ABSTRACT

DELLINGER, ANDREW EVERETTE. Computational Biology of Ras Proteins. (Under the direction of William R. Atchley.)

In this research, computational biology is used to elucidate how evolutionary history has changed roles of structure and function among Ras proteins, with a focus on the Ras family. This dissertation begins with phylogenetic analyses of the Ras superfamily and Ras family. Phylogenetic trees of the Ras family were estimated using Neighbor-Joining, Weighted Neighbor-joining, Parsimony, Quartet Puzzling, Maximum Likelihood and Bayesian methods. In nearly all cases, each clade represented a subfamily. Clade members and clade divisions were consistent among all the trees, increasing the probability of a correct estimation of the evolutionary history.

Further investigation into the evolution of sequence involved decomposing sequence covariation into its respective components. The roles of the functional and structural components of covariation were the focus of several multivariate analyses. Decision tree analysis, a data mining method, found that sequence divergence in critical sites of the hydrophobic core, dimerization regions and ligand binding regions were sufficient to divide Ras subfamilies. Alignments of GDP-bound and GTP-bound crystal structures revealed that only Ral and M-Ras proteins have structural variation in the effector binding switch I regions, while all Ras structures vary in the protein binding switch II region. Di-Ras2-GDP was shown to have a unique C-terminal loop which binds to the interswitch region. Last, a common factor analysis was computed. The factors contain the set of sites that both discriminate among the subfamilies and have a unique functional or structural role, such as Ral tree-determinant sites.

Finally, sequence signatures were developed for each of the families of the Ras superfamily using Boltzmann-Shannon entropy. This method was compared to the PROSITE signature, profile hidden Markov model and MEME position-specific scoring matrix methods. The Entropy method identified approximately 8% fewer proteins than the best of the other methods, MEME. Comparative analyses of these sequence signatures determined which sites and amino acids played important roles in the changes in protein function and structure among Ras families.

Computational Biology of Ras Proteins

by
ANDREW EVERETTE DELLINGER

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

Bioinformatics

Raleigh

2006

APPROVED BY:

_____ _____

Chair of Advisory Committee

_____ _____

DEDICATION

This dissertation is dedicated to the glory of God, my Creator who gave me all my talents and abilities. By His grace He helped me to go beyond my abilities in order to write this dissertation and pass my oral exam and defense.

BIOGRAPHY

I was born on March 5, 1974 in Hickory, North Carolina. I was an honor student throughout school and was especially interested in the sciences. I was valedictorian of Saint Stephens High School out of a class of 205. I struggled to decide whether to attend North Carolina State University or the University of North Carolina at Chapel Hill. I began at UNC-Chapel Hill as a Chemistry major and switched to Pharmacy school because Pharmacy seemed like a good-paying job in the sciences. I graduated with a B. S. Pharmacy in 1997 and subsequently worked about 9 months as a Pharmacist in Shelby, N.C.

Working as a pharmacist was stressful because I had to deal with shoplifters, drug-seekers and scam artists plus every person expecting their prescription right now. I lost 20 pounds, becoming a 105 pound person. I chose to quit and attend North Carolina State University to get a second B.S. in Computer Science. I had computers since about the second grade, where I wrote programs on my Commodore 64 in BASIC. I had other programming classes in school such as for the Pascal language. So I hoped to combine my love for computers with Pharmacy. During my undergraduate study, I received a job at the National Institute of Environmental Health Sciences (NIEHS) working for Drs. Doug Bell and Robert Boissy on a genetic database project. Dr. Bell encouraged me to attend microarray seminars. There I met Dr. Bruce Weir, who at the time was the head of the Bioinformatics program at NC State. He encouraged me to apply for the Bioinformatics Ph. D. program, which I did after receiving my B.S. in Computer Science in December 2000. I chose Dr. Atchley as an adviser because I was interested in working with Bioinformatics methods to analyze protein

families. It's been a long journey but I believe that the purpose of all my education is to make me an excellent and fruitful researcher in the future. I hope my current and future research will be able to be used for the practical benefit of others.

ACKNOWLEDGMENTS

First, I would like to acknowledge God's help, grace, hope and strength which was required in order to achieve a Ph. D. Second, I appreciate all of Dr. Atchley's advice, revision suggestions and support these last few years. I thank all the Atchley lab members, past and present, who have given me presentation and research suggestions as well as encouragement. I thank Andrew Fernandes for setting up several research and computational tools for the lab to use.

Third, I would like to acknowledge all the help, support, and prayer of my church, family and friends in writing this dissertation. I thank my wife Heather for helping me to do all the things I never had time to do throughout my Ph. D. study. I thank her for her love and spiritual and emotional support. I thank her for grammatical correction of my writing. I thank my parents who encouraged me throughout this process even though I am still getting my education at 32 years old.

I thank my Bible teacher and leader of our church, Sh. John Martin and his wife Sh. Grace Martin for their prayers, spiritual guidance, inspiration and practical advice about scheduling, post-doc applications, etc. I would also like to thank our church members and my family for their prayers.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

Introduction

Sequencing genomes from a wide variety of species has produced a huge number of predicted protein sequences. Indeed, the number of predicted sequences far exceeds the number of well-annotated sequences, which in turn far exceeds the number of available solved protein structures. At the beginning of 2006, there were over 4.22 million protein sequences in the Entrez search engine (Wheeler et al. 2006). However, there are only about 200,000 annotated sequences in the Swissprot database and about 34,400 crystal structures in the protein database PDB (Berman et al. 2000). The lag in producing structural and functional data is due to the time- and labor-intensive experimental processes required to generate these data. There is a desperate need for methods that would more quickly generate structural and functional data from sequences.

Advances in computational biology offer an important means to fulfill this need for detailed information of functional and structural aspects of proteins. Computational biology incorporates the fields of biology, statistics, biomathematics, and computer science. The *in silico* methods used in computational biology have the potential to give deeper insight into aspects of proteins structure and function, and in some cases, to do so more quickly and less-expensively than many *in vitro* and *in vivo* experimental methods.

The primary goal of computational biology is to be able to quickly infer detailed biological knowledge from raw sequence data. Computational biology has the capacity to provide extensive information about proteins not available from typical crystal structures, MRI, and related methods. Through analyses of large numbers of related sequences,

computational biology can provide important information about evolutionary relationships, patterns of multidimensional covariation among amino acids, relationships of these patterns to experimentally known structure, delimitation of information about sequence variability and its relevant underlying causes that are not available from crystal structure and MRI studies. Methods in computational biology offer a large and powerful battery of tools that can be used in conjunction with molecular methods.

Experiments to determine the functional roles of protein sites are usually limited to mutagenesis of a few critical amino acid sites either one at a time or in sets of two or three sites. However, computational approaches can generate experimentally testable hypotheses which direct experiments which are more complex in their assessment of functional and structural interactions.

Herein, computational biology is used to estimate the phylogenetic relationships among Ras superfamily proteins, explore the set of functionally and structurally significant sites that define the Ras families and analyze the functional, structural and phylogenetic components of covariance in the amino acid sites of Ras family proteins.

The Ras superfamily is a large, diverse and important group of proteins that provide excellent material for development and implementation of new and existing methods to investigate these topics. Ras superfamily proteins are found in eukaryotes but not in prokaryotes and Archaea (Jékely, 2003). They have a common core structure and common function as GTPases despite having as little as 12% sequence identity among the proteins of its six families (Sprang, 1997). The Ras superfamily is fairly well known structurally, having several solved crystal structures in each family except the Rad-Gem-Kir (RGK) family.

Many proteins are also functionally known given the important roles of these proteins in disease and cell function. However, there is still much to learn about Ras evolutionary history, structural and functional interactions among amino acid sites, and other aspects of Ras biology.

The Ras superfamily is biologically significant because its proteins are involved in a number of human diseases. The most frequently observed oncogenes in human cancer are Ras family genes (Giehl, 2005). Rac proteins are targets for the botulinum toxin (Didsbury et al. 1989). Arf activates cholera toxin (Zeghouf et al. 2005). And Rab9 is a viral transport protein for HIV, Ebola and other important disease-causing viruses (Chen et al. 2004). Because of their importance in disease and in cell biology, there is a large literature about the function of Ras superfamily proteins.

Ras superfamily proteins differ in their specific biological function. Sar/Arf family proteins initiate vesicle budding and provide binding sites for coat proteins (COPs), which form the vesicle membrane. Sar proteins are part of the COP-II vesicles which transport proteins from the Endoplasmic Reticulum to the Golgi. Arf proteins are part of the COP-I vesicles which transport proteins between the cisternae of the Golgi from *trans* to *cis* and from the Golgi to the ER (Lewin, 2000). Rab proteins control vesicular trafficking (Schimmöller et al. 1998). Rho family proteins act on the actin cytoskeleton to assemble stress fibers (Ridley and Hall, 1992), form lamellipodia and filopodia, control cellular adhesion (Hall, 1998) and cell shape (Majumdar et al. 1998) and assist in phagocytosis (Chimini and Chavrier). They are also important in G1 to S phase cell cycle progression (Olson et al. 1995). Ras family proteins control cell proliferation, differentiation,

transformation, growth, exocytosis and transcription (Feig, 2003; Reuther and Der, 2000; Kinbara, 2003, Saucedo, et al. 2003). Some functions of Ran family proteins include: control of nuclear import and export of proteins, regulation of spindle assembly in mitosis and regulation of DNA replication (Wilde and Zheng, 1999; Ciciarello and Lavia, 2005; Lounsbury et al. 1996; Li et al. 2003). And RGK family proteins regulate calcium channel activity, morphology and cytoskeleton reorganization (Finlin et al. 2003; Beguin et al. 2005; Piddini et al. 2001; Pan et al. 2000).

Though their specific functions vary, Ras superfamily proteins share a general biological function as GTPases that act as molecular switches. Ras superfamily proteins have two major binding states. When the proteins are bound to GTP, they are in an active state. They can bind to their effector proteins and perform their specific biological function. Effector binding is enabled in the GTP-bound state because the effector loop L2, also known as switch I, is in the proper conformation. The switch II region, which is composed of L4 and part of $\alpha 2$, is also in the active conformation and can bind accessory proteins like GTPase Activating Proteins (GAPs). The intrinsic rate of hydrolysis in the H-Ras protein is 0.03 min^{-1} which is too unresponsive for some biological tasks. GAPs increase the rate in which Ras proteins hydrolyze GTP into GDP and an inorganic phosphate. Upon GTP hydrolysis, the switch regions change conformation. In this conformation, switch I cannot bind to effectors but switch II can still bind to accessory proteins such as Guanine nucleotide Exchange Factors (GEFs). A Mg^{2+} ion is hexacoordinated in Ras superfamily proteins. In the GTP-bound state it is coordinated to the conserved Threonine in switch I, the conserved Serine or Threonine in the P-loop, the β- and γ-phosphates of GTP and two water molecules. In the

GDP-bound state the switch I and γ-phosphate are replaced by water molecules. The Mg^{2+} ion decreases the rate of spontaneous GDP and GTP dissociation by four orders of magnitude to 10^{-8} and 10^{-9} s^{-1} respectively (Sprang, 1997). Thus GDP dissociation is unlikely to occur spontaneously. After the proper biological stimulus, GEFs are used to stimulate the dissociation of GDP. After the dissociation, GDP is often replaced with GTP since the intracellular ratio of GTP to GDP is 10:1 (Giglione and Parmeggiani, 1998). Upon GTP binding the switch regions change to their active-state conformation and the cycle is complete.

Ras superfamily proteins have a common core structure of five α-helices, a six stranded β-sheet and 10 loops (Nicely et al. 2004; Sprang, 1997). There is also structural variation in the superfamily. For example, the Arf family has an N-terminal helix that forms a hasp over the interswitch region in the GDP-bound form of the protein. Di-Ras2 proteins have a C-terminus which also interacts with the interswitch region. And Ran proteins have a β7 strand and a division in helix 1, forming the α1a and α1b helices (Sprang, 1997; Pasqualato et al. 2002; PDB ID: 2ERX; Papagrigoriou et al. 2005 *unpublished*).

Chapter 1 of this dissertation is a phylogenetic analysis of Ras superfamily proteins with a focus on the Ras family. The estimation of the superfamily tree provides insight into the order in which the families arose. For example, Arf appears to have arisen before Sar, Rab before Ran, RGK before Ras, and Ras and Rho came from a common ancestor. Ran shared an ancestor with two Rab-like proteins, suggesting that the ancestor of the Ran family was an ancestral Rab protein.

Several methods of phylogenetic estimation were used on a subset of 98 Ras proteins

from the well-annotated Swissprot database. Neighbor-joining, weighted neighbor-joining, parsimony, quartet puzzling, maximum likelihood and Bayesian methods were used to increase the confidence in the relationships among the Ras family proteins according to their clade memberships. The subfamilies of the Ras family consistently formed distinct clades. Few if any members of a subfamily were missing from the clade. The relationships among the clades were inconsistent due to small bootstrap values in the deeper nodes of the tree. In trees which did attempt to estimate these relationships, the subfamilies responsible for cell growth were nearest to the prokaryotic G-protein outgroup while the subfamilies responsible for cell division were deeper in the tree.

The sequence of the Ras family ancestral protein was estimated using both parsimony and maximum likelihood techniques. In a BLAST search, both the ancestral sequence computed by parsimony and the sequence computed by maximum likelihood were most related to a Rheb protein (Altschul et al. 1990). The Rheb clade is nearest to the outgroup in the phylogenetic trees where the bootstrap values were large enough to make such an estimation.

In chapter 2, a set of multivariate analyses were conducted to explore the relationships among sequence, structure and function in Ras family proteins. Decision tree analyses were conducted to determine which residues in which sites are critical for discriminating function. The sites chosen by the decision tree analyses were involved in protein binding, dimerization, maintaining the hydrophobic core and physiochemical interactions with important functional and structural sites. Second, the crystal structures of Ras family proteins were aligned in the GTP-bound state and in the GDP-bound state. More structural variation was found in the

completely disordered switch II region than in the partially disordered switch I region (Nicely, et al. 2004). Structural variation was also found in the interswitch region, loop L7 and adjacent to the three indels. The alignment revealed a unique structural component to Di-Ras2-GDP, where the C-terminus interacts with the interswitch region. Third, a common factor analysis was performed on the multiple alignment of the Ras family. The sites of the resulting factors discriminate subfamilies of the Ras family and share a common functional or structural role. Sites within factors 4 and 5 form two clusters of interacting sites involving the C-terminal ends of switch I and switch II.

In chapter 3, entropy-based sequence signatures were generated for each family of the Ras superfamily. The ability of these signatures to detect and classify superfamily proteins were compared to the abilities of profile hidden Markov models (Eddy, 1998), MEME's position specific scoring matrices (Bailey and Elkan, 1994) and PROSITE's signatures (Sigrist et al. 2002). Entropy signatures retrieved fewer sequences on average than MEME and profile hidden Markov models. However, Entropy signatures have the advantage over these methods that they describe the important sites in the diversification of structure and function, require fewer sites and provide a simple and interpretable way to delimit and define a family.

Literature Cited

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403-410.

Bailey, T. L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California.

Beguin, P., R. N. Mahalakshmi, K. Nagashima, D. H. Cher, A. Takahashi, Y. Yamada, Y. Seino, and W. Hunziker. 2005. 14-3-3 and calmodulin control subcellular distribution of Kir/Gem and its regulation of cell shape and calcium channel activity. J. Cell Sci. **118**:1923-1934.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Research **28**:235-242.

Chen, L., E. DiGiammarino, X. E. Zhou, Y. Wang, D. Toh, T. W. Hodge, and E. J. Meehan. 2004. High Resolution Crystal Structure of Human Rab9 GTPase A NOVEL ANTIVIRAL DRUG TARGET. J. Biol. Chem. **279**:40204–40208.

Chimini, G., and P. Chavrier. 2000. Function of Rho family proteins in actin dynamics during phagocytosis and engulfment. Nature Cell Biology **2**:E191 – E196.

Ciciarello, M., and P. Lavia. 2005. New CRIME plots: Ran and transport factors regulate mitosis. EMBO reports **6**:714–716.

Didsbury, J., R. F. Weber, G. M. Bokoch, T. Evans, R. Snyderman. 1989. rac, a novel ras-related family of proteins that are botulinum toxin substrates. J. Biol. Chem. **264**:16378-82.

Eddy, S. R. 1998. Profile Hidden Markov Models. Bioinformatics **14**:755-763.

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

Finlin, B. S., S. M. Crump, J. Satin, and D. A. Andres. 2003. Regulation of voltage-gated calcium channel activity by the Rem and Rad GTPases. Proc. Natl. Acad. Sci. U S A. **100**:14469-14474.

Giehl, K. 2005. Oncogenic Ras in tumour progression and metastasis. Biol. Chem. **386**:193–205.

Giglione, C. and A. Parmeggiani. 1998. Raf-1 Is Involved in the Regulation of the Interaction between Guanine Nucleotide Exchange Factor and Ha-Ras: EVIDENCES FOR A FUNCTION OF Raf-1 AND PHOSPHATIDYLINOSITOL 3-KINASE UPSTREAM TO Ras. J. Biol. Chem. **273**:34737-34744.

Hall, A. 1998. Rho GTPases and the actin cytoskeleton. Science **279**:509-514.

Jékely, G. 2003. Small GTPases and the evolution of the eukaryotic cell. BioEssays **25**:1129–1138.

Kinbara, K., L. E. Goldfinger, M. Hansen, F. L. Chou, and M. H. Ginsberg. 2003. Ras GTPases: integrins' friends or foes? Nat. Rev. Mol. Cell Biol. **4**:767-776.

Lewin, B. 2000. Genes VII. Oxford University Press, Oxford, UK.

Li, H.-Y., K. Cao, and Y. Zheng. 2003. Ran in the spindle checkpoint: a new function for a versatile GTPase. Trends in Cell Biology **13**:553-557.

Lounsbury, K. M., S. A. Richards, K. L. Carey, and I. G. Macara. 1996. Mutations within the Ran/TC4 GTPase. Effects on regulatory factor interactions and subcellular localization. J. Biol. Chem. **271:**32834-32841.

Majumdar, M., T. M. Seasholtz, D. Goldstein, P. de Lanerolle, and J. H. Brown. 1998. Requirement for Rho-mediated Myosin Light Chain Phosphorylation in Thrombin-stimulated Cell Rounding and Its Dissociation from Mitogenesis. J. Biol. Chem. **273**:10099-10106.

Nicely, N. I., J. Kosak, V. De Serrano, and C. Mattos. 2004. Crystal Structures of Ral-Gppnhp and Ral-Gdp Reveal Two Binding Sites that are Also Present in Ras and RAP. Structure **12**:2025-2036. PDB ID: 1U8Y, 1U8Z.

Olson, M. F., A. Ashworth, and A. Hall. 1995. An essential role for Rho, Rac, and Cdc42 GTPases in cell cycle progression through G1. Science **269**:1270-1272.

Pan, J. Y., W. E. Fieles, A. M. White, M. M. Egerton, and D. S. Silberstein. 2000. Ges, A human GTPase of the Rad/Gem/Kir family, promotes endothelial cell sprouting and cytoskeleton reorganization. J. Cell Biol. **149**:1107-1116.

Papagrigoriou, E., X. Yang, J. Elkins, F. E. Niesen, N. Burgess, E. Salah, O. Fedorov, L. J. Ball, F. von Delft, M. Sundstrom, A. Edwards, C. Arrowsmith, J. Weigelt, and D. Doyle.

2005. *unpublished.* PDB ID: 2ERX.

Pasqualato, S., L. Renault, and J. Cherfils. 2002. Arf, Arl, Arp and Sar proteins: a family of GTP-binding proteins with a structural device for 'front–back' communication. EMBO Reports **3**:1035–1041.

Piddini, E., J. A. Schmid, R. de Martin, C. G. Dotti. 2001. The Ras-like GTPase Gem is involved in cell shape remodelling and interacts with the novel kinesin-like protein KIF9. EMBO J. **20**:4076-4087.

Reuther, G. W., and C. J. Der. 2000. The Ras branch of small GTPases: Ras family members don't fall far from the tree. Current Opinion in Cell Biology **12**:157-165.

Ridley, A. J., and A. Hall. 1992. The small GTP-binding protein rho regulates the assembly of focal adhesions and actin stress fibers in response to growth factors. Cell **70**:389-399.

Saucedo, L. J., X. Gao, D. A. Chiarelli, L. Li, D. Pan, and B. A. Edgar. 2003. Rheb promotes cell growth as a component of the insulin/TOR signalling network. Nature Cell Biology **5**:566-571.

Schimmöller, F.. I. Simon, and S. R. Pfeffer. 1998. Rab GTPases, Directors of Vesicle Docking. J. Biol. Chem. **273**:22161-22164.

Sigrist, C. J. A., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief. Bioinform. **3**:265-274.

Sprang, S. R. 1997. G proteins, effectors and GAPs: structure and mechanism. Curr. Opin. Struct. Biol. **7**:849-56.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. 2006. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. **34**:D173-D180.

Wilde, A., and Y. Zheng. 1999. Stimulation of microtubule aster formation and spindle assembly by the small GTPase Ran. Science **284**:1359–1362.

Zeghouf, M., B. Guibert, J.-C. Zeeh, and J. Cherfils. 2005. Arf, Sec7 and Brefeldin A: a model towards the therapeutic inhibition of guanine nucleotide exchange factors. Biochemical Society Transactions **33**:1265-1268.

Chapter 1

Phylogenetic Analysis of the Ras Family of GTPases

by

Andrew E. Dellinger[1] and William R. Atchley[2]

[1]Bioinformatics Research Center and [2]Department of Genetics

North Carolina State University

Raleigh, NC 27695

Send editorial correspondence to Andrew E. Dellinger (aedellin@ncsu.edu)

Abstract

A comprehensive phylogenetic analysis of the Ras family is presented. A total of 476 Ras superfamily proteins including 98 Ras family proteins were used to analyze the phylogeny and summarize the evolutionary history of the Ras family. To establish the accuracy of the resultant tree, Neighbor-Joining, weighted Neighbor-Joining, Maximum Parsimony, Quartet Puzzling, Maximum Likelihood and Bayesian methods of tree building were used. All of these procedures gave nearly identical subfamily clades indicating a stable evolutionary classification among the Ras family proteins. The proteins in these various clades appear orthologous and functionally similar. Thus, tree-building algorithms have a potential role in assigning a subfamily and a biological function to unknown Ras family proteins. Reconstruction of ancestral sequences by parsimony and maximum likelihood revealed that the Ras family ancestral protein was most similar to the Rheb proteins in multicellular eukaryotic species.

Introduction

The Ras superfamily of proteins is composed of small GTPases that function as molecular switches. Ras superfamily proteins occur in eukaryotes but not in prokaryotes or Archaea (Jékely, 2003). Previous works have divided them into 6 families: Ras, Rho, Rab, Ran, Sar/Arf and RGK (Rad-Gem-Kir) (Oxford and Theodorescu, 2003; Crespo and León, 2000). Members of all families except RGK are found in even the most primitive modern eukaryotes, suggesting that the origin of these families occurred before the rise of modern-day eukaryotes.

Understanding the evolutionary history of the Ras superfamily is fundamental. Its divergence from the ancestral protein has led to cellular functions unique to eukaryotes, e.g., nuclear protein transport, vesicle trafficking, cell proliferation controlled by signaling, and control of the actin cytoskeleton (Garcia-Ranea and Valencia, 1998; Devos, et al. 2004). Ras superfamily proteins control essential cell functions such as differentiation, division, growth, morphology and adhesion (Wennerberg, et al. 2005). Some Ras family proteins are oncogenes involved in 20-30% of all human cancers (Isoldi, et al. 2005). Thus, studying the evolutionary history of the Ras superfamily can give insight into the rise of the eukaryotes and some of their unique cellular functions. It may lead to new insights into the mechanisms and treatment of Ras-related cancers.

The Ras superfamily proteins are small GTPases that bind GTP, hydrolyzing it into GDP and an inorganic phosphate (Sprang, 1997; Nixon, et al. 1995). They are inactive when bound to GDP and active when bound to GTP (Sprang, 1997). They often require GTPase activating proteins (GAPs) to speed hydrolysis of GTP and guanine exchange factors (GEFs) to stimulate release of GDP (Oxford and Theodorescu, 2003). Superfamily conserved sites occur exclusively in the 5 loop regions that are instrumental in coordinating Mg^{2+} and binding the guanine nucleotide (Sprang, 1997). Since these are the only Ras superfamily conserved amino acid sites (Sprang, 1997) at the sequence level, evolutionary conservation in the superfamily appears to be limited to GTPase activity alone. Similarly, at the biological level each member of the Ras superfamily plays a unique set of cellular roles, but all bind GTP/GDP and Mg^{2+}, act as molecular switches and change conformation according to their binding state (Sprang, 1997).

Structurally speaking, the core tertiary structure of the Ras superfamily protein is composed of a six stranded β-sheet, five α-helices, and ten loops (Nicely, et al. 2004). In the Ran family, the α1-helix is split into α1a and α1b (Sprang, 1997). Outside the core structure, Arf family proteins have an extra N-terminal helix and Ran-GDP proteins have an extra C-terminal helix (Sprang, 1997). Ras superfamily proteins have two switch regions which control binding activity and specificity. These switch regions change conformation according to the binding state of the protein (Crespo and León, 2000). Switch I binds effectors and GAPs (Crespo and León, 2000). Switch II binds GAPs and GEFs and stabilizes effector binding through its interactions with γ-phosphate of GTP (Crespo and León, 2000; Scheffzek, et al. 1997).

Within the Ras superfamily is a subgroup of proteins designated as the "Ras family" in the more restricted sense. These "Ras family proteins" control cell proliferation, growth, differentiation, adhesion and morphology as well as transcription and other cellular functions (Feig, 2003; Reuther and Der, 2000; Kinbara, 2003, Saucedo, et al. 2003). Members of the Ras family include the Rit, Ral, Rap, Ras, N-, K-, H-, M-, and R-Ras proteins. Rit controls cell differentiation and transformation (Hoshino, et al. 2005). Rap1 negatively regulates cell cycle and controls cell differentiation and adhesion (Stork, 2003). Rap2 putatively regulates Rho proteins (Myagmar, et al. 2005) and the actin cytoskeleton (Taira, et al. 2004) . It inhibits AMPA synaptic transmission (Zhu, et al. 2005). Ral proteins are activated by Ras subfamily proteins and possibly by Rap (Feig, 2003). Some functions of Ral include exocytosis and endocytosis, inhibition of neurotransmitter release, changes in the actin cytoskeleton and enhancement of cell proliferation and oncogenesis (Feig, 2003). The N-, K-,

H-, M- and R-Ras genes are oncogenes whose mutated form is present in 20-30% of all cancers (Isoldi, et al. 2005). Their proteins, along with other Ras family proteins, are posttranslationally geranylgeranylated, farnesylated and/or palmitoylated. This enables the proteins to attach to the plasma membrane, which is essential for their function (Reuther and Der, 2000; Sun, et al. 1998). Therefore, farnesyltransferase (FTase) and geranylgeranyltransferase (GGTase) inhibitors have been developed to fight cancer by preventing membrane attachment of these constitutively active proteins (Sebti and Hamilton, 2000).

Ras family proteins DexRas, Di-Ras and Rheb control cell growth. Saucedo et. al (Saucedo, et al. 2003) showed that Rheb promotes cell growth and accelerates the cell cycle from G1 to S without increasing the rate of cell division. Di-Ras1 suppressed growth of astrocytoma cells (Ellis, et al. 2002). Mice with Rhes (DexRas2) double knockout mutations weighed significantly less than wild-type mice (Spano, et al. 2004). This could be caused either by less cell division or less cell growth. Vargiu et. al (Vargiu, et al. 2004) showed that Rhes is not transforming. Thus, cell growth is more likely affected than cell division.

This is the first of a series of articles exploring the phylogenetic history of the Ras superfamily. Herein, we delve into the phylogeny of the Ras family proteins. The proteins analyzed here come from a wide variety of organisms ranging from humans to slime mold. Among the questions being explored include: 1) How did the Ras family ancestral protein evolve into many divergent present-day proteins? 2) How are the subfamilies of the Ras family interrelated? 3) What was the ancestral protein of the Ras family and how does it relate to other Ras family and superfamily proteins?

Materials and Methods

The Ras superfamily is composed of six families (Ras, Ran, Rho, Rab, Sar/Arf, and RGK) and all were included in the superfamily analysis. All 581 Ras superfamily sequences and all 119 Ras family sequences from Swissprot release 47 (Bairoch, et al. 2004) were aligned using CLUSTALX (Thompson, et al. 1997) with minimal improvement by eye. The analyses reported here focus on the Ras core that includes amino acid sites 5 to 164 in the protein core (Ras family numbering scheme as in Nicely, et al. 2004). Sequence fragments were deleted from the analysis. To accommodate software limitations of a maximum of 500 proteins, highly similar sequences were eliminated, leaving a total of 476 proteins. In eliminating proteins, an effort was made to minimize sample size bias among the families in the superfamily.

After a preliminary analysis elucidated the overall phylogenetic structure of the entire Ras superfamily (Figure 1.1), the Ras family was explored in a series of phylogenetic analyses. Duplicate Ras family proteins and proteins missing over 10% of the core were eliminated, leaving 98 proteins.

Tree Estimation Methods

Choosing a phylogenetic tree estimation procedure is sometimes controversial. As a consequence, we have used a variety of tree estimation algorithms to evaluate the robustness of the various approaches used for phylogenetic analyses. If the phylogenetic analyses using different approaches are all concordant across the different methods, then there is greater

confidence in the accuracy of the results and the biological conclusions one might draw from them.

Maximum parsimony (MP), neighbor-joining (NJ), weighted neighbor-joining (W-NJ), quartet puzzling (QP), maximum likelihood (ML) and Bayesian methods were used to estimate the Ras family phylogenetic tree. The assumptions and optimality criteria of these methods vary. For example, MP searches for trees with a minimal number of changes while NJ builds trees by minimizing distances between pairs or groups of sequences. W-NJ improves NJ by increasing the accuracy of long branch placement. QP and ML define the most likely tree given the data and the evolutionary model. The Bayesian method uses prior distributions on tree-building and evolutionary parameters and samples the posterior probability distribution using Markov-Chain Monte Carlo (MCMC) until the chains' likelihoods converge (Ronquist and Huelsenbeck, 2003). Discussions of these methods are summarized in (Felsenstein, 2004; Bruno, et al. 2000) for weighted neighbor-joining, (Felsenstein, 2004; Schmidt, et al. 2002) for quartet puzzling, (Ronquist and Huelsenbeck, 2003; Felsenstein, 2004) for Bayesian analysis and (Felsenstein, 2004; Mount, 2001) for parsimony, neighbor-joining and maximum likelihood.

Ras Superfamily Tree:

Because of the large number of proteins analyzed, a somewhat superficial Ras superfamily tree was estimated using only the Neighbor-Joining method built into the MEGA 3.1 (Kumar, 2004) package. The options for the Neighbor-Joining model used in this study were pairwise deletion for dealing with gaps and Poisson distance correction for multiple

hits. 1000 non-parametric bootstrap replications were used to determine the statistical confidence in the tree topology. This model uses gamma distributed rates of evolution among sites where $\alpha = 1$.

To estimate the true value of α, Ziheng Yang's program PAML (Phylogenetic Analysis of Maximum Likelihood) was used with this preliminary NJ tree and the multiple sequence alignment as input data (Yang, 1997). PAML estimates branch lengths and ancestral states and builds a phylogenetic tree with the maximum likelihood given the data. PAML estimated α to be 0.83452 by maximizing the likelihood of the value given the alignment and tree. Eight rate categories were used. The Neighbor-Joining topology was calculated with $\alpha = 0.83452$ and the branch lengths were recalculated using maximum likelihood in PAML (Yang, 1997). The resulting tree is shown in Figure 1.1.

Ras Family Trees:

The more restricted Ras family trees were estimated using 98 Swissprot Ras family proteins with G-proteins as an outgroup. Ras superfamily proteins are called small G-proteins because they are related to the prokaryotic G-proteins (Takai, 2001; Sprang, 1997). G-proteins, like Ras proteins, are GTPases. They share a core structure and conserved GTP-binding sites (Sprang, 1997). Thus, G-proteins were used as an outgroup for phylogenetic calculations of the Ras family proteins. Six G-proteins from Swissprot (Bairoch, et al. 2004) including Elongation factor Tu and G, translation initiation factor 2 and peptide chain release factors were used when the software allowed, otherwise only Elongation factor Tu (Swissprot id: EFTU_STRPY) was used.

Two NJ trees were estimated using the Ras superfamily tree-building algorithm, one using the Poisson model (equal amino acid frequencies and correction for multiple substitutions at a site) and another using the JTT substitution matrix (Jones, et al. 1992).

Next, parsimony trees were estimated using PAUP* 4.0 beta 10 for Windows (Swofford, 1998). A heuristic search for the optimal tree was performed. Ten random-addition sequence replications were performed to determine the score of the most parsimonious tree. Each tree underwent tree bisection-reconnection to find the optimal topology. Tree bisection-reconnection is a method in which an interior branch is broken and the two resulting subtrees are reconnected at every possible branch. This process is repeated for each interior branch (Felsenstein, 2004). Branches with less than 50% support were collapsed. This heuristic search was iterated until the parsimony score decreased to a minimum and that most parsimonious score was obtained over three iterations.

Weighted neighbor-joining trees were generated using Weighbor (Bruno, et al. 2000), which downweights large distances so this method doesn't suffer from long branch attraction problems, as found with parsimony (Bruno, et al. 2000). Pairs of sequences are chosen to be joined by maximizing the likelihood that the resulting branches are both additive and positive. Because of both of these changes in the neighbor-joining algorithm, the resulting trees are thought to be more frequently correct in the presence of distant taxa compared to results from Neighbor Joining and parsimony methods (Bruno, et al. 2000). SEQBOOT from PHYLIP 3.5 (Felsenstein, 1989) was used to generate 1000 datasets from the multiple alignment. The distance matrices for these datasets were computed using PROTDIST in PHYLIP 3.5 (Felsenstein, 1989). The Jones-Thornton-Taylor (JTT) substitution matrix

(Jones, et al. 1992) was used with a coefficient of variation ($1/\sqrt{(\alpha)}$) among sites of 0.943, where alpha was calculated to be 1.1251 in PAML using 8 categories of rate variation. Weighbor was run on the 1000 datasets. Because of unequal amino acid frequencies, an effective alphabet size of 14 was used as recommended by Swofford and Olsen in *Molecular Systematics* (1990) (Hillis, et al. 1990). There were 276 varying sites in the alignment. A consensus tree was generated using PAUP*.

Three different maximum likelihood trees were generated using PHYLIP's PROML (Felsenstein, 1989) program with a JTT substitution matrix (Jones, et al. 1992), 8 variable + 1 invariant gamma distributed rate categories, randomized sequence addition, a coefficient of variation of 0.943, and the G-protein with Swissprot id EFTU_STRPY as an outgroup. The problem was of sufficient complexity that insufficient computational and time resources were available to estimate ML trees by other methods. Quartet Puzzling trees were generated using Tree-Puzzle 5.2 (Schmidt, et al. 2002). The JTT substitution matrix (Jones, et al. 1992) and 8 variable + 1 invariant categories of gamma distributed rates were used.

Bayesian-derived trees were generated using MrBayes (Ronquist and Huelsenbeck, 2003; Huelsenbeck, et al. 2001). For computational expediency, sequences with small differences (0.006 substitutions per site) were deleted, leaving 92 Ras family sequences and 6 outgroup sequences. Priors were set using the JTT substitution matrix (Jones, et al. 1992) and 4 categories of gamma-distributed site rates. All other settings were of their default value. The analysis was terminated after 2,050,000 generations, at which point the average standard deviation of split frequencies had fluctuated between 0.0243 and 0.025 for 250,000 generations.

Ancestral Reconstruction

Reconstruction of ancestral states was performed using both maximum likelihood and parsimony methods. PAML (Yang, 1997) was used to perform a marginal reconstruction of ancestral states via the maximum likelihood method. Parsimony derived ancestral states were derived simultaneously with the generation of parsimony trees in PAUP* (Swofford, 1998).

Some sites have more than one most parsimonious state. Consequently, to run BLAST (Altschul et al. 1990) on the reconstructed sequence, one residue was chosen for each site. The selected residue was the one with the greatest probability of being substituted for the residue in the same site of the maximum likelihood reconstruction according to the JTT substitution matrix (Jones, et al. 1992). Each reconstructed Ras family ancestor was input into BLAST (Altschul, et al. 1990) using the nr database (Benson, 2005) and the BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992) to retrieve the match with the smallest e-value.

Results

Ras Superfamily:

The unrooted radiation tree in Figure 1.1 provides an overview of the evolution of the Ras superfamily. A radiation tree is an unrooted phylogenetic tree where the major clades radiate from a central point. Each family is distinct and no overlap occurs from members of other families. The only exception is the Arf family, which includes the Sar family as a

subset.

Even with this unrooted radiation tree, there is important information about the evolution of the families of the Ras superfamily. Figure 1.1 shows the Sar family evolved from ancestral Arf proteins. The Sar family shares an ancestor with the Arf1 protein from the fungus *Encephalitozoon cuniculi*. The Ran family ancestor is related to Rab-like proteins 2A and 2B from humans. This tree suggests that several subfamilies evolved early in the history of the Ras family. The Rheb subfamily ancestor arose first, followed by the ancestor of the RGK family and the Di-Ras subfamily. The ancestor of the other subfamilies, including the Ras subfamily, appeared later.

Ras Family:

The Ras family trees were estimated using 98 Swissprot Ras family proteins plus a G-protein outgroup. Ras superfamily proteins are called small G-proteins because they are related to the prokaryotic G-proteins (Takai, 2001; Sprang, 1997). G-proteins, like Ras proteins, are GTPases. They share a core structure and conserved GTP-binding sites (Sprang, 1997). Thus, G-proteins were used as an outgroup for phylogenetic calculations of the Ras family proteins. Six G-proteins from Swissprot including elongation factors (Swissprot ids EFTU_STRPY, TETM_UREUR and EFG_STRP6) , translation initiation factor 2 (IF2G_PYRKO) and peptide chain release factors (ERF2_CANAL and RF3_STRP3) were used when the software allowed, otherwise only Elongation factor Tu (EFTU_STRPY) was used.

The clades of the trees in Figure 1.2 and Appendix Figures A.1-A.6 are

predominantly subfamilies of the Ras family. Note that the "fungal Ras" and "distant Ras" clades are not subfamilies but groups of organisms containing Ras subfamily proteins. With the exception of the H,N,K-Ras clade and the *C. elegans* Rheb protein, no subfamily protein is outside of a clade. No subfamily clade contains a protein known to belong to another subfamily. These facts indicate the stability and biological relevance of the clade structure. Table 1.1 shows that all the methods yield nearly equivalent topologies. However, there is variability among the clades of the trees.

First, the trees vary in their estimation of the evolutionary history at the level of the subfamily clade. The W-NJ and QP trees (Figures A.3 and A.5) provide no information about the evolution of the subfamilies at the 50% bootstrap threshold. All of the subfamilies come from the same multifurcating branch. At this same 50% bootstrap threshold, the NJ tree divides the family into the cell division promoting subfamilies (Rit, Ral, and Ras) and the cell growth controlling protein subfamilies (Rheb, DexRas and Di-Ras) together with the Rap subfamilies, and the Bayesian tree divides the family into: Rheb; DexRas, Di-Ras, and Rap; Rit, Ral, and Ras. The ML and MP trees are bifurcating and yield unique subfamily histories.

The bootstrap threshold must be less than the 30% used for the NJ and W-NJ trees and less than the 50% used for the QP and Bayesian trees for a more detailed reconstruction of the evolutionary history of the Ras family to be achieved. Atchley, et al. 2001 (44) showed a similar result where 36% and 40% of PSD cliques and intron/exon groups, respectively, had bootstrap values of less than 75% and required a lower cutoff for the trees to reveal the evolutionary history of the serpin proteins.

Second, the subfamily proteins may be present in anywhere from one to four clades

(Table 1.1). For example, the H,N,K-Ras subfamily proteins are in four clades in the NJ, W-NJ and QP trees, two clades in the ML and Bayesian trees and one clade in the MP tree. The M,R-Ras, Rheb, "fungal Ras" and "distant Ras" proteins are also present in a varying number of clades.

Third, the clades have varying compositions. Some subfamily members are not present in a clade. The Rheb protein from the worm *Caenorhabditis elegans* was only present in the Rheb clade in the ML tree. Some K-Ras proteins in the QP and W-NJ trees are not present in the H,N,K-Ras subfamily clades because they failed to reach the 50% bootstrap threshold. Also, the H-Ras protein in *Gallus gallus* (chicken) is missing from the subfamily clades of the ML tree. The *Schizosaccharomyces pombe* Ras protein and *Neurospora crassa* Ras2 protein are the most frequent fungal Ras proteins not present in a clade. The "distant Ras" clade does not contain the same number of proteins in any two of the trees**.**

H,N,K-Ras and M,R-Ras subfamily clades vary in the composition of their members that are not confirmed subfamily proteins as described in Table 1.1. Ras2 proteins in the fruit fly *Drosophila melanogaster* and the hydra *Hydra magnipapillata* are in the M,R-Ras subfamily clade in all except the W-NJ tree where only the *D. melanogaster* protein is present in the clade. The Ras proteins of the fish *Carassius auratus* and *Limanda limanda* are in the H,N,K-Ras clades in all but the ML tree.

The previous method for building Neighbor-Joining trees was used to estimate the phylogenetic tree of the switch regions from residues 30-38 in Switch I and residues 60-76 in Switch II (Ras family numbering) (Nicely et al. 2004; Milburn et al. 1990). The tree is shown in Figure A.6. These sites were sufficient to yield the same subfamily clades produced by the

full alignment in each of the methods, allowing for a more lenient bootstrap threshold.

The Bayesian tree results (Figure 1.2) are included here as indicative of the results produced by all the methods. This decision was made due to the fact that, all other things being equal, the Bayesian method has the advantage that it accounts for multiple substitution and rate variation among sites and doesn't suffer from errors due to long branch attraction like the NJ method (Bruno, et al. 2000). The latter is important since the Ras family tree contains some longer branches that can potentially cause problems. The results using ML (Figure A.6) are dependent on what order the sequences were added to the algorithm (data not shown). They are not consistent in the order of subfamily evolution suggesting they are not reliable estimations of phylogeny. There is certainly no consensus as to which phylogenetic method is best, but the Bayesian method is an appropriate representative in this venue.

In the Bayesian tree (Figure 1.2), the Rheb proteins are the nearest to the root, as in the MP tree. They are followed by the Rap, DexRas and Di-Ras subfamilies. The NJ tree supports this at a bootstrap threshold of 30%. In Genbank there is a Rheb-like protein from the slime mold *Dictyostelium discoideum* (Benson, 2005). Since *D. discoideum* is one of the most primitive eukaryotes, this is evidence that the placement of the Rheb subfamily near the base of the tree is reasonable.

Reconstruction of the Ras Family Ancestor

The ancestor of the Ras family proteins was reconstructed using both parsimony and maximum likelihood methods. The extent of congruence of these two estimates can be seen

in the aligned reconstructions shown in Table 1.2. The maximum likelihood ancestor was derived from the Neighbor-Joining tree, while the parsimony ancestor was derived from one of the most parsimonious trees. In the case where a site had multiple possible residues, the residue closest physiochemically to the ML ancestral residue in that site was chosen.

BLAST (Altschul, et al. 1990) searches were performed on these reconstructed ancestral proteins in order to reveal the extant proteins that are most similar to them. The similarity criterion is based on the extant protein's e-value, which is computed in part using the BLOSUM62 amino acid substitution matrix (Henikoff and Henikoff, 1992). By matching the Ras family ancestor to the most similar extant protein, we gain information about the possible function and structure of the ancestral protein.

A BLAST search of the parsimony-derived Ras family ancestor against the NR database resulted in a hypothetical protein of the zebrafish *D. rerio* as the match with the smallest e-value (9e-61). The next 12 best matches are known Rheb proteins, suggesting that the hypothetical protein is also a Rheb protein. The switch regions of the parsimony ancestor are more similar to the Rheb from the yeast *Saccharomyces cerevisiae*, however. A BLAST search of the maximum likelihood derived family ancestor resulted in the Rheb protein in *D. melanogaster* as the match with the smallest e-value (2e-59). This protein is from the well annotated database Swissprot (Bairoch et al. 2004), so the protein's identity as a Rheb is more assured than that of the *D. rerio* protein. The switch regions of the ML ancestor more closely match the *C. elegans* and *S. cerevisiae* Rheb proteins than the *D. melanogaster* protein.

Rheb is the subfamily nearest the outgroup in the Bayesian, NJ and MP trees. Rheb

proteins, like the Di-Ras and DexRas proteins that adjoin them in the phylogenetic tree, promote cell growth- a different function from the cell-division controlling Rap1, Ras, Ral, Rit, H-Ras, N-Ras, K-Ras, M-Ras and R-Ras proteins that comprise the bulk of the Ras family members.

Both BLAST searches of the ML ancestor and of RHEB_DROME, its closest match, gave matches in all subfamilies of the Ras family. The ML ancestor gave 56 more Ras family results with e-values smaller than that of the first non-Ras family protein. Thus, by this method the ML ancestor is only a marginally better predictor of Ras family proteins than an extant Rheb protein. Previous authors (Atchley and Fernandes, 2005) found that the estimated ancestral sequence is very similar to the computed "sequence signature" in basic helix-loop-helix (bHLH) proteins (Atchley and Fernandes, 2005). The Ras family of proteins, however, is much less conserved than the bHLH proteins in the Atchley study and cannot be separated from the other Ras superfamily proteins using a "sequence signature" with a single amino acid at each site.

Discussion

Ras superfamily

Despite being an unrooted superfamily tree, Figure 1.1 gives indications of the order in which families arose: Arf before Sar, Rab before Ran, RGK before Ras, and Ras and Rho from a common ancestor.

Ran shares an ancestor with two Rab-like proteins in Figure 1.1, suggesting that the

ancestor of the Ran family is an ancestral Rab protein. This is reasonable as long as either ancestral Rab proteins could import and export proteins through the nuclear membrane or there was a different nuclear structure in the first eukaryotes that allowed proteins to pass by some other mechanism. Mans, et al. (2004) showed that some nuclear envelope proteins evolved at different points in eukaryotic evolution and others arose from existing eukaryotic proteins. This leads to the more likely inference that the latter premise, that there was a nuclear structure allowing protein import and export before the origin of Ran.

The superfamily tree in Figure 1.1 suggests that the Sar family evolved from an Arf family ancestral protein. Both families have representatives throughout extant eukaryotes. There are more Arf proteins than Sar proteins per organism. Humans have just 2 Sar family proteins, Sar1a and Sar1b, while we have at least 6 Arf and 8 Arf-like proteins. So either Arf family proteins evolved first and had more time to accept gene duplications or they evolved at a similar time or later than the Sar family proteins and have less pressure from purifying selection than Sar family proteins.

It is reasonable to postulate a common Rho/Ras family ancestor. This common ancestry is supported not only at the sequence level, but also at the functional level. (Kozminski, et al. 2003; Nobes and Hall, 1995; Sahai, et al. 2001; Olson, et al. 1998; Danen, et al. 2000) Rho and Ras family proteins seem to have coevolved to interact with each other in several ways. Kozminski, et al. 2003 showed that a Ras family member, RSR1, and a Rho family member, CDC42, interact to enable polarized cell growth in yeast. Ras and Rho family proteins can interact hierarchically as well as simultaneously. Oncogenic Ras induces cytoskeletal changes. Since Rho proteins rearrange the cytoskeleton, this indicates that Ras

family proteins can activate Rho family proteins (Nobes and Hall, 1995). Ras and Rho have also been shown to play opposite roles in the control of the p21 cyclin-dependent kinase inhibitor. Finally, Ras induces and Rho inhibits p21 transcription (Sahai, et al. 2001; Olson, et al. 1998; Danen, et al. 2000).

The superfamily proteins are present in modern-day eukaryotes from humans to unicellular eukaryotes (Bairoch, et al. 2004). Even the primitive eukaryote *D. discoideum*, a slime mold, has members in all families but the RGK family (data not shown). The clearly divided family clades in Figure 1.1 not only tell us that the proteins are correctly assigned to their families but also that there were clearly established family ancestral proteins before the modern eukaryotes came into existence. If the ancestral proteins were established later, primitive eukaryotes would be missing proteins of some of the families of the Ras superfamily. This is evidence for the evolution of the Ras superfamily ancestor occurring between the prokaryotic/eukaryotic split (~2500 Mya) (Gu, 1997) and the origin of extant eukaryotes (1500+ Mya) (Feng, et al. 1997).

Ras family

The phylogenetic methods used in this study embrace a broad range of assumptions and techniques. Selection of one method over another is often controversial; however these various methods produced subfamily clades that were stable with regard to protein composition. This suggests that the clades themselves and the overall results of the phylogenetic methods are stable in this case.

In fact, the clades are stable even with much less data in the alignment. Figure A.6

was estimated using the Switch I and Switch II regions of the Ras family proteins. These sites were sufficient to yield seven of eight of the subfamily clades and a proportion of the fungal and distant Ras clades given a bootstrap threshold of 30%. Thus, we conclude that the switch regions have a strong influence on Ras family phylogeny. It is because of this influence that the Ras family phylogenetic trees in this study are remarkably consistent in composition of their clades. Switches bind effectors and subfamilies bind similar effectors. Ras proteins that bind similar effectors have highly conserved sequences in their switches. Therefore, subfamily proteins tend to form clades.

The clade composition of the N-, H-, K-Ras and M-, R-Ras subfamilies has biological relevance. The switch regions, which determine the biological function of Ras family proteins, give evidence that even the proteins of unknown function in these clades belong to the proper subfamily. The Ras family proteins RAS_LIMLI, RAS_CARAU, RAS_ARTSA, LET60_CAEEL, and RAS1_DROME are present in various combinations in the H,N,K-Ras clade. To our knowledge, their function and proper subfamily assignment are unknown. The sequence of their switch regions are exact matches to the H-Ras, N-Ras and K-Ras proteins except for two viral H-Ras proteins. The proteins RAS2_DROME and RAS2_HYDMA are present in the M- and R-Ras clade. The RAS2_DROME switch sequences have two sites with residues in the same functional group as M-Ras and R-Ras proteins. The other residues are identical. Sites 62-64 in Switch II of RAS2_HYDMA are Glu-Glu-Phe, a sequence unique to M-Ras and R-Ras proteins. So all these proteins contain switch regions which match the subfamily clade to which they belong.

