# STUDIES ON CORRELATED MUTATIONS ALGORITHMS OF PROTEINS PROVIDING STRUCTURAL, SPATIAL, AND ALLOSTERY INFORMATION FROM MULTIPLE SEQUENCE ALIGNMENTS

by

Kyle Everett Kreth

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2013

Approved by:

______________________________
Dr. Anthony A. Fodor

______________________________
Dr. Donald J. Jacobs

______________________________
Dr. Dennis R. Livesay

______________________________
Dr. Cem Saydam

______________________________
Dr. Evan G. Houston

# ABSTRACT

KYLE EVERETT KRETH. Studies on correlated mutations algorithms of proteins providing structural, spatial, and allostery information from multiple sequence alignments. (Under the direction of DR. ANTHONY A. FODOR)

Proteins provide innumerable cellular functions and benefits for all kingdoms in the domains of life. Advancements in the high throughput collection and analysis of proteins have led to ever-deeper understanding of biological pathways, evolution, and coding biases. Most protein functional and/or structural analysis that is carried out in an *in vitro* manner is not amenable to high throughput technologies. With the incredible growth of sequences to study, we have capabilities to further refine algorithms that work *in silico*, using the work done *in vitro* as a benchmark. There has been a renaissance of the study of proteins using new approaches that are largely possible because of the amount of data now available for analysis. The research in this dissertation investigates some of the new techniques available in this field, to find the limitations of these techniques as well as improve upon them.

Chapter 1 presents both an overview of generalized techniques at the disposal of researchers looking for links between protein sequence covariance and allostery. The methods most commonly used including mutual information, chemical similarity matrixes, phylogenetic perturbation, and chi-square analysis are reviewed as well as the limits of such approaches to detecting allostery. Chapter 2 explores using a recent phylogenetic correction that has been successful for improving the efficacy of mutual information to predict special contact on the other algorithm types introduced in the first chapter. Chapter 3 is an attempt to detect bias of covariance algorithms on the rigid

bodies found in protein structures.  Chapter 4 is the description of a novel algorithm, termed COvariance By Sections (COBS), that in many ways is a combination of the methodologies used in Chapter 2 and Chapter 3, whereby we leverage a phylogenetic correction on groups of MSA columns rather than individual columns.

# ACKNOWLEDGEMENTS

Bank of America for supporting my efforts for a graduate career at UNC Charlotte, in particular several managers: Robert Noriega, Stephanie Mackie, Scott Reese, and Henry Tsang. In particular Robert Noriega supported my efforts early on, which led to flexibility in work hours being possible in the earliest years of course work that made meeting the graduate school requirements for full time status possible.

An invaluable collaborator for Chapters 1 and 4, Dr. Gregory Gloor from the Schulich School of Medicine and Dentistry at The University of Western Ontario has been gracious with his knowledge and time. Dr. Gloor's enthusiasm for the study and understanding of covariance of amino acids in proteins proved to be contagious when he visited UNCC.

I would also like to acknowledge Dr. Lawrence Mays, Bioinformatics and Computational Biology, University of North Carolina at Charlotte for his continuous support, starting before there even was such a department. My Doctoral committee members Dr. Dennis Livesay, Dr. Donald Jacobs, Dr. Cem Saydam, and Dr. Evan Houston for their considerable time in agreeing to be on my committee, with all the meetings, mentoring, and reviewing effort that comes with such an honorific. My advisor Dr. Anthony Fodor from whom I have learned an enormous amount of knowledge on how to think carefully about scientific problems, when to start fresh with new approaches, and helping me whiteboard, brainstorm, and act as a sounding board.

