in a file called “SolventSuperPos1.log”.

```

#!/usr/bin/perl
# *****

```

```

#
# SameAtomSolventSuperPos.pl
# Author: Michelle Dechene
#
# Perl script to identify solvent superpositions (with other solvents)
# in all PDB files in the current directory. Structures must be
# superimposed.
#
# Solvents must be identified by HETATM instead of ATOM
#
# There is no error checking for Atom types, etc
#
# Assumes the following PDB format
# ATOM   ###   xxx xxx x   ###   ##.###   ##.###   ##.###   #.##   ##.##
# i.e. assumes that there is a defined chain between the residue name
# and the residue number
#
# *****

opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

# open output file and echo user input
open RESULTS, ">", "SolventSuperPos1.log";
select RESULTS;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        print "\nPDB File: $PDBFileName\n";
        open PDBFILE, "<", $PDBFileName;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Search for HETATMs
            if ($Line =~ m/^HETATM/){
                @Atom = split /\s+/, $Line;
                # SolventID is ResidueName + Chain ID + Residue #
                $SolventID = $Atom[3] . $Atom[4] . $Atom[5];
                $SolventHash{$SolventID}{@Atom[2]} = [@Atom];
            }
        }
        close PDBFILE;
    }
}

for $ResID1 ( sort keys %SolventHash ) {
    for $AtomName1 ( sort keys %{ $SolventHash{$ResID1} } ) {
        for $ResID2 ( sort keys %SolventHash ) {
            for $AtomName2 ( sort keys %{ $SolventHash{$ResID2} } ) {
                # Don't compare atoms from the same residue
                if ($ResID1 ne $ResID2){
                    @Atom1 = @{ $SolventHash{$ResID1}{$AtomName1} };
                    @Atom2 = @{ $SolventHash{$ResID2}{$AtomName2} };

                    # Calculate the distance between 2 atoms
                    $X1 = @Atom1[6];
                    $Y1 = @Atom1[7];
                    $Z1 = @Atom1[8];

                    $X2 = @Atom2[6];
                    $Y2 = @Atom2[7];
                    $Z2 = @Atom2[8];
                }
            }
        }
    }
}

```

```

$XDiff = $X2 - $X1;
$YDiff = $Y2 - $Y1;
$ZDiff = $Z2 - $Z1;
$SumOfSquares = $XDiff**2 + $YDiff**2 + $ZDiff**2;
$Distance = sqrt $SumOfSquares;

# If the atom centers are closer than 1.0 Angstroms, they are stored with their
# distances
if ($Distance <= 1.0) {
  if ( ($ResID1 =~ m/^ACN/) || ($ResID1 =~ m/^DOX/) || ($ResID1 =~ m/^DMF/)
    || ($ResID1 =~ m/^DMS/) || ($ResID1 =~ m/^HEZ/) || ($ResID1 =~ m/^IOH/)
    || ($ResID1 =~ m/^BIR/) || ($ResID1 =~ m/^TBU/) || ($ResID1 =~ m/^ETF/)
    || ($ResID1 =~ m/^TMA/)
    || ($ResID2 =~ m/^ACN/) || ($ResID2 =~ m/^DOX/) || ($ResID2 =~ m/^DMF/)
    || ($ResID2 =~ m/^DMS/) || ($ResID2 =~ m/^HEZ/) || ($ResID2 =~ m/^IOH/)
    || ($ResID2 =~ m/^BIR/) || ($ResID2 =~ m/^TBU/) || ($ResID2 =~ m/^ETF/)
    || ($ResID2 =~ m/^TMA/) ) {

    # Assign Atom to Atom 1
    if ($AtomName1 =~ m/^C/){
      $Name1 = "C";
    }
    elsif ($AtomName1 =~ m/^N/){
      $Name1 = "N";
    }
    elsif ($AtomName1 =~ m/^O/){
      $Name1 = "O";
    }
    elsif ($AtomName1 =~ m/^S/){
      $Name1 = "S";
    }
    elsif ($AtomName1 =~ m/^P/){
      $Name1 = "P";
    }
    elsif ($AtomName1 =~ m/^F/){
      $Name1 = "F";
    }
    elsif ($AtomName1 =~ m/^V/){
      $Name1 = "V";
    }
    elsif ($AtomName1 =~ m/^G/){
      if ($AtomName1 =~ m/^.C/){
        $Name1 = "C";
      }
      elsif ($AtomName1 =~ m/^.N/){
        $Name1 = "N";
      }
      elsif ($AtomName1 =~ m/^.O/){
        $Name1 = "O";
      }
      elsif ($AtomName1 =~ m/^.S/){
        $Name1 = "S";
      }
      elsif ($AtomName1 =~ m/^.P/){
        $Name1 = "P";
      }
      elsif ($AtomName1 =~ m/^.F/){
        $Name1 = "F";
      }
      elsif ($AtomName1 =~ m/^.V/){
        $Name1 = "V";
      }
    }
  }
  else {
    $Name1 = "-1";
  }
}

```



```

        printf "Distance between $AtomID1 and $AtomID2: %2.3f\n" ,
$AtomOverlap{$AtomID1}{$AtomID2};
    }
}

close RESULTS;

```

SortAtomDist.pl

This program sorts the distances from the output file generated by AtomDist.pl. The user is asked to input the filename of the file to sort, and the distances are then sorted into the low third, the middle third, and the high third.

