SMITH, MYCHAL DAIJON. Probing the Hotspot Binding Sites of Proteins: P22TSP, Ras and BAF. (Under the direction of Dr. Carla Mattos.)

The Multiple Solvent Crystal Structure’s (MSCS) method is a powerful tool for studying protein surfaces, interactions and their function. MSCS is an experimental method that utilizes small organic solvents to map these protein characteristics. More recently it was discovered that these different organic solvent environments allow you to observe water molecules of structural and functional importance. Many studies have shown evidence that water is essential to protein-protein interactions, protein-ligand interactions and even catalytic mechanisms. Given this powerful potential of the MSCS method for characterizing protein binding sites, our group set out to analyze surfaces of proteins that bind to different types of ligands: peptides, other proteins, sugars, lipids, RNA and DNA. Previous research shows results for Elastase, Lysozyme, RNAse A, Chymotrypsin, Trypsin and Ras. While these proteins represent a multitude of categories my study seeks to expand the analysis sets.

This thesis describes results of the MSCS method applied to the protein surfaces of Phage 22 Tailspike (P22TSP), RasG12V and RasQ61L, and Barrier-to-autointegration Factor (BAF). The analysis of P22TSP demonstrates the method applied to a large sugar binding protein. The results show the ability of the method to depict the active site, sites delineating longer sugar binding pockets, a site thought to be a vestigial site, and comprehensive water analysis including the catalytic water and surrounding water molecules that interact with the ligand. This is one of the first applications of MSCS to a protein family where the active site
has been well studied but little is known of sites elsewhere. Analysis of RasG12V and RasQ61L gives the first example of an MSCS comparative analysis between wild type and disease-causing mutants. Results show that oncogenic RasG12V, RasQ61L have different areas of organic solvent overlap with the wild type protein. In RasG12V MSCS was able to depict areas of known protein complex interactions that were previously only observed computationally; while RasQ61L MSCS shows a unique area of organic solvent binding not seen in either RasG12V or wild type. We were also able to depict differences and similarities in water networks between both mutants and wild type. The BAF MSCS analysis demonstrated the first time this method was applied to a DNA-binding protein. MSCS depicts areas of protein-protein interactions supporting published data. In addition, analysis of surface water molecules revealed how mutated residues affect BAF-DNA interactions through water mediated networks instead of direct interactions. This is an important novel contribution to understanding the mechanism through which mutations affects the binding of BAF to DNA. Overall we are able to show the ability of MSCS to map protein binding surfaces in agreement with published data, further validating the method. Most excitingly, we go on to use MSCS to discover new binding sites and water networks in these three proteins of interest, enhancing our understanding of their structural features.
© Copyright 2013 by Mychal Daijon Smith

All Rights Reserved
Probing the Hotspot Binding Sites of Proteins: P22TSP, Ras and BAF

by
Mychal Daijon Smith

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 Biochemistry

Raleigh, North Carolina
2013

APPROVED BY:

Dr. Clay Clark
Committee Chair

Dr. Carla Mattos

Dr. Robert Rose

Dr. Alexander Deiters
DEDICATION

I would like to dedicate this dissertation to my friends and family, especially my mom Shelia Smith for always supporting me no matter what, whether financially or mentally. You are always available to talk or for whatever else I needed. No matter what you are dealing with, you helped me to become the man I am today. Also to my dad Donald Smith and sister Andrea Smith for always welcoming me home and treating me like a son and brother no matter how long I have to be away. Next to my grandmother Edna Tutt for being essential in how I was raised into the man I am today. I am grateful for all the times she was there for me and the family, whether it was watching me when I was younger or even letting us stay at her house when we first moved to NC. You are one of the biggest reasons I have made it as far as I have, even influencing me in my future research. For that I am forever grateful and even though there is no current cure or reversal process for neurodegenerative diseases, one day I hope that I can say I made a contribution to the field, thanks to you. Lastly, I would like to dedicate this to my current girlfriend/future wife, Fantasy Lozada, for everything she has done for me throughout graduate school. Everything you’ve done, from pushing me to work harder, motivating me to finish my experiments and writing, and giving 100% to everything I do, is invaluable.
BIOGRAPHY

Mychal Daijon Smith was born on July 10, 1985 in San Diego California where he was adopted by Donald and Shelia Smith. He has a younger sister Andrea Smith. Due to his dad’s job in the Navy, Mychal spent his first 5 years in California, Guam, and Maryland before finally settling in North Carolina. From second grade on, Mychal lived in NC where he graduated in the spring of 2003 from Pamlico County High School. Mychal then attended Fayetteville State University, a Historically Black College in Fayetteville, North Carolina. Throughout his studies at Fayetteville State, he conducted undergraduate research at his home institution and multiple summer internships at North Carolina State University. During this time he also became a member of Alpha Phi Alpha Fraternity, Incorporated. In the fall of 2007 Mychal graduated from Fayetteville State University, receiving a Bachelor of Science Degree in Chemistry and began working at Hexion (a chemical company in Fayetteville, NC). In the spring of 2008 Mychal was accepted to the graduate program for Biochemistry at North Carolina State University.

At NC State, Mychal became a full time member of Dr. Carla Mattos’ lab, where he worked with the Multiple Solvent Crystal Structures (MSCS) method. Mychal was exposed to MSCS through the multiple summer internships within the Mattos lab. This method would become the main focus of his dissertation work. In January 2012 Mychal moved to Northeastern University with Dr. Mattos to continue research within the lab. In May 2013 he returned to Durham, NC to finish his dissertation work. He will receive his Ph.D. in Biochemistry from North Carolina State University this fall.
ACKNOWLEDGEMENTS

First I would like to thank Dr. Carla Mattos for being an invaluable asset as a mentor, providing help with writing or analysis and always being my biggest advocate throughout my graduate career.

Thanks to all the past and present members of the Mattos group for supporting me throughout my career. Special thanks to Dr. Bradley Kearney, Dr. Gen Holzapfel and Dr. Susan Fetics who were there from the beginning. Dr. Kearney and Dr. Fetics were invaluable in helping me to finish this process. Thanks for the countless emails and discussions regarding life in general and lab work as well. Our continued friendship has given me great peace in going to the next level. Thanks for Dr. Michelle Dechene for allowing me to work on her MSCS project and even helping me to publish from it. Thanks to Christian Johnson for all of the great experimental and biological talks. I know that the future of the lab is in great hands. Thanks to Melissa Young for being a wonderful undergraduate student and mentee who was essential in moving the BAF project along. Thanks to Dr. Greg Buhrman for all of your help in the lab whether it was giving advice on how to approach an experiment or even teaching methods themselves. Thanks to Dr. Paul Swartz for being a great mentor and overall life advisor from teaching me the essentials of the MSCS method to always keeping my eye on the bigger picture. Also thanks for being there for me when preparing for my dissertation defense. The countless practices and talks were essential to securing my doctoral future; I owe a lot of my success to you. Thanks to the rest of my advisory committee including Dr. Clay Clark, Dr. Bob Rose, and Dr. Alexander Deiters for their help.
in giving research ideas and grooming me into a true scientist. Thanks to the rest of the Biochemistry Department especially Pam Francis, Curt Moore, Nancy House, Angie Barefoot, and the rest of the graduate students for support and making me feel like a part of the family throughout my studies.

Thanks to the newest members of the Mattos lab, including graduate students Ryan Knihtila, Kendra Marcus, Jillian Parker and undergraduates including Patrick Donohue and Austin Hendricks. I know the lab will continue to do great work with you all as members. Special thanks to undergraduate mentees Nancy Ho and Nicole Angel, who while at Northeastern tried to help me with a difficult project. Also thanks to the fantasy football crew, specifically Joslynn Lee, Sara Sadler, Hanlon Miller and Kara Strickland for making me feel I had a home away from home. Thanks to the members of CrossFit Durham for allowing me to continue to obtain physical goals and giving me an outlet to keep me mentally sane. Also thanks to the members of Grace Church on NC State campus specifically, Pastor Berk for helping me to increase my spiritual growth throughout my tenure at NC State.

Thanks to Dr. Sherrice Allen for being instrumental in giving me the tools I needed to become a great researcher and getting me started in research during my undergraduate career. Thanks to Ms. Alison Al-Baati for giving me the opportunity to conduct summer research in the Alliances for Graduate Education and the Professoriate; this opportunity catapulted my graduate research career. Thanks to Dr. Erin Banks for not only funding me through the Initiative for Maximizing Student Diversity grant my first two years but providing advice about life and school throughout my tenure. Thanks to Dr. David Schafer for providing extra funding to cover external expenses associated with being in graduate school especially when
dealing with the move to Boston. Also thanks to Dr. David Schafer and Mr. Brett Locklear for the countless opportunities within career advancement and opportunities to talk to minority students to help fuel my passion in helping the youth see higher degrees as an obtainable goal. Thanks to the NSF (MCB - 1237512) grant for providing the remaining financial support needed to finish my dissertation work.

Special thanks to all of my friends who have supported me no matter what. To Charlie Lawrence III for always treating me as a brother no matter the circumstances; Morika Williams for being a great friend and inspiring me to study more; Nicole Joyner for being a true friend whether we talk every day or every once in a while; and Thaddeus Ned for keeping me down to earth and being there since high school. Thanks from the bottom of my heart to Fantasy Lozada for supporting me no matter the circumstances. Whether it is me having to be up all times of the night writing to moving to another state, you have always stood by me. I could not ask for a more intelligent, caring, influential, and dedicated woman in my life. Thank you for really being there when I was writing my prelim and dissertation. For that I am forever grateful and hopefully can continue to give All of Me. I would also like to thank my immediate family Sheila Smith, Donald Smith, Andrea Smith, Edna Tutt, and my cousins for their continued support throughout life. To anyone I may have missed please forgive me and blame it on my head not my heart.
# TABLE OF CONTENTS

**LIST OF FIGURES** ................................................................................................................. ix
**LIST OF TABLES** ...................................................................................................................... xvi

## CHAPTER 1: INTRODUCTION ................................................................................................. 1
- Protein Interactions .................................................................................................................. 1
- Protein Interaction Experimental Methods .............................................................................. 4
- Protein Hot Spot Definition ...................................................................................................... 5
- FTMAP ..................................................................................................................................... 6
- Drug Discovery Methods .......................................................................................................... 7
- Multiple Solvent Crystal Structures Method ........................................................................... 9
- DRoP Program .......................................................................................................................... 11
- P22TSP ..................................................................................................................................... 12
- Ras ........................................................................................................................................ 15
- BAF .......................................................................................................................................... 23
- Conclusions .............................................................................................................................. 26

## CHAPTER 2: THE MSCS OF P22TSP .................................................................................... 28
- Introduction .............................................................................................................................. 28
- Experimental Methods ............................................................................................................ 34
- Results and Discussion ........................................................................................................... 41
- Water Analysis ....................................................................................................................... 48
- Conclusions .............................................................................................................................. 55
- Acknowledgments ................................................................................................................... 57
- Contribution ............................................................................................................................. 57

## CHAPTER 3: THE MSCS OF THE RAS G12V MUTANT ....................................................... 58
- Introduction .............................................................................................................................. 58
- Experimental Methods ............................................................................................................ 62
  - Mutagenesis, Expression and Purification ............................................................................ 62
  - Crosslinking and Organic Solvent Soaks ............................................................................. 62
  - Data Collection and Structure Refinement ........................................................................... 63
- Results ...................................................................................................................................... 64
- Conclusions .............................................................................................................................. 87
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wild type P22TSP with natural ligand (PDB Code 1TYX). Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>2</td>
<td>MAPK/ERK pathway including key protein members.</td>
</tr>
<tr>
<td>3</td>
<td>MSCS of Ras with clusters 1 – 8 shown in red spheres and organic solvents represented as sticks. Green depicts the effector lobe and catalytic machinery. (a) front side view of Ras (b) back side view of Ras. Grey depicts the allosteric lobe. Reprinted with permission from (Buhrman, O’Connor, et al., 2011).</td>
</tr>
<tr>
<td>4</td>
<td>MSCS versus FTMAP results. Clusters in red spheres (MSCS) and clusters in purple spheres (FTMAP). Effector lobe shown in green and allosteric lobe shown in grey. (a) front side view of Ras (b) back side view of Ras. Reprinted with permission from (Buhrman, O’Connor, et al., 2011).</td>
</tr>
<tr>
<td>5</td>
<td>Published BAF-DNA complex (PDB code 2BZF). Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>6</td>
<td>Monomers and trimers of all three members of the tailspike family with their respective bound ligands. A – P22TSP, B – HK620TSP and C – SF6TSP. Catalytic residues are shown within the respective active sites. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>7</td>
<td>Cube-like crystals of P22TSP.</td>
</tr>
<tr>
<td>8</td>
<td>Diffraction pattern of P22TSP diffracting better than 1.86 Å with detector at 250mm. Taken on home source at NC State University.</td>
</tr>
<tr>
<td>9</td>
<td>DRoP output of organic solvent cluster from P22TSP structures.</td>
</tr>
<tr>
<td>10</td>
<td>DRoP output of water clusters from P22TSP.</td>
</tr>
<tr>
<td>11</td>
<td>Ligand binding cavity organic clusters. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>12</td>
<td>Clusters further from the active site. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>13</td>
<td>Vestigial Active site. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>14</td>
<td>Remaining highly conserved clusters.</td>
</tr>
<tr>
<td>15</td>
<td>Overall water conservation view. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>16</td>
<td>Catalytic Water. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>17</td>
<td>Waters interacting with the natural ligand. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>18</td>
<td>Ligand bound vs. unbound water analysis. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>19</td>
<td>Glu359 alternate confirmation. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>20</td>
<td>Water interactions near vestigial active site. Figure created in PyMOL (Schrodinger, 2010).</td>
</tr>
<tr>
<td>21</td>
<td>All RasG12V structures resulting from the various soaks. The 2RGE model in yellow represents the “off” state of the allosteric switch and the 3K8Y model in green.</td>
</tr>
</tbody>
</table>
represents the “on” state. The structures soaked in organic solvents are represented as follows: ACT (pale green), DMF (tv_yellow), DMS (pale yellow), GOL (yellow orange), HEX (limon), RSR (wheat), SRS (sand), TBU (smudge), TFE (magenta), TMO (chartreuse) and X-link (split pea). P-loop (black) Switch I (blue), Switch II (red) and Loop7 (orange). Colors based on PyMOL settings. Figure created in PyMOL (Schrodinger, 2010). Figure 2. Switch II levels of disorder. All RasG12V structures soaked in organic solvents, with 2RGE (yellow) and 3K8Y (green) as defined in Figure 1. ACT (pale green), DMF (tv_yellow), DMS (pale yellow), GOL (yellow orange), HEX (limon), RSR (wheat), SRS (sand), TBU (smudge), TFE (magenta), TMO (chartreuse) and X-link (split pea). Figure created in PyMOL (Schrodinger, 2010). Figure 3. Wild type Ras hotspots (blue) versus RasG12V hotspots (red). P-loop (black) Switch I (blue), Switch II (red) and Loop7 (orange). Figure created in PyMOL (Schrodinger, 2010). Figure 4. Two RasG12V clusters in similar locations to wild type Ras FTMAP results. Figure created in PyMOL (Schrodinger, 2010). Figure 5. Zoom in of RasG12V Cluster 4 interaction. Figure created in PyMOL (Schrodinger, 2010). Figure 6. RasG12V Cluster 5 and 7 interactions. Figure created in PyMOL (Schrodinger, 2010). Figure 7. RasG12V Cluster 3 interactions. Figure created in PyMOL (Schrodinger, 2010). Figure 8. Effector Lobe (green) and Allosteric Lobe (black) with highly conserved waters. P-loop (black), Switch I (blue), Switch II (red) and Loop7 (orange). Figure created in PyMOL (Schrodinger, 2010). Figure 9. AllostERIC network 1 in wild type Ras (PDB code 3K8Y) (green) vs. RasG12V (cyan). Loop 7 is in orange. Figure created in PyMOL (Schrodinger, 2010). Figure 10. Allosteric hydrogen bond network 2 “on” (green) vs. “off” (cyan). Figure created in PyMOL (Schrodinger, 2010). Figure 11. Active site comparison “on” state for the wild type structure (green – PDB code 3K8Y) versus “off” state for the RasG12V structures as in the MSCS set (cyan). Figure created in PyMOL (Schrodinger, 2010). Figure 12. Beginning of Helix 5 to N85 network with 3K8Y (green) versus G12V X-link (cyan). Figure created in PyMOL (Schrodinger, 2010). Figure 13. Continuation of network with 3K8Y (green) versus RasG12V X-link (cyan). Figure created in PyMOL (Schrodinger, 2010). Figure 14. Network ending in the active site with 3K8Y (green) versus G12V X-link (cyan). Figure created in PyMOL (Schrodinger, 2010). Figure 1. RasQ61L models (in various colors) versus 2RGE (yellow) and 3K8Y (green). The GTP analogue GppNHp is bound to wild type and mutant Ras in the three models. The bound calcium and acetate ions are shown in green sphere and in stick respectively. P-loop (black), Switch I (blue), Switch II (red) and Loop7 (orange). Figure created in PyMOL (Schrodinger, 2010).
Figure 2. Organic Cluster overlap in effector lobe. P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 3. Cluster overlap between the lobes. P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 4. Only cluster not in overlap. P-loop (orange in effector lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 5. Highly conserved water molecules (red spheres) in the effector lobe (green) and in the allosteric lobe (black). P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 6. Active site water comparison “on” wild type (green – PDB code 3K8Y) versus “off” RasQ61L (cyan – as in the MSCS set). Figure created in PyMOL (Schrodinger, 2010).

Figure 7. Allosteric water network 2. Figure created in PyMOL (Schrodinger, 2010).

Figure 8. A) Beginning of Helix 5 network. Figure created in PyMOL (Schrodinger, 2010).

Figure 8. B) Continuation of helix 5 to active site water network. Figure created in PyMOL (Schrodinger, 2010).

Figure 8. C) Hydrogen bonding network ending in the active site. Figure created in PyMOL (Schrodinger, 2010).

Figure 1. His-Trap results graph.

Figure 2. Gel Filtration results graph.

Figure 3. Rod-like shape of high diffraction quality BAF crystals.

Figure 4. Example of diffraction pattern from BAF crystals.

Figure 5. BAF dimer with organic solvent spheres. Figure created using Pymol (Schrodinger, 2010).

Figure 6. BAF DNA structure (pdb code 2BZF) superimposed with soaked structures in the MSCS set. Figure created using Pymol (Schrodinger, 2010).

Figure 7. BAF/LEM-domain and BAF/DNA structures superimposed with soaked structures.

Figure 8. Zoom in of BAF/LEM-domain structure with soaked structures.

Figure 9. Overall water analysis of BAF. A – Front view of BAF and B – Back view of BAF.

Figure 10. BAF-DNA structure superimposed using coloring function. Amino acid residues are colored based on the conservation level associated with interacting water molecules. Residues that interact with 100% conserved waters are blue, 90% cyan, and 80% green.

Figure 11. BAF/DNA structure superimposed with water analysis showing waters interacting with DNA backbone groups.
Figure 12. Water mediated network between BAF and DNA. Figure created using Pymol (Schrodinger, 2010).

Figure 13. Water mediated network between BAF and DNA involving bases 3, 4 and 5. Figure created using Pymol (Schrodinger, 2010).

Figure 1. Wild type P22TSP with natural ligand (PDB Code 1TYX). Figure created in PyMOL (Schrodinger, 2010).

Figure 2. MAPK/ERK pathway including key protein members.

Figure 3. MSCS of Ras with clusters 1 – 8 shown in red spheres and organic solvents represented as sticks. Green depicts the effector lobe and catalytic machinery. (a) front side view of Ras (b) back side view of Ras. Grey depicts the allosteric lobe. Reprinted with permission from (Buhrman, O’Connor, et al., 2011).

Figure 4. MSCS versus FTMAP results. Clusters in red spheres (MSCS) and clusters in purple spheres (FTMAP). Effector lobe shown in green and allosteric lobe shown in grey. (a) front side view of Ras (b) back side view of Ras. Reprinted with permission from (Buhrman, O’Connor, et al., 2011).

Figure 5. Published BAF-DNA complex (PDB code 2BZF). Figure created in PyMOL (Schrodinger, 2010).

Figure 1. Monomers and trimers of all three members of the tailspike family with their respective bound ligands. A – P22TSP, B – HK620TSP and C – SF6TSP. Catalytic residues are shown within the respective active sites. Figure created in PyMOL (Schrodinger, 2010).

Figure 2. Cube-like crystals of P22TSP.

Figure 3. Diffraction pattern of P22TSP diffracting better than 1.86 Å with detector at 250mm. Taken on home source at NC State University.

Figure 4. DROp output of organic solvent cluster from P22TSP structures.

Figure 5. DROp output of water clusters from P22TSP.

Figure 6. Ligand binding cavity organic clusters. Figure created in PyMOL (Schrodinger, 2010).

Figure 7. Clusters further from the active site. Figure created in PyMOL (Schrodinger, 2010).

Figure 8. Vestigial Active site. Figure created in PyMOL (Schrodinger, 2010).

Figure 9. Remaining highly conserved clusters.

Figure 10. Overall water conservation view. Figure created in PyMOL (Schrodinger, 2010).

Figure 11. Catalytic Water. Figure created in PyMOL (Schrodinger, 2010).

Figure 12. Waters interacting with the natural ligand. Figure created in PyMOL (Schrodinger, 2010).

Figure 13. Ligand bound vs. unbound water analysis. Figure created in PyMOL (Schrodinger, 2010).

Figure 14. Glu359 alternate confirmation. Figure created in PyMOL (Schrodinger, 2010).

Figure 15. Water interactions near vestigial active site. Figure created in PyMOL (Schrodinger, 2010).
Figure 1. All RasG12V structures resulting from the various soaks. The 2RGE model in yellow represents the “off” state of the allosteric switch and the 3K8Y model in green represents the “on” state. The structures soaked in organic solvents are represented as follows: ACT (pale green), DMF (tvyellow), DMS (pale yellow), GOL (yellow orange), HEX (limon), RSR (wheat), SRS (sand), TBU (smudge), TFE (magenta), TMO (chartreuse) and X-link (split pea). P-loop (black) Switch I (blue), Switch II (red) and Loop7 (orange). Colors based on PyMOL settings. Figure created in PyMOL (Schrodinger, 2010).

Figure 2. Switch II levels of disorder. All RasG12V structures soaked in organic solvents, with 2RGE (yellow) and 3K8Y (green) as defined in Figure 1. ACT (pale green), DMF (tvyellow), DMS (pale yellow), GOL (yellow orange), HEX (limon), RSR (wheat), SRS (sand), TBU (smudge), TFE (magenta), TMO (chartreuse) and X-link (split pea). Figure created in PyMOL (Schrodinger, 2010).

Figure 3. Wild type Ras hotspots (blue) versus RasG12V hotspots (red). P-loop (black) Switch I (blue), Switch II (red) and Loop7 (orange). Figure created in PyMOL (Schrodinger, 2010).

Figure 4. Two RasG12V clusters in similar locations to wild type Ras FTMAP results. Figure created in PyMOL (Schrodinger, 2010).

Figure 5. Zoom in of RasG12V Cluster 4 interaction. Figure created in PyMOL (Schrodinger, 2010).

Figure 6. RasG12V Cluster 5 and 7 interactions. Figure created in PyMOL (Schrodinger, 2010).

Figure 7. RasG12V Cluster 3 interactions. Figure created in PyMOL (Schrodinger, 2010).

Figure 8. Effector Lobe (green) and Allosteric Lobe (black) with highly conserved waters. P-loop (black), Switch I (blue), Switch II (red) and Loop7 (orange). Figure created in PyMOL (Schrodinger, 2010).

Figure 9. Allosteric network 1 in wild type Ras (PDB code 3K8Y) (green) vs. RasG12V (cyan). Loop 7 is in orange. Figure created in PyMOL (Schrodinger, 2010).

Figure 10. Allosteric hydrogen bond network 2 “on” (green) vs. “off” (cyan). Figure created in PyMOL (Schrodinger, 2010).

Figure 11. Active site comparison “on” state for the wild type structure (green – PDB code 3K8Y) versus “off” state for the RasG12V structures as in the MSCS set (cyan). Figure created in PyMOL (Schrodinger, 2010).

Figure 12. Beginning of Helix 5 to N85 network with 3K8Y (green) versus G12V X-link (cyan). Figure created in PyMOL (Schrodinger, 2010).

Figure 13. Continuation of network with 3K8Y (green) versus RasG12V X-link (cyan).

Figure 14. Network ending in the active site with 3K8Y (green) versus G12V X-link (cyan).

Figure 1. RasQ61L models (in various colors) versus 2RGE (yellow) and 3K8Y (green). The GTP analogue GppNHp is bound to wild type and mutant Ras in the three models. The bound calcium and acetate ions are shown in green sphere and in stick respectively. P-
loop (black), Switch I (blue), Switch II (red) and Loop7 (orange). Figure created in PyMOL (Schrodinger, 2010).

Figure 2. Organic Cluster overlap in effector lobe. P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 3. Cluster overlap between the lobes. P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 4. Only cluster not in overlap. P-loop (orange in effector lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 5. Highly conserved water molecules (red spheres) in the effector lobe (green) and in the allosteric lobe (black). P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 6. Active site water comparison “on” wild type (green – PDB code 3K8Y) versus “off” RasQ61L (cyan – as in the MSCS set). Figure created in PyMOL (Schrodinger, 2010).

Figure 7. Allosteric water network 2. Figure created in PyMOL (Schrodinger, 2010).

Figure 8. A) Beginning of Helix 5 network. Figure created in PyMOL (Schrodinger, 2010).

Figure 8. B) Continuation of helix 5 to active site water network. Figure created in PyMOL (Schrodinger, 2010).

Figure 8. C) Hydrogen bonding network ending in the active site. Figure created in PyMOL (Schrodinger, 2010).