The difference between the sequences of these unknown proteins and their putative

subfamilies causes the variation in the clade membership of the N-, H-, K-Ras and M-, R-Ras subfamilies. These differences occur in the loops and α-helices on the side of the protein opposite that of the switch regions. The same is true for the *S. pombe* protein RAS_SCHPO, which is often missing from the fungal Ras clade. The biological reason for the variation in clade membership is thus nonfunctional but may be structural and environmental because some of the sites cluster together in the tertiary structure and most of the sites are solvent exposed. The Rheb protein from the worm *C. elegans* (RASL_CAEEL) and the Ras protein from the fungus *N. crassa* (RAS2_NEUCR) do not often belong to a clade because the sequence in their switches, and thus their function, differs from the rest of their subfamily. Their differing function most likely necessitated sequence changes to compensate structurally and physiochemically and thus increased their branch lengths, which also influenced their clade membership. Of course, ancestry and the different ways the phylogenetic tree-building methods treat sequence differences also contribute to variation.

The main source of variability among the phylogenetic trees is in how they treat the evolutionary history of the subfamilies. For example, the trees disagree on an evolutionary history for the Rheb/DexRas/Di-Ras group, the Rap1/Rap2/RSR1 group, and the Ras/Ral/Rit group. The NJ, MP, and Bayesian trees do agree that the cell-division promoting proteins arose after the cell-growth promoting proteins and the inhibitory proteins. Only the ML tree disagrees and asserts that the division proteins evolved first. Essentially, the conclusions that can be drawn about the evolutionary history of the subfamilies are this: there are three groups whose subfamily history is indistinguishable and these three groups evolved in the following order: Rheb group first, Rap group second, Ras group last.

Perhaps the confusion among the methods derives from the fact that there is only one member of each group in the oldest extant eukaryotes. *D. discoideum* and the fungi, including the completely sequenced fungi *S. cerevisiae*, *Encephalitozoon cuniculi,* and *Cryptococcus neoformans var. neoformans JEC21* genomes only contain proteins from the Rheb, Ras, and Rap1 subfamilies of the Ras family. Therefore, there should also be three pre-extant eukaryotic Ras proteins and three points of origin in the Ras tree. The tree is thus a combination of three unrooted trees. Since unrooted trees are directionless, the variation in clade order within groups is not unexpected.

It is possible that the three ancestral proteins existed in the prokaryotes. So a BLAST (Altschul, et al. 1990) search using the *D. discoideum* Rap, Rheb and Ras proteins was performed to determine the most related prokaryotic protein. The best matches are different proteins, but they are all GTPases that are closest to the Rab protein family instead of the Ras family, meaning that they are not direct relatives of the Ras family. This gives us some evidence that the Rab family may have evolved before the Ras family and that the direct ancestors of these proteins were not prokaryotic but pre-extant Ras superfamily proteins in eukaryotes.

The presence of three subfamilies in the oldest extant eukaryotes leads one to conclude that: there has been considerable Ras evolution since the extant fungal taxa arose and both cell growth and cell division control is part of the Ras family function in all extant eukaryotes. We hypothesize that both cell-growth and cell-division control were in fact performed by Ras family proteins before the extant eukaryotes arose. Also, since there has been considerable evolution in the Ras family, we should be able to gain a greater

understanding of the evolutionary history of the three Ras family subgroups (Rap and RSR; Rheb, DexRas and Di-Ras; Ras, Ral, Rit, and H-, N-, K- , M-, and R-Ras).

The ancestral protein in the node nearest the outgroup in the NJ and MP trees was reconstructed. The resulting proteins were most similar to Rheb proteins. These estimated Ras family ancestors can be synthesized in the lab and their structure and function determined. Chang, 2003 reconstructed an ancestral rhodopsin protein, synthesized in and compared its function to the extant rhodopsin proteins. Other estimated ancestral proteins have been synthesized and tested (Chang et al. 2005; Thornton et al. 2003; Jermann ***et al.*** 1995). It would be of interest to synthesize one or both of the reconstructed Ras family ancestors to determine their structure and function. Successful experimentation on these proteins would give us more insight into the evolution of structure and function in the Ras family.

There seems to be no discernible pattern to the sites in which the MP and ML reconstructions of the Ras family ancestor disagree. Sites 69-71 in Switch II and sites 102-103 face outward and are in close proximity to the RasGEF SOS but the side chains of many sites face inward (data not shown). The sites that vary are mainly in α-helices, though several sites are in loops and the β-sheet (data not shown). Neither method decides consistently on the residue of a particular extant Rheb protein or the extant proteins in adjacent clades. The primary source of variation between the reconstructions is probably the methods' interpretation of the data rather than strictly biological phenomena.

The Ras superfamily tree in Figure 1.1 also shows that the Rheb subfamily was the first Ras family member to evolve, which agrees with the Ras family Bayesian, NJ and MP

trees. According to this information, if a single prokaryotic G-protein ancestor does exist for the Ras family, it is likely that it first evolved into the ancestral Rheb protein.

Since the clades of the Ras family phylogenetic trees are so consistent across tree-estimating methods and since the clades represent subfamily and function, tree estimation can be used to identify unknown Ras family proteins. Unknown Ras family proteins are added to existing data and the tree is estimated. The clades that unknown proteins are assigned to identify what subfamily the protein belongs to and give a reasonable hypothesis as to its function. Of course, some mutants and pseudogenes will slip through, but this method provides a good way to direct experiments for the purpose of determining the biological function of unknown Ras superfamily proteins.

There are crystal structures for the cell-division controlling proteins (H-Ras, M-Ras, R-Ras2, and Ral), cell-growth controlling proteins (Rheb, Di-Ras2), and the inhibitory Rap proteins (Rap1a and Rap2a). These structures cover the major functions of the proteins in the phylogenetic tree. Incorporating structural and phylogenetic data can be used to model the relationships between evolution of structure and the evolution of sequence and function. The pattern of subfamily clade evolution, and thus evolution of function, can be matched to the pattern of structural evolution using ancestral reconstructions of both sequence and structure. Sites that vary in their three-dimensional location from structure to structure can be matched to functional residues that vary from subfamily to subfamily. And patterns of site change among clades can be matched to patterns of structural change using common factor analysis.

Conclusions

Six models of phylogenetic estimation using different methods and assumptions divided the Ras family into the same clades (functional groups). Ten of eleven functional groups have little or no variation in their membership. This gives confidence in the accuracy of the clades, which divide the Ras family by subfamily and thus by function. The phylogenetic tree was used to reconstruct the Ras family ancestral sequence, which best matches the Rheb subfamily. This is consistent with the topologies of the Ras superfamily and family trees where Rheb is the first subfamily of the Ras family to evolve. The robust topology of the Ras family phylogenetic tree estimated in this study along with the sequence and functional information it contains can be a powerful tool for the study of the origin and evolution of the Ras family.

We thank Heather Dellinger for her encouragement during the development of this study. This work was supported by a NSF Genomics IGERT fellowship, the NCSU Functional Genomics fellowship and WRA's NIH grant.

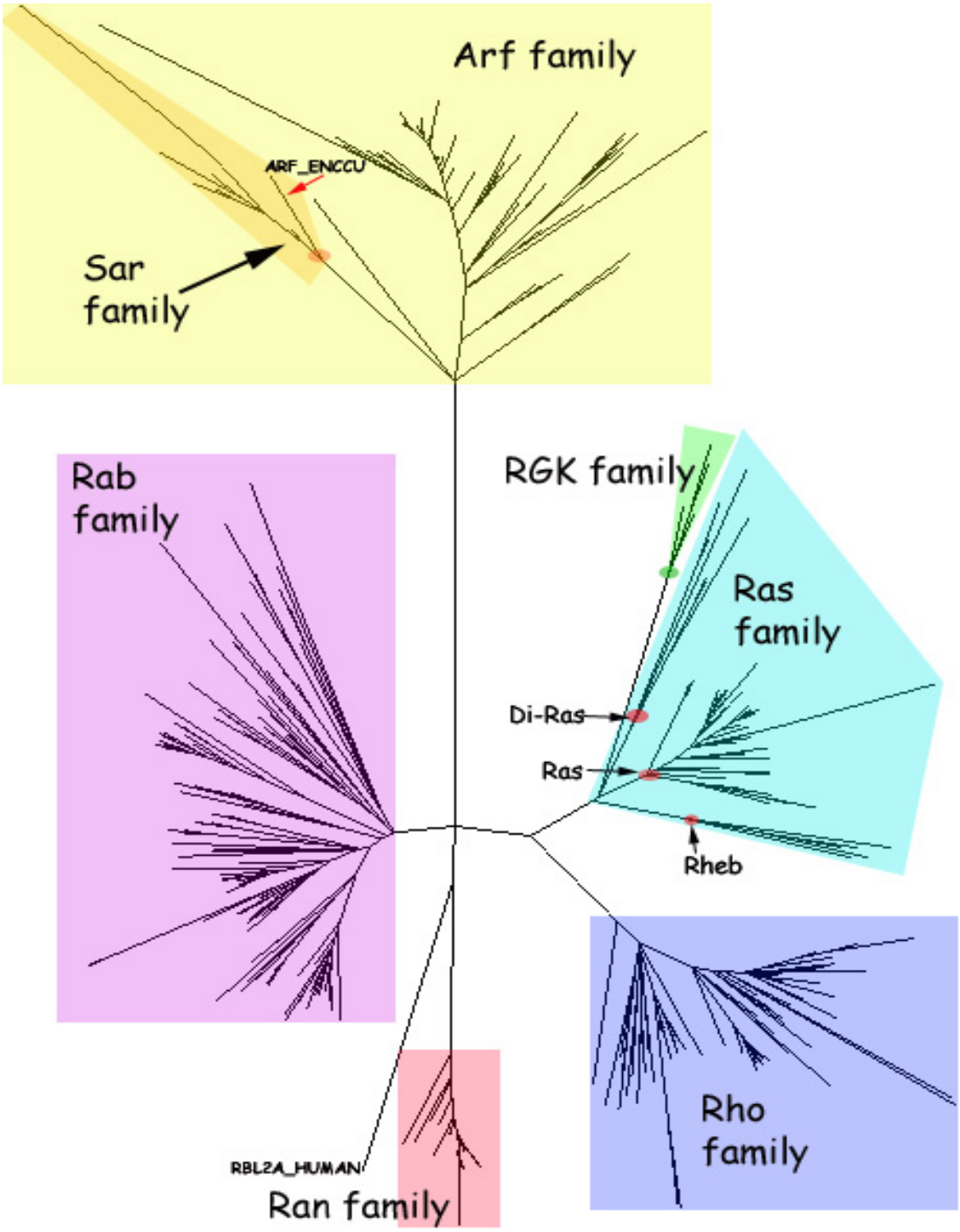


Figure 1.1 Radiation tree of the Ras superfamily.
Branch ordering of Figures 2-8 was determined by bootstrap consensus. Families are denoted by shaded shapes. Ellipses denote ancestral family or subfamily sequences. The Ras family within the light blue diamond is the major focus of this paper.

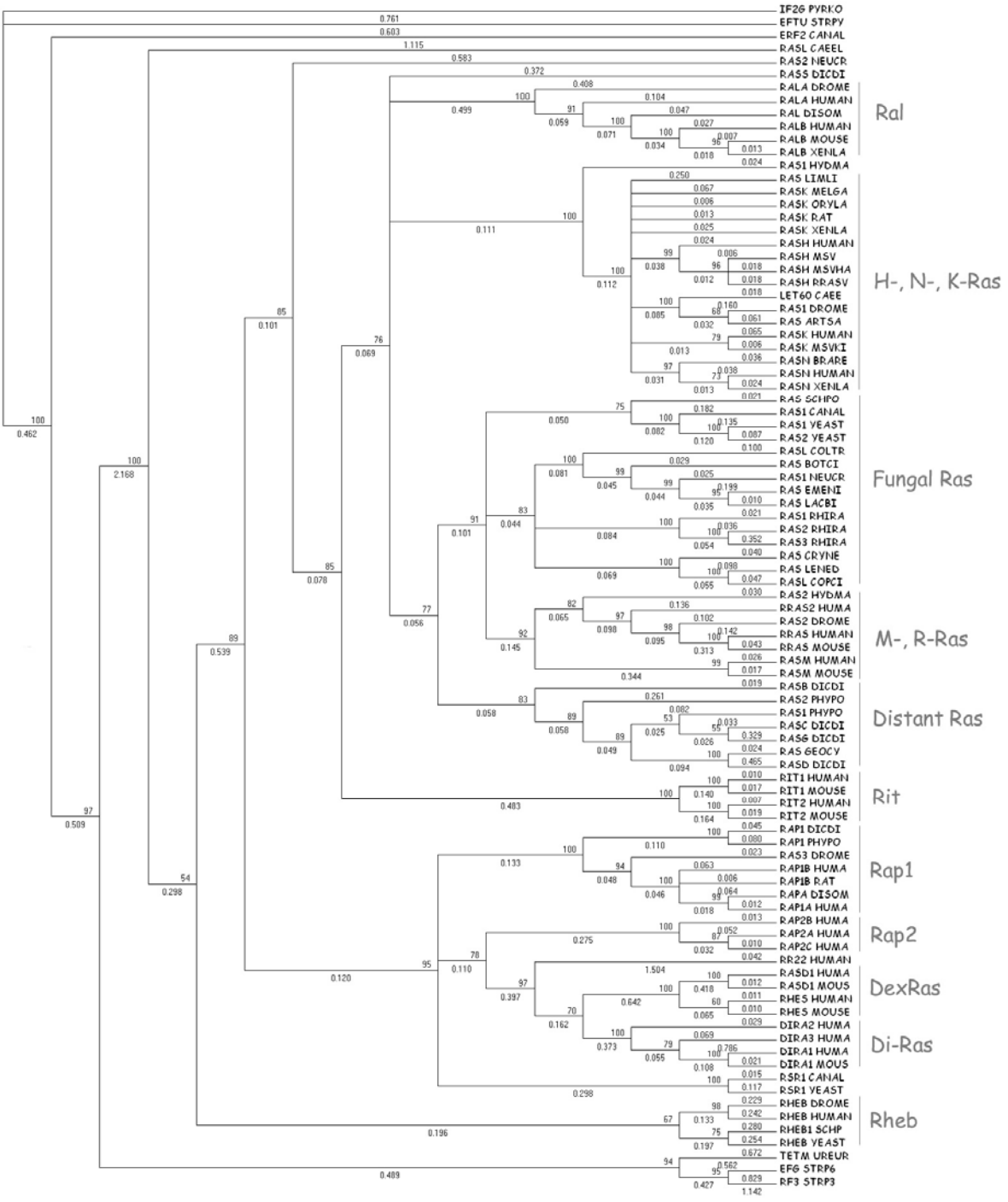
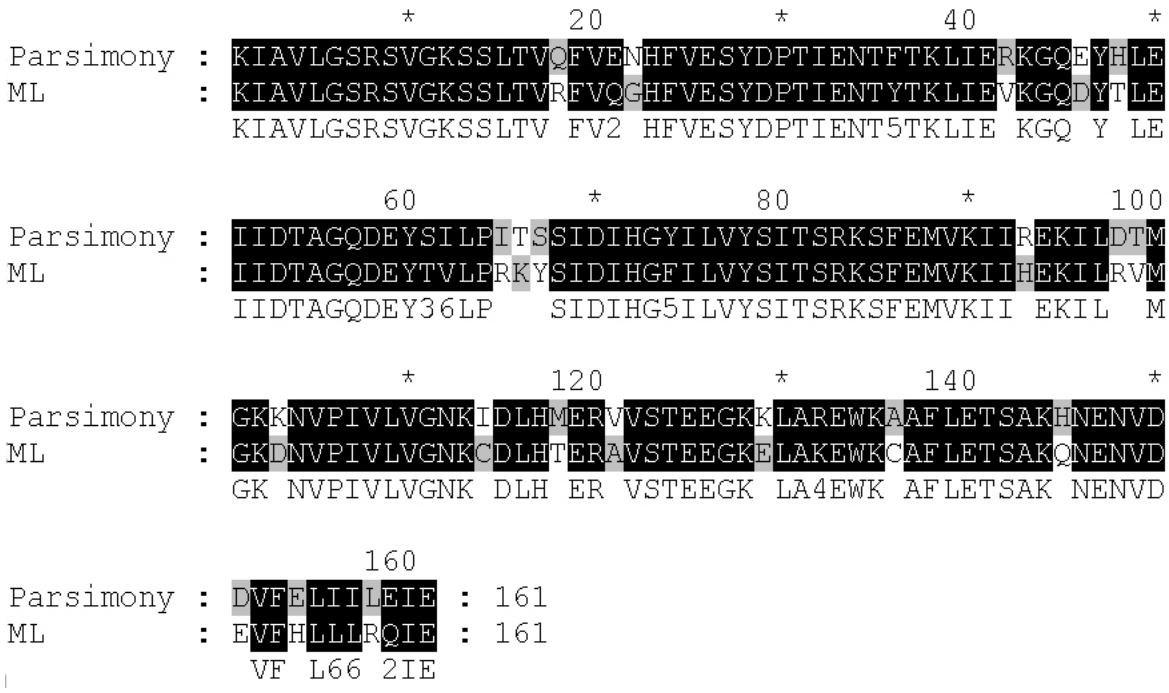


Figure 1.2. Ras Family Phylogeny Constructed by Bayesian Statistics.
Clades are labeled by subfamily. The Ras subfamily is divided by organism into "Distant Ras" and "Fungal Ras". Branches with less than 30% support have been collapsed. Bootstrap values are adjacent to the internal nodes and branch lengths are on the branches.

Table 1.1 Repeatability of Clades among Phylogenetic Methods
The data used in this table are from the Appendix figures A.1-A.5. Each major subfamily has a clade in each of the phylogenetic trees calculated in this study. The fraction of subfamily proteins in the subfamily clade(s) are shown. The bootstrap value for clades above the 50% threshold is shown in brackets. The maximum likelihood tree was not bootstrapped. The Bayesian tree contains fewer proteins so subfamilies may have a smaller denominator than in the other trees. Notes: [1]RASL_CAEEL is a Rheb subfamily protein. [2]RAS3_DROME is a Rap1 protein. RSR1_CANAL and RSR1_YEAST are Rap1 homologs but are not included as members of the Rap1 clade. Some clades have additional proteins that are not confirmed as subfamily proteins. The additional proteins are: [3]RAS_LIMLI and RAS_CARAU, [4]RAS2_DROME and RAS2_HYDMA, [5]RSR1_YEAST, RSR1_CANAL (Rap1 homologs) [6]RAS_CARAU, RAS_LIMLI, RAS_ARTSA, RAS1_DROME and LET60_CAEEL, [7]RAS2_DROME, [8]RAS_LIMLI, RAS_ARTSA, RAS1_DROME and LET60_CAEEL.

Clade	**NJ-Poisson**	**MP**	**QP**	**ML**	**W-NJ**	**Bayesian**
Rheb[1]	5/6 [67]	5/6 [69]	5/6 [58,76] (2 clades)	6/6	5/6 [50]	4/5 [67]
DexRas	4/4 [92]	4/4 [100]	4/4 [85]	4/4	4/4 [78]	4/4 [100]
Di-Ras	7/7 [96]	7/7 [100]	7/7 [69]	7/7	7/7 [84]	7/7 [100]
Rap1[2]	7/7 [96]	7/7 [100]	7/7[5] [53]	7/7	7/7 [88]	7/7 [100]
Rap2	3/3 [98]	3/3 [100]	3/3 [95]	3/3	3/3 [93]	3/3 [100]
Ral	8/8 [97]	8/8 [100]	8/8 [79]	8/8	8/8 [80]	6/6 [100]
Rit	4/4 [99]	4/4 [100]	4/4 [90]	4/4	4/4 [97]	4/4 [100]
H-Ras, N-Ras, K-Ras	19/19[3] [52]	19/19[3] [100]	14/19[3] [82,78,77,57] (4 clades)	18/19[6] (2 clades)	18/19[3] [80,82] (2 clades)	13/13[8] [100]
M-Ras, R-Ras	5/5[4] [70]	5/5[4] [100]	5/5[4] [67]	5/5[4]	5/5[7] [53,94,96] (3 clades)	5/5[4] [92]
Fungal Ras	12/16 [92,86,95,96] (4 clades)	15/16 [52]	13/16 [94,70,65,53] (4 clades)	14/16 (2 clades)	12/16 [91,81,77,85] (4 clades)	15/16 [75,83] (2 clades)
Distant Ras	5/9 [85]	6/9 [69]	2/9 [72]	7/9	4/9 [87,89] (2 clades)	8/9 [83]

Table 1.2. Comparison of the estimated ancestral sequences of Ras using Maximum Parsimony and Maximum Likelihood procedures.
An alignment of the Ras family ancestral sequence reconstructed by parsimony and ML. Sites with a black background indicate matching residues between the reconstructions. Sites 5 to 165 (Ras family numbering (Nicely, et al. 2004)) were reconstructed. Image from GeneDoc (Nicholas, et al. 1997)



Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403-410.

Atchley, W. R., T. Lokot, K. Wollenberg, A. Dress, and H. Ragg. 2001. Phylogenetic Analyses of Amino Acid Variation in the Serpin Proteins. Molecular Biology and Evolution **18**: 1502.

Atchley, W. R., and A. D. Fernandes. 2005. Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network. Proc. Natl. Acad. Sci. USA **102**:6401–6406.

Bairoch, A., B. Boeckmann, S. Ferro, and E. Gasteiger. 2004. Swiss-Prot: Juggling between evolution and stability. Brief. Bioinform. **5**:39-55.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2005. GenBank. Nucleic Acids Res. **33(**Database issue):D34-8.

Bruno, W. J., N. D. Socci, and A. L. Halpern. 2000. Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction. Mol. Biol. Evol. **17**:189-197.

Chang, B. S. W. 2003. Ancestral Gene Reconstruction and Synthesis of Ancient Rhodopsins in the Laboratory. Integrative and Comparative Biology **43**:500–507.

Chang, B. S., J.A. Ugalde, and M.V. Matz. 2005. Applications of ancestral protein reconstruction in understanding protein function: GFP-like proteins. Methods Enzymol. **395**:652-70.

Crespo, P., and J. León. 2000. Ras proteins in the control of the cell cycle and cell differentiation. Cell. Mol. Life Sci. **57**:1613-1636.

Danen, E. H., P. Sonneveld, A. Sonnenberg, and K. M. Yamada. 2000. Dual stimulation of Ras/mitogen-activated protein kinase and RhoA by cell adhesion to fibronectin supports growth factor-stimulated cell cycle progression. J. Cell Biol. **151:** 1413-1422.

Devos, D., S. Dokudovskaya, F. Alber, R. Williams, B. T. Chait, A. Sali, and M. P. Rout. 2004. Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture. PLOS Biology **2**:2085-2093.

Ellis, C. A., M. D. Vos, H. Howell, T. Vallecorsa, D. W. Fults, and G. J. Clark. 2002. Rig is a novel Ras-related protein and potential neural tumor suppressor. Proc. Nat. Acad. Sci. **99**:9876-9881.

Feig, L. A. 2003. Ral-GTPases: approaching their 15 minutes of fame. Trends Cell Biol. **13**:419-25.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics **5:** 164-166.

Feng, D. F. , G. Cho, and R. F. Doolittle. 1997. Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA **94**:13028-13033.

Garcia-Ranea, J. A., and A. Valencia. 1998. Distribution and functional diversification of the ras superfamily in Saccharomyces cerevisiae. FEBS Letters **434**:219-225.

Gu, X. 1997. The age of the common ancestor of eukaryotes and prokaryotes: statistical inferences. Mol. Biol. Evol. **14**:861-866.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**:10915-10919.

Hillis, D. M., C. Moritz, and B.K. Mable, ed. 1990. Molecular Systematics. Sinauer Associates, Sunderland, Mass.

Hoshino, M., T. Yoshimori, and S. Nakamura. 2005. Small GTPase proteins Rin and Rit Bind to PAR6 GTP-dependently and regulate cell transformation. J. Biol. Chem. **280**:22868-22874.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310-2314.

Isoldi, M. C., M. A. Visconti, and A. M. de Lauro Castrucci. 2005. Anti-cancer drugs: molecular mechanisms of action. Mini Rev. Med. Chem. **5**:685-95.

Jékely, G. 2003. Small GTPases and the evolution of the eukaryotic cell. BioEssays **25**:1129–1138.

Jermann, T. M., J.G. Opitz, J. Stackhouse, and S.A. Benner. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature **374**:57-59.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

Kinbara, K., L. E. Goldfinger, M. Hansen, F. L. Chou, and M. H. Ginsberg. 2003. Ras GTPases: integrins' friends or foes? Nat. Rev. Mol. Cell Biol. **4**:767-76.

Kumar S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment. Briefings in Bioinformatics **5**:150-163.

Kozminski, K. G., L. Beven, E. Angerman, A. H. Y. Tong, C. Boone, and H.-O. Park. 2003. Interaction between a Ras and a Rho GTPase Couples Selection of a Growth Site to the Development of Cell Polarity in Yeast. Mol. Biol. Cell. **14**:4958–4970.

Mans, B. J., V. Anantharaman, L. Aravind, E. V. Koonin. 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. Cell Cycle **3**:1612-1637.

Milburn, M. V., L. Tong, A. M. deVos, A. Brunger, Z. Yamaizumi, S. Nishimura, and S. H. Kim. 1990. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. Science **247**:939-45.

Mount, D. W. 2001. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Myagmar, B. E., M. Umikawa, T. Asato, K. Taira, M. Oshiro, A. Hino, K. Takei, H. Uezato, and K. Kariya. 2005. PARG1, a protein-tyrosine phosphatase-associated RhoGAP, as a putative Rap2 effector. Biochem. Biophys. Res. Commun. **329**:1046-52.

Nicely, N. I., J. Kosak, V. de Serrano, and C. Mattos. 2004. Crystal structures of Ral-GppNHp and Ral-GDP reveal two binding sites that are also present in Ras and Rap. Structure (Camb) **12**:2025-36.

Nicholas, K. B., H. B. Nicholas Jr., and D.W. Deerfield II. 1997. GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW.NEWS **4**:14.

Nixon A. E., M. Brune, P. N. Lowe, and M. R. Webb. 1995. Kinetics of inorganic phosphate release during the interaction of p21ras with the GTPase-activating proteins, p120-GAP and neurofibromin. Biochemistry **34**:15592-8.

Nobes, C. D., and A. Hall. 1995. Rho, rac, and cdc42 GTPases regulate the assembly of multimolecular focal complexes associated with actin stress fibers, lamellipodia, and filopodia. Cell **81:**53-62.

Olson, M. F., H. F. Paterson, and C. J. Marshall. 1998. Signals from Ras and Rho GTPases interact to regulate expression of p21Waf1/Cip1. Nature **394:**295-299.

Oxford, G., and D. Theodorescu. 2003. Ras superfamily monomeric G proteins in carcinoma cell motility. Cancer Letters **189**:117-128.

Reuther, G. W., and C. J. Der. 2000. The Ras branch of small GTPases: Ras family members don't fall far from the tree. Current Opinion in Cell Biology **12**:157-165.

Ronquist, F., and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Sahai, E., M. F. Olson, and C. J. Marshall. 2001. Cross-talk between Ras and Rho signalling pathways in transformation favours proliferation and increased motility. EMBO J. **20:**755-766.

Saucedo, L. J., X. Gao, D. A. Chiarelli, L. Li, D. Pan and B. A. Edgar. 2003. Rheb promotes cell growth as a component of the insulin/TOR signalling network. Nature Cell Biology **5**:566-571.

Scheffzek, K., M. R. Ahmadian, W. Kabsch, L. Wiesmuller, A. Lautwein, F. Schmitz, and A. Wittinghofer. 1997. The Ras-RasGAP Complex: Structural Basis for GTPase Activation and Its Loss in Oncogenic Ras Mutants. Science **277**:333-338.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502-504.

Sebti, S. M., and A. D. Hamilton. 2000. Farnesyltransferase and geranylgeranyltransferase I inhibitors and cancer therapy: lessons from mechanism and bench-to-bedside translational studies. Oncogene. **19**:6584-93.

Spano, D., I. Branchi, A. Rosica, M. T. Pirro, A. Riccio, P. Mithbaokar, A. Affuso, C. Arra, P. Campolongo, D. Terracciano, V. Macchia, J. Bernal, E. Alleva, and R. Di Lauro. 2004. Rhes is involved in striatal function. Mol. Cell Biol. **24**:5788-96.

Sprang, S. R. 1997. G PROTEIN MECHANISMS: Insights from Structural Analysis. Annu. Rev. Biochem. **66**:639–78.

Stork, P.J.S. 2003. Does Rap1 deserve a bad Rap? Trends in Biochemical Sciences **28:**267-275.

Swofford, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods) Version 4. Sinauer Associates, Sunderland, Mass.

Sun, J., Y. Qian, A. D. Hamilton, and S. M. Sebti. 1998. Both farnesyltransferase and geranylgeranyltransferase I inhibitors are required for inhibition of oncogenic K-Ras prenylation but each alone is sufficient to suppress human tumor growth in nude mouse xenografts. Oncogene **16**:1467-73.

Taira, K., M. Umikawa, K. Takei, B. E. Myagmar, M. Shinzato, N. Machida, H. Uezato, S. Nonaka, and K. Kariya. 2004. The Traf2- and Nck-interacting kinase as a putative effector of Rap2 to regulate actin cytoskeleton. J. Biol. Chem. **279**:49488-96.

Takai, Y., T. Sasaki, and T. Matozaki. 2001. Small GTP-Binding Proteins. Physiological Reviews **81**:153-208.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research **24**:4876-4882.

Thornton, J. W., E. Need, and D. Crews. 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science **301**:1714-1717.

Vargiu, P., R. De Abajo, J. A. Garcia-Ranea, A. Valencia, P. Santisteban, P. Crespo, and J. Bernal. 2004. The small GTP-binding protein, Rhes, regulates signal transduction from G protein-coupled receptors. Oncogene **23**:559-68.

Wennerberg, K., K. L. Rossman, and C. J. Der. 2005. The Ras superfamily at a glance. Journal of Cell Science **118**:843-846.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in BioSciences **13**:555-556.

Zhu, Y., D. Pak, Y. Qin, S. G. McCormack, M. J. Kim, J. P. Baumgart, V. Velamoor, Y. P. Auberson, P. Osten, L. van Aelst, M. Sheng, and J. J. Zhu. 2005. Rap2-JNK removes synaptic AMPA receptors during depotentiation. Neuron **46**:905-16.

Chapter 2

Multivariate Analysis of the Ras Family Proteins

by

Andrew E. Dellinger[1] and William R. Atchley[2]

[1]Bioinformatics Research Center and [2]Department of Genetics

North Carolina State University

Raleigh, NC 27695

Send editorial correspondence to Andrew E. Dellinger (aedellin@ncsu.edu)

Abstract

Protein families covary in their sequence, structure and function. Covariation in amino acid sites is composed of functional, structural, phylogenetic, interaction, and stochastic components. This study explores the functional and structural components of covariation and their interactions in the Ras family of proteins through multivariate analyses. Decision tree analyses showed that sites with significant functional and structural roles discriminate among subfamilies and between cell-growth and cell-division controlling proteins. Structural alignments revealed that the structure of Ras family proteins is most variable in switch II and is most conserved in nucleotide binding loops. Common factor analysis was used to detect covariation of sequence, structure, and function. The covarying amino acid sites within each factor had a visible phylogenetic component- they discriminated subfamilies. The sites also had a common structural or functional role in the protein. Two clusters of sites were found that may have significance in maintaining protein structure during changes in switch conformations.

Introduction

A fundamental part of understanding the structural, functional, and evolutionary mechanisms acting on a protein family is understanding the role of the underlying components of variation and covariation. Multivariate analyses are useful tools for detection and quantification of variation and covariation and for exploration of their underlying components. The underlying components of amino acid covariation were described in Atchley et al. (2000), which proposed that covariation among amino acid sites could be

decomposed into phylogeny, structure, function, interactions of phylogeny, structure, and function, and stochastic processes. Thus, covariation between amino acid sites *i* and *j* (C_{ij}) can be expressed mathematically (Atchley et al. 2000) as

$$C_{ij} = C_{\text{phylogeny}} + C_{\text{structure}} + C_{\text{function}} + C_{\text{interactions}} + C_{\text{stochastic}}.$$

(1)

Felsenstein (2004) pointed out that we must not forget the phylogenetic component of covariation ($C_{\text{phylogeny}}$) among amino acid sites in looking for structural information or for correlated evolution. Covariation among amino acid sites may be due in part or in whole to their common evolutionary history. Statistical methods such as likelihood ratio tests (Pagel, 1994; Pollock, Taylor and Goldman, 1999) and parametric bootstrapping (Wollenberg and Atchley, 2000) have been used to account for the phylogenetic component of covariation. However, they cannot completely separate the interactions between the phylogenetic and the structural and functional components.

Proteins must maintain their folded structures to the extent that they retain biological function. The structural component of covariation ($C_{structure}$) comes from the limited number of combinations of amino acid substitutions that are allowed by this constraint (Atchley et al. 2000). Physiochemical interactions among structurally important sites in a protein are critical and an amino acid substitution in one site often requires parallel substitutions, conformational change, or a change in side chain orientation to maintain the interaction and thus the structure (Lesk and Chothia, 1982). The functional component ($C_{function}$) comes from interactions that occurred either to maintain a protein's biological role or to change its function (Atchley et al. 2000).

Structural and functional components of covariation are not necessarily linear. Rather, they interact with each other as well as with phylogeny, which necessitates the $C_{interaction}$ component (Atchley et al. 2000). For example, protein binding and ligand binding regions must maintain a specific conformation in order for the protein to function. And sites not essential to function may covary to properly position a functional site.

By understanding the roles of the components of covariation among amino acid sites, evolutionary and structural mechanisms within related proteins can be better understood. For example, structural change may change protein function. Such change may trigger purifying selection, which must reverse or compensate for the change or the change may trigger directional selection, which could eventually complete the change. Further, covarying sites with a large structural component can reveal structural constraints and give better insight into the proteins' structural dynamics. The structural consequence of the substitution of just a single residue varies significantly. It might be as minimal as changing the local environment around the residue (Jackson et al. 1993) or as big as changing the orientation of α-helices or even breaking them (Liu et al. 2004; Konvicka et al. 1998). Understanding the structural component can aid the understanding and prediction of any compensatory substitutions and structural changes.

Although the goal of many methods of structural comparison is determination of structural similarity, these methods, including measurement of the root mean square deviation (RMSD) of a structural alignment (Humphrey et al. 1996; Russell and Barton, 1992; Deiderichs, 1995) and differences in interatomic distances (Holm and Sander, 1993) can also be used to describe structural variation. Herein, a hybrid method which measures

interatomic distances in a structural alignment is used to find regions of significant structural variation.

Functional variation is difficult to measure as differences in the functions of proteins in a family such as the Ras protein family are primarily qualitative in nature. Variation of function is nearly inseparable from the variation of structure and sequence. Understanding the covariation among these elements of function, sequence and structure is perhaps the best way to understand the variation of function.

Proteins that share a common fold or active site conformation also share a common function given a certain amount of sequence identity. Protein function has been predicted because of this covariance of structure with function and sequence (Fetrow et al. 1999; Wilson et al. 2000). Understanding the covariation within a set of homologous proteins can yield greater knowledge of the level of identity, fold sharing requirements and the sites that must change together in sequence or structure to maintain a certain function.

Analysis of covariation between structure and function begins with the structural alignment of homologous proteins with varying functions. Structural alignment can reveal multidimensional structural constraints which maintain the fundamental relationship among the proteins. For example, Ras family proteins are involved in cellular proliferation, growth, adhesion and exocytosis but they all act as molecular switches and share guanine nucleotide binding structure and GTPase function (Wennerberg et al. 2005; Feig, 2003; Kontani et al. 2002). One might expect that in Ras proteins sequence and structure must covary to maintain the proper functioning of protein binding, nucleotide binding and GTP hydrolysis. Herein, structural alignment of proteins with a common general fold and differing function is

conducted to explore this covariation of sequence, structure and function.

Sequence variation can be measured with Boltzmann-Shannon entropy using letter representations of amino acid residues (Atchley et al. 1999, 2000), or numerical representations through physiochemical indices or factor scores (Atchley et al. 2005). Previous studies measured the relationship between sequence and crystal structure by comparing protein structures using procedures like the RMS (root-mean-square) separation measure, while their sequences themselves are compared using percent identity (Chothia & Lesk 1986,1987; Flores et al. 1993; Russell & Barton, 1994; Russell et al. 1997). With RMS, the amount of structural evolution is an exponential function of the amount of sequence evolution. Other approaches use the minimal level of sequence identity or maximum amount of sequence evolution between proteins and from that imply structural homology (Doolittle, 1987; Murzin et al. 1995; Brenner et al. 1998). In addition, existing protein crystal structures will provide data on structural compensations for different sets of sequence changes as well as allowable and prohibited structural changes.

Common factor analysis has been used to predict protein folding time and secondary structure (Ortiz and Skolnick, 2000; Lee et al. 1990; Wi et al. 1998). Spectra from infrared analyses of proteins were used as input in order to predict secondary structure (Lee et al. 1990; Wi et al. 1998). Ortiz and Skolnick (2000) used protein sequences as input for factor analysis to identify clusters of sites that control folding rate. A multiple sequence alignment can also be input for factor analyses, where the variables are amino acid sites. Common factor analysis groups these variables into clusters of covarying sites (factors). Variation in such clusters may arise from a common functional or structural role in the protein such as the

control of folding time and secondary structure in previous studies (Ortiz and Skolnick, 2000; Lee et al. 1990; Wi et al. 1998).

Covariation of protein sequence, structure, and function is a difficult problem to address. Many related protein sequences are needed to accurately detect and describe the covariation of sites within related proteins. There are over 4.22 million protein sequences available in the Entrez search engine (Wheeler et al. 2006), but reliable sequences in well annotated databases such as Swissprot and TrEMBL are fewer in numbers- 5% and 58%, respectively (Bairoch et al. 2004; Bairoch et al. 2005). An even bigger problem is the lack of crystal structures within groups of related proteins. Solving tertiary structures of proteins is difficult. There are about 34,500 solved crystal structures in the Protein Data Bank (PDB) (Berman et al. 2000) and this number is growing more slowly than the number of protein sequences. A significant number of solved crystal structures within a group of related proteins is necessary to represent the structural variation within that group. Only with sufficient representation of structural variation can structural and functional covariation be related to the covariation among amino acid sites.

Ras Proteins

Herein, we explore the variation and covariation in the Ras protein family. Ras proteins are involved in a variety of cellular functions such as differentiation, proliferation, growth, adhesion and exocytosis (Wennerberg et al. 2005; Feig, 2003). As such, they provide an excellent source of biological data to study the components of covariation among amino acid sites. They include proteins of common evolutionary origin and closely related protein

sequences as well as proteins of varying function. The crystal structures of the Ras family are well represented in PDB (Berman et al. 2000). There are GDP-bound and GTP-bound structures for proteins of 7 different functions. In total there are structures for 18 nonmutant proteins, including proteins bound to GDP, GTP, GTP and effector, GTP and GTPase activating protein (GAP), and GTP and guanine nucleotide exchange factor (GEF). Further, there are also structures for 12 mutant proteins available.

Ras family proteins are involved in 20-30% of all human cancers (Isoldi, et al. 2005), making the understanding of the structural, functional and interaction components of their site covariance highly relevant. Revelations about the structural and functional components of amino acid covariation in the Ras family may give additional insight into the mechanisms of Ras proteins in cancer and into possible treatments. As a consequence of their role in cancer, Ras family proteins have a large amount of literature detailing biological, structural and functional information, which can be used to supplement computational analyses and to match significantly covarying amino acid sites to their functional and structural roles.

Protein taxonomy can be slightly confusing. For example, there are Ras superfamily, Ras family and Ras subfamily proteins. The Ras superfamily is composed of 6 protein families including the Ras family. The Ras family is composed of several subfamilies including Rap, Ral, Rheb and Ras. In this study, covariation was analyzed at the level of the Ras family. So for clarity, the word Ras will be used to denote the Ras family in the strict sense unless otherwise noted.

The Ras family includes Rap, Ral, Rit, Rheb, DexRas, Di-Ras, Ras, H-Ras, N-Ras, K-Ras, M-Ras, and R-Ras. They are all GTPases that act as molecular switches and participate

in cellular functions such as differentiation, division, growth, exocytosis, endocytosis, and adhesion (Wennerberg et al. 2005; Feig, 2003). Most notably, H-Ras, N-Ras, K-Ras, M-Ras and R-Ras are proto-oncogenes (McCormick, 1995; Valencia et al. 1991).

Eight Ras family proteins have solved 3-dimensional structures in one or more nucleotide binding states in PDB (Berman et al. 2000): H-Ras (PDB ID 121p; Wittinghofer et al. 1991), M-Ras (PDB ID 1X1S; Ye et al. 2005), Rheb (PDB ID 1XTS; Yu et al. 2005), R-Ras2 (PDB ID 2ERY; *unpublished*), Di-Ras2 (PDB ID 2ERX; *unpublished*), Rap1a (PDB ID 1C1Y; Nassar, 1995), Rap2a (PDB ID 2RAP; Cherfils et al. 1997) and Ral-A in *Homo sapiens* (PDB ID 1UAD; Fukai et al. 2003) and Ral-A in *Saguinus oedipus* (PDB ID 1U8Y; Nicely et al. 2004). They are structurally characterized by their possession of 5 α-helices, a 6 strand β-sheet and 10 loops in their core tertiary structure (Bateman et al. 2004). Ras proteins contain two switch regions. Switch I is composed of sites 30-38 in L2, called the effector loop because it binds effector proteins (Nicely et al. 2004; Sprang, 1997). Switch II is composed of sites 60-76 in L4 and part of α2 (Milburn et al. 1990) and binds GAPs (Sprang, 1997) and GEFs (Crechet et al. 1996). In the active state, Ras family proteins are GTP-bound and the switch regions are in the proper conformation to bind effectors (Sprang, 1997). GTPase activating proteins (GAPs) bind to Ras proteins and increase the rate of the hydrolysis of GTP to GDP and an inorganic phosphate (Oxford and Theodorescu, 2003). Upon GTP hydrolysis, the switch regions change conformation (Crespo and León, 2000) and the protein is switched off, i. e. can no longer bind to its effector molecules (Sprang, 1997).

Ras proteins have considerable potential for studying sources of covariation. Switch regions are critical for function (Sprang, 1997; Crechet et al. 1996). Sequence covariation in

these regions probably has a large functional component. Switch regions must properly bind effectors, GAPs and GEFs for the protein to function and conformation is critical to this binding. The nucleotide binding regions participate in the functional, structural and interaction components of covariation because these regions are important for maintaining nucleotide binding, GTP hydrolysis and protein conformation. Last, there are at least nine groups of Ras proteins with unique sets of functions. Thus, there are at least nine instances when the interplay among sequence, structure and function can be assessed.

We examine the following questions: What subset of sites and residues discriminate Ras family functions and can a functional component of variation be found in these sites? How does structure covary with function? How does structure covary with sequence? What do clusters of covarying sites reveal about the structural and functional components of covariation among amino acid sites in Ras proteins?

Materials and Methods

Herein, three methods were used to characterize the components of variation and covariation in the Ras family. First, decision tree analyses (Breiman et al. 1998) were performed to investigate covariation between sequence and function. Ras family function was divided into two functional groups for one analysis and ten functional groups in the other analyses. The hypothesis that switch sites are sufficient to build a DTA that discriminates functional groups was also tested. Second, two structural alignments were performed on GDP-bound Ras proteins and two were performed on GTP-bound Ras proteins to analyze structural variation (Gibrat et al. 1996; Madej et al. 1995; Humphrey et al. 1996). Pairs of

related sites in the aligned structures which surpassed an assigned distance threshold were further analyzed to define the relationship of the structural variation to variation of sequence and function. Third, common factor analysis (Johnson and Wickern, 1992) was used to partition covariation among amino acid sites into its structural, functional, and phylogenetic components.

A total of 98 Ras family proteins were retrieved from Swissprot (Bairoch et al. 2004) and multiply aligned using CLUSTALX (Thompson et al. 1997) with minimal adjustments by eye. The analyses focused on the Ras core, which is the functional part of the protein that is conserved in sequence and structure throughout the Ras protein family (Sprang, 1997). The core begins at site 5 and ends at site 164, the C-terminal site of the guanine binding region (Valencia et al. 1991; Chardin, 1993). Herein, the numbering of the sites of these proteins follow the Ras family numbering convention (Valencia et al. 1991).

Solved crystal structures of Ras family proteins were retrieved from the Protein Data Bank (PDB) (Berman et al. 2000). A structure is available in at least one binding state for H-Ras, M-Ras, R-Ras2, Rap1a, Rap2a, Rheb, Di-Ras2, and Ral proteins. The PDB database was queried with the names of Ras family proteins using the PDB search engine (www.pdb.org/pdb/Welcome.do). BLAST was used in order to identify and select structures. BLAST (Altschul et al. 1990) searches were conducted using H-Ras, M-Ras, and Rheb proteins in *Homo sapiens*. Ras subfamily names, i. e. Rit, were input into the PDB search engine. Structures were selected with a preference for nonmutants and for X-ray crystallography over nuclear magnetic resonance (NMR) as the technique for structural determination. For a list of retrieved structures, see Table 1.

Decision Tree Analysis

Decision tree analysis (DTA) is a method which chooses a subset of variables and their values to partition a set into *a priori* defined subsets. The chosen variables show what information is important in making this decision. The resulting decision tree can be used to decide which subset a new set member belongs to. For example, DTA has been used previously to assign protein function using amino acid sites as variables (Wang et al. 2001), to locate protein coding regions (Salzberg, 1995), to predict protein interactions (Zhang et al. 2004), and to allocate proteins into specific DNA binding groups (Atchley and Zhao, 2006). Herein, DTA was used to determine the sites and residues that discriminate Ras family functions. This discrimination can give insight into the functional component of variation in these sites.

The sites in the multiple alignment were variables input into the DTA. Amino acid sites that contained more than 50% gaps were removed from the analysis. Highly gapped sites typically reflect insertions involving a single subfamily. For other subfamilies, they are uninformative and do not covary with other sites since there is no variance outside of the subfamily containing the insertion. Thus, these sites were removed. The remaining sites were input into SAS® Enterprise Miner™ (www.sas.com), where DTAs were performed.

The Ras family was partitioned into *a priori* subsets of proteins according to their functional groups. Use of DTA to determine functionally discriminatory sites assumes that one or more variables, sites in the aligned data, can discriminate among the sequences of the *a priori* defined groups. The first grouping involved protein function, i. e. each protein was

classified as either cell growth or cell division controlling according to its documented function in the literature.

Second, proteins were classified by their functional group, which was defined as one or more clades in a Bayesian phylogenetic tree of the Ras family (Dellinger and Atchley, 2006). Proteins in each clade of the Bayesian tree putatively belong to the same subfamily and share both function and common ancestry. Only Ras and Rap subfamily proteins are present in more than one clade. Ras subfamily proteins in many of the fungi are present in two clades, whose taxa compose the functional group "fungal Ras". A second clade of Ras subfamily proteins, was assigned to the functional group named "Distant Ras" because of their long branch lengths. The proteins come from the slime molds *Dictyostelium discoideum* and *Physarum polycephalum* and the sponge *Geodia cydonium* (Dellinger and Atchley, 2005). Rap1, Rap2 and the Ras-related (RSR) proteins homologous to Rap1, though of differing function (Stork, 2003; Myagmar et al. 2005; Taira et al. 2004), were assigned to one functional group because of their probable common ancestry and the common GEF of Rap1 and Rap2 (Pellis-van Berkel et al. 2005). The other functional groups are: M-Ras and R-Ras, Rit, Ral, Rheb, Di-Ras, DexRas and H-Ras, N-Ras, and K-Ras.

Twelve Ras proteins in this analysis were not included in the Bayesian tree: RalA in the mouse *Mus musculus*, RalB in the rat *Rattus norwegicus*, K-Ras in *Mus musculus*, K-Ras in the carp *Cyprinus carpio*, K-Ras in the possum *Monodelphis domestica*, H-Ras in the chicken *Gallus gallus*, H-Ras in *Mus musculus*, H-Ras in the Murine sarcoma virus ns.c58, Ras in the goldfish *Carassius auratus*, Rheb in *Mus musculus*, Di-Ras2 in *Mus musculus*, Di-Ras2 in the macaque *Macaca fascicularis*, and Di-Ras2 in the orangutan *Pongo pygmaeus*.

All but Ras in *Carassius auratus* have known subfamilies and were assigned to their subfamily's functional group. Ras2 in the fungus *Neurospora crassa*, RasS in the slime mold *Dictyostelium discoideum*, a Ras-like protein in the worm *Caenorhabditis elegans* and a Related-Ras protein in *Homo sapiens* chromosome 22 did not belong to the clade of a functional group. These proteins and the Ras protein in *Carassius auratus* were assigned to a functional group using a DTA as described below.

Three methods of decision tree-building are available in Enterprise Miner™: 1) the χ^2-test at significance levels 0.2, 0.1, and 0.05; 2) the entropy reduction test; 3) the Gini reduction test. The χ^2-test measures the value -log(p-value) for each possible split. A split is the choice of a child node (a site) and the subsets of residues that best match the criteria. The χ^2-test chooses the site and residues with the smallest value to -log(p-value).

The entropy and Gini reduction tests choose sites and residues with the maximum worth, where the calculation of worth depends on the test. The general formula for worth is:

$$I(node) - \Sigma_{branches}(P(b) * I(b)) \quad (2)$$

where *P(b)* is proportion of observations in the node assigned to branch *b* and *I(node)* and *I(b)* are the entropy or Gini values in the node and branch, respectively. The Gini measure for each node is

$$I(node) = \left(1 - \sum_{i=1}^{groups} \left(\frac{group_i}{group_{all}}\right)^2\right) \quad (3)$$

where $group_i$ is the number of proteins of functional group *i* in the decision tree *node* and $group_{all}$ is the total number of proteins in the *node.* For the entropy reduction

$$I(node) = -\Sigma_{aa}(P_{aa} * \log_2 P_{aa}) \quad (4)$$

where *aa* = one of the 20 amino acid residues and P_{aa} = the probability of seeing residue *aa* in the node (SAS 9.1 manual). Worth increases as entropy decreases, thus sites with the greatest identity tend to be chosen.