Results will be written to a file called "SortAtomDist.log".

```

#!/usr/bin/perl

# *****
#
# SortAtomDist.pl
# Author: Michelle Dechene
#
# Perl script to sort the output from the AtomDist programs into three
# groups.
#
# *****

# user input for filenames
print "Enter Filename to sort: ";
chomp($FileName = <STDIN>);

# Read file into array
open FILETOSORT, "<", $FileName;
chomp(@FileLines = <FILETOSORT>);

# open output file and echo user input
open RESULTS, ">", "SortAtomDist.log";
select RESULTS;

$high = -1;
$low = -1;

foreach $Line (@FileLines) {
    # Read in PDB filename, Chain Name, and Calculated distance
    if ($Line =~ m/^\bPDB/){
        $PDBFile = $Line;
    }
    if ($Line =~ m/^\bChain/){
        $Chain = $Line;
    }
    if ($Line =~ m/^\bDistance\b/){
        @Distance = split /\s+/, $Line;
    }
}

```

```

# Store high and low distance values
if ($high == -1){
    $high = $Distance[3];
}
if ($low == -1){
    $low = $Distance[3];
}
if ($high < $Distance[3]){
    $high = $Distance[3];
}
if ($low > $Distance[3]){
    $low = $Distance[3];
}

}

# Store values in a hash
if ($PDBFile && $Chain && @Distance){
    $AtomDistances{$Distance[3]} .= $PDBFile . ", " . $Chain . "; ";
    @Distance = ();
    $Chain = undef;
}
}

# Calculate the borders of the distance range for a high, medium, and low
$span = $high - $low;
$thirds = $span/3;
$lowmiddle = $low + $thirds;
$highmiddle = $high - $thirds;

# Using calculated borders, sort into high, medium, and low values and print
# filename, chain name, and distance
print "Lowest distance is $low and Highest distance is $high\n\n";
print "Molecules with a measure in the lower third ($low to $lowmiddle Angstroms):\n\n";

$middle_done = undef;
$high_done = undef;

foreach $DistanceValue (sort keys %AtomDistances){
    if (!($middle_done) && ($DistanceValue > $lowmiddle) && ($DistanceValue <= $highmiddle)){
        print "\nMolecules with a measure in the middle third ($lowmiddle to $highmiddle
            Angstroms):\n\n";
        $middle_done = 1;
    }
    elseif (!($high_done) && ($DistanceValue > $highmiddle)){
        print "\nMolecules with a measure in the high third ($highmiddle to $high
            Angstroms):\n\n";
        $high_done = 1;
    }
    print "Distance $DistanceValue: $AtomDistances{$DistanceValue}\n";
}

# print ranges and data points in each range

#*****
# This would be more useful if all the data points were sorted by distance!!! perhaps use a
# hash with a key of distance?
#*****

close FILETOSORT;

close RESULTS;

```

SplitFiles.pl

This program splits PDB files into files containing protein, water, and ligands. By parsing through the PDB file line by line, each line is then sorted into the appropriate file. Lines with ATOM tag and amino acids for the residue name will be sorted into a protein file. Lines with either the ATOM or HETATM tag, and a residue name of "HOH" will be sorted into a water file. All remaining lines with a HETATM tag are sorted into a ligand file. Split files are saved in a new directory called "Split_PDB" and filenames have a suffix of "-protein.pdb", "-water.pdb", or "-ligand.pdb" appended to the original filename as appropriate.