Without the support of my wife Ashley and our three daughters Kaitlyn, Mallory, and Rachel I never would have completed the program. They have seen this entire process from the vantage point of lost personal time. My family's endearing support has

brightened some of the darkest hours of frustrations and failures, and it is my sincere hope that these pages honor their sacrifices.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF EQUATIONS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| APC | average product correction |
| CASP | critical assessment of methods of protein structure prediction |
| COBS | covariance by sections |
| COBSp | covariance by sections with phylogenetic correction |
| ELSC | explicit likelihood of subset co-variation |
| FIRST | floppy inclusion and rigid substructure topography |
| GOC | groups of columns |
| MI | mutual information |
| MIp | mutual information with the phylogenetic correction |
| McBASC | McLachlan based substation correlation |
| McBASCp | McLachlan based substation correlation with phylogenetic correction |
| MSA | multiple sequence alignment |
| PFAM | protein family database |
| ROC | receiver operating characteristic |
| SCA | statistical coupling analysis |
| SCAp | statistical coupling analysis with phylogenetic correction |
| SS | secondary structure |

# CHAPTER 1: CAN COVARIANCE PREDICT ALLOSTERIC MECHANISMS?[1]

## 1.1 Abstract

The notion of using the evolutionary history encoded within multiple sequence alignments to predict allosteric mechanisms is appealing. In this approach, correlated mutations are expected to reflect coordinated changes that maintain intramolecular coupling between residue pairs. Despite much early fanfare, the general suitability of correlated mutations to predict allosteric couplings has not yet been established. Lack of progress along these lines has been hindered by several algorithmic limitations including phylogenetic artifacts within alignments masking true covariance and the computational intractability of consideration of more than two correlated residues at a time. Recent progress in algorithm development, however, has been substantial with a new generation of correlated mutation algorithms that have made fundamental progress towards solving these difficult problems. Despite these encouraging results, there remains little evidence to suggest that the evolutionary constraints acting on allosteric couplings are sufficient to be recovered from multiple sequence alignments. In this thesis, we argue that due to the exquisite sensitivity of protein dynamics, and thus allosteric mechanisms, allosteric mechanisms vary widely within protein families. If it turns out to be generally true that even very similar homologs display a wide divergence of allosteric mechanisms, then even a perfect correlated mutation algorithm could not be reliably used as a general mechanism for discovery of allosteric pathways.

## 1.2 Introduction

Starting with the proposal by Horovitz *et al.* in 1994 [2], there has been a deep interest in predicting allosteric couplings within proteins based on coevolutionary processes. The intuitive approach is based on multiple sequence alignment column pairs displaying correlated mutations, which have been interpreted to reflect coordinated changes that maintain pairwise intra-molecular couplings. The premise is that when a mutation occurs within a protein, a compensating mutation can occur elsewhere, and conserved patterns of these pairs across a multiple sequence alignment are interpreted as a signal of co-evolutionary processes. The seminal paper by Lockless and Ranganathan [3] argued that such an approach predicts thermodynamic coupling in proteins. Specifically, they demonstrated that extent of binding energy nonadditivity from double mutant cycles within PDZ domains is linearly related to the strength of the correlated mutation signal. Since then, the explosion of publically available protein sequences has ensured that co-evolutionary analysis has continued to develop as a staple within the field of protein bioinformatics. However, despite more than ten years of subsequent vigorous research, there remains no statistically significant demonstration of the ability for correlated mutation algorithms to predict intra-molecular couplings over long distances [4-6].

In their study, Lockless and Ranganathan utilized a novel algorithm (named SCA) for detecting correlated mutations. In the decade since publication of their paper, extraordinary progress has been made in the study of correlated mutation algorithms, which is highlighted by dramatic improvements in the ability of correlated mutations to predict structure contacts. In the first part of this thesis, we summarize this progress and the remaining algorithmic problems that need to be addressed. In the second part of our review, we present our arguments for why we believe that despite this progress the enterprise of predicting allosteric couplings from correlated mutations may be based on flawed assumptions of allostery. We suggest that there is no underlying reason to believe that coevolution of allosteric mechanisms actually occurs routinely. While allosteric mechanisms can be conserved within close taxonomic groups [7], anecdotal reports indicating that allostery is not a strong evolutionary driving force are increasingly commonplace [8-10]. Surprisingly, diversity of response is even evident in hemoglobin [11, 12], an archetype of long-range intramolecular communication. Clearly, conservation of allostery and intramolecular couplings is nowhere near that of structure and/or function [13]. Moreover, allosteric pathways are both frequent [14] and mechanistically plastic [15, 16]. As a consequence, even if the algorithms were perfect, we remain to be convinced of the underlying notion that allosteric couplings can be recovered from correlated mutations.