Each test was performed in both the functional group assignment, growth vs. division and functional group analyses. DTA assessed the model by the proportion of misclassified proteins in the tree and subtrees. DTA uses this proportion to determine if the proteins described by a branch of a tip node in the tree have a small enough proportion of mismatches to define an *a priori* subset. If they do not, another node is formed to further split the proteins into members and nonmembers of the subset. The final decision trees were evaluated by eye using the proportion of misclassified proteins as a measure of the quality of each tip of the tree. By this measure, the entropy tree was chosen in all analyses.

Decision trees using entropy were built as follows. The entropy test was performed on each combination of residue subsets in each site, where only the residues which occur in a site are tested. For example, the worth of {KR} vs. {LIQT} may be compared to the worth of {QT} and {LIKR}, where each letter in the subset is the standardized one-letter representation of a specific residue. In bifurcating decision trees, two subsets are tested at a time. The test chose the site and the subsets of residues in that site with the maximum worth. This site was the root of the tree. The residue subsets determined the branches emanating from the root. For each Ras protein, the method looked at the residue in the chosen site of each protein, decided which subset the residue belonged to, and associated the protein with the appropriate branch.

A separate entropy test was performed on the proteins associated with each branch.

Previously chosen sites are included in all future entropy tests, so the same site can appear more than once in a tree. The chosen site in each test became a child node of the root and the proteins associated with the branch were divided according to the residue subsets. For each branch, a prediction was made as to which *a priori* subset the proteins in the branch represent. The process of testing, splitting, and predicting was repeated for each new branch until that branch met one of the cutoff criteria. One cutoff criteria was that the proportion of proteins misclassified by a prediction is sufficiently small. This proportion could not be modified or identified using the DTA program. Another cutoff criteria was maximum tree depth, which is a limit on the number of nodes from the root to the tip of the decision tree.

The first DTA was conducted on all 98 Ras proteins. The proteins were assigned *a priori* to the cell growth or cell division category according to protein function. The second DTA was conducted on proteins with the nomenclature or subfamily classification in their Swissprot record that identified them as members of a functional group (i. e. RHEB_HUMAN). Exceptions are the "fungal Ras" functional group where the Ras but not Ras-like proteins in fungi were included.

The purpose of this DTA was to assign proteins to the correct functional group which were not in the clade of an existing group or which had uncertain functional similarity to its assigned functional group. Each assignment began at the root of the decision tree. An unclassified protein followed the branch that contained the residue in the site associated with the root. The process was repeated for the node that branch led to until a tip branch was reached. Proteins that successfully reached that branch were classified as a member of the subfamily the branch predicted. Proteins which did not successfully reach such a branch were

assigned using additional information from the tree. For each subfamily, the residues present in each chosen site in the tree were stored. The residues in the same sites in the unclassified proteins were compared to these residues. The proteins were assigned to the subfamily with which they had the greatest identity. The assignments were confirmed by analyzing the percent identity of the switch regions of each protein to the switch regions of the subfamilies.

A third DTA was performed which included the newly assigned proteins. Finally, DTA was performed to test the hypothesis that switch sites alone can discriminate among the functional groups of Ras proteins. The DTA was performed by eye. A site was selected from the Ras family alignment if the site's residues were diagnostic of at least one functional group without false positives or false negatives. Sites which discriminated more than one functional group were preferred. The model evaluation criteria was proportion of misclassified proteins, where no misclassified proteins were allowed.

Structural Analysis with VAST and VMD

Comparison of crystal structures can be performed using structural alignment algorithms. These algorithms attempt to superimpose protein structures as closely as possible (Gibrat et al. 1996; Madej et al. 1995; Russell and Barton, 1992). Herein, structural alignment algorithms in the Vector Alignment Search Tool (VAST) and Visual Molecular Dynamics (VMD) program are used on Ras structures to analyze structural variation, covariation of structure and function and covariation of structure and sequence (Gibrat et al. 1996; Madej et al. 1995; Humphrey et al. 1996; Russell and Barton, 1992).

VAST includes a database of structural neighbors which are prealigned using the

VAST algorithm. Structural neighbors are aligned protein structures with statistically significant alignments of their secondary structural elements (SSEs). The significance test is -log(p-value) < 4, where the p-value is computed by multiplying the number of independent protein structures in the database by the probability that the superposition score of aligned SSEs would be seen by chance in drawing SSE pairs at random. Computation of the superposition score is based on the number of residues and secondary structure elements aligned and the RMSD (Gibrat et al. 1996).

A protein's structural neighbors can be retrieved by entering the identification number of its crystal structure in PDB. To find the structural neighbors of GTP-bound Ras family structures, the VAST database (Gibrat et al. 1996; Madej et al. 1995) was queried using the identification number of the crystal structure of a *Homo sapiens* H-Ras protein bound to a GTP analog (PDB ID 121p; Wittinghofer et al. 1991). This structure was selected because the H-Ras protein is bound to a GTP analog, binds no other proteins and has no mutations in its sequence. Structural neighbors of this protein were chosen if they were bound to a GTP analog and had no sequence mutations. GTP analogs are not hydrolyzable and so negate the possibility of hydrolysis to GDP during structural determination by X-ray crystallography or nuclear magnetic resonance (NMR). Ras family structures were preferred but not required to be unbound to another protein. Where multiple structures met these criteria, a convenient structure was selected. Only one representative of each subfamily was chosen. The aligned PDB structures selected in addition to H-Ras (PDB ID 121p; Wittinghofer et al. 1991) were: M-Ras (PDB ID 1X1S; Ye et al. 2005), Rap2a (PDB ID 2RAP; Cherfils et al. 1997), Rap1a (PDB ID 1C1Y; Nassar, 1995), Ral (PDB ID 1U8Y; Nicely et al. 2004) and Rheb (PDB ID

1XTS; Yu et al. 2005). These aligned structures were further analyzed in Cn3D (Hogue, 1997). The unaligned, crystal structures of these proteins were input into VMD.

A second VAST database query was conducted for aligned, GDP-bound Ras family structures. All structures were chosen according to the criteria in the previous search with the exception that binding of GDP is required instead of binding a GTP analog. The GDP-bound H-Ras structure from *Homo sapiens* was chosen as the subject of the query (PDB ID 1IOZ; Kigawa et al. 2002). Rheb (PDB ID 1XTQ; Yu et al. 2005), Di-Ras2 (PDB ID 2ERX; *unpublished*), M-Ras (PDB ID 1X1R; Ye et al. 2005), Rap2a (PDB ID 1KAO; Cherfils et al. 1997) and Ral (PDB ID 1U8Z; Nicely et al. 2004) were its structural neighbors. The p-value of the alignment of the R-Ras2 (PDB ID 2ERY; *unpublished*) with H-Ras was above the cutoff. Proteins which are not structural neighbors cannot be included in the VAST alignment and thus cannot be included in the analysis (Gibrat et al. 1996; Madej et al. 1995). A VMD alignment of GDP-bound structures was also performed (Humphrey et al. 1996; Russell and Barton, 1992). The alignment included the R-Ras2 structure (PDB ID 2ERY; *unpublished*) in addition to the six structures included in the VAST alignment.

The prealigned structures from the VAST database were used in the VAST analysis. VAST alignment uses a Gibbs sampling algorithm beginning from an alignment of two equivalent secondary structural elements in different structures. The optimal alignment is the one that is least probable to be drawn from a distribution of alignments of random structural fragments (Gibrat et al. 1996; Madej et al. 1995).

The VMD analysis began with the unaligned, crystal structures. These structures were input into VMD and aligned using the STAMP (STructural Alignment of Multiple Proteins)

module (Humphrey et al. 1996; Russell and Barton, 1992). STAMP compares pairs of structures using a least squares fit on the alpha-carbon atoms. These pairwise comparisons are used to generate a structural tree, which guides the superimposition of the alpha-carbons in all structures of the multiple structure alignment. The resulting PDB files were merged into a single file.

The VAST alignment only displays the alpha-carbons (Gibrat et al. 1996; Madej et al. 1995). The VMD alignment, on the other hand, provides coordinates for all of the atoms in the structures (Bryant and Hogue, 1996). However, the variation among the positions of related alpha-carbons was the focus of both analyses. The VAST alignment was visualized in Cn3D (Hogue, 1997) and the VMD alignment was visualized in Protein Explorer (Martz, 2002).

Protein Explorer, a graphical interface for PDB structures, was used to calculate distances between related alpha-carbons, compare secondary and tertiary structure and perform other structural comparisons. Distances were calculated for all pairs of related alpha-carbons in the alignment of GTP-bound proteins. Distances of 1.0Å, 1.5Å, and 2.0Å between related alpha-carbons were tested as cutoffs for inclusion of structural sites in further analysis. A cutoff distance of 1.0Å resulted in too many significantly distant alpha-carbon pairs to distinguish patterns of structural and functional roles. A significance cutoff distance of 2.0Å was too restrictive (data not shown). So alpha-carbons in the same site which were greater than 1.5Å apart were considered significantly different. This threshold was kept for the analysis of GDP-bound proteins without calculation of the 1.0Å and 2.0Å distances. Structural comparisons of the VAST alignments were performed by eye since no method was

found that compared distances between alpha-carbons in angstroms.

Common Factor Analysis

Common factor analysis (Johnson and Wickern, 1992) is a multivariate statistical procedure that elucidates multidimensional, complex patterns of covariation among variables. Previous authors have used common factor analysis to predict protein folding time and secondary structure (Ortiz and Skolnick, 2000; Lee et al. 1990; Wi et al. 1998). Buck and Atchley (2005) used factor analysis to explore multidimensional covariation in serpin proteins and relate the covariance patterns to protein structure. The goal of our factor analysis is to describe total covariation among the amino acid sites in the Ras family protein alignment in terms of biologically interpretable patterns. The factor analyses herein will reveal clusters of covarying sites.

These patterns of covariation can be related to underlying structural, functional, or phylogenetic causes. Common factor analysis elucidates these patterns by combining a large number of observed variables into a smaller number of common factors. As outlined in Atchley et al. (2005), the common factors computed by the common factor analysis describe the underlying "latent structure" of the sets of interrelated variables they comprise.

Each site in a multiple sequence alignment can be considered as an observation. The general factor analysis model postulates that the variability among the sites is linearly dependent on a much smaller number of common factors. Common factors are unobservable random variables which describe the structure of highly correlated observed variables (Johnson and Wichern, 1982; Atchley et al. 2005). Each observation can be written as the

linear function

$X_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + ... + \varepsilon_i$

(5)

where F_i is factor *i*, l_{ij} is the loading of the *i*th variable on the *j*th factor and e_i is the unique variation of observation *i*. Unlike principal components analysis, in common factor analysis the unique variation is distinguished from the common variation (Johnson and Wichern, 1982). Adding the assumptions

$E(\mathbf{F}) = 0, Cov(\mathbf{F}) = \mathbf{I}$ (6)

$E(\boldsymbol{\varepsilon}) = 0, Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$ (7)

and **F** and **ε** are independent, where **F** is the *m*X1 vector of factors and *m* is the number of factors, **I** is the *m*X*m* identity matrix, **ε** is the *p*X1 vector of unique variances and *p* is the number of observations, and **Ψ** is a *p*X*p* diagonal matrix we can build the orthogonal factor model, which implies a covariance structure for the set of observations **X**. The orthogonal factor model is

$\mathbf{X} = \boldsymbol{\mu} + \mathbf{LF} + \boldsymbol{\varepsilon}$ (8)

where **μ** is the *p*X1 vector of the means of the variables and **L** is the *p*X*m* loading matrix for the *p* observations and *m* factors (Johnson and Wichern, 1982). In different terms, the elements of matrix **L** are factor coefficients which quantify the contribution of the site to the factor.

Factor coefficients are orthogonally rotated to "simple structure" to improve interpretability. If **L** is the *p*X*m* matrix of factor loadings and **T** is an orthogonal *m*X*m* matrix then

L* = **LT**, where **TT'** = **T'T** = **I** (9)

is the *p*X*m* matrix of rotated factor loadings. Simple structure includes having a large load on one factor and small loadings on the other factors for each variable and increasing the trichotomy of sites with factor coefficients near -1, 0 and 1 (Johnson and Wichern, 1982). For more information about common factor analysis see Johnson and Wichern (1982).

Factor coefficients range from -1 to 1 and quantify the relationship between the variables and the common factors. They can be used to determine which variables are significantly related to the factor. Herein, a factor coefficient of an absolute value of 0.6 or greater was considered significant. This cutoff was chosen to: minimize the likelihood of a site having a significant coefficient with more than one factor, yield a sufficiently large number of sites in each factor to determine the roles of the components of covariation, and yield a sufficiently small number of sites in each factor that the complexities of the patterns of the components' roles would be interpretable.

The multiple sequence alignment of the 98 Ras family Swissprot proteins was used to calculate a covariance matrix. This matrix was the input for the common factor analysis. The equation for calculating the covariance between sites *x* and *y* is

$$\sum_{i=1}^{20}\sum_{j=1}^{20} p(x_i, y_j) * \log_{20} \frac{(p(x_i, y_j))}{(p(x_i) \cdot p(y_j))} \quad (10)$$

where $p(x_i)$ is the probability of seeing amino acid *i* in site *x*, $p(y_j)$ is the probability of seeing amino acid *j* in site *y* and $p(x_i,y_j)$ is the probability of seeing amino acid *i* in site *x* and amino acid *j* in site *y* in the same protein. The log base 20 normalizes the covariance to a value between zero and unity. In the first factor analysis, a matrix of pairwise mutual information values was constructed that described the covariation in residues between sites in the

multiple alignment (Buck and Atchley, 2005). The covariance matrix of mutual information values was analyzed by common factor analysis. The resulting factors were orthogonally rotated. Many orthogonal rotations resulted in a large fraction of sites having high factor coefficients with factor 1 or factor 2. This made the patterns of covariation uninterpretable. The equamax rotation (Harman, 1976) was chosen because of its simple structure and because no resulting factor was overloaded as far as quantity of sites, thus patterns of covariation were more highly interpretable.

The number of factors was first estimated with a scree plot. A scree plot has the number of factors on the x-axis and the eigenvalue on the left. Factor 1 is represented by the number one (1) at the point (1, eigenvalue) where eigenvalue is factor 1's eigenvalue. Factor analyses were conducted beginning with the number of factors when the curve begins to flatten out. The number of factors was set at the maximum number of factors for which the last factor had at least five sites with a factor coefficient greater than the cutoff of 0.6.

The structural component of covariation was evaluated by using solved tertiary structures to determine sites' roles in protein structures. Physiochemical interactions may be reflected as significant covariances. Such interactions include hydrogen bonding, salt bridges and hydrophobic interactions. The functional component of covariation was evaluated by finding the roles of a factor's sites in protein function through a literature search or through evaluating the structures for physiochemical interactions with ligands, proteins, and residues within the protein.

Results and Discussion

Decision Tree Analyses (DTA)

Two decision tree analyses were performed. First, each Ras family protein was classified as either a cell-growth or a cell-division controlling protein. DTA was performed to find those sites and residues that distinguish these functional classes. Decision tree analysis divided cell-growth and cell-division controlling proteins on only a single site (Figure 1). Site 69 was the optimal choice and sites 62, 73, 105, and 149 were secondary choices. Sites 69 and 73 in switch II, the GEF and GAP protein binding region, have 100% accuracy in distinguishing between growth and division proteins (Table 2). The sets of residues in growth and division proteins do not share any members.

In proteins controlling cell division, site 69 has a negatively charged Aspartic or Glutamic Acid with the exception of Ras in the sponge *Geodia cydonium*. The cell growth controlling Di-Ras and DexRas proteins have the positively charged Arginine in site 69. Rheb proteins, which are also cell growth controlling proteins, have both polar and aliphatic side chains in site 69. In cell division proteins, site 73 has a positively charged Arginine or Lysine, except in a Ras-related protein in chromosome 22 in *Homo sapiens* and Ras in *G. cydonium* . Site 73 residues in cell growth proteins are uncharged. In both cases, a charge difference is the diagnostic attribute between growth and division controlling proteins. This difference apparently has a profound impact on GEF and GAP protein binding and thus on protein function and binding state dynamics.

For example, Rheb is actually a target of its GAP, Tsc2, which is a tumor suppressor

protein (Zhang et al. 2003). So switch II is important for Rheb function and a charge difference in site 69 or 73 would probably create a functional difference. Also, sites 69 and 73 in H-Ras-GTP are sufficiently close in the folded structure to sites 879-881 and 884 of the RasGEF SOS to suggest hydrophobic interactions. Site 73 has a potential H-bond with SOS as the nitrogen of R73 in H-Ras is 2.414Å from the side chain oxygen of N879 in SOS (PDB ID: 1BKD; Berman et al. 2000). There are, unfortunately, no crystal structures in PDB where Ras-related growth controlling proteins are bound to another protein. However, given that sites 69 and 73 change charge states as they change function, it is likely that these sites contribute to the change in biological function between the growth and division controlling proteins. In conclusion, it seems that the functional component of variation- particularly switch II binding is most important in dividing Ras proteins by their function in cell growth or cell division and in uniting the proteins within these categories. Perhaps this is because switch II has less variability than switch I across subfamilies.

There is a built-in phylogenetic component to variation in the sites of the DTAs in this manuscript. The cell growth proteins are near the outgroup in the Bayesian phylogenetic tree of the Ras family in Dellinger and Atchley (2005). The cell division promoting proteins are deeper in the tree and the two groups do not overlap. Rap proteins are in the cell division functional group but are part of a larger clade containing the cell growth proteins Di-Ras and DexRas. Further, the functional groups in the other DTAs were derived from the clades because most clades are subfamilies and thus have common function.

Next, each Ras family protein was classified by its functional group from the Bayesian phylogenetic tree in Dellinger and Atchley (2005). Four proteins did not clearly

belong to a functional group's clade in the phylogenetic tree: a Ras-like protein in the worm *Caenorhabditis elegans*, Ras2 in the fungus *Neurospora crassa*, RasS in the slime mold *Dictyostelium discoideum* and a Related-Ras protein in *H. sapiens* chromosome 22. These proteins' functional groups were assigned using a DTA which excluded proteins for which the subfamily was unidentified in Swissprot (Table 3). The DTA assigned all of these unidentified proteins to the functional group of the clade they belonged to. All assignments were confirmed by analyzing the percent identity of the switch regions of each protein to the switch regions of the clades (data not shown).

The RasS protein in *D. discoideum* followed the path in the DTA which led to the "distant Ras" functional group. The Ras-like protein in *C. elegans* followed the path in the DTA which led to the Rheb functional group. The Ras2 protein in *N. crassa* followed the path in the DTA which led to the Rit1 proteins in the Rit functional group. However, the Ras2 protein's residues did not match Rit's residues in sites 29, 101 and 105, where Rit has residues unique to its functional group. The Ras2 protein's switch regions have greater identity to the proteins of the "fungal Ras" functional group than the proteins of the Rit functional group. Further, they match other Ras2 fungal proteins in a BLAST search (Altschul et al. 1990). So the protein was included in the "fungal Ras" functional group. The Ras-related protein in *Homo sapiens* was deleted from the final analysis because it could not be assigned to a clade by any of the above methods.

Next, several DTAs were run to determine the best method of discriminating the functional groups. A bifurcating tree was chosen because analyses allowing multifurcating nodes were of lesser or equal accuracy than those restricted to bifurcating trees. The entropy

method was chosen because it had the lowest proportion of misclassified proteins (data not shown). Table 4 shows the results of this DTA.

Not all sites in the pathway to a functional group were considered to be discriminatory. For example, for the Di-Ras1 and Di-Ras2 pathway, decisions made in sites 88, 73, 31 and 9 all led to residues shared between Di-Ras and another functional group while decisions made in sites 7 and 21 led to unique residues to one or both of the Di-Ras1 and Di-Ras2 groups of proteins. Nondiscriminatory sites will not be mentioned in this section. Further, not all functional groups have structural representatives, so ascribing any structural or functional role to these proteins beyond that in the literature would be too speculative. Discussion of the pathways and discriminating sites of the DTAs will thus primarily involve the Ral, Rap, Rheb, Di-Ras, M-Ras and R-Ras and H-Ras, N-Ras and K-Ras functional groups.

DTA again chose sites 69 and 73 as tree nodes. These sites helped split the growth and division controlling proteins as in the previous analysis. Site 69 residues were diagnostic of some of the fungal Ras functional group. No specific structural or functional information could be found about the role of this site in switch II of fungal Ras. The residues of switch II site 73 are diagnostic of the Rap functional group. There are no PDB structures of a Rap protein with a protein-bound switch II region. Both sites 69 and 73, however very probably have important functional and structural components involved in GEF and GAP binding.

Other sites also have significant roles in maintaining structure and function. Sites 7 and 21 discriminate the Di-Ras functional group from the other groups. Site 7 is part of the binding site between switch II and $\alpha 3$ in the Ral, H-Ras and Rap functional groups (Nicely et

al. 2004). But A7 in Di-Ras2 and the M-Ras group is buried and almost certainly cannot participate in a binding site, thus discriminating the Di-Ras and M-Ras groups from the Ral, H-Ras and Rap groups (PDB ID 2ERX; *unpublished*). Further, site 7 is part of the hydrophobic core in Ral-GTP and has van der Waals' contacts with F72 (4.4Å) as well as potential contacts with F78 (3.846Å) and L56 (4.249Å) (PDB ID: 1U8Y; Nicely et al. 2004). V7 in H-Ras-GTP has potential van der Waals' contacts with sites Y71 (3.790Å) and G75 (3.795Å) (PDB ID: 121p; Wittinghofer et al. 1991). V7 in Rap2a-GTP has a potential van der Waals' contact with site 78 (3.672Å) (PDB ID: 2RAP; Cherfils et al. 1997). A7 in Rheb-GTP has a potential van der Waals' contact with site 75 (3.391Å) (PDB ID: 1XTS; Yu et al. 2005). However, the A7 of Di-Ras2-GDP does not allow these contacts, though V7 of Di-Ras1 might (PDB ID 2ERX; *unpublished*).

Di-Ras2, R-Ras2, and Ral crystallize as dimers (PDB ID 2ERX; *unpublished*; PDB ID 2ERY; *unpublished*; PDB ID 1U8Y; Nicely et al. 2004). I21 in R-Ras2-GDP does not interact with the dimerization region, but L21 in Di-Ras2-GDP and Ral-GDP do. They main-chain H-bond to site 25 (2.832Å and 2.772Å, respectively). In both structures, site 25 has potential hydrophobic interactions with the opposite chain- 3.822Å with switch I site I37 of Di-Ras2 and parallel rings of switch I site Y36 in Ral (≥3.590Å). Thus site 21 discriminates Di-Ras from the M-Ras and R-Ras group and the Rheb group and so sites 7 and 21 totally discriminate Di-Ras1 and Di-Ras2 from the rest of the functional groups. Here, the functional component of variation in site 7 is participation in a binding site and the structural component is participation in the hydrophobic core. The functional and structural component of variation in site 21 is the interaction with the dimerization region.

Site I7 is diagnostic of the Ral functional group. As mentioned above, I7 in the Ral group is part of a binding site, while site 7 in the M-Ras and R-Ras group and Di-Ras group is buried. Also, the van der Waals' contacts of site 7 differ between the H-Ras, Ral, Rheb and Rap groups, probably because of the differing switch II conformations among the proteins of the four groups (PDB ID: 1U8Y; Nicely et al. 2004; PDB ID: 121p; Wittinghofer et al. 1991; PDB ID: 1XTS; Yu et al. 2005; PDB ID: 2RAP; Cherfils et al. 1997). So site 7 is clearly diagnostic based on structural and functional components as well as the phylogenetic component.

K88 and M6 or L6 are diagnostic of the H-Ras, N-Ras and K-Ras functional group. The K88 in H-Ras-GTP has two H-bonds not found in the Ral-GTP or Rap1a-GTP structures: the main-chain H-bond with Q86 (3.322Å) and the side-chain H-bond with D92 (3.119Å). Additionally, K88 is within H-bonding range of the main-chain oxygen of T791 in the RasGAP (3.019Å) (PDB ID: 1WQ1; Scheffzek et al. 1997). Residues in site 88 of other functional groups may also form H-bonds with their GAPs but the different residues would help different GAPs to bind, thus sequence may vary with function in site 88.

H-, N-, K-Ras and non-fungal Rheb proteins split at site 6. L6 in H-Ras-GTP appears to have van der Waals' contact with site 163 (3.479Å) in α5, which influences L10 loop stability and thus the protein's affinity for GDP (PDB ID: 121p; Wittinghofer et al. 1991; Zhang and Matthews, 1998a,b). L6 is in main-chain H-bond range of L56 (2.860Å) and D54 (2.863Å) in the interswitch (PDB ID: 121p; Wittinghofer et al. 1991). These H-bonds may influence the position of the structurally and functionally critical D57 and G60 residues and Q61. D57 coordinates Mg^{2+} and the γ-phosphate of GTP (Paduch et al. 2001; Sprang, 1997).

G60 is a critical pivot point for the change in switch II conformation between the GDP-bound and GTP-bound states (Sprang, 1997). Q61 has been touted as the catalytic base in the S_N2 reaction by which GTP hydrolysis proceeds (Paduch et al. 2001). So site 6 has both functional and structural components to its variation that relate to nucleotide binding and hydrolysis.

Sites 7, 9 and 31 discriminate the M-Ras and R-Ras functional group from the other groups. No solved structure in this functional group is bound to a GAP or effector, so the role of D31 in switch I is unknown. However, in solved structures of other Ras family proteins, the residue in site 31 does not bind to effectors (PDB ID: 1HE8; Pacold et al. 2000; PDB ID: 1C1Y; Nassar, 1995; PDB ID: 1UAD; Fukai et al. 2003). V7 of M-Ras-GTP is buried and cannot participate in a binding site as it does in Ral, H-Ras and Rap. V9 is also buried and has potential van der Waals' contact with T58 (4.158Å), F96 (4.043Å) and L80 (3.674Å). V9 is in β1 and has potential main-chain H-bonds with sites L79 (2.957Å) and V81 (2.894Å) in the adjacent β4 strand (PDB ID: 1X1R; Ye et al. 2005). Thus, sites V7 and V9 have a large structural component of variation as they are important in maintaining the hydrophobic core and the β-sheet.

Other proteins have no structural information in PDB with which to describe their discriminating sites. However, information is available in the literature about their functional and/or structural role in Ras family proteins. For example, site 20 in H-Ras is in H-bonding range of K16**,** which H-bonds with the α and β-phosphates of GTP and GDP (PDB ID 121p; Wittinghofer et al. 1991). Site 5 is conserved in the Ras superfamily and is the first site in the protein core (Valencia et al. 1991). In H-Ras-GTP, site 5 has main-chain H-bonds with the

residues in sites 76 (2.805Å) and 77 (3.391Å) (PDB ID 121p; Wittinghofer et al. 1991). These residues could influence the flexibility of the switch II hinge residue G75 (Díaz et al. 2000).

In summary, the decision tree analysis performed by the entropy method primarily chose sites with detectable functional and structural components of variation. Sites with conserved residues unique to one functional group, such as the Ral or Rheb group offer the best reduction in the entropy and thus are the best choice for this method. This is the same concept used in Casari et al. (1995), which used principal components analysis to explore the sequence space of related proteins. The goal of that analysis was to find sites containing residues which are conserved within only one subfamily. Such residues were named tree-determinant residues, an apt name for DTA sites (Casari et al. 1995).

Tree-determinant sites and DTA sites tend to be under functional constraint (Casari et al. 1995). Such sites are conserved because they are necessary for the proteins to maintain their unique function in the group. The mechanism of this constraint is purifying selection, which eliminates mutations and so maintains sequence conservation. Sites under structural constraint are conserved because they are necessary for the protein to maintain the proper structure. Maintenance of structure is often important for function as well. Residues in sites under structural constraint may be required to change among groups because of variation in the structural and physiochemical environment of the sites in differing groups. For example, residue changes in other sites may change the structure such that a compensatory mutation is required. Structurally constrained sites may have physiochemical interactions with functionally relevant sites (sites 20 and 21), position functional sites for proper binding (site

6) or maintain protein structure (sites 5 and 7).

Last, a DTA was performed by eye using the multiple alignment (Table 5). Building the tree by eye will not always generate an ideal mathematical result, but the role of functionally and structurally significant sites in defining protein groups can be more directly explored. Sites were preferred to reside in or near a switch region and have the ability to discriminate among more than one functional group.

Although some sites outside the switch regions can discriminate more than one functional group, sites 33 and 70 in the switch regions and site 39 just C-terminal to switch I were found to be the best choices when the final tree was evaluated. Sites 30, 33, 39 and 70, which are in or adjacent to the switch regions, are nodes in the DTA (Table 5). Sites 26 and 83 are the other nodes. They discriminate among the Fungal, Distant and H-, N-, K-Ras groups. The proteins in these three groups have nearly identical switch I and switch II regions and likely have similar functions.

This tree has only one pathway for each functional group, while the Rheb, Di-Ras and "fungal Ras" functional group in the entropy-derived tree above required two pathways to describe the clade. They provided a tree with fewer decisions than non-switch candidate sites.

The major discriminatory factor in this DTA is probably function, given the protein binding function of the switch regions. Unfortunately, PDB has Ras family structures with protein-protein interactions only in the case of H-Ras, Ral and Rap. Site 70 in switch II discriminates between the Ral and Rap functional groups. The significant change in hydrophobicity from N70 in Ral and Q70 in H-Ras to L70 in Rap is probably the

discriminating factor in site 70 for Rap proteins. Specific changes in effector binding are unknown because L70 in Rap does not interact with its effector Raf1 (PDB ID: 1C1Y; Nassar et al. 1995). Ral is the only protein which has a solved structure where site 70 interacts with another protein. Specifically, N70 in Ral H-bonds to R226 (2.742Å) in the exocyst complex protein Exo84 (PDB ID: 1ZC3; Jin et al. 2005). N70 does not interact with Ral's effector, the exocyst protein Sec5 (PDB ID: 1UAD; Fukai et al. 2003). Q70 in H-Ras does not bind P120GAP, the GEF named SOS, or its effectors byr2 or PI3Kγ (PDB ID: 1WQ1; Scheffzek et al. 1997; PDB ID: 1BKD; Boriak-Sjodin et al. 1998; PDB ID: 1K8R; Scheffzek et al. 2001; PDB ID: 1HE8; Pacold et al. 2001).

Site 33 residues are diagnostic of the Di-Ras functional group. I33 in each chain of the Di-Ras2 dimer has a unique structural role through its potential hydrophobic interactions with T31 (3.493Å) and K29 (3.633Å) in the opposite chain (PDB ID 2ERX; *unpublished*). This interaction is not present in the Ral dimer or R-Ras2 dimer (PDB ID 2ERY; *unpublished*; PDB ID 1U8Y; Nicely et al. 2004). Site 39 discriminates between the Rheb and Rit functional groups. No intraprotein H-bonds or hydrophobic interactions were found in Rheb-GTP (PDB ID: 1XTS; Yu et al. 2005). Site 30 residues are diagnostic of the M-Ras and R-Ras group. Structural alignment of Ras proteins showed that P30 gives M-Ras a unique switch I conformation and thus effector and GAP binding. S30 in R-Ras and T30 in R-Ras2 probably has a more direct functional role through physiochemical interactions with effectors and GAPs. Site 83 residues are diagnostic of the H-Ras, K-Ras and N-Ras functional group. In the effector-bound mutant H-RasG12V, A83 nearly has van der Waal's contact with K117 (3.783Å) which H-bonds to guanine (PDB ID: 1HE8; Pacold et al. 2000). In effector-bound

Ral, the side-chain carbon of S83 is not as close to having contact with K117 (4.016Å) (PDB ID: 1UAD; Fukai et al. 2003).

This DTA constructed by eye reveals less about the functional and structural components of variation in the Ras family than the other DTAs did. The DTA has a lot of potential for describing the functional component of variation when solved structures of protein-protein interactions become available for more of the functional groups.

Structural Variation in the Ras Family

Figure 2 displays the results of the VAST alignment of GTP-bound Ras proteins (Gibrat et al. 1996; Madej et al. 1995). The structurally aligned H-Ras (PDB ID: 121p; Wittinghofer et al. 1991), M-Ras (PDB ID: 1X1S; Ye et al. 2005), Rap1a (PDB ID: 1C1Y; Nassar, 1995), Rap2a (PDB ID: 2RAP; Cherfils et al. 1997), Ral (PDB ID: 1U8Y; Nicely et al. 2004) and Rheb (PDB ID: 1XTS; Yu et al. 2005) structures were used. Differences could not be measured in angstroms and so were visualized. Backbone structures in nucleotide binding regions were more tightly clustered than in other sites. Nucleotide binding regions in Ras are important to structure as well as to GTPase function. Although the switch regions change conformation according to the nucleotide binding state, the rest of the structure changes very little if at all. However, Ras proteins which are not bound to a nucleotide are structurally unstable (Sprang, 1997). They may change in their general structure. Therefore, the interaction of nucleotide binding sites with the nucleotide are critical for maintaining general protein structure. Nonbinding sites in these regions are important as well. They maintain the proper conformation for nucleotide binding sites to physiochemical interact with

the nucleotide.

From equation 1, these dual roles of the nucleotide binding sites in structure and function are the interaction component $C_{interaction}$. Since the interaction component among these sites is probably large, the functional and structural components of covariation are probably small. Thus, related sites in the nucleotide binding regions of Ras proteins should have the same functional and structural roles in the various GTP-bound Ras proteins. A normalized mutual information matrix reveals that the total covariation among these sites is small since there is little or no sequence variation in these sites.

Switch I is composed of sites 30-38 and is the effector binding loop. Even though switch I is partially disordered when unbound (Nicely et al. 2004), the aligned switches are closely overlapping except in the case of M-Ras. Sequence distance among Ras family switch I sites was calculated using the JTT amino acid substitution matrix (Jones, Thornton, and Taylor, 1992) (data not shown). Ral has the greatest average sequence distance from other Ras family proteins but has similar structure to other switch I regions. Switch I in M-Ras has the greatest structural distance because site 30 of M-Ras contains a proline residue, while other Ras family proteins contain an acidic residue. P30 gives M-Ras switch I a sharper turn than the other Ras family proteins, causing the switch residues to be significantly displaced (Figure 2 and Figure 3) (PDB ID: 1X1S; Ye et al. 2005).

Unlike switch I, which has relatively little structural variation, switch II has clear variation among all sites and all Ras structures (Figure 2). Though the switch regions change conformation according to the protein's nucleotide binding state, they are unimportant in maintaining protein structure as a whole and thus have a small structural component to

covariation. However, they play a large role in function and that role varies between each aligned structure. This functional variation is reflected in the variation of switch I and II sequence and switch II structure.

Changes in main-chain location and conformation occur in sites adjacent to and including three indels (Figure 4). First, an insertion in Rap1 between sites 137 and 138 caused the insertion site and site 138 to significantly differ in their spatial coordinates (PDB ID: 1C1Y; Nassar, 1995). Second, a deletion in H-Ras between sites 120 and 121 caused a shift in the location of sites 121 and 122 (PDB ID: 121p; Wittinghofer et al. 1991). H-Ras site numbering is identical to Ras family site numbering, so any deletions occur between consecutive site numbers. Third, an insertion in M-Ras between sites 148 and 149 shifted site 148 and the insertion site (PDB ID: 1X1S; Ye et al. 2005).

Finally, the conformation of Rheb loop L7 varies from that of other Ras proteins (Figure 5). Sites 106-109 are visibly different in their spatial location, perhaps caused by Rheb's different main-chain angles between sites 103 and 105 (PDB ID: 1XTS; Yu et al. 2005). Unlike the five conserved nucleotide binding loops in Ras family proteins, L7 is not known to be critical to function or structure. Its structural variation seems to be primarily due to variation in sequence.

Next, the VMD alignment (Russell and Barton, 1992) of GTP-bound Ras proteins was analyzed in Protein Explorer (Martz, 2002). In the VMD alignments, all differences or variations discussed have surpassed the significance threshold of 1.5Å apart. The regions of structural variation were similar in the GDP and GTP-bound VMD alignments (Tables 6 and 7). However, M-Ras-GTP had large regions of structural differences, whereas in the VAST

alignment M-Ras-GTP closely aligned with other GTP-bound proteins except in the switch regions.

As in the VAST alignment, switch I had few significant variations among the proteins (Figure 6). Switch I in Rap1a, Rap2a and Rheb had no related sites more distant than the 1.5Å significance threshold and H-Ras only surpassed this threshold in site 32 (PDB ID: 1C1Y; Nassar, 1995; PDB ID: 2RAP; Cherfils et al. 1997; PDB ID: 1XTS; Yu et al. 2005). Nearly all Ral and M-Ras switch I sites were significantly distant from all other Ras proteins (Table 6) (PDB ID: 1U8Y; Nicely et al. 2004; PDB ID: 1X1S; Ye et al. 2005). H-Ras and M-Ras are both proto-oncogenes whose proteins promote cell division, while Rap1a inhibits cell division and Rheb controls cell growth instead of cell division (Stork, 2003; Saucedo et. al 2003). So though functional variation probably contributes to structural variation, there does not appear to be a direct relationship between the amount of structural difference and the amount of functional difference in switch I. Switch I structure appears to depend more on sequence variation. As mentioned above, Ral has the largest average sequence distance and M-Ras has a proline in site 30, which gives switch I a unique conformation (PDB ID: 1X1S; Ye et al. 2005).

The indels were sources of structural variation (Figure 7). The Rap1 insertion caused a shift in Rap1 numbered site 139 instead of sites 139 and 140 as in the VAST alignment (PDB ID: 1C1Y; Nassar, 1995). The H-Ras deletion caused the same shift as in VAST and the M-Ras insertion was in the middle of a region of structural shift, so no effects were noted (PDB ID: 121p; Wittinghofer et al. 1991; Ye et al. 2005). Interestingly, the deletion in H-Ras and the insertion in M-Ras are two sites and one site away from nucleotide binding sites,

respectively. In these cases the sequence appears to have evolved to maintain the physiochemical interactions with GTP which are so critical to structure and function (Sprang, 1997).

Loop L7 is again a region of structural variation (Figure 8b). The VMD analysis found a new region of significant structural variability in sites 42-50 in the C-terminal half of β2, L3, and the N-terminal half of β3 (Figure 8a). None of these sites match the interswitch cluster of covarying and interacting sites found by the common factor analysis which follows in the next section. The cluster of sites includes sites 40, 41, 51, 53 and 55, which are both N-terminal and C-terminal to this structurally variable region. We propose that this cluster of sites stabilizes Ras protein structure during conformational changes. Thus the variation in the interswitch region is probably limited to sites 42-50 because of structural constraints.

Next, the VAST alignment (Gibrat et al. 1996; Madej et al. 1995) of GDP-bound R-Ras2 (PDB ID: 2ERY; *unpublished*), M-Ras (PDB ID: 1X1R; Ye et al. 2005), H-Ras (PDB ID: 1IOZ; Kigawa et al. 2002), Ral (PDB ID: 1U8X; Nicely et al. 2004), Rap2a (PDB ID: 1KAO; Cherfils et al. 1997), Rheb (PDB ID: 1XTQ; Yu et al. 2005) and Di-Ras2 (PDB ID: 2ERX; *unpublished*) structures was analyzed (Figure 9). Again, the switch regions have structural variation as do the interswitch region (Figure 10a), L7 (Figure 9) and the sites adjoining the indels (Figure 10b and 10c).

In addition to the previous indels, Di-Ras2 has an insertion between sites 108 and 109. The relevant sites are missing from the Di-Ras2 structure (PDB ID 2ERX; unpublished). Another change is found in the conformations of M-Ras and Rheb switch II. They are different from their GTP-bound counterparts and from the other Ras family switch II

conformations. The sites in M-Ras switch II, which arch upward in Figure 9, are the most distant from the other Ras family proteins, while the C-terminal half of switch II in Rheb forms a U shape and is also distant from the related sites in the other Ras structures (PDB ID: 1X1R; Ye et al. 2005; PDB ID: 1XTQ; Yu et al. 2005).

Rheb and Rap2 begin the α5 helix in sites 148 and 149, respectively, while the other structures begin α5 in site 151. These other structures twist in the opposite way of α5, leading to their significant distance from the related alpha-carbons in Rheb and Rap2 (Figure 10b). There is a functional constraint imposed by A146, just two sites N-terminal to the differing twists. A146 H-bonds with the guanine base (Valencia et al. 1991). The L10 loop conformation must change as needed in order to maintain this interaction that stabilizes GDP binding. The position of GDP and the conformation of the surrounding structure in all Ras proteins are tightly overlapping. So in this case, structure appears to have changed to maintain the GDP binding functionality because of sequence variation.

In the VMD alignment of GDP-bound Ras structures (Figure 11) the switch regions are again sources of structural variation (Humphrey et al. 1996; Russell and Barton, 1992). M-Ras and Ral switch I sites are more closely aligned to the other Ras structures than in the alignment of GTP-bound structures (PDB ID: 1X1R; Ye et al. 2005; PDB ID: 1U8X; Nicely et al. 2004). This closer alignment results in fewer pairs of significantly distant alpha-carbons. Perhaps this is because switch I requires no function-specific conformation in the inactive GDP-bound state. Also, while all switch I sites differ between H-Ras and Ral in the GTP-bound state, none differ in the GDP-bound state. H-Ras and R-Ras2 as well as R-Ras2 and Rap2 are other protein pairs with no significantly differing sites (PDB ID: 1IOZ; Kigawa

et al. 2002; PDB ID: 2ERY; *unpublished*; PDB ID: 1KAO; Cherfils et al. 1997). In conclusion, there is a general decrease in switch I variability going from GTP-bound to GDP-bound proteins. However, one pair of proteins, Rap2 and Rheb gained differences in sites 31-35 (PDB ID: 1KAO; Cherfils et al. 1997; PDB ID: 1XTQ; Yu et al. 2005). H-Ras switch II is also more closely aligned, resulting in a decrease of 18 differences among 4 other proteins. Other pairs of proteins increased in their number of differences in sites 68-74. In switch II, GDP-bound Ras proteins unlike GTP-bound Ras proteins are probably required to adopt different switch II conformations in order to bind different GEFs. Thus, there is an additional source of structural variability in switch II.

As in the alignment of GTP-bound proteins, loop L7, indels and the interswitch region were sources of structural variation (Figure 12). Loop L7 variability and the site shifts due to indels remained the same, though the variability in interswitch sites 42-50 and 52-53, as measured by the number of pairs of significantly distant alpha-carbons, decreased in the GDP alignment. The decrease in variability in the C-terminus of switch I and the N-terminus of switch II may contribute to the decrease in variability in the interswitch region.

Sites in α5 had structural differences among the proteins (Figure 12). The orientation and intrahelix interactions in the α5 helix influence Ras family proteins' affinity for GDP by stabilizing L10 and positioning A146 to H-bond to GDP (Zhang and Matthews, 1998a, b). Structural differences in α5 probably occurred to maintain these interactions in the face of changing sequence. And in Di-Ras2, sites C-terminal to site 164 are structurally unique. These sites form two loops joined by a two residue β7 strand in sites 170-171 (PDB ID: 2ERX; *unpublished*) . The structure of these sites is discussed in subsequent paragraphs.

Finally, there were scattered differences in single residues of one protein where there was no discoverable functional cause. Sequence variation is likely to be the primary contributor here.

Di-Ras2 (PDB ID: 2ERX; *unpublished*) has unique conformations in sites 43 to 53 of the interswitch region and in the sites C-terminal to α5 (Figure 13). The interswitch region is composed of β2, L3 and β3. First, the conformation is unique because of the disulfide bond between the cysteine residues in sites 46 and 51 of chain B in the Di-Ras2 dimer. Second, the unique conformation is necessitated by the unique interactions of the interswitch region in Di-Ras2. The interactions of this region with β1 are similar to that of the same region in H-Ras. However, interactions with the C-terminal region of the protein are different.

In H-Ras, the α5 helix is composed of sites 152 to 168 (PDB ID: 1IOZ; Kigawa et al. 2002). In Di-Ras2, a unique conformation occurs C-terminal to Ras numbered site 163. The α6 helix terminates at site 163, 3 residues earlier than the homologous α5 helix in H-Ras (PDB ID: 2ERX; *unpublished*) . Other Ras structures have the α5 helix and one or two loop sites at the C-terminus. But C-terminal to α6, Di-Ras sites 164-169 compose L12, sites 170-171 compose β7 and site 172 is the C-terminal residue. Site 169 H-bonds to site 47 (2.783Å) and site 171 forms two H-bonds with site 45 (2.895Å, 3.040Å) (PDB ID: 2ERX; *unpublished*). Thus, the loop restricts the movement of the interswitch region, as does the hasp formed by the N-terminal helix in some Arf proteins (Pasqualato et al. 2002). Both the Di-Ras2 loop and the Arf hasp interact with the β2 and β3 strands in the GDP-bound form. The proteins are also structurally similar in that they both dimerize at β2 (PDB ID 1ERX; Pasqualato et al. 2002). The structural differences in β2 of Di-Ras2 from other Ras family proteins are greater than 1.5Å. This structural change probably contributes to functional

change as it does in Arf, where two proteins become one functional unit (PDB ID 1ERX; Pasqualato et al. 2002).

The purpose of the C-terminal loop/interswitch interaction in Di-Ras2 is probably different from the purpose of the N-terminal helix/interswitch interaction in Arf. Di-Ras2 is 52% GTP-bound and has very little GTPase activity compared to H-Ras. Residues A59 and Q61 are critical for GTP hydrolysis, so the A59T and Q61S substitutions in Di-Ras2 seriously impair hydrolysis. The resulting long half-life of GTP-bound Di-Ras2 enables it to function as a constant inhibitor of cell growth in brain tissue (Kontani et al. 2002). Thus, the purpose of the Di- Ras2 interaction is not likely to be maintenance the structure in the inactive GDP-bound state as it is in the Arf interaction (Pasqualato et al. 2002; 1ERX).

Patterns of Amino Acid Covariation

For computational expediency, many algorithms assume independence among nucleotide and amino acid sites (Huelsenbeck and Crandall, 1997; Tillier and Collins, 1995). This is not a biologically justifiable assumption in most cases. Sites may have dependence due to phylogeny, stochastic processes, and functional or structural constraints (Atchley et al. 2000). In order to maintain the structural geometry necessary for biological function, changes at one site may result in parallel change at other sites to maintain tertiary structure (Creighton, 1983). Such covarying sites may be detected using multivariate statistical processes such as common factor analysis on the between-site correlation matrix.

The normalized mutual information value, a measure of covariance used in information theory, was computed for every pair of sites in the multiple alignment of Ras

family proteins. The resulting mutual information matrix was considered to be a correlation matrix and used as input for a common factor analysis. The factors were orthogonally rotated to simple structure using the equamax method (Harman, 1976). A scree plot, total variance explained and the number of sites with correlation coefficients greater than the designated cutoff of 0.6 were used to choose six factors for further analysis. Table 8 shows the resulting factor coefficients. Sites in each factor that had a factor coefficient greater than an absolute value of 0.6 were further analyzed in their structural and functional relationship.

Sites in factor 1 (sites 94, 98, 108, 113, 114, 127, 130, 131, 136 and the insert between sites 121 and 122) are located on the opposite side of the switch regions and do not participate in protein-protein interactions. There are no perceivable patterns to the residues' physiochemical properties, the sites' arrangement in secondary structure, or the variation in residues among Ras subfamilies.

Factor 2 sites 5, 20, 22, 59, 61, 71, 77 and 120 discriminate DexRas and Di-Ras proteins. These sites have factor coefficients greater than 0.7, while sites 90, 123, and 157 which have factor coefficients between 0.6 and 0.7 do not clearly discriminate DexRas or Di-Ras proteins. In these three sites, DexRas and/or Di-Ras proteins share residues with at least one entire subfamily.

Site 59 is essential for proper switch II conformation change. A59 is part of the hinge that moves switch II and its mutation is usually found in oncogenic proteins (Diaz et al. 1997). The small volume of the alanine side chain enhances the flexibility that comes from the glycine in site 60, enabling loop L4 to bend in a way that places switch II in an active conformation (Kontani et al. 2002). The otherwise Ras family conserved A59 residue is

substituted with serine in DexRas and threonine in Di-Ras. These substitutions decrease the flexibility of switch II. The A59T substitution begins the unique conformation of switch II in Di-Ras2. The substitution also enables interactions not found in other Ras family proteins in the GDP-bound state: an H-bond with Q68 (2.991Å) and potential van der Waals' contact with site 68 (3.756Å) (PDB ID: 2ERX; *unpublished*).

Site 59 plays a critical role in GTP hydrolysis (Kontani, 2002) and the ability to signal and bind (Shirouzu et al. 1994). Site 61 is also necessary for GTP hydrolysis and nucleotide binding functions. Q61 activates water for nucleophilic attack on the γ-phosphate of GTP, a key part of the S_N2 reaction that hydrolyzes GTP (Sprang, 1997; Pai et al. 1990). Unlike most Ras family proteins, Di-Ras2 is primarily (52%) GTP-bound. This is a known effect of the Q61T substitution in Rap2, which is parallel to the Q61S substitution in Di-Ras. In addition, the substitution contributes to the much higher rate of GTPγS dissociation compared to the rest of the Ras family where GTPγS is very tightly bound (Kontani et al. 2002). In conclusion, sequence variation in sites 59 and 61 leads to variation in Di-Ras function and structure.

Site 120 affects GDP binding in an unknown way. Compared to H-Ras-GDP (PDB ID: 1IOZ; Kigawa et al. 2002), the L120E substitution in Di-Ras2-GDP adds a potential H-bond from E120 to the guanine base (3.343Å), lengthens the H-bond from K117 to the base from 3.125Å to 3.276Å and shortens the H-bonds of D119 with the base from 2.842Å to 2.828Å and from 2.950Å to 2.872Å. Interactions in α5 stabilize L10 for guanine binding (Zhang and Matthews, 1998a,b). Substitutions in sites 157 and 163 of α5 make little difference in the interaction of L10 with GDP, but may change the interactions with GTP.