All PDB files in the current folder will be split and saved in the new directory. This program assumes that the residue name for water molecules is called "HOH" in the PDB file.

```
#!/usr/bin/perl
# *****
#
# SplitFiles.pl
# Author: Michelle Dechene
#
# Perl script to split PDB files into protein, water, and ligand.
# This was done for ease of manipulation in Pymol.
#
# This script works on all PDB files in the current directory.
#
# Old files are left alone and new split files are written in a new
# directory called Split_PDB
#
# *****

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

mkdir "Split_PDB", 0755;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        open PDBFILE, "<", $PDBFileName;
        @FileName = split /\./, $PDBFileName;

        open PDBPROTEIN, ">", "Split_PDB/" . $FileName[0] . "-protein.pdb";
```

```

open PDBWATER, ">", "Split_PDB/" . $FileName[0] . "-water.pdb";
open PDBLIGAND, ">", "Split_PDB/" . $FileName[0] . "-ligand.pdb";

# Read in PDB file into array
chomp(@FileLines = <PDBFILE>);

foreach $Line (@FileLines) {
    # Locate lines of non-protein and non-water atoms
    if (($Line =~ m/^ATOM/)&&((substr($Line,17,3) eq "ALA")|| (substr($Line,17,3) eq
        "CYS")|| (substr($Line,17,3) eq "ASP")|| (substr($Line,17,3) eq "GLU")||
        (substr($Line,17,3) eq "PHE")|| (substr($Line,17,3) eq "GLY")||
        (substr($Line,17,3) eq "HIS")|| (substr($Line,17,3) eq "ILE")||
        (substr($Line,17,3) eq "LYS")|| (substr($Line,17,3) eq "LEU")||
        (substr($Line,17,3) eq "MET")|| (substr($Line,17,3) eq "ASN")||
        (substr($Line,17,3) eq "PRO")|| (substr($Line,17,3) eq "GLN")||
        (substr($Line,17,3) eq "ARG")|| (substr($Line,17,3) eq "SER")||
        (substr($Line,17,3) eq "THR")|| (substr($Line,17,3) eq "VAL")||
        (substr($Line,17,3) eq "TRP")|| (substr($Line,17,3) eq "TYR"))){
        select PDBPROTEIN;
        print "$Line\n";
    }
    elsif (((($Line =~ m/^ATOM/)|| ($Line =~ m/^HETATM/))&&(substr($Line,17,3) eq "HOH"))){
        select PDBWATER;
        print "$Line\n";
    }
    elsif ($Line =~ m/^HETATM/){
        select PDBLIGAND;
        print "$Line\n";
    }
}
select PDBPROTEIN;
print "END\n";
select PDBWATER;
print "END\n";
select PDBLIGAND;
print "END\n";

close PDBPROTEIN;
close PDBWATER;
close PDBLIGAND;
close PDBFILE;
}

```

SplitFilesA.pl

This program removes molecule B from PDB files. Edited files are saved in a new directory called “FirstCutA” and filenames have a suffix of “-A.pdb”.

All PDB files in the current folder will be edited and saved in the new directory. If no chain B is present, the entire file will be copied over to the new location.

```

#!/usr/bin/perl

# *****
#
# SplitFilesA.pl
# Author: Michelle Dechene
#
# Perl script to remove protein molecule B from the PDB file.
#
# This script works on all PDB files in the current directory.
#
# Old files are left alone and new split files are written in a new
# directory called FirstCutA
#
# *****

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

mkdir "FirstCutA", 0755;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        open PDBFILE, "<", $PDBFileName;
        @FileName = split /\./, $PDBFileName;

        open PDBA, ">", "FirstCutA/" . $FileName[0] . "-A.pdb";
        select PDBA;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Locate lines of non-protein and non-water atoms
            if (((!($Line =~ m/^ATOM/)) || (substr($Line,21,1) ne "B")) && ((!($Line =~ m/^TER/))
                || (substr($Line,21,1) ne "B"))){
                print "$Line\n";
            }
        }
        close PDBA;
        close PDBFILE;
    }
}

```

SplitFilesB.pl

This program removes molecule A from PDB files. Edited files are saved in a new directory called “FirstCutB” and filenames have a suffix of “-B.pdb”.

All PDB files in the current folder will be edited and saved in the new directory. If no chain A is present, the entire file will be copied over to the new location.

```
#!/usr/bin/perl

# *****
#
# SplitFilesB.pl
# Author: Michelle Dechene
#
# Perl script to remove protein molecule A from the PDB file.
#
# This script works on all PDB files in the current directory.
#
# Old files are left alone and new split files are written in a new
# directory called FirstCutB
#
# *****

# get all file names from the current directory
opendir MYDIR, ".";
@files = readdir MYDIR;
closedir MYDIR;

mkdir "FirstCutB", 0755;

foreach $PDBFileName (@files) {
    if ($PDBFileName =~ m/pdb$/){
        open PDBFILE, "<", $PDBFileName;
        @FileName = split /\./, $PDBFileName;

        open PDBB, ">", "FirstCutB/" . $FileName[0] . "-B.pdb";
        select PDBB;

        # Read in PDB file into array
        chomp(@FileLines = <PDBFILE>);

        foreach $Line (@FileLines) {
            # Locate lines of non-protein and non-water atoms
            if (((!($Line =~ m/^ATOM/)) || (substr($Line,21,1) ne "A")) && (!($Line =~ m/^TER/))
                || (substr($Line,21,1) ne "A")){
                print "$Line\n";
            }
        }
        close PDBB;
        close PDBFILE;
    }
}
}
```