Figure 9. Overall water analysis of BAF. A – Front view of BAF and B – Back view of BAF. Figure created using Pymol (Schrodinger, 2010).

Figure 10. BAF-DNA structure superimposed using coloring function. Amino acid residues are colored based on the conservation level associated with interacting water molecules. Residues that interact with 100% conserved waters are blue, 90% cyan, and 80% green. Figure created using Pymol (Schrodinger, 2010).
Figure 11. BAF/DNA structure superimposed with water analysis showing waters interacting with DNA backbone groups. Figure created using Pymol (Schrodinger, 2010). 

Figure 12. Water mediated network between BAF and DNA. Figure created using Pymol (Schrodinger, 2010).

Figure 13. Water mediated network between BAF and DNA involving bases 3, 4 and 5. Figure created using Pymol (Schrodinger, 2010).
LIST OF TABLES

Table 1. Refinement statistics for P22TSP. Parentheses equal highest resolution shell........ 40
Table 2. Clusters with number of organic molecules in parentheses and their protein
interactions.......................................................................................................................... 42
Table 1. Refinement Statistics for RasG12V. Parentheses equal highest resolution shell. .... 65
Table 2. List of all protein contacts within 3.5 Å between RasG12V and associated organic
solvent clusters. In parentheses are the number of organic molecules found in each
cluster within the MSCS set.................................................................................................... 70
Table 1. Refinement Statistics of RasQ61L-GppNHp. Parentheses equals highest resolution
shell........................................................................................................................................ 97
Table 2. Cluster numbers and their interacting residues listed. In parentheses are the numbers
of organic molecules found in each cluster within the MSCS set. .................................. 101
Table 1. Refinement Statistics for BAF crystals. Parentheses equal highest resolution shell.
............................................................................................................................................... 132
Protein Interactions

Proteins are the essential work machines of the cell and are involved in virtually all aspects of function. Between 20,000 and 30,000 genes encode for over 500,000 types of proteins, 10,000 of which can be produced in a cell at any time (Berggård, Linse, & James, 2007). It is thought that 80% of these proteins form some type of complex; including protein-protein interactions defined as physical contacts with molecular docking occurring between proteins within a living cell or organism (Berggård et al., 2007) (De Las Rivas & Fontanillo, 2010). These physical contacts regulate cell growth, morphology, motility, gene expression, proliferation, apoptosis, and intercellular communication (De Las Rivas & Fontanillo, 2010). These many processes are kept meticulously in check, but can harm the cell when deregulated, leading to diseases such as cancer. Understanding the rules that govern protein-protein interactions is an important component in being able to disrupt complexes in dysfunctional cellular systems and could lead to higher success in drug development for treatment of disease. However, due to their characteristics such as size, shape and surface complementarity protein-protein interactions are very complex and difficult to modulate (A. Fernández & Scheraga, 2003).

The standard size of protein-protein interfaces vary greatly with larger interfaces typically occurring in G proteins and other members of signal transduction pathways (Horton & Lewis, 1992). By studying these interfaces the types of interactions between complexes can be determined. The classification of complexes includes the homo-complexes, that occur
through dimerization of identical proteins and tend to be fairly stable, and hetero-complexes that are usually highly sensitive the cellular environment and/or external factors (Cardinale et al., 2010; Jones & Thornton, 1996). Understanding the factors responsible for formation of these complex types contributes greatly to the study of protein-protein interactions. Proteins that are involved in homo-complexes are usually not found as monomers in solution and have their hydrophobic surfaces permanently buried within the complex (Jones & Thornton, 1996). In contrast hetero-complexes can exist separately as monomers in solution, allowing a large exposure of its protein surface (Jones & Thornton, 1996). The more transient nature of the hetero-complexes requires a less hydrophobic surface and energetically weaker interactions compared to homo-complexes (Jones & Thornton, 1996). This is an important distinction when studying protein-protein interactions and the forces that contribute to complex formation, such as those involved in hydrophobic and electrostatic interactions.

A gain in free energy resulting from the transfer of residues from a polar to nonpolar environment promote complex interactions (Dill, 1990). These interactions cause tight packing of residues causing organized patches protruding from the protein surface (Moreira, Fernandes, & Ramos, 2007). These areas also cause expulsion of water molecules increasing entropy and contributing favorably to energy for complex formation (Dill, 1990). Proteins tend to form favorable complexes when energy is most favorable. One of the most important electrostatic interactions contributing to complex formation, even defining the length of complex interaction, is hydrogen bonding (Camacho, Weng, Vajda, & DeLisi, 1999; Vijayakumar et al., 1998). These forces are essential to maintaining protein interactions.
Specific regions or amino acid sequences also contribute to protein-protein interactions within cells (Cordomí & Perez, 2009; Roy, Martinez, Platero, Lane, & Werner-Washburne, 2009). Amino acid mutations cause alterations in protein surfaces resulting in dysfunction (Cordomí & Perez, 2009; Roy et al., 2009). This dysfunction can cause complexes to either disassociate too quickly (or not bind at all) or to stay bound too long, leading to diseases. These areas of dysfunction not only get contributions from the immediate area but the outside regions as well. Interacting forces from outside these regions create favorable surfaces dictating the interactions within, with conformational changes occurring in one or both protein complex members (Kumar & Litwack, 2009). Examples include the three proteins included in the present thesis: proteins interacting to form a virus complex to further infect cells (P22TSP) (Baxa et al., 1996), proteins interacting to further propagate a signal within a pathway (Ras GTPase) (Buhrman, de Serrano, & Mattos, 2003), and proteins interacting to prevent degradation of a virus cell by integrating its own DNA (BAF) (Cai et al., 1998). Proteins interact not only with other proteins, but also with DNA, RNA, sugars and other ligands, including a variety of small molecules.

Proteins interact with DNA specifically or non-specifically to regulate the biological function of DNA such as gene expression. Proteins also interact with sugar moieties. One example of this is from the family of lectin proteins. Lectins recognize sugar moieties and play a key role in proteins recognizing the surface of a cell (Rutishauser & Sachs, 1975). This allows these types of proteins to have many biological functions such as regulation of control of protein levels in the blood, cell adhesion, and glycoprotein synthesis (Rutishauser & Sachs, 1975). Another important interaction that is heavily studied is the protein-ligand
interaction. Some proteins within the human body have natural ligands to help regulate processes in the body. Examples in metabolism and in the nervous system come to mind (Isin & Guengerich, 2007; Okuno & Yokomizo, 2011). Other proteins are the targets of man-made ligands. Many proteins have become drug targets over the years. Understanding these protein interactions is essential to making drugs that are safe and effective to the intended targets. All these types of interactions add various levels of complexity to how proteins interact with their environments. Because of this a plethora of experimental methods have been devised to study the different types of protein interactions.

Protein Interaction Experimental Methods

There are many different methods for isolating and characterizing protein interactions in vitro such as blue native gel electrophoresis, protein cross-linking, in vitro binding assays, and co-immunoprecipitation (Miernyk & Thelen, 2008). However, one of the oldest and most widely used methods is alanine scanning mutagenesis. This method systematically substitutes amino acid residues with an alanine, removing all the side chain atoms past the β-carbon (K. L. Morrison & Weiss, 2001). Alanine allows for the protein to maintain the backbone dihedral angles without introducing conformational flexibility (K. L. Morrison & Weiss, 2001). This reveals whether a specific residue is important for binding while minimizing the chances of perturbing the overall structure of the protein. Assays done in vitro can then determine the contribution of that side chain to the overall binding energy. One of the drawbacks to this method is the amount of work involved. First mutations must be made and then mutant proteins are expressed and purified, which sometimes involves a refolding step (K. L. Morrison & Weiss, 2001). In spite of being labor intensive, this method helped
Redefine a framework for studying individual residue contributions to intermolecular interactions and led to the new categorization of hot spots as areas on the protein surface that contribute significantly to affinity (Clackson & Wells, 1995).

**Protein Hot Spot Definition**

Originally, a protein hot spot was determined to be a site where an alanine mutation causes a significant increase in the binding free energy of at least 2.0 kcal/mol (Thorn & Bogan, 2001). Previously, residues were considered to have a strong impact on protein binding if it contributed a free energy greater than 4.0 kcal/mol (Moreira et al., 2007). However, this binding energy is highly unusual and the limit for a hot spot had to be lowered to 2.0 kcal/mol in order to get more accurate statistical analyses (Moreira et al., 2007). This led to the idea that within a protein-protein interface only a small subset of the buried residues contributes to the majority of the binding affinity (Moreira et al., 2007). Analysis show that the composition of the residues within these hot spots are not random but distinct (Lichtarge, Bourne, & Cohen, 1996), with tyrosine (12.3%), arginine (13.3%), and tryptophan (21%) being considered the most fundamental (Moreira et al., 2007). These amino acids contribute to the complementarity of the binding pocket which normally contains a plethora of structurally conserved residues (Fau & Argos, 1994). A complementary pocket is defined as a pocket with similarity in both the position and shape of hydrophilic and hydrophobic hot spots (Arkin & Wells, 2004). Complementary pockets have characteristics of hydrophobic residues from one protein surface fitting into pockets on the opposite surface or the forming of salt bridges from buried charged residues (Arkin & Wells, 2004). This led to the idea that the face of one hot spot packs against the face of another hot
spot when complex binding occurs (I. S. Moreira, P. A. Fernandes, & M. J. Ramos, 2006; Irina S. Moreira, Pedro A. Fernandes, & Maria J. Ramos, 2006) and has become a driving force in drug discovery (Li, Keskin, Ma, Nussinov, & Liang, 2004). Overall complementarity can be affected by the number of buried water molecules, packing densities of atoms involved in the interface, positioning of polar and nonpolar residues and the size of the buried surface (Lawrence & Colman, 1993). Complementarity is thought to be consistent across pockets because they are enriched with conserved residues that coevolve (del Sol Mesa, Pazos, & Valencia, 2003; Goh & Cohen, 2002) and also correlate within the protruding residues that follow a similar trend (Li et al., 2004). These ideas of hot spots and pockets of complementarity have driven the methods of drug discovery for decades. Alanine scanning is not sufficient when looking at these types of protein interactions. However, more efficient methods are recently developed to determine areas of protein interactions.

FTMAP

Experimental methods are often used to locate areas of protein interactions, yet one drawback to experimental methods is that the experiments take a long time to complete. Computational methods can be used in combination with experiments to expedite the process. FTMAP is a computational tool developed to reproduce NMR and X-ray binding site mapping results (Brenke et al., 2009). The method uses organic probe molecules by employing an algorithm that generates over 2000 bound molecules using rigid body docking, refining of positions through use of energy minimization, clustering of the conformational results, and ranking the clusters based on average free energy (Brenke et al., 2009). This program allows for the use of millions of probes covering the entire protein with organic
solvents computationally to determine binding sites. The ability to probe the protein surface regardless of crystal symmetry is one of the advantages of FTMAP. One disadvantage of FTMAP is that it may not always be biologically relevant since it does not take water, ion, or ligand molecules into account. For this reason, it is essential we combine experimental and computational methods to get a better idea of what is happening within protein complexes. Two such experimental methods are SAR by NMR and phage display.

**Drug Discovery Methods**

Drug discovery is a process in which identification and optimization of drugs lead to uncovering of interactions with a target to alter its activity. This process relies heavily on the properties of protein-ligand interactions. Because of the plethora of compounds available for drug targets, high throughput methods are sought out by many companies. More recently, because of the amount of resources required, more efficient fragment-based methods are currently being refined. Fragment-based methods allows for the screening of fewer compounds, use of lesser resources than high throughput, use of smaller compounds that are weak but efficient binders, and development of more potent compounds by merging or linking while keeping high ligand efficiency (Erlanson, 2006). SAR by NMR is a method that fits nicely in this category because of its ability to screen a robust library of weak to tight binding compounds in a high throughput fashion. SAR by NMR accomplishes these goals by testing if the ligand interacts with target by measuring changes in target chemical shifts and testing if the ligand interacts with target by measuring changes in ligand NMR parameters upon addition of the target (Coles, Heller, & Kessler, 2003). This allows for the screening of 100,000 or more compounds a week while being very cost efficient. Even though this method
is highly effective, disadvantages such as the limits of protein size, the need to breakdown compound mixtures, and the large amount of soluble and isotopically labeled protein required (C. Fernández & Jahnke, 2004) increases the need to explore other methods of drug discovery.

Another method being highly used in the drug discovery realm is phage display. Phage display is a screening combinatorial peptide library technique used to study protein-protein, protein-peptide, and protein-DNA interactions using bacteriophages connecting proteins with genetic information in which they are encoding (Huang, Pershad, Kokoszka, & Kay, 2011). This technique allows for the generating of libraries including $10^{10}$ or more different peptides (Huang et al., 2011). This vast number of peptide targets confers many advantages to the phage display method, such as the ability of the libraries to be regenerated and stored indefinitely, the short time it takes to identify ligands, and the ease with which a selected peptide can be identified (Huang et al., 2011). While this method is great for screening large amounts of targets for use in drug discovery there are some limitations. The molarity of the peptides is low causing the need for synthetic versions to be made and because of some limitations on biological selections of certain residues within the peptide the results can become biased (Huang et al., 2011). This results in phage display becoming more of a surveying tool rather than a tool to analyze specifics about the interactions. Both SAR by NMR and phage display are tools for fragment-based high throughput experiments to identify potential interaction targets. However, when considering the analysis of protein interactions, more information such as a pocket composition, protein conformational changes, and water
analysis are needed in order to determine the true effects of binding. The Multiple Solvent Crystal Structures (MSCS) method allows for this deeper, more detailed, analysis.

**Multiple Solvent Crystal Structures Method**

MSCS was first published as an experimental method for use in determining and characterizing protein binding sites (Carla Mattos & Ringe, 1996). It was proposed that developing a good method for locating and characterizing binding site hot spots on protein surfaces would lead to a more efficient method for drug design (Carla Mattos & Ringe, 1996). The early experiments revealed that the proteins retained a heavily hydrated first shell while only a few organic solvent molecules bound within the active site (M. Dechene, G. Wink, M. Smith, P. Swartz, & C. Mattos, 2009; Carla Mattos et al., 2006). This initial work focused on the development and validation of the MSCS method as a general means to study binding sites on protein surfaces. First high quality crystals that diffract to greater than 2.0 Å resolution (preferably between 1.3 Å – 1.7 Å) must be obtained. Next these crystals are cross-linked so that they can be transferred to a variety of solvent conditions with minimal damage. Although gluteraldehyde is the standard crosslinking agent for MSCS experiments, any type of cross-linker can be used as long as it does not perturb the structure of the protein. After cross-linking, the crystals are soaked in either neat organic solvent or a solvent/water mixture for at least an hour. The crystals are then dipped in a cryoprotectant to be collected using x-ray diffraction. The resulting structures are superimposed and analyzed. The organic solvents cluster in areas of protein-protein interactions or protein-ligand interactions that were determined to coincide with hot spots (Carla Mattos et al., 2006). The organic molecules within these hot spots are thought to reveal the shape and chemical properties of
the surrounding interacting region (Michelle Dechene, Glenna Wink, Mychal Smith, Paul Swartz, & Carla Mattos, 2009). By knowing these characteristics we then should be able to develop functional groups that bind to the surface within a larger drug ligand. In the MSCS method we consider a hot spot to bind at least two different kinds of organic solvent molecules that cluster when the relevant structures are superimposed. The MSCS method has been successfully applied to many proteins, providing new ways to analyze structures.

After MSCS was validated through its application to Elastase (serine protease) (Carla Mattos et al., 2006) we used the MSCS RNAse A (a RNA binding protein) to study subtle changes in protein conformations (protein plasticity) and hydration characteristics (Michelle Dechene et al., 2009). This was the first time a thorough water analysis was presented through the use of MSCS using a semi-automated method. Water is essential to protein structure, function and dynamics (Ball, 2007). Crystallographic water molecules can be split into four classes “inside”, “contact”, “first-layer” and “second-layer” depending on the location relative to the solvent-accessible surface (SAS) of protein molecules and interaction modes (Connolly, 1983). The “inside” class includes waters inside the SAS occupying cavities within the protein molecules and are most likely confined to the folding process and exchange with bulk solvent water (Nakasako, 2004). “Contact” class waters are usually located outside the SAS interacting with adjoining molecules related by crystallographic symmetry and mediating intermolecular interactions (Nakasako, 2004). Both “first-layer” and “second-layer” waters are located outside the SAS with “first-layer” waters interacting directly with protein surface atoms and “second-layer” waters making no direct protein interactions (Nakasako, 2004). These classifications of water molecules are important in
analysis of water structure through protein crystallography. With the use of MSCS we are able to identify each of these types of waters; these waters have important implications for being integral to catalytic mechanisms and are essential for a natural ligand binding in an active site. Since we are looking at so many different structures, each with their own crystallographic waters and organic solvents, our group needed to develop an automated way to classify and analyze solvent structures. Instead of doing visual determination of water and organic solvent clusters for a protein in an MSCS set, we now use an automated organization of the clusters for easy analysis of results with the Detection of Related Solvent Positions (DRoP) program.

**DRoP Program**

DRoP was created to make a more overall consistent analysis method for analyzing solvent structure in multiple structures of proteins (Kearney, 2012). One previous problem in determining solvent clusters in multiple crystal structures is symmetry. Waters and organic solvents would sometimes be placed in different symmetry-related positions within the structures in the MSCS set. One would then have to determine by visual inspection that all water molecules with equivalent positions are located consistently in the asymmetric unit. DRoP does this automatically for all relevant solvent molecules (Kearney, 2012). DRoP is also able to give a cluster quality report and a way to visualize clusters using PyMOL (Kearney, 2012). This program allows for more accurate organic solvent and water molecule location identification, which yields more accurate structural analyses. MSCS has been applied to Elastase (serine protease) (Carla Mattos et al., 2006), Lysozyme (Enzyme) (Nicely, 2005), and RNAse A (RNA binding enzyme) (Michelle Dechene et al., 2009), while
both MSCS and DRoP have been used on Lysozyme (glycoside hydrolase) (Donohue, Unpublished Results), wild type Ras (GTP/GDP binding protein) (Buhrman, O’Connor, et al., 2011), Trypsin (serine protease) (Kearney, 2012), Chymotrypsin (serine protease) (Kearney, 2012), and RNAse A (Kearney, 2012). These previous sets do not encompass all types of proteins. My research adds to these previous MSCS sets by expanding the types of protein analyzed by MSCS to P22TSP (sugar binding protein) and, BAF (DNA binding protein). I also do a comparative analysis between the published MSCS for wild type Ras and for the new MSCS sets of two of its potent oncogenic mutants, Ras G12V (mutant GTP/GDP binding protein) and Ras Q61L (mutant GTP/GDP binding protein). Each MSCS set is analyzed using DRoP for solvent analysis, including conserved water and organic solvent locations. Below I describe the proteins relevant to the present studies.

**P22TSP**

Salmonella phage P22 is a widely studied virus in the detection and application of better drug therapy methods and in understanding the bacteriophage infection process. Salmonella phage P22 is a dsDNA phage in the *Podoviridae* family (Steinbacher et al., 1997). This bacteriophage is lysogenic and infects Salmonella bacteria by packaging the DNA into the capsid, which fills the cell, causing lysis and releasing the phage to infect the surrounding cells (Susskind & Botstein, 1978). Bacterial infection with Salmonella affects humans when one comes in contact with infected uncooked or cooked food, animals and even other people. Live salmonellae enter the body eventually ending up in the gastrointestinal tract and passing through the lymphatic system of the intestine into other organs. This can lead to diarrhea, abdominal cramps, vomiting and nausea in healthy adults
or can be fatal in humans with compromised immune systems. Recently, it was reported that bacteria resistant to antibiotics is on the rise and the tailspike portion of phage 22 (P22TSP) is a strong candidate for use as a drug therapy because of its high resistance to GI tract protease enzymes (Waseh et al., 2010). P22TSP was orally administered to chickens and found to reduce Salmonella colonization within the gut (Waseh et al., 2010).

P22TSP is studied as a model for large protein folding systems and its role in the bacteriophage’s infection process. It is a good model system for our purposes because little is known about the binding of large complex oligosaccharides to proteins, since structural and thermodynamic information for these proteins are rarely available (Baxa, Cooper, Weintraub, Pfeil, & Seckler, 2001). However, there is plenty of data on P22TSP available for comparison across multiple experimental methods. Phages use tail spike proteins to establish first contact with host cell surfaces, which makes understanding structure as it pertains to function that much more important (Andres et al., 2012). P22TSP is a 666 amino acid residue protein that is biologically relevant as a homotrimer (Steinbacher et al., 1994). The monomer (seen in the asymmetric unit of crystal structures, where through a three-fold axis of symmetry the biological trimer is obtained) has an overall shape of a fish composed of six main segments: the body, mouth, dorsal fin, and the first, second and third segments of the caudal fin (Steinbacher et al., 1994). The N-terminus interacts with the viral capsid and is highly flexible in the isolated protein. Therefore it is not seen in the crystallographic model. The C-terminus has three integrated subunits causing high thermal stability (Steinbacher et al., 1994). Figure 1 shows the triangular shape of the β-helix within the main body forming
the long groove of the solvent-exposed side in which the active site is contained (Steinbacher et al., 1994).

Figure 1. Wild type P22TSP with natural ligand (PDB Code 1TYX). Figure created in PyMOL (Schrodinger, 2010).

The active site is composed of three catalytic residues including one glutamic acid (E359) and two aspartic acids (D392 & D395) with mutations showing a decrease in cleavage activity (Baxa et al., 1996). These residues help to cleave the ligand that P22TSP recognizes, the O-antigen moiety of the lipopolysaccharide of Salmonella, belonging to serotype A, B, or D1 (Baxa et al., 1996). The multitude of serotypes allows for different binding modes of ligands in the active groove pocket, enabling the phage to infect multiple
strains of host cells (Baxa et al., 1996). P22TSP also has an endorhamosidase activity producing two repeating octasaccharide units (Baxa et al., 1996). This process is required for the virus to achieve infection and may help allow access to the cell surface for DNA injection (Bayer, Takeda, & Uetake, 1980). Because of P22TSP’s characteristics of high thermostability, protease resistance, resistance to reversibly unfolding in concentrated chemical denaturants and resistance to high levels of urea and SDS, P22TSP is ideal for bacteriophage therapy in humans (Waseh et al., 2010). However, in order to drive these applications in humans a greater understanding of binding interactions between P22TSP, its ligand and the rest of the bacteriophage complex is needed. Elucidating these interactions will increase the efficiency of piecing together functional groups to make viable drugs. Through use of MSCS we are able to discover unreported binding sites of potential importance including a subsite in the active site as presented in Chapter two.

**Ras**

Ras is important for controlling cell proliferation, differentiation and basic cell survival (Fernandez-Medarde & Santos, 2011). There are three isoforms of Ras: H-Ras, N-Ras and K-Ras. The three isoforms are highly similar, excluding the posttranslational modifications and amino acids within the hypervariable region (Hancock, 2003). The sequence is 100% identical in the N-terminal lobe 1 (residues 1-86) and show 90% identity in the C-terminal lobe 2 (residues 87-171) (Hancock, 2003). These are the main regions where catalytic and protein-protein interaction activities occur, which justify the use of one form as a model for the all the isoforms. H-Ras mutations are within about 20% of cancers and occur mainly in the G12 and Q61 amino acid residues (Fernandez-Medarde & Santos, 2011). H-
Ras shows 9% of the mutations in codon 13, 34% of mutations in codon 61 and 54% of mutations occurring in codon 12 (Fernandez-Medarde & Santos, 2011). These mutations affect Ras so strongly because they interfere with its GTPase function, which is essential for the molecular switch mechanism that regulates signal transduction pathways within the cell (Bourne, Sanders, & McCormick, 1990). Ras has two nucleotide-bound states: an inactive form bound to GDP and an active form bound to GTP (Bourne et al., 1990). GDP is exchanged for GTP through the action of guanine nucleotide exchange factors (GEFs) and GTP is hydrolyzed to GDP through the enhancement of the intrinsic GTPase activity of Ras by GTPase activating proteins (GAPs) (D. K. Morrison & Cutler Jr, 1997). GEFs promote Ras to the signaling active state through the nucleotide exchange, while GAPs promote Ras to the inactive state by enhancing the hydrolysis of GTP to GDP by $10^3$ – $10^5$ fold relative to the rate measured in vitro in wild type Ras (Campbell, Khosravi-Far, Rossman, Clark, & Der, 1998).
Figure 2. MAPK/ERK pathway including key protein members.
As seen in Figure 2, one pathway that plays a significant role in several biological processes related to Ras is the Ras/Raf/MEK/ERK pathway (Campbell et al., 1998). The pathway begins as the epidermal growth factor receptor (EGFR) is activated by an external ligand. Binding of this external ligand (epidermal growth factor EGF) causes phosphorylation of tyrosine residues on EGFR allowing for the binding of growth factor receptor-bound protein 2 (GRB2) (Schulze, Deng, & Mann, 2005). GRB2 is then poised to interact with the guanine nucleotide exchange factor SOS (sons of sevenless) causing its activation (Zarich et al., 2006). SOS then binds the membrane bound form of Ras inside the cell, promoting the exchange of GDP for GTP (Campbell et al., 1998). The activated Ras-GTP then interacts with Raf, recruiting it to the cell membrane and propagating the signal (D. K. Morrison & Cutler Jr, 1997; Vojtek, Hollenberg, & Cooper, 1993). Activated Raf binds and phosphorylates MEK1 and MEK2, leading to stimulation of ERK1 and ERK2 by modifying their tyrosine and threonine residues (Campbell et al., 1998; D. K. Morrison & Cutler Jr, 1997). Further propagation of the signal leads to translocation of ERK to the nucleus where it interacts with transcription factors to modulate a multitude of cell processes (Roux & Blenis, 2004). This stimulation also in turn activates mitogen-activated protein kinase (MAPK) through phosphorylation (Avruch et al., 2001). MAPK can in turn phosphorylate many proteins including 40S ribosomal protein kinase (RSK), C-myc, and MNK causing activation leading to phosphorylation of CREB. These phosphorylation events affect many processes within the cell such as proliferation or apoptosis, which can cause diseases such as cancer. Understanding the structure of Ras is crucial to understanding how mutations within these different interacting proteins affect the pathway.
The Ras construct we use for crystallographic purposes has 23 residues truncated from the C-terminus known as the hyper variable region. The resulting construct is a 166 amino acid residue protein with important structural features including the P-loop (residues 10-17), switch I (residues 30 – 40) and switch II (residue 60 – 76) (Milburn et al., 1990). Switch I and switch II are important in Ras because of the conformational changes that occur when GTP is hydrolyzed to GDP, leading to the name “molecular switch” (Milburn et al., 1990). The hydrolysis of GTP decreases the affinity for down-stream effector proteins, thus terminating the signal (Bourne, Sanders, & McCormick, 1991). Recently our group discovered a new allosteric site that modulates distinct conformational states associated with Ras-GTP (Buhrman, Holzapfel, Fetics, & Mattos, 2010). The “off” state of the allosteric switch in Ras has an empty allosteric site and a disordered switch II within the active site containing catalytic residue Q61 (Buhrman et al., 2010). The “on” state contains calcium and acetate bound at the allosteric site, leading to a shift in helix3/loop7 towards helix 4 that orders switch II, placing Q61 in the catalytic center (Buhrman et al., 2010). The acetate is thought to mimic an allosteric modulator with a possibly negatively charged membrane head group which together with calcium, can cause a catalytically active state which would promote intrinsic hydrolysis (Buhrman et al., 2010). The discovery of this site allowed us to further characterize Ras in the context of its two lobes, the effector lobe and allosteric lobe respectively. The effector lobe contains the main catalytic machinery including switch I, switch II, P-loop and most of the nucleotide binding pocket (Buhrman, O’Connor, et al., 2011). It also contains the binding sites for effector proteins (Buhrman, O’Connor, et al., 2011). The allosteric lobe, equivalent to lobe 2 in the second half, contains Ras regions that
interact with the membrane, allosteric site including residues R97, D107 and Y137 and the allosteric switch elements including helix3, helix4 and loop 7 (Buhrman, O'Connor, et al., 2011). This is considered the allosteric lobe because it connects the allosteric site through helix 3, at the edge of interlobal region, and switch II (Buhrman, O'Connor, et al., 2011).