Site 33 in switch I is also a factor 2 site. The D33I substitution eliminates interaction with the Ras effector PI3Kγ and the Ras and Rap1 effector Raf (Kontani et al. 2002) and almost certainly changes interactions with other effectors and GAPs the other Ras family proteins bind. The D33I and V29R substitutions contribute to the unique Di-Ras2 switch I conformation. Factor 2 sites 68, 71 and 73 are in switch II. In H-Ras, the residues in these sites H-bond to the RasGEF SOS (PDB ID: 1BKD; Boriack-Sjodin et al. 1998). In Di-Ras2, they may also have a functional role in binding GEFs and GAPs.

Site 5 at the beginning of the protein core is conserved in the Ras superfamily as a lysine or arginine residue. In Di-Ras2, R5 H-bonds to H80 at the N-terminus of switch II (2.816Å), K74 in switch II (2.651Å) and D3 in the N-terminus (2.992Å) (PDB ID: 2ERX; *unpublished*). Site 77 may affect the flexibility of the switch II hinge at site 75 (Díaz et al. 2000). Site 77 changes from glycine to alanine in DexRas and to valine in Di-Ras1 and Di-Ras2. The bigger side chain in DexRas and Di-Ras proteins would lessen the flexibility in this region and change the dynamics of conformational change. Sites 20, 22, 62, 90, 102 and 103 are the other factor 2 sites. They have unknown structural and functional roles in Di-Ras and DexRas proteins.

In addition to distinguishing the DexRas and Di-Ras subfamilies, factor 2 provides insight into the covariation of sites critical to the function and structure of the Ras family as a whole. Table 9 summarizes the roles of factor 2 sites in Ras function and structure. The sites primarily bind or influence the binding of nucleotides or proteins.

Sites 26, 80, 96, 101 and 151 in factor 3 discriminate the Rit subfamily. There is no solved Rit structure, but in Rheb these sites are important for positioning GTP-binding

residues. In the Rheb structure, G26 is the N-terminal site of L2, the effector loop, and begins the turn to position switch I for effector interaction and GTP binding. L80 is within H-bond range of sites 112 (2.820Å) and 114 (2.901Å), which position sites in L8 to bind the guanine ring of GDP and GTP. Site 96 may have van der Waals' contact with site 10 (3.832Å) in the phosphate binding loop and with site 61 (3.841Å) which is critical to GTP hydrolysis. T151 has H-bonds with Q149 (2.753Å), which is also in the α5 helix (PDB ID: 1XTS; Yu et al. 2005). Physiochemical interactions among sites in the α5 helix stabilize L10 and enable GDP binding (Paduch et al. 2001). Finally, site 101 has unknown structural implications.

Factor 4 sites (sites 7, 36, 37, 46, 53, 67, 70, 72 and 112) distinguish the Ral subfamily from the rest of the Ras family. They also reveal sites important for effector binding and specificity. A sequence space analysis (Casari et al. 1995) predicted fourteen tree-determinant sites to be responsible for the differences between Ral and Ras effector specificity (Bauer et al. 1999). Tree-determinant sites contain residues which are conserved within a subfamily and differ from the other subfamilies (Bauer et al. 1999). In factor 4, seven of these sites have a correlation coefficient greater than 0.6 and thirteen sites have coefficients greater than 0.5. However, site 33 has a greater correlation coefficient for factor 2 (0.633) than for factor 4 (0.543). Five sites are either in the effector loop (sites 36 and 37) or switch II (sites 67, 70 and 72) and six sites (sites 7, 46, 53, 67, 70 and 160) are part of the effector binding pockets predicted in Nicely et al. 2004.

Tree-determinant sites in factor 4 are critical for Ral structure and function. Sites 36 and 37 are responsible for effector binding in Ral proteins (Bauer et al. 1999) and have been shown to interact with Sec5 (Nicely et al. 2004). Site 53 changes from an α-helix site to a β-

turn site in the transition from the GDP to GTP-bound states. Sites 46, 53, and 160 are in the putative binding site adjacent to switch I and sites 7, 67 and 70 are in the putative binding site near switch II and α3 (Nicely et al. 2004). I7 H-bonds to L79 (3.081Å) and G77 (2.865Å) and has van der Waals' contact with I78 (3.56Å), thus stabilizing the β-strand C-terminal to switch II. Site 72 also has van der Waals' contact with I7 (4.4Å) and with I78 (3.5Å) (PDB ID: 1U8Y; Nicely et al. 2004). Both sites 7 and 72 are part of the hydrophobic core of Ral proteins (Nicely et al. 2004). Site 67 has van der Waals' contact with R226 in Exo84 (3.607Å, 3.921Å) (PDB ID: 1ZC3; Jin et al. 2005). Exo84 is an effector of Ral and is part of the exocyst complex (PDB). Site 70 H-bonds (2.953Å) with the main chain nitrogen of F247 in Exo84. Sites 46 and 112 have interactions of unknown importance.

Figure 14 shows the locations of two clusters of Factor 4 and Factor 5 sites in the tertiary structure of SOS-bound H-Ras-GTP (PDB ID: 1BKD; Boriack-Sjodin et al. 1998). One cluster is composed of sites 36-38, 40, 41, 55 and 56, which are in the C-terminus of switch I and the interswitch region. The second cluster is composed of sites 7, 66, 67, 70, 72, 75, 78, 104 and 112, which are primarily in Switch II and helix α3. In each cluster, the sites form a network of H-bonds and hydrophobic interactions.

In H-Ras and Ral proteins, the clusters appear to stabilize the C-terminal ends of the switches through their conformational changes. Extensive switch I cluster interactions are limited to the GEF-bound state of H-Ras. In this state, H-Ras site 37 near the C-terminus of switch I and site 40 C-terminal to switch I appear to be stabilized with three intraprotein interactions each (Figure 15; PDB ID: 1BKD; Boriack-Sjodin et al. 1998). These interactions do not exist in the effector-bound or GDP-bound states (Figure 16; PDB ID: 1HE8; Pacold et

al. 2000; PDB ID: 1IOZ; Kigawa et al. 2002).

Many of the interactions in the switch II cluster of H-Ras remain intact among the three states (Figures 17 and 18). In the SOS-bound H-Ras-GTP structure, H-Ras site 78 just C-terminal to switch II has a potential H-bond with site 112 (2.911Å) and potential van der Waals' contacts with sites 7 (3.659Å) and 104 (3.769Å). Site 75, the putative switch II hinge (Díaz et al. 2000), has a potential H-bond with site 104 (2.609Å) and a potential van der Waals' contact with site 7 (3.632Å) (Figure 17; PDB ID: 1BKD; Boriack-Sjodin et al. 1998). In effector bound H-Ras and Ral proteins, there are also several interactions of sites 75 and 78 with other cluster sites. In effector bound H-Ras, there are potential H-bonds between site 75 and site 104 (2.854Å) and between site 78 and site 112 (3.152Å). Potential van der Waals' contacts are between sites 7 and 75 (3.842Å), 7 and 78 (3.458Å) and 104 and 78 (3.732Å) (PDB ID: 1HE8; Pacold et al. 2000). In effector bound Ral, there is a potential H-bond between site 78 and site 112 (3.057Å). Potential van der Waals' contacts are between sites 7 and 75 (3.275Å), 7 and 78 (3.748Å) and 104 and 78 (3.748Å) (PDB ID: 1UAD; Fukai et al. 2003). The interactions present in the effector-bound proteins remain in H-Ras-GDP and Ral-GDP (PDB ID: 1IOZ; Kigawa et al. 2002; PDB ID: 1U8Z; Nicely et al. 2004).

Factor 5 is composed of sites 13, 38, 40, 41, 47, 55, 56, 66, 68, 75, 78, 99 and 104. These sites discriminate the Rheb subfamily. In addition, they are all important for the maintenance of protein structure. All sites except sites 13, 47, 68 and 99 are part of the two switch anchoring clusters discussed above. Site 13 is in the phosphate binding loop. It has potential H-bonds with GTP (3.158Å) and GDP (3.108Å) (PDB ID: 1XTS; Yu et al. 2005). Sites 47, 68 and 99 have unknown functional and structural roles.

Finally, factor 6 has five sites: 81, 123, 143, 152 and 156. These sites discriminate the Ras subfamily protein in *Geodia cydonium*. GTP-bound H-Ras (PDB ID: 121p; Wittinghofer et al. 1991) was used as a structural model for the Ras protein in *G cydonium*. In H-Ras, V152 and F156 have a potential main-chain H-bond in α5 (3.142Å) which may contribute to L10 stability and thus promote GDP binding. R123 has two H-bonds with E143 (2.712Å, 3.097Å) and may help position A146 to H-bond to GDP. The structural and functional role of V81 is unknown.

In conclusion, common factor analysis was a useful tool for determining phylogenetic relationships within the Ras family. Common factor analysis was also a useful tool for discovering the relationships among the structural, functional and phylogenetic components of sequence covariation in the Ras family.

Conclusions

The multivariate analyses performed herein gave insight into the structural, functional, and phylogenetic components of amino acid covariation in the Ras family. Decision tree analyses and common factor analysis gave insights into the phylogenetic component of covariation, as they were able to accurately divide the Ras family into the clades of the Bayesian phylogenetic tree in Dellinger and Atchley (2005). Sequence covariation among amino acid sites thus was found to have a strong phylogenetic component.

Common factor analysis elucidated the role of the structural and functional components of amino acid covariation as well as the phylogenetic component. The sites in each factor not only discriminated a clade but also had a common functional or structural role

in the proteins of the clade. Proteins within each clade are also subfamilies that share common function, so it is not unexpected that functional and structural components of covariation are confounded with the phylogenetic component.

Structural alignments showed that structure covaries with sequence and function in the switches. The interaction between the structural and functional components of variation in the nucleotide binding regions kept the variation of sequence and structure small. DTAs showed that variation in function among Ras family proteins could be discriminated using sites with known structural or functional components to their variation. Additional inquiries into the covariation of sequence, structure and function could be made using other multivariate analyses such as common factor analysis of sites' physiochemical properties.

AD thanks Heather Dellinger for her encouragement during this study. This work was supported by an NIH Genomics IGERT fellowship, the NCSU Functional Genomics fellowship, and WRA's NIH grant.

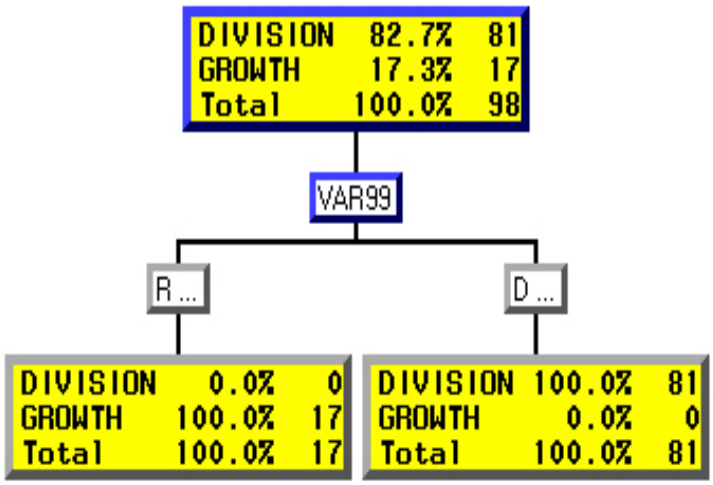


Figure 2.1. Decision Tree Between Cell-Growth and Cell-Division Controlling Proteins of the Ras Family.
The variable VAR99 corresponds to Ras numbered site 69.

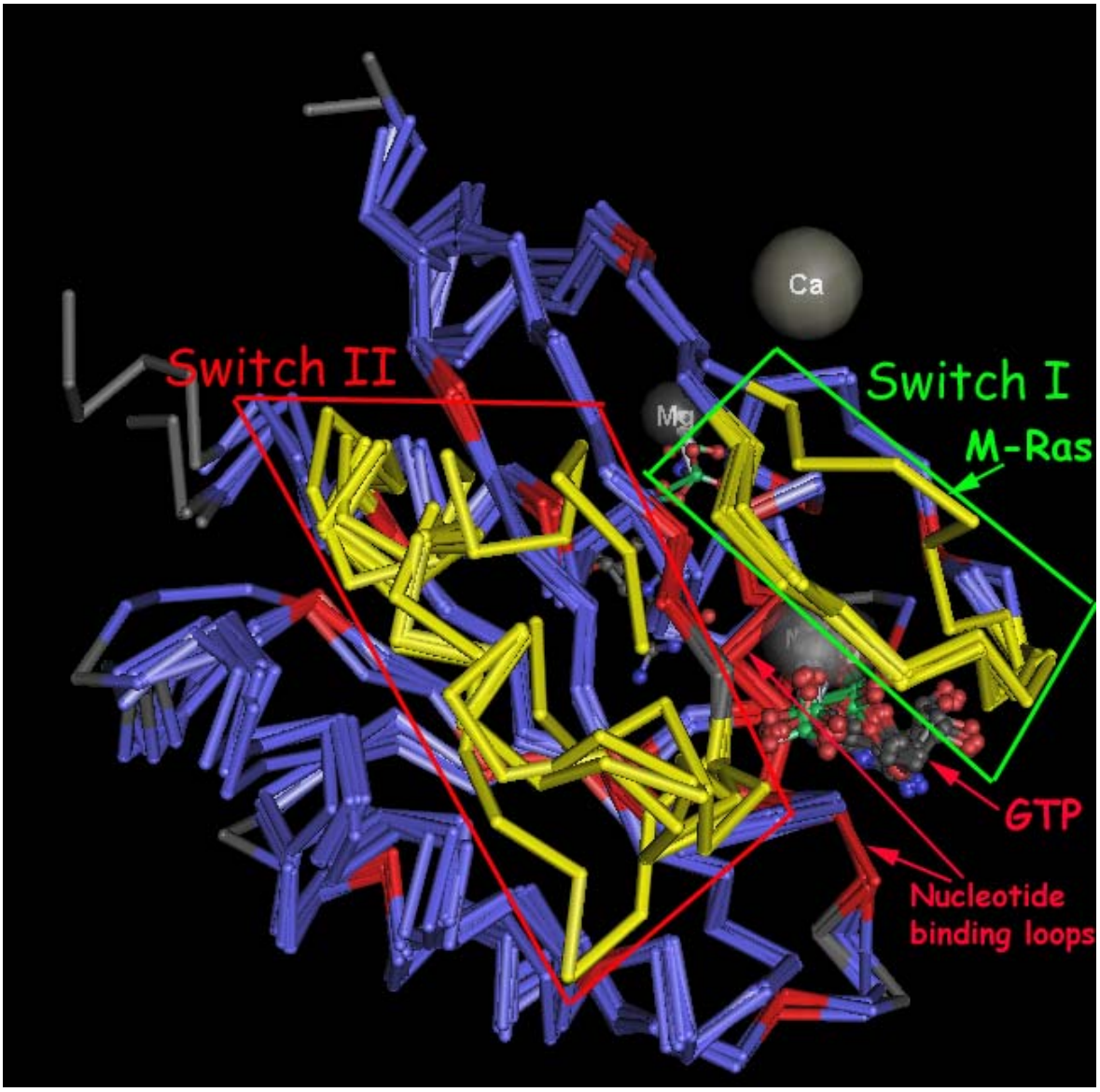


Figure 2.2. VAST Alignment of GTP-bound Ras Family Structures.
VAST alignment of GTP-bound H-Ras (121p; Wittinghofer et al. 1991), M-Ras (1X1S; Ye et al. 2005), Rap2a (2RAP; Cherfils et al. 1997), Rap1a (1C1Y; Nassar, 1995), Ral (1U8Y; Nicely et al. 2004) and Rheb (1XTS; Yu et al. 2005) structures. Switch II is the major region of structural differences. Nucleotides, Mg^{2+} and Ca^{2+} ions overlap in the same way as the protein structures.

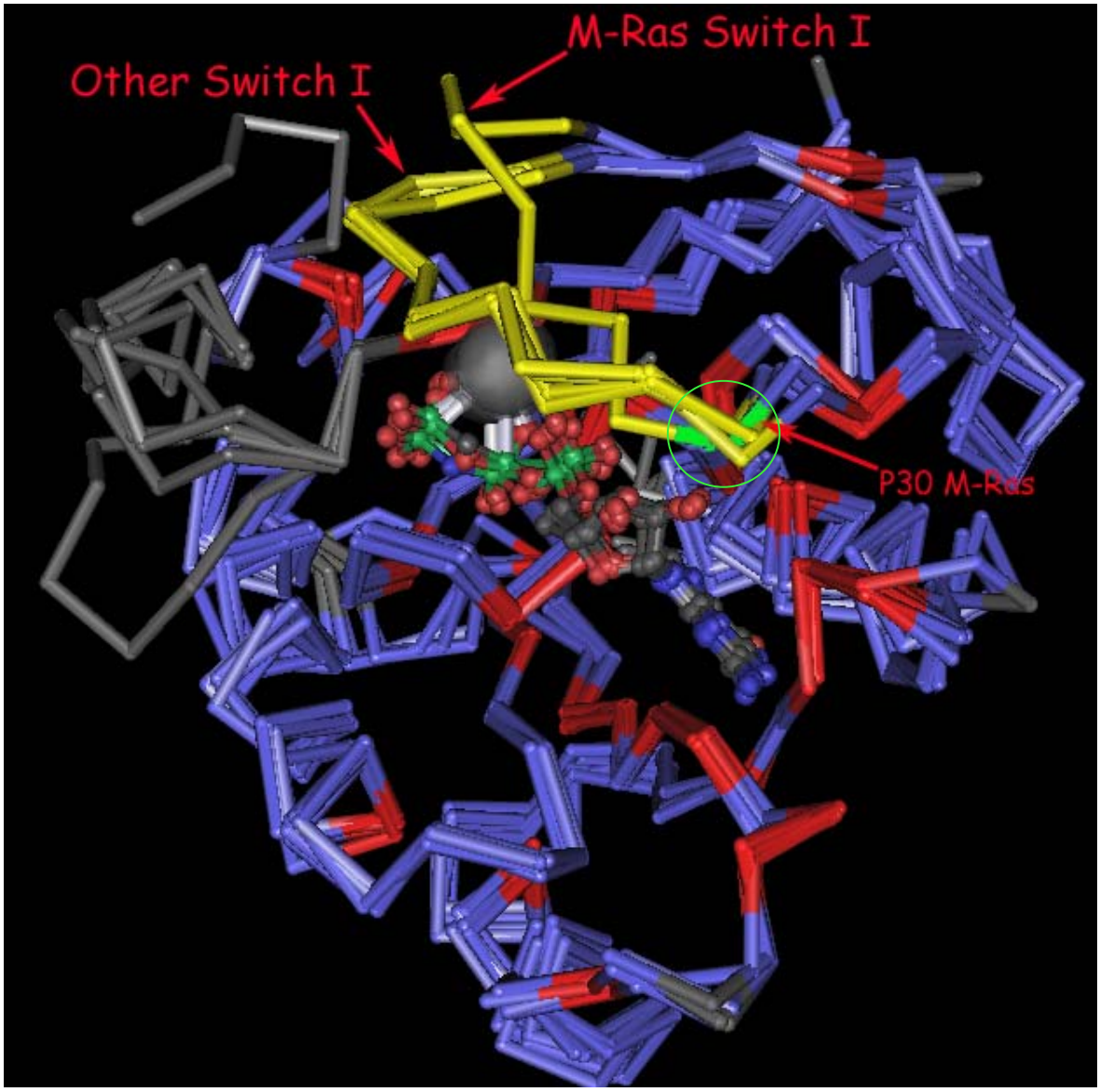


Figure 2.3. P30 in M-Ras changes Switch I conformation.
Switch I residues 30-38 are highlighted in yellow with the exception of M-Ras P30. M-Ras residue P30 is highlighted and circled in green. Structural variation in M-Ras switch I proceeds C-terminally from P30.

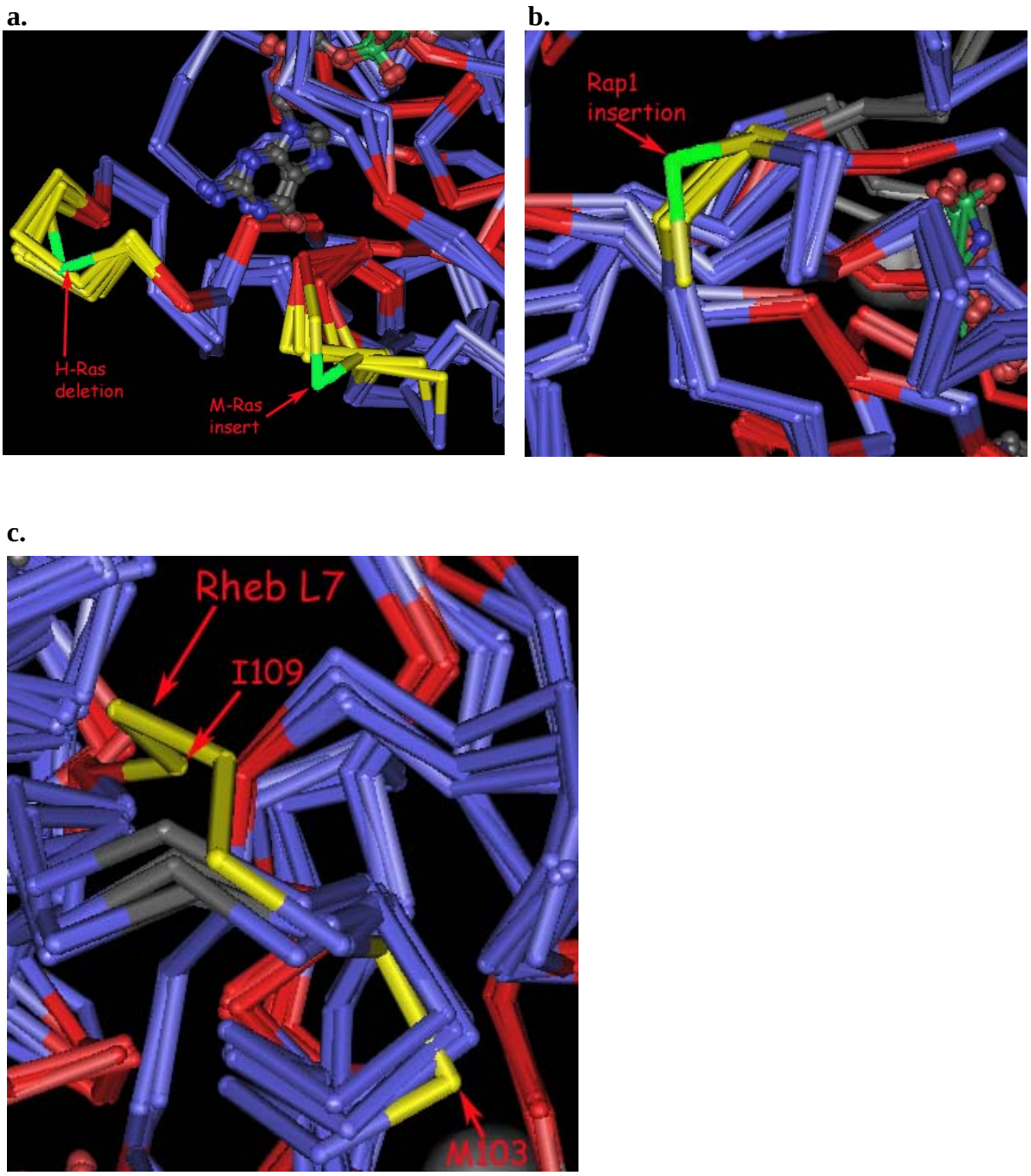


Figure 2.4. L7 and Indels Cause Variation in the VAST alignment of GTP-bound Ras Proteins.

a. M-Ras insertion in L10 between Ras numbered sites 141 and 142. H-Ras deletion after site 124. **b.** Rap1a insertion in L9 between Ras numbered sites 141 and 142. **c**. L7 is the most variable loop. Rheb's L7 is depicted in yellow

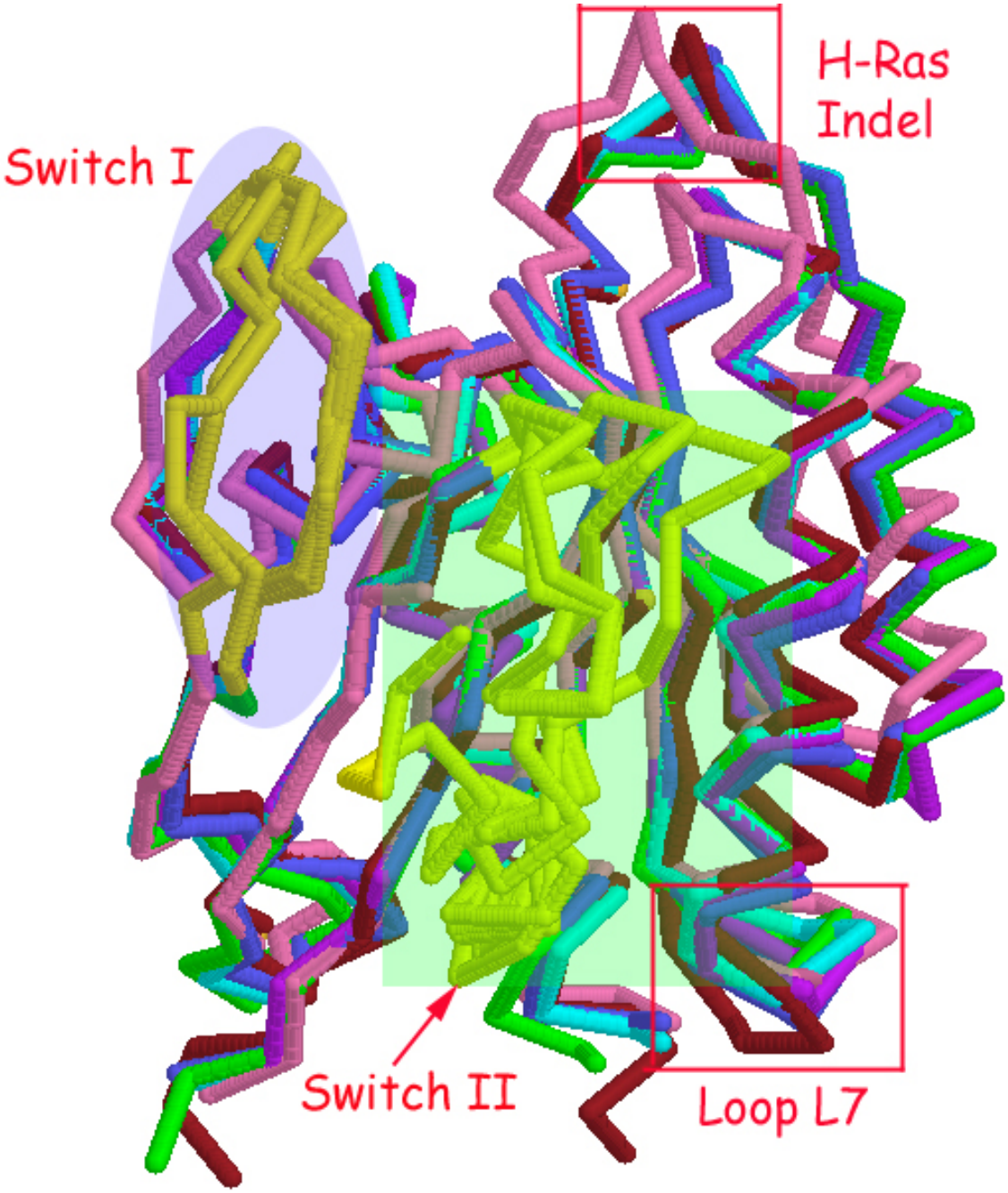


Figure 2.5. VMD Alignment of GTP-bound Ras Family Structures.
VAST alignment of GTP-bound H-Ras (121p; Wittinghofer et al. 1991), M-Ras (1X1S; Ye et al. 2005), Rap2a (2RAP; Cherfils et al. 1997), Rap1a (1C1Y; Nassar, 1995), Ral (1U8Y; Nicely et al. 2004) and Rheb (1XTS; Yu et al. 2005) structures. Structural differences are present in the switches, indels, and L7.

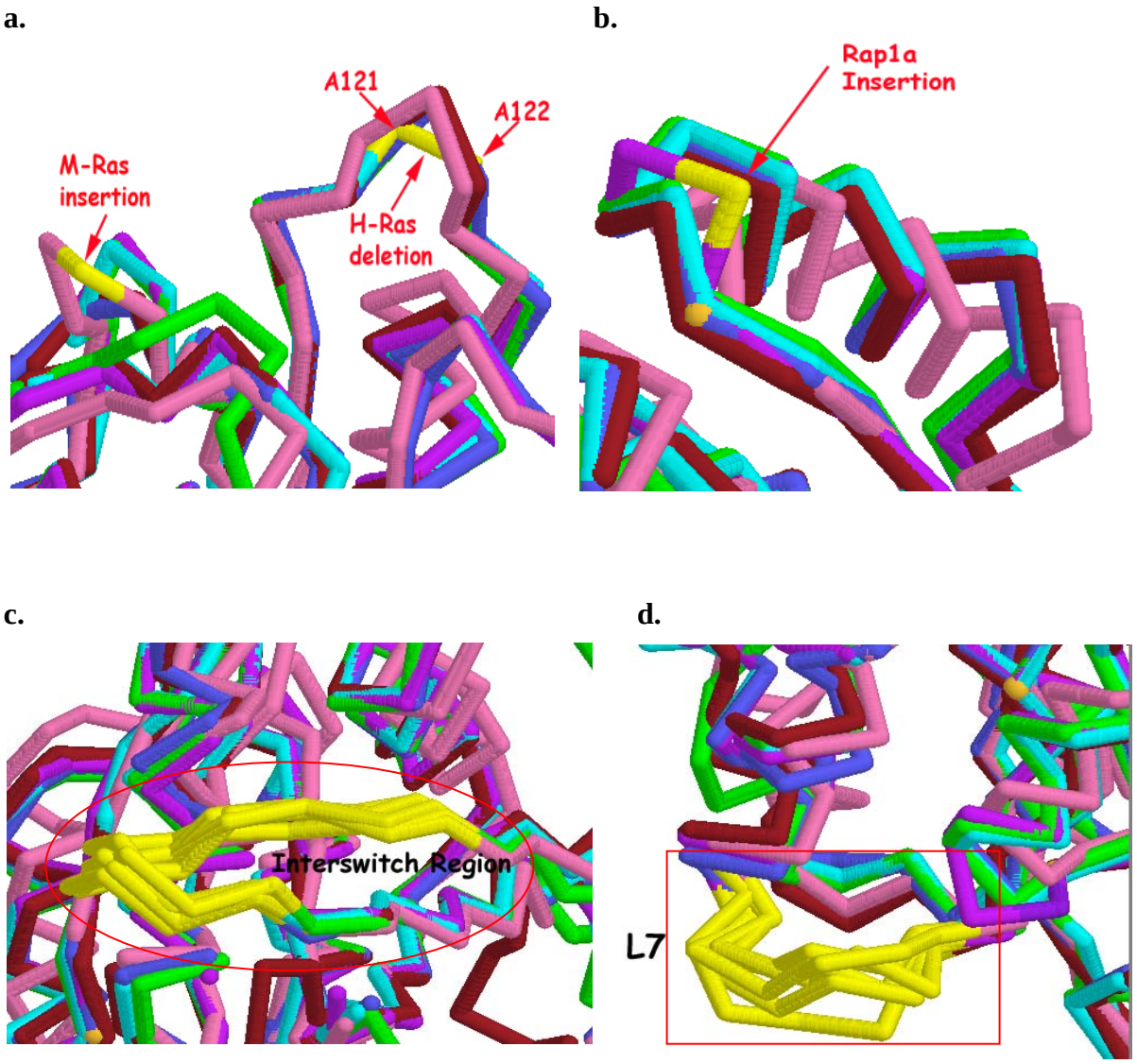


Figure 2.6. Structural Divergence in the VMD alignment of GTP-bound Ras Proteins. *Notes*: Each site in *c.* and *d.* has at least one pair of residues ≥ 1.5Å apart. For a detailed list of these pairs, see Table 1. **a**. A M-Ras insertion between sites 147 and 148 and a H-Ras deletion contribute to structural variation in nucleotide binding loops. **b**. A Rap1a insertion causes a shift in the position of Rap numbered sites 139 and 140. **c.** Sites 42-50 in the interswitch region **d.** Sites 104-109 of L7.

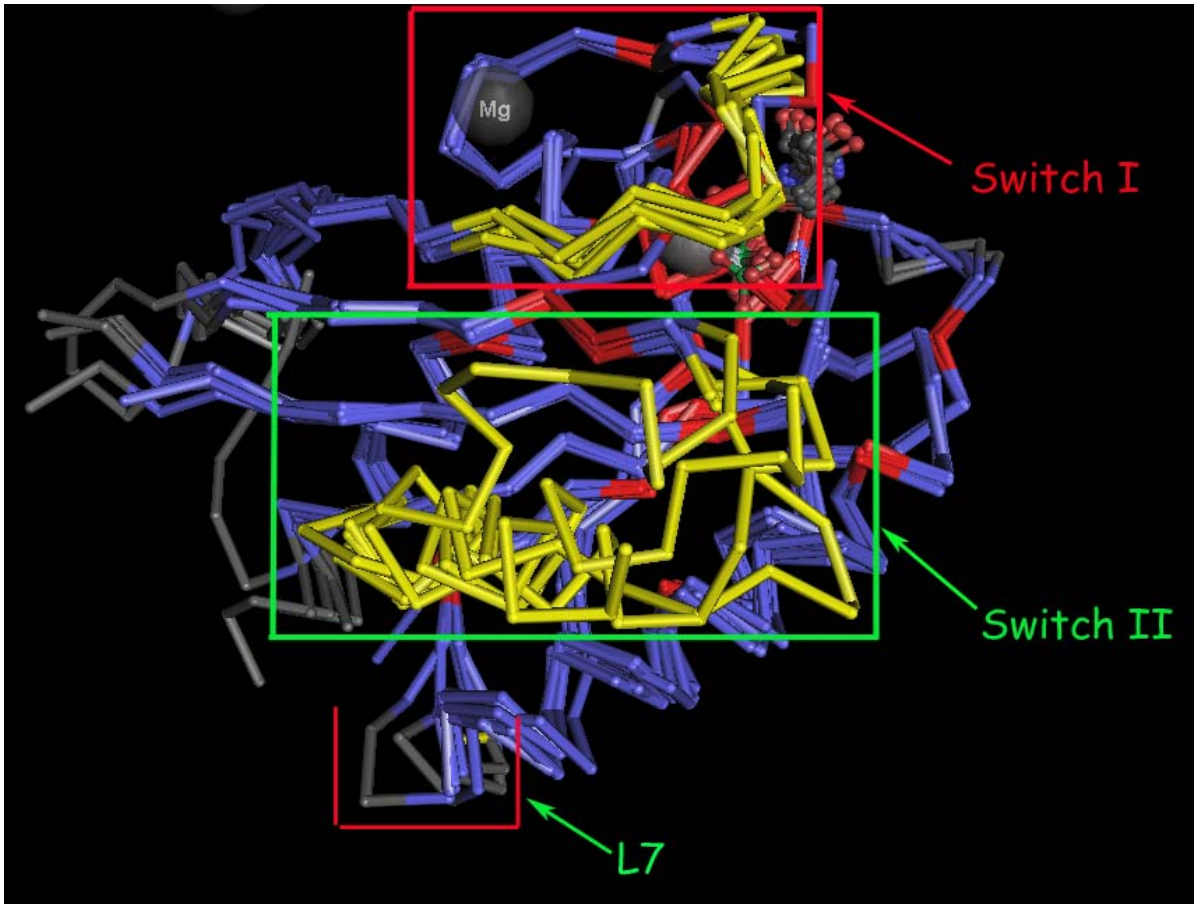
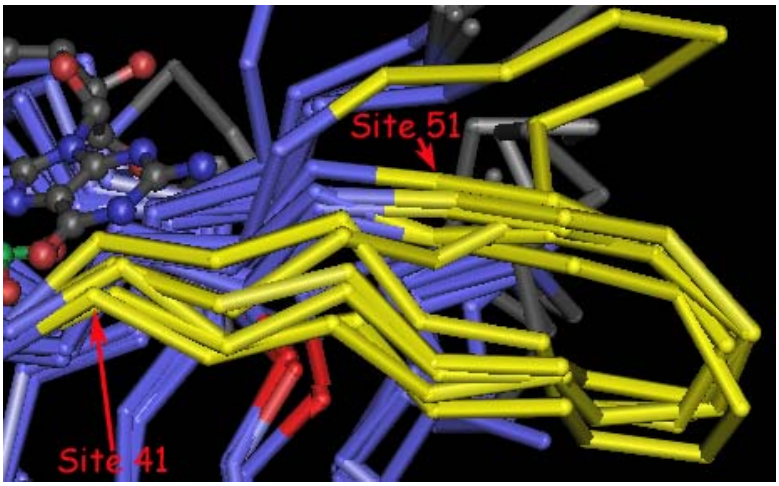
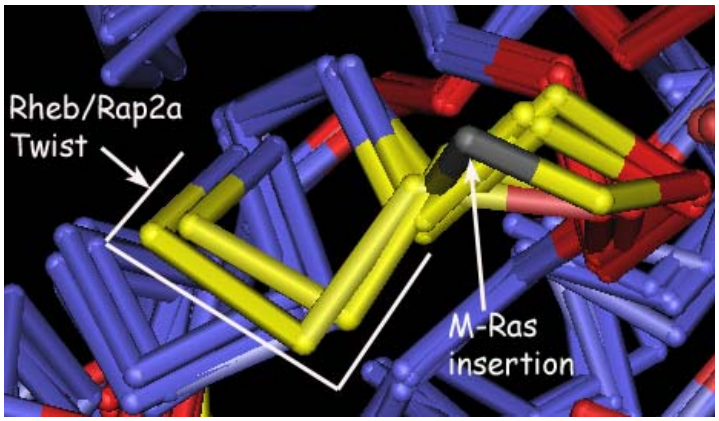


Figure 2.7. VAST Alignment of GDP-bound Ras Family Structures.
GDP-bound R-Ras2 (2ERY; unpublished), M-Ras (1X1R; Ye et al. 2005), H-Ras (1IOZ; Kigawa et al. 2002), Ral (1U8X; Nicely et al. 2004), Rap2a (1KAO; Cherfils et al. 1997), Rheb (1XTQ; Yu et al. 2005) and Di-Ras2 (2ERX; unpublished) structures were aligned using VAST (Gibrat et al. 1996; Madej et al. 1995). Major regions of structural divergence include the switch II region and loop L7 pictured here. Switch I is fairly conserved among Ras family structures.

a.



b.



c.

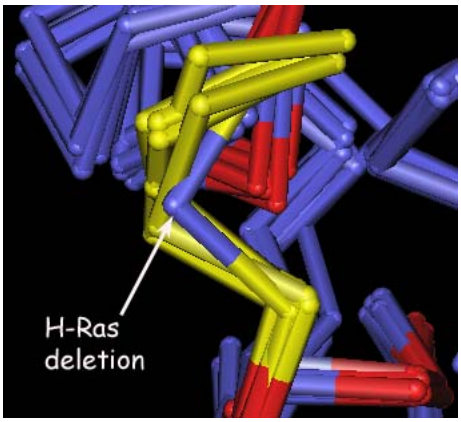


Figure 2.8. Sites of Structural Variation in VAST Aligned GDP-bound Ras Proteins.
a) Variability in Sites 41-51 of the Interswitch Region b) Insertion in M-Ras and a Different Twist c) Deletion in H-Ras

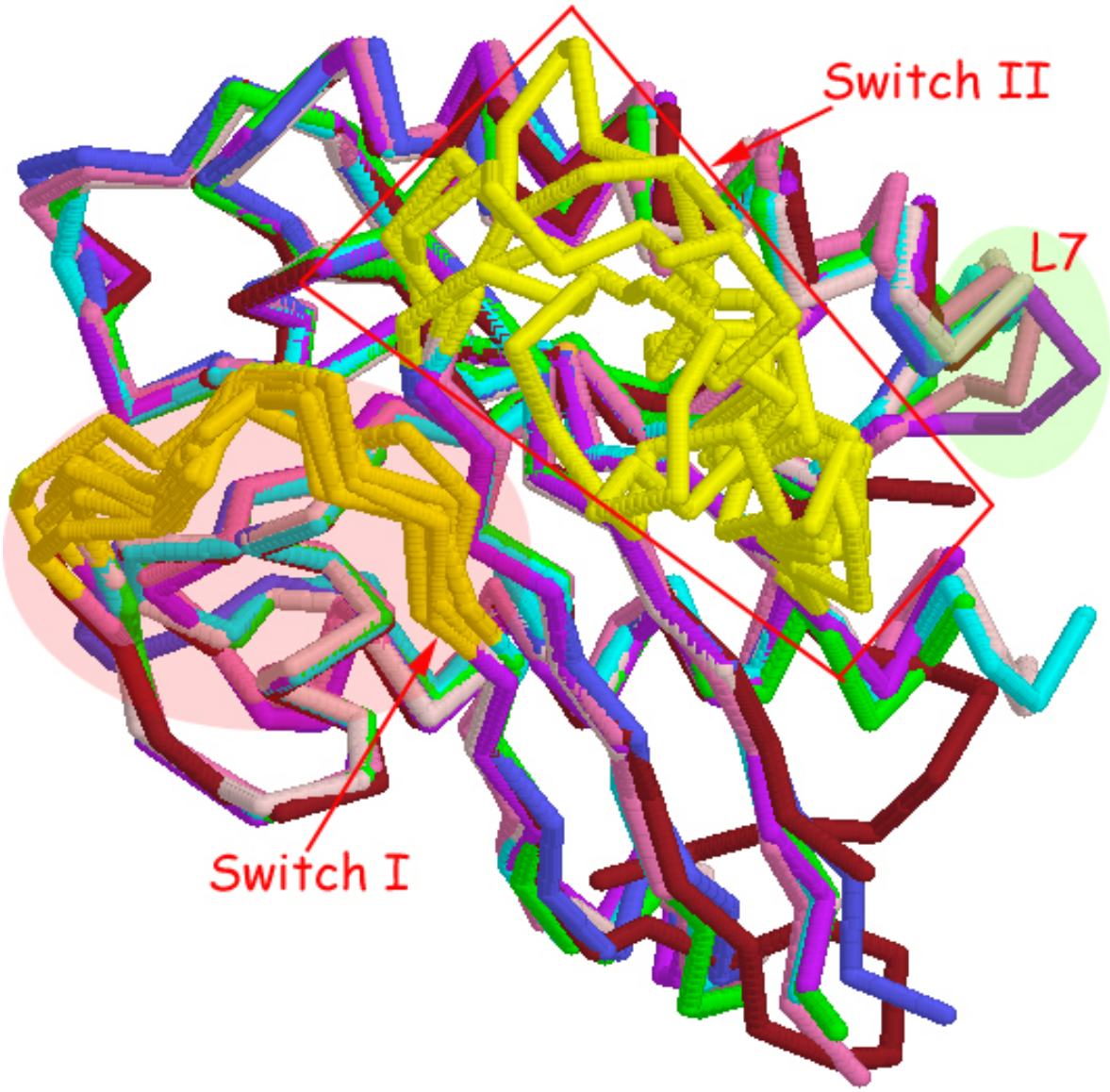


Figure 2.9. VMD Alignment of GDP-bound Ras Family Structures.

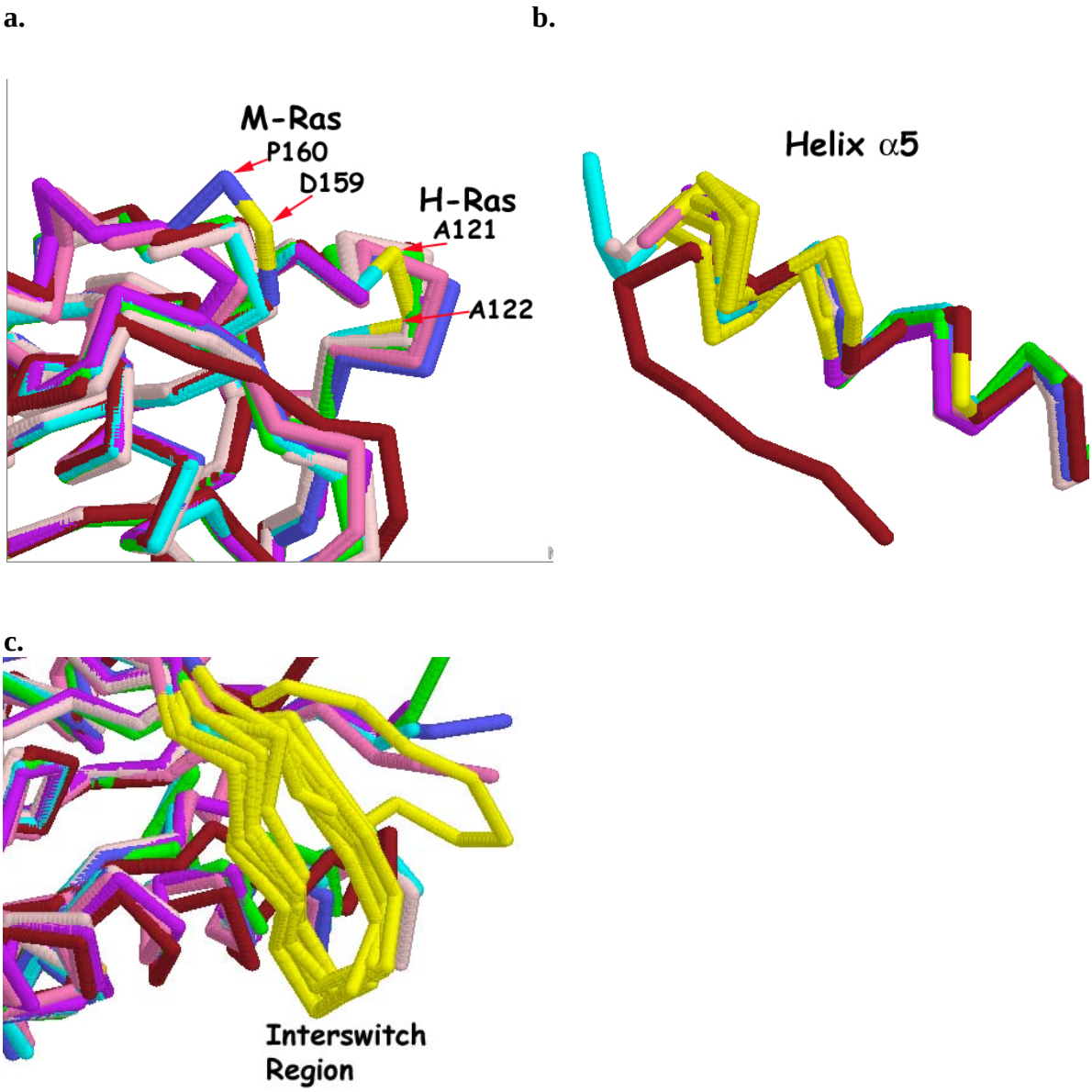


Figure 2.10. Sites of Variation in the VMD Alignment of GDP-bound Ras Proteins. a) M-Ras Insertion and H-Ras Deletion b) Structural Variability in α5 c) Variation in Sites 42-52 of the Interswitch Region.

Notes: **a.** An insertion at M-Ras numbered site 159 causes displacement sites 159 and 160. A deletion in H-Ras between sites 121 and 122 increases variability within those sites. **b.** Pairs of significantly distant residues in α5 are highlighted in yellow. **c.** At least one pair of residues in each highlighted site are ≥ 1.5Å apart.

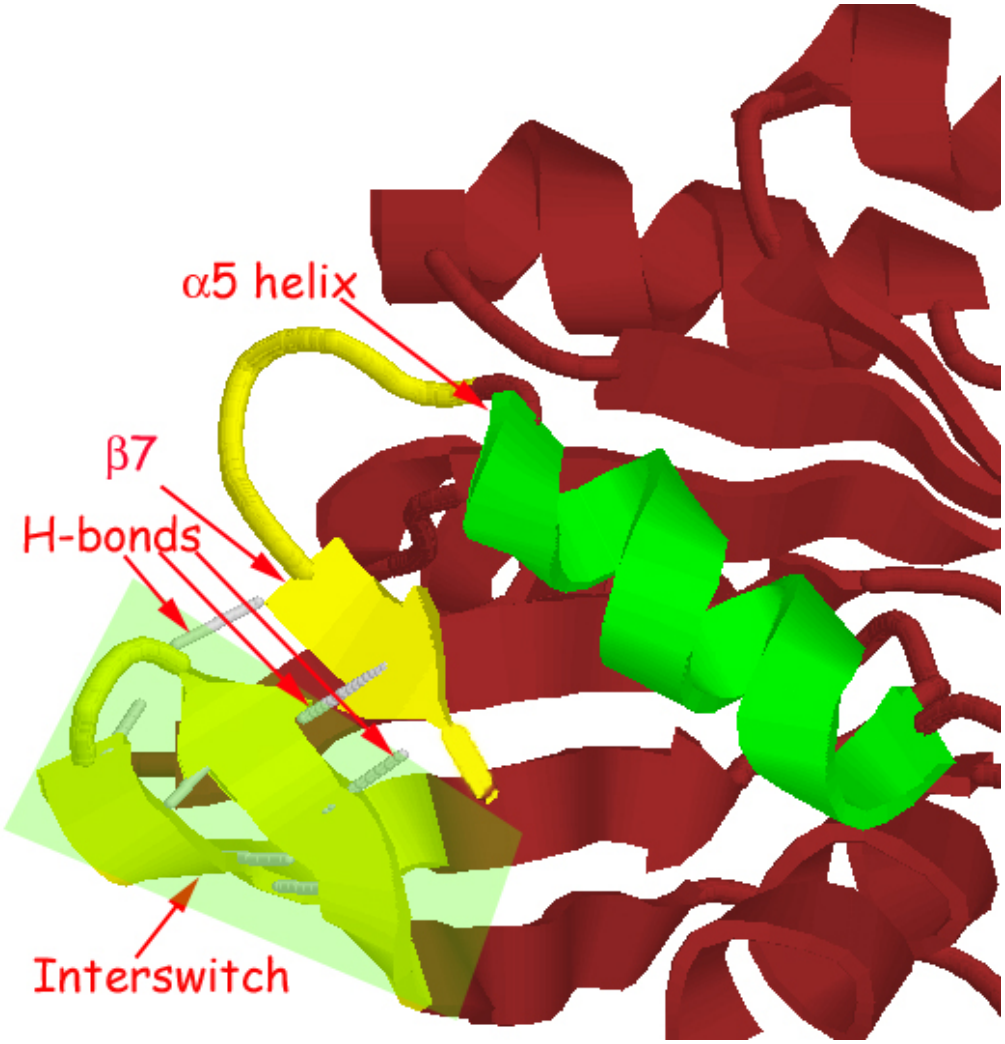


Figure 2.11. β2-β7 Interaction in the Di-Ras2-GDP structure.
The Di-Ras2 structure C-terminal to α5 adopts a unique conformation and performs a unique function. Sites colored yellow are greater than 1.5A from related sites in other Ras family structures. The interaction of β7 with β2 in the interswitch region is similar to the interaction of the N-terminal helix with β2 in Arf proteins.