APPENDIX B: Downloaded PDB Files

Table 1. Downloaded PDB files

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1AFK	C 1 2 1	DD Leonidas, R Shapiro, LI Irons, N Russo, KR Acharya	Crystal structures of ribonuclease A complexes with 5'-diphosphoadenosine 3'-phosphate and 5'-diphosphoadenosine 2'-phosphate at 1.7 Å Resolution	Biochemistry (1997) 36:5578-5588	A	162	0.211	0.266	1.7	3'-Phosphate-adenosine-5'-diphosphate	P1, B2, P2	Inhibitor	
1AFL	C 1 2 1	DD Leonidas, R Shapiro, LI Irons, N Russo, KR Acharya	Crystal structures of ribonuclease A complexes with 5'-diphosphoadenosine 3'-phosphate and 5'-diphosphoadenosine 2'-phosphate at 1.7 Å Resolution	Biochemistry (1997) 36:5578-5588	A	122	0.217	0.278	1.7	2'-Monophospho adenosine-5'-diphosphate	P1, B2	Inhibitor	Citric Acid
1EOS	C 1 2 1	L Vitagliano, A Merlino, A Zagari, L Mazzarella	Productive and nonproductive binding to ribonuclease A: x-ray structure of two complexes with uridylyl(2',5')guanosine	Protein Science (2000) 9:1217-1225	A	97	0.178	n/a	2.0	Uridylyl-2'-5'-phospho-guanosine	B1, P1 (molecule A only)	Inhibitor	
1EOW	P 1 21 1	L Vitagliano, A Merlino, A Zagari, L Mazzarella	Productive and nonproductive binding to ribonuclease A: x-ray structure of two complexes with uridylyl(2',5')guanosine	Protein Science (2000) 9:1217-1225	A	72	0.187	n/a	2.0	Uridylyl-2'-5'-phospho-guanosine	B1 (retro-binding)	Inhibitor	Sulfate Ion (active site)

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1FS3	P 32 2 1	E Chatani, R Hayashi, H Moriyama, T Ueki	Conformational strictness required for maximum activity and stability of bovine pancreatic ribonuclease A as revealed by crystallographic study of three Phe120 mutants at 1.4 Å resolution	Protein Science (2002) 11: 72-81	A	115	0.217	0.255	1.4	No		P3(2)21	
1JN4	C 1 2 1	AM Jardine, DD Leonidas, JL Jenkins, C Park, RT Raines, KR Acharya, R Shapiro	Cleavage of 3',5'-pyrophosphate-linked dinucleotides by ribonuclease A and angiogenin	Biochemistry (2001) 40:10262-10272	A	265	0.225	0.295	1.8	Adenosine-5'-[trihydrogen diphosphate] P ⁻ -3'-ester with 2'-deoxyuridine	B1, P1, B2 (molecule A only)	Inhibitor	
1JVU	C 1 2 1	L Vitagliano, A Merlino, A Zagari, L Mazzarella	Reversible substrate-induced domain motions in ribonuclease A	Proteins: Structure, Function, and Genetics (2002) 46:97-104	A	84	0.190	0.240	1.78	Cytidine-2'-monophosphate	B1, P1 (molecule A only)	Inhibitor	
1O0F	C 1 2 1	DD Leonidas, GB Chavali, NG Oikonomakos, ED Chrysina, MN Kosmopoulou, M Vlasi, C Frankling, KR Acharya	High-resolution crystal structures of ribonuclease A complexed with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors	Protein Science (2003) 12:2559-2574	A	389	0.229	0.271	1.5	Adenosine-3'-5'-diphosphate	P1, B2, P2	Inhibitor	

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
100H	C 1 2 1	DD Leonidas, GB Chavali, NG Oikonomakos, ED Chrysina, MN Kosmopoulou, M Vlassi, C Frankling, KR Acharya	High-resolution crystal structures of ribonuclease A complexed with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors	Protein Science (2003) 12:2559-2574	A	541	0.190	0.221	1.2	Adenosine-5'-diphosphate	P1, B2	Inhibitor	
100M	C 1 2 1	DD Leonidas, GB Chavali, NG Oikonomakos, ED Chrysina, MN Kosmopoulou, M Vlassi, C Frankling, KR Acharya	High-resolution crystal structures of ribonuclease A complexed with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors	Protein Science (2003) 12:2559-2574	A	399	0.211	0.244	1.5	Phosphoric acid mono-[2-(2,4-dioxo-3,4-dihydro-2H-pyrimidin-1-yl)-4-hydroxy-5-hydroxymethyl-tetrahydro-furan-3-yl] ester	B1, P1	Inhibitor	
100N	C 1 2 1	DD Leonidas, GB Chavali, NG Oikonomakos, ED Chrysina, MN Kosmopoulou, M Vlassi, C Frankling, KR Acharya	High-resolution crystal structures of ribonuclease A complexed with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors	Protein Science (2003) 12:2559-2574	A	415	0.220	0.24	1.5	3'-Uridine-monophosphate	B1, P1	Inhibitor	