### 1.3 Improved Correlated Mutation Algorithms

The limiting factor of correlated mutation analysis is a low signal-to-noise ratio [17]. In a 2004 comparison of four different covariance algorithms, Fodor and Aldrich found that Mutual Information (MI) was the worst performing algorithm in predicting residue contacts [18]. Based on their most widespread application, the predictive power of correlated mutation algorithms was assessed by the ability to predict structural contacts. The poor performance of MI was ascribed in part to its tendency to give high scores to random (or poorly conserved) columns. Nevertheless, MI remains an appealing approach because it is stated in the simple, formal language of the application of information theory to entropy and is therefore easy to understand and calculate. As such, a good deal of recent research has found ways to dramatically improve the performance of MI. As a result, the past five or so years has produced a number of updated MI methods that are now the most powerful correlated mutation algorithms. However, it is interesting to note that some of the algorithmic improvements applied to MI are also applicable to other approaches[17]. Below we summarize this recent exciting progress.

### 1.4 Reducing Phylogenetic Noise Improves Predictive Power

A critical factor limiting the predictive power of covariance methods originates from tangling of phylogenetic and correlated mutation relationships [19]. Consider two subfamilies in which the sequences within each are closely related, but the inter-subfamily relationships are more distant. An alignment that consists of sequences from both subfamilies will have a great deal of apparent covariation due to the many changes in each column that are correlated with changes in other columns. However, these changes reflect a phylogenetic artifact of the way the alignment was constructed and do not reflect the underlying structure or functional constraints on the protein. This problem has been widely recognized but is difficult to correct because the true phylogenetic history of a family is unknown. Several algorithms have attempted to reduce bias by using rigorous phylogenetic approaches that take into account evolutionary distances within the family [20-22]. These methods have been demonstrated to improve structural contact identification, but they are computationally intensive and are therefore not appropriate for the many cases where alignments are made up of thousands much less tens of thousands of protein sequences. Simple *ad-hoc* methods, such as removing overly similar sequences from the alignment, can be easily employed, but this is a blunt approach with arbitrary parameters. Similarly, owing to disparate evolutionary forces, others have attempted to remove paralog sequences; however, the presence of paralogs has been shown to actually improve correlated mutation identification in some cases [23]. Alternatively, others have developed methods based on physiochemical properties of the identified residue pairs [24] or a complicated number of algorithmic filters [25]. In both cases, improvements in contact prediction accuracy have been reported. However, the *ad hoc* nature of these methods suggests that they are particularly tuned to prediction of structural contacts, and it is unclear the degree to which they actually filter phylogenetic biases.

In 2008 [26], a simple and computationally efficient method to suppress phylogentic bias that dramatically improves contact prediction performance was introduced. Rather than attempt a phylogenetic reconstruction, this method normalized the observed covariance of a pair of columns by the background covariance of the columns, where background covariance is measured as the average covariance score of a column with all other columns. With this correction in place, MI went from the worst performing algorithm [18] to the best, easily outperforming previously described methods [26]. For the rest of this chapter, we will refer to MI with this correction as $MI_p$. Application of the $MI_p$ correction to other correlated mutation algorithms also improves predictive power [17], highlighting that phylogenetic bias is a general problem and is not limited to just MI.

### 1.5 No Generally Accepted Method To