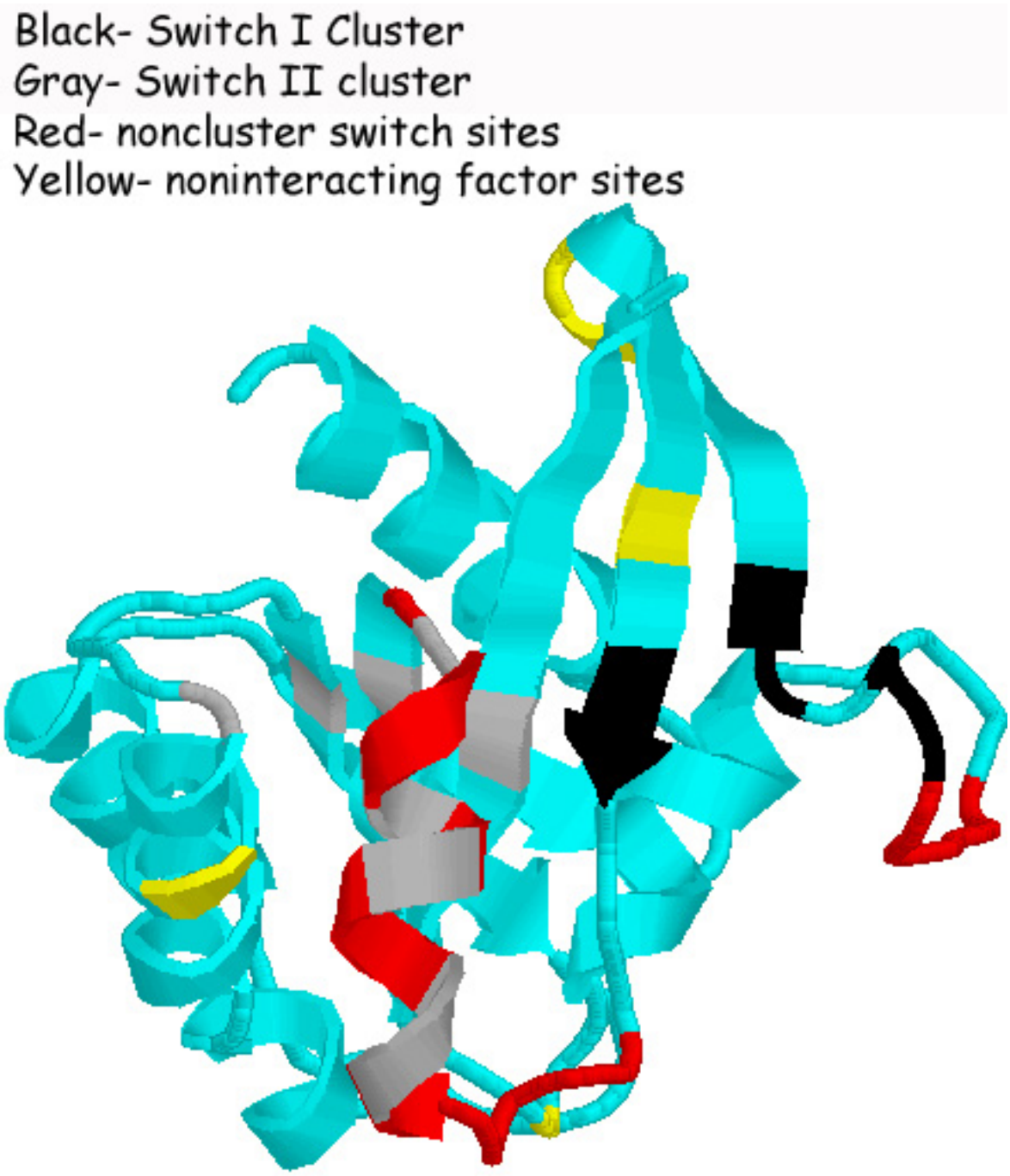


Figure 2.12. Locations of Factor 4 and Factor 5 Sites in H-Ras-GTP.
(PDB ID: 1BKD; Boriack-Sjodin et al. 1998).

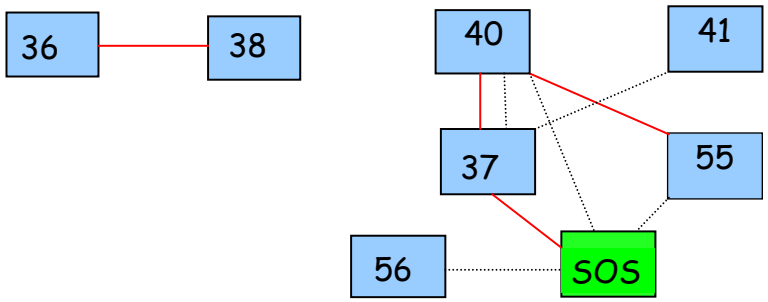


Figure 2.13. Physiochemical Interactions in the Switch I Cluster of H-Ras-GTP.
Solid Red Line = Potential H-bond
Dashed Black Line = Potential van der Waals' contacts
(PDB ID 1BKD; Boriack-Sjodin et al. 1998)

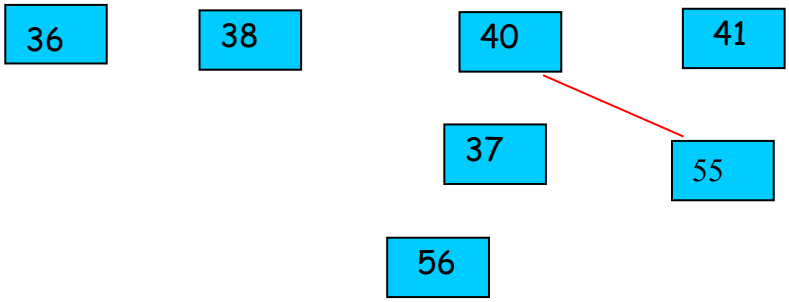


Figure 2.14. Physiochemical Interactions in the Switch I Cluster of H-Ras-GDP.
Solid Red Line = Potential H-bond
Dashed Black Line = Potential van der Waals' contacts
(PDB ID 1IOZ; Kigawa et al. 2002)

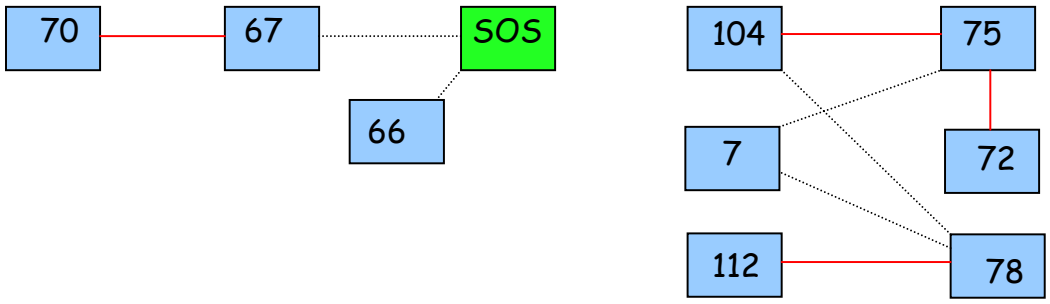


Figure 2.15. Physiochemical Interactions in the Switch II Cluster of H-Ras-GTP.
Solid Red Line = Potential H-bond
Dashed Black Line = Potential van der Waals' contacts
(PDB ID 1BKD; Boriack-Sjodin et al. 1998)

Figure 2.16. Physiochemical Interactions in the Switch II Cluster of H-Ras-GDP.
Solid Red Line = Potential H-bond
Dashed Black Line = Potential van der Waals' contacts
(PDB ID: 1IOZ; Kigawa et al. 2002)

Table 2.1. Solved Ras Family PDB Structures Used in the Analyses.
Notes: 1) The ligand, if any, is not contained in the PDB file.

Protein	**PDB ID**	**Reference**
H-Ras-GTP	121P	Wittinghofer et al. 1991
H-Ras and SOS[1]	1BKD	Boriack-Sjodin et al. 1998
H-Ras G12V-GNP and PI3Kγ	1HE8	Pacold et al. 2000
H-Ras-GNP and Bry2 Ras Binding Domain	1K8R	Scheffzek et al. 2001
Ras E31K-GNP and RalGDS	1LFD	Huang et al. 1998
Rap1a-G5'TP and C-Raf1 Kinase Ras Binding Domain	1C1Y	Nassar, 1995
Rap2a-G5'TP	2RAP	Cherfils et al. 1997
M-Ras-GNP	1X1S	Ye et al. 2005
Ral-GNP	1U8Y	Nicely et al. 2004
Ral-GNP and Sec5	1UAD	Fukai et al. 2003
Ral-GNP and Exo84	1ZC3	Jin et al. 2005
Rheb-G5'TP	1XTS	Yu et al. 2005
H-Ras-GDP	1IOZ	Kigawa et al. 2002
H-Ras-GDP and P120GAP	1WQ1	Scheffzek et al. 1997
Rap2a-GDP	1KAO	Cherfils et al. 1997
M-Ras-GDP	1X1R	Ye et al. 2005
R-Ras2-GDP	2ERY	*Unpublished*
Ral-GDP	1U8Z	Nicely et al. 2004
Rheb-GDP	1XTQ	Yu et al. 2005
Di-Ras2-GDP	2ERX	*Unpublished*

Table 2.2. Sites Distinguishing Growth and Division Controlling Proteins.
The sites are numbered according to the Ras standard in Valencia et al. (1991). Amino acid residues are represented by their one-letter IUPAC code.

Site	Growth Protein Residues	Division Protein Residues	Accuracy
62	D, H, Q	D, E, Q, A	97.96%
69	V, I, Q, R, S	D, E, M	100%
73	A, L, I, M, S, T	R, K, Q, W	100%
105	G, S	A, S, D, E, N, R, H, P	98.98%
149	D, N, S	R, K, Q, N, A, P	97.96%

Table 2.3. Initial DTA to Determine the Functional Group Membership of Unknown Proteins.

The entropy method was used. Sites are numbered according to the Ras family standard (Valencia et al. 1991). The decisions necessary to reach each functional group's leaf or leaves in the decision tree are listed in the Sites(Residues) column. The Accuracy column is calculated for each leaf as the number of correctly predicted proteins divided by the total number of proteins. **Notes**: 1) The actual division of residues in the tree covers all of the residues in the node. For example, the site 73 node has K on the Rap branch and R,S,T,A,L,I,M,W on the branch with Ral, DexRas, Di-Ras, Rheb and Distant Ras proteins. 2) Some sites may appear twice. 3) The 19th HNK protein is in the DexRas leaf.

Functional Group	**Sites (Residues)[1]**	**Accuracy**
Rap	101(LIV)→73(K)	12/12
Ral	101(L)→73(R)→69(D)→7(I)	8/8
Rit2 Rit1	101(F)→73(R)→69(E)→31(Y)→6(V) 101(Y)→73(R)→69(D)→7(V)→105(R)→29(P)→7(V)→30(E)[2]	2/2 2/2
DexRas	101(L)→73(L)→69(R)→7(V)→105(S)→29(E)→8(V)→30(D)	4/5[3]
Rheb	101(LV)→73(TLIM)→69(RQSIV)→7(A)→105(G)→29(VP)→20(T)→7(A)	6/6
Di-Ras	101(VC)→73(SA)→69(R)→7(AV)→105(G)→29(R)	7/7
M,R-Ras	101(L) →73(R)→69(E)→31(D)→6(L)	5/5
H,N,K-Ras	101(K)	18/18
Fungal Ras	101(RQLCA)→73(R)→69(E)→31(E)	13/13
Distant Ras	101(LI)→73(RW)→69(DM)→7(V)→105(D)	7/7

Table 2.4. Pathways of the Entropy Model DTA.
The entropy method was used. Sites are numbered according to the Ras family standard (Valencia et al. 1991). The decisions necessary to reach each functional group's leaf or leaves in the decision tree are listed in the Sites(Residues) column. The Accuracy column is calculated for each leaf as the number of correctly predicted proteins divided by the total number of proteins. **Notes**: 1) The actual division of residues in the tree covers all of the residues in the node. For example, the site 73 node has K on the Rap branch and R,S,T,A,L,I,M,W on the branch with Ral, DexRas, Di-Ras, Rheb and Distant Ras proteins. 2) Some sites may appear twice. 3) The leaf containing the Rit clade also contains a fungal Ras protein.

Functional Group	**Sites (Residues)[1]**	**Accuracy**
Rap	88(QNSI)→73(K)	12/12
Ral	88(E)→73(R)→31(D)→7(I)	8/8
Rit	88(RQ)→73(R)→31(DY)→7(V)→9(L)→21(M)→20(T) →7(V)[2,3]	4/5
DexRas	88(DE)→73(L)→31(QA)→7(V)→9(L)→21(S)→20(V)	4/4
Rheb	88(SA)→73(TLIM)→31(RS)→7(A)→9(LIM)→18(LIV)→ 20(T)→7(A) 88(K) →6(VI)	 2/2 4/4
Di-Ras	88(Q)→73(S)→31(ST)→7(AV)→9(F)→21(L) 88(E) →73(A)→31(E)→69(R)→5(R)	6/6 1/1
M,R-Ras	88(QSAG) →73(R)→31(D)→7(V)→9(V)	7/7
H,N,K-Ras	88(K)→6(ML)	25/25
Fungal Ras	88(HENQSL)→73(AR)→31(E)→69(E) 88(S)→73(R)→31(T)→7(V)→9(L)→21(I)→20(T)→7(V)	15/15 1/1
Distant Ras	88(QSTA)→73(RW)→31(E)→69(DM)→5(K)	8/8

Table 2.5. Pathways of the Multifurcating DTA Generated by Eye.
Sites are numbered according to the Ras family standard (Valencia et al. 1991). The decisions necessary to reach each functional group's leaf or leaves in the decision tree are listed in the Sites(Residues) column. The Accuracy column is calculated for each leaf as the number of correctly predicted proteins divided by the total number of proteins .

Functional Group	**Sites (Residues)**	**Accuracy**
Rap	33(D)→39(SF)→70(L)	12/12
Ral	33(E)→39(S)→70(N)	8/8
Rit	33(D)→39(A)	2/2
DexRas	33(T)	4/4
Rheb	33(DEY)→39(EQT)	6/6
Di-Ras	33(I)	7/7
M,R-Ras	33(D)→39(S)→70(EQ)→30(QSTP)	6/6
H,N,K-Ras	33(D)→39(SI)→70(HQ)→30(DE)→26(N)→83(A)	25/25
Fungal Ras	33(D) →39(S)→70(RQ)→30(DE)→26(ESG)	16/16
Distant Ras	33(D)→39(S)→70(DQ)→30(DEA)→26(NQ)→83(DNS)	8/8

Table 2.6. Sites of Structural Differences in VMD Aligned GTP-bound Ras Proteins. Notes: 1) Because of indels between sites 104 and 105, 121 and 122, 138 and 139 and 147 and 148, sites are numbered by the representative PDB structure of the subfamily. Sites in each row are numbered by the PDB numbering of the protein listed in the row's first column. H-Ras and Rap2 are numbered by the family standard (Valencia et al. 1991). Site numbering N-terminal to the indels is equivalent in Rap1, 11 greater in Ral and R-Ras2 (i. e. site 5 in H-Ras is site 16 in Ral), 10 greater in M-Ras, 4 greater in Di-Ras2 and 3 greater in Rheb. 2) M-Ras sites 69-73 are missing from the structure.

Subfamily	**Differing sites (>1.5Å)[1]**
M-Ras[2]	Rap1- 27-29, 31-49, 52, 57, 75-81, 84, 95, 114, 118, 127-149, 151-167, 170-172; Rap2- 27-29, 31-32, 35-48, 52-54, 56-57, 75-81, 84, 95-96, 105, 108, 116-118, 126-149, 154-162; H-Ras- 28, 30-48, 74-81, 84-86, 127-139, 155-167, 169-170, 177-178; Ral- 27-29, 33-44, 46-49, 56-57, 59, 75-81, 94-98, 117-118, 127-149, 151-167, 171-175; Rheb- 28-49, 52-58, 74-82, 94-96, 106, 109-112, 113-114, 114-119, 121-122, 127-162
Rap1	Rap2- 62, 64-66, 95, 98, 139-140, 150-153; Ral- 29, 31-38, 42, 61-68, 70, 105-108, 129, 131, 139-140, 150, 152; Rheb- 61-69, 71-72, 74, 104, 106-109, 111, 139-140, 150-153; M-Ras- 17-19, 21-39, 42, 47, 65-71, 74, 85, 104, 108, 117-140, 142-157, 160-162; H-Ras- 32, 62, 121-122, 139-140
Rap2	Rap1- 62, 64-66, 95, 98, 138, 148-151; M-Ras- 17-19, 21-22, 25-38, 42-44, 46-47, 65-71, 74, 85-86, 95, 98, 106-108, 116-139, 144-151; H-Ras- 32, 44, 62, 64-66, 95, 98, 106-108, 121-122, 148-152; Ral- 29-38, 42, 46-49, 61, 63-64, 66-68, 70-71, 95, 98, 149-153, 167-168; Rheb- 61-70, 72, 74, 95-109
H-Ras	M-Ras- 18, 20-38, 64-71, 74, 84-86, 117-138, 143-155, 157-158, 165-166; Rap1- 32, 62, 121, 138; Rap2- 32, 44, 62, 64-66, 95, 98, 106-108, 121, 147-151; Ral- 29-38, 61-68, 70, 105-108, 121, 164-166; Rheb- 32, 43-46, 50, 61-69, 71-74, 106-109, 148-151
Ral	M-Ras- 28-30, 34-45, 47-50, 57-58, 60, 76-82, 95-97, 118-119, 128-150, 152-167, 171-175; H-Ras- 40-49, 72-79, 81, 116-119, 132-133, 176-178; Rap1- 40, 42-49, 53, 72-79, 81, 116-119, 140, 142; Rap2- 40-49, 53, 57-60, 72, 74-75, 77-79, 81-82, 106, 109, 160-164, 178-179; Rheb- 40, 42-50, 53-61, 72-79, 82-83, 85, 114-120, 132-133, 138, 140, 152, 160-163, 173-178
Rheb	M-Ras- 21-42, 45-51, 67-75, 87-89, 99, 102-103, 106-107, 109-112, 114-115, 120-154; H-Ras- 35, 46-49, 53, 64-72, 74-77, 109-112, 125, 151-154; Rap1- 64-72, 74-75, 77, 107, 109-112, 114, 142, 152-155; Rap2- 64-73, 75, 77, 98-112; Ral- 32, 34-42, 45-53, 64-71, 73-75, 77, 106-112, 124-125, 130, 132, 144, 152-155, 165-170

Table 2.7. Sites of Structural Differences in VMD Aligned GDP-bound Ras Proteins.
Notes: 1) Because of indels between sites 104 and 105, 121 and 122, 138 and 139 and 147 and 148, sites are numbered by the representative PDB structure of the subfamily. Sites in each row are numbered by the PDB numbering of the protein listed in the row's first column. H-Ras and Rap2 are numbered by the family standard (Valencia et al. 1991). Site numbering N-terminal to the indels is equivalent in Rap1, 11 greater in Ral and R-Ras2 (i. e. site 5 in H-Ras is site 16 in Ral), 10 greater in M-Ras, 4 greater in Di-Ras2 and 3 greater in Rheb. 2) H-Ras sites 61-67 are missing from the structure. 3) Ral sites 72-74 are missing from the structure. 4) Di-Ras2 sites 110-112 are missing from the structure.

Subfamily	**Differing sites (>1.5Å)[1]**
M-Ras	H-Ras- 41-44, 48, 70, 78-81, 132-133, 159-160; Di-Ras2- 37, 39-42, 45-47, 54-62, 69-82, 102-103, 113-115, 118-120, 131-132, 159-160, 169, 175; Ral- 41, 60, 70, 75-81, 84, 115-118, 131- 132, 159-160; Rheb- 40-42, 46, 51-54, 61, 70-84, 109-110, 112-114, 117-118, 159-162; Rap2- 42-45, 52-56, 58, 60-62, 70-82, 84, 112-113, 116-118, 131-132, 158-162, 177-178; R-Ras2- 41, 51-55, 70-82, 84, 131-132, 159-160
R-Ras2	Di-Ras2- 40-44, 46-47, 54-63, 71-80, 104, 115-116, 120-121, 169, 174-176; Ral- 53, 57-61, 75-81, 116-119, 140, 176-178; Rheb- 40-42, 58-59, 71-86, 106-107, 110-111, 114, 118-119, 149, 159-162; H-Ras- 79-80, 132, 148; Rap2- 71-82, 117-119, 149, 159-162; M-Ras- 42, 52-56, 71-83, 85, 132-133, 160
Rap2	H-Ras- 33-34, 48-49, 60, 68, 70-71, 74, 106-108, 121, 147-148, 150-152; Di-Ras2- 27, 29-36, 43-52, 59-68, 105, 108-110, 139, 148-151, 160, 162-165; Ral- 32-34, 44, 46-50, 60, 64-71, 98, 102, 106-108, 148-152, 166; Rheb- 31-35, 47-48, 61-70, 72-76, 102, 106-108; M-Ras- 32-35, 42-46, 48, 50-52, 60-72, 74, 102-103, 106-108, 121-122, 148-151, 166-167; R-Ras2- 60-71, 106-108, 138, 148-151
H-Ras[2]	Rap2- 33-34, 48-49, 60, 68, 70-71, 74, 106-108, 121, 147-148, 150-151; Ral- 49, 105-107, 121; M-Ras- 31, 38, 41-44, 60, 68-71, 122, 148; Rheb- 30-31, 60, 68-74, 99-100, 103, 106-108, 122, 148-150; Di-Ras2- 27, 29-33, 35-36, 43-53, 60, 68, 121, 136, 157, 161-165; R-Ras2- 68-70, 121, 136
Ral[3]	Di-Ras2- 40-48, 55-63, 71, 75-82, 119-120, 169, 173-176; H-Ras- 60, 116-118, 132; Rap2- 43-45, 57-61, 71, 75-82, 109, 113, 117-119, 159-163, 177; M-Ras- 42, 61, 71, 76-82, 85, 116-119, 132-133, 160; R-Ras2- 53, 57-61, 75-81, 116-119, 140, 176-178; Rheb- 41, 71, 75-78, 80-85, 106, 110, 114-115, 117-119, 140, 144, 160-163, 174-178
Rheb	H-Ras- 33-34, 63, 71-77, 102-103, 106, 109-111, 124-126, 152-154; Di-Ras2- 32-36, 38-39, 45-55, 63-77, 97-99, 102-103, 105-108, 111-113, 124-125, 153-154, 167-168; Ral- 33, 63, 67-70, 72-77, 98, 102, 106-107, 109-111, 132, 152-155, 166-170; Rap2- 34-38, 50-51, 64-73, 75-79, 105, 109-111; M-Ras- 33-35, 39, 44-47, 54, 63-77, 102-103, 105-107, 110-111, 152-154; R-Ras2- 32-34, 50-51, 63-78, 98-99, 102-103, 106, 110-111, 141, 151-154
Di-Ras2[4]	H-Ras- 31, 33-37, 39-40, 47-57, 64, 72, 126-127, 142, 163, 167-171; R-Ras2- 33-37, 39-41, 47-56, 64-73, 97, 108-109, 113-115, 163, 168-170; Rap2- 31, 33-40, 47-56, 63-72, 109, 113-115, 144, 153-156, 165, 167-170; M-Ras- 31, 33-36, 39-41, 48-56, 63-76, 96-97, 107-109, 113-115, 126-127, 154, 163, 169; Ral- 33-41, 48-56, 64, 68-75, 113-114,163, 167-170; Rheb- 33-37, 39-40, 46-56, 64-78, 98-100, 103-104, 106-109, 113-115, 126-127, 154, 156, 169-170

Table 2.8. Equamax Rotated Factor Pattern of the Ras Family Alignment
Notes: 1) VARN, where N represents an integer, represents the site number in the Ras family multiple sequence alignment. **2)** The normal typeface numbers in parentheses represent the corresponding Ras numbered sites (Valencia et al. 1991). Numbers are only given to sites with factor coefficients greater than 0.6. **3)** Sites with italic numbers are Ral tree-determinant sites with factor coefficients less than 0.6.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
VAR162[1] (113)[2]	**0.67331**	0.29917	0.20788	0.12710	0.16015	0.26109
VAR182 (127)	**0.63677**	0.14901	0.48850	0.40075	0.13930	0.25143
VAR185 (130)	**0.63558**	0.21628	0.29159	0.29413	0.12159	0.30514
VAR186 (131)	**0.62460**	0.21907	0.49747	0.29944	0.22450	0.23773
VAR124 (94)	**0.62328**	0.16428	0.59166	0.36525	0.12264	0.19194
VAR128 (98)	**0.61756**	0.23731	0.44009	0.29611	0.31899	0.19637
VAR171 (121)	**0.61403**	0.16129	0.54895	0.35569	0.23556	0.25261
VAR163 (114)	**0.60984**	0.49802	-0.04861	0.00291	0.08426	0.44055
VAR191 (136)	**0.60955**	0.32958	0.49004	0.32157	0.10832	0.27213
VAR138 (108)	**0.60114**	0.30591	0.47087	0.30536	0.29764	0.23812
VAR187	0.59955	0.33976	0.46162	0.39969	0.17975	0.20175
VAR111	0.59440	0.03580	0.31174	0.24958	0.11306	0.21000
VAR117	0.59433	0.17744	0.56506	0.39783	0.10659	0.27455
VAR177	0.59248	0.17027	0.56196	0.41939	0.13506	0.21442
VAR204	0.59220	0.18153	0.13347	0.43201	0.17063	0.19148
VAR212	0.59177	0.28609	0.55997	0.31523	0.22446	0.23250
VAR125	0.58776	0.23318	0.58401	0.40230	0.15549	0.19787
VAR167	0.58647	0.28249	0.37430	0.36099	0.36413	0.30616
VAR189	0.58395	0.46692	0.06535	0.13019	0.21528	0.40729
VAR183	0.58218	0.20615	0.57205	0.34656	0.14250	0.25740
VAR121	0.58190	0.14326	0.49092	0.44058	0.08977	0.20485
VAR181	0.57821	0.27648	0.55495	0.36582	0.13170	0.20670
VAR113	0.57705	0.35832	0.19293	0.14652	0.14510	0.44857
VAR223	0.57520	0.09535	0.17930	0.11405	0.34778	0.10100
VAR73	0.56822	0.26030	0.38450	0.50501	0.27012	0.21949
VAR192	0.56341	0.30465	0.47526	0.41822	0.16534	0.28603
VAR190	0.55804	0.24986	0.43756	0.45937	0.24977	0.24321
VAR114	0.55130	0.31256	0.43264	0.32001	0.21363	0.36481
VAR202	0.55026	0.12305	0.48885	0.47400	0.16508	0.26778
VAR201	0.54812	0.42976	0.30522	0.17235	0.08046	0.20992
VAR184	0.54624	0.12594	0.51836	0.32266	0.12704	0.23049
VAR207	0.54597	0.45393	0.26434	0.22297	0.33726	0.30451
VAR49	0.53971	0.21261	0.46648	0.46167	0.05192	0.15062
VAR71	0.52088	0.45154	0.46315	0.34068	0.26158	0.20464

Table 2.8 (continued)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
VAR170	0.52081	0.34895	0.48739	0.35750	0.33544	0.25532
VAR221	0.52056	0.33401	0.46608	0.43867	0.29747	0.22421
VAR193	0.51889	0.46118	0.48236	0.27470	0.24768	0.23069
VAR203	0.50130	0.30669	0.49072	0.47427	0.18337	0.22509
VAR2	0.50026	0.47355	0.23303	0.45463	0.27598	0.24346
VAR64	0.49953	0.47430	0.44442	0.31675	0.28864	0.19170
VAR205	0.49778	0.33721	0.49033	0.43467	0.15449	0.30185
VAR219	0.49122	0.47825	0.42961	0.46120	0.14990	0.21651
VAR50	0.48150	0.43733	0.45892	0.43550	0.27507	0.20704
VAR18 (22)	0.25779	**0.86436**	0.14298	0.14104	0.23191	0.14281
VAR81 (59)	0.33376	**0.84619**	0.11658	0.14321	0.15819	0.21857
VAR169 (120)	0.29674	**0.83694**	0.16700	0.19272	0.20361	0.23731
VAR106 (77)	0.27093	**0.83433**	0.11141	0.17733	0.17754	0.32554
VAR16 (20)	0.24776	**0.81793**	0.13419	0.21973	0.33502	0.19429
VAR1 (5)	0.28056	**0.80771**	0.12424	0.18450	0.28346	0.16926
VAR100 (71)	0.28164	**0.74366**	0.21849	0.24547	0.32869	0.29043
VAR83 (61)	0.39470	**0.71862**	0.32462	0.29833	0.16033	0.23250
VAR225 (157)	0.33541	**0.68094**	0.18954	0.24275	0.48465	0.20812
VAR227 (163)	0.38215	**0.67068**	0.33614	0.28175	0.27487	0.26919
VAR120 (90)	0.29077	**0.66333**	0.19279	0.22008	0.08335	0.28784
VAR48 (29)	0.41161	**0.63846**	0.39611	0.17669	0.23775	0.31584
VAR91 (62)	0.28906	**0.63530**	0.30664	0.20708	0.46811	0.30232
VAR102 (73)	0.31835	**0.63392**	0.32269	0.34868	0.40745	0.25264
VAR52 (33)	0.31863	**0.63372**	0.20156	0.54393	0.28315	0.22475
VAR132 (102)	0.32038	**0.63361**	0.25254	0.25823	0.50183	0.21172
VAR133 (103)	0.31418	**0.61008**	0.17131	0.26630	0.56270	0.22336
VAR14	0.33007	0.59848	0.28466	0.26595	0.44958	0.09223
VAR5	0.10732	0.59052	0.49952	0.35401	0.35342	0.18940
VAR76	0.29451	0.58664	0.35205	0.38671	0.35580	0.20790
VAR137	0.39633	0.55604	0.36804	0.36716	0.33149	0.30228
VAR58	0.28724	0.55567	0.38945	0.24375	0.51586	0.26842
VAR213	0.44908	0.55265	0.41751	0.33278	0.31545	0.13195
VAR108	0.37176	0.54612	0.45347	0.29130	0.25562	0.21361
VAR105	0.48451	0.52502	0.38165	0.31989	0.30307	0.27111
VAR226	0.48594	0.52103	0.47851	0.33958	0.14323	0.22354
VAR98	0.40188	0.51990	0.32904	0.30238	0.45607	0.24744
VAR228	0.45987	0.51832	0.34940	0.33922	0.39637	0.17939
VAR17	0.35962	0.50870	0.44153	0.46260	0.19026	0.29591
VAR94	0.39280	0.50743	0.44850	0.49714	0.10395	0.26675
VAR93	0.10790	0.46384	0.45256	0.38152	0.29227	0.39308

Table 2.8 (continued)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
VAR82	-0.10584	0.43830	0.06886	0.07512	0.05754	0.01189
VAR80	-0.06495	0.41592	0.09217	0.23932	0.27949	0.21347
VAR168	0.01304	-0.02815	0.01774	-0.00405	-0.01006	-0.01616
VAR10	0.01818	-0.04771	0.01819	0.00484	0.04105	-0.03543
VAR131 (101)	0.26422	0.21495	**0.75416**	0.35075	0.21057	0.25851
VAR109 (80)	0.22012	0.11583	**0.74017**	0.23210	0.17921	0.26897
VAR45 (26)	0.26453	0.18596	**0.66810**	0.56086	0.13738	0.22931
VAR215 (151)	0.30262	0.09861	**0.64602**	0.34562	0.36026	0.32567
VAR126 (96)	0.27015	0.32801	**0.62789**	0.46120	0.26545	0.17941
VAR68	0.46787	0.39498	0.59456	0.33632	0.17535	0.15262
VAR214	0.50154	0.23217	0.58960	0.45042	0.20004	0.24307
VAR7	0.30947	0.16541	0.58187	0.52373	0.30137	0.18498
VAR127	0.48861	0.27857	0.58049	0.22896	0.27836	0.23417
VAR116	0.54403	0.11524	0.57819	0.47359	0.13410	0.22556
VAR115	0.50291	0.14775	0.57750	0.50125	0.10598	0.17814
VAR112	0.26466	0.05113	0.57231	0.36703	0.11383	0.37232
VAR65	0.42210	0.42498	0.56867	0.31098	0.30784	0.27139
VAR222	0.51499	0.29240	0.56862	0.35905	0.28147	0.25164
VAR74	0.44582	0.36384	0.56041	0.46625	0.23417	0.22147
VAR136	0.55163	0.21168	0.55864	0.46702	0.12592	0.25603
VAR72	0.54778	0.26008	0.55664	0.45123	0.20205	0.19852
VAR179	0.54102	0.23657	0.55602	0.38861	0.20750	0.22765
VAR19	0.15866	0.34604	0.55142	0.45536	0.23149	0.18784
VAR61	-0.03530	0.51137	0.52408	0.20121	0.32478	0.40094
VAR46	0.44694	0.37353	0.51792	0.51593	0.20067	0.22887
VAR158	0.47371	0.32934	0.51380	0.35291	0.26350	0.23713
VAR51	-0.13047	0.00785	0.47610	-0.04726	0.33992	0.41756
VAR217	0.25203	0.46282	0.47596	0.44822	0.43359	0.17266
VAR118	0.11070	0.26125	0.40928	0.25474	-0.00400	0.04505
VAR166	0.00708	-0.03465	0.04054	0.00675	-0.03367	-0.01947
VAR56 (37)	0.10198	0.03897	0.05991	**0.90439**	0.12967	0.20257
VAR75 (53)	0.10836	0.05679	0.04381	**0.89252**	0.31078	0.12038
VAR55 (36)	0.08287	0.27684	0.19310	**0.84258**	0.14632	0.27498
VAR3 (7)	0.10602	0.22810	0.17648	**0.80213**	0.43404	0.11807
VAR96 (67)	0.13281	0.15947	0.17309	**0.74385**	0.51115	0.29063
VAR161 (112)	0.20030	0.32556	0.41697	**0.67376**	0.18743	0.17324
VAR101 (72)	0.16575	0.38955	0.28111	**0.65658**	0.42937	0.28536
VAR99 (70)	0.20218	0.35825	0.42192	**0.63679**	0.32579	0.26906
VAR66 (46)	0.32318	0.45597	0.34771	**0.62675**	0.32448	0.12508
VAR122 *(92)*[3]	0.29850	0.08671	0.50520	0.58827	0.41110	0.22689

Table 2.8 (continued)

	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
VAR20 *(24)*	0.43327	0.43914	0.36419	0.58661	0.18785	0.21983
VAR160	0.23368	0.16054	0.49134	0.58457	0.25498	0.25213
VAR188	0.57162	0.23857	0.26824	0.58360	0.09040	0.27133
VAR218	0.33261	0.11978	0.52103	0.57937	0.19155	0.24599
VAR4	0.19405	0.06151	0.38168	0.57659	0.28398	0.35260
VAR224 *(160)*	0.38027	0.54699	0.22644	0.56970	0.30433	0.20370
VAR123 *(93)*	0.43154	0.31497	0.44477	0.56770	0.14458	0.21723
VAR92	0.39265	0.51580	0.32043	0.54782	0.08057	0.29295
VAR21 *(25)*	0.35044	0.45468	0.38718	0.52182	0.36945	0.24942
VAR103	0.37644	0.41868	0.49322	0.50627	0.28852	0.20349
VAR135	0.37003	0.45531	0.38703	0.50418	0.33823	0.24118
VAR62 *(43)*	0.45046	0.45560	0.38816	0.48347	0.31988	0.26445
VAR107 (78)	0.05017	0.08223	-0.02109	0.17116	**0.91733**	0.09069
VAR104 (75)	0.04943	0.16666	0.07080	0.28255	**0.88522**	0.17484
VAR95 (66)	0.12368	0.17873	0.16347	0.30009	**0.86385**	0.16402
VAR134 (104)	0.02695	0.12028	0.31757	0.17765	**0.86023**	0.26036
VAR57 (38)	0.08154	0.17349	0.05281	0.28517	**0.81069**	0.19031
VAR129 (99)	0.06650	0.17447	0.34018	0.28396	**0.78020**	0.25836
VAR59 (40)	0.38082	0.41262	-0.03428	0.16975	**0.72924**	0.27654
VAR67 (47)	0.36981	0.47183	0.06079	0.18825	**0.70390**	0.21587
VAR9 (13)	0.39275	0.40970	-0.00172	0.14455	**0.68228**	0.37170
VAR97 (68)	0.01246	0.55477	0.20644	0.35858	**0.62494**	0.27087
VAR78 (56)	0.03187	0.56154	0.23885	0.37176	**0.62170**	0.15590
VAR60 (41)	0.14593	0.16660	0.51556	0.27369	**0.61884**	0.30817
VAR77 (55)	0.22116	0.07423	0.30198	0.32903	**0.61192**	0.24537
VAR8	0.45215	0.46575	0.06493	0.21943	0.57926	0.26312
VAR130	0.48041	0.13168	0.00297	0.12946	0.55904	0.19239
VAR165	-0.02051	0.01848	-0.01768	0.00597	0.53357	0.06175
VAR13	0.13820	0.31264	-0.02869	0.11221	0.40394	0.35539
VAR208	0.01012	0.03389	0.02744	-0.02634	0.39401	-0.12032
VAR53	0.00946	0.03351	0.02814	-0.02564	0.39029	-0.12124
VAR220 (156)	-0.11280	0.00106	-0.12774	0.01374	-0.11347	**0.76139**
VAR206 (143)	-0.01674	-0.02946	-0.02111	0.09018	-0.09886	**0.73216**
VAR178 (123)	0.23119	0.27678	0.09111	0.22132	0.20746	**0.68344**
VAR110 (81)	-0.07768	0.00901	0.47332	-0.01421	0.28467	**0.66120**
VAR216 (152)	0.11605	0.14297	0.35754	0.06925	0.48311	**0.61039**
VAR180	0.54514	0.34200	-0.00689	0.12864	0.28606	0.58471
VAR164	0.09772	-0.00680	0.20086	0.16024	0.00169	0.58387
VAR47	0.39672	0.15765	-0.17485	0.02913	0.08461	0.54533

Table 2.8 (continued)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
VAR15	0.39446	0.33585	0.32053	-0.02255	0.23799	0.50762
VAR6	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR11	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR12	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR22	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR23	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR24	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR25	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR26	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR27	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR28	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR29	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR30	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR31	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR32	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR33	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR34	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR35	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR36	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR37	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR38	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR39	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR40	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR41	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR42	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR43	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR44	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR54	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR63	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR69	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR70	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR79	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR84	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR85	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR86	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR87	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR88	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR89	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR90	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR119	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Table 2.8 (continued)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
VAR139	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR140	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR141	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR142	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR143	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR144	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR145	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR146	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR147	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR148	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR149	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR150	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR151	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR152	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR153	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR154	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR155	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR156	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR157	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR159	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR172	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR173	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR174	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR175	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR176	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR194	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR195	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR196	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR197	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR198	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR199	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR200	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR209	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR210	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
VAR211	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Table 2.9. Factor 2 Sites Affect Ras Structure and Function
References: 1) Valencia et al. 1991. 2) Wittinghofer et al. 1991. 3) Kuppens et al. 1999. 4) Nassar, 1995. 5) Shirouzu et al. 1994. 6) Mapelli et al. 2005. 7) Paduch et al. 2001. 8) Créchet et al. 1996. 9) Ma and Karplus, 1997. 10) Neuwald et al. 2003. 11) Boriack-Sjodin et al. 1998. 12) Zhang and Matthews, 1998a,b.

Site(s)	Role in Ras Family Proteins
5	Superfamily conserved site[1]; H-bonds to site 77[2]
29	Hinge for switch I conformational change[3]
33	H-bond with Rap1a effector[4]
59	GTP hydrolysis, switch I conformational change, Raf binding[5,6]
61	Activates water for nucleophilic attack in GTP hydrolysis[7]
62	Decreases GEF binding affinity[8]
71	Facilitates conformational change upon hydrolysis[9]
77	Interaction with K5[2]
90	Unknown; part of switch mechanism in the Ran family[10]
102	H-bond with the RasGEF SOS[11]
120	Interaction with GTP binding site K117[1]
161	Intrahelix interactions may contribute to GDP binding stability[12]
162	Intrahelix interactions may contribute to GDP binding stability[12]
22, 103	unknown

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403-410.

Atchley, W. R., K. R. Wollenberg, W. M. Fitch, W. Terhalle, and A. W. Dress. 2000. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol. Biol. Evol. **17**:164-78.

Atchley, W. R., and A. D. Fernandes. 2005. Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network. Proc. Natl. Acad. Sci. USA **102**:6401–6406.

Atchley, W. R., J. Zhao, A. D. Fernandes, and T. Drüke. 2005. Solving the protein sequence metric problem. Proc. Natl. Acad. Sci. USA **102**:6395–6400.

Bairoch, A., B. Boeckmann, S. Ferro, and E. Gasteiger. 2004. Swiss-Prot: Juggling between evolution and stability. Brief. Bioinform. **5**:39-55.

Bairoch, A., R. Apweiler, C. H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. 2005. The Universal Protein Resource (UniProt*)***.** Nucleic Acids Res.

33:D154-159.

Barbacid, M. 1987. Ras Genes. Ann. Rev. Biochem. **56**:779-827.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. 2004. The Pfam Protein Families Database. Nucleic Acids Research Database Issue **32**:D138-D141.

Bauer, B., G. Mirey, I. R. Vetter, J. A. García-Ranea, A. Valencia, A. Wittinghofer, J. H. Camonis, and R. H. Cool. 1999. Effector Recognition by the Small GTP-binding Proteins Ras and Ral. J. Biol. Chem. **274**:17763-17770.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Research **28**:235-242.

Boriack-Sjodin, P.A., S. M. Margarit, D. Bar-Sagi, and J. Kuriyan. 1998 . The structural basis of the activation of Ras by Sos. Nature **394**:337-343. PDB ID: 1BKD.

Breiman, L. 1998. Classification and regression trees. Chapman & Hall/CRC, Boca Raton, Florida.

Bryant, S. H., and C. W. V. Hogue. 1996. Structural Neighbors and Structural Alignments: The Science Behind Entrez/3D. Presented at the IUCr Macromolecular Crystallography Computing School.

Buck, M. J., and W. R. Atchley. 2005. Networks of coevolving sites in structural and functional domains of serpin proteins. Mol Biol Evol. **22**:1627-1634.

Casari, G., C. Sander, and A. Valencia. 1995. A method to predict functional residues in proteins. Nat. Struct. Biol. **2**:171-178.

Castillo-Davis, C. I., F. A. Kondrashov, D. L. Hartl, and R. J. Kulathinal. 2004. The Functional Genomic Distribution of Protein Divergence in Two Animal Phyla: Coevolution, Genomic Conflict, and Constraint. Genome Research **14**:802-811.

Caughey, B. W., A. Dong, K. S. Bhat, D. Ernst, S. F. Hayes, and W. S. Caughey. 1991. Secondary Structure Analysis of the Scrapie-Associated Protein PrP 27-30 in Water by Infrared Spectroscopy. Biochemistry **30**:7672-7680.

Chardin, P. 1993. *Structural conservation of ras-related proteins and its functional implications*, p. 159-176. In F. D. Burton, and B. Lutz (ed.), GTPases in biology I, vol. 108. Spring-Verlag KG, Berlin, Germany.

Cherfils, J., J. Menetrey, G. Le Bras, I. Janoueix-Lerosey, J. de Gunzburg, J. R. Garel, and I. Auzat. 1997. Crystal structures of the small G protein Rap2A in complex with its substrate GTP, with GDP and with GTPgammaS. EMBO J. **16**:5582-5591. PDB ID: 2RAP, 1KAO.

Chou, P. Y., and G. D. Fasman. 1974. Prediction of protein conformation. Biochemistry **13**:222-245.

Connolly, D. 1993. Constructing hidden variables in Bayesian networks via conceptual learning. Proceedings of 10th International Conference on Machine Learning (ICML-93), Amherst, MA, USA, 65-72.

Créchet, J. B., A. Bernardi, and A. Parmeggiani. 1996. Distal switch II region of Ras2p is required for interaction with guanine nucleotide exchange factor. J Biol Chem. **271**:17234-17240.

Creighton, T. E. 1983. Proteins, Structure and Molecular Principles. W. H. Freeman and Company, New York, New York.

Crespo, P., and J. León. 2000. Ras proteins in the control of the cell cycle and cell differentiation. Cell. Mol. Life Sci. **57**:1613-1636.

Díaz, J. F., B. Wroblowski, J. Schlitter, and Y. Engelborghs. 1997. Calculation of pathways for the conformational transition between the GTP- and GDP-bound states of the Ha-ras-p21 protein: calculations with explicit solvent simulations and comparison with calculations in vacuum. Proteins **28**:434-51.

Díaz, J. F., M. M. Escalona, S. Kuppens, and Y. Engelborghs. 2000. Role of the switch II region in the conformational transition of activation of Ha-ras-p21. Protein Science **9**:361–368.

Diederichs, K. 1995. Structural superposition of proteins with unknown alignment and detection of topological similarity using a six-dimensional search algorithm. Proteins. **23**:187-95.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U. S. A.* **95**:14863–14868.

Feig, L. A. 2003. Ral-GTPases: approaching their 15 minutes of fame. Trends Cell Biol. **13**:419-25.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

Fetrow, J. S., and J. Skolnick. 1998. Method for Prediction of Protein Function from Sequence using the Sequence-to-Structure-to-Function Paradigm with Application to Glutaredoxins/Thioredoxins and T1 Ribonucleases. J. Mol. Biol. **281**:949-968.

Fetrow, J. S., N. Siew, and J. Skolnick. 1999. Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. The FASEB Journal **13**:1866-1874.

Fiegen, D., L. Blumenstein, P. Stege, I. R. Vetter, and M. R. Ahmadian. 2002. Crystal Structure of Rnd3/Rhoe: Functional Implications. FEBS Lett. **525**:100-104. PDB ID: 1M7B.

Foster, R., K.-Q. Hu, Y. Lu, K. M. Nolan, J. Thissen, and J. Settleman. 1996. Identification of a Novel Human Rho Protein with Unusual Properties: GTPase Deficiency and In Vivo Farnesylation. Mol. Cell. Biol. **16**:2689–2699.

Fukai, S., H. T. Matern, J. R. Jagath, R. H. Scheller, and A. T. Brunger. 2003. Structural basis of the interaction between RalA and Sec5, a subunit of the sec6/8 complex. EMBO J. **22**:3267-78. PDB ID: 1UAD.

Gaucher, E. A., X. Gu, M. M. Miyamoto, and S. A. Benner. 2002. Predicting functional

divergence in protein evolution by site-specific rate shifts. TRENDS in Biochemical Sciences **27**:315-321.

Gibrat, J. F., T. Madej, and S. H. Bryant. 1996. Surprising similarities in structure comparison. Current Opinion in Structural Biology **6**:377-385.

Harman, H. H. 1976. *Modern Factor Analysis*. University of Chicago Press, Chicago.

Heringa, J. 2000. Computational Methods for Protein Secondary Structure Prediction Using Multiple Sequence Alignments. Current Protein and Peptide Science **1**:273-301.

Hobohm, U., and C. Sander. 1995. A sequence property approach to searching protein databases. J. Mol. Biol. **251**:390-399.

Hogue, C. W. 1997. Cn3D: a new generation of three-dimensional molecular structure viewer. Trends Biochem. Sci. **22**:314-316.

Holm, L., and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233**:123-138.

Hsia, C. C., and W. McGinnis. 2003. Evolution of transcription factor function. Current Opinion in Genetics & Development **13**:199–206.

Huang, L., F. Hofer, G. S. Martin, and S. H. Kim. 1998. Structural basis for the interaction of Ras with RalGDS. Nat. Struct. Biol. **5**:422-426. PDB ID: 1LFD.

Huelsenbeck, J. P., and K. A. Crandall. 1997. PHYLOGENY ESTIMATION AND HYPOTHESIS TESTING USING MAXIMUM LIKELIHOOD. Annual Review of Ecology and Systematics **28**:437-466.

Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD - Visual Molecular Dynamics. J. Molec. Graphics **14**:33-38.

Isoldi, M. C., M. A. Visconti, and A. M. de Lauro Castrucci. 2005. Anti-cancer drugs: molecular mechanisms of action. Mini Rev. Med. Chem. **5**:685-95.

Jackson, S. E., M. Moracci, N. elMasry, C. M. Johnson, and A. R. Fersht. 1993. Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. Biochemistry. **32**:11259-11269.

Jin, R., J. R. Junutula, H. T. Matern, K. E. Ervin, R. H. Scheller, and A. T. Brunger. 2005. Exo84 and Sec5 are Competitive Regulatory Sec6/8 Effectors to the Rala Gtpase. Embo J. **24**:2064-2074. PDB ID: 1ZC3.

Johnson, R. A., and D. W. Wickern. 1992. ***Applied Multivariate Statistical Analysis, 3rd*** edition. Prentice Hall, Englewood Cliffs, New Jersey.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

Kigawa, T., E. Yamaguchi-Nunokawa, K. Kodama, T. Matsuda, T. Yabuki, N. Matsuda, R. Ishitani, O. Nureki, and S. Yokoyama. 2002. Selenomethionine Incorporation Into a Protein by Cell-Free Synthesis J. Struct. Funct. Genom. **2**:29-35. PDB ID: 1IOZ.

Kim, S., E. R. Dougherty, Y. Chen, K. Sivakumar, P. Meltzer, J. M. Trent, and M. Bittner. 2000. Multivariate measurement of gene expression relationships. Genomics **67**:201–209.