Recently the MSCS of wild type Ras was published and compared to FTMAP results. Figure 2 shows the experimental MSCS results.

Figure 3. MSCS of Ras with clusters 1 – 8 shown in red spheres and organic solvents represented as sticks. Green depicts the effector lobe and catalytic machinery. (a) front side view of Ras (b) back side view of Ras. Grey depicts the allosteric lobe. Reprinted with permission from (Buhrman, O'Connor, et al., 2011).

Ras in complex with the GTP analogue GppNHp was crystallized and soaked in 9 different organic solvents with an unsoaked crystal being used as a control. The Ras crystals are solved in space group R32, a crystal form where switch I is stabilized by crystal contacts and is in the same conformation seen in the Ras/Raf-RBD complex (Fetics, 2012) and where
switch II is free of crystal contacts. This leads to an ordered switch I region and usually a disordered switch II region in the crystal structures with the allosteric switch in the “off” state. In the R32 crystal form, helix 4 and helix 5 are in regions of crystal contacts, precluding binding of all but the smallest solvent molecules in that region. Clusters 1 – 8 are in the interlobal region between helix 3 and switch II, in the allosteric lobe between helix 3 and 4, in the effector lobe opposite switch I relative to GppNHp, in the interlobal region between P-loop and N-terminus of helix 3, in the effector lobe near N-terminus of switch I, in the interlobal region near C-terminus of helix 3, and in the interlobal region between P-loop and switch I (Buhrman, O’Connor, et al., 2011). Many of these sites are in areas of protein-protein interactions including where importin-β binds the homologous GTPase Ran, membrane interacting region, Raf-RBD binding site, Raf-CRD binding site, and Ras GAP binding site. This further validated the usefulness for MSCS to depict hot spots that are of importance to a GTPase. However, as stated earlier crystal contacts can be a drawback and thus we used FTMAP to do computational solvent mapping and obtain a collection of predicted functional group clusters on the surface of Ras. The FTMAP results are highly complementary to the MSCS results.

Figure 4 shows the clusters that not only matched MSCS but other clusters that are undiscovered.
Figure 4. MSCS versus FTMAP results. Clusters in red spheres (MSCS) and clusters in purple spheres (FTMAP). Effector lobe shown in green and allosteric lobe shown in grey. (a) front side view of Ras (b) back side view of Ras. Reprinted with permission from (Buhrman, O'Connor, et al., 2011).

One advantage of FTMAP is its ability to not be constrained by crystal contacts. It also allows us to sample the “on” state of the allosteric switch of Ras since only the “off” state is accessible experimentally (Buhrman et al, 2011). Two FTMAP clusters overlapped with the most highly conserved MSCS clusters 1 and 2. However, as seen in figure 4 there are many clusters determined computationally that appear in regions of crystal contact and therefore are not observed experimentally. FTMAP clusters R128 and R135 are located between helices 4 and 5 (Buhrman, O'Connor, et al., 2011). Molecules in these clusters interact with residues R128 and R135 which are known to make salt bridges with membrane phospholipids (Buhrman, O'Connor, et al., 2011). FTMAP also picked out the allosteric site, which is expected since it can sample the catalytically “on” conformation of Ras. There was also the Y71 cluster in the effector lobe near switch II and the loop 7 cluster in the allosteric
lobe near loop 7, opposite the allosteric site (Buhrman, O'Connor, et al., 2011). MSCS and FTMAP are complementary methods, allowing us to pick out sites of membrane interactions as well protein-protein interactions. This study leads to the idea that oncogenic mutations may result in changes to the binding surface of Ras, leading to different positions for organic solvent clusters and water molecule networks. This was further explored by applying MSCS to Ras G12V and Q61L as detailed in Chapters three and four.

**BAF**

Barrier-to-autointegration factor (BAF) is an essential protein to understanding protein-DNA interactions. Its known function is to block autointegration of retroviral DNA and is usually within the cell cytoplasm and nucleus (Umland, Wei, Craigie, & Davies, 2000). The integration of the DNA copy of the viral genome into the cell host genome is essential to the virus life cycle. This integration process is mediated by the preintegration complex (PIC). The PIC was discovered in cells infected by Moloney murine leukemia virus (Mo-MLV) and human immunodeficiency virus type 1 (HIV-1) (Ellison, Abrams, Roe, Lifson, & Brown, 1990; Farnet & Haseltine, 1990; Fujiwara & Mizuuchi, 1988). This discovery led to the study of the process of autointegration, which happens when the virus integrates its own viral DNA and destroys the viral genome and itself (Umland et al., 2000). Viruses evolved over time to be able to prevent and protect against this process, which is why the BAF protein is necessary (Chen & Engelman, 1998; M S Lee & Craigie, 1994; Myung Soo Lee & Craigie, 1998). BAF plays an important role in many areas throughout the cell such as in nuclear assembly, organization in metazoans, and gene expression (Segura-Totten & Wilson, 2004). To determine these important areas a knock-down experiment of
BAF was conducted in Caenorhabditis elegans (Margalit, Segura-Totten, Gruenbaum, Wilson, & Kornberg, 2005; Zheng et al., 2000) and Drosophila melanogaster (Furukawa et al., 2003) which show the knock-down to be lethal. Even with all of these functional studies the endogenous function of BAF is still unknown. To understand the function of BAF one must not only understand its binding interactions but its overall structure as well.

BAF binds double stranded DNA, but not single stranded DNA, in a non-specific manner independent of the bases (Myung Soo Lee & Craigie, 1998). Non-specific binding gives BAF the ability to bridge multiple DNA segments together which is believed to be responsible for BAF’s autointegration protection function (Myung Soo Lee & Craigie, 1998). BAF is a biological homodimer, with each monomer containing 89 amino acid residues. This dimer appears in the asymmetric unit of the available crystal structure (Bradley, Ronning, Ghirlando, Craigie, & Dyda, 2005). Even though BAF has no known similarity to any other protein it is highly conserved across species (Umland et al., 2000). Figure 5 shows the published crystal structure of BAF, in which there is a 7-bp piece of DNA bound to each monomer on opposite ends of the homodimeric interface (Bradley et al., 2005).
The BAF monomer itself is made up of a Helix-hairpin-Helix (HhH) motif (Val20 – Glu35), pseudo HhH motif (Leu63 - Phe78) and a connecting helix (dimer interface helix) making up the five-helix (HhH)$_2$ motif (Shao & Grishin, 200). BAF makes contact with the minor groove of the DNA phosphate backbone regulated by the four mentioned motifs and N-terminus of the $\alpha$-helix within each monomer (Bradley et al., 2005). The DNA makes contacts with members of the HhH motif including the amide nitrogen atoms of Gly25, Gly27, Val29, and Leu30 hydrogen bonding to the adjacent phosphates in nucleotides 5 and 6 (Bradley et al., 2005). The Val29 also makes extra hydrophobic interaction with the sugar of nucleotide 5 (Bradley et al., 2005). The main chain carbonyl groups of residues Gly21, Ile26, and Leu23 form a pocket in which the $\varepsilon$-amino group of Lys6 is stabilized to hydrogen bond to the phosphate of nucleotide 6 (Bradley et al., 2005). The pseudo HhH motif has few interactions with the DNA however there is a hydrogen bond between the amide group of Ala71 and the phosphate group of nucleotide 1' and between the guanidine group of Arg75.
and phosphate of nucleotide 2' (Bradley et al., 2005). The N-terminus of α1 helix also forms hydrogen bonds between the amide groups of Gln5 and Lys6 and the phosphate within nucleotide 7 (Bradley et al., 2005). This published structure also has an example of a water mediated contact between Ser4 and the phosphate of nucleotide 6 (Bradley et al., 2005). All of these findings are consistent with reported mutagenesis data and a predicted computational model (Umland et al., 2000). The interacting regions of BAF are well characterized but it would be interesting to determine other areas of protein interactions. Water molecules tend to mediate contacts between DNA and proteins (Reddy, Das, & Jayaram, 2001) but this has not been heavily studied in BAF. MSCS can be used as a tool to perform such analysis and is presented in Chapter 5.

Conclusions

MSCS is a powerful tool for determining properties of binding sites and overall characteristics of proteins. My research contributes to both the development and validation of MSCS as well as to obtaining new information on P22TSP (sugar binding protein), Ras G12V and Ras Q61L (mutant GTP/GDP binding proteins with extensive sites of protein-protein interactions) and BAF (DNA binding protein). These analyses will increase our overall understanding of the strengths and limitations of MSCS considering these protein types have not been studied. Previous MSCS sets demonstrated the ability to not only pick out protein plasticity (Carla Mattos et al., 2006) and important areas of protein-protein interactions (Buhrman, O'Connor, et al., 2011), but also discovered new water networks never before seen in Ras (Kearney, 2012). This type of analysis has been applied to the current study. Chapter two describes P22TSP with discovery of a never before seen vestigial
active site and thorough water analysis in high agreement with published water molecule descriptions. Chapters three and four detail Ras G12V and Q61L MSCS comparison to the wild type with some cluster overlaps and a few novel clusters as well. A thorough water network analysis was also performed comparing old and newly discovered water networks. Chapter five describes BAF and its MSCS depiction of a site with a known protein-protein interaction. The thorough water analysis also depicts areas of DNA interactions and gives clues to general features of non-specific protein/DNA interactions. All of these projects help to expand the knowledge on the type of structural information obtained by using MSCS. By contributing multiple types of proteins we can have a broad MSCS overview to show that the method works for proteins in general, as long as it can be crosslinked with retention of diffraction after soaking in organic solvents. This could help to encourage MSCS as a useful method to probe protein surfaces. The current studies present the many different protein analysis types that can be accomplished through using the MSCS method with the hopes of eventually moving forward towards the bigger picture of drug design.
CHAPTER 2: THE MSCS OF P22TSP

Introduction

Tailspike proteins play an important role in the bacteriophage’s infection process (Steinbacher et al., 1994) and have become classic models for the study of folding of large proteins. These proteins recognize and cleave sugars on the membrane of bacteria, thus allowing for the injection of phage DNA into the cell. The understanding of tailspike protein-carbohydrate interactions is crucial for the understanding of biological recognition processes. In the case of these larger proteins in general very little is known of the interaction processes because of the lack of structural and thermodynamic data. The work presented here contributes structural information by exploring the interactions of phage P22 tailspike protein (P22TSP) not only with its natural ligand but other unnatural ligands as well. By more fully understanding the protein interface available for interactions with ligands it is possible to better understand how P22TSP affects the overall phage P22 virus formation. The tailspike protein in this study comes from salmonella phage P22, a 44 kilobase long dsDNA phage in the Podoviridae family (Steinbacher et al., 1997). The virus is a lysogenic bacteriophage that infects Salmonella. Its DNA is packaged into the capsid, once the cell is filled with viruses lysis occurs and releases the phage to infect other cells (Susskind & Botstein, 1978). The tailspike family does not contain binding sites rich in aromatic residues (Steinbacher et al., 1994). Instead, tailspike proteins contain a large interaction site recognizing oligosaccharides with high linkage specificity and sequence (Steinbacher et al., 1994). Through structural
studies the interactions of the ligand within these sites can be further characterized for better understanding of function.

P22TSP is biologically relevant as a trimer composed of three right-handed β-helices connected by its C-terminal trimerization domain (Steinbacher et al., 1996). Three to six homotrimers of P22TSP have to be attached to each phage for the virus particle to be infectious (Baxa et al., 2001). One glutamic acid and two aspartic acids make up the catalytic residues in the active site (Baxa et al., 1996). The active sugar-binding site is on the solvent exposed groove along the parallel β-helix (Steinbacher et al., 1996). P22TSP recognizes and cleaves the O-antigen moiety of the lipopolysaccharide of Salmonella belonging to serotype A, B, or D1 (Baxa et al., 1996). These different serotypes allow for various binding modes of ligands in a pocket within the active groove which enables phage activity on a different number of hosts (Baxa et al., 1996). A common characteristic of the tailspike family is O-antigens containing the main chain trisaccharide repeating unit α-D-mannose-(1→4)-α-L-rhamnose-(1→3)-α-D-galactose-(1→2) (Baxa et al., 2001). However, they show specificity in the 2,6-dideoxy-hexose substituent at C-3 of the mannose (Baxa et al., 2001). Phage 22 tailspike has endorhamnosidase activity and cleaves the α-L-Rha-(1→3)-α-D-Gal glycosidic bond in turn producing two repeating octasaccharide units (Baxa et al., 1996). The cleavage is required for infectivity and may facilitate virus access to the cell surface for DNA injection (Bayer et al., 1980). The function of P22TSP is well understood but there are still specific pieces to the puzzle that need further characterization. My study of the structural interactions will help to give clues to what structural changes occur during environmental changes. An isothermal calorimetry study shows binding to P22TSP is driven by electrostatics and
hydrophobic interactions between the protein and its sugar (Baxa et al., 2001). My study looks to further characterize these hydrophobic interactions using multiple types of ligands. These discoveries could lead to a better understanding of the overall tailspike family. Recent studies about genomes and characteristics of different phages have led to the discovery of other tailspikes, HK620TSP and Sf6TSP, from phages evolutionarily related to P22 (Chua, Manning, & Morona, 1999; Clark, Inwood, Cloutier, & Dhillon, 2001).

HK620TSP and Sf6TSP are native trimers with strong stability against proteases and heat (Barbirz, Becker, Freiberg, & Seckler, 2009). Crystal structures of Sf6TSP and HK620TSP show that they are structural homologues of P22TSP; and fold into large central $\beta$-helixes flanked by a C-terminal $\beta$-sandwich domain, which is responsible for stabilization of the trimer (Barbirz et al., 2008; Muller et al., 2008). Comparing the genome of the tailspike family SF6 and HK620 we see they are more evolutionarily related to each other than P22 (Casjens et al., 2004). The genomes of all three family members each code for a single tailspike protein (Barbirz et al., 2008). All three phages also share 70% sequence identity within the N-terminus while no major sequence similarity is within the C-terminus (Barbirz et al., 2008). Although all three phages have low sequence similarity within the active site of the C-terminus body they still have similar binding and cleavage activities. Sf6TSP binds and cleaves the O-antigen of the outer lipopolysaccharide of Shigella flexneri producing octasaccharides, while HK620TSP is an endo-N-acetylglucosaminidase that produces hexasaccharides in Escherichia coli (Barbirz et al., 2008; Muller et al., 2008). The sugars binding to Sf6TSP are in a non-canonical position within the long grooves at the subunit interfaces of the trimers (Muller et al., 2008). The catalytic residues of Sf6TSP (E366
and D399) are on two separate peptide chains flanking the binding groove (Muller et al., 2008). In comparison HK620TSP like P22TSP sugar binding sites are located on the parallel \( \beta \)-helices in a groove formed by a single polypeptide chain between two sets of loops flanking one of the \( \beta \)-sheets (Barbirz et al., 2008). An aspartate and glutamate are the catalytic residues of HK620TSP (D339 and E372) located along the solvent exposed groove of the \( \beta \)-helix of a single monomer where sugars bind (Barbirz et al., 2008; Muller et al., 2008). The binding grooves of P22TSP, Sf6TSP and HK620TSP all have similar dimensions as shown in Figure 1 (Barbirz et al., 2009).
When the sugars bind to the proteins, water is displaced from the binding site and about 60% of the sugar surface becomes buried. These sugars make direct and water-mediated hydrogen bond interactions. One of the overall characteristics of the family is the minimal amount of aromatic residues within the binding sites. However, the few aromatic residues that are present are thought to facilitate the positioning of the glycosidic bond for hydrolytic attack. These interactions and characterizations are pertinent to understanding the overall effectiveness of bacteriophage infection. Through using crystal structures we are able to observe details of the protein/ligand interactions at atomic resolution. These details help to
further identify structural properties that may affect the virus protein formation. A construct containing residues 109 to 666 (the C-terminal fragment) was used in the Multiple Solvent Crystal Structures method (MSCS) to probe the binding surface of the sugar-interacting domain. This construct maintains the enzymatic activity and lipopolysaccharide binding properties of P22TSP.

The MSCS method uses organic solvents to map protein surfaces (Allen et al., 1996). Only one neat organic solvent solution or an organic solvent water mixture is used at a time. By superimposing the structures solved in different organic solvents, protein-protein interactions, protein-ligand interactions, protein plasticity, and hydration characteristics can be visualized. In previous studies, MSCS correctly located known substrate binding sites for Elastase (Carla Mattos et al., 2006), RNAse A (M. Dechene et al., 2009), Ras (Buhrman, O'Connor, et al., 2011), Trypsin (Kearney, 2012) Chymotrypsin (Kearney, 2012), and RNase A (Kearney, 2012). Not only does this method allow us to find known active sites but also other binding pockets that may not be as well characterized. Analysis of the P22TSP MSCS was performed using DRoP to further study the binding sites for evolutionary comparison.

One of the advantages of the MSCS method is that in addition to providing a venue for the study of hot spots in protein binding sites, it also provides ample data for analysis of crystallographic water molecules. The Detection of Related Solvent Positions (DRoP) program (Kearney, 2012) was written to produce a better representation of conservation between structures in different environments. DRoP allows for consistent analysis by addressing of crystallographic symmetry, giving cluster quality reporting and allowing
cluster visualization (Kearney, 2012). Through the use of DRoP we are able to better analyze the effect of water and organic solvent molecules on protein surfaces. This analysis is pertinent when trying to relate binding sites and water networks to functions within the protein. Using the truncated version of P22TSP we look to further validate the MSCS method specifically within a large sugar binding protein.

**Experimental Methods**

P22TSP was obtained from our collaborator Stephanie Barbirz in the Steinbacher group as a purified protein solution (14.6. mg/ml of p22tspΔN in 10 mM HEPES buffer pH 7) (Steinbacher et al., 1996). P22TSP crystals are grown using hanging drop vapor diffusion at 4°C in 8 μl drops containing 5 μl of protein solution and 3μl of crystallization buffer over a 750 μl reservoir. Crystallization buffer contained 0.1 M Na₂HPO₄, 1.5 M (NH₄)₂SO₄, while reservoir solution contained 0.1 M Na₂HPO₄, 1.0 M (NH₄)₂SO₄ both adjusted with 50% NaOH to pH 10. Two weeks passed before high diffraction crystals formed. A picture of the formed crystals is shown in Figure 2.
P22TSP crystals are transferred with a cryo loop to a 10 µl drop of .08% gluteraldehyde (8.26% w/v in distilled water, pH 4.25) in stabilization buffer (0.1 M Na₂HPO₄, 1.0 M Na₂SO₄, 125 microliters of 1M HEPES, pH = 7.2) placed over 500µl reservoir (the same drop mixture) and crosslinking occurred at room temperature for 45 minutes. Cross-linked crystals are transferred to a new drop containing an organic solvent and reservoir solution to soak for 1 hour at room temperature. Crystal soaks include 60% acetone (ACT), 50% dimethylformamide (DMF), 50% glycerol (GOL), neat 1,6-hexandiol (HXD), 50% isopropanol (IPR), 50% t-butanol (TBU) 40% tetrafluoroethylene (ETF), 1.5M trimethylamine oxide (TMO) and 1.5M urea (URE). Successfully soaked crystals are collected and placed in cryo-protectant containing reservoir solution with 30% glycerol, and flash frozen using liquid nitrogen.
Diffraction data are collected at APS the SERCAT ID-22 beamline at 100 K using 1 Å wavelength radiation and a Mar300 CCD detector at a distance of 150 mm. The data are scaled and processed using HKL2000 (Otwinowski & Minor, 1997). Figure 3 shows an example of the diffraction pattern obtained from x-ray.

![Diffraction pattern](image)

Figure 3. Diffraction pattern of P22TSP diffracting better than 1.86 Å with detector at 250mm. Taken on home source at NC State University.

A model of P22TSP (PDB code 2VFM) with water molecules and alternate conformations removed was used to obtain an initial fit for all models. Refinement was accomplished using PHENIX (Adams et al., 2010). Coot was used to rebuild models and identify water and organic solvent molecules with the Fo-Fc electron density map contoured.
at 3σ and 2Fo-Fc electron density map contoured at 1σ (Emsley, Lohkamp, Scott, & Cowtan, 2010). Organic solvent molecules are placed in areas of well-defined electron density. All models are refined once without waters and twice with waters before placing in any organic molecules. Omit map calculations are used to confirm placement of organic solvents. All organic molecules had average B-factors similar to that of the overall model. All completed structures are run through phenix.xtriage and DRoP to statistically check reliability of the models. The first step of DRoP is optimization. Optimization calculates symmetry related molecules and scores the importance of those molecules in relation to protein contacts (Kearney, 2012). Next DRoP splits pdb files if it contains a dimer or higher order oligomers and then superimposes the structures through use of the Ce-align algorithm in PyMOL (Kearney, 2012). Lastly, DRoP groups clusters based on amount of conservation and then condenses clusters to ensure that molecule is in the most optimized position (Kearney, 2012). Figures 4 and 5 show examples of the DRoP output from previously run P22TSP structures.

<table>
<thead>
<tr>
<th>Res Num</th>
<th>Average Position</th>
<th>RMSD</th>
<th>Files</th>
<th>#files</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.779 30.174 -9.214</td>
<td>2.347</td>
<td>1</td>
<td>4 5 6 8 9 10 11 (10)</td>
</tr>
<tr>
<td>2</td>
<td>36.539 52.088 2.210</td>
<td>0.231</td>
<td>1</td>
<td>3 4 5 6 8 9 11 (8)</td>
</tr>
<tr>
<td>3</td>
<td>34.630 21.745 5.714</td>
<td>0.429</td>
<td>1</td>
<td>4 5 6 8 9 10 11 (8)</td>
</tr>
<tr>
<td>4</td>
<td>36.705 13.629 -2.896</td>
<td>0.725</td>
<td>1</td>
<td>3 4 6 7 8 9 11 (8)</td>
</tr>
<tr>
<td>5</td>
<td>14.963 -12.386 -18.076</td>
<td>0.480</td>
<td>1</td>
<td>3 4 6 8 9 11 (7)</td>
</tr>
<tr>
<td>6</td>
<td>45.102 25.969 35.597</td>
<td>0.565</td>
<td>2 3</td>
<td>6 8 9 11 (6)</td>
</tr>
<tr>
<td>7</td>
<td>27.006 11.306 -12.010</td>
<td>0.760</td>
<td>3</td>
<td>5 6 7 8 9 (6)</td>
</tr>
</tbody>
</table>

Figure 4. DRoP output of organic solvent cluster from P22TSP structures.
Figures 4 and 5 demonstrate that the DRoP program gives an output of numbered clusters at the average position of all the molecules making up that cluster. If a water or organic solvent contains the number 1 it is relative to Cluster 1. This cluster is the most highly conserved. For example, Cluster 1 from the organic output contains an organic molecule in 8 out of the 10 pdb files in that position. We consider clusters with at least two organic molecules to be of high importance within our analysis. The same idea can be seen in the water output where the most highly conserved cluster contains water molecules in 11 out of the 11 pdb files. For analysis of water molecules the cut off for highly conserved is around 75% or in our case a water at a given position in 8 out of the 11 pdb files. After the structures are run through
DROPCone more round of refinement is run and the resulting crystallographic statistics are shown in Table 1.
Table 1. Refinement statistics for P22TSP. Parentheses equal highest resolution shell.