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1000	C 1 2 1	DD Leonidas, GB Chavali, NG Oikonomakos, ED Chrysina, MN Kosmopoulou, M Vlassi, C Frankling, KR Acharya	High-resolution crystal structures of ribonuclease A complexed with adenylic and uridylic nucleotide inhibitors. Implications for structure-based design of ribonucleolytic inhibitors	Protein Science (2003) 12:2559-2574	B A	558	0.19	0.23	1.2	Adenosine-2'-5'-diphosphate	P1 (molecule A) P1, B2 (molecule B)	Inhibitor	
1QHC	C 1 2 1	DD Leonidas, R Shapiro, LI Irons, N Russo, KR Acharya	Toward rational design of ribonuclease inhibitors: high-resolution crystal structure of a ribonuclease A complex with a potent 3',5'-pyrophosphate-linked dinucleotide inhibitor	Biochemistry (1999) 38:10287-10297	A	137	0.200	0.260	1.7	Adenylate-3'-phosphate-[[2'-deoxy-uridine-5'-phosphate]-3'-phosphate]	P0, B1, P1, B2, P2	Inhibitor	
1RAR	P 3 2 2 1	S Baudet-Nessler, M Jullien, MP Crosio, J Janin	Crystal structure of a fluorescent derivative of RNase A	Biochemistry (1993) 32:8457-8464	A	104	0.172	n/a	1.9	5-(1-Sulfonaphthyl)-acetylamine-ethylamine	P1	Inhibitor P3(2)21	3 Chloride Ions
1RAS	P 3 2 2 1	S Baudet-Nessler, M Jullien, MP Crosio, J Janin	Crystal structure of a fluorescent derivative of RNase A	Biochemistry (1993) 32:8457-8464	A	100	0.203	n/a	1.7	5-(1-Sulfonaphthyl)-acetylamine-ethylamine	P1	Inhibitor P3(2)21	
1RBJ	P 4 1 2 1 2	TP Ko, R Williams, A McPherson	Structure of a ribonuclease B+d(pA) ₄ complex	Acta Crystallographica Section D (1996) 52:160-164	A	0	0.163	n/a	2.7	DNA (5'-D(*AP*AP*AP*A)-3')	P0, B1, P1, B2, P2	Inhibitor	
1RCA	P 1 2 1 1	JN Lisgarten, D Maes, L Wyns, CF Aguilar, RA Palmer	Structure of the crystalline complex of deoxycytidylyl-3',5'-guanosine (3',5'-dCpdG) cocrystallized with ribonuclease at 1.9 Å resolution	Acta Crystallographica Section D (1995) 51:767-771	A	58	0.218	n/a	1.9	2'-Deoxycytidine-2'-deoxyguanosine-3',5'-monophosphate	B1, P1 (retro-binding)	Inhibitor	Phosphate Ion

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1RCN	P 21 21 21	JC Fontecilla-Camps, R de Llorens, MH le Du, CM Cuchillo	Crystal structure of ribonuclease A d(ApTpApApG) complex. Direct evidence for extended substrate recognition	Journal of Biological Chemistry (1994) 269:21526-21531	A	68	0.172	n/a	2.32	DNA (5'-D(*AP*TP*AP*A)-3')	P0, B1, P1, B2, P2	Inhibitor	
1RHA	P 1 21 1	RV Kishan, NR Chandra, C Sudarsanakumar, K Suguna, M Vijayan	Water-dependent domain motion and flexibility in ribonuclease A and the invariant features in its hydration shell. An x-ray study of two low-humidity crystal forms of the enzyme.	Acta Crystallographica Section D (1995) 51:703-710	A	145	0.176	n/a	1.8	No		Sadasivan	
1RHB	P 1 21 1	RV Kishan, NR Chandra, C Sudarsanakumar, K Suguna, M Vijayan	Water-dependent domain motion and flexibility in ribonuclease A and the invariant features in its hydration shell. An x-ray study of two low-humidity crystal forms of the enzyme.	Acta Crystallographica Section D (1995) 51:703-710	A	160	0.173	n/a	1.5	No		Sadasivan	
1RNC	P 1 21 1	CF Aguilar, PJ Thomas, A Mills, DS Moss, RA Palmer	Newly observed binding mode in pancreatic ribonuclease	Journal of Molecular Biology (1992) 224:265-267	A	30	0.210	n/a	1.5	Guanosine-5'-monophosphate	B1 (retro-binding)	Inhibitor	Sulfate Ion (active site)
1RND	P 1 21 1	CF Aguilar, PJ Thomas, A Mills, DS Moss, RA Palmer	Newly observed binding mode in pancreatic ribonuclease	Journal of Molecular Biology (1992) 224:265-267	A	96	0.190	n/a	1.5	2'-Deoxyguanosine-5'-monophosphate	B1 (retro-binding)	Inhibitor	Sulfate Ion (active site)