Kontani, K., M. Tada, T. Ogawa, T. Okai, K. Saito, Y. Araki, and T. Katada. 2002. Di-Ras, a Distinct Subgroup of Ras Family GTPases with Unique Biochemical Properties. J. Biol. Chem. **277**:41070-41078.

Konvicka, K., F. Guarnieri, J. A. Ballesteros, and H. Weinstein. 1998. A Proposed Structure for Transmembrane Segment 7 of G Protein-Coupled Receptors Incorporating an Asn-Pro/Asp-Pro Motif. Biophys. J. **75**:601-611.

Kuppens, S., J. F. Diaz, and Y. Engelborghs. 1999. Characterization of the hinges of the effector loop in the reaction pathway of the activation of *ras*-proteins. Kinetics of binding of beryllium trifluoride to V29G and I36G mutants of Ha-*ras*-p21. Protein Science **8**:1860–1866.

Lee, D. C., P. I. Haris, D. Chapman, and R. C. Mitchell. 1990. Determination of protein secondary structure using factor analysis of infrared spectra. Biochemistry **29**:9185-9193.

Lesk, A. M., and C. Chothia. 1982. Evolution of proteins formed by β-sheets II. The core of the immunoglobulin domains. J. Mol. Biol. **160**:325-342.

Liu, W., M. Eilers, A. B. Patel, and S. O. Smith. 2004. Helix Packing Moments Reveal Diversity and Conservation in Membrane Protein Structure. J. Mol. Biol. **337**:713–729.

Ma, J., and M. Karplus. 1997. Molecular switch in signal transduction: Reaction paths of the conformational changes in *ras* p21. Proc. Natl. Acad. Sci. USA **94**:11905-11910.

Madej, T., J. F. Gibrat, and S.H. Bryant. 1995. Threading a database of protein cores. Proteins **23**:356-369.

Mapelli, V., S. Fantinato, E. Accardo, L. De Gioia, and M. Vanoni. 2005. Structure-based

hypothesis on active role of RasGEF αG-helix. FEBS Journal 272 (s1), E4-006.

Martz, E. 2002. Protein Explorer: Easy Yet Powerful Macromolecular Visualization. Trends in Biochemical Sciences **27**:107-109. http://proteinexplorer.org

McCormick, F. 1995. Ras-related proteins in signal transduction and growth control. Mol. Reprod. Dev. **42**:500-6.

Milburn, M. V., L. Tong, A. M. deVos, A. Brunger, Z. Yamaizumi, S. Nishimura, and S. H. Kim. 1990. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. Science **247**:939-45.

Myagmar, B. E., M. Umikawa, T. Asato, K. Taira, M. Oshiro, A. Hino, K. Takei, H. Uezato, and K. Kariya. 2005. PARG1, a protein-tyrosine phosphatase-associated RhoGAP, as a putative Rap2 effector. Biochem. Biophys. Res. Commun. **329**:1046-52.

Nassar, N. 1995. The 2.2 A Crystal Structure of the Ras-Binding Domain of the Serine/Threonine Kinase C-Raf1 in Complex with RAP1A and a GTP Analogue. Nature **375**:554-560. PDB ID: 1C1Y.

Neuwald, A. F., N. Kannan, A. Poleksic, N. Hata, and J. S. Liu. 2003. Ran's C-terminal, Basic Patch, and Nucleotide Exchange Mechanisms in Light of a Canonical Structure for

Rab, Rho, Ras, and Ran GTPases. Genome Research **13**:673-692.

Nicely, N. I., J. Kosak, V. De Serrano, and C. Mattos. 2004. Crystal Structures of Ral-Gppnhp and Ral-Gdp Reveal Two Binding Sites that are Also Present in Ras and RAP. Structure **12**:2025-2036. PDB ID: 1U8Y, 1U8Z.

Ortiz, A. R., and J. Skolnick. 2000. Sequence Evolution and the Mechanism of Protein Folding. Biophysical Journal **79**:1787-1799.

Oxford, G., and D. Theodorescu. 2003. Ras superfamily monomeric G proteins in carcinoma cell motility. Cancer Letters **189**:117-128.

Pacold, M. E., S. Suire, O. Perisic, S. Lara-Gonzalez, C. T. Davis, E. H. Walker, P. T. Hawkins, L. Stephens, J. F. Eccleston, and R. L. Williams. 2000. Crystal structure and functional analysis of Ras binding to its effector phosphoinositide 3-kinase gamma. Cell **103**:931-43. PDB ID: 1HE8.

Paduch, M., F. Jeleñ, and J. Otlewski. 2001. Structure of small G proteins and their regulators. Acta Biochim. Pol. **48**:829-850.

Pagel, M. D. 1994. Detecting correlated evolution of phylogenies: A general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London,

Series B **255**:37-45.

Pai, E. F., U. Krengel, G. A. Petsko, R. S. Goody, W. Kabsch, and A. Wittinghofer. 1990. Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: implications for the mechanism of GTP hydrolysis. EMBO J. **9**:2351-9.

Papagrigoriou, E., X. Yang, J. Elkins, F. Niesen, N. Burgess, E. Salah, O. Fedorov, L. J. Ball, F. Von Delft, M. Sundstrom, A. Edwards, C. Arrowsmith, J. Weigel, and D. Doyle. Crystal Structure of Diras2. *To be Published* PDB ID: 2ERX.

Pasqualato, S., L. Renault, and J. Cherfils. 2002. Arf, Arl, Arp and Sar proteins: a family of GTP-binding proteins with a structural device for 'front–back' communication. EMBO Reports **3**:1035–1041.

Pellis-van Berkel, W., M. H. Verheijen, E. Cuppen, M. Asahina, J. de Rooij, G. Jansen, R. H. Plasterk, J. L. Bos, and F. J. Zwartkruis. 2005. Requirement of the Caenorhabditis elegans RapGEF pxf-1 and rap-1 for epithelial integrity. Mol. Biol. Cell **16**:106-16.

Pollock, D. D., W. R. Taylor, and N. Goldman. 1999. Coevolving Protein Residues: Maximum Likelihood Identification and Relationship to Structure. J. Mol. Biol. **287**:187-198.

Pollock, D. D. 2002. Genomic biodiversity, phylogenetics and coevolution in proteins. Appl. Bioinformatics **1**:81-92.

Russell, R. B., and G. J. Barton. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 14:309-323.

Salah, E., G. Schoch, A. Turnbull, E. Papagrigoriou, M. Soundararajan, N. Burgess, J. Elkins, C. Gileadi, O. Gileadi, F. Von Delft, A. Edwards, C. Arrowsmith, J. Weigel, M. Sundstrom, and D. Doyle. The Crystal Structure of the Ras Related Protein Rras2 (Rras2) in the Gdp Bound State *To be Published* PDB ID: 2ERY.

Salzburg, S. 1995. Locating protein coding regions in human DNA using a decision tree algorithm. J. Comput. Biol. **2**:473-485.

Saucedo, L. J., X. Gao, D. A. Chiarelli, L. Li, D. Pan, and B. A. Edgar. 2003. Rheb promotes cell growth as a component of the insulin/TOR signalling network. Nature Cell Biology **5**:566-571.

Scheffzek, K., M. R. Ahmadian, W. Kabsch, L. Wiesmuller, A. Lautwein, F. Schmitz, and A. Wittinghofer. 1997. The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. Science **277**:333-8. PDB ID: 1WQ1.

Scheffzek, K., P. Grunewald, S. Wohlgemuth, W. Kabsch, H. Tu, M. Wigler, A. Wittinghofer, and C. Herrmann. 2001. The Ras-Byr2RBD complex: structural basis for Ras effector recognition in yeast. Structure **9**:1043-50. PDB ID: 1K8R.

Shirouzu, M., H. Koide, J. Fujita-Yoshigaki, H. Oshio, Y. Toyama, K. Yamasaki, S. A. Fuhrman, E. Villafranca, Y. Kaziro, and S. Yokoyama. 1994. Mutations that abolish the ability of Ha-Ras to associate with Raf-1. Oncogene **9**:2153-2157.

Sprang, S. R. 1997. G proteins, effectors and GAPs: structure and mechanism. Curr. Opin. Struct. Biol. **7**:849-56.

Stork, P. J. S. 2003. Does Rap1 deserve a bad Rap? Trends in Biochemical Sciences **28:**267-275.

Taira, K., M. Umikawa, K. Takei, B. E. Myagmar, M. Shinzato, N. Machida, H. Uezato, S. Nonaka, and K. Kariya. 2004. The Traf2- and Nck-interacting kinase as a putative effector of Rap2 to regulate actin cytoskeleton. J. Biol. Chem. **279**:49488-96.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research **24**:4876-4882.

Tillier, E. R. M., and R. A. Collins. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. Mol. Biol. Evol. **12**:7-15.

Valencia, A., P. Chardin, A. Wittinghofer, and C. Sander. 1991. The ras protein family: evolutionary tree and role of conserved amino acids. Biochemistry **30**:4637-48.

Wang, D., X. Wang, V. Honavar, and D. Dobbs. 2001. Data-Driven Generation of Decision Trees for Motif-Based Assignment of Protein Sequences to Functional Families. Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems & Technology.

Wennerberg, K., K. L. Rossman, and C. J. Der. 2005. The Ras superfamily at a glance. Journal of Cell Science **118**:843-846.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. 2006. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. **34**:D173-D180.

Wi, S., P. Pancoska, and T. A. Keiderling. 1998. Predictions of protein secondary structures using factor analysis on Fourier transform infrared spectra: Effect of Fourier self-deconvolution of the amide I and amide II bands. Biospectroscopy **4**:93-106.

Wilson, C. A., J. Kreychman, and M. Gerstein. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J. Mol. Biol. **297**:233-249.

Wittinghofer, F., U. Krengel, J. John, W. Kabsch, and E. F. Pai. 1991. Three-dimensional structure of p21 in the active conformation and analysis of an oncogenic mutant. Environ. Health Perspect. **93**:11-15. PDB ID: 121p.

Wollenberg, K. R., and W. R. Atchley. 2000. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc. Natl. Acad. Sci. USA **97**:3288-91.

Ye, M., F. Shima, S. Muraoka, J. Liao, H. Okamoto, M. Yamamoto, A. Tamura, N. Yagi, T. Ueki, and T. Kataoka. 2005. Crystal Structure of M-Ras Reveals a GTP-Bound "Off" State Conformation of Ras Family Small Gtpases. J. Biol. Chem. **280**:31267-31275. PDB ID: 1X1S, 1X1R.

Yu, Y., S. Li, X. Xu, Y. Li, K. Guan, E. Arnold, and J. Ding. Structural Basis for the Unique Biological Function of Small Gtpase Rheb. 2005 J. Biol. Chem. **280**:17093-17100. PDB ID: 1XTS, 1XTQ.

Zhang, J. and C. R. Matthews. 1998. The role of ligand binding in the kinetic folding mechanism of human p21(H-ras) protein. Biochemistry **37**:14891-14899.

Zhang, J. and C. R. Matthews. 1998. Ligand binding is the principal determinant of stability for the p21(H)-ras protein. Biochemistry **37**:14881-14890.

Zhang, L. V., S. L. Wong, O. D. King, and F. P. Roth. 2004. Predicting co-complexed protein pairs using genomic and proteomic data integration BMC Bioinformatics **5**:38.

Zhang, Y., X. Gao, L. J. Saucedo, B. Ru, B. A. Edgar, and D. Pan. 2003. Rheb is a direct target of the tuberous sclerosis tumour suppressor proteins. Nature Cell Biology **5:**578-581.

Chapter 3

Sequence Signatures For Delimiting Families of Ras Proteins

by

Andrew Dellinger[1] and William R. Atchley[2]

[1]Bioinformatics Research Center and [2]Department of Genetics

North Carolina State University

Raleigh, NC 27695

Send editorial correspondence to Andrew E. Dellinger (aedellin@ncsu.edu)

Abstract

The Ras protein superfamily is composed of six families: Rab, Ran, Ras, Rho, Sar/Arf and Rad/Gem/Kir (RGK). We delimit and define the families of the Ras superfamily using Entropy-based sequence signatures (motifs). The sequence signatures correctly classified the most proteins when composed of superfamily-conserved sites and uniquely-conserved family sites. We discuss what the sequence signatures reveal about Ras biology and evolution. Comparison of signatures revealed that the sites that describe essential structural and functional changes among protein families have physiochemical interactions with protein- and nucleotide-binding sites. The quantity of known and unknown family proteins retrieved by the Entropy-based sequence signature, PROSITE sequence signature, profile HMM and position-specific scoring matrix methods of protein classification were compared. The Entropy signatures identified more known and previously unclassified proteins than the PROSITE signatures but fewer than the profile HMM and position-specific scoring matrix methods.

Introduction

A sequence signature or "predictive motif" is a small set of amino acids that when considered together uniquely identify a particular set of proteins (Atchley and Fernandes, 2005; Atchley et al. 1999). These sequence combinations are useful for quantitatively examining the limits of a particular protein family and exploring the question "What sequence elements define a particular group of proteins?" Sequence signatures are typically composed of sites which are highly conserved within the target group of proteins, a feature

that facilitates the use of site-selection criteria like the Boltzmann-Shannon entropy value. Comparative analyses of these sequence signatures can determine which sites and amino acids played important roles in the changes in function and structure among protein families. Further, sequence signatures can be used to speed database searches, delimit sets of related proteins and classify unknown proteins (Atchley and Fernandes, 2005; Banerjee-Basu and Baxevanis, 2001; Tang et al. 2005).

Atchley et al. (1999) provided a 19 element sequence signature to accurately delimit the members of the basic helix-loop-helix protein family, an important and diverse set of transcriptional regulators. More recently, Atchley and Fernandes (2005) used sequence signatures to accurately identify the members of the Myc/Mad/Max network of bHLH transcription factors. Atchley et al (1999) and Atchley and Fernandes (2005) incorporated Boltzmann-Shannon entropy and fuzzy logic searching since there is considerable *a priori* knowledge available about the proteins in question. Many other methods have been proposed to produce sequence signatures. Their efficacy is dependent on how much *a priori* knowledge is available about the sequences under discussion. Gaurav et al. (2005) used position specific scoring matrices (PSSMs) to classify proteins with discontinuous domains or circular permutations. Pellegrini-Calace and Thornton (2005) used hidden Markov models (HMMs) to find DNA binding helix-turn-helix proteins. Henikoff and Henikoff (1994) developed BLOCKS, a method to classify proteins using highly conserved regions. BLAST can be used to search for related proteins and classify unknown proteins by their sequence similarity (Altschul et al. 1990).

Entropy-based sequence signatures have several advantages over these methods.

Protein classification methods such as BLOCKS, PSSMs, and profile HMMs use large stretches of sites in their identification models, while entropy signatures use only a few select sites. Unlike the sites chosen by the other methods, these sites are the most important in discriminating among protein families. They also tend to be functionally or structurally significant and are informative about the diversification of related proteins, while such information is lost in the larger and more contiguous sets of sites required by other methods.

Herein, we generate a series of sequence signatures for the Ras protein superfamily, an important collection of GTPases well-known to act as molecular switches. The Ras superfamily is composed of six families: Rab, Ras, Ran, Rho, Sar/Arf, and Rad/Gem/Kir (RGK) (Oxford and Theodorescu, 2003; Crespo and León, 2000). Structurally, these proteins have a common catalytic domain involving 5 α-helices, a 6 stranded β-sheet, and 10 loops. Five of these loops are functional (Paduch et al. 2001; Milburn et al. 1990). Loops L1, L2, L4, L8, and L10 have nucleotide binding sites. Loops L2 and L4 are part of the protein binding switch regions. Switch I is composed of sites 30-38 in L2 (Nicely et al. 2004) and switch II is composed of sites 60-76 in L4 and part of α2 (Milburn et al. 1990). The switches change conformation according to the binding state of the protein. GTP-bound superfamily proteins are active because the two switch regions are in the proper conformation for effector protein binding. After the protein hydrolyzes GTP into GDP, the switch regions change conformation and the protein is no longer active, being unable to bind effectors (Sprang, 1997).

Ras superfamily proteins play important and diverse roles in the cell, including regulation of proliferation, differentiation, vesicle trafficking, nuclear import and export of

proteins, adhesion, and exocytosis (Wennerberg, et al. 2005). They also play important roles in human diseases. For example, Ras family proteins are involved in 20-30% of human cancers (Isoldi, et al. 2005) and Rab9 is a viral transport protein for HIV, Ebola and other important disease-causing viruses (Chen et al. 2004).

Herein we generate a set of sequence signatures in order to ask the following questions: 1) What sequence elements define these important groups of proteins? 2) Can we find small subsets of amino acids that will accurately identify the various families within the Ras Superfamily? 3) Do the components of these sequence signatures reflect amino acids with known functional and structural roles? 4) Are these signature components confined to structural and functional domains within the proteins? 5) What is the relative efficacy of various methods for generating sequence signatures of the Ras superfamily proteins?

Materials and Methods

A dataset of 490 unique and complete Ras superfamily protein sequences was obtained from Swissprot (Boeckmann et al. 2003). These proteins were classified into families and clades using the Ras superfamily tree of Dellinger and Atchley (2006) and their family identification in Swissprot. The proteins were divided into Rab, Rad-Gem-Kir (RGK), Ran, Ras, Rho, Sar, and Arf (Oxford and Theodorescu, 2003; Crespo and León, 2000). The traditional Sar/Arf family was divided because of the early split between the two subfamilies in the Ras superfamily phylogenetic tree in Dellinger and Atchley (2006a). Sequences from each family were aligned using CLUSTALX (Jeanmougin et al. 1998) with minimal adjustment by eye. The Swissprot version 48.7 and TrEMBL version 31.7 databases were

used to assess the ability of protein classification methods to identify family proteins. The nr-aa database, excluding Swissprot and TrEMBL sequences, was used to optimize the identification power of the entropy-based sequence signatures (Boeckmann et al. 2003; Barker et al. 1999; http://www4.prf.or.jp/en/seqsrch.html, protein sequence database of the Protein Research Foundation, Osaka, Japan; Benson et al. 2000).

Since each family of the Ras superfamily has its own numbering system, the Ras family numbering system will be used whenever possible. The Ras family number system is based on Ras subfamily proteins, which have a four residue N-terminus and a superfamily-conserved site 5 (Valencia et al. 1991). Residues will be denoted in the form G60, where G is the one-letter amino acid code for Glycine and 60 is the location in the protein.

Definition of Known and Unknown Proteins

Evaluation of probabilistic identification procedures with proteins require both "known" and "unknown" data. The Uniprot database (Bairoch et al. 2005) and the family's phylogenetic tree were used to classify proteins as "known" or "unknown" family members. Detection of unknown proteins is a measure of the efficacy of a method since it determines the method's ability to accurately assign distantly related and unclassified proteins to a family.

A protein was classified as a "known" family protein if it met one of two criteria: 1) it was found to be within a well-established family clade in the phylogenetic analyses 2) it was clearly identified and placed in a given protein family in the Uniprot database. Proteins which met neither of these criteria were classified as "unknown". Clearly, there are situations

where this dichotomy is somewhat subjective. For example, proteins that did not meet the second criteria were classified as either “known” or “unknown” based on their location in their clade. Proteins that arose before the proteins classified as “known” by the second criteria were defined as “unknown” based on the possibility that they are more related to proteins ancestral to the family than to the family itself. Proteins that arose after the proteins classified as “known” by the second criteria were defined as “unknown” based on the possibility that their sequences have changed such that they are no longer family proteins.

The following procedure of phylogenetic tree estimation was repeated for each of the families. First, four classification analyses were used as described below. The amino acid sequences of the proteins classified by the analyses were obtained from Uniprot (Bairoch et al. 2005). To classify proteins with sequence characteristics similar to more than one group, these sequences included non-group proteins with e-values more significant than the e-value of the least significant true positive group protein. An e-value is the number of sequences one would expect to find by chance with a score greater than or equal to the current match. E-values are linearly dependent on database size such that the same alignment in a database twice as big will give an e-value twice as big. To assist in the identification of non-family proteins, a sample of Swissprot (Boeckmann et al.2003) proteins from each of the families were also retrieved and added to each family's set of proteins.

The family's set of proteins was aligned. DIALIGN2, a local alignment program, was used to generate alignments for the RGK, Ran, Ras, and Sar sets (Morgenstern, 1999). The Arf, Rho and Rab sets contained too many sequences to be computationally feasible using

DIALIGN2. They were aligned using CLUSTALW (Gibson, 1994) with a GONNET250 substitution matrix (Gonnet et al. 1992). The GONNET substitution matrix has been reported to be superior to the PAM and BLOSUM matrices in both global and local alignments (Vogt, 1995; Henikoff, 1996). All alignments were minimally improved by eye if necessary. The N-terminal and C-terminal regions outside the functional core of the proteins were deleted. Further, autapomorphies were deleted to increase the computational feasibility of phylogenetic tree estimation. Branch lengths were not considered in the assignment of proteins as "known" or "unknown". Phylogenetic trees were estimated using the neighbor-joining method with a GONNET substitution matrix, allowance for multiple substitutions and 3000 nonparametric bootstraps. Five thousand nonparametric bootstraps were used in estimating the Rab tree in order to compensate for the greater number of sequences.

Entropy-based Sequence Signatures

The entropy-based sequence signature method, hereafter referred to as the Entropy Method, was proposed by Atchley et al. (1999) to generate sequence signatures for the entire family of basic helix-loop-helix proteins and was then used to construct sequence signatures for the Myc-Mad-Max network. Unlike fully automatic signature building methods such as those used by PROSITE, MEME and profile HMMs, the Entropy Method requires manual integration of entropy values and known structural and functional information by the user (Sigrist el al. 2002; Bailey and Elkan, 1994; Eddy, 1998). Thus, the Entropy Method is best used on a set of related proteins for which there is significant knowledge about the roles of amino acid sites in structure and function.

The disadvantage of this method is that it is more time intensive. However, the quality of the signature is potentially much better. For example, each site in the signature gives insight into the constraints on protein structure or function. Further, there are usually fewer sites required to delimit the set of proteins. And comparison of related families' Entropy signatures gives insight into the pertinent variability that controls the differences in the families' structure and biological function.

Boltzmann-Shannon entropy is calculated as:

$$E = -\sum_{i=1}^{20} (P_i * \log_2(P_i))$$

(1)

where P_i is the frequency of amino acid *i* observed in the site. Entropy is zero when the site is completely conserved ($P_i = 1$ for the conserved amino acid) in a multiple alignment. Entropy was calculated for the twenty amino acids using $\log_{20}$, which scales the values from 0 to 1. Twenty amino acids were divided into the eight functional groups previously described in Atchley et al. (1999). Entropy values were also calculated by summing over the eight functional groups of amino acids using $\log_8$ to normalize the values. Functional groups are physiochemically related amino acids which are substituted with an amino acid within the group more often than with an amino acid outside of the group. Using the IUPAC one letter amino acid abbreviations (IUPAC, 1968), the functional groups are: FWY, AGLIVM, HKR, DE, NQ, ST, C and P.

Sites with small entropy values are good candidates for sequence signatures. They are more evolutionarily conserved and should have a smaller chance of exhibiting a different residue in a family protein outside of the Swissprot dataset. Thus, these sites should yield

fewer false negatives.

Swissprot gives a small, species-biased sample of protein families. Such a sample may lead to underestimation of a site's true entropy value. Two measures were taken to correct for this sampling problem. First, a literature search and analyses of protein crystal structures in PDB (Berman et al. 2000) were conducted to find sites with both small entropy values and structural or functional constraints. Constraints provide additional evidence of the restriction of allowable residues in a site. A subset of these sites were used to build the signature. Because biological classification is hierarchical, signature sites may be from either family-conserved sites or superfamily-conserved sites. Ras superfamily-conserved sites are nucleotide binding sites. They are also conserved in the related G-proteins, which like the Ras proteins are GTPases. Family-defining sites have a subset of amino acids unique to that family. Both classes of sites were used to build entropy-based signatures. Sites with minimal overlap with amino acid composition at the corresponding sites in other families were candidates for this category. Second, signatures were tested against the PIR and PRF protein databases for permissible residues not included in the Swissprot-built signature sites (Barker et al. 1999; http://www4.prf.or.jp/en/seqsrch.html, protein sequence database of the Protein Research Foundation, Osaka, Japan). The scaled entropy values of each Ras family's amino acid sites and the chosen signature sites are shown in Appendix B figures B.1-B.7.

The ideal signature is like the Myc-Mad-Max signature of Atchley and Fernandes (2005), which involves 28 contiguous sites that are involved in a structurally and evolutionarily important function. Longer and noncontiguous sites risk being affected by indels. Indels increase the range of the number of sites between potential signature sites.

They also increase the probability that other protein families will match the signature because there are a greater number of valid combinations of sites for the signature. A signature was built for each family except the Sar/Arf family, where both Sar and Arf related proteins received their own signatures. The structural or functional significance of each site in the signature was analyzed.

The Entropy Method has several advantages over other signature-building methods. Calculation of entropy values is simple compared to methods which use position specific scoring matrices, profile hidden Markov models, or other complex probabilistic models. Also, the results of the Entropy Method are straightforward compared to these signature-building methods. Signature sites often reflect sites having important phylogenetic, structural, or functional constraints. Finally, the qualitative structural and functional constraints of a site can be determined using experimental and structural data.

Several signatures were evaluated for each family to obtain the signature with the most true positives and fewest false positives. Each signature was input into the Prosite search engine and searched against the NR-AA database, a nonredundant protein database built from the Swissprot, PIR, PRF, and GenPept databases (motif.genome.jp/MOTIF2.html; Boeckmann et al. 2003; Barker et al. 1999; http://www4.prf.or.jp/en/seqsrch.html, protein sequence database of the Protein Research Foundation, Osaka, Japan; Benson et al. 2000). The results of each test signature's NR-AA database search were analyzed to confirm that the signature retrieved no false positives. False positives were considered to be those proteins whose annotation in the Uniprot database identified them as non-family members. Signatures which returned no false positives were compared based on the number of true positives.

Final versions of each family's signature were compared to other methods of discriminating among protein families: PROSITE's sequence signatures (Sigrist el al. 2002), profile hidden Markov models (HMMs) (Eddy, 1998) and MEME's position-specific scoring matrices (PSSMs) (Bailey and Elkan, 1994).

PROSITE signatures

PROSITE is an annotated database of signatures and patterns for protein families and domains. For each family, PROSITE first performed a multiple alignment of known family proteins. The signatures were built using conserved sites in the alignment which were also within regions of biological function (Sigrist et al. 2002). These signatures are available for the Sar, Arf and Ran families of the Ras superfamily.

In this paper, both PROSITE and Entropy Method signatures are written as follows. An example signature is [ALIVT]-x(2)-[GS]-[LI]-[DQ]-x(2)-G-K-[ST]-[ST]-x-[LIVM]-x(13-22)-T-x-G. Amino acids are written in their one-letter IUPAC code (IUPAC, 1967). Signature sites are separated by a dash. [KR] signifies that a site may contain either Lysine (K) or Arginine (R). *x* signifies that a site may contain any of the 20 amino acids. x(*i*) signifies that *i* adjacent sites may contain any amino acid. And *x(i-j)* is used when between i and j adjacent sites may contain any amino acid. A period terminates the signature.

The PROSITE signatures for the Sar, Arf and Ran families were input into Windows Grep (Mullington, 2000), a tool used to search files for regular expressions, to search the Swissprot and TrEMBL databases (Boeckmann et al. 2003). The number of unknown proteins, true positives, and false positives were recorded for each search.

Profile HMMs

Profile HMMs build statistical models from sequence alignments. For each site in the sequence alignment the models provide information about the amount of conservation and the likelihood of each amino acid residue's presence in that site (Bateman et al. 2002). The alignment of each family's Swissprot sequences were input into the HMMer (Eddy, 2003) module *hmmbuild*, which built a profile HMM. The module *hmmcalibrate* was used to generate an extreme value distribution which was used to calculate more accurate e-values. An estimate of the mean protein length was entered to increase accuracy. Finally, *hmmpfam* was used to search the TrEMBL (Boeckmann et al. 2003) database for matches with e-values less than 1e-05. The number of proteins in the TrEMBL database was entered into the *hmmpfam* module to correctly calculate the e-values.

Matches were determined by the score and e-value of the pairwise alignment of the HMM and the protein. The resulting matches were scanned by eye to find the smallest e-value within the set of true negatives. Matches with e-values less than this number were considered possible true positives and kept for further analysis. These proteins were divided into known, unknown, and non-family proteins.

MEME

Multiple Expectation maximization for Motif Elicitation (MEME) builds signatures using position-specific scoring matrices (PSSMs). In the matrix, rows are sites and columns are amino acids. The elements of the matrix are the log-odds calculated by

$$\log_2 (p_i/f_i) \qquad (2)$$

where p_i is the probability of amino acid *i* at that amino acid site, and f_i is the background frequency of the amino acid (Bailey and Elkan, 1994). MEME weights each site's importance by bits of information (Durbin et al. 1998). Bits are a measure from information theory and are calculated as

$$\sum_{i=1}^{20} (q_i * \log_2(\frac{q_i}{p_i})) \quad (3)$$

where q_i is the observed proportion of amino acid *i* in the site. Equation 3 computes a site's variation from the background amino acid frequencies, while Equation 1 used by the Entropy Method computes a site's variation from complete conservation. Gaps are not allowed in MEME signatures. MEME solves this problem by creating a signature on either side of a variable-length region, (Durbin et al. 1998).

Each family's unaligned Swissprot sequences were input into the MEME server. For each family, MEME was allowed to build up to five signatures of 6-50 sites. MEME returned PSSMs for each family, which were subsequently submitted to MAST (Motif Alignment and Search Tool) (Bailey and Gribskov, 1998). MAST was run to search the Uniprot database for significant matches to MEME's PSSMs. Matching proteins with an e-value less than that of the false positive with the greatest e-value were considered true positives because matches with an e-value greater than that of a false positive must be considered suspect as to the accuracy of their classification.

Results and Discussion

Sequence signatures were built to: classify all of the Ras superfamily proteins in the TrEMBL database, find a minimal set of discriminatory sites, identify the sites that give

unique functional and structural properties to the Ras families, and find patterns of functional and structural change that occurred in the diversification of the Ras superfamily.

Entropy Signatures

Sequence signatures built using the Entropy Method are given in Table 1. The location of the signature sites in the protein are shown in Appendix Figures B.8-B.14. Ran, Rho and Rab signature sites are primarily located in switch II. Sar and Arf signature sites are primarily in the P-loop and switch I. RGK and Ras signature sites are spread out from the P-loop to switch II.

Sar and Arf: The structural and functional significance of Sar and Arf family signature sites are summarized in Tables 3.2 and 3.3. Residues L11 and D12 (Ras numbering) are diagnostic for the Sar and Arf family proteins. L11 denotes a Leucine in site 11. In both Sar and Arf crystal structures, D12 forms salt bridges with R65 in helix α2 and R89 in helix α3 (Tables 2 and 3). In Sar and Arf proteins, the Leucine in site 11 has possible van der Waals' contact with Ras numbered site 61, which is the catalytic site for GTP hydrolysis in Arf (Boehm et al. 2001) and probably in Sar as well. The salt bridge in site 12 and the possible van der Waals' contact of sites 11 and 61 are not present in other families. Switch I sites 37 and 38 separate Sar and Arf family proteins. The hydrogen side-chain of G37 in Arf family proteins is probably the switch I hinge, which allows switch I to achieve an active conformation. Site P38 in Sar proteins in addition to P34 may provide an alternate mechanism to achieve this conformation. The prolines may bend the switch into the active conformation in the GTP-bound form of the protein.

Ras: Table 3.4 describes the structural and functional significance of Ras family signature sites. Site 56 is the primary diagnostic site. Ras family proteins typically have aliphatic residues in this site, while the other families of the superfamily almost exclusively have aromatic residues. A L56F substitution was modeled in order to determine if the structure and physiochemistry of the surrounding environment prohibits the presence of aromatic amino acids in site 56. The H-Ras-GTP crystal structure was used (PDB ID: 121p; Wittinghofer et al. 1991).

First, L56 and F78 of Ras have approximately the same angle among the alpha, beta and gamma carbons of their side chains (115.1° and 112.8°, respectively), thus the γ-carbon of L56 is in a nearly equivalent spatial position to the γ-carbon of F56. The position of the other carbon atoms in the aromatic ring may be computed beginning from the γ-carbon. Second, the aromatic ring of Phenylalanine maintains the slope of the line between the β-carbon and γ-carbon, thus the position of the ζ-carbon in F78 is easily computed. This line in site 56 intersects the δ-carbon of E37 in switch I. The distance along this line between the δ-carbon of E37 and the γ-carbon of L56 is 4.367Å. The distance along this line between the δ-carbon of E37 and ζ-carbon in F78 is only 2.773Å (PDB ID: 121p; Wittinghofer et al. 1991), leaving only 1.594Å between the nuclei of these two carbon atoms. The van der Waals' radius of the E37 δ-carbon is approximately 1.74Å and the radius of the ζ-carbon of F56 would be approximately 1.82Å (Li and Nussinov, 1998)-- an overlap of 1.966Å. Physics does not allow much, if any overlap of the van der Waals' spheres of two atoms. Thus switch I conformation would have to change, which would disrupt binding of H-Ras to its effectors and GAPs. Site 56 of the RasS protein does have a Tyrosine in the slime mold *Dictyostelium*

discoideum, which arose early among the eukaryotes. Unlike other Ras proteins, the structural environment around site 56 in *D. discoideum* probably has the characteristics of other Ras superfamily proteins, which allow residues with bulky side chains.

Ran: Table 3.5 describes the structural and functional significance of Ras family signature sites. Site 81 is the only true diagnostic site in the Ran signature. The Ran family is unique in having Isoleucine or Valine in this site (Boeckmann et al. 2003). I81 and V81 give Ran van der Waals' contact with proteins associated with Ran's unique function- the nuclear import and export of proteins. Ran site 81 has potential van der Waals' contact with Karyopherins, also known as Importins. Specifically, I81 has contacts with Karyopherin-β2 (3.256Å) and Importin-β (3.592Å, 3.632Å) (Table 5; PDB ID: 1QBK; Chook and Blobel, 1999); PDB ID: 1IBR; Vetter et al. 1999). Karyopherin-β2 binds to a substrate and imports it into the nucleus. Then Ran-GTP binds Karyopherin-β2, causing the substrate to dissociate. In nuclear export systems Ran-GTP, Karyopherin-β2 and the substrate cooperate to export the substrate into the cytoplasm (PDB ID: 1QBK; Chook and Blobel, 1999). I81 also has potential van der Waals' contact with nuclear transport factor 2 (3.548Å) (PDB ID: 1A2K; Stewart et al. 1998), which binds to Ran-GDP and is essential for efficient nuclear protein import. The Ran signature retrieved one false positive from TrEMBL.

Rho: Table 3.6 describes the structural and functional significance of Rho family signature sites. In the Rho signature, the Cysteine, Alanine, Glycine and Serine residues in site 83 are found only in Rit proteins of the Ras family and in a few Rab family proteins. Most other superfamily proteins have Leucine, Isoleucine, Valine or Methionine in this site. In the RhoA structure, this site has potential interactions with P-loop site 13 and switch II site

70 (PDB ID: 1CZX; Maesaki et al. 1999). Rho numbered site V83 in H-Ras has no interactions with the P-loop or switch II and V83 in Arf1 interacts with Rho numbered site K18, which H-bonds to GTP and GDP β and γ-phosphates (Sprang, 1997; PDB ID: 1HE8; Pacold et al. 2000; PDB ID: 1J2J; Shiba et al. 2003). Site W99 excludes the Sar, Arf, RGK and Ras families. Rho site W99 has possible van der Waals' contact with signature site 82. Rho numbered site R99 in Ras does not have contact with this site. Site 99 in Arf1 and H-Ras share a H-bond with site 95 (2.996Å and 3.042Å respectively) and have potential van der Waals' contact with differing sites in α4 (PDB ID: 1HE8; Pacold et al. 2000; PDB ID: 1J2J; Shiba et al. 2003). Site K70 excludes Sar, Arf, RGK and most Rab proteins. K70 forms a salt bridge with D67 (Å) and H-bonds to A61, which is critical for GTP hydrolysis and switch II conformational change (Paduch et al. 2001; Diaz et al. 1997). Rho numbered site W70 in Arf1 and A70 in Rab5 also H-bond to G61 (2.947Å and 2.950Å respectively). Arf 1 has an H-bond with R67 (2.966Å) but Rab does not. Rab site A70 H-bonds to other switch II sites (PDB ID: 1J2J; Shiba et al. 2003; PDB ID: 1TU3; Zhu et al. 2004)**.** The Rho signature retrieved one false positive from TrEMBL, Q34UL8_RHOPA, an ABC-type branched-chain amino acid transport protein (Boeckmann et al. 2003).

Rab: The Rab signature is shown in Table 3.7. No single site is diagnostic of the Rab family. The unique composition of Rab numbered site 62 in Ras, as discussed above distinguishes the Ras and Rab family proteins. Residues in Rab site 88 distinguish Rab from Sar family proteins. In Rab, site 88 H-bonds to N121, which H-bonds to the guanine ring of GDP and GTP (PDB ID: 1TU3; Zhu et al. 2004). In Sar these sites are too distant to H-bond (PDB ID: 1J2J; Shiba et al. 2003; PDB ID: 1M2O; Bi et al. 2002). Residues in site 79

distinguish Rab from Ran and RGK family proteins. Ran diagnostic site 81 is the equivalent site to site 79 in Rab. Both sites are protein binding. Site 79 in Rab family protein YPT1 H-bonds to its GDP dissociation inhibitor (GDI), while site 81 in Ran has potential van der Waals contact with Importin-β (PDB ID: 1TU3; Zhu et al. 2004; PDB ID: 1IBR; Vetter et al. 1999). Rho family proteins are distinguished from Rab family proteins through site 77, except for RhoD proteins which are distinguished through site 79. Both Rho and Rab have a main-chain H-bond with Rab numbered site 74 (2.796Å and 3.348Å respectively) in switch II but differ in their other interactions (PDB ID: 1CC0; Longenecker et al. 1999; PDB ID: 1TU3; Zhu et al. 2004).

RGK: Sites 42 and 53 in Table 3.8 discriminate the RGK family. R42 is the primary diagnostic site in the RGK signature. Also, site 53 contains Leucine or Isoleucine, which rarely appears in this site in other superfamily proteins. There are no solved crystal structures for RGK proteins in the PDB structural database (Berman et al. 2000). However, these sites in other Ras superfamily proteins are in the interswitch region, which is not directly involved in the proteins' biological function.

The Entropy Method was compared to profile hidden Markov models (HMMs), position-specific scoring matrices (PSSMs) and Prosite signatures, which can also delimit families of a superfamily and classify unknown superfamily proteins. Table 3.9 compares the results of these methods with the Entropy Method. The Rho Entropy signature matched one false positive, an ABC-type branched-chain amino acid transport system protein (Bairoch et al. 2005). The Rab Entropy signature identified the most known proteins and the Arf entropy signature identified the most unknown proteins. Otherwise, the Entropy signatures identified

fewer proteins than the best method. Excluding the families where the Entropy signatures classified the most proteins, they identified 4.4% fewer known proteins and 2.9% fewer unknown proteins than the method that classified the most known proteins overall. MEME classified the most known proteins and tied profile HMMs in the classification of the most unknown proteins. One advantage of probability based methods is their ability to detect protein fragments. Sequence signatures are limited to the region of the chosen sites.

Although Entropy signatures did not classify the most proteins in six of seven families (Table 3.9), they met all other criteria of this study. They found a minimal set of sites. Nearly all of the sites have critical functional or structural roles. Where structural information was available, a subset of these sites in each signature was shown to have unique structural or functional roles. Patterns of change were not extensively studied. However, comparison of discriminatory sites showed that changes occurred to provide unique switch conformations (Ras, Rho, Sar, Arf), maintain essential interactions with guanine nucleotide binding sites in the face of changing structure (Rab and Rho) and create a unique binding interaction (Ran).

MEME

MEME selects the most conserved contiguous sets of sites in each alignment to form the position-specific scoring matrices (PSSMs) by which it defines a set of proteins (Bailey and Elkan, 1994; Durbin et al. 1998). Table 3.10 lists the PSSMs for each of the families of the Ras superfamily. The PSSMs covering the nucleotide-binding P-loop and the protein binding switch II region have the most significant e-values (Sprang, 1997). Switch II binds

GTPase activating proteins (GAPs), which increase the hydrolysis rate and guanine nucleotide exchange factors (GEFs), which increase the GDP dissociation rate (Sprang, 1997; Crechet et al. 1996).

The switch II containing PSSM covers Ras numbered sites 56-68 in all families. These sites include the DXXG motif in Ras numbered sites 57-60. D57 coordinates Mg^{2+} and H-bonds to the γ-phosphate of GTP (Paduch et al. 2001; Sprang, 1997). The small methyl side-chain of Ras residue A59 and the hydrogen side-chain of G60 allow switch II loop L4 to change to the active conformation (Sprang, 1997; Diaz et al. 1997). A59 is also critical for GTP hydrolysis (Diaz et al. 1997), G60 is the critical site in the hinge that allows switch II conformation to change (Sprang, 1997) and Q61 is thought to be the catalytic base in GTP hydrolysis (Paduch et al. 2001).

MEME uses information theory to measure the contribution of each PSSM site in determining which proteins match the signatures. The amount of information in a site is determined by both the amount of conservation in the site and the background frequency of the conserved residues. The more conserved a site and the less frequent the conserved residue in the set of proteins, the more information the site yields. Sites with large information values are not necessarily biologically informative. For example, some sites in small datasets such as the Ran, Sar and RGK families are well conserved because of small sample size and species bias and not because they are critical to structure or function.

However, some informative sites, such as those described in the switch II PSSM, are important for structure or function. The most informative sites in the Arf switch I PSSM (data not shown) are Ras numbered sites W68, W56, and Y71 followed by Q61, which are

also functionally or structurally important sites. In Arf1-GTP, W68 has a potential H-bond with A59 (2.947Å) and thus has a structural role in positioning A59 for hydrolysis. W68 also has potential main-chain H-bonds with R65 (2.966Å) and with Y71 (2.896Å). Y71 is another informative signature site that appears to have van der Waals' contact with L182 of Arf1's effector protein GGA1 (3.935Å). W56 appears close to having van der Waals' contact with L190 of GGA1 (3.757Å) (PDB ID: 1J2J; Shiba et al. 2003).

Also, Ras numbered sites P34 and W56 are consistently informative across the families. The proline in site 34 bends T35 into the proper conformation to bind the γ-phosphate of GTP and to coordinate Mg^{2+} (Sprang, 1997). Site 56 is conserved as an aromatic residue in the Rab, Ran, Rho, Sar and Arf families and probably has structural significance in positioning residues D57 and G60. Superfamily conserved sites G10, G15, K16, N116, K117 and D119, which are critical to structure and nucleotide binding, are less informative than the aforementioned sites but are frequently present in the PSSMs of each family (Table 3.10).

MEME was the best performing method, classifying fewer known proteins in only two families and fewer unknown proteins in only one family (Table 3.9). Its ability to use multiple signatures is an advantage in protein classification but increases the difficulty in finding functionally and structurally discriminatory sites and the patterns of sequence change that led to diversification of the Ras superfamily. Further, MEME retrieved 103 known Arfs and 1231 known Rabs with e-values below the threshold of the first false positive (data not shown). These proteins were not included in the comparison in Table 3.9 because of their uncertain family membership. However, with evidence from database annotation,

phylogenetic trees, references and other sources, this method has much greater potential for family protein detection.

The Ran, RGK and Sar families, which have small datasets, have PSSMs of up to the maximum size allowed in this study, 50 sites (Table 3.10). In the Ras, Rho, Rab and Arf families, PSSMs encompass nearly all sites in the functional regions of the protein, which are the switch regions and nucleotide binding regions. So MEME does not find the minimal set of discriminatory sites. The PSSMs of these four families cover approximately the same sites and mostly have the same order when ranked by their e-values. Thus, patterns of differentiation among the families are difficult to detect. Finally, MEME's information measure is not a measure of unique structural or functional properties so there is no inherent way in the method to determine which sites are either biologically unique or relevant .

Profile HMMs

Profile HMMs were generated for each of the families. HMMs have an advantage over MEME's PSSMs in that they can handle gaps, thus using more of the information contained in the alignment (Durbin et al. 1998). HMMs are more automated and more probabilistically formal (Eddy, 1998) than the entropy method. They are also more probabilistically complex, so it is difficult to find biological meaning within the probability distribution of the profile sites.

The profile HMM method performed second best to the MEME method in detecting known and unknown family proteins. It retrieved greater than or equal to the number of known and unknown proteins retrieved by the Entropy Method except in the number of

known Rabs and unknown Arfs (Table 3.9). The Rab HMM retrieved one false positive, a hypothetical protein, which according to evidence in the phylogenetic tree is probably a Rap2 protein of the Ras family .

Sites which are significant to the detection of a family are not easily found in an HMM model. Such sites may be defined as those sites which have positive log-odds scores in only a few residues. It is preferred but not necessary that these residues be physiochemically related because such a relationship indicates a further structural or functional constraint within the site. Groups of physiochemically related residues include the aliphatic (A, G, L, I, V, and M), aromatic (F, W, and Y), acidic (D and E) and basic (H, K, and R) residues. In the Ras families these sites include the superfamily conserved sites, non-superfamily conserved P-loop sites in Sar and Arf, many of the functionally and structurally significant sites used by the Entropy Method and other physiochemically conserved sites.

Profile HMMs use every site in the protein. They measure the significance of the amount of conservation of each amino acid in each site. The profile HMM does not reveal discriminatory sites. So unique structural or functional sites cannot be found and patterns of functional and structural change cannot be found. At best, with additional analyses, they can be used to find the most conserved sites, which have the same limitations as the MEME sites of high information.

PROSITE

The PROSITE signatures retrieved the smallest number of known and unknown proteins of the four methods (Table 3.9). PROSITE signatures are only available for the Sar,

Arf and Ran families. Compared to the Entropy signatures, the PROSITE signature for the Sar family retrieved 19% fewer known family proteins and identified 32% fewer unknown family proteins. The signature has 20 sites not counting the sites where any amino acid is allowed. The Entropy signature for the Sar family has 12 sites. The longer Prosite signature contains sites that are not very informative about the structure and function of Sar. Also, a longer signature allows less flexibility in identifying new members of the family. One advantage to a longer signature, however, is a greater assurance of validity at present when many predicted proteins from newly sequenced genomes are being added to the protein databases.

The PROSITE signature for Sar is "R-x-[LIVM]-E-[LV]-F-[MPT]-C-S-[LIVM]-[LIVMY]-x-[KRQ]-x-G-Y-x-[DE]-[AG]-[FI]-x-W-[LIVM]-x-[NQK]-Y.", where each site is separated by a hyphen, *[LV]* means that the site can contain residues *L* or *V*, *x* means that the site can contain any amino acid, and a period terminates the signature. The signature starts N-terminal to β6 and ends within a few residues of the C-terminus. This includes the amphipathic helix α5, which covers the hydrophobic core (Huang et al. 2001). The significant functional interaction is the H-bond of V173 with the exocyclic oxygen of GTP (3.047Å). Significant structural interactions include the potential H-bonds between V173 and R176 (2.804Å) and between S172 and N177 (2.801Å and 2.941Å), which probably contribute to loop stability so the interaction of V173 and GTP can take place (PDB ID: 1M2O; Bi et al. 2002). The potential main-chain H-bonds between α5 sites Y179 and F183 (3.215Å), L180 and Q184 (2.921Å), E181 and W185 (2.843Å), F183 and S187 (2.994Å) and W185 and Q188 (2.869Å) may also contribute to loop stability as they do in Ras family

proteins (Zhang and Matthews, 1998a,b). Signature sites also interact with other parts of the protein. For example there is a salt bridge between sites R176 and D46 (3.100Å) and a potential H-bond between R164 and F117 (3.230Å). The signature sites in the α5 helix reveal a unique set of interactions that stabilize GTP binding.

The PROSITE signature for the Arf family retrieved 35% fewer known family proteins and identified 36% fewer unknown family proteins than the Entropy signature. Four false positives from the TrEMBL database matched the PROSITE signature: Q4QFH4_LEIMA's closest match in BLAST (Altschul et al. 1990) is a Dimethylaniline monoxygenase, Q4ARD5_9BURK is a UDP-N-acetylglucosamine pyrophosphorylase, Q3QTU1_9RHOB is a Cyclic nucleotide-binding ABC transporter and Q82X47_NITEU is an Ammonium transporter family (Rh-like) protein (Bairoch, et al. 2005).

The signature has 10 sites not counting the sites where any amino acid is allowed. The Entropy signature for the Arf family has 12 sites. The PROSITE signature for Arf is "[HRQT]-x-[FYWI]-x-[LIVM]-x(4)-A-x(2)-G-x(2)-[LIVM]-x(2)-[GSA]-[LIVMF]-x-[WK]-[LIVM].", where *x(i)* means any amino acid is allowed in *i* consecutive sites. The signature starts N-terminal to β7 and ends in the C-terminal helix, approximately the same region as the Sar signature. The significant functional interaction is the H-bond of Arf numbered site A160 with the exocyclic oxygen of GTP (2.941Å). Significant structural interactions include the potential H-bonds between Arf1 numbered sites C159 and S162 (3.187Å), C159 and G163 (2.912Å), and C159 and D164 (3.098Å). These interactions probably contribute to loop stability so the interaction of A160 and GTP can take place. The potential main-chain H-bonds between α5 sites G165 and G169 (3.015Å), L166 and L170 (2.995Å), Y167 and T171

(2.843Å), G169 and L173 (3.090Å) and E168 and W172 (3.079Å) may also contribute to loop stability as they do in Ras family proteins (PDB ID: 1E0S; Menetrey et al. 2000). The signature sites in the C-terminal helix reveal a unique set of interactions that stabilize GTP binding.

The PROSITE signature for the Ran family retrieved two fewer known family proteins and three fewer unknown family proteins. The signature has 16 sites not counting the sites where any amino acid is allowed. The Entropy signature for the Ran family has 8 sites. The PROSITE signature for Ran is "D-T-A-G-Q-E-[KR]-[LFY]-G-G-L-R-[DE]-G-Y-[YF].". The signature starts with the superfamily-conserved DXXG motif and ends in switch II.