<table>
<thead>
<tr>
<th>Data collection</th>
<th>Cross-linked</th>
<th>ACT 60%</th>
<th>DMF 50%</th>
<th>GOL 50%</th>
<th>HXD neat</th>
<th>IPR 50%</th>
<th>TBU 50%</th>
<th>TFE 40%</th>
<th>TMO 1.5M</th>
<th>URE 1.5M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>120.45, 119.81</td>
<td>119.87, 119.81</td>
<td>119.80, 120.35</td>
<td>119.84, 120.45</td>
<td>120.00, 120.45</td>
<td>120.78, 120.05</td>
<td>120.05, 120.05</td>
<td>120.05, 120.05</td>
<td>120.05, 120.05</td>
<td></td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>35-1.75</td>
<td>35-1.58</td>
<td>35-1.88</td>
<td>35-1.82</td>
<td>35-1.90</td>
<td>34-2.13</td>
<td>35-1.78</td>
<td>35-1.61</td>
<td>35-2.15</td>
<td>50-1.71</td>
</tr>
<tr>
<td>Rsym or Rmerge</td>
<td>0.116</td>
<td>0.158</td>
<td>0.188</td>
<td>0.182</td>
<td>0.190</td>
<td>0.213</td>
<td>0.178</td>
<td>0.161</td>
<td>0.215</td>
<td>0.171</td>
</tr>
<tr>
<td>I / σ</td>
<td>0.833</td>
<td>0.777</td>
<td>0.836</td>
<td>0.816</td>
<td>0.959</td>
<td>0.967</td>
<td>0.968</td>
<td>0.905</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>56.2</td>
<td>59.3</td>
<td>38.8</td>
<td>43.8</td>
<td>55.7</td>
<td>30.2</td>
<td>48.5</td>
<td>58.4</td>
<td>33.8</td>
<td>36.3</td>
</tr>
<tr>
<td>Redundancy</td>
<td>9.99</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R.M.S.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>1.114</td>
<td>1.08</td>
<td>1.05</td>
<td>1.10</td>
<td>1.09</td>
<td>1.09</td>
<td>1.18</td>
<td>1.14</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>Bond angles (Å)</td>
<td>1.114</td>
<td>1.08</td>
<td>1.05</td>
<td>1.10</td>
<td>1.09</td>
<td>1.09</td>
<td>1.18</td>
<td>1.14</td>
<td>1.11</td>
<td></td>
</tr>
</tbody>
</table>
Results and Discussion

The crystallized form of P22TSP is shortened and consists of amino acid residues 109 to 666 (Steinbacher et al., 1994), in cubic space group P2\textsubscript{1}3. P22TSP is biologically relevant as a homotrimer (Steinbacher et al., 1994). Only one protein molecule is seen in the asymmetric unit. However, the homotrimer is formed through the three-fold crystallographic symmetry. P22TSP monomer has an overall fish shape, composed of six main segments: the main body, mouth, dorsal fin, and first, second, and third segments of the caudal fin (Steinbacher et al., 1994). The main body of P22TSP is characterized by its large right-handed parallel β-helix of 13 complete turns (Steinbacher et al., 1994). Near the carboxyl-terminal of tailspike the three subunits are interdigitated which cause high thermal stability (Steinbacher et al., 1994). Looking at the cross section of the β-helix a triangular shape is seen forming a long groove on the solvent-exposed side (Steinbacher et al., 1994). This solvent exposed area is between an irregular domain and loops within turns 5, 7, and 8 of the β-helix (Steinbacher et al., 1994). Within this groove is where the polysaccharide binding site of tailspike is contained. The main active site of P22TSP contains an endoglycosidase mechanism with Asp392 as the acid and a catalytic water molecule being activated by Asp395 and Glu359 as the bases (Steinbacher et al., 1994). This is seen within the active site of all solved structures. All of the structures are superimposed for cluster analysis. The resulting structures contained 19 unique clusters with 12 of them containing at least two molecules and therefore considered significant. Table 2 shows the 12 clusters with the amino acid residues present in the respective binding pocket interactions.
Table 2. Clusters with number of organic molecules in parentheses and their protein interactions.

<table>
<thead>
<tr>
<th>Cluster Number</th>
<th>Residue Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (8)</td>
<td>Lys317, Asn340, Asn378, Asp411, Tyr412, Pro423, Asn425,</td>
</tr>
<tr>
<td>2 (8)</td>
<td>His420, Leu466, Ser493, Gly494</td>
</tr>
<tr>
<td>3 (8)</td>
<td>Asp395, Asp399, Gly442, Gly443, Asp444</td>
</tr>
<tr>
<td>4 (7)</td>
<td>Glu359, Lys363</td>
</tr>
<tr>
<td>5 (7)</td>
<td>Pro131, Arg148, Asn151</td>
</tr>
<tr>
<td>6 (6)</td>
<td>Ser557, Lys561</td>
</tr>
<tr>
<td>7 (5)</td>
<td>Trp365</td>
</tr>
<tr>
<td>8 (3)</td>
<td>Asp314</td>
</tr>
<tr>
<td>9 (3)</td>
<td>Asn520</td>
</tr>
<tr>
<td>10 (2)</td>
<td>Arg497</td>
</tr>
<tr>
<td>11 (2)</td>
<td>Asp612, Arg614</td>
</tr>
<tr>
<td>12 (2)</td>
<td>Lys291</td>
</tr>
</tbody>
</table>

In Figure 6 Cluster 7 (DMF, GOL, TBU and TFE) and Cluster 4 (ACT, DMF, GOL and TBU) are seen within the ligand binding cavity aligning with sugar rings 3,6-dideoxy-alpha-D-ribo-hexopyranose (PZU) and alpha-L-rhamnose (RAM) respectively. Cluster 4 (DMF, GOL, IPR, TFE and TBU) also interacts with catalytic site of P22TSP, more specifically catalytic residue Glu359, with all organic solvents being within 3.4 Å or less.
Plasticity can be seen in Glu359 with solvent exposed structures maintaining the conformations opposite the sugar soaked structure (green). Catalytic residues ASP395 and ASP392 have the same conformation in all structures. Further along the ligand binding cavity Cluster 3 (GOL, IPR, TMO, TFE, and URE) is seen within a deep pocket leading out of the active site. Cluster 3 is also making direct interactions with catalytic residue Asp395.

Previous research only allowed for the binding of a sugar containing 8 rings. This is because P22TSP is still catalytically active even in its truncated version. Cluster 3 is believed to be delineating the pocket where a longer sugar ring would reside in P22TSP if cleavage did not occur. A crystal structure of a catalytically inactive P22TSP in complex with a longer sugar
moiety would give further insights into this active site and such studies are planned for the near future.

As we look further from the active site moving towards the more solvent exposed side of P22TSP, Cluster 2 (GOLs), Cluster 9 (GOL and IPR) and Cluster 10 (DMF and TBU) are seen in Figure 7.

![Figure 7](image_url)

Figure 7. Clusters further from the active site. Figure created in PyMOL (Schrodinger, 2010).

Cluster 2 is within a pocket and comprised of only Glycerol, possibly picking out a site at which a sugar ring interacts. While Clusters 9 and 10 are not interacting with any residues of importance within the literature, they could possibly be further mapping how the ligand
moves through the binding cavity. Currently, a structure of tailspike with its O-antigen receptor only contains two repeating units (Steinbacher et al., 1996). This smaller unit doesn’t allow us to see how the sugars fully fit within the active site. However, studies have shown that P22TSP can bind octasaccharide and dodecasaccharide fragments (Baxa et al., 1996). These longer sugar fragment rings and our organic solvent clusters shown in Figures 2 and 3 of P22TSP should align. The current published alignment of the ligand shows the power of the MSCS method in picking known active sites that reside in hydrophobic regions of the protein. An alignment of a longer ligand would give clues to how the sugar interacts within the active site and further along the active site.

Another area of interest shown in Figure 8 is near a solvent exposed area of tailspike opposite the active site. Cluster 1 contains a glycerol, hexandiol, isopropanol, t-butanol, TFE and TMAO, while Cluster 8 only contains glycerol.
Figure 8. Vestigial Active site. Figure created in PyMOL (Schrodinger, 2010).

Again, these organic solvents are contained within a hydrophobic pocket. This subsite currently has no known function in P22TSP; however, this site is located where the active site is located in tailspike family member SF6. The HK620 and P22TSP family members’ endoglycosidase active sites are intramolecularly located while SF6 has its active site at the interface between subunits (Barbirz et al., 2008). If Cluster 8 is to be superimposed on Sf6TSP this cluster in P22STP would be within the known active site of Sf6TSP. This is similar to those in the pocket within the P22TSP active site shown in Figure 6 Cluster 7. Cluster 1 would also superimpose within Sf6TSP’s active site and be comparable to Cluster 4 within P22TSP seen in Figure 6. The location of these organic molecules between subunits
could be indicative of a vestigial active site that disappeared in P22TSP during the evolution of the protein. Alternatively, the Sf6TSP was the one that diverged and it would be interesting to test whether a vestigial site exists in this protein at the location where the active site is located in the P22TSP and HK620 family. Figure 9 shows locations of the remaining cluster sites in various locations throughout the protein surface.

Figure 9 shows the locations of the remaining highly conserved clusters. Cluster 5 (contains a glycerol molecule from ACT, DMS, HXD, IPR, TBU, TFE, TMO and URE) is located near the N-terminus in the mouth region of P22TSP. Cluster 12 (contains a glycerol molecule
from GOL and HXD) is located on the backside of P22TSP closest to the ventral fin. Clusters 6 (contains DMF, GOL, IPR, URE, TBU and TFE organic molecules) and 11 (contains a TBU and URE organic molecule) are located in areas interacting with the caudal fin II of P22TSP. Even though these clusters are conserved within the family, there is no known function in the current literature for the residues with which they interact. However cluster 5 is near the N-terminus known to interact with the head binding domain which is essential for non-covalently attaching P22TSP to the DNA injection apparatus (Steinbacher et al., 1997). Clusters 6 and 11 are near the caudal fin that help mediate interactions between the virus capsid and the outer membrane of the pre-infected cell (Andres et al., 2010). More experiments are needed to determine if these are areas of protein-protein interactions (cluster 5) or even areas of protein-membrane interactions (clusters 6 and 11).

**Water Analysis**

Water analysis using the DRoP program was conducted to view protein hydration in different organic solvent environments. There is no previous research that focuses on the overall analysis of P22TSP’s waters within different environmental conditions. Overall there are 579 unique clusters with 244 water-binding sites that maintain 80% or better conservation within the structures (conserved waters). This means that out of the 10 structures at least 8 structures contained water in that position. As seen in Figure 10, the waters are color coded by percent conservation: 100% in red, 90% in orange, and 80% in yellow.
Figure 10 shows most of P22TSP is hydrated near the surface of the protein except in places of organic solvent molecules bind, where there are no conserved waters. These findings are in line with what is usually seen within a MSCS (Carla Mattos et al., 2006; C. Mattos & Ringe, 2001). The structural waters are always highly conserved within the protein surface. These waters are believed to be a part of the protein structure and are important for mediating hydrogen bond interactions. These interactions can include involvement in catalysis and protein-protein interactions. The displaceable waters are not as highly conserved. These waters are believed to easily displaced and important for protein-ligand interactions as seen in my P22TSP structures. The waters may mediate the first contacts between the organic solvent molecules and the protein then move to allow direct organic solvent protein interactions. The MSCS structures are run in DRoP to get a feel for how the organic solvent environments affected the overall water networks. Then DRoP was run again on the set of MSCS structures along with the published structure containing the natural
ligand (1TYU). By running the analyses with the ligand structure we can compare water networks present when P22TSP is in the natural bound ligand conformation with the presence of other organic solvent induced conformations. This allows us to compare the changes when the ligand is bound within the entire active site to the smaller more local changes that occur when small organic molecules are bound within different pockets.

Figure 11 shows that the active site water 104 that mediates the sugar cleavage reaction is 100% conserved within this set of 11 structures. This catalytic water is essential for the cleavage of sugars within the active site. The fact that it is 100% conserved shows that
this water molecule is present even in the absence of a substrate as an integral part of the active site.

Figure 12. Waters interacting with the natural ligand. Figure created in PyMOL (Schrodinger, 2010).

In Figure 12 a more detailed analysis of the active site containing the ligand shows 8 water molecules that interact with the ligand, including some that mediate contacts with protein. The waters shown are 101, 539, 104, 88, 49, 300, 78 and 542 with 100%, 0%, 100%, 100%, 100%, 50%, 100% and 0% conservation respectively. These waters are shown to be essential for the placement of the ligand within the active site (Steinbacher et al., 1997).
Water molecules that are 0% conserved are those that appear in the sugar-bound structures but are not present in the MSCS set.

In Figure 13 water 542 is only present when the natural ligand is bound to the protein. The Asp309 residue shown in green is in this conformation when the ligand is bound and makes direct contact with water 542. The unbound conformation is seen in all other structures. Water 542 shifts 1.6 Å to water 322 (50% conversation), accompanying the side chain to which it remains hydrogen bonded. This is an example of how residues change conformations depending on interactions. When the ligand is present this residue swings
inward for direct hydrogen bond interactions. However, when the ligand is absent the residue swings outward to form better hydrogen bond interactions within the protein itself.

Figure 14. Glu359 alternate confirmation. Figure created in PyMOL (Schrodinger, 2010).

In Figure 14 more alternate conformations between the ligand bound and unbound structures can be seen. When the ligand is bound the catalytic residue Glu359 interacts directly with the water 539 and the ligand. Water 539 is only seen in the ligand bound structure directly interacting with the catalytic residue. Upon the ligand leaving the binding pocket Glu359 shifts 4.6 Å with the hydroxyl group replacing water 539. Water 542 and 539
are important for protein-ligand interactions. This is evident in the fact that these two water binding sites are only occupied when the ligand is present.

![Image](image1.png)

**Figure 15.** Water interactions near vestigial active site. Figure created in PyMOL (Schrodinger, 2010).

Taking a closer look at the side view of P22TSP a hydrophobic pocket can be seen in which the organic solvents fit nicely, as shown in Figure 15. There are only a few conserved water molecules within and surrounding the hydrophobic pocket. This is similar to what is seen in other MSCS projects where many organic binding sites outside of the active site are in locations of easily displaced waters. This is different than what is seen in the active site where many highly conserved waters are seen mediating contacts between the ligand and the
organic molecules. There is currently no known function of this pocket in P22TSP. Therefore, further experiments are needed to relate the discovery of these binding sites to function within the protein.

Conclusions

Due to the increasing number of viral genomes it is important to continue studying the bacteriophage infection machinery such as tailspike. However, the high sequence diversity within homologous and evolutionarily conserved tailspike families causes difficulty when trying to identify function from sequence alone (Casjens et al., 2004). This increases the importance of the need for structural information. Through use of MSCS we are able to not only add to structural information but also to do an analysis of the binding sites on the protein surface. Phages have many advantages most importantly as models for human viruses. When comparing phages to human viruses, animal viruses, plant viruses, and insect viruses, phages have the ability to be grown fast and easy with less demands on space, equipment, and facilities (Pelczar MJ (Jr), 1988). This ease of use allowed for the virus reproduction cycle to be first modeled through work with phages (Ackermann & Nguyen, 1983). MSCS was used to add a model system such as P22TSP to the current experimental data. This is the first time MSCS has been applied to a large sugar binding protein, and was not only able to pick out the known active site but a vestigial site as well. Further mutation experiments changing this vestigial site to be similar to SF6TSP’s active site will be needed to determine its properties. There are multiple clusters picking out the sugar binding pockets and even interacting with the catalytic residues. Moving further along the active site additional clusters are seen. The next step for this project is to bind a longer sugar ligand
within this pocket to delineate the sugar binding pockets moving towards the bulk solvent. This would give us an idea of how P22TSP interacts with its ligand through the entire active site and should be comparable to the MSCS organic solvents. This could also lead to further experiments to better determine the importance of tailspike proteins in virus infection due to its interactions with the sugar moiety. The vestigial site was located in a similar location to the active site of tailspike family member SF6. This discovery supports that MSCS can reveal sites that are evolutionarily changed across family members. These results could give evidence to an evolutionary divergence within a family where both sites are still present but have different functions across family members. This also gives further evidence to the published data that suggests each of the tailspike members are derived from a common ancestor (Barbirz et al., 2008). More research is needed to determine if this vestigial site plays an important role within P22TSP. Water analysis allowed us to see that the catalytic water is 100% conserved no matter the crystal environment. Through water analysis we are able to also pick out waters of high importance to the ligand binding pocket. Many of the waters are highly conserved. Several water molecules that are not highly conserved are replaced with amino acid side chains substituting the water interaction. This analysis allowed us to pick out the subtle changes within the pocket of P22TSP with no ligand bound compared to the ligand bound structure. The vestigial site showed an example of a standard MSCS pocket, showing a hydrophobic pocket surrounded by highly conserved water molecules. This pocket is able to be traced back to the active site through hydrogen bond interactions. With the emergence of more experimental evidence of allosteric sites within proteins future studies are needed to determine if this vestigial site can somehow affect the
active site during the infection process. MSCS successfully mapped the protein surface of P22TSP. This is the first time in which results are not only used from a test set but also from a real world application. Although we know the active site of P22TSP, little is known about other binding sites of importance. We hope these discoveries will lead to experiments showing that our newly discovered sites have functional roles in P22TSP.

Acknowledgments

Thanks to Dr. Stefanie Barbirz for purified P22TSP and giving a presentation that helped me realize the significance of my results. Thanks to Dr. Paul Swartz for countless suggestions on how to perform crosslinking and soaking of P22TSP crystals. Thanks to Dr. Bradley Kearney for all the help on using the DRoP program including running sets and teaching me how to run sets on my own.

Contribution

I grew crystals, soaked, collected data and analyzed all the structures presented here. Protein was provided by Dr. Stefanie Barbirz. Data are collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-BM 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. This research was supported by the NSF (MCB - 1237512) grant.
CHAPTER 3: THE MSCS OF THE RAS G12V MUTANT

Introduction

Ras is a monomeric GTPase that functions as a molecular switch in signal transduction pathways within the cell (Bourne et al., 1990). Ras is in its inactive form when bound to GDP and active form when bound to GTP (Bourne et al., 1990). The nucleotides are regulated through binding partners such as GTPase activating proteins (GAPs) and guanine nucleotide exchange factors (GEFs) (D. K. Morrison & Cutler Jr, 1997). GAPs cause Ras to switch to the inactive state by greatly enhancing the hydrolysis rate of GTP to GDP, while GEFs promote Ras activation by catalyzing the exchange of GDP for GTP (Campbell et al., 1998). This is seen within all Ras isoforms including H-Ras, N-Ras, and K-Ras. These three isoforms differ in the types of posttranslational modifications and sequence of the hypervariable region (Hancock, 2003). The catalytic domain of Ras has two lobes. Lobe 1 contains the first half of the molecule (residues 1-86) and is 100% conserved among the three isoforms; Lobe 2 consists of the C-terminal half (residues 87-171) and is 90% conserved (Hancock, 2003). Lobe 2 contains the structural elements of an allosteric switch mechanism that promotes a disorder-to-order transition in the active site (Buhrman et al., 2010). This mechanism has been associated with the Ras/Raf/MEK/ERK pathway that controls cell proliferation (Buhrman, Kumar, Cirit, Haugh, & Mattos, 2011). Since the three Ras proteins are very similar we use H-Ras to make inferences that are biologically relevant toward the entire family.
The Ras/Raf/MEK/ERK pathway is highly conserved and plays a role in several biological processes (Campbell et al., 1998). SOS (son of sevenless) is recruited when a growth factor binds to its receptor; this process converts the membrane bound form of Ras-GDP to Ras-GTP (Campbell et al., 1998). In the GTP-bound form Ras interacts with Raf, recruiting it to the cell membrane for propagation of the signal. (D. K. Morrison & Cutler Jr, 1997; Vojtek et al., 1993). This activation in turn causes Raf to bind and phosphorylate MEK1 and MEK2 stimulating ERK1 and ERK2 by modifying their phosphorylation on tyrosine and threonine residues (Campbell et al., 1998; D. K. Morrison & Cutler Jr, 1997). ERK translocates into the nucleus where it interacts with transcription factors, resulting in modulation of various cellular processes associated with the cytoskeleton, cytoplasm, nucleus and membranes (Roux & Blenis, 2004). Thus, Ras ultimately affects multiple targets such as membrane proteins, protein kinases, and transcription factors (Sacks, 2006). Mutations of proteins along this pathway cause constitutive signal activation, leading to the development of cancer.

Ras is essential for controlling differentiation, proliferation and the basic survival of eukaryotic cells and is mutated in about 20% of all cancers (Fernandez-Medarde & Santos, 2011). The mutations are located in two general hotspots near G12 and Q61 (Fernandez-Medarde & Santos, 2011). In H-Ras the mutations show a pattern with 54% of mutations in codon 12, 35% of mutations in codon 61, and 9% of mutations in codon 13 (Fernandez-Medarde & Santos, 2011). These mutations lead to a decrease in GTP hydrolysis, essentially leaving Ras in the active conformation that leads to disease. Since Ras is prevalent in many diseases it is highly studied for drug therapy. However, considering the flexibility of the
active site, the ability to make useful drugs is limited. Previous research has shown that Ras has several areas of protein/ligand interaction away from the active site (Buhrman, O'Connor, et al., 2011). These findings are based on wild type Ras and thus in order to discern changes in mutant structures we must first delve into the wild type structure.

The “molecular switch” in Ras is based on conformation changes within the switch I (residues 30-40) and switch II (residues 60-76) regions that occur when GTP is hydrolyzed to GDP (Milburn et al., 1990). Once GTP is hydrolyzed the upstream signal is terminated and affinity for down-stream effector proteins is lost (Bourne et al., 1991). Recently, an allosteric switch mechanism was discovered and believed to result from the equilibrium between two conformational states of Ras-GTP (Buhrman et al., 2010). The “off” state has an “empty” allosteric site and disordered active site, mainly switch II, containing the catalytic residue Q61 (Buhrman et al., 2010). The “on” state contains a calcium acetate bound at a remote site, a shift in helix3/loop7 toward helix 4 and ordered active site with Q61 near catalytic center (Buhrman et al., 2010). This calcium acetate is thought to mimic an allosteric modulator within the cell causing a catalytically active state that would promote intrinsic hydrolysis (Buhrman et al., 2010). These structural characteristics allow us to further categorize Lobe 1 and Lobe 2 into the effector lobe and allosteric lobe, respectively. Lobe 1 (effector lobe) mostly contains the catalytic site including switch I, switch II, the P-loop (residues 10-17) and most of the nucleotide binding pocket (Buhrman, O’Connor, et al., 2011). This lobe makes many protein-protein interactions with effector proteins of Ras (Buhrman, O’Connor, et al., 2011). Lobe 2 (allosteric lobe) contains the parts of Ras interacting with the membrane, allosteric site with residues R97, D107 and Y137, and the allosteric switch components
(helix 3/loop 7/helix 4) (Buhrman, O'Connor, et al., 2011). This lobe connects the active and allosteric sites of Ras through coupling helix 3 at the edge of the interlobal region and switch II (Buhrman, O'Connor, et al., 2011). The allosteric site/lobe is currently under intense investigation in our lab. Thus it is timely to also study the similar sites and networks within the most important Ras mutants. The Multiple Solvent Crystal Structures (MSCS) method was applied to wild type Ras-GppNHp. This method identified several binding site hot spots through analysis of organic solvent clustering on the surface of Ras. We have also applied the MSCS method to study the surfaces of the oncogenic mutants G12V and Q61L and present here the results for RasG12V in comparison to those of wild type Ras.

The MSCS method (Allen et al., 1996) was applied to a truncated version of Ras G12V containing 166 amino acids. Crystals of Ras are soaked in either a water organic solvent mixture or neat solvent. This allows us to see areas of protein-protein interactions, protein-ligand interactions, protein plasticity, and differences in water networks throughout a family. Previous research successfully applied this method to Elastase (Carla Mattos et al., 2006), RNAse A (M. Dechene et al., 2009) and Ras (Buhrman, O'Connor, et al., 2011). Previous MSCS of the wild type Ras picked out interactions sites with the Raf Ras Binding Domain (Raf-RBD), the Raf Cysteine Rich Domain (Raf-CRD) (Buhrman, O'Connor, et al., 2011) and with GAP. These areas of interaction are on areas of the protein away from the active site. This paper shows the power of MSCS in using small molecules to find possible druggable sites on different areas on the surface of Ras. By using a similar experiment with the mutant proteins we can determine whether there are hot spots that only occur when the mutation is present. To better organize our structures we employ the use of the Detection of
Related Solvent Positions (DRoP) program. This program was written to allow a consistent analysis, cluster quality reporting and visualization, and account for crystallographic symmetry (Kearney, 2012). This consistent analysis allows us to make more accurate conclusions across a family of proteins. For our study Ras, G12V mutant was made using PCR within a PET 21 vector, purified and crystalized for organic solvent soaks.

**Experimental Methods**

*Mutagenesis, Expression and Purification*

The G12V mutant was made using site-directed mutagenesis of wild type H-Ras (residues 1-166), using the Quick Change kit from Stratagene, following manufacturer’s instructions. H-RasG12V was then expressed in *Escherichia coli* BL21 cells and purified with GDP bound as previously published (Buhrman et al., 2003). The GDP was exchanged for the non-hydrolyzable GTP analogue GppNHp and then crystallized in 200mM CaCl₂ and 20% PEG 3350 as previously described (Buhrman, Wink, & Mattos, 2007).