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1RNM	P 32 2 1	AA Federov, D Josef-Mccarthy, I Graf, D Anguelova, EV Federov, SC Almo	Structure of the crystalline complex of bovine pancreatic ribonuclease A and cytidylic acid	To be published (released 1996)	A	80	0.158	n/a	2.0	Cytidine-5'-monophosphate	P0, B1	Inhibitor P3(2)21	2 Sulfate Ions (1 in active site)
1RNN	P 32 2 1	AA Federov, D Josef-Mccarthy, I Graf, D Anguelova, EV Federov, SC Almo	Structure of the crystalline complex of bovine pancreatic ribonuclease A and cytidylic acid	To be published (released 1996)	A	106	0.170	n/a	1.8	Cytidine-5'-monophosphate	P0, B1	Inhibitor P3(2)21	Formic Acid (active site)
1RNO	P 32 2 1	AA Federov, D Joseph-McCarthy, E Fedorov, D Sirakova, I Graf, SC Almo	Ionic interactions in crystalline bovine pancreatic ribonuclease A	Biochemistry (1996) 35: 15962-15979	B	112	0.165	n/a	1.9	No		P3(2)21	Sulfate Ion (active site), Sulfate Ion (Ala 4)
1RNQ	P 32 2 1	AA Federov, D Joseph-McCarthy, E Fedorov, D Sirakova, I Graf, SC Almo	Ionic interactions in crystalline bovine pancreatic ribonuclease A	Biochemistry (1996) 35: 15962-15979	A	95	0.161	n/a	2.0	No		P3(2)21	Formic Acid (active site, B1, P1)
1RNW	P 32 2 1	AA Federov, D Joseph-McCarthy, E Fedorov, D Sirakova, I Graf, SC Almo	Ionic interactions in crystalline bovine pancreatic ribonuclease A	Biochemistry (1996) 35: 15962-15979	B	110	0.167	n/a	1.8	No		P3(2)21	Sulfate Ion (active site), Sulfate Ion (Ala 4)

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1RNX	P 32 2 1	AA Federov, D Joseph-McCarthy, E Fedorov, D Sirakova, I Graf, SC Almo	Ionic interactions in crystalline bovine pancreatic ribonuclease A	Biochemistry (1996) 35: 15962-15979	A	103	0.162	n/a	1.9	No		P3(2)21	4 Chloride ions: 2 active site, 2 outside
1RNY	P 32 2 1	AA Federov, D Joseph-McCarthy, E Fedorov, D Sirakova, I Graf, SC Almo	Ionic interactions in crystalline bovine pancreatic ribonuclease A	Biochemistry (1996) 35: 15962-15979	A	99	0.175	n/a	2.0	No		P3(2)21	5 Chloride ions: 2 active site, 3 outside
1RNZ	P 32 2 1	AA Federov, D Joseph-McCarthy, E Fedorov, D Sirakova, I Graf, SC Almo	Ionic interactions in crystalline bovine pancreatic ribonuclease A	Biochemistry (1996) 35: 15962-15979	A	95	0.167	n/a	1.9	No		P3(2)21	3 Chloride ions: 1 active site, 2 outside
1ROB	P 1 21 1	JN Lisgarten, V Gupta, D Maes, L Wyns	Structure of the crystalline complex of cytidylic acid (2'-CMP) with ribonuclease at 1.6 Å resolution. Conservation of solvent sites in RNase-A high-resolution structures	Acta Crystallographica Section D (1993) 49:541-547	A	101	0.170	n/a	1.6	Cytidine-2'-monophosphate	B1, P1	Inhibitor Zegers	
1RPF	P 32 2 1	I Zegers, D Maes, M-H Dao-Thi, F Poortmans, R Palmer, L Wyns	The structures of RNase A complexed with 3'-CMP and d(CpA): Active site conformation and conserved water molecules	Protein Science (1994) 3:2322-2339	B	115	0.155	n/a	2.2	Cytidine-3'-monophosphate	B1, P1	Inhibitor Zegers P3(2)21	