The small methyl side-chain of A67 and the hydrogen side-chain of G68 allow switch II to change to the active conformation (Sprang, 1997; Diaz et al. 1997). A67 and Q69 are critical for GTP hydrolysis (Diaz et al. 1997; Paduch et al. 2001). D65 coordinates Mg^{2+} and H-bonds to the γ-phosphate of GTP (Paduch et al. 2001; Sprang, 1997). Binding sites include: K71 (2.757Å, 3.313Å) and R76 (2.659Å, 2.727Å), which form salt bridges with nuclear transport factor 2; site D77 (2.990Å), which potentially H-bonds to a nuclear transport factor; sites A67 (3.689Å), Q69 (3.500Å) and F72 (3.756Å), which have potential van der Waals' contact with a nuclear transport factor; G74 (3.292Å), which has a potential H-bond with the RanGEF RCC1 and L75 (3.239Å); and G78 (3.651Å), which has potential van der Waals' contacts with Importin-β (Stewart et al. 1998; Renault et al. 2001; Vetter et al. 1999). These binding sites reveal some of the unique functional properties of the Ran family.

PROSITE signatures classify approximately 65-80% as many proteins as Entropy

signatures. The PROSITE Sar and Ran signatures have eight more sites than the Entropy signature. While the Arf Entropy signature has two more sites. Elimination of the 5 C-terminal sites of the Sar signature results in the identification of eight additional Sar proteins. Thus PROSITE signatures do not have the minimal subset of sites and contain sites unnecessary for the discrimination of family proteins. Determination of which sites are essential for discrimination would require systematic trials of every subset of signature sites of reasonable size which is a difficult prospect. PROSITE's signature sites are functionally or structurally relevant and a subset of them were identified as having unique functional and structural properties. The sites were used to find patterns of functional and structural change. However, PROSITE signatures are only available for the more conserved Ras families, for which signatures are easier to obtain. It is uncertain if this is a limitation of the method. PROSITE profiles can be built for the Ras, Rho, Rab and RGK families, although these profiles are similar to MEME's PSSMs in their difficulties with the criteria of this study.

Conclusions

Entropy-based sequence signatures are an accurate way to identify both known and previously unknown proteins of the Ras families. Unlike the profile HMM and MEME methods of protein classification, the Entropy Method is a simple but powerful tool for the discovery of sites with unique functional and structural roles. Profile HMMs used all amino acid sites in the alignment MEME's PSSMs used five sets of 10-50 amino acid sites. The sites with the most information were identified using information theory. However, only a small subset of these sites uniquely identify the family in either sequence or biology and this

subset cannot be determined using the MEME method. But Entropy signatures used thirteen or fewer sites, of which an easily identifiable subset of uniquely conserved sites were structurally or functionally significant. The sites which identified the unique biological characteristics of the family belonged to the nucleotide-binding P-loop and switch I regions and the protein binding switch I and switch II regions. These sites revealed uniquenesses in effector binding, the Ran GTPase cycle, and sites with critical structural or physiochemical interactions with nucleotide-binding sites.

Entropy signatures were easily compared to gain insight into the diversification of structure and function in the Ras superfamily. The Isoleucine and Valine in Ran site 81 enable Importin-β binding, part of the unique Ran GTPase cycle. The aliphatic residues in Ras site 56 allow the switch I conformation necessary to bind the family's effectors. Residues in Rho site 83 have unique structural interactions with nucleotide and protein binding regions. In Sar and Arf proteins, Leucine and Isoleucine give Ras numbered site 11 a unique interaction with the catalytic site Q61. Rab sites 62, 77, 79, and 88 give Rab a unique set of interactions. In summary, comparative analyses of Entropy signature sites provided functional and structural information on the key changes that separate the families of the Ras superfamily. Use of these signatures in future studies will assist in understanding the evolution and function of Ras superfamily proteins.

We thank Heather Dellinger for her encouragement during the development of this study. This work was supported by a NSF Genomics IGERT fellowship, the NCSU Functional Genomics fellowship and WRA's NIH grant.

Table 3.1. Identification of TrEMBL Proteins Using Entropy-based Sequence Signatures.

Notes: 1) Signatures are written in the following format: Allowed residues are denoted by their one-letter IUPAC codes (IUPAC, 1968). Signature sites are separated by a dash (-). Sites that allow more than one residue contain the allowed residues in square brackets ([ ]). For example, the first site in the Rab signature allows F (Phenylalanine), W (Tryptophan), and R (Arginine). Sites that exclude a set of residues are denoted in curly braces ({ }). Sites that allow any residue are denoted by an x. x(i) denotes i adjacent sites that allow any residues and x(i-j) denotes that between i and j sites are allowed. 2) The site number of the first signature site. For consistent numbering, sites are numbered using the Ras family standard (Valencia et al. 1991). 3) Known proteins have evidence for their family membership in the Uniprot database (Bairoch et al. 2005) or the family's phylogenetic tree. 4) The Rho signature returned one false positive.

Family	Signature[1]	Starting Site[2]	Ending Site	Known proteins identified[3]	Unknown proteins identified
Arf	[ALIVT]-x(2)-[GS]-[LI]-[DENQ]-x-[AS]-G-K-[ST]-x(2)-[LIVM]-x(12-21)-[PSTA]-T-x-G.	7	37	376	52
Rab	[FWR]-D-[STIM]-[AGS]-G-x(3)-[FYGL]-x(2-3)-[LIVMH]-x(3)-[FYLQ]-x-[HKRETY]-x(8)-[FYH].	56	82	1336	198
Ras	T-[LIVMKR]-[AGEQ]-[DEN]-x-[FWYH]-x(15)-[LIVMTY]-D-[STA]-G.	35	60	336	25
Ran	D-T-x-G-x(6)-L-x(3)-Y-[FY]-[IV]-x(9)-D.	57	83	63	4
RGK	G-x(4)-G-K-[ST]-x-L-x(3)-F-x(17-20)-R-x-[LI]-x-V-x(2)-E-x(3)-[LI]-[LIVM].	10	54	12	0
Rho	D-x(2)-G-x(6,7)-[KR]-x(2-3)-[FY]-x(7)-[LIVM]-[CAGS]-[FY]-x(6)-[ST]-x(7)-W.	57	97	392[4]	58
Sar	G-L-D-N-A-G-K-[ST]-T-L-x(13)-P-x(3)-P.	10	38	53	25

Table 3.2. Functional and Structural Importance of Arf Family Sequence Signature Sites
Notes: 1) Arf1 site numbering according to PDB file 1J2J 2) PDB ID: 1J2J; Shiba et al. 2003 3) Sprang, 1997 4) 1J2J is a Q61L constitutively active mutant. There is no nonmutant GTP-bound structure with a related function (Arf or Arl1) in PDB. 4) Residues are divided into groups with common physiochemical properties: aliphatic (A, G, L, I, V and M), aromatic (F, W and Y), acidic (D and E), basic (H, K and R), hydroxyl (S and T), other hydrophilic (N and Q), Cysteine (C) and Proline (P).

Site[1]	Residue (%)	Functional/Structural Role[2,3]	Residue (Group) Entropy[4]
21	A (0), L (80), I (5), V (14), T (1)	H-bond with the potential switch II hinge G87 (2.899Å)[2]	0.22 (0.02)
24	G (99), S (1)	Flexibility in P-loop[3]	0.02 (0.02)
25	L (100), I (0)	H-bonds with A28 (3.320Å); possible van der Waals' contact with switch II sites L71 (3.661)[4] and Y81 (3.727Å)[2]	0 (0)
26	D (100), Q (0)	H-bond with switch II hinge G70 (3.174Å), switch II site R75 (3.301Å)[2] and R99 (2.932Å)	0 (0)
28	A (95), S (5)	H-bond with N126 (2.994Å), which H-bonds to the guanine ring; Potential H-bond with β-PO_4 of GTP (3.230Å)	0.069 (0.10)
29	G (100)	H-bond to GTP/GDP α-PO_4 [3]	0 (0)
30	K (100)	H-bond to GTP/GDP β,γ-PO_4 [3]	0 (0)
31	S (2), T (98)	H-bond to γ-PO_4; coordinates Mg^{2+} [3]	0.03 (0)
34	L (95), I (2), V (2), M (1)	H-bond to K30 (2.882Å); possible van der Waals' contact with N52 (3.599Å), E54 (3.895Å) and V65(3.799) in β2 and β3	0.09 (0)
47	A (0), P (96), S (0), T (4)	Structural positioning of site 48 to bind GTP and Mg^{2+}	0.05 (0.07)
48	T (100)	H-bond to γ-PO_4; coordinates Mg^{2+} [3]	0 (0)
50	G (100)	May be switch I hinge	0 (0)

Table 3.3. Functional and Structural Importance of Sar Family Sequence Signature Sites
Notes: 1) Sites numbered by PDB structure 1M2O (Sar1-GTP). 2) PDB ID: 1M2O; Bi et al. 2002 3) Sprang, 1997 4) Assuming the Oxygen is in the same position as the Nitrogen between β and γ-PO_4 of GNP. 5) Residues are divided into groups with common physiochemical properties: aliphatic (A, G, L, I, V and M), aromatic (F, W and Y), acidic (D and E), basic(H, K and R), hydroxyl (S and T), other hydrophilic (N and Q), Cysteine (C) and Proline (P).

Site [1]	Residue (%)	Functional/Structural Role[2,3]	Residue (Group) Entropy
30	G (100)	Critical for P-loop structure	0 (0)
31	L (100)	Possible van der Waals' contact with switch II sites R81(3.666Å), H77 (3.758Å), G76 (3.673Å). G76 is the hinge. Sites 31-35 are the P-loop.[2]	0 (0)
32	D (100)	Salt bridges with R81(3.429Å) and R105 (3.083Å, 3.117Å)[2]	0 (0)
33	N (100)	Backbone N H-bonds to β-PO_4 (2.872Å)[2,4]	0 (0)
34	A (100)	H-bond with L31 (3.356Å); van der Waals' contact with L97 (3.811Å)[2]	0 (0)
35	G (100)	H-bond to GTP/GDP α-PO_4 [3]	0 (0)
36	K (100)	H-bond to GTP/GDP β,γ-PO_4 [3]	0 (0)
37	T (94), S (6)	H-bond to γ-PO_4; coordinates Mg^{2+} [3]	0.07 (0)
38	T (100)	Backbone N H-bond to α-PO_4 (2.669Å)[2]	0 (0)
39	L (100)	H-bond with G35 (3.069Å); possible van der Waals' contact with V173 (3.799Å)[2]	0 (0)
53	P (100)	Switch I; brings T30 into position for GTP, Mg^{2+} binding	0 (0)
57	P (100)	Switch I; Unknown structural importance	0 (0)

Table 3.4. Functional and Structural Importance of Ras Family Sequence Signature Sites
Notes: 1) Sites numbered by PDB structure 121p (H-Ras-GTP). 2) PDB ID: 121p; Wittinghofer et al. 1997 3) Sprang, 1997 4) PDB ID: 1HE8; Pacold et al. 2000 5) PDB ID: 1BKD; Boriack-Sjodin et al. 1998 5) Díaz et al. 1997; Kontani et al. 2002 6) Residues are divided into groups with common physiochemical properties: aliphatic (A, G, L, I, V and M), aromatic (F, W and Y), acidic (D and E), basic (H, K and R), hydroxyl (S and T), other hydrophilic (N and Q), Cysteine (C) and Proline (P).

Site[1]	**Residue (%)**	**Functional/Structural Role[2,3]**	**Residue (Group) Entropy[7]**
35	T (100)	H-bond to γ-PO_4; coordinates Mg^{2+} [3]	0 (0)
36	L (2), I (86), V (4), M (0), K (8), R (0)	Switch I site; possible van der Waals' contact with effector PI3Kγ (3.275Å)[4]	0.18 (0.13)
37	G (0), A (8), E (91), Q (1)	Switch I site; H-bond with effector PI3Kγ (3.301Å)[4] and GEF SOS1 (3.078Å)[5]	0.11 (0.15)
38	D (91), E (1), N (6)	Switch I site	0.14 (0.17)
40	F (5), W (0), Y (89), H (6)	Possible van der Waals' contact with GEF SOS1 (3.378Å)[5]	0.15 (0.11)
56	L (89), I (2), V (2), M (0), T (6), Y (1)	Aromatic residues and long chains do not appear to fit in Ras' conformation[2]	0.16 (0.13)
57	D (100)	Coordinates Mg^{2+} in GTP and H_2O to Mg^{2+} in GDP[3]	0 (0)
58	T (98)	Possible van der Waals' contact with Y71 (3.777Å)[2]	0.03 (0.02)
59	A (88), S (5), T (6)	Essential for GTP hydrolysis and switch II active conformation[6]	0.16 (0.19)
60	G (100)	Switch II hinge; allows active conformation[3]	0 (0)

Table 3.5. Functional and Structural Importance of Ran Family Sequence Signature Sites
Notes: 1) Sites numbered by Ran's PDB structure. 2) PDB ID: 1IBR; Vetter et al. 1999 3) Sprang, 1997 4) Residues are divided into groups with common physiochemical properties: aliphatic (A, G, L, I, V and M), aromatic (F, W and Y), acidic (D and E), basic(H, K and R), hydroxyl (S and T), other hydrophilic (N and Q), Cysteine (C) and Proline (P).

Site[1]	**Residue (%)**	**Functional/Structural Role[2,3]**	**Residue (Group) Entropy[4]**
65	D (100)	Coordinates Mg^{2+} in GTP and H_2O to Mg^{2+} in GDP[3]	0 (0)
66	T (100)	unknown	0 (0)
68	G (100)	Switch II hinge; allows active conformation[3]	0 (0)
75	L (100)	Switch II; possible van der Waals' contact with Importin-β (3.239Å)[2]	0 (0)
79	Y (100)	Switch II; possible H-bond to L75 (3.157Å)	0 (0)
80	F (3), Y (97)	Switch II; unknown	0.04 (0)
81	I (95), V (5)	Switch II; possible van der Waals' contacts with Importin-β (3.592Å, 3.632Å)[2]	0.07 (0)
91	D (100)	H-bonds to K123 (2.669Å). K123 H-bonds to the guanine ring[3]	0 (0)

Table 3.6. Functional and Structural Importance of Rho Family Sequence Signature Sites

Notes: 1) Sites numbered by the PDB structure 1CXZ (RhoA). 2) PDB ID: 1CXZ; Maesaki et al. 1999 3) Sprang, 1997 4) Paduch et al. 2001; Diaz et al. 1997 5) Residues are divided into groups with common physiochemical properties: aliphatic (A, G, L, I, V and M), aromatic (F, W and Y), acidic (D and E), basic(H, K and R), hydroxyl (S and T), other hydrophilic (N and Q), Cysteine (C) and Proline (P).

Site[1]	Residue (%)	Functional/Structural Role[2,3]	Residue (Group) Entropy[5]
59	D (100)	Coordinates Mg^{2+} in GTP and H_2O to Mg^{2+} in GDP[3]	0 (0)
62	G (100)	Switch II hinge; allows active conformation[3]	0 (0)
70	K (1), R (99)	H-bonds to A61 (2.710Å), which is critical for GTP hydrolysis and flexibility in Ras family proteins[4]; salt bridge with switch II site D67 (2.628Å)[2]	0.02 (0)
74	F (2), Y (98)	H-bond with switch II site T77 (3.164Å), possible van der Waals' contact with switch II site P71 (3.444Å)[2]	0.03 (0)
82	L (28), I (43), V (16), M (13)	H-bonds with I113 (2.988Å), possible van der Waals' contacts with I112 (3.735Å) and W99 (3.862Å)[2]	0.43 (0)
83	C (85), A(13), G (2), S (0)	H-bond with P-loop site D13 (2.767Å); possible van der Waals' contact with switch II site R70 (3.693Å)[2]	0.17 (0.22)
84	F (91), Y (9)	Possible van der Waals' contacts with L114 (3.655Å), V115 (3.387Å) and G116 (3.958Å)[2] which position the NKxD motif for guanine binding[3]	0.10 (0)
91	S (94), T (6)	Unknown; H-bonds with S85 (2.808Å) and S88 (3.103Å)[2]	0.08 (0)
99	W (100)	possible van der Waals' contact with M82 (3.862Å)[2]	0 (0)

Table 3.7. Functional and Structural Importance of Rab Family Sequence Signature Sites

Notes: 1) Sites are numbered by the PDB structure 1UKV (YPT1-GDP). 2) PDB ID: 1UKV; Rak et al. 2003 3) Sprang, 1997 4) Paduch et al. 2001 5) Diaz et al. 1997 6) Residues are divided into groups with common physiochemical properties: aliphatic (A, G, L, I, V and M), aromatic (F, W and Y), acidic (D and E), basic (H, K and R), hydroxyl (S and T), other hydrophilic (N and Q), Cysteine (C) and Proline (P).

Site[1]	Residue (%)	Functional/Structural Role	Residue (Group) Entropy[6]
62	W (99), F (1), R (0)	Possible van der Waals' interaction with GDP dissociation inhibitor (3.709Å) and K10 (3.755Å), the first superfamily conserved residue[2]	0.02 (0.01)
63	D (100)	Coordinates Mg^{2+} in GTP and H_2O to Mg^{2+} in GDP[3]	0 (0)
64	S (1), T (96), I (3), M (0)	H-bonds to Q67 (3.035Å), the putative catalyst for GTP hydrolysis[4]	0.06 (0.07)
65	A (96), G (2), S (2)	Switch II flexibility and GTP hydrolysis in Ras family proteins[4,5]	0.06 (0.05)
66	G (100)	H-bonds to γ-PO_4; critical pivot for Switch II[3]	0 (0)
70	F (70), Y (29), G (1), L (0)	Possible van der Waals' interaction with GDP dissociation inhibitor (3.788Å and 3.779Å)[2]	0.23 (0.04)
73	L (30), I (51), V (10), M (9), H (0)	Possible van der Waals' interaction with Q67 (3.500Å), the putative catalyst for GTP hydrolysis[4,5]	0.39 (0.01)
77	F (13), Y (86), L (1), Q (0)	Possible van der Waals' interaction with switch II site I73 (3.621Å); H-bond with switch II site T74 (3.348Å)[2]	0.15 (0.04)
79	H (1), K (2), R (95), E (1), T (0), Y (1)	H-bonds to GDP dissociation inhibitor (2.985Å and 3.031Å)[2]	0.09 (0.06)
88	F (17), Y (83), H (0)	H-bonds to N121 (2.796Å), which H-bonds to the guanine ring[2,3]	0.16 (0.01)

Table 3.8. Functional and Structural Importance of RGK Family Sequence Signature Sites

Notes: 1) Sites are numbered by the Ras family standard (Valencia et al. 1991). 2) There is no crystal structure for a RGK protein. No functional data was found in the literature except for site 17. 3) Sprang, 1997 4) Zhu et al. 1999 5) Residues are divided into groups with common physiochemical properties: aliphatic (A, G, L, I, V and M), aromatic (F, W and Y), acidic (D and E), basic (H, K and R), hydroxyl (S and T), other hydrophilic (N and Q), Cysteine (C) and Proline (P).

Site[1]	Residue (%)	Functional/Structural Role[2]	Residue (Group) Entropy[5]
10	G (100)	Superfamily conserved; critical for P-loop structure	0 (0)
15	G (100)	H-bond to GTP/GDP α-PO_4[3]	0 (0)
16	K (100)	H-bond to GTP/GDP β,γ-PO_4 [3]	0 (0)
17	S (82), T (18)	H-bond to γ-PO_4; coordinates Mg^{2+} [3]; essential for GTP binding[4]	0.16 (0)
19	L (100)	Unknown	0 (0)
23	F (100)	Unknown	0 (0)
42	R (100)	Unknown	0 (0)
44	L (45), I (55)	Unknown	0.23 (0)
46	V (100)	Unknown	0 (0)
49	E (100)	Unknown	0 (0)
53	L (73), I (27)	Unknown	0.20 (0)
54	L (9), I (45), V (27), M (18)	Unknown	0.4142 (0)

Table 3.9. Comparison of the Four Methods In Identification of Family Proteins
Notes: 1) Known proteins are defined as described in Materials and Methods. 2) PROSITE has only defined motifs for the Sar, Arf and Ran proteins.

Family	MEME Known[1]	Entropy Known	HMM Known	PROSITE Known[2]	MEME Unknown	Entropy Unknown	HMM Unknown	PROSITE Unknown
Arf	362	376	383	240	46	52	46	33
Ras	343	336	305	N/A	26	25	23	N/A
Ran	73	63	67	61	5	4	5	3
Rab	1214	1336	1286	N/A	203	198	207	N/A
RGK	15	12	14	N/A	0	0	0	N/A
Rho	436	392	433	N/A	59	58	59	N/A
Sar	57	53	53	44	26	25	25	10

Table 3.10. MEME PSSMs and Their Significance
Notes: 1) Sites are numbered beginning from the first superfamily conserved site, which is Ras family site 5 (Sprang, 1997). Subsequent sites are numbered according to the family standard seen in the multiple alignment. 2) The sites cover all or part of the listed regions. 3) MEME generates signatures in the order they are found, not necessarily according to the significance of the e-value (Bailey and Gribskov, 1998).

Family	Signature Number	Sites[1]	Significant Region(s)[2]	E-value[3]
Arf	1	48-68	Interswitch and switch II	3.3e-1858
	2	5-25	P-loop	2.5e-1653
	3	26-46	Switch I and interswitch	8.3e-1486
	4	114-124	NKxD motif loop	2.9e-845
	5	71-85		3.0e-1075
Sar	1	12-61	P-loop, switch I, interswitch	1.5e-643
	2	147-175	[ST]A motif loop	6.9e-374
	3	62-102	Switch II	3.5e-398
	4	105-115	NKxD motif loop	1.0e-111
	5	117-137		6.1e-129
RGK	1	109-158	NKxD and [ST]A motif loops	1.5e-350
	2	65-105	Switch II	1.6e-248
	3	33-60	Switch I and interswitch	1.8e-131
	4	5-25	P-loop	1.6e-94
	5	160-167		1.3e-32
Ran	1	57-106	Switch I	1.7e-1875
	2	107-156	NKxD and [ST]A motif loops	2.5e-1726
	3	5-54	P-loop, switch I, interswitch	2.2e-1538
	4	157-164		2.5e-162
Ras	1	13-29	P-loop	1.8e-1465
	2	54-68	Interswitch and switch II	1.8e-1341
	3	32-42	Switch I	1.7e-1079
	4	70-90	Switch II	1.6e-1682
	5	109-124	NKxD motif loop	7.3e-1191
Rho	1	55-69	Switch II	6.5e-1539
	2	10-24	P-loop	1.0e-1381

Table 3.10 (Continued)

Family	Signature Number	Sites[1]	Significant Region(s)[2]	E-value[3]
Rho	3	32-42	Switch I and interswitch	2.8e-914
	4	70-84	Switch II	2.6e-1144
	5	96-105		1.3e-683
Rab	1	52-66	Interswitch and switch II	1.5e-3269
	2	9-19	P-loop	5.6e-1815
	3	72-82	Switch II	1.7e-1668
	4	110-120	NKxD motif loop	1.3e-1465
Rab	5	142-152	[ST]A motif loop	1.9e-1411

Literature Cited

http://motif.genome.jp/MOTIF2.html

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403-410.

Atchley, W. R., W. Terhalle, and A. Dress. 1999. Positional Dependence, Cliques, and Predictive Motifs in the bHLH Protein Domain. J. Mol. Evol. **48**:501–516.

Atchley W. R., and A. D. Fernandes. 2005. Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network. Proc. Natl. Acad. Sci. USA **102**:6401–6406.

Bailey, T. L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California.

Bailey, T. L., and M. Gribskov. 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**:48-54.

Bairoch, A., R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. 2005. The Universal Protein Resource (UniProt). Nucleic Acids Res. **33**:D154-D159.

Banerjee-Basu, S., and A. D. Baxevanis. 2001. Molecular evolution of the homeodomain family of transcription factors. Nucleic Acids Research **29**:3258-3269.

Barker, W. C., J. S. Garavelli, P. B. McGarvey, C. R. Marzec, B. C. Orcutt, G. Y. Srinivasarao, L. L. Yeh, R. S. Ledley, H. Mewes, F. Pferffer, A. Tsugita, and C. Wu. 1999. The PIR-International Protein Sequence Database. Nucleic Acid Res. **27**:39–43.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. 2002. The Pfam Protein Families Database. Nucleic Acids Res. **30**:276-280.

Benson, D. A., Ilene Karsch-Mizrachi, D. J. Lipman, J. Ostell, and David L. Wheeler. 2004. GenBank: update. Nucleic Acids Res. **32**:D23–D26.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N.Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Research 28:235-242.

Bi, X., R. A. Corpina, and J. Goldberg. 2002. Structure of the Sec23/24-Sar1 pre-budding complex of the COPII vesicle coat. Nature. **419**:271-277. PDB ID: 1M2O.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**:365–70.

Boehm, M., R. C. Aguilar, and J. S. Bonifacino. 2001. Functional and physical interactions of the adaptor protein complex AP-4 with ADP-ribosylation factors (ARFs). EMBO J. **20**:6265-6276.

Boriack-Sjodin, P.A., S. M. Margarit, D. Bar-Sagi, and J. Kuriyan. 1998 . The structural basis of the activation of Ras by Sos. Nature **394**:337-343. PDB ID: 1BKD.

Chen, L., E. DiGiammarino, X. E. Zhou, Y. Wang, D. Toh, T. W. Hodge, and E. J. Meehan. 2004. High Resolution Crystal Structure of Human Rab9 GTPase A NOVEL ANTIVIRAL DRUG TARGET. J. Biol. Chem. **279**:40204–40208.

Chook YM, and Blobel G. 1999. Structure of the nuclear transport complex karyopherin-beta2-Ran x GppNHp. Nature. **399**:230-237. PDB ID: 1QBK

Créchet, J. B., A. Bernardi, and A. Parmeggiani. 1996. Distal switch II region of Ras2p is required for interaction with guanine nucleotide exchange factor. J Biol Chem. **271**:17234-17240.

Crespo, P., and J. León. 2000. Ras proteins in the control of the cell cycle and cell differentiation. Cell. Mol. Life Sci. **57**:1613-1636.

Dellinger, A. E., and W. R. Atchley. 2006. Phylogenetic Analysis of Evolution in the Ras Superfamily. *Unpublished.*

Díaz, J. F., B. Wroblowski, J. Schlitter, and Y. Engelborghs. 1997. Calculation of pathways for the conformational transition between the GTP- and GDP-bound states of the Ha-ras-p21 protein: calculations with explicit solvent simulations and comparison with calculations in vacuum. Proteins **28**:434-51.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK.

Eddy, S. R. 1998. Profile Hidden Markov Models. Bioinformatics **14**:755-763.

Eddy, S. R. 2001. HMMER: Profile hidden Markov models for biological sequence analysis. http://hmmer.wustl.edu.

Gaurav, K., N. Gupta, and R. Sowdhamini. 2005. FASSM: Enhanced Function Association in whole genome analysis using Sequence and Structural Motifs. In Silico Biol. **5**:0040.

Gonnet, G. H., M. A. Cohen, and S. A. Benner. 1992. Exhaustive matching of the entire protein sequence database. Science **256**:1433-1445.

Henikoff, S., and J. G. Henikoff. 1994. Protein family classification based on searching a database of blocks. Genomics. **19**:97-107.

Henikoff, S. 1996. Scores for sequence searches and alignments. Current Opinion in Structural Biology **6**:353-360.

Huang, M., J. T. Weissman, S. Béraud-Dufour, P. Luan, C. Wang, W. Chen, M. Aridor, I. A. Wilson, and W. E. Balch. 2001. Crystal structure of Sar1-GDP at 1.7 Å resolution and the role of the NH_2 terminus in ER export. The Journal of Cell Biology **155**:937-948. PDB ID: 1M2O

Isoldi, M. C., M. A. Visconti, and A. M. de Lauro Castrucci. 2005. Anti-cancer drugs: molecular mechanisms of action. Mini Rev. Med. Chem. **5**:685-95.

IUPAC-IUB Commission on Biochemical Nomenclature (CBN). 1968. A One-Letter Notation for Amino Acid Sequences. Arch. Biochem. Biophys. **125**:i-v.

Jeanmougin F., J. D. Thompson, M. Gouy, D. G. Higgins, and T. J. Gibson. 1998. Multiple sequence alignment with Clustal X. Trends Biochem. Sci. **23**:403-5.

Kontani, K., M. Tada, T. Ogawa, T. Okai, K. Saito, Y. Araki, and T. Katada. 2002. Di-Ras, a Distinct Subgroup of Ras Family GTPases with Unique Biochemical Properties. J. Biol. Chem. **277**:41070-41078.

Li, A.-J., and R. Nussinov. 1998. A Set of van der Waals and Coulombic Radii of Protein Atoms for Molecular and Solvent-Accessible Surface Calculation, Packing Evaluation, and Docking. PROTEINS: Structure, Function, and Genetics **32**:111–127.

Longenecker, K., P. Read, U. Derewenda, Z. Dauter, X. Liu, S. Garrard, L. Walker, A. V Somlyo, R. K. Nakamoto, A. P. Somlyo, and Z. S. Derewenda. 1999. How RhoGDI binds Rho. Acta Crystallogr. D. Biol. Crystallogr. **55**:1503-1515.

Maesaki, R., K. Ihara, T. Shimizu, S. Kuroda, K. Kaibuchi, and T. Hakoshima. 1999. The structural basis of Rho effector recognition revealed by the crystal structure of human RhoA complexed with the effector domain of PKN/PRK1. Mol. Cell. **4**:793-803. PDB ID: 1CXZ.

Meinnel, T., C. Lazennec, S. Villoing, and S. Blanquet. 1997. Structure-function relationships within the peptide deformylase family. Evidence for a conserved architecture of the active site involving three conserved motifs and a metal ion. J. Mol. Biol. **267**:749-761.

Menetrey, J., E. Macia, S. Pasqualato, M. Franco, and J. Cherfils. 2000. Structure of Arf6-GDP suggests a basis for guanine nucleotide exchange factors specificity. Nat Struct Biol. **7**:466-469. PDB ID: 1E0S.

Milburn, M. V., L. Tong, A. M. deVos, A. Brunger, Z. Yamaizumi, S. Nishimura, and S. H. Kim. 1990. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. Science **247**:939-45.

Morgenstern, B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15:**211-218.

Mullington, H. 2000. Windows Grep 2.2.1.2222

Myers, G. A. 1992. Four Russians Algorithm for Regular Expression Pattern Matching. Journal of the Association for Computing Machinery. **39**:431-448.

Newton, A. C. 1995. Protein Kinase C: Structure, Function, and Regulation. J. Biol. Chem. **270**:28495-28498.

Nicholas, K. B., H. B. Nicholas Jr., and D. W. Deerfield II. 1997. GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW.NEWS **4**:14.

Oxford, G., and D. Theodorescu. 2003. Ras superfamily monomeric G proteins in carcinoma cell motility. Cancer Letters **189**:117-128.

Pacold, M. E., S. Suire, O. Perisic, S. Lara-Gonzalez, C. T. Davis, E. H. Walker, P. T. Hawkins, L. Stephens, J. F. Eccleston, and R. L. Williams. 2000. Crystal structure and functional analysis of Ras binding to its effector phosphoinositide 3-kinase gamma. Cell **103**:931-43. PDB ID: 1HE8.

Paduch, M., F. Jeleñ, and J. Otlewski. 2001. Structure of small G proteins and their regulators. Acta Biochim. Pol. **48**:829-850.

Pasqualato, S., L. Renault, and J. Cherfils. 2002. Arf, Arl, Arp and Sar proteins: a family of GTP-binding proteins with a structural device for 'front–back' communication. *EMBO Reports* **3**:1035-1041.

Pellegrini-Calace, M., and J. M. Thornton. 2005. Detecting DNA-binding helix-turn-helix structural motifs using sequence and structure information. Nucleic Acids Res. **33**:2129-40.

Protein Research Foundation, Osaka, Japan. http://www4.prf.or.jp/en/seqsrch.html

Rak, A., O. Pylypenko, T. Durek, A. Watzke, S. Kushnir, L. Brunsveld, H. Waldmann, R. S. Goody, and K. Alexandrov. 2003. Structure of Rab GDP-dissociation inhibitor in complex with prenylated YPT1 GTPase. Science. **302**:646-650. PDB ID: 1UKV.

Renault, L., J. Kuhlmann, A. Henkel, and A. Wittinghofer. 2001. Structural basis for guanine nucleotide exchange on Ran by the regulator of chromosome condensation (RCC1). Cell. **105**:245-255. PDB ID: 1I2M.

Shiba T., M. Kawasaki, H. Takatsu, T. Nogi, N. Matsugaki, N. Igarashi, M. Suzuki, R. Kato, K. Nakayama, and S. Wakatsuki. 2003. Molecular mechanism of membrane recruitment of GGA by ARF in lysosomal protein transport. Nat. Struct. Biol. **10**:386-393. PDB ID: 1J2J.

Sigrist, C. J. A., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. **3**:265-274.

Sprang, S. R. 1997. G proteins, effectors and GAPs: structure and mechanism. Curr. Opin. Struct. Biol. **7**:849-56.

Stewart, M., H. M. Kent, and A. J. McCoy. 1998. Structural basis for molecular recognition between nuclear transport factor 2 (NTF2) and the GDP-bound form of the Ras-family GTPase Ran. J. Mol. Biol. **277**:635-646. PDB ID: 1A2K.

Tang, X., S. Orlicky, Q. Liu, A. Willems, F. Sicheri, and M. Tyers. 2005. Genome-Wide Surveys for Phosphorylation-Dependent Substrates of SCF Ubiquitin Ligases. Methods Enzymol. **399**:433-58.

Thompson, J. D., D. G. Higgins, and T. J. Gibson 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**: 4673–4680.

Valencia, A., P. Chardin, A. Wittinghofer, and C. Sander. 1991. The ras protein family: evolutionary tree and role of conserved amino acids. Biochemistry **30**:4637-4648.

Vetter, I. R., A. Arndt, U. Kutay, D. Gorlich, and A. Wittinghofer. 1999. Structural view of the Ran-Importin beta interaction at 2.3 A resolution. Cell. **97**:635-646. PDB ID: 1IBR.

Vogt, G., T. Etzold, and P. Argos. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J. Mol. Biol. **249**:816-831.

Wennerberg, K., K. L. Rossman, and C. J. Der. 2005. The Ras superfamily at a glance. Journal of Cell Science **118**:843-846.

Wittinghofer, F., U. Krengel, J. John, W. Kabsch, and E. F. Pai. 1991. Three-dimensional structure of p21 in the active conformation and analysis of an oncogenic mutant. Environ. Health Perspect. **93**:11-15. PDB ID: 121p.

Zhang, J., and C. R. Matthews. 1998. The role of ligand binding in the kinetic folding mechanism of human p21(H-ras) protein. Biochemistry **37**:14891-14899.

Zhang, J., and C. R. Matthews. 1998. Ligand binding is the principal determinant of stability for the p21(H)-ras protein. Biochemistry **37**:14881-90.

Zhu, G., P. Zhai, J. Liu, S. Terzyan, G. Li, and X. C. Zhang. 2004. Structural basis of Rab5-Rabaptin5 interaction in endocytosis. Nat Struct Mol Biol. **11**:975-83.

Zhu, J., Y.-H. Tseng, J. D. Kantor, C. J. Rhodes, B. R. Zetter, J. S. Moyers, and C. R. Kahn. 1999. Interaction of the Ras-related protein associated with diabetes Rad and the putative tumor metastasis suppressor NM23 provides a novel mechanism of GTPase regulation. Cell Biology **96**:14911-14918.

Appendices

Appendix A: Ras Family Phylogenetic Trees and Ancestral Reconstructions

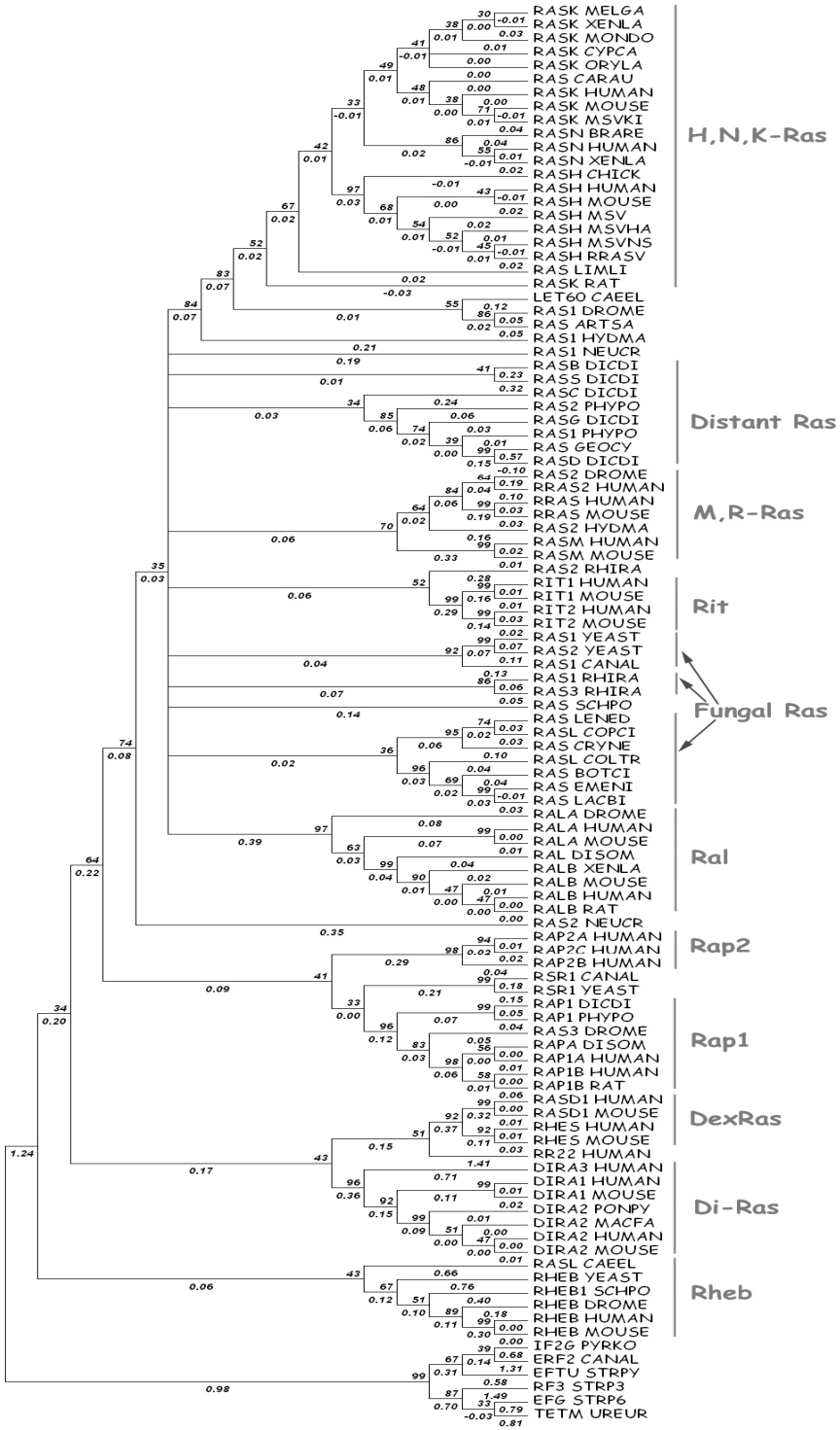


Figure A.1. Neighbor-Joining tree of the Ras Family Using the Poisson Process. The tree was condensed at a bootstrap threshold of 30%.

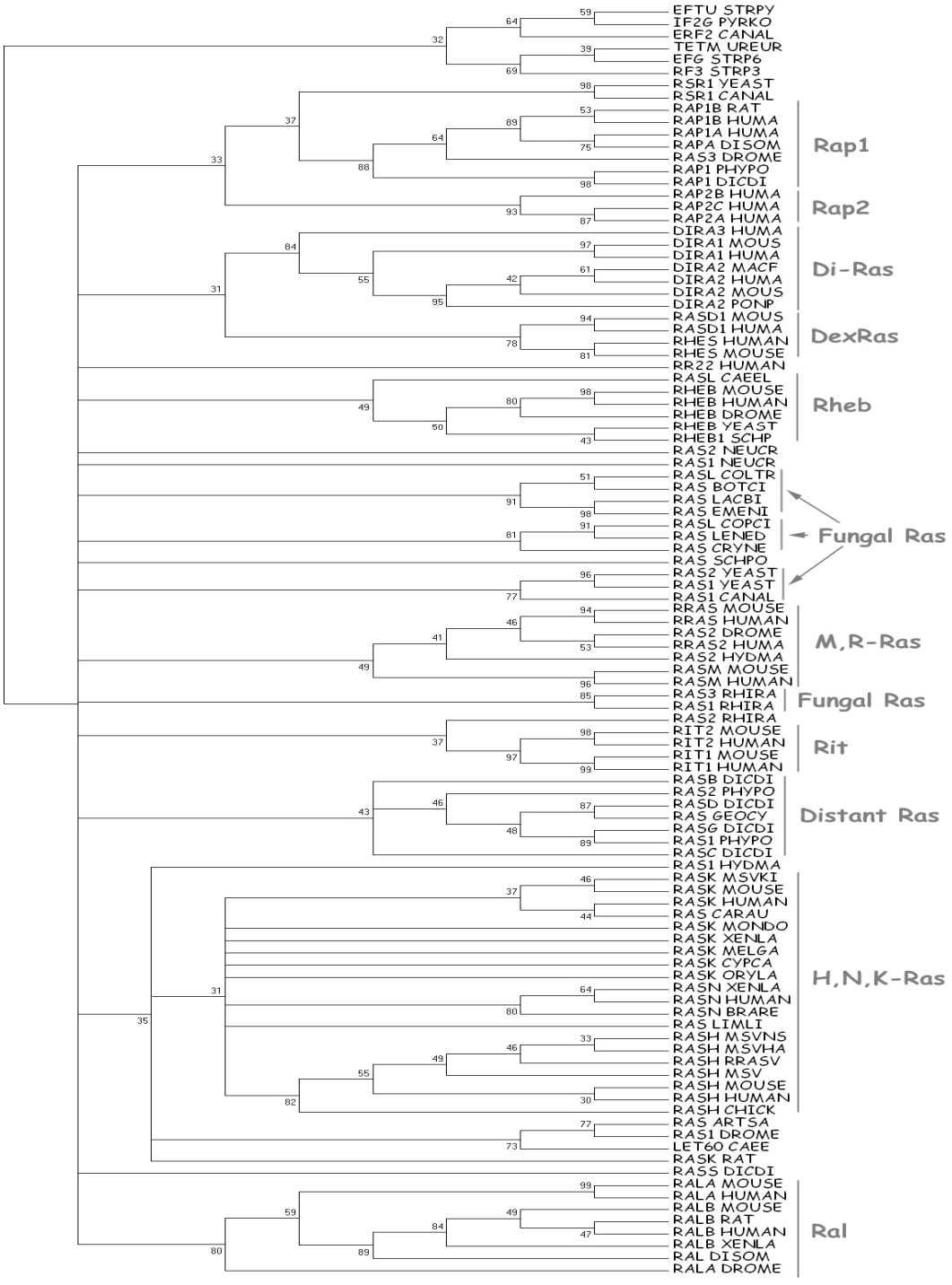


Figure A.2. Neighbor-Joining Tree of the Ras Family Using the JTT Substitution Matrix.

The tree was condensed at a bootstrap threshold of 30%.

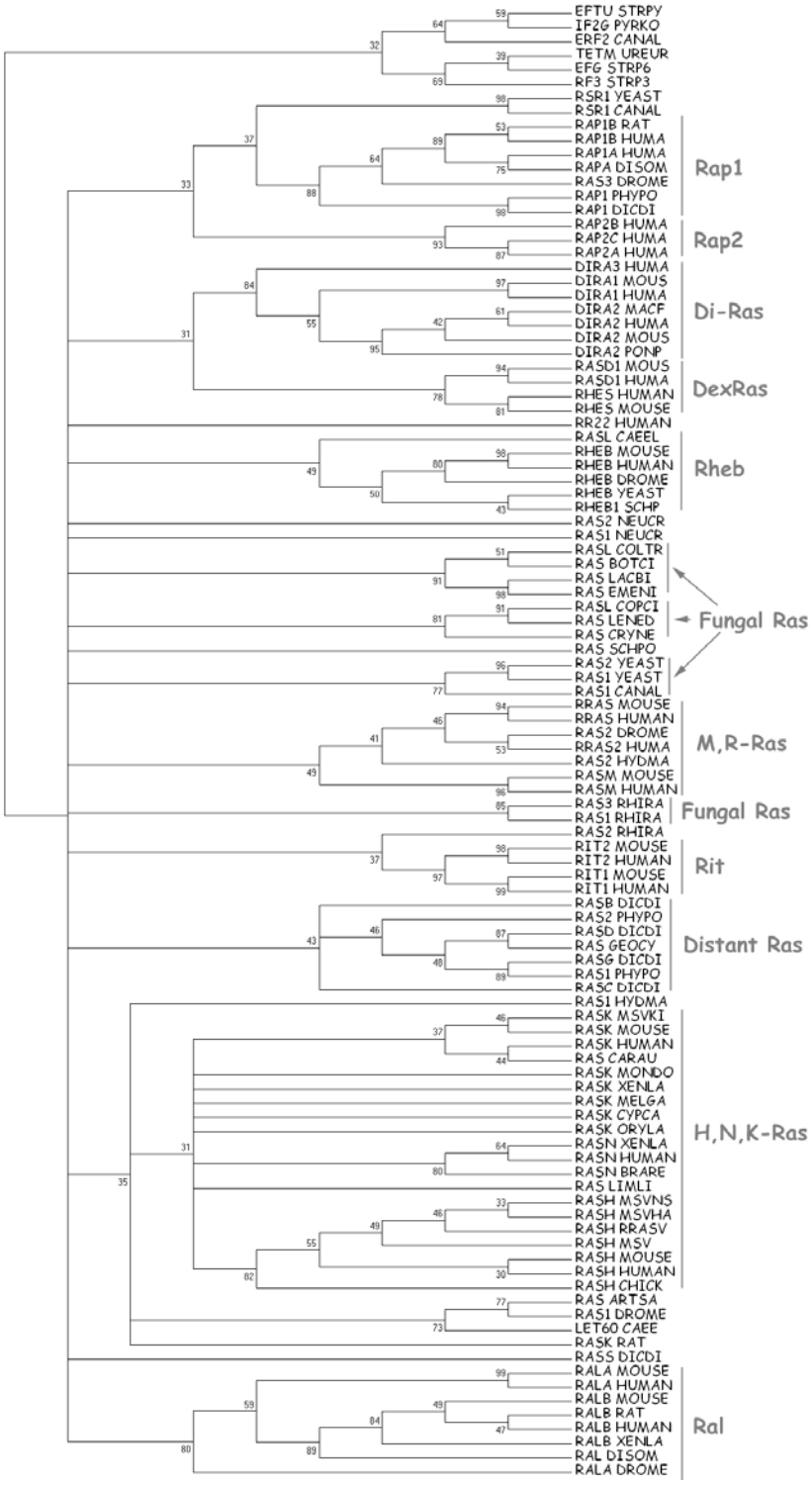


Figure A.3. Weighbor Consensus Tree.
A 30% consensus of 1000 trees built from bootstrapped datasets.

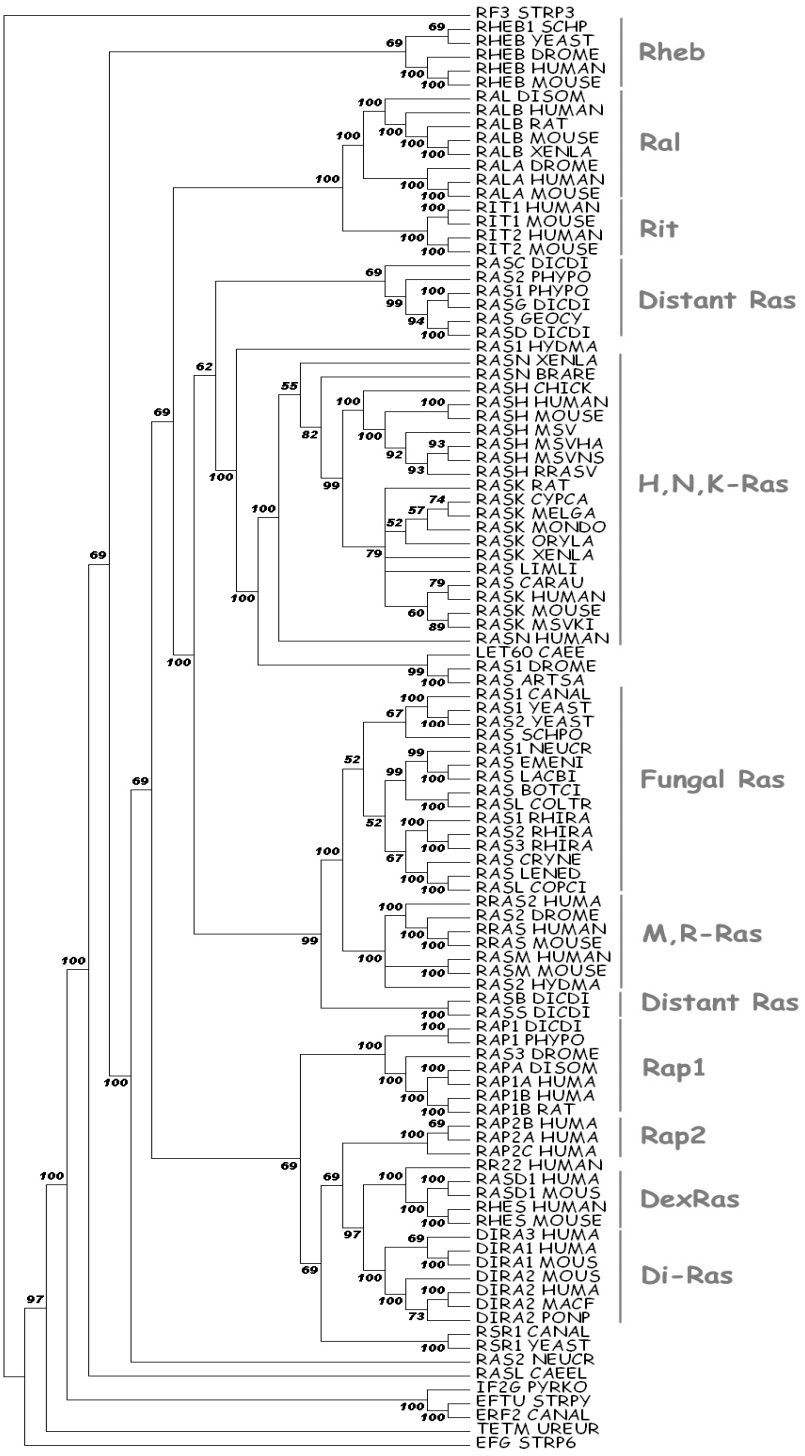


Figure A.4. Maximum Parsimony Consensus Tree.
A 50% Majority-rule consensus of 2555 trees.