*Crosslinking and Organic Solvent Soaks*

H-Ras-GppNHp crystals are transferred to 9-well sitting-drop glass plates with 50 µl of stabilization buffer [20 mM Hepes (pH 7.5), 50 mM NaCl, 5 mM MgCl₂, 1 mM DTT, 200 mM CaCl₂ and 25% PEG 3350]. The buffer was then exchanged multiple times with only stabilization buffer and finally with stabilization buffer containing 1% gluteraldehyde. The cross-linking reaction was allowed to go for 30 minutes to 1 hour before fresh stabilization buffer was added to remove unreacted gluteraldehyde. Crystals are then soaked in the organic solvent solutions for at least 1 hour as specified in the results.
Data Collection and Structure Refinement

Data were collected at the Synchrotron on the Southeast Regional Collaborative Access Team 22-ID beamline at the Advanced Photon Source (Argonne National Labs). The source had an X-ray wavelength of 1.0 Å and a MAR CCD detector was used to collect reflections. The published structure of wild type Ras with PDB code 2RGE was used as a molecular replacement model for phasing. Python-based Hierarchical ENvironment for Integrated Xtallography (PHENIX) was used for reciprocal space refinement with 10% of the unique reflections set aside for calculation of $R_{\text{free}}$ (Adams et al., 2010). Coot was used for visualization of electron density maps with the Fo-Fc contoured at $3\sigma$ and 2Fo-Fc contoured at $1\sigma$ (Emsley et al., 2010). Crystallographic water molecules are added for a couple of rounds of refinements, then organic solvent molecules are added and validated using omit maps. PDB and .cif files are made for the organic molecules using the PHENIX function for ligands. Finished structures are then run through DROp for analysis. This allowed the structures to have the organic solvent and water molecules numbered based on levels of conservation providing a better way to consistently discuss clusters. The structures are then superimposed back on the original structures and a final round of refinement was run for consistency. The results of the analysis are given below.
Results

Each RasG12V-GppNHp crystal was cross-linked in gluteraldehyde and soaked in a high concentration of organic solvents. These crystals are soaked in acetone (ACT), dimethylformamide (DMF), dimethyl sulfoxide (DMS), 2,2,2-trifluoroethanol (ETF), glycerol (GOL), neat hexane (HEX), R,S,R-bisfuranol (RSF), S,R,S-bisfuranol (RSG), t-butanol (TBU), and trimethylamine N-oxide (TMO). The cross-linked structure (X-Link) in aqueous solution was used as a control. Table 1 shows the overall data collection and structure refinement statistics for RasG12V crystals soaked in organic solvents.
<table>
<thead>
<tr>
<th>Data Collection</th>
<th>X-Link</th>
<th>ACT 70%</th>
<th>DMF 50%</th>
<th>DMS 50%</th>
<th>ETF 40%</th>
<th>GOL 50%</th>
<th>HEX Neat</th>
<th>RSF 80%</th>
<th>RSG 35%</th>
<th>TBU 40%</th>
<th>TMO 2M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td></td>
</tr>
<tr>
<td>Cell dimensions</td>
<td>a, b, c (Å)</td>
<td>88.91, 88.91, 135.09</td>
<td>89.13, 87.82, 133.96</td>
<td>87.89, 89.10, 134.29</td>
<td>89.10, 88.34, 134.27</td>
<td>89.33, 89.33, 134.46</td>
<td>88.34, 89.33, 135.38</td>
<td>89.17, 88.17, 135.24</td>
<td>88.57, 88.17, 134.97</td>
<td>87.87, 87.87, 134.43</td>
<td>87.95, 87.95, 134.09</td>
</tr>
<tr>
<td>a, b, c, γ (°)</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td>90.90, 120, 60</td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>50 - 1.29, 1.70</td>
<td>50 - 1.70, 1.72</td>
<td>50 - 1.64, 1.78</td>
<td>50 - 1.35, 1.67</td>
<td>50 - 1.67, 1.76</td>
<td>50 - 1.76, 1.76</td>
<td>50 - 1.64, 1.75</td>
<td>50 - 1.55, 1.64</td>
<td>50 - 1.53, 1.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R_{sym} or R_{merge}</td>
<td>(1.35 - 1.29)</td>
<td>(1.35 - 1.29)</td>
<td>(1.35 - 1.29)</td>
<td>(1.35 - 1.29)</td>
<td>(1.35 - 1.29)</td>
<td>(1.35 - 1.29)</td>
<td>(1.35 - 1.29)</td>
<td>(1.35 - 1.29)</td>
<td>(1.35 - 1.29)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>l &amp; σ</td>
<td>64.01, 46.2, 7.7, 10.8</td>
<td>318, 28.5, 9.0, 24.8</td>
<td>290, 24.8, 317, 51.4</td>
<td>290, 24.8, 317, 51.4</td>
<td>290, 24.8, 317, 51.4</td>
<td>290, 24.8, 317, 51.4</td>
<td>290, 24.8, 317, 51.4</td>
<td>290, 24.8, 317, 51.4</td>
<td>290, 24.8, 317, 51.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>(96.3)</td>
<td>(98.5)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(100)</td>
<td>(99.6)</td>
<td>(98.3)</td>
<td>(91.6)</td>
<td>(99.9)</td>
<td></td>
</tr>
<tr>
<td>Redundancy</td>
<td>9.6</td>
<td>9.4</td>
<td>4.9</td>
<td>8.8</td>
<td>4.9</td>
<td>8.8</td>
<td>4.9</td>
<td>8.8</td>
<td>4.9</td>
<td>8.8</td>
<td></td>
</tr>
<tr>
<td>Refinement</td>
<td>Resolution (Å)</td>
<td>1.29</td>
<td>1.70</td>
<td>1.72</td>
<td>1.64</td>
<td>1.78</td>
<td>1.35</td>
<td>1.67</td>
<td>1.76</td>
<td>1.64</td>
<td>1.55</td>
</tr>
<tr>
<td>No. reflections</td>
<td>51224</td>
<td>22709</td>
<td>26401</td>
<td>25154</td>
<td>22802</td>
<td>50066</td>
<td>20160</td>
<td>22239</td>
<td>27696</td>
<td>31067</td>
<td>35549</td>
</tr>
<tr>
<td>R_{work} / R_{free}</td>
<td>20.39/ 21.86</td>
<td>18.18/ 21.09</td>
<td>17.76/ 18.84</td>
<td>17.87/ 20.00</td>
<td>17.04/ 19.06</td>
<td>18.79/ 17.84</td>
<td>18.84/ 22.27</td>
<td>18.34/ 20.53</td>
<td>17.11/ 19.43</td>
<td>18.96/ 20.47</td>
<td>19.98/ 19.09</td>
</tr>
<tr>
<td>No. atoms</td>
<td>Protein</td>
<td>1311</td>
<td>1335</td>
<td>1327</td>
<td>1327</td>
<td>1326</td>
<td>1368</td>
<td>1326</td>
<td>1319</td>
<td>1334</td>
<td>1345</td>
</tr>
<tr>
<td>Magnesium/Ca 2+/Cl</td>
<td>0/ 2/2</td>
<td>1/ 2/2</td>
<td>4/ 2/2</td>
<td>3/ 2/2</td>
<td>1/ 2/2</td>
<td>9/ 2/2</td>
<td>2/ 2/2</td>
<td>1/ 2/2</td>
<td>1/ 2/2</td>
<td>1/ 2/2</td>
<td>1/ 2/2</td>
</tr>
<tr>
<td>GppNHp</td>
<td>1/ 1</td>
<td>1/ 1</td>
<td>1/ 1</td>
<td>1/ 1</td>
<td>1/ 1</td>
<td>1/ 1</td>
<td>1/ 1</td>
<td>1/ 1</td>
<td>1/ 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>134</td>
<td>76</td>
<td>100</td>
<td>97</td>
<td>96</td>
<td>107</td>
<td>63</td>
<td>107</td>
<td>63</td>
<td>63</td>
<td>74</td>
</tr>
<tr>
<td>RMSD</td>
<td>Bond lengths (Å)</td>
<td>.007</td>
<td>.006</td>
<td>.007</td>
<td>.007</td>
<td>.006</td>
<td>.007</td>
<td>.007</td>
<td>.007</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td>Bond angles (Å)</td>
<td>1.03</td>
<td>1.14</td>
<td>1.15</td>
<td>1.17</td>
<td>1.11</td>
<td>1.21</td>
<td>1.19</td>
<td>1.18</td>
<td>1.24</td>
<td>1.18</td>
<td>1.10</td>
</tr>
</tbody>
</table>
As seen in Figure 1, all of the RasG12V structures are similar to the published structure with PDB Code 2RGE in terms of the helix 3/loop7 structural elements associated with the allosteric switch mechanism. This means that these structures are all in the ‘off” state of the allosteric switch.

Figure 1. All RasG12V structures resulting from the various soaks. The 2RGE model in yellow represents the “off” state of the allosteric switch and the 3K8Y model in green represents the “on” state. The structures soaked in organic solvents are represented as follows: ACT (pale green), DMF (tv_yellow), DMS (pale yellow), GOL (yellow orange), HEX (limon), RSR (wheat), SRS (sand), TBU (smudge), TFE (magenta), TMO (chartreuse) and X-link (split pea). P-loop (black) Switch I (blue), Switch II (red) and Loop7 (orange). Colors based on PyMOL settings. Figure created in PyMOL (Schrodinger, 2010).
This state is believed to be catalytically slow with disorder in the switch II region. In our RasG12V structures there are varying degrees of helical disorder between residues 60 – 76 of switch II. In Figure 2 we can see that the switch II helix goes from being totally disordered to fully ordered, based on the clarity of electron density (checked by COOT) in that region.

Figure 2. Switch II levels of disorder. All RasG12V structures soaked in organic solvents, with 2RGE (yellow) and 3K8Y (green) as defined in Figure 1. ACT (pale green), DMF (tv_yellow), DMS (pale yellow), GOL (yellow orange), HEX (limon), RSR (wheat), SRS (sand), TBU (smudge), TFE (magenta), TMO (chartreuse) and X-link (split pea). Figure created in PyMOL (Schrodinger, 2010).

As seen in the figure above the RasG12V structures obtained from soaks in DMF and GOL had a fully ordered switch II region, while those soaked in TBU, TMAO, and the X-link
structure had a semi-ordered switch II. The remaining structures, including acetone, DMSO, HXD, RSR, and SRS, had a disordered switch II. The TFE structure had a fully disordered switch II to the point that it could not be built in to the crystallographic model. Switch I however, was fully ordered in all structures. This is typical of all Ras structures obtained from crystals with symmetry R32, since switch I is stabilized by crystal contacts in this crystal form. This is the crystal form we used for the published MSCS of wild type Ras (Buhrman, O’Connor, et al., 2011) and it is the one used here for the MSCS of the oncogenic mutants. In order to see if networks are different for wild type versus mutant, a comparison of hotspot locations was done in Figure 3.
As seen in Figure 3, wild type Ras hotspot locations are shown as blue spheres while RasG12V hotspot locations are shown as red spheres. We define a hotspot as an area on the protein that contains 2 or more organic solvents. We have also distinguished between the effector lobe in green and the allosteric lobe in grey. Overall, we had a total of 15 unique clusters with 8 of those being considered conserved. Table 2 gives a list of all 15 clusters with their protein interactions.
Table 2. List of all protein contacts within 3.5 Å between RasG12V and associated organic solvent clusters. In parentheses are the number of organic molecules found in each cluster within the MSCS set.

<table>
<thead>
<tr>
<th>Cluster Number</th>
<th>Protein Contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (6)</td>
<td>Asn85, Lys117, GNP (nucleotide)</td>
</tr>
<tr>
<td>2 (5)</td>
<td>Cys118, Glu143, Gln150</td>
</tr>
<tr>
<td>3 (3)</td>
<td>Arg97, Asp108, Val109, Tyr137, Ile139</td>
</tr>
<tr>
<td>4 (2)</td>
<td>Gln131, Asp132, Arg135</td>
</tr>
<tr>
<td>5 (2)</td>
<td>Arg41, Gln43</td>
</tr>
<tr>
<td>6 (2)</td>
<td>Ser17, Asp33, Pro34, Thr35, Asp38, Tyr40, Asp57</td>
</tr>
<tr>
<td>7 (2)</td>
<td>Glu37, Asp38, Ser39</td>
</tr>
<tr>
<td>8 (1)</td>
<td>Glu3, Glu76</td>
</tr>
<tr>
<td>9 (1)</td>
<td>Gly138</td>
</tr>
<tr>
<td>10 (1)</td>
<td>Cys118, Leu120, Ala121, Arg123</td>
</tr>
<tr>
<td>11 (1)</td>
<td>Leu23, Asn26, Lys42, Arg149</td>
</tr>
<tr>
<td>12 (1)</td>
<td>Gln22, Asn26, Lys147, Arg149</td>
</tr>
<tr>
<td>13 (1)</td>
<td>Gly48, Thr50</td>
</tr>
<tr>
<td>14 (1)</td>
<td>Asp30, GNP (nucleotide)</td>
</tr>
<tr>
<td>15 (1)</td>
<td>Val145, Tyr157</td>
</tr>
</tbody>
</table>

There are many more different hotspot locations in wild type compared to the G12V mutant. However, there are two locations that have significant overlap. Cluster 1 from RasG12V is in a similar location to Cluster 4 and Cluster 8 in wild type while Cluster 6 from RasG12V is in
a similar location to Cluster 3 in Wild Type. Cluster 1 from RasG12V was the strongest cluster containing six organic solvents (ACT, DMF, DMS, GOL, ETF, and TBU). Cluster 4 and Cluster 8 from wild type only contained two organic solvents RSF and RSG (Buhrman, O’Connor, et al., 2011). These clusters are located in the interlobal region near the p-loop and switch I. These clusters pick out an area of protein-protein interaction where Ras is believed to interact with Ras-GAP (Buhrman, O’Connor, et al., 2011). Cluster 6 from RasG12V contains two organic solvents DMF and HEX. This is in close proximity with Cluster 3 from wild type which contains GOL, DMF, and RSF (Buhrman, O’Connor, et al., 2011). These clusters are located in the effector lobe, opposite of switch I. The clusters are in a location that is experimentally determined to interact with Raf-RBD (Nassar et al., 1996). This further establishes the ability of MSCS to accurately pick out areas of protein-protein contacts. FTMAP, a purely computational solvent mapping method is also useful for studying hotspots in proteins that may have crystallographic restraints due to crystal contacts (Brenke et al., 2009). Figure 4 shows two hotspots that did not match experimental clusters but are seen by FTMAP in the wild type protein results.
Cluster 2 (DMS, GOL, HEX, RSF and RSG) and Cluster 4 (DMF and GOL) are in the allosteric lobe of Ras. Both clusters 2 and 4 are 2.6 Å from residues GLU 143 and GLN 131 respectively. These clusters are between helix 4 and 5 that are unavailable in the wild type protein due to crystal contacts. This area is believed to interact directly with the cell membrane within Ras (Buhrman, O’Connor, et al., 2011). Focusing on Cluster 4, we can see DMF and GOL make direct interactions with R135, which interact directly with the membrane (Gorfe, Hanzal-Bayer, Abankwa, Hancock, & McCammon, 2007).
The carbonyl group of DMF and the hydroxyl group of GOL are hydrogen bonding directly to the δ-N group of R135 within 2.3 Å and the ω-N group of R135 within 2.9 Å, confirmed by electron density. Previous published research using molecular dynamic simulations supported by biological experiments show R128 and R135 sites to interact with the membrane (Abankwa et al., 2008; Gorfe et al., 2007). This is critical in that these results support that MSCS is able to successfully pick out areas of protein-membrane interactions as well. This area could be a possible drug target since it is away from the active site but still able to affect Ras. Whether these sites directly affect hydrolysis is still under investigation. Figure 6 shows two more organic solvent clusters.
Cluster 5 containing DMF, DMS and HEX, is located within 2.8 Å of the backbone amino group of Q43 in the effector lobe. The side chain of Q43 makes hydrogen bonds with the backbone carbonyl group of L919 and sidechain hydroxyl group of Y974 contained within the Cdc25 domain of SOS (Boriack-Sjodin, Margarit, Bar-Sagi, & Kuriyan, 1998). Cluster 7 containing GOL and TBU is located within 3.0 Å and 2.7 Å (respectively) of serine 39 (switch I residue) in the effector lobe of Ras. This is within the Y71 site which is a pocket made up of L56 from β-strand 3 and Y71 from switch II (Buhrman, O'Connor, et al., 2011). Even though there are no crystal contacts within this site the wild type structures had discontinuities in the electron density of switch II (Buhrman, O'Connor, et al., 2011). These
discontinuities made it difficult for organic solvents to bind in the pocket. However, in the RasG12V structures the Ras crystals soaked in glycerol (GOL) resulted in a structure with a fully ordered switch II and those soaked in t-butanol (TBU) had a semi-ordered switch II, allowing for the binding of organic solvent molecules. Figure 7 shows a detailed look at cluster 3 interactions.

Cluster 3 is composed of a t-butanol and two separate glycerol molecules. This cluster makes direct interactions with R97 (2.8 Å), D107 (2.7 Å), D108 (2.8 Å), and Q162 (2.5 Å). This is in the allosteric lobe near helix 3 and loop 7, although it does not overlap entirely with the
allosteric site containing calcium acetate in the structure with PDB code 3K8Y. In the previous experimental wild type structures His 166 blocked the availability of this pocket while FTMAP used structures with the His 166 swung away from the pocket (Buhrman, O'Connor, et al., 2011). This allowed the computational method to pick out the pocket while the experimental MSCS could not. However, in the RasG12V GOL structure the His 166 is swung away allowing organic solvents into the pocket. Also, looking at the RasG12V TBU structure there is a dual position of His166. The His166 residue can be in either position, but it appears that the glycerol molecule has a high enough affinity for the site to stabilize the His166 conformation swung away from the pocket. These differences give some insight into how pockets on proteins can be available based on whether it is wild type or mutated. While some hotspots overlapped with experimental methods performed on wild type many aligned with computational methods. This further validates the need to not only study the wild type for possible druggable targets but to also study the mutants using multiple experimental approaches. Another area of importance is the water networks within Ras. Water networks contribute greatly to the ability of Ras to crosstalk between lobes (Kearney, 2012).

Given the importance of water networks in Ras, we also checked if the water networks are similar in wild type versus the RasG12V mutant. Figure 8 shows the crystallographic water molecules network in the mutant.
RasG12V contains 204 unique water clusters throughout the protein with 69 of those waters being highly conserved (7/11 or 72% of the structures). Water conservation is related to the number of water molecules found in a particular position (cluster) when all of the structures are superimposed. Looking at the effector lobe in green we can see that most of the highly conserved waters are located near switch I and the active site of Ras containing the nucleotide. Switch II does not have any highly conserved waters. This is expected since only four of the structures had a semi to highly ordered switch II region, which allowed us to see and place waters in the area. Looking at the allosteric lobe in black we can see that most of the highly conserved waters are near the loop regions. There are a few highly conserved waters near the helices but due to crystal contacts in the area, they are limited. Previous
research suggests there is a hydrogen bonding network that connects the allosteric site to the active site in Ras (Buhrman et al., 2010).

Hydrogen bonding network 1 is in the allosteric site where the calcium and acetate ions bind. Figure 9 gives a comparison of the loaded allosteric site “on” in the wild type Ras versus the empty allosteric site “off” located in RasG12V.

![Figure 9. Allosteric network 1 in wild type Ras (PDB code 3K8Y) (green) vs. RasG12V (cyan). Loop 7 is in orange. Figure created in PyMOL (Schrodinger, 2010).](image)

Seen in green is the previously published structure of wild type Ras (PDB code 3K8Y) with the loaded allosteric site, while in cyan is a representation of our RasG12V structures with an empty site. As seen in RasG12V many residues are in different orientations compared to the
wild type. Loop 7 is shifted causing D107, V103, K101, Q99, and E98 to shift as well. In the “off” state structures D107 makes a direct hydrogen bond to K101, thus facilitating an interaction between K101 and R97. This shift in residues causes E98 to flip out leaving the side chain to satisfy hydrogen bonds using water 154 (9% conserved) and water 181 (9% conserved). Interactions between residue E98 and protein atoms in the 3K8Y model are replaced in the “off” state by water interactions. R97 in the RasG12V structures hydrogen bonds to water 99 (27% conserved) which hydrogen bonds to water 181 (9% conserved), which in turn makes a hydrogen bonding interaction with E98. This process is an example of the importance of waters in their ability to replace residue interactions when conformational shifts occur. H94 is still able to maintain its hydrogen bonding to Y137 but an extra water 185 (9% conserved) is needed to stabilize the ε² NH group. Even though the RasG12V structures have an empty site there are still waters present. Waters 137 (9% conserved) and 131 (9% conserved) connect the oxygen backbone atoms of Y137 to D107. In wild type this same hydrogen-bonding network is present. However it occurs through the calcium ion instead of waters. This is yet another example of how waters are used to maintain hydrogen bonding for structural integrity.

Hydrogen bonding network 2 moves closer toward the active site including switch II residues, water molecules, and is centered around residue R68 (Buhrman et al., 2010). Figure 10 shows a comparison of hydrogen bonding network 2 in the “on” (wild type Ras PDB code 3K8Y) versus “off” (RasG12V soaks) conformations.
In the “on” conformation Q99 hydrogen bonds to water 384, this in turn hydrogen bonds to R68. In the “off” state the bonds are rearranged and instead of going through a water-mediated network Q99 binds directly to R68. The R68 hydrogen bonding to Y96 is conserved through water interactions. However, water 367 (“on” state) shifts 2.64 Å to water 105 (“off” state, 27% conserved). The Y96 to Q61 direct hydrogen bonding network cannot be elucidated since the residue is disordered in the “off” conformation. However, Y96 hydrogen bonds to water 16 (100% conserved), which in turn hydrogen bonds to water 29 (100% conserved). Water 29 then hydrogen bonds to water 155 (9% conserved), completing the H-bonding network to the amine backbone of E62. This helps to maintain the hydrogen-
bonding network from Y96 to E62. The hydrogen bonding between E62 and Q61 is not present in the “off” state structures. This is the case even when the E62 side chain is present in the electron density maps. The remaining network involving S65, Q99, and R68 is also broken in RasG12V. This is expected since these networks are thought to help place Q61 in the precatalytic confirmation and the V12 residue causes steric hindrance that does not allow this to happen (Buhrman, O’Connor, et al., 2011). Since all of our structures are in the “off” state of the allosteric switch, it is not surprising that the allosteric switch networks are disrupted. Another area of interest is the active site comparison between the “on” and “off” state structures.

Figure 11 shows the active site comparison between the “on” versus “off.”
In the “on” conformation E62 hydrogen bonds to water 176, water 176 hydrogen bonds to Q61, Q61 then hydrogen bonds to water 189 (bridging water molecule), and completing the hydrogen bond network is water 189 hydrogen bonding to Y32 and GppNHp. This allows Ras to be in the correct conformation for hydrolysis. In contrast, in the “off” conformation, as found in the RasG12V mutant, E62 is disordered but could potentially hydrogen bond to water 155 (9% conserved), Q61 hydrogen bonds to water 67 (81% conserved) and water 56 (catalytic water/81% conserved), water 56 hydrogen bonds directly to GppNHp while water 67 hydrogen bonds to Y32, and completing this hydrogen bond network is Y32 directly hydrogen bonding to water 67.
hydrogen bonding to GppNHp. This puts Q61 in the catalytically “off” conformation since it does not interact with a bridging water molecule to promote intrinsic hydrolysis. We proposed the alternative mechanism of intrinsic hydrolysis involving the bridging water and nucleophilic water. In this mechanism a proton from the catalytic water is transferred through the γ-phosphate to the bridging water between the γ-phosphate and Y32 leading to delivery of the proton to the GDP leaving group (Buhrman et al., 2010). Recently, it was reported that there is a Ras water network connecting helix 5 to N85, providing a communication pathway between the membrane interacting region on the surface of Ras and the active site (Kearney, 2012).

Until recently there was no known research to suggest how regions of Ras interacting with the membrane are able to communicate with the active site. However, it was recently hypothesized that this is possible through a hydrogen bonding network including waters and residues (Kearney, 2012). Figure 12 shows the beginning of this network in the “off” state of RasG12V.
The beginning of the network includes: the NH1 group of R164 hydrogen bonding to water 22 (100% conserved/water 181 3K8Y), and water 22 then hydrogen bonds to water 24 (100% conserved/water 179 3K8Y). Next water 24 hydrogen bonds to the oxygen backbone atom of R161, then the NH1 group of R161 hydrogen bonds to water 9 (100% conserved/water 213 3K8Y), water 9 then hydrogen bonds to the hydroxyl group of T158 and lastly the hydroxyl group of T158 hydrogen bonds to the oxygen backbone atom of D154. We then continue the network in Figure 13.
Figure 13. Continuation of network with 3K8Y (green) versus RasG12V X-link (cyan). Figure created in PyMOL (Schrodinger, 2010).

The network continues with the amino backbone group of D154 hydrogen bonding to water 8 (100% conserved) while in the 3K8Y structure water 306 not only interacts with the backbone but the hydroxyl group as well. Continuing along water 8 then hydrogen bonds to water 35 (100% conserved/ water 246 3K8Y), next water 35 hydrogen bonds to the amino backbone group of Q150 which directly hydrogen bonds to C118. Cysteine 118 goes through an important S-glutathiolation which increases Ras activity as the Ras/Raf-1 complex enters into the membrane fraction (Clavreul et al., 2006). The network ends at N85 in RasG12V as seen in Figure 14.
Figure 14. Network ending in the active site with 3K8Y (green) versus G12V X-link (cyan). Figure created in PyMOL (Schrodinger, 2010).