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1RPG	P 1 21 1	I Zegers, D Maes, M-H Dao-Thi, F Poortmans, R Palmer, L Wyns	The structures of RNase A complexed with 3'-CMP and d(CpA): Active site conformation and conserved water molecules	Protein Science (1994) 3:2322-2339	A	172	0.172	n/a	1.4	2'-Deoxycytidine-2'-deoxyadenosine-3',5'-monophosphate	B1, P1, B2	Inhibitor Zegers	(4S)-2-Methyl-2,4-pentanediol
1RPH	P 32 2 1	I Zegers, D Maes, M-H Dao-Thi, F Poortmans, R Palmer, L Wyns	The structures of RNase A complexed with 3'-CMP and d(CpA): Active site conformation and conserved water molecules	Protein Science (1994) 3:2322-2339	A&B	99	0.158	n/a	2.2	No		Sadasivan Zegers P3(2)21	Sulfate Ion (active site)
1RSM	P 21 21 21	PC Weber, S Sheriff, DH Ohlendorf, BC Finzel, FR Salemme	The 2-Å resolution structure of a thermostable ribonuclease A chemically cross-linked between lysine residues 7 and 41	Proceedings of the National Academy of Sciences USA (1985) 82:8473-8477	A	75	0.184	n/a	2.0	Dinitrophenylene		Inhibitor Zegers	
1RTA	P 21 21 21	DL Birdsall, A McPherson	Crystal structure disposition of thymidylic acid tetramer in complex with ribonuclease A	Journal of Biological Chemistry (1992) 267:22230-22236	A	0	0.235	n/a	2.5	DNA (5'-D(*TP*TP*TP*T)-3')	B1, P1, B2	Inhibitor	
1RUV	P 1 21 1	JE Ladner, BD Wladkowski, LA Svensson, L Sjölin, GL Gilliland	X-ray structure of a ribonuclease A-uridine vanadate complex at 1.3 Å resolution	Acta Crystallographica Section D (1997) 53:290-301	A	131	0.197	n/a	1.25	Uridine-2',3'-vanadate	B1, P1	Inhibitor	Tertiary-butyl alcohol
1U1B	P 1 21 1	H Beach, R Cole, ML Gill, JP Loria	Conservation of μ s-ms enzyme motions in the apo- and substrate-mimicked state	Journal of the American Chemical Society (2005) 127:9167-9176	A	72	0.218	0.258	2.0	5'-Phosphothymidine(3'-5')-pyrophosphate adenosine 3'-phosphate	P0, B1, P1, B2, P2	Inhibitor	

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1W4O	C 1 2 1	CL Jenkins, N Thiyagarajan, RY Sweeney, MP Guy, BR Kelemen, KR Acharya, RT Raines	Binding of non-natural 3'nucleotides to ribonuclease A	FEBS Journal (2005) 272:744-755	A	200	0.232	0.248	1.6	Uracil arabinose-3'-phosphate	B1, P1 (molecule A only)	Inhibitor	
1W4P	C 1 2 1	CL Jenkins, N Thiyagarajan, RY Sweeney, MP Guy, BR Kelemen, KR Acharya, RT Raines	Binding of non-natural 3'nucleotides to ribonuclease A	FEBS Journal (2005) 272:744-755	A	290	0.217	0.226	1.69	2'-Deoxyuridine 3'-monophosphate	B1, P1	Inhibitor	
1W4Q	C 1 2 1	CL Jenkins, N Thiyagarajan, RY Sweeney, MP Guy, BR Kelemen, KR Acharya, RT Raines	Binding of non-natural 3'nucleotides to ribonuclease A	FEBS Journal (2005) 272:744-755	A B	308	0.209	0.241	1.68	2'-Fluoro-2'-deoxyuridine 3'-monophosphate	B1, P1	Inhibitor	
1WBU	C 1 2 1	MJ Hartshorn, CW Murray, A Cleasby, M Frederickson, IJ Tickle, H Jhoti	Fragment-based lead discovery using x-ray crystallography	Journal of Medicinal Chemistry (2005) 48:403-413	A B	222	0.207	0.270	1.9	5-Amino-1H-pyrimidine-2,4-dione	B1	Inhibitor	
1XPS	P 1 21 1	C Sadasivan, HG Nagendra, M Vijayan	Plasticity, hydration and accessibility in ribonuclease A. The structure of a new crystal form and its low-humidity variant.	Acta Crystallographica Section D (1998) 54:1343-1352	A	246	0.175	0.248	1.8	No		Sadasivan	

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
1XPT	P 1 21 1	C Sadasivan, HG Nagendra, M Vijayan	Plasticity, hydration and accessibility in ribonuclease A. The structure of a new crystal form and its low-humidity variant.	Acta Crystallographica Section D (1998) 54:1343-1352	A	280	0.162	0.222	1.9	No		Sadasivan	
1Z6D	C 1 2 1	GN Hatzopoulos, DD Leonidas, R Kardakaris, J Kobe, NG Oikonomakos	The binding of IMP to ribonuclease A	FEBS Journal (2005) 272:3988-4001	A	360	0.189	0.234	1.54	Inosinic Acid	B1 (retro-binding), P1, B2	Inhibitor	
1Z6S	C 1 2 1	GN Hatzopoulos, DD Leonidas, R Kardakaris, J Kobe, NG Oikonomakos	The binding of IMP to ribonuclease A	FEBS Journal (2005) 272:3988-4001	A	330	0.195	0.231	1.5	Adenosine monophosphate	P1, B2	Inhibitor	
2AAS	NMR	J Santoro, C Gonzalez, M Bruix, JL Neira, JL Nieto, J Herranz, M Rico	High-resolution three-dimensional structure of ribonuclease A in solution by nuclear magnetic resonance spectroscopy	Journal of Molecular Biology (1993) 229:722-734		0	n/a	n/a		No		NMR	
2BLP	P 32 2 1	MH Nanao, GM Sheldrick, RB Ravelli	Improving radiation-damage substructures for RIP	Acta Crystallographica, Section D (2005) 61:1227-1237	B	143	0.147	0.176	1.4	No		P3(2)21	Chloride Ion (active site)
2BLZ	P 32 2 1	MH Nanao, GM Sheldrick, RB Ravelli	Improving radiation-damage substructures for RIP	Acta Crystallographica, Section D (2005) 61:1227-1237	B	143	0.151	0.181	1.4	No		P3(2)21	Chloride Ion (active site)