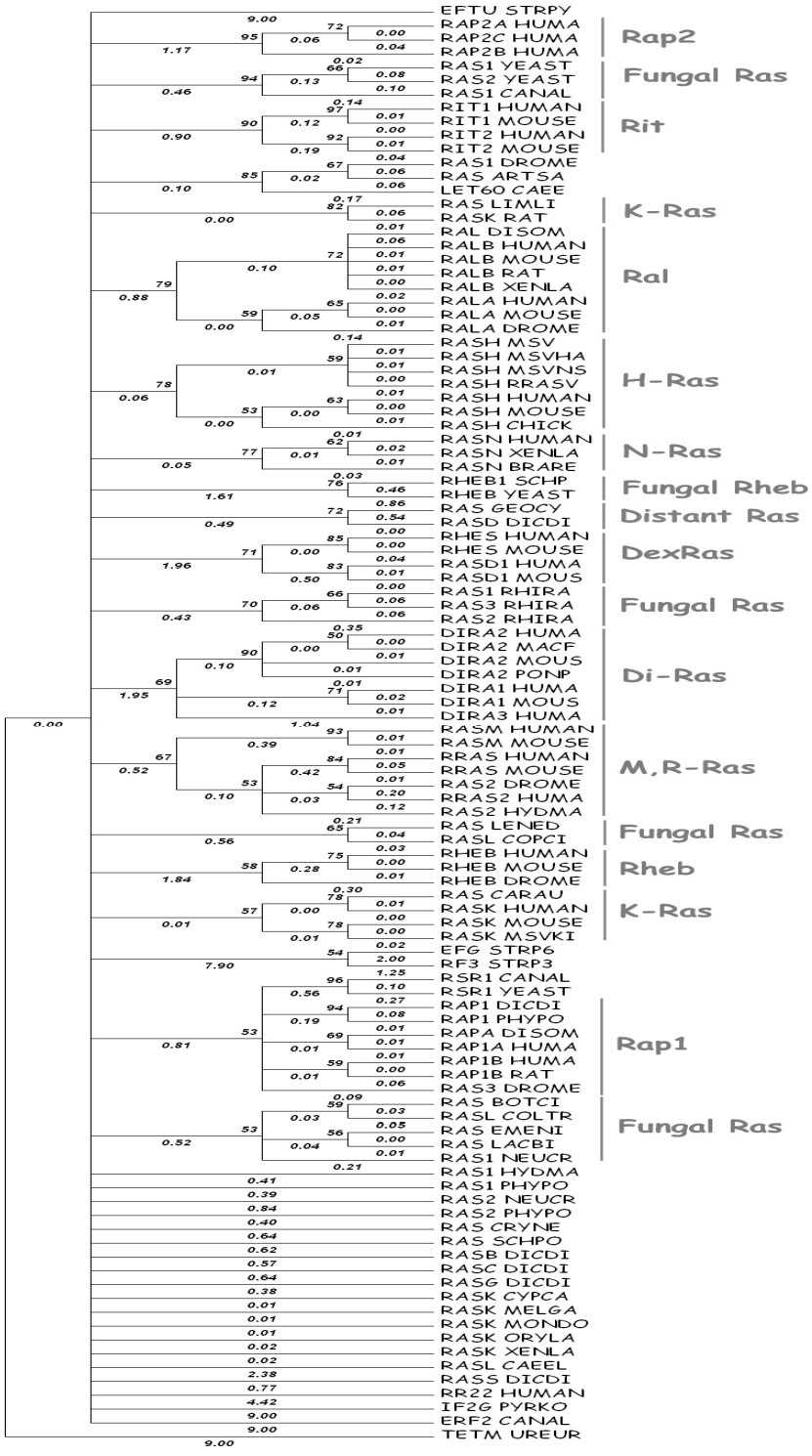


Figure A.5. Quartet-Puzzling Tree.
The tree was condensed at a threshold of 50%.

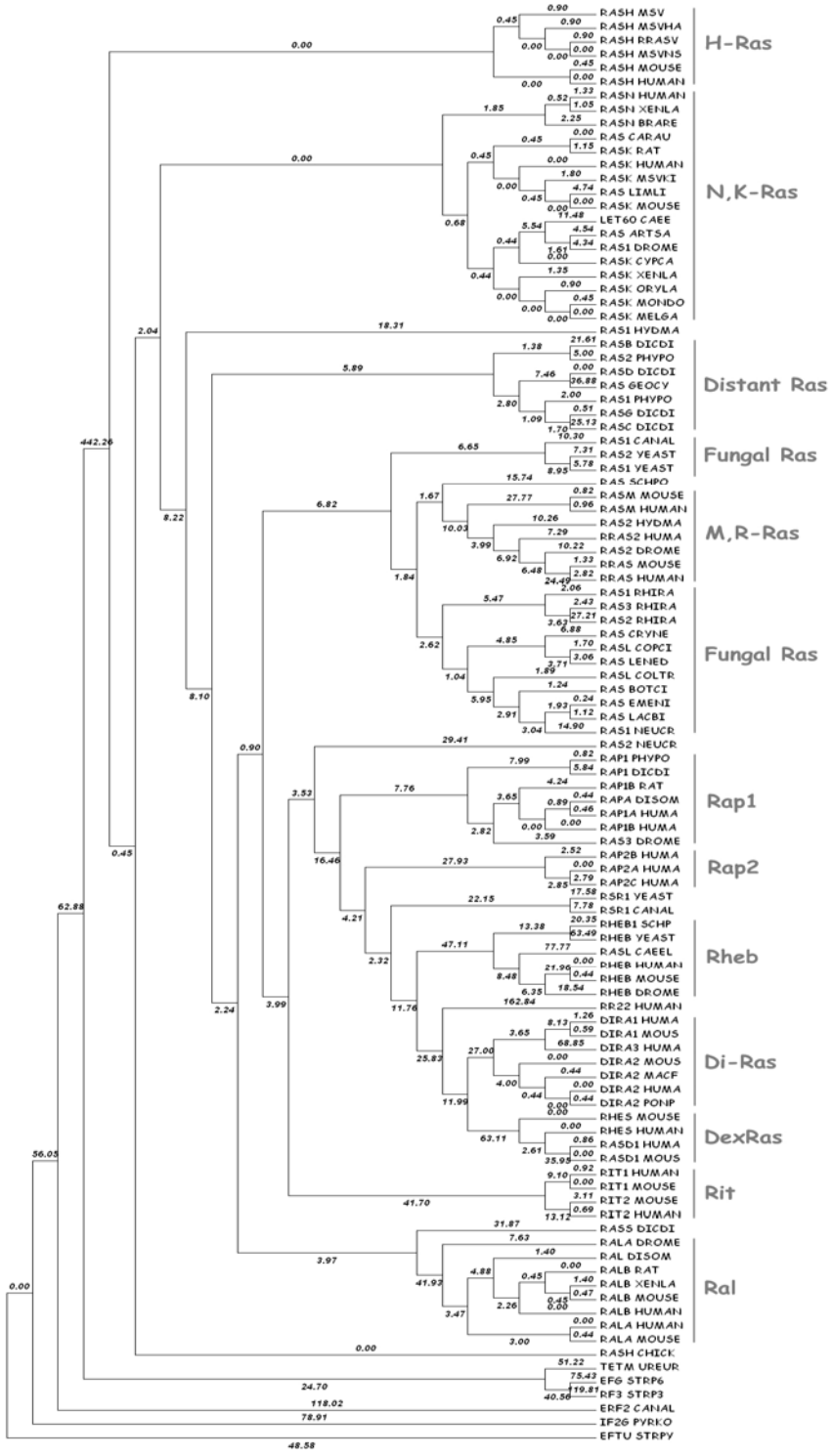


Figure A.6. Maximum-Likelihood Tree computed by Stepwise Addition

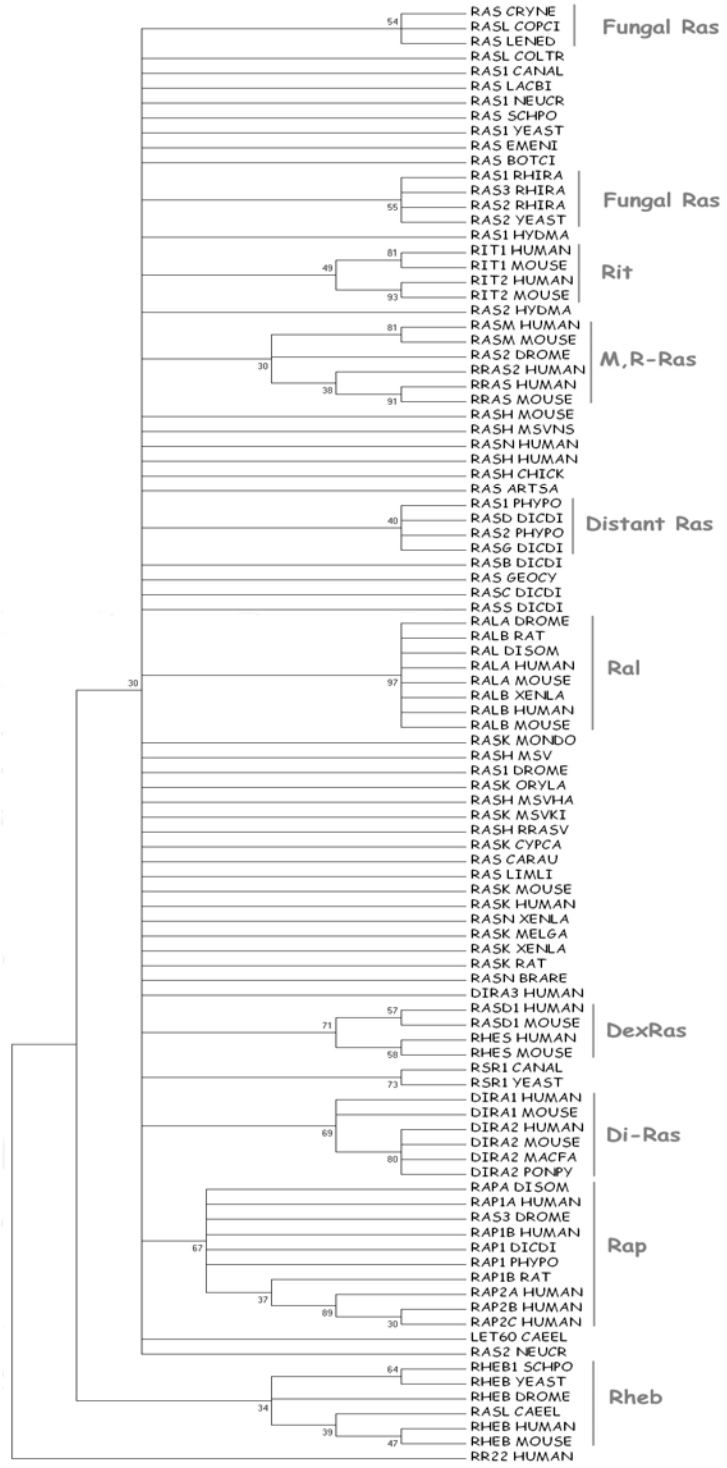


Figure A.7. Neighbor-Joining Tree of Ras Family Switch Sites

Table A.1. Reconstructions of the Ras Family Ancestral Sequence.
Bold residues indicate a difference between the parsimony reconstruction's best match to the ML reconstruction and the actual ML reconstruction. [*bxz*] indicates that residues *b*, *x* or *z* are allowed in the site.

Reconstruction Method	**Reconstructed Sequence**
Parsimony (MP)	KIA[IMV][LM]G[ASY]RSVGKS[ST]LT[AIV]QFVEN[DEHR]FV[DE]SY[DY]PTIENTFTK[FHILNQV]IERKGQE[CFY]HL[EKQ]IIDTAGQDEYSI[FL][NP]ITSSI[DG]IHGY[IV]LVYSITS[IKQR]KSFE[MV]VKI[IL][YR][DEG]K[IL]LD[HQT][MVY]GKK[KNQSWY][IV]PIVLVGNKIDLHM[EQ]RVVS[TA]EEGK[AK]LA[ER][SE]W[RNGK]AAF[LT]E[ITAC]SA[EKR]HNE [NST]V[DG]DVFELIILEIE
Parsimony (Best match to ML)	KIAVLGSRSVGKSSLTVQFV**EN**HFVESYDPTIENT**F**TKLIE**R**KGQ**E**Y**H**LEIIDTAGQDEY**SI**LP**ITS**SIDIHG**Y**ILVYSITSRKSFEMVKII**R**EKIL**DT**MGK**K**NVPIVLVGNK**I**DLH**M**ER**V**VSTEEGK**K**LA**R**EWK**A**AFLETSAK**H**NENVD**D**VF**E**L**IILE**IE
Maximum Likelihood (ML)	KIAVLGSRSVGKSSLTVRFV**QG**HFVESYDPTIENT**Y**TKLIE**V**KGQ**D**Y**T**LEIIDTAGQDEY**TV**LP**R****KY**SIDIHG**F**ILVYSITSRKSFEMVKII**H**EKIL**RV**MGK**D**NVPIVLVGNK**C**DLH**T**ER**A**VSTEEGK**E**LA**K**EWK**C**AFLETSAK**Q**NENVD**E**VF**H**L**LLRQ**IE

Appendix B: Development of Entropy Signatures of the Ras Families

Figure B.1. Scaled Boltzmann-Shannon Entropy of Arf Family Sites.
A histogram of Boltzmann-Shannon entropy values of the amino acid sites of the Arf family. The twenty amino acids were divided into their eight functional groups (AGLIVM, FWY, DE, NQ, HKR, ST, C, P) as previously described in Atchley et al. (1999). The entropy was scaled to [0,1] using $\log_8$. Sites of low entropy values are good candidates for building the entropy signature. Signature sites are denoted by an asterisk above the sites' columns. Sites were numbered using the Arf1 crystal structure (PDB ID: 1J2J; Shiba et al. 2003).

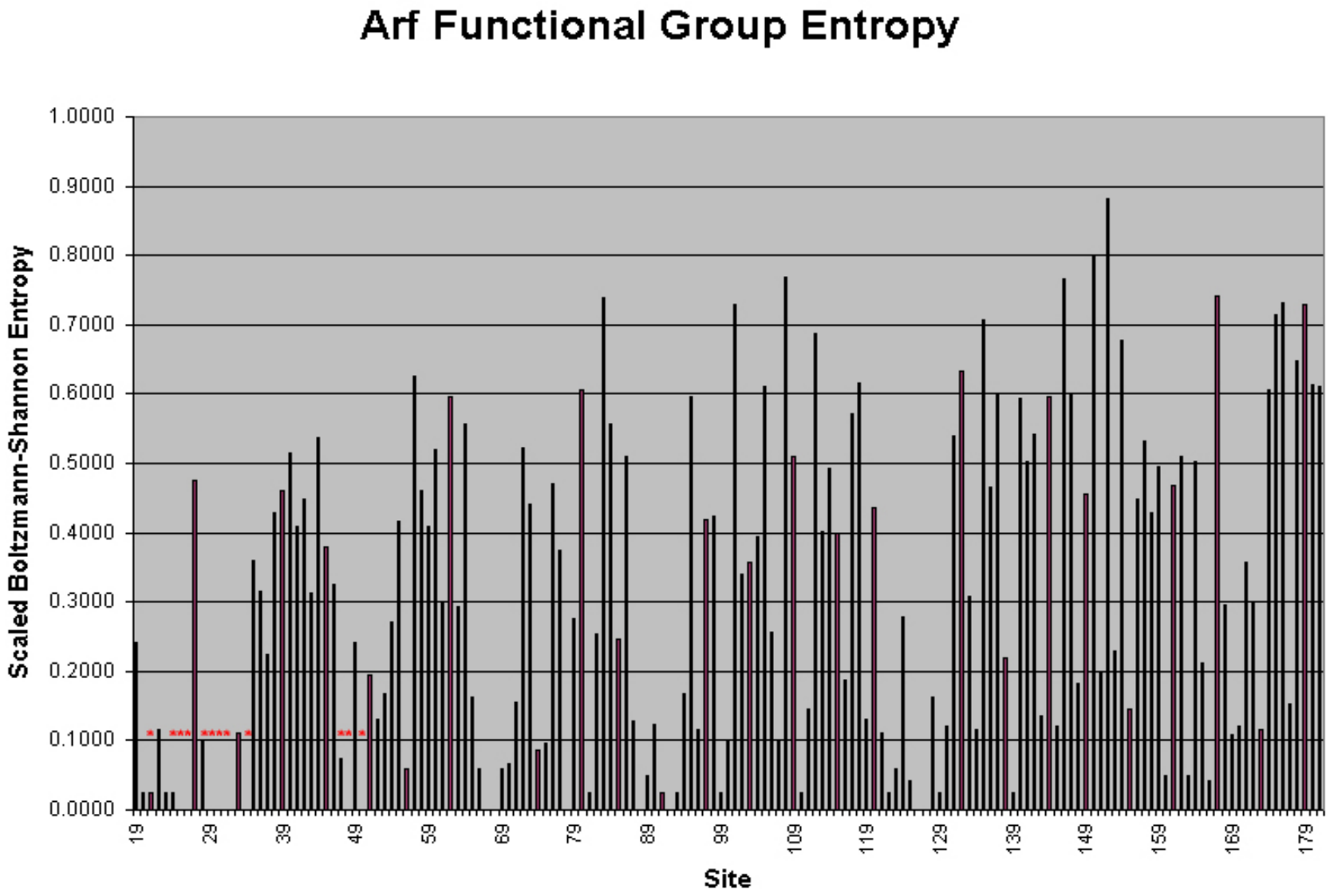

Arf Functional Group Entropy
Scaled Boltzmann-Shannon Entropy
1.0000
0.9000
0.8000
0.7000
0.6000
0.5000
0.4000
0.3000
0.2000
0.1000
0.0000
19
29
39
49
59
69
79
89
99
109
119
129
139
149
159
169
179
Site

Figure B.2. Scaled Boltzmann-Shannon Entropy of Sar Family Sites.
A histogram of Boltzmann-Shannon entropy values of the amino acid sites of the Sar family. The twenty amino acids were divided into their eight functional groups (AGLIVM, FWY, DE, NQ, HKR, ST, C, P) as previously described in Atchley et al. (1999). The entropy was scaled to [0,1] using $\log_8$. Sites of low entropy values are good candidates for building the entropy signature. Signature sites are denoted by an asterisk above the sites' columns. Sites were numbered using the Sar1 crystal structure (PDB ID: 1M2O; Bi et al. 2002).

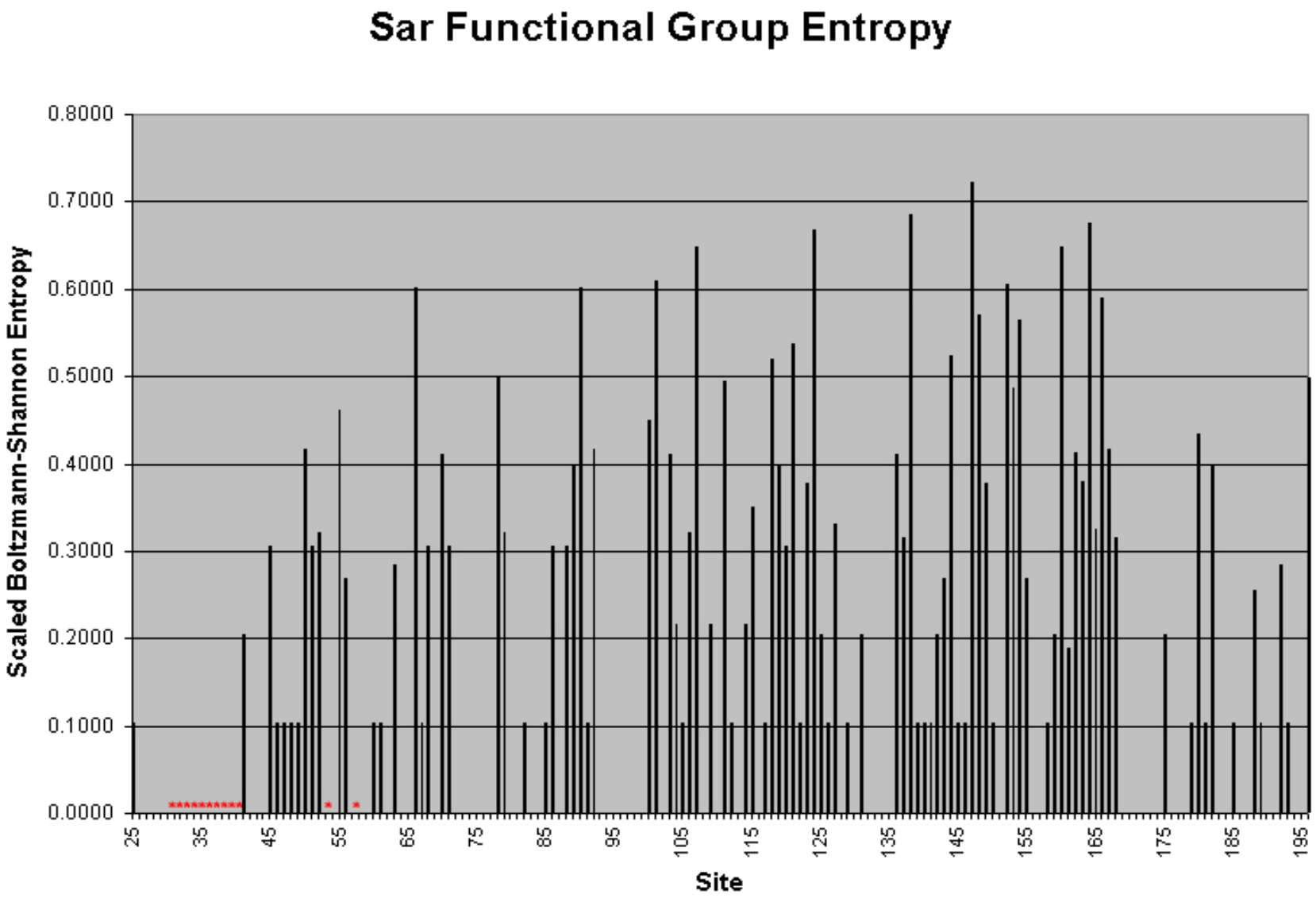

Sar Functional Group Entropy
Scaled Boltzmann-Shannon Entropy
0.8000
0.7000
0.6000
0.5000
0.4000
0.3000
0.2000
0.1000
0.0000
25
35
45
55
65
75
85
95
105
115
125
135
145
155
165
175
185
195
Site

Figure B.3. Scaled Boltzmann-Shannon Entropy of Rab Family Sites.
A histogram of Boltzmann-Shannon entropy values of the amino acid sites of the Rab family. The twenty amino acids were divided into their eight functional groups (AGLIVM, FWY, DE, NQ, HKR, ST, C, P) as previously described in Atchley et al. (1999). The entropy was scaled to [0,1] using $\log_8$. Sites of low entropy values are good candidates for building the entropy signature. Signature sites are denoted by an asterisk above the sites' columns. Sites were numbered using the YPT1 crystal structure (PDB ID: 1UKV; Rak et al. 2003).

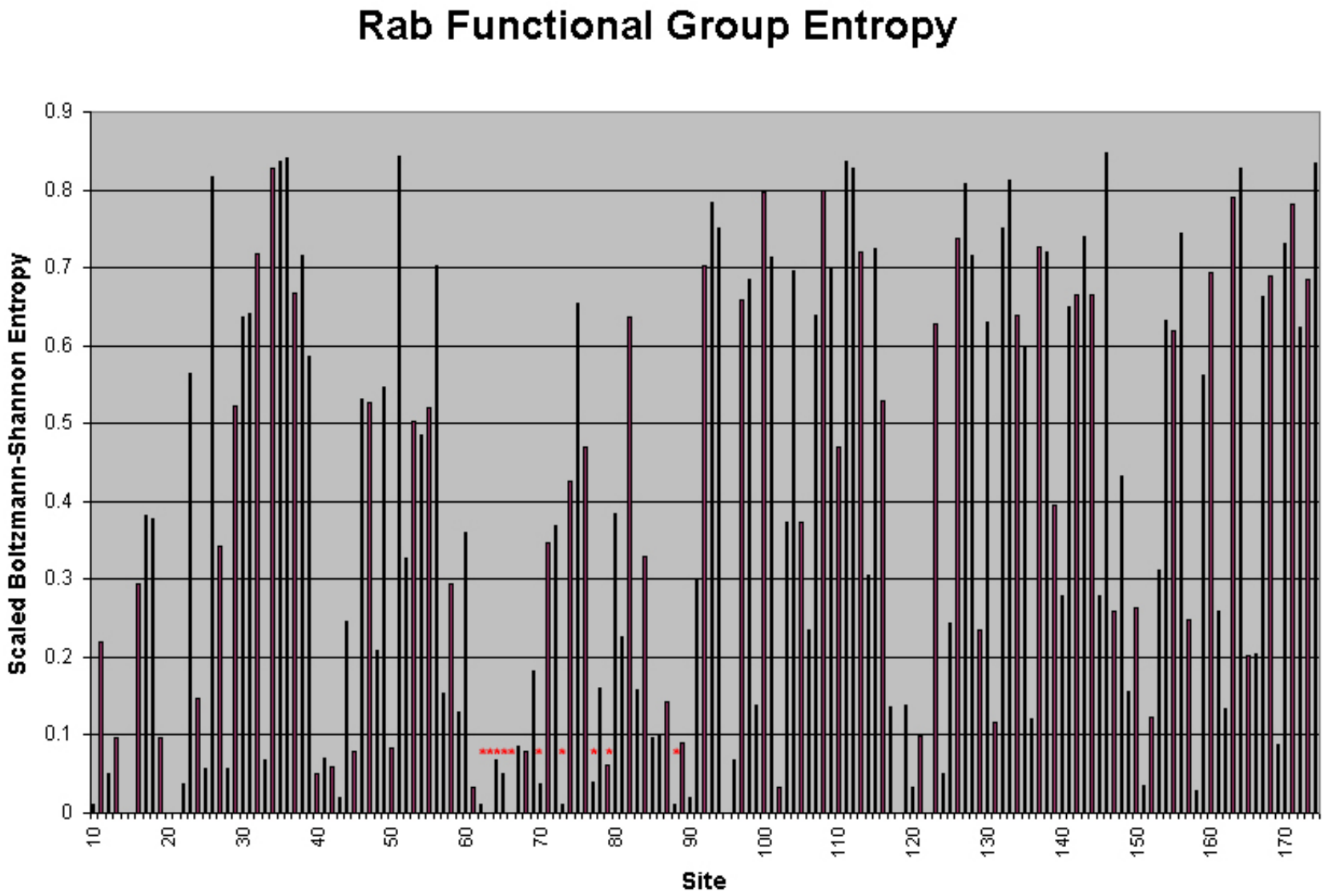

Rab Functional Group Entropy
Scaled Boltzmann-Shannon Entropy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
Site

Figure B.4. Scaled Boltzmann-Shannon Entropy of Ran Family Sites.
A histogram of Boltzmann-Shannon entropy values of the amino acid sites of the Ran family. The twenty amino acids were divided into their eight functional groups (AGLIVM, FWY, DE, NQ, HKR, ST, C, P) as previously described in Atchley et al. (1999). The entropy was scaled to [0,1] using $\log_8$. Sites of low entropy values are good candidates for building the entropy signature. Signature sites are denoted by an asterisk above the sites' columns. Sites were numbered using the Ran crystal structure (PDB ID: 1IBR; Vetter et al. 1999).

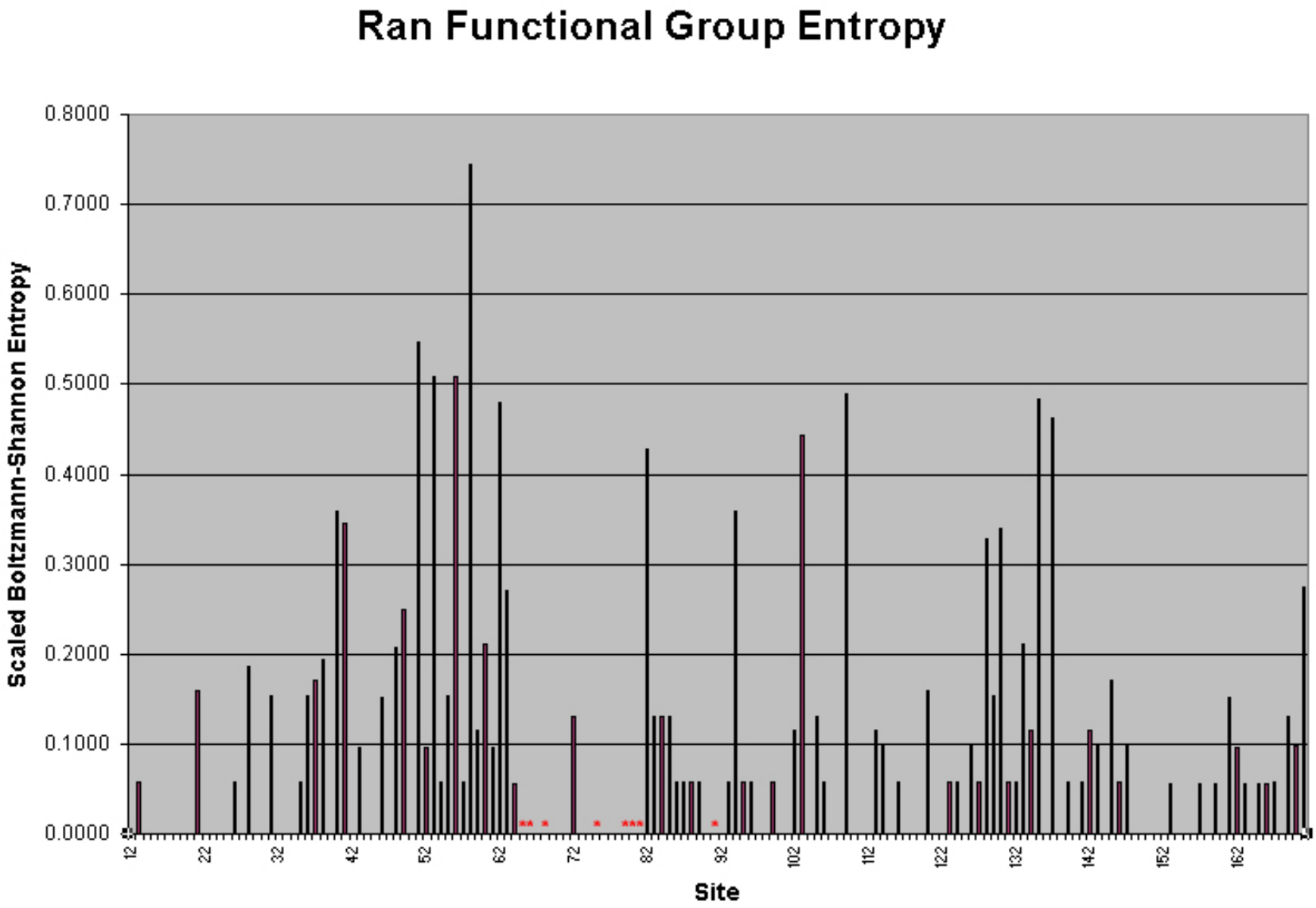

Ran Functional Group Entropy
Scaled Boltzmann-Shannon Entropy
0.8000
0.7000
0.6000
0.5000
0.4000
0.3000
0.2000
0.1000
0.0000
12
22
32
42
52
62
72
82
92
102
112
122
132
142
152
162
Site

Figure B.5. Scaled Boltzmann-Shannon Entropy of Ras Family Sites.
A histogram of Boltzmann-Shannon entropy values of the amino acid sites of the Ras family. The twenty amino acids were divided into their eight functional groups (AGLIVM, FWY, DE, NQ, HKR, ST, C, P) as previously described in Atchley et al. (1999). The entropy was scaled to [0,1] using $\log_8$. Sites of low entropy values are good candidates for building the entropy signature. Signature sites are denoted by an asterisk above the sites' columns. Sites were numbered using the H-Ras crystal structure (PDB ID: 121P; Wittinghofer et al. 1997).

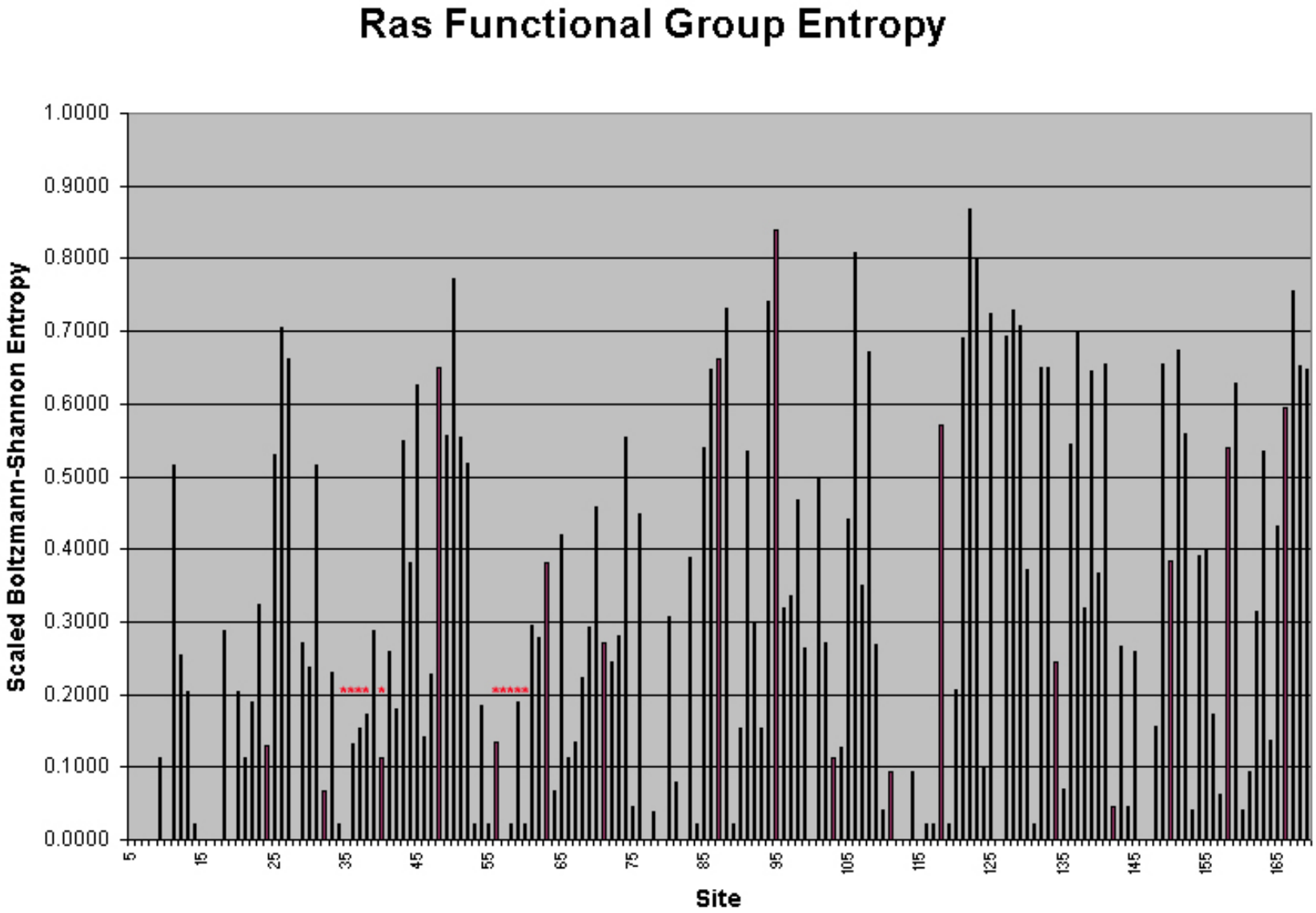

Ras Functional Group Entropy
Scaled Boltzmann-Shannon Entropy
1.0000
0.9000
0.8000
0.7000
0.6000
0.5000
0.4000
0.3000
0.2000
0.1000
0.0000
5
15
25
35
45
55
65
75
85
95
105
115
125
135
145
155
165
Site

Figure B.6. Scaled Boltzmann-Shannon Entropy of RGK Family Sites.
A histogram of Boltzmann-Shannon entropy values of the amino acid sites of the RGK family. The twenty amino acids were divided into their eight functional groups (AGLIVM, FWY, DE, NQ, HKR, ST, C, P) as previously described in Atchley et al. (1999). The entropy was scaled to [0,1] using $\log_8$. Sites of low entropy values are good candidates for building the entropy signature. Signature sites are denoted by an asterisk above the sites' columns. Sites are numbered by the Ras family standard (Valencia et al. 1991).

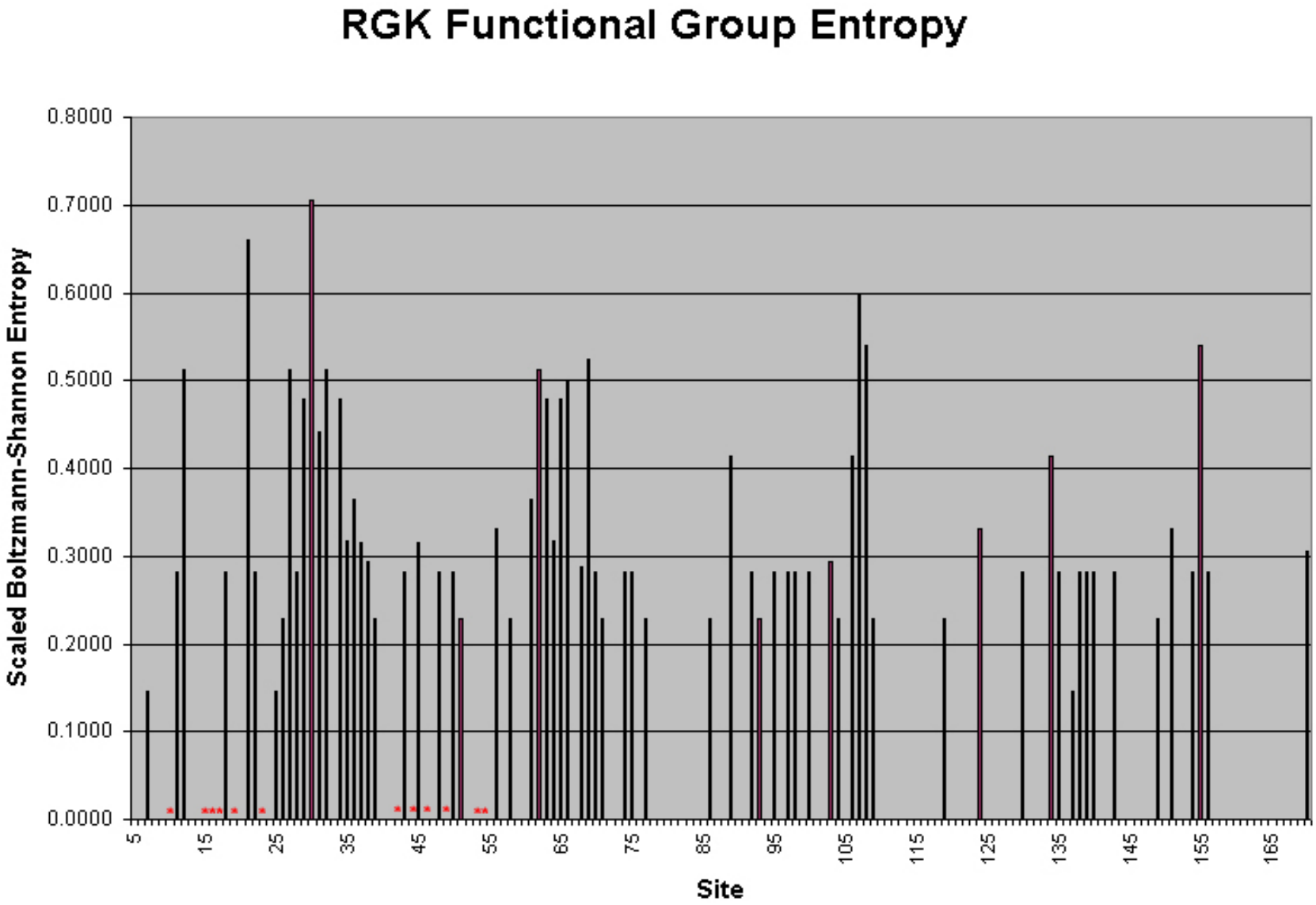
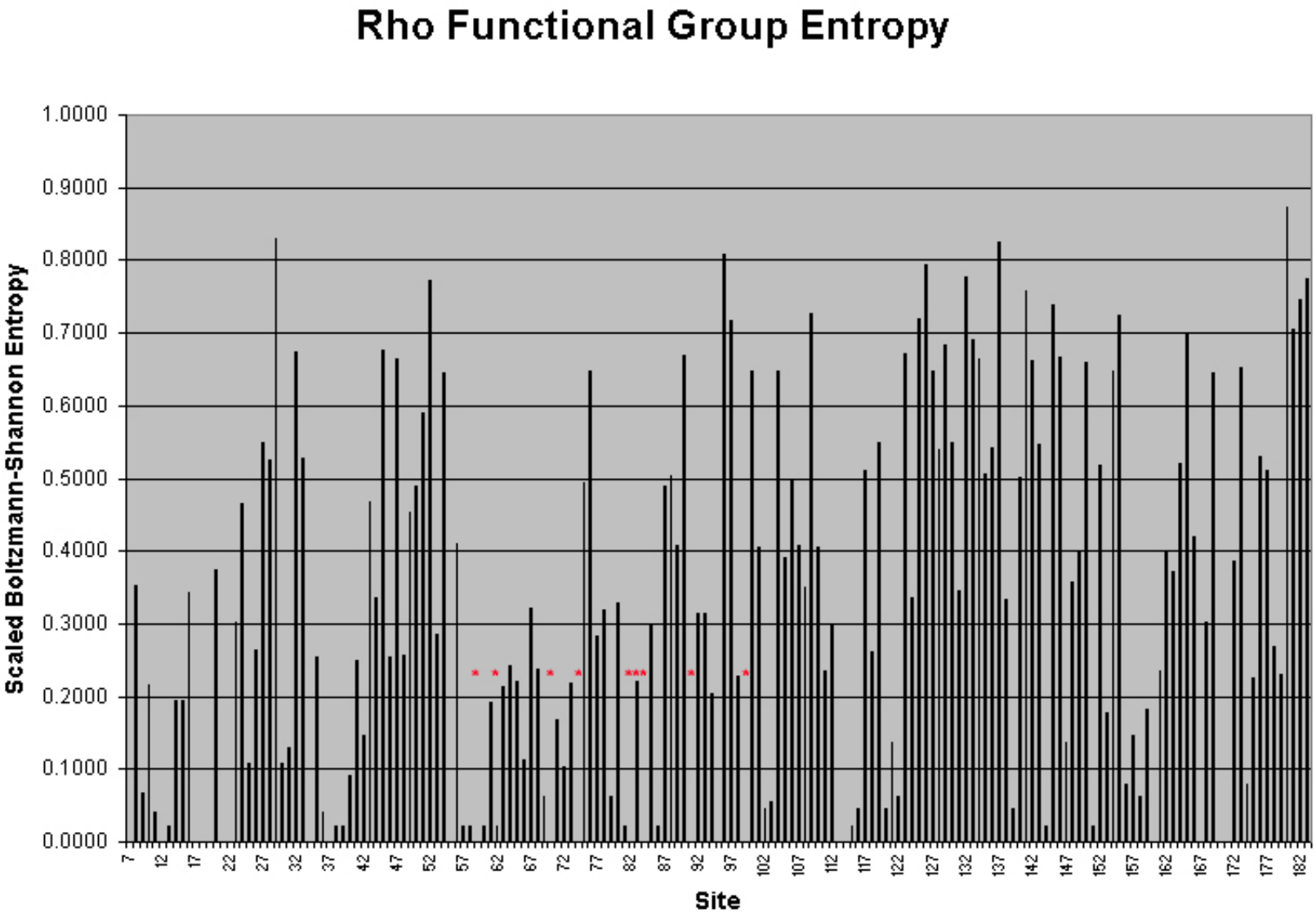

RGK Functional Group Entropy
Scaled Boltzmann-Shannon Entropy
0.8000
0.7000
0.6000
0.5000
0.4000
0.3000
0.2000
0.1000
0.0000
5
15
25
35
45
55
65
75
85
95
105
115
125
135
145
155
165
Site

Figure B.7. Scaled Boltzmann-Shannon Entropy of Rho Family Sites.
A histogram of Boltzmann-Shannon entropy values of the amino acid sites of the Rho family. The twenty amino acids were divided into their eight functional groups (AGLIVM, FWY, DE, NQ, HKR, ST, C, P) as previously described in Atchley et al. (1999). The entropy was scaled to [0,1] using $\log_8$. Sites of low entropy values are good candidates for building the entropy signature. Signature sites are denoted by an asterisk above the sites' columns. Sites are numbered by the RhoA crystal structure (PDB ID: 1CZX; Maesaki et al. 1999).


Rho Functional Group Entropy
Scaled Boltzmann-Shannon Entropy
1.0000
0.9000
0.8000
0.7000
0.6000
0.5000
0.4000
0.3000
0.2000
0.1000
0.0000
7
12
17
22
27
32
37
42
47
52
57
62
67
72
77
82
87
92
97
102
107
112
117
122
127
132
137
142
147
152
157
162
167
172
177
182
Site

RI**L**MV**GLD**A**AGKT**TI**L**YKLKLGEIVTTI**PTIG**FNVETVEYKNISFTVWDVGGQDKI

RPLWRHYFQNTQGLIFVVDSNDRERVNEAREELMRMLAEDELRDAVLLVFANKQD

LPNAMNAAEITDKLGLHSLRHRNWYIQATCATSGDGLYEGLDWLSNQLR

Figure B.8. Location of the Arf Family Entropy Signature
The functional core of the Swissprot protein Arf1 in *Homo sapiens* (Boeckmann et al. 2003) is shown. The sites covered by the signature, including the sites for which any amino acid is allowed, are in bold and boxed.

KLLLIGDSGVGKSCLLLRFADDTYTESYISTIGVDFKIRTIELDGKTIKLQI**WDTAGQ**

ER**F**RT**I**TSS**Y**Y**R**GAHGIIVV**Y**DVTDQESFNNVKQWLQEIDRYASENVNKLLVGNK

CDLTTKKVVDYTTAKEFADSLGIPFLETSAKNATNVEQSFMTMAAEIKKR

Figure B.9. Location of the Rab Family Entropy Signature
The functional core of the Swissprot protein Rab1a in *Homo sapiens* (Boeckmann et al. 2003) is shown. The sites covered by the signature, including the sites for which any amino acid is allowed, are in bold and boxed.

KLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFHTNRGPIKFNVW

DTA**G**QEKFGG**L**RDG**YYI**QAQCAIIMF**D** VTSRVTYKNVPNWHRDLVRVCENIPIVL

CGNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLEF

VAMPALAPPEVVMDPALAAQYEHDLEVAQ

Figure B.10. Location of the Ran Family Entropy Signature
The functional core of the Swissprot protein Ran in *Homo sapiens* (Boeckmann et al. 2003) is shown. The signature sites are in bold. All sites covered by the signature, including the sites for which any amino acid is allowed, are boxed.

KLVVVGAGGVGKSALTIQLIQNHFVDEYDP**TIEDSY**RKQVVIDGETCLLDI**LDTAG**

QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGN

KCDLAARTVESRQAQDLARSYGIPYIETSAKTRQGVEDAFYTLVREIR

Figure B.11. Location of the Ras Family Entropy Signature
The functional core of the Swissprot protein H-Ras in *Homo sapiens* (Boeckmann et al. 2003) is shown. The signature sites are in bold. All sites covered by the signature, including the sites for which any amino acid is allowed, are boxed.

RVVLL **G**DPGV**GKTSL**ASL**F**AGKQERDLHEQLGEDVYE**R**T**L**T**V**DG**E**DTT**LV** VVD

TWEAEKLDKSWSQESCLQGGSAYVIVYSIADRGSFESASELRIQLRRTHQADHVPIIL

VGNKADLARCREVSVEEGRACAVVFDCKFIETSATLQHNVAELFEGVVRQLR

Figure B.12. Location of the RGK Family Entropy Signature
The functional core of the Swissprot protein Rem1 in *Homo sapiens* (Boeckmann et al. 2003) is shown. The signature sites are in bold. All sites covered by the signature, including the sites for which any amino acid is allowed, are boxed.

KLVIVGDGACGKTCLLIVFSKDQFPEVYVPTVFENYVADIEVDGKQVELALW

DTA**G**QEDYDRL**R**PLS**Y**PDTDVIL**MCF**SIDSPD**S**LENIPEK**W** TPEVKHFCPNVPIILV

GNKKDLRNDEHTRRELAKMKQEPVKPEEGRDMANRIGAFGYMECSAKTKDGVRE

VFEMATRAAL

Figure B.13. Location of the Rho Family Entropy Signature
The functional core of the Swissprot protein RhoA in *Homo sapiens* (Boeckmann et al. 2003) is shown. The signature sites are in bold. All sites covered by the signature, including the sites for which any amino acid is allowed, are boxed.

KLVFL **GLDNAGKTTL**LHMLKDDRLGQHV**P**TLH**P**TSEELTIAGMTFTTFDLGGHE

QARRVWKNYLPAINGIVFLVDCADHSRLVESKVELNALMTDETISNVPILILGNKIDR

TDAISEEKLREIFGLYGQTTGKGNVTLKELNARPMEVFMCSVLKRQGYGEGFRWLS

Figure B.14. Location of the Sar Family Entropy Signature
The functional core of the Swissprot protein Sar1a in *Homo sapiens* (Boeckmann et al. 2003) is shown. The signature sites are in bold. All sites covered by the signature, including the sites for which any amino acid is allowed, are boxed.