The network continues with K117 hydrogen bonding to water 40 (100% conserved/water 192 3K8Y), which in turn hydrogen bonds to N85. N85 connects to N86, reported to be connected to Y32 through a water hydrogen bonding network in the “on” state of Ras (Buhrman et al., 2010). Y32 is thought to be important for the intrinsic hydrolysis mechanism of Ras as it interacts with the previously mentioned bridging water molecule (Buhrman, Kumar, et al., 2011). In our RasG12V structures the network continues through the amino group of N85 hydrogen bonding to water 199 (1% conserved/water 308 3K8Y). Next, water 199 hydrogen bonds to water 153, which is also 1% conserved and in similar position as water 309 from the 3K8Y structure. This is a similar to a reported trend where the
waters that are close to N86 in the catalytically “off” structures are barely conserved (Kearney, 2012). Looking at the catalytically “on” structure 3K8Y, the hydrogen bond network continues through water 396 and finally hydrogen bonding to Y32. In our RasG12V structures this network is destroyed and there is no connection to Y32. This is typical of catalytically competent Ras structures where the Y32 side chain interacts with a water molecule and in the catalytically incompetent Ras structures where Y32 interacts directly with the nucleotide. All of these observations are expected with our structures being in the “off” state conformation. This could be a possible mechanism for Ras at the membrane to be able to tell if GTP or GDP is bound within the active site. The water network is completed when Ras is in the “on” state, allowing the catalytic residues to be in proper alignment and destroyed when Ras is in the “off” state, changing the alignment of the catalytic residues. More experiments are needed to confirm the importance of these network findings.

**Conclusions**

The current study presents one of the first successful applications of MSCS to a comparison of a mutant to a wild type protein within the same family. While two clusters in RasG12V overlap with two on the surface of wild type Ras, many clusters are in unique areas. The two areas of overlap are in sites of protein-protein interactions, more specifically Ras-GAP and Ras/Raf-RBD. Even though there are many sites in RasG12V that are not identified experimentally in the wild type, they seemed to overlap well with FTMAP results for wild type Ras. These areas of crystal contact result in less access to the organic solvent molecules. In RasG12V there was just enough space to allow smaller organic molecules to bind. These clusters also elucidated areas of protein-protein and protein membrane
interaction known to occur in Ras. In this case, the combined MSCS results for wild type and mutant Ras resulted in a more complete sampling of the hotspots on the surface of Ras.

MSCS also has a very powerful water analysis component. The allosteric switch hydrogen bond networks 1 and 2 presented in the “on” state structures are different in our RasG12V structures in the “off” state. One interesting feature is the ability to pick out subtle changes in water networks. In allosteric network 1, even though calcium acetate was missing and residues are configured differently, waters are present to replace ions and protein residues that maintain similar hydrogen bonding network. In allosteric network 2 we are able to see how waters use an alternate water network route to stabilize similar hydrogen bonding interactions in the “off” conformations. The active site water network shows waters needed for catalytic mechanisms are usually highly conserved. The 100% conserved catalytic water is present no matter the environment of the Ras crystals. Also, results show the active site water networks break down in the “off” state conformation. This was also true for the wild type protein. Lastly, a recently discovered a water network that connects Helix 5 at the membrane interacting area of Ras to the active site and show that this network is mostly conserved in the mutant. This is not surprising given our hypothesis that this network is a sensor for the nucleotide bound state, which is the same in both wild type and mutant proteins. Overall we have discovered many different types of information using MSCS, from conserved water networks to new organic solvent cluster hotspots. In summary, while we see unique organic solvent clusters in RasG12V that are not present in the wild type protein, there are no differences observed for the water networks in the allosteric off states of wild
type and mutant Ras-GppNHp. Further research is needed to determine if this method is useful for depicting differences between wild type and oncogenic mutant proteins.

Acknowledgements

Thanks to Dr. Senthil Kumar for collecting the data on the RasG12V structures in the MSCS set. Thanks to Dr. Greg Buhrman for rechecking structures and giving initial ideas of ways to analyze the set. Thanks to Dr. Bradley Kearney for help in using the DRoP program and answering any related questions. This research was supported by the NSF (MCB - 1237512) grant.

Contribution

All structures are checked and analyzed by me. Also some structures are redone by me to enhance accuracy.
CHAPTER 4: THE MSCS OF RASQ61L

Introduction

Ras is an important member within the superfamily of small monomeric GTPase proteins which function as “molecular switches” in many signaling pathways within the cell (Barbacid, 1987). Ras has two forms including inactive when bound to GDP and active when bound to GTP. Nucleotide cycling causes conformational changes mainly in the switch I (residues 30 – 38) and switch II (residues 59 – 72) regions (Campbell et al., 1998). These regions bind downstream effector proteins as well as regulator proteins that help control the state of the nucleotide. GTPase-Activating Proteins (GAPs) help to enhance the low intrinsic rate of GTPase activity by at least $10^3$-fold causing the signal transduction to be turned off (Scheffzek et al., 1997). Guanine Nucleotide Exchange Factors (GEFs) accelerate the rate of nucleotide release allowing GTP to bind and turning the signal on (Sprang, 1997). Problems in the cycling between GTP and GDP are caused by mutations in Ras that lead to cancer. The three main mutations occur in codons 12, 13, or 61 which convert the Ras gene into an oncogene (Adari, Lowy, Berthe, Der, & McCormick, 1988). These mutations lead to the protein being permanently turned on causing unregulated cell growth and tumors. RasQ61L has one of the highest transformation efficiency of any mutant. While other residue substitutions at position 61 are not as effective they all still decrease in vitro GTPase activity by about 10-fold (Der, Finkel, & Cooper, 1986). Even though the effects of the Q61L mutant are well known, the resulting effects on the intrinsic catalytic mechanism are still being elucidated.
Effector proteins bind Ras in the GTP-bound state to propagate signaling through diverse Ras-associated pathways. Termination of the signal through GAP-catalyzed GTPase activity has been well studied with the assumption that intrinsic hydrolysis is biologically irrelevant. However, our group recently discovered an allosteric switch mechanism associated with the effector Raf, a key interaction in the Ras/Raf/MEK/ERK cascade that controls cell proliferation (Buhrman, Kumar, et al., 2011). This mechanism involves the increase of GTP hydrolysis rate through a GAP independent mechanism that we are still researching. Previous research suggests a two-water model in which the γ-phosphate of GTP abstracts a proton from W189 activating the catalytic W175 for nucleophilic attack on the γ-phosphate during hydrolysis (Scheidig, Burmester, & Goody, 1999). Based on our more recent crystal structures, we proposed an alternate two-water mechanism where a proton from the catalytic water molecule is transferred via the γ-phosphate to a bridging water molecule between the γ-phosphate and Y32 eventually being delivered to the GDP leaving group (Buhrman et al., 2010). During the transition state of the reaction the bridging water molecule, that donates an H-bond to both Y32 and Q61, is proposed to accrue a partial positive charge that stabilizes developing negative charge on the oxygen atom between the □ and □ phosphates of GTP. This mechanism is thought to be more relevant to representing hydrolysis within the Ras/Raf complex since it is based on the structure of Ras from crystals with symmetry of the space group R32 (Buhrman et al., 2007). In this crystal form switch II is free of crystal contacts and shows switch I with water molecules in the active site as seen in the Raps/Raf complex (Nassar et al., 1996). This mimic of the active site of Ras in complex with Raf provides an excellent model for GTP hydrolysis catalyzed intrinsically by
Ras (Buhrman et al., 2010). Raf interacts with the GTP bound form of Ras through the Ras-binding domain (RBD) and the cysteine-rich domain (CRD) (Bondeva, Balla, Várnai, & Balla, 2002). Q61L structures published in the R32 space group show the conserved switch I residue Y32 hydroxyl group interacting directly with the $\gamma$-phosphate of GTP, rather than through a bridging water molecule as in the wild type structure (Buhrman et al., 2007; Wittinghofer & Nassar, 1996). This structure of RasQ61L shows that L61 becomes part of a hydrophobic cluster that closes over the nucleotide, isolating it from bulk solvent and abrogating hydrolysis of GTP in the presence, but not in the absence of Raf-RBD ((Buhrman, Kumar, et al., 2011; Buhrman et al., 2007)). Given how relevant RasQ61L is to cancer the next logical step is to determine possible drugs to disrupt the pathway. One such method for identifying hotspots as possible targets is the Multiple Solvent Crystal Structure’s (MSCS) method.

The MSCS method uses organic solvents for the mapping of protein surfaces (Allen et al., 1996). This method employs either an organic water mixture or neat solvent applied to one crystal at a time. This allows us to locate areas of protein-ligand interactions, protein-protein interactions, protein plasticity, and comprehensive water analysis (Allen et al., 1996; Carla Mattos et al., 2006). This method was recently applied to wild type H-Ras in the R32 space group, combined with the FTMAP computational method for locating hot spots on protein surfaces (Buhrman, O’Connor, et al., 2011). This paper highlighted the efficiency of MSCS by successfully locating hotspots that picked out Ras interactions with Raf-RBD, RasGap, Raf-CRD, and membrane interactions (Buhrman, O’Connor, et al., 2011). The FTMAP revealed additional hot spots in areas that are in extensive crystal contacts in Ras
crystallized in the R32 crystal form. The experimental and computational methods combined are essential for locating hotspots that could be possible drug sites away from the active site, which is intrinsically disordered (Ito et al., 1997; O’Connor & Kovrigin, 2012) in the absence of protein binding partners and thus difficult to target with small molecules. In addition to the possibility that oncogenic mutations may affect the locations of hot spots on the surface of Ras, it could also disrupt the water-mediated H-bonding networks critical to the allosteric switch mechanism associated with the Ras/Raf/MEK/ERK pathway (Buhrman et al., 2010). The MSCS method was therefore applied to RasG12V and RasQ61L. In RasG12V two hotspots overlapped with wild type. Interestingly however, most of the hotspots identified by MSCS for this mutant overlapped with those discovered by FTMAP in the wild type protein. In order to better visualize results we use the DRoP program. This program was created to allow consistent analysis across a large number of structures, address crystallographic symmetry, allow cluster visualization, and give a quality measurement of the clusters (Kearney, 2012). DRoP allows us to be able to accurately assess the differences and similarities between wild type, RasG12V, and RasQ61L. The present chapter focuses on the MSCS results and analysis of surface features for the RasQ61L oncogenic mutant.

**Experimental Methods**

*Expression, Purification and Crystallization*

Truncated wild type Ras (residues 1-166) was mutated at codon 61 (Q61L) using a QuickChange II Site-Directed Mutagenesis Kit from Stratagene following the manufacturer’s instructions. The DNA was then cloned into the PeT21A (+) vector (Novagen) and transformed into *E. coli* BL21 cells. These cells are used for purification as previously
described (Buhrman et al., 2003). The GDP was then exchanged for GppNHp using the published procedure (Stumber et al., 2002). Ras-GppNHp solution contained between 10–15 mg/mL of protein in stabilization buffer (20 mM Hepes pH 7.5, 50 mM NaCl, 20 mM MgCl$_2$ and 10 mM dithioerythritol). Crystals are grown at 18 °C in a sitting drop tray containing 5 µl of protein solution and 5 µl of reservoir solution for about a week. The reservoir solution contained 500 µl of 200 mM calcium acetatehydrate, 20% w/v PEG 3350 and 0.05% n-Octyl-β-D-glucopyranoside diluted with 50 µl of stabilization buffer for crystals grown in the presence of calcium acetate. For those grown in the presence of calcium chloride the reservoir solution consisted of 500 µl of 200 mM CaCl$_2$, 25% w/v PEG 3350 and 1 mM DTT.

**Cross-linking and organic solvent soaks**

H-RasQ61L-GppNHp crystals are transferred to a 9-well sitting-drop glass plate containing 50 µl of stabilization buffer (20 mM Hepes pH 7.5, 50 mM NaCl, 5 mM MgCl$_2$, 1 mM DTT, 200 mM CaCl$_2$ or 200 mM Calcium Acetate and 25% PEG 3350). The buffer was then exchanged with stabilization buffer containing 1% gluteraldehyde. The cross-linking reaction was allowed to occur for 30 minutes to 1 hour before exchanging with fresh stabilization buffer to remove gluteraldehyde. The crystals are then successfully soaked in neat organic solvents or organic solvent/water mixtures as seen in the results section.

**Data collection and structure refinement**

High-resolution data were collected on the Ser-CAT ID-22 beamline at APS at 100K using the Mar CCD detector. The X-rays had a wavelength of 1.0 Å and were exposed from 1 to 3 seconds per frame with an oscillation angle of 1° and a crystal to detector distance of 120
mm. The collected data are processed using HKL2000 (Otwinowski & Minor, 1997). The published structure of H-Ras (PDB code 1CTQ) with the removal of all non-protein atoms and switch II deleted from the model with atomic B-factors set at 30 Å was used as an initial search model for molecular replacement using Crystallography and NMR System (CNS) (Brünger et al., 1998). CNS was used for all reciprocal space refinement with 10% of the unique reflections set aside for R_free calculations (Kleywegt & Brünger, 1996). The molecular replacement solution with the best statistics was applied to generate a model using a rigid body refinement at 2.5 Å, then a rigid body refinement at 2.0 Å, simulated annealing, energy minimization and group B-factor refinement in CNS. Next the 2Fo - Fc and Fo - Fc electron density mappers are generated. Building of the model was accomplished by COOT (Emsley et al., 2010). CNS was then used in multiple rounds of energy minimization and individual B-factor refinements. The GppNHp molecule was added to the model early while water, ions and organic solvents are added later in the refinement process. All structures are then run through DRoP and the resulting analysis is presented below.

**Results**

Ras Q61L crystals are grown in either calcium acetate or calcium chloride and then soaked in multiple solvents. The crystal soaks grown in the presence of calcium acetate are in 50% dimethylformamide (DMF), and 60% 1,6-hexanediol (HEZ). Crystals grown in the presence of calcium chloride included the X-link structure (no organic solvent), 50% DMF, neat hexane (HEX), and 70% glycerol (GOL). Note that there are two structures obtained from crystals soaked in 50% DMF. One originates from crystals grown in the presence of calcium acetate and the other in the presence of calcium chloride. We will differentiate these
two structures as DMF-Ca(OAc)$_2$ and DMF-CaCl$_2$. The X-link structure was cross-linked and maintained in stabilization solution derived from the crystal growth mother liquor and is used as a control to make sure that the crosslinking itself does not significantly affect the protein structure. The statistics for the crystal structures are given below in Table 1.
### Table 1. Refinement Statistics of RasQ61L-GppNHp. Parentheses equals highest resolution shell.

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>X-Link</th>
<th>DMF 50%</th>
<th>DMF 50%</th>
<th>HEX</th>
<th>HEZ 60%</th>
<th>GOL 70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
<td>R32</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>88.77</td>
<td>88.73</td>
<td>88.57</td>
<td>88.23</td>
<td>87.76</td>
<td>89.16</td>
</tr>
<tr>
<td></td>
<td>88.77</td>
<td>88.73</td>
<td>88.57</td>
<td>88.23</td>
<td>87.76</td>
<td>89.16</td>
</tr>
<tr>
<td></td>
<td>134.66</td>
<td>133.86</td>
<td>134.46</td>
<td>134.38</td>
<td>134.62</td>
<td>135.09</td>
</tr>
<tr>
<td>α,β,γ (°)</td>
<td>90, 90, 120</td>
<td>90, 90, 120</td>
<td>90, 90, 120</td>
<td>90, 90, 120</td>
<td>90, 90, 120</td>
<td>90, 90, 120</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>50 - 1.81</td>
<td>50 - 1.80</td>
<td>50 - 2.20</td>
<td>50 - 1.45</td>
<td>50 - 1.55</td>
<td>50 - 2.18</td>
</tr>
<tr>
<td></td>
<td>(1.89 - 1.81)</td>
<td>(1.86 - 2.00)</td>
<td>(2.07 - 1.51)</td>
<td>(1.51 - 1.45)</td>
<td>(1.62 - 1.55)</td>
<td>(2.26 - 2.18)</td>
</tr>
<tr>
<td>R&lt;sub&gt;sym&lt;/sub&gt; or R&lt;sub&gt;merge&lt;/sub&gt;</td>
<td>0.088 (0.456)</td>
<td>0.053 (0.570)</td>
<td>0.108 (0.786)</td>
<td>0.075 (0.584)</td>
<td>0.067 (0.471)</td>
<td>0.093 (0.497)</td>
</tr>
<tr>
<td>I/σ</td>
<td>37.4 (3.52)</td>
<td>52.6 (3.21)</td>
<td>36.7 (3.95)</td>
<td>41.7 (3.88)</td>
<td>33.2 (3.64)</td>
<td>19.8 (3.35)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>97.3 (94.0)</td>
<td>99.9 (99.1)</td>
<td>99.7 (98.1)</td>
<td>99.9 (100)</td>
<td>99.5 (99.1)</td>
<td>98.9 (97.5)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>9.4 (6.0)</td>
<td>10.4 (7.9)</td>
<td>9.3 (6.5)</td>
<td>6.0 (8.9)</td>
<td>9.7 (8.0)</td>
<td>9.7 (6.7)</td>
</tr>
<tr>
<td>Refinement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>1.81</td>
<td>1.80</td>
<td>2.20</td>
<td>1.45</td>
<td>1.55</td>
<td>2.18</td>
</tr>
<tr>
<td>No. reflections</td>
<td>19484</td>
<td>18439</td>
<td>10043</td>
<td>38311</td>
<td>30623</td>
<td>11649</td>
</tr>
<tr>
<td>R&lt;sub&gt;work&lt;/sub&gt; / R&lt;sub&gt;free&lt;/sub&gt;</td>
<td>19.08/ (21.01)</td>
<td>20.88/ (23.64)</td>
<td>19.11/ (22.74)</td>
<td>19.33/ (20.00)</td>
<td>18.39/ (19.75)</td>
<td>19.39/ (21.94)</td>
</tr>
<tr>
<td>No. atoms</td>
<td>1313</td>
<td>1250</td>
<td>1310</td>
<td>1304</td>
<td>1289</td>
<td>1308</td>
</tr>
<tr>
<td>Protein</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Organcics</td>
<td>2/1</td>
<td>2/2</td>
<td>2/2</td>
<td>1/4</td>
<td>1/4</td>
<td>2/2</td>
</tr>
<tr>
<td>Magnesium/ Calcium Ions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GppNHp</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Water</td>
<td>127</td>
<td>89</td>
<td>103</td>
<td>135</td>
<td>127</td>
<td>89</td>
</tr>
<tr>
<td>RMSD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>.005</td>
<td>.006</td>
<td>.007</td>
<td>.004</td>
<td>.004</td>
<td>.013</td>
</tr>
<tr>
<td>Bond angles (Å)</td>
<td>1.19</td>
<td>1.17</td>
<td>1.22</td>
<td>1.16</td>
<td>1.10</td>
<td>1.82</td>
</tr>
</tbody>
</table>
The space group was R32 with the crystals diffracting to a resolution of 2.2 Å or better. We show that in this crystal form Ras-GppNHp structures can be obtained in the two states associated with the allosteric switch mechanism: one is obtained in the presence of calcium chloride and has a structure where switch II is completely disordered (PDB code 2RGE) and intrinsic hydrolysis is slow (Buhrman et al., 2007); the other is obtained in the presence of calcium acetate and shows a shift in helix 3/loop 7 away from the active site and towards helix 4, resulting in a highly ordered active site in which intrinsic hydrolysis is expected to be greatly accelerated (PDB code 3K8Y) (Buhrman et al., 2010). The binding of calcium and acetate at the remote allosteric site is responsible for this shift. We propose that the binding of calcium and a negatively charged membrane component (mimicked by the acetate) to Raf-bound Ras is at the core of the allosteric switch mechanism responsible for turning off signaling through the Ras/Raf/MEK/ERK signaling pathway.

Figure 1 shows the overall comparison of the RasQ61L crosslinked structure to those with PDB codes 3K8Y (with the shift in helix3/loop7 and an ordered active site) and 2RGE (with a disordered active site).
Figure 1. RasQ61L models (in various colors) versus 2RGE (yellow) and 3K8Y (green). The GTP analogue GppNHp is bound to wild type and mutant Ras in the three models. The bound calcium and acetate ions are shown in green sphere and in stick respectively. P-loop (black), Switch I (blue), Switch II (red) and Loop7 (orange). Figure created in PyMOL (Schrodinger, 2010).

All of the RasQ61L structures resembled the 2RGE model, containing no shift in helix 3/loop 7 and with a partially ordered or disordered switch II region. This is because even when crystals are grown in the presence of calcium acetate, the calcium and acetate ions are soaked out once the crystals are transferred to solutions containing high concentrations of organic solvents. For this same reason, all of the structures in the MSCS of wild type Ras (Buhrman, O'Connor, et al., 2011) and RasG12V are also in the conformation observed in the model with PDB code 2RGE, corresponding to the “off” state of the allosteric switch where
catalysis is expected to be slow. In crystals of RasQ61L soaked in 50% DMF CaAc the switch II region was totally disordered and could not be built due to lack of electron density. The structure solved in 60% HEZ had a partially ordered switch II with most of the residues containing sufficient electron density to build at least the main chain for that region. The structures obtained in 50% DMF CaCl₂, 70% GOL, neat HEX and X-link had an ordered switch II region as confirmed by electron density. This conformation is the same as observed in our recently published structure of wild type H-Ras bound to DTT near switch II (Holzapfel, Buhrman, & Mattos, 2012) and consists of the anti-catalytic conformation observed in the RasQ61L mutant (Buhrman et al., 2007). Given these structures are similar to RasG12V and wild type we compared the water and organic solvent structure networks in the three structures.

There are 14 unique organic molecule clusters within RasQ61L and 3 of those clusters are highly conserved. Table 2 gives a list of all the cluster numbers and the residue interactions within 3.5 Å.
Table 2. Cluster numbers and their interacting residues listed. In parentheses are the numbers of organic molecules found in each cluster within the MSCS set.

<table>
<thead>
<tr>
<th>Cluster Number</th>
<th>Protein Contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (3)</td>
<td>GLN25, HIS27</td>
</tr>
<tr>
<td>2 (2)</td>
<td>GLN95</td>
</tr>
<tr>
<td>3 (2)</td>
<td>GLU126, GLN129</td>
</tr>
<tr>
<td>4 (1)</td>
<td>GLU3, TYR4, GLU76</td>
</tr>
<tr>
<td>5 (1)</td>
<td>GLU162, GLN165</td>
</tr>
<tr>
<td>6 (1)</td>
<td>CYS118, LEU120, ARG123</td>
</tr>
<tr>
<td>7 (1)</td>
<td>TYR32, ASP33</td>
</tr>
<tr>
<td>8 (1)</td>
<td>GLY12, ASN86</td>
</tr>
<tr>
<td>9 (1)</td>
<td>ILE21, SER17, ASP33, THR35, ASP38</td>
</tr>
<tr>
<td>10 (1)</td>
<td>PRO34, ILE36, GLU37, ASP38</td>
</tr>
<tr>
<td>11 (1)</td>
<td>ASP92, HIS94, GLN95</td>
</tr>
<tr>
<td>12 (1)</td>
<td>Nucleotide, ASN 85, LYS117, LEU120</td>
</tr>
<tr>
<td>13 (1)</td>
<td>ASN26, LYS42, VAL44, ARG149, GLU153, TYR157</td>
</tr>
<tr>
<td>14 (1)</td>
<td>GLU91, HIS94, TYR137</td>
</tr>
</tbody>
</table>

Figure 2 shows a comparison of hotspots obtained from MSCS analysis for wild type and the two oncogenic Ras mutants G12V and Q61L.
Figure 2. Organic Cluster overlap in effector lobe. P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 2 shows many clusters that are in different areas of the protein and are represented by spheres in place of the organic solvent clusters that appear after superposition of all structures in the respective MSCS sets. Clusters for RasQ61L are shown in red spheres, those for RasG12V are shown in blue spheres and those for wild type are shown in yellow spheres. For the purposes of this analysis if there are two organic solvent molecules bound at the same site in an MSCS set, that site is considered to contain a cluster indicative of a hotspot of protein/protein or protein/ligand interaction.

Ras can be thought of as containing two lobes. The first half of the molecule is the effector lobe, shown in green in Figure 2. The effector lobe contains residues 1 – 86.
including switch I, switch II, the P-loop and most of the nucleotide binding pocket (Buhrman, O'Connor, et al., 2011). It is called the effector lobe because it contains the protein/protein interaction sites with effectors (Buhrman, O'Connor, et al., 2011). The second half of the catalytic domain is the allosteric lobe shown in black with gray surface in Figure 2. The allosteric lobe contains residues 87 – 171 including the portions of Ras that interact with the membrane, the allosteric site and allosteric switch components involving helix 3, loop 7 and helix 4 (Buhrman, O'Connor, et al., 2011). One area of cluster overlap occurs in the effector lobe. Figure 2 shows RasQ61L Cluster 1 (DMF-Ca(OAc)$_2$, DMF-CaCl$_2$, HEX) overlaps with wild type Ras Cluster 3 (DMF, GOL, and RSF) and is in close proximity to Cluster 6 (DMF and GOL) in RasG12V. These clusters are on the other side of switch I relative to the nucleotide. This interaction site depicts a protein/protein interaction associated mainly with Ras/Raf-RBD in wild type (Buhrman, O'Connor, et al., 2011). Another site of overlap is in an area between the effector lobe and allosteric lobe as seen in wild type shown in Figure 3.
Figure 3. Cluster overlap between the lobes. P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).

Figure 3 shows an area of overlap between lobes specifically Cluster 2 (HEX and HEZ) from RasQ61L and Cluster 1 (DMF, ETF, HEX and HEZ) from wild type. These clusters are between helix 3 and switch II near the active site, considered the interlobal region. This site is another area of protein/protein interaction. This is where importin-β binds the homologous Ran GTPase when it binds GTP causing a stabilization in an identical catalytically incompetent conformation of switch II (Vetter, Arndt, Kutay, Görlich, & Wittinghofer, 1999). Another cluster is in the allosteric lobe not in overlap with either RasG12V or wild type as seen in Figure 4.
As seen in the Figure 4 above Cluster 3 in RasQ61L is in the allosteric lobe. This cluster contains DMF-CaAc and DMSO (collection data unavailable) organic solvent molecules. These organic molecules bound in a site near Q129 interacting with its side chain amide group. As of now there are no known important interactions that occur at residue 129 in H-Ras. Although the organic solvent clusters are of high importance water molecules also play an important role in mediating protein/protein interactions. A comprehensive water comparison of RasQ61L to the wild type with an ordered active site (PDB code 3K8Y) is used to determine whether the water mediated H-bonding networks are conserved in the mutant.
For this analysis we used our in-house program DRoP (Kearney, 2012). DRoP takes coordinates for multiple structures of a given protein and renumbers all of the solvent molecules according to how conserved they are in the set and how tightly the molecules are in the cluster. Thus water molecule number 1 is conserved in all 7 structures of RasQ61L in the various solvent conditions and water molecules at that position has the smallest RMSD of all 100% conserved clusters, and so forth. RasQ61L structures contain 228 unique crystallographic water-binding sites. However, only 85 of those clusters are highly conserved, meaning that they appeared in at least 5 out of the 7 structures at that given position as seen in Figure 5.