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
2G4X	P 32 2 1	C Mueller-Dieckmann, S Panjikar, A Schmidt, S Mueller, J Kuper, A Geerlof, M Wilmanns, RK Singh, PA Tucker, MS Weiss	On the routine use of soft X-rays in macromolecular crystallography. Part IV. Efficient determination of anomalous substructures in biomacromolecules using longer X-ray wavelengths.	Acta Crystallographica, Section D (2007) 63:366-380	B	88	0.176	0.224	1.95	No		P3(2)21	5 Chloride ions: 1 active site, 4 outside, Sulfate Ion (Ala 4)
2G8R	C 1 2 1	DD Leonidas, TK Maiti, A Samanta, S Dasgupta, T Pathak, SE Zographos, NG Oikonomakos	The binding of 3'- <i>N</i> -piperidine-4-carboxyl-3'-deoxy- <i>ara</i> -uridine to ribonuclease A in the crystal	Bioorganic & Medicinal Chemistry (2006) 14:6055-6064	A A&B	262	0.193	0.230	1.7	1-[3-(4-Carboxypiperidin-1-yl)-3-deoxy-beta-D-arabinofuranosyl]pyrimidine-2,4(1H,3H)-dione	P0, B1, B2 (molecule A only)	Inhibitor	
2RNS	P 31 2 1	EE Kim, R Varadarajan, HW Wyckoff, FM Richards	Refinement of the crystal structure of ribonuclease S. Comparison with and between the various ribonuclease A structures	Biochemistry (1992) 31:12304-12314	B	83	0.174	n/a	1.6	No		Zegers	Sulfate Ion (active site)
3RN3	P 1 21 1	B Howlin, DS Moss, GW Harris	Segmented anisotropic refinement of bovine ribonuclease A by the application of the rigid-body TLS model	Acta Crystallographica Section A (1989) 45:851-861	A&B	107	0.223	n/a	1.45	No		Sadasivan	Sulfate Ion (active site)
4SRN	P 32 2 1	VS deMel, PD Martin, MS Doscher, BF Edwards	Structural changes that accompany the reduced catalytic efficiency of two semisynthetic ribonuclease analogs	Journal of Biological Chemistry (1992) 267:247-256	B (D 121 A)	111	0.172	n/a	2.0	No		Zegers P3(2)21	Sulfate Ion (active site)

Table 1 continued.

PDB ID	Crystal Form	Reference Authors	Title	Reference	119 Conf	# of HOH	R	R free	Resolution (Å)	Inhibitor	Occupies what pockets?	Structure Set	Additional Molecule
5RSA	P 1 21 1	A Wlodawer, N Borkakoti, DS Moss, B Howlin	Comparison of two independently refined models of ribonuclease A	Acta Crystallographica Section B (1986) 42:379-387	A	128	n/a	n/a	2.0	No		Sadasivan	Phosphate Ion (active site), Deuterated water
7RSA	P 1 21 1	A Wlodawer, LA Svensson, L Sjölin, GL Gilliland	Structure of phosphate-free ribonuclease A refined at 1.26 Å	Biochemistry (1988) 27:2705-2717	A	188	0.150	n/a	1.25	No		Sadasivan Zegers	Tertiary-butyl alcohol, Deuterated water
8RSA	P 21 21 21	J Nachman, M Miller, GL Gilliland, R Carty, M Pincus, A Wlodawer	Crystal structure of two covalent nucleoside derivatives of ribonuclease A	Biochemistry (1990) 29:928-937	B	246	0.162	n/a	1.8	3'-Deoxy-3'-acetamido-thymidine	P1	Inhibitor Zegers	
9RSA	P 21 21 21	J Nachman, M Miller, GL Gilliland, R Carty, M Pincus, A Wlodawer	Crystal structure of two covalent nucleoside derivatives of ribonuclease A	Biochemistry (1990) 29:928-937	A	181	0.196	n/a	1.8	3'-Deoxy-3'-acetamido-uridine		Inhibitor	