Figure 5. Highly conserved water molecules (red spheres) in the effector lobe (green) and in the allosteric lobe (black). P-loop (black in effector lobe) Switch I (blue in effector lobe), Switch II (red in effector lobe) and Loop7 (orange in allosteric lobe). Figure created in PyMOL (Schrodinger, 2010).
As seen in the figure the effector lobe in green has a higher hydration level, focused mainly near the nucleotide and switch I. However, looking at the switch II region there are virtually no conserved crystallographic water positions. This is due to the disordering of switch II, which will inevitably result in a correspondingly disordered hydration shell. The allosteric lobe, shown in black, has much higher hydration in the loop regions than near helices 3, 4 or 5. This is because this area is near crystal contacts that limit accessibility to this region. Looking within the active site we can see that the water networks are different between the catalytically “on” (ordered active site with water molecules, Y32 and Q61 in place for catalysis) and “off” (anticatalytic conformation) states of the allosteric switch in Ras.
In Figure 6 a representative RasQ61L structure (“off”) is in cyan, with cyan spheres representing its associated water molecules; and the published structure representing the “on” state (PDB code 3K8Y) is shown in green. Of water molecules from the RasQ61L MSCS set, only the nucleophilic water molecule, number 14, is 100% conserved. This water molecule is in a similar position to Water 175 from the published (PDB code 3K8Y) in the precatalytic state. The catalytic water is essential to the intrinsic hydrolysis of Ras and is activated by the \( \gamma \)-phosphate of GTP abstracting one of its protons and shuttling it to the bridging water molecule 189 (Buhrman et al., 2010). As seen in the Figure 6, Water 189 is not present in our RasQ61L structures. This is due to the fact that the structures are in the “off” conformation.
with Y32 interacting directly with the GppNHp, replacing the bridging water molecule. Water 396 in the wild type “on” conformation is shifted in the “off” conformations in the RasQ61L structures represented by the cyan Water 118 (43% conserved). Water 396 links Y32 to N86 at the N-terminal end of helix 3 through a hydrogen bond network involving waters 162 and 101 (Buhrman et al., 2010). In our structures this hydrogen bond network is broken as Water 118 is too far to link the H-bonding network originating at N86 to the active site Y32, now shifted toward the Ð-phosphate of the nucleotide. Water 176 in wild type interacts directly with the side chain oxygen atom of Q61 linking it through hydrogen bonds to E62. This water works in conjunction with E62 to position Q61 near the bridging water molecule 189 (Buhrman et al., 2010). Water 108 (43% conserved) from the RasQ61L structures is in a similar position but instead makes interactions with the backbone of L61 with E62 being too disordered to participate in the interactions. In previous research hydrogen bonding networks labeled network 1 and 2 connect the allosteric site to Q61 in the active site (Buhrman et al., 2010).

Allosteric network 1 includes calcium acetate which involves H94, R97, E98, K101 (located on helix 3), D107 (located on loop 7), and Y137 (located on helix 4) (Buhrman et al., 2010). In our RasQ61L structures this network is completely destroyed as it is in the RasG12V structures. All of the residues are in different orientations compared to those in the model with PDB code 3K8Y, with waters replacing some of the hydrogen bonding interactions in the network. Figure 7 shows a comparison of allosteric network 2 in the structure of the “on” state conformation of the allosteric switch (PDB code 3K8Y) versus our organic solvent soaked RasQ61L structures in the “off” state conformation.
Helix 3 residues Y96, Q99, and R102 are in position to contribute to network 2 along with switch II residues M72, Y71, D69, R68, S65, E62, and multiple water molecules (Buhrman et al., 2010). Conditions that form networks 1 and 2 are caused through a shift in Ras creating the networks that allow the ordering of switch II and placing Q61 in a precatalytic state for hydrolysis of GTP (Buhrman et al., 2010). Figure 7 shows water 176 (3K8Y) and 108 (RasQ61L soaks and 43% conserved) shift 1.4 Å in going from the “on” to “off” state. Comparing Water 367 (3K8Y) and Water 76 (RasQ61L soaks and 71% conserved) the shift is 1.6 Å. Lastly, Water 373 (3K8Y) and Water 132 (RasQ61L soaks and 29% conserved) show a 0.9 Å shift. Although these water molecules are in close proximity to important
waters in the precatalytic state the roles are very different. This is due to the fact that the
 glutamine at position 61 is changed to a leucine, breaking the hydrogen bond network.
 Overall there is a shift is several key residues important for the network 2 and the water
 molecules that interact with these residues shift along with them (Water 76 shifts with Y96,

 Recently, another important water network was discovered in Ras linking the
 membrane interaction helixes to the active site (Kearney, 2012), elucidating for the first time
 another important communication network between the membrane interaction region in the
 allosteric lobe (helices 4 and 5) and the active site. This network involves several water
 molecules that are highly conserved in the wild type MSCS set. It turns out that this network
 is also highly conserved in the RasQ61L MSCS set presented here. Figure 8 gives a depiction
 of those residues and water molecules involved.
The network begins in Figure 8A with the NH1 group of R164 hydrogen bonding to two 100% conserved waters 20 and 24 that in turn link to the backbone oxygen atom of R161. R161 then hydrogen bonds to 100% conserved water 78, which in turn H-bonds to the hydroxyl oxygen atom of the side chain of residue T158. Figure 8B continues the network.
Figure 8. B) Continuation of helix 5 to active site water network. Figure created in PyMOL (Schrodinger, 2010).

The hydroxyl group of the T135 side chain then hydrogen bonds to the backbone carbonyl of D154 which in turn hydrogen bonds to 100% conserved water 25. Next Water 25 hydrogen bonds to the amine backbone group of Q150, which in turn hydrogen bonds to C118. Cysteine 118 can be S-glutathionylated increasing Ras activity (Clavreul et al., 2006). The experiment shows the time course of this modification coincided with Ras binding to Raf-1 and the complex entering into the membrane fraction (Clavreul et al., 2006). Looking further along the active site the network continues through K117 as seen in Figure 8C.
Lastly, the amine group of K117 hydrogen bonds to 100% conserved water 11, which in turns hydrogen bonds to N85. K117 binds directly to the nucleotide while N85 connects to N86. N86 connects to Y32 through a water network in the “on” state of the allosteric switch in Ras (Buhrman et al., 2010). In our Q61L organic solvent soaked “off” state structures N86 does not hydrogen bond any water molecules that connect to Y32 as seen previously stated above.
Conclusions

With the growing difficulty to make viable drugs that target mutated proteins, experimental methods that can enhance and speed up this process are vital. If a method is able to depict differences between wild type and mutant protein, then the chances of creating a viable drug targeting only the rogue protein would be greatly increased. This would have huge impacts on the field of drug targeting as a whole. In the work presented here we use MSCS to start to answer these drug targeting questions by finding a novel result of RasQ61L.

Two highly conserved clusters are in similar locations of published wild type clusters. These clusters are in important locations of protein/protein interactions in Ras. One cluster novel to the RasQ61L structures was near Q129, but its significance for Ras function has not yet been determined. There are several positions associated with a single organic solvent molecule on the surface of RasQ61L. However, these are not considered to be hot spots and therefore are not analyzed in this chapter. The MSCS water analysis revealed 286 unique crystallographic water positions in the set, with 85 of those being 100% conserved. Many of the low conservation water positions are within the active site and allosteric site, which is expected since the protein is in the “off” conformation of the allosteric switch. One area of interest was the connection of helix 5 to N85. This connection was achieved through several 100% conserved water molecules and residues that connect the membrane-binding allosteric lobe to the active site and is conserved in wild type and mutant Ras in the GTP-bound state. This water network is present as it is in RasG12V and wild type, for the link between helix 5 and N85 at the C-terminal end of helix 3. The connection to the γ-phosphate in the active site is
dependent on the position of Y32 interacting with the bridging water molecule to complete
the network. It is thus present only in the wild type protein and absent in both Ras-G12V and
Ras-Q61L. These results led to the idea of further hydrolysis experiments creating a double
mutant at N85/N86 to determine if disrupting this water network affects RAS hydrolysis.
Also by crystallizing this mutant we can hopefully show the overall structure to be
unperturbed proving this is a direct result of an impaired water network. By comparing the
MSCS sets published for wild type with those obtained for the two oncogenic mutants
RasG12V and RasQ61L we not only see agreement with published binding pockets but also
novel sites that are unique to the oncogenic mutants. However, we are not allowed we do not
know at this time if these sites are truly novel. In order to answer this question more
experimentation, such as making a list of all organics across all sets and then resoaking
missing structures, is needed. After reanalysis of these new structures we hope this will not
only allow us to present this method as a way to tease out differences between wild type and
mutant proteins but also better design inhibitors by tailoring them to specific pockets away
from the nucleotide binding site.
Acknowledgements

Thanks to Dr. Greg Buhrman for collection of data and solving of structures. Thanks to Dr. Bradley Kearney for help with the DRoP program and answering any related questions. This research was supported by the NSF (MCB - 1237512) grant.

Contributions

Dr. Greg Buhrman soaked crystals and collected all relevant data. I conducted all analysis.
CHAPTER 5: THE MSCS OF THE BARRIER-TO-AUTOINTEGRATION FACTOR (BAF) PROTEIN

Introduction

Understanding protein-protein and protein-ligand interactions within different types of proteins is essential for furthering disease related research. Even though proteins have been highly studied over time many still have unknown functions such as barrier-to-autointegration factor (BAF). BAF is a small DNA binding protein and in the context of our work on MSCS is it an ideal representative to study surface features in protein/DNA interactions. BAF is located within the cytoplasm and the nucleus of cells. Although its endogenous function is not yet known, it is used to block autointegration of retroviral DNA (Umland et al., 2000). In the life cycle of retroviruses the integration of a DNA copy of the viral genome into the host cell genome is an important process. The integration process is mediated by the preintegration complex (PIC). PIC is found in cells infected by human immunodeficiency virus type 1 (HIV-1) and Moloney murine leukemia virus (Mo-MLV) (Ellison et al., 1990; Farnet & Haseltine, 1990; Fujiwara & Mizuuchi, 1988). When the virus has completed reverse transcription there can be a delay of as much as hours before the viral DNA are integrated (M S Lee & Craigie, 1994). During this time the virus can go through a lethal process resulting in destruction of the viral genome, autointegration (M S Lee & Craigie, 1994). Autointegration is the process in which the virus integrates its own viral DNA into itself, causing damage to the virus replication cycle (Umland et al., 2000). The virus tends to favor intermolecular integration into the cellular DNA avoiding intramolecular
integration of the viral DNA (Cai et al., 1998). If the virus were to integrate its own DNA it would destroy the viral DNA before it can integrate into the host genome, causing abortion of the viral replication cycle (Cai et al., 1998). BAF protein protects viruses against autointegration (Chen & Engelman, 1998; M S Lee & Craigie, 1994; Myung Soo Lee & Craigie, 1998); so understanding the structural features of the BAF interacting surfaces is crucial to the research of the virus life cycle. BAF has been shown to be important in gene expression, nuclear assembly, and chromosomal organization in metazoans (Segura-Totten & Wilson, 2004). Knock-down experiments of BAF in Drosophila melanogaster (Furukawa et al., 2003) and Caenorhabditis elegans (Margalit et al., 2005; Zheng et al., 2000) are shown to be lethal. Despite the progress in determining the overall processes in which BAF is involved, its molecular mechanisms are still not entirely known.

BAF is able to bind double stranded but not single-stranded DNA in a non-specific manner with respect to the sequence of bases involved in the interaction (Myung Soo Lee & Craigie, 1998). BAF also possesses the ability to bridge together multiple segments of DNA. This activity is believed to be responsible for BAF’s function of protecting viral DNA from autointegration (Otwinowski & Minor, 1997). BAF is a biological dimer, with one molecule related to the other in a “head to tail” arrangement through a two-fold symmetry axis (Bradley et al., 2005). Its crystal structure in complex with a 7 base-pair DNA is solved and shows duplexed DNA bound to opposite ends of the BAF dimer (Bradley et al., 2005). BAF only forms contacts with the minor groove of the phosphate backbone of DNA (Bradley et al., 2005). BAF and DNA interactions mainly occur within motifs in the dimers termed HhH motif, related pseudo HhH motif, and the N-terminus of α-helix within each monomer
These findings are consistent with mutagenesis data and the computationally predicted model (Umland et al., 2000). Furthermore, comparison with a structure of the dimer in the absence of DNA shows that BAF does not change conformation upon binding to DNA (Bradley et al., 2005). This is interesting considering that many proteins have slight conformational changes upon ligand binding. In addition to interacting with DNA BAF interacts with LEM proteins.

Lamin-associated polypeptide 2 (LAP2) is composed of a family of spliced proteins associated with the inner nuclear membrane (Dechat, Vlcek, & Foisner, 2000; Foisner & Gerace, 1993). A conserved LEM motif is located within the constant N-terminal region (residues 1 – 187) of LAP2 (Harris Ca Fau - Andryuk et al.; Lin et al., 2000). This region is also conserved in many other proteins including emerin and MAN1 thus leading to the term LEM or LAP2-emerin-MAN1 proteins (Lin et al., 2000). Emerin is a multidomain protein containing a N-terminal LEM domain of approximately 50 residues (Lin et al., 2000), a polyserine segment and a C-terminal transmembrane region (Cai et al., 2007). Loss of emerin causes the x-linked recessive form of Emery-Dreifuss muscular dystrophy (Nagano et al., 1996). The discovery led to NMR experiments of LEM protein interactions between the LEM motifs of LAP2 and emerin binding to BAF as a model.

The NMR solution of the BAF-LAP2 complex shows BAF interacting with helix 1, the loop connecting helices 1 and 2, and the N-terminal residues of helix 2 representing one LEM domain near the C-terminal end of LAP2 (Cai et al., 2001). The BAF interaction site is located in the cleft bridging the two monomers and are comprised of a central hydrophobic patch surrounded by hydrophilic residues including Glu35, Phe39, Asp40, Gly47, Gln48,
Leu50, Val51 and Trp62 (Cai et al., 2001). These experiments show that the C-terminal LAP2 LEM domain contained a convex interaction surface that complements BAF’s concave interaction surface in both shape and composition (Cai et al., 2001). Later the NMR solution of BAF bound to emerin showed the binding surface to consist of a central hydrophobic portion surrounded by charged and polar residues (Cai et al., 2007). The key interactions include Val51, Leu52, and Leu58 of one monomer within the dimer of BAF plus Gly38, Phe39, and Val51 of the other BAF monomer with Leu23, Gly24, Phe25, and Val26 of the LEM domain of emerin (Cai et al., 2007). Electrostatic interactions are seen between Arg37, Glu61 and Asp65 of one BAF monomer with Asp9, Lys 36, and Lys38 of emerin’s LEM domain and Glu36 of the other BAF monomer with Arg17 of the LEM domain of emerin (Cai et al., 2007). An example of water interactions can also be seen as water molecules bridge contacts between Gln48 of one monomer of BAF and Ser29 and Thr30 of the LEM domain of emerin (Cai et al., 2007). These interactions are consistent with published mutational studies detailing mutated residues that greatly reduce BAF-emerin binding (K. K. Lee et al., 2001). All of these interactions are also similar to what we see from structures obtained using the Multiple Solvent Crystal Structures (MSCS) method.

The features of protein surfaces that bind peptides, sugars or RNA in a base specific manner as substrates have been well characterized by the MSCS method (Michelle Dechene et al., 2009). MSCS employs either organic solvent/water mixtures or neat organic solvents to probe the binding surfaces of proteins (Carla Mattos et al., 2006). We demonstrated organic solvent molecules are able to displace water primarily in regions of protein surfaces that evolved to interact with ligands (Carla Mattos et al., 2006). Superposition of the models
of a given protein obtained in a variety of solvent environments shows that the organic solvent molecules cluster in hot spots of protein/ligand interactions. Furthermore, due to the fact that the protein adjusts its structures to the different environments, we can also observe the range of plasticity and hydration characteristics associated with the protein of interest (Carla Mattos et al., 2006). Previous studies using MSCS correctly located known substrate binding sites for Elastase (Carla Mattos et al., 2006), RNAse A (M. Dechene et al., 2009) and Ras (Buhrman, O'Connor, et al., 2011), showing that it can identify location of hot spots in proteins where binding sites have been well characterized and whose specificity is determined by a mix of hydrophobic and H-bonding interactions with the ligand. We have more recently begun to use MSCS to locate binding sites for proteins for which not all sites of interactions are known. The MSCS has now matured to a point where it can be reliably used to make new discoveries. For this, our group has developed a powerful computational tool for the analysis of MSCS data sets. The program Detection of Related Solvent Positions (DRoP) was written to produce a more consistent and automated description of the clustering associated with organic solvent molecules in MSCS and to rank the conservation levels of crystallographic water molecules in the set. DRoP allows for the addressing of crystallographic symmetry, gives cluster quality reporting and allows for cluster visualization (Kearney, 2012). By using DRoP we can better analyze how organic and water molecules affect the protein and thus be more confident about our interpretations regarding binding site locations and hydration properties.

BAF is of great interest to our study of protein surfaces because it contains non-specific sites for protein/DNA interactions. Thus we applied the MSCS method to human
BAF, which is a homodimeric, 89 amino acid, protein with no known similarity to other proteins and a molecular weight of about 10.1 kDa per subunit (Umland et al., 2000). The mouse BAF protein has 86 out of 89 residues identical to the human protein (Umland et al., 2000). BAF is highly conserved among other species as well. This allows us to compare experiments using BAF across different species to draw conclusions for the entire family. Based on the MSCS data set we identify a single cluster of organic solvent molecules per monomer at a site that interacts with LEM proteins. Interestingly we find a series of conserved water molecules at the DNA binding site, which we compare extensively to the structure of the BAF/DNA complex, gaining insight on possible roles of water in this non-specific protein/DNA interface.

**Experimental Methods**

*Purification of the BAF Protein*

For our study the Human BAF plasmid was received in a PET vector that was relevant for the protein purification process. BAF was purified using a previously published method with a few modifications. A fresh transformation must be performed before the *E. coli* cells containing the plasmid for BAF expression can be grown. A transformation of BAF in BL21 cells was performed by placing 3 µl of DNA into 1 tube of pre- aliquoted cells. Then the cells are incubated for 30 minutes, heat shocked for 40 sec at 42 °C and 200 µL of autoclaved LB are added, then placed in 37 °C shaker for 1 hour. Next the solution was plated on ampicillin resistant plates and placed overnight into 37 °C chamber. The following day 15 mg of ampicillin are added to 300 mL of autoclaved LB; a colony was taken from the plate previously grown and shaken overnight at 37 °C. Next day 50 mL of the overnight was
added to each of the 6 L of LB. The cells are grown to an OD of between .8 and 1 measured at 595 nm on the spectrometer. Once the desired OD is reached the protein was induced with IPTG (.93 g of IPTG was dissolved in 8 mL of water) adding 1 mL to each flask. The protein was then shaken for 4 hours and after it was done spinning down for 20 minutes at 7000 rpm the cells were collected, the pellet scrapped, placed in a specimen cup and put in -80 °C until ready for purification.

Cells are thawed on ice in 100 mL of Buffer 1 (10 mM Hepes pH 7.4, 150 mM KCl, 1 mM EDTA, 0.1 % Triton X 100, and a pinch of Lysozyme). Add Lysozyme and protease inhibitors right before use on cells. Protease inhibitors are added with the final concentrations of: 5 mM Benzamidine, 1 mM Pefabloc, 1 mM Antipain, 1 µg/mL Leupeptin, 1 µg/mL Pepstatin A, 1 µg/mL E64. The cells are resuspended using spatulas and 20 mL pipettes. The cells are pipetted up and down from the beaker to the sonication cup until all the frozen chunks of cells are gone. Sonication at power 16 was used to break the cells apart for 30 seconds on, 30 seconds off, 10 times on ice. The cell lysate was then spun in a centrifuge for 30 minutes at 18,000 rpm and 120 mL of Solubilization buffer (20 mM Hepes pH 7.4, 5 mM Imidazole, 6 M Guanidine HCl and 5 mM 2-Mercaptoethanol (BME)) was used to resuspend the pellets. The pellets are best resuspended by pipetting with the 5 mL pipette, then 10 mL and finally the 20 mL pipette. Using a 5 mL syringe, with a needle on it, all the cells are transferred to the sonication cup. The cells are again sonicated 10 times at power 16 (30 seconds on and 30 second off) keeping on ice and spun down at 18,000 rpm for 30 minutes. One liter of wash buffer (150 mM KCl, Hepes pH 7.4, 20 mM Imidazole, 6 M Guanidine HCl and 2-Mercaptoethanol (BME)) was made and split into two 500 mL aliquots. One
aliquot was used to make the elution buffer (10% Glycerol, 5 mM BME, and 2 M Imidazole) for use in later purification steps. The supernatant was then filtered using the 5 micron, 1 micron, .8 micron, and finally the .4 micron filters.

Before using the Nickel column the color was checked. If it is not blue, but rather it is a white or brownish color the His Trap recharging protocol must be followed. The recommended stripping buffer (20 mM sodium phosphate, 0.5 M NaCl, 50 mM EDTA at pH 7.4) was used for this protocol. First, the column was stripped by washing with 5–10 column volumes of stripping buffer; then followed by washing with 5–10 column volumes of binding buffer (stripping buffer without EDTA) and 5–10 column volumes of distilled water before re-charging the column. The water-washed column was recharged by loading 2.5 ml of 0.1 M Ni(NO$_3$)$_2$·6H$_2$O solution in distilled water on the HiTrap 5 mL column. The column was washed with 5 column volumes of distilled water and 5 column volumes binding buffer.

After the Ni column was stripped the column was again cleaned with ddH$_2$O and then loaded with the protein. The HPLC with a previously made program was used. Using a flow rate of 1-2 mL/minute, the column was washed with wash buffer. The UV was then zeroed and the protein was eluted using 100% of elution buffer. The fractions (5 – 9) of the peaks are collected for further use.
The collected protein fractions are then placed in dialysis tubing and set in dialysis buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM BME and .1 mM EDTA in 2 L) overnight in the refrigerator. The following morning BAF was placed into a new dialysis buffer and left to dialyze for 4 more hours. The protein concentration was then checked using the Bradford assay. 1mL of Bradford solution was added to 4 mL of water; then 999 μl of Bradford mixture was added to a cuvette and blanked at 595 nm. Then 1 μl of protein was added and slightly mixed. The reading was then taken and plugged into the formula for protein concentration, (OD - .12)/.05252/# of μl in 1 mL (divided by 1 since using 1 mL). After the dialysis the protein shows a lot of precipitate, which is fine. The protein was then pipetted up
and down using a 5 mL pipette to resuspend the protein. BAF was spun down at 17,000 rpm for 20 minutes and the supernatant was collected. Thrombin cleavage buffer (50 mM Tris-HCl pH-7.4, 150 mM NaCl, 2.5 mM CaCl2 and .1% BME in 1L) was made and 20 units of thrombin/mg of protein are added directly to the dialyzed protein. The protein with thrombin was then placed in a dialysis bag and dialyzed in the 18°C refrigerator overnight. Using a table top centrifuge the thrombin cleaved protein was concentrated to 5 mL. Next dialysis buffer was used to wash the Gel Filtration column using the wash program. Next the Gel Filtration program was run and the fractions of the last peak are collected for further use.

Figure 2. Gel Filtration results graph.
The final purified protein was then concentrated to 1.5 mL and the concentration was checked. The concentration was left around 10 mg/mL, flash frozen and placed in the -80°C until time for crystal growing.

**BAF Crystallization**

A reservoir solution was made using 10 mL of Hampton Crystal Screen 2 #44 (0.1 M Tris pH 8.5 and 20% w/v Ethanol) to 2.5 mL of 50% w/v PEG 1450 (in water) and then 500 μL was added to each well. The protein solution contained 20 mM Tris pH 7.4, 150 mM NaCl, .05 M EDTA and 10 mM TCEP. 10 μL of 100 mM TCEP was directly added to the protein solution before setting crystal trays. To set crystal trays 4 μl (from reservoir solution) x 4 μl (protein solution) drops are used with the hanging drop vapor diffusion method. Once the crystals are formed a cross-linking solution was made by first making stabilization buffer. First 360 μl of 1 M HEPES was added to 1440 μl of H₂O; then 360 μl of Ethanol was added to 1440 μl of the previously made 2. M HEPES (Solution A). Second 5 g of PEG 1450 was added to 10 mL of .2M HEPES (Solution B). To complete the stabilization buffer 1680μl of Solution A was added to 420μl of Solution B to make a solution with the ratio of 80% A to 20% B. After the stabilization buffer was made 25 μl of gluteraldehyde was added to 975 μl of stabilization buffer to make a .2% crosslinking solution. The crystals are allowed to crosslink for 35 – 40 minutes before removing for organic solvent soaks. Organic solvent soak solutions are made as follows: 70% - 250 μl of Tris-HCL pH 7.4, 50 μl of PEG 1450 and 700 μl of organic solvent; 60% - 334 μl of Tris-HCL pH 7.4, 66 μl of PEG 1450 and 600 μl of organic solvent; 50% - 418 μl of Tris-HCL pH 7.4, 82 μl of PEG 1450 and 500 μl of organic solvent; 40% - 501 μl of Tris-HCL pH 7.4, 99 μl of PEG 1450 and 400 μl of organic solvent.
solvent. The crystals are allowed to soak in the organic solvents for 1 hour. Before freezing in liquid nitrogen the crystals are briefly placed in a cryo solution made up of 25% 2-Methyl-2,4-pentanediol (MPD) and 75% stabilization buffer. The crystals are then shot with X-rays to check diffraction quality.

Figure 3. Rod-like shape of high diffraction quality BAF crystals.

Good diffraction quality crystal data are collected at APS the SERCAT ID-22 beamline at 100 K using 1 Å wavelength radiation and a Mar300 CCD detector at a distance of 150 mm. The data are scaled and processed using HKL2000 (Otwinowski & Minor, 1997).
A model of BAF (PDB code 1CI4) with water and ligand molecules removed was used to obtain an initial fit for all models. Refinement was accomplished using PHENIX (Adams et al., 2010). Coot was used to rebuild models and identify water and organic solvent molecules with the Fo-Fc electron density map set at 3σ and 2Fo-Fc electron density map set at 1σ (Emsley et al., 2010). Organic solvent molecules are placed in areas of well-defined electron density. All models are refined once without waters and twice with waters before placing in any organic molecules. Omit map calculations are used to confirm placement of organic solvents. All organic molecules had average B-factors similar to that of
the overall model as shown in Table 1. All completed structures are run through phenix.xtriage and DRoP to statistically check reliability of the models.

**Results and Discussion**

BAF crystallized in the space group P4_{3}2_{1}2 and diffract to high resolution. The organic solvent soaks included acetone 50% (ACT), dioxane 50% (DOX), dimethyl sulfoxide 50% (DMS), ethanol 50% (ETF), isobutanol (IBU), isopropanol (IPR), tert-Butanol 40% (TBU), 1.0 M trimethylamine N-oxide (TMO) and 1.5 M urea (URE). All soaked crystals collected have a resolution of 1.92 or better as shown in the table below.
<table>
<thead>
<tr>
<th>Data collection</th>
<th>X-linked ACT 50%</th>
<th>X-linked DOX 50%</th>
<th>X-linked DMS 50%</th>
<th>X-linked ETF 50%</th>
<th>X-linked IBU 50%</th>
<th>X-linked IPR 50%</th>
<th>X-linked TBU 40%</th>
<th>X-linked TMO 1.0M</th>
<th>X-linked URE 1.5M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
<td>P43212</td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>41.88</td>
<td>41.78</td>
<td>41.74</td>
<td>41.75</td>
<td>41.78</td>
<td>41.77</td>
<td>41.73</td>
<td>41.70</td>
<td>41.82</td>
</tr>
<tr>
<td></td>
<td>41.88</td>
<td>41.78</td>
<td>41.74</td>
<td>41.75</td>
<td>41.78</td>
<td>41.77</td>
<td>41.73</td>
<td>41.70</td>
<td>41.82</td>
</tr>
<tr>
<td></td>
<td>213.10</td>
<td>213.88</td>
<td>215.38</td>
<td>213.70</td>
<td>214.47</td>
<td>214.04</td>
<td>214.07</td>
<td>214.232</td>
<td>214.56</td>
</tr>
<tr>
<td>aβγ (°)</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
<td>90, 90, 90</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>50 - 1.75</td>
<td>50 - 1.79</td>
<td>50 - 1.92</td>
<td>50 - 1.87</td>
<td>50 - 1.87</td>
<td>50 - 1.83</td>
<td>50 - 1.87</td>
<td>50 - 1.92</td>
<td>50 - 1.75</td>
</tr>
<tr>
<td></td>
<td>(1.78 - 1.75)</td>
<td>(1.82 - 1.79)</td>
<td>(1.95 - 1.92)</td>
<td>(1.90 - 1.87)</td>
<td>(1.90 - 1.87)</td>
<td>(1.90 - 1.83)</td>
<td>(1.90 - 1.87)</td>
<td>(1.90 - 1.92)</td>
<td>(1.78 - 1.83)</td>
</tr>
<tr>
<td>Rmerge or Rfree</td>
<td>0.061</td>
<td>0.048</td>
<td>0.056</td>
<td>0.075</td>
<td>0.091</td>
<td>0.050</td>
<td>0.051</td>
<td>0.065</td>
<td>0.080</td>
</tr>
<tr>
<td></td>
<td>(0.380)</td>
<td>(0.275)</td>
<td>(0.595)</td>
<td>(0.816)</td>
<td>(0.555)</td>
<td>(0.409)</td>
<td>(0.321)</td>
<td>(0.341)</td>
<td>(0.431)</td>
</tr>
<tr>
<td>I/σ</td>
<td>64.0</td>
<td>54.7</td>
<td>44.2</td>
<td>55.6</td>
<td>56.6</td>
<td>67.0</td>
<td>49.6</td>
<td>49.5</td>
<td>60.4</td>
</tr>
<tr>
<td></td>
<td>(6.04)</td>
<td>(5.45)</td>
<td>(2.51)</td>
<td>(3.89)</td>
<td>(5.23)</td>
<td>(5.86)</td>
<td>(4.35)</td>
<td>(4.50)</td>
<td>(5.65)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.5</td>
<td>99.5</td>
<td>98.7</td>
<td>99.7</td>
<td>99.6</td>
<td>97.5</td>
<td>98.9</td>
<td>95.4</td>
<td>99.4</td>
</tr>
<tr>
<td></td>
<td>(99.3)</td>
<td>(97.3)</td>
<td>(92.5)</td>
<td>(99.9)</td>
<td>(99.9)</td>
<td>(96.9)</td>
<td>(90.1)</td>
<td>(95.4)</td>
<td>(97.9)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>9.6</td>
<td>9.8</td>
<td>10.9</td>
<td>13.1</td>
<td>11.7</td>
<td>13.8</td>
<td>10.2</td>
<td>13.0</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>(9.0)</td>
<td>(8.6)</td>
<td>(6.4)</td>
<td>(11.6)</td>
<td>(9.7)</td>
<td>(11.7)</td>
<td>(7.9)</td>
<td>(8.8)</td>
<td>(9.8)</td>
</tr>
<tr>
<td>Refinement</td>
<td>1.75</td>
<td>1.75</td>
<td>1.92</td>
<td>1.87</td>
<td>1.87</td>
<td>1.83</td>
<td>1.87</td>
<td>1.92</td>
<td>1.75</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>1.75</td>
<td>1.79</td>
<td>1.92</td>
<td>1.87</td>
<td>1.87</td>
<td>1.83</td>
<td>1.87</td>
<td>1.92</td>
<td>1.75</td>
</tr>
<tr>
<td>No. reflections</td>
<td>20104</td>
<td>18823</td>
<td>15404</td>
<td>16647</td>
<td>16728</td>
<td>17394</td>
<td>16455</td>
<td>15391</td>
<td>20259</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17798</td>
</tr>
<tr>
<td></td>
<td>23.68</td>
<td>24.49</td>
<td>25.91</td>
<td>24.80</td>
<td>23.76</td>
<td>25.39</td>
<td>25.25</td>
<td>26.01</td>
<td>25.14</td>
</tr>
<tr>
<td>No. atoms</td>
<td>1406</td>
<td>1406</td>
<td>1406</td>
<td>1406</td>
<td>1406</td>
<td>1406</td>
<td>1406</td>
<td>1406</td>
<td>1406</td>
</tr>
<tr>
<td>Protein Org.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Water</td>
<td>75</td>
<td>96</td>
<td>76</td>
<td>89</td>
<td>82</td>
<td>103</td>
<td>98</td>
<td>94</td>
<td>96</td>
</tr>
<tr>
<td>R.M.S.D. Bond lengths (Å)</td>
<td>.007</td>
<td>.007</td>
<td>.008</td>
<td>.007</td>
<td>.008</td>
<td>.007</td>
<td>.007</td>
<td>.007</td>
<td>.008</td>
</tr>
<tr>
<td>Bond angles (Å)</td>
<td>.952</td>
<td>1.02</td>
<td>.950</td>
<td>.981</td>
<td>.967</td>
<td>.915</td>
<td>.913</td>
<td>.939</td>
<td>.978</td>
</tr>
</tbody>
</table>
BAF was seen as the biologically relevant dimer within the asymmetric unit as shown in Figure 5.

![BAF dimer with organic solvent spheres. Figure created using Pymol (Schrodinger, 2010).](image)

Figure 5. BAF dimer with organic solvent spheres. Figure created using Pymol (Schrodinger, 2010).

Also seen in Figure 5 BAF contained two organic solvent clusters between the interfaces of the monomers. Cluster 1 contained a representation of all organic solvents while cluster 2 contained all organic solvents except for acetone. In both clusters the organic molecules are interacting with the backbone oxygen atom where most of the interaction occurs and directly hydrogen bonding Val51. No organic molecules are in the DNA binding pockets of BAF. This may be in part because the DNA binding region of BAF is located in areas of crystal contacts, but it also has to do with the charged nature of that interface. Figure 6 shows the
organic solvent soaked structures superimposed with the published structure of the BAF/DNA complex (pdb code 2BZF).

As seen here there are no organic solvents where DNA is normally bound. Even though there are only two hotspots, they are highly important.

Seven mutants (25E, 25Q, 46E, 47E, 51E, 53E, and 54E) show little or no binding to emerin (Segura-Totten, Kowalski, Craigie, & Wilson, 2002). BAF interacts specifically with emerin within the LEM (LAP2, emerin, MAN1) domain. Emerin is an important protein within the inner nuclear membrane in vertebrates (Clements, Manilal, Love, & Morris, 2000). Emerin is crucial for anchoring proteins to the membrane which may cause emerin to have direct or indirect roles in nuclear assembly, regulating a functional actin cortical network at the nuclear envelope, and gene expression (Bengtsson & Wilson, 2004). This is of significance because the organic molecules in the MSCS structures are seen to align with a
residue in the published NMR structure (pdb code 2ODG) of the BAF/LEM domain as seen in Figure 7.

Figure 7. BAF/LEM-domain and BAF/DNA structures superimposed with soaked structures. Figure created using Pymol (Schrodinger, 2010).

Figure 7 also shows an example of how BAF can interact with a domain of a protein and DNA at the same time. This is important since BAF is thought to be essential in many processes within the cell including nuclear assembly. It is essential that emerin binds BAF for reassembly at the nuclear envelope after mitosis (Bengtsson & Wilson, 2004). During telophase it was shown that BAF and emerin co-localize at the ‘core regions’ of segregated chromosomes (Bengtsson & Wilson, 2004). If emerin proteins are unable to bind BAF they
are not recruited to the core regions thus failing to assemble at the inner nuclear membrane (Bengtsson & Wilson, 2004).

Figure 8 provides a closeup view of cluster 1, showing excellent overlap between the organic solvent molecules and Pro25 from the published NMR structure (pdb code 2ODG).

This is direct evidence for the ability of the MSCS to be able to pick out areas of protein-protein interactions. This is the first time there is crystallographic evidence for an important hotspot on the BAF surface that is not the DNA binding site, with previously published
structures being either simulation or NMR. Another advantage of the MSCS method is the ability to analyze water through the DRoP program.

There are 161 unique clusters with 66 of those being considered as highly conserved. Figure 9 shows the overall positions of the highly conserved waters within BAF.

Figure 9. Overall water analysis of BAF. A – Front view of BAF and B – Back view of BAF. Figure created using Pymol (Schrodinger, 2010).
Figure 9 only contains the highly conserved waters. 100% conservation is defined as water seen in a particular position in all structures. In the figure 100% conserved waters are red, 90% conserved are orange and 80% conserved are yellow. These waters of high conversation are mostly seen on the back part of BAF and the areas of DNA interactions. Where the organic solvents are bound BAF binds emerin leading to waters of low conservation, which makes sense because areas of protein interactions need to have easily displaceable waters. There are also no crystal contacts on the front or backside of the BAF dimer. Using the coloring function of DRoP as seen in Figure 10 we can easily locate hydration patterns throughout a protein.

Figure 10. BAF-DNA structure superimposed using coloring function. Amino acid residues are colored based on the conservation level associated with interacting water molecules. Residues that interact with 100% conserved waters are blue, 90% cyan, and 80% green. Figure created using Pymol (Schrodinger, 2010).
DRoP has a visualization tool with the ability to generate an additional set of PDB files of the analysis set that reassigns the B-factors of the amino acid residues interacting with water molecules according to how often the water molecule appears in the set (Kearney, 2012). It is accomplished by placing a value in the B-factor column in the pdb file and then using the color by B-factor option in Pymol (Kearney, 2012). As seen in Figure 10, areas that contact 100% conserved waters are blue, 90% conserved cyan and 80% conserved green. Both DNA molecules overlay with areas of protein contacting highly conserved waters. However, comparing both sides of the dimer, the side shown in Figure 10a contains more conserved water molecules than the side in Figure 10b. This is because the side of the dimer shown in Figure 10b is in an area of substantial crystal contact as confirmed by electron density. Therefore, it would be expected that while both sides of the dimer probably make similar protein-water interactions, they are different in the two molecules in the asymmetric unit of the crystal due to different crystal packing. Looking further into the water analysis we can see that Figure 11 gives an example of how waters can delineate ligand groups of the interacting molecule.
As seen in Figure 11 there are two 100% conserved waters (3 and 14) overlaid on the oxygen groups of the phosphate backbone connecting bases 5 and 6. This is a prime example of how even if organic solvents are unable to find hotspots the conserved waters can be used as a secondary tool. This tool is also useful because of the size of organic molecules versus waters. Because of crystal contacts sometimes the larger molecules are not able to fit within these compact areas. However, waters are able to easily fit within these small areas, giving us another useful mapping tool.

Looking deeper at the plethora of conserved waters in the DNA binding pocket shows that while the waters are not directly aligning with the DNA they are in areas where they could mediate the protein/DNA interaction. This is typical of protein-ligand interactions where waters help to mediate initial contacts. This could also explain how BAF makes only
minimal contact with the DNA. Most of the conserved waters, however, do not mediate contacts between protein and DNA but act as a placeholder until ligand binding. Once DNA binds the waters shift but stay close enough to the DNA to supplement the phosphate backbone hydrogen bonding requirements. BAF only interacts with three phosphates on one strand of the DNA and two on the other strand (Bradley et al., 2005). This minimal binding mode makes no contacts with DNA bases and only uses four side chains of BAF (Farnet & Haseltine, 1990). This ensures that the BAF-DNA interaction is sequence independent.

Figure 12 shows the contacts between the DNA and BAF through water molecules based on our MSCS sets.

Figure 12. Water mediated network between BAF and DNA. Figure created using Pymol (Schrodinger, 2010).
Figure 12 shows an example of protein and DNA water mediated contacts. The carbonyl group of the side chain of Asn70 hydrogen bonds with water 34 (100% conserved), which then hydrogen bonds water 46 (90% conserved) that directly contacts the phosphate backbone of the DNA between bases 5 and 6. This phosphate backbone oxygen group hydrogen bonds to water 110 (30% conserved), which in turn hydrogen bonds to Lys72. Mutations of Lys72 results in undetectable DNA binding of BAF (Umland et al., 2000) even though it is not known to make direct DNA interactions. This water network reveals the mechanism through which Lys72 affects DNA binding without direct interactions. Asn70 has no direct implications in any mutational studies. However, Asn70 is part of a region that resembles a conserved HhH motif in sequence and structure but is not able to be classified as one because Asn70 (most significant differing residue) replaces a conserved glycine residue of the motif (Umland et al., 2000). The involvement of this side chain in the water network may justify the exception to the observed pattern (Umland et al., 2000). Another example of a protein DNA water mediated network can be seen in Figure 13.
Figure 13 shows an example of a mini water network involving multiple residues mediating contacts between BAF and DNA and also within the DNA bases. Lys33 hydrogen bonds water 99 (30% conserved), then water 99 hydrogen bonds water 50 (90% conserved), next water 50 hydrogen bonds water 67 (70% conserved), which interacts directly with the oxygen of the DNA phosphate group between bases 3 and 4 leading to a direct hydrogen bond between the phosphate backbone oxygen group and Gly27. Alternatively, water 50 hydrogen bonds water 69 which then hydrogen bonds the amino backbone group of Asn70. Lastly water 50 also has the ability to directly hydrogen bond the phosphate backbone oxygen between bases 4 and 5. When Lys33 and Gly27 are mutated the BAF binding to
DNA is undetectable (Umland et al., 2000). While Gly27 directly interacts with the backbone of DNA, Lys33 makes no such interactions. However, one possible explanation for Lys33 affecting DNA binding could be this water network. Asn70 is a common residue between both water-mediated networks. However, more experiments are needed to determine if this residue affects BAF and DNA interactions.

Conclusions

The MSCS method is a powerful tool for studying protein characteristics. Seen here is an example of how this tool can be applied to a protein that binds DNA. The organic solvents successfully picked out an area of protein-protein interactions. The proline of the LEM-domain of emerin aligned with the organic solvents within the pocket. These organic molecules are also seen to interact with the same Val 51 of BAF, mimicking residues in emerin. Previous research shows this residue is essential for the proteins’ interactions. Deficiency within the binding of BAF and emerin leads to defects in DNA condensation during mitosis. This is just one example of how MSCS identifies important areas within a protein. Another advantage of MSCS is its water analysis. We successfully demonstrated the ability of MSCS to pick out ligand interaction sites through water analysis instead of organic solvent analysis. Even though one monomer of BAF had its DNA binding pocket in an area of considerable crystal contacts we are still able to pick out important characteristics. Highly conserved waters are seen to align with the oxygen groups of the phosphate backbone of DNA. There are also many waters within the binding cavity believed to be examples of how waters can act as placeholders for the ligand. Finally, our analysis revealed water networks connected to residues whose mutations disrupt DNA binding. This water analysis is a
wonderful secondary tool for finding important interaction information where organic solvents may not be available. These results lead to further ideas of finding BAF within a different crystal symmetry. This would allow us to possibly soak in organic solvents within the DNA binding cavity to show direct correlations between MSCS and DNA-protein interactions. Regardless, this research further validates the MSCS method as an exemplary tool for finding areas of protein-protein interactions, protein-ligand interactions and conducting a consistent comprehensive water analysis. In addition, it reveals important potential water-mediated protein DNA interactions that are not available due to the low resolution of the structure of the BAF/DNA complex.

**Acknowledgements**

Thanks to Indhira De la Rosa for initially getting the project started in our lab. Thanks to Klim Verba for taking the time to show me the initial protocol for BAF purification. Thanks to Melissa Young for setting multiple crystal trays and discovering TCEP would be a better reducing agent. Thanks to Dr. Susan Fetics for answering questions about DNA sequencing and cleaving protocols in regards to BAF. Thanks to Dr. Greg Buhrman for showing and giving ideas on how to better the BAF purification process. Thanks to Dr. Paul Swartz for figuring out the crystallization, soaking and cryo conditions. Thanks to Dr. Bradley Kearney for all the help in using the DRoP program and answering related questions.

Data are collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-BM 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. This research was supported by the NSF (MCB - 1237512) grant.

Contributions

I purified protein, grew crystals, soaked crystals, plus collected and analyzed all the data presented here.
OVERALL CONCLUSIONS

Presented here are the many results and applications of analysis that can be achieved through use of the MSCS method. One important aspect of MSCS is the need to have a representative set of the multiple protein types. My research expanded this set by incorporating results for a large sugar binding protein, a mutant/wild type protein comparison, and a DNA binding protein. Through use of these different protein types we are able to begin to observe some overall characteristics of binding sites. Organic solvents in Ras and BAF bound pockets that were implicated in complex formation through insertion of an amino acid residue. These complex formations were experimentally shown to greatly affect the cell in multiple ways with disruptions leading to diseases. These findings are in overall agreement with previously published results in RNAse A and Elastase. P22TSP and Lysozyme demonstrate how the MSCS method can detect the sugar binding pockets when the active site is empty. These results helped to expand our database proving the ability of MSCS to pick out active sites even when the proteins interact with different size sugar moieties such as P22TSP (octasaccharide or greater) and Lysozyme (hexasaccharide). BAF shows an example of how a protein interacting with the backbone of DNA nonspecifically utilizes water networks more so than direct interactions. This was a trend also seen in P22TSP where the sugar moieties interact through highly conserved waters to make protein contacts. In both P22TSP and RAS water molecules essential to catalysis are highly conserved no matter the organic solvent environment. Overall this work shows the ability of MSCS in mapping protein-protein, protein-ligand and water interactions. By adding to our
experimental database we will be able to show MSCS as a viable method for the beginning of drug discovery as long as a crystal is obtainable and can be crosslinked. With the Protein Data bank having thousands of proteins with no known function this method could also be used to elucidate areas of high importance. Since a high-resolution structure would already be able to be obtained MSCS could pick out sites in which to start experiments. This could save time and resources allowing protein functional studies to be more streamlined. Lastly, as other experimental sets are added to the overall body of work MSCS could also be used to determine more detailed characteristics of binding pockets. My research helped to add to this experimental database and hopefully in the future MSCS will be seen as a method to help drive forward the protein drug targeting field.
REFERENCES


retroviral DNA from autointegration. [10.1038/2345]. Nat Struct Mol Biol, 5(10), 903-909.


Fetics, S. (2012). *Structural Analysis of Ras Bound to Raf Kinase and Implications of Ras Allosteric Site.* (Doctoral Dissertation), North Carolina State University. Retrieved from [http://repository.lib.ncsu.edu/ir/bitstream/1840.16/7882/1/etd.pdf](http://repository.lib.ncsu.edu/ir/bitstream/1840.16/7882/1/etd.pdf)


Goh, C.-S., & Cohen, F. E. (2002). Co-evolutionary Analysis Reveals Insights into Protein–Protein Interactions. *Journal of Molecular Biology, 324*(1), 177-192. doi: [http://dx.doi.org/10.1016/S0022-2836(02)01038-0](http://dx.doi.org/10.1016/S0022-2836(02)01038-0)


Harris Ca Fau - Andryuk, P. J., Andryuk Pj Fau - Cline, S., Cline S Fau - Chan, H. K., Chan Hk Fau - Natarajan, A., Natarajan A Fau - Siekiera, J. J., Siekierka J Fau - Goldstein, G., & Goldstein, G. Three distinct human thymopoietins are derived from alternatively spliced mRNAs. (0027-8424 (Print)). doi: D - NLM: PMC44185 EDAT-1994/07/05 MHDA- 1994/07/05 00:01 CRDT- 1994/07/05 00:00 PST - ppublish


31P Solid State NMR Spectroscopy. *Journal of Molecular Biology*, 323(5), 899-907. doi: [http://dx.doi.org/10.1016/S0022-2836(02)01010-0](http://dx.doi.org/10.1016/S0022-2836(02)01010-0)


Appendix A Copyright Permissions
This is a License Agreement between Mychal D Smith ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier: Elsevier Limited
The Boulevard, Langford Lane
Kidlington, Oxford, OX5 1GB, UK

Registered Company Number: 1982084

Customer name: Mychal D Smith
Customer address: 104 South Street
Bridgewater, MA 02324

License number: 3095720062549
License date: Feb 25, 2013

Licensed content publisher: Elsevier
Licensed content publication: Journal of Molecular Biology
Licensed content title: Analysis of Binding Site Hot Spots on the Surface of Ras GTPase
Licensed content author: Greg Buhrman, Casey O'Conner, Brandon Zerbe, Bradley M. Kearney, Raeanne Napoleon, Elizaveta A. Kovrigina, Sandor Vajda, Dima Kozakov, Evgenii L. Kovrigin, Carla Matos
Licensed content date: 4 November 2011
Licensed content volume number: 413
Licensed content issue number: 4
Number of pages: 17
Start Page: 773
End Page: 789
Type of Use: reuse in a thesis/dissertation
Portion: figures/tables/illustrations
Number of figures/tables/illustrations: 2
Format: both print and electronic
Are you the author of this: No

https://1s100.copyright.com/CustomerAdmin/PLF.jsf?ref=671e2099-7e68-4f0c-b63f-0aad634e65
INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol./edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC’s Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC’s Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC’s Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher’s written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher’s behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC’s Billing and Payment terms and conditions. These terms and conditions, together with CCC’s Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC’s Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to
LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world English rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article. If this license is to re-use 1 or 2 figures then permission is granted for non-exclusive world rights in all languages.

16. **Website**: The following terms and conditions apply to electronic reserve and author websites:

   **Electronic reserve**: If licensed material is to be posted to website, the web site is to be password-protected and made available only to bona fide students registered on a relevant course if.

   This license was made in connection with a course,

   This permission is granted for 1 year only. You may obtain a license for future website posting.

   All content posted to the web site must maintain the copyright information line on the bottom of each image,

   A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com, and

   Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

17. **Author website** for journals with the following additional clauses:

   All content posted to the web site must maintain the copyright information line on the bottom of each image, and the permission granted is limited to the personal version of your paper. You are not allowed to download and post the published electronic version of your article (whether PDF or HTML, proof or final version), nor may you scan the printed edition to create an electronic version.

   A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx. As part of our normal production process, you will receive an e-mail notice when your article appears on Elsevier’s online service ScienceDirect (www.sciencedirect.com). That e-mail will include the article’s Digital Object Identifier (DOI). This number provides the electronic link to the published article and should be included in the posting of your personal version. We ask that you wait until you receive this e-mail and have the DOI to do any posting.

   Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

18. **Author website** for books with the following additional clauses:

   Authors are permitted to place a brief summary of their work online only.

   A hyper-text must be included to the Elsevier homepage at http://www.elsevier.com. All content
posted to the web site must maintain the copyright information line on the bottom of each image. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version.

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

19. Website (regular and for author): A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx, or for books to the Elsevier homepage at http://www.elsevier.com

20. Thesis/Dissertation: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

21. Other Conditions:

v1.6

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLKN500963755. Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratiss licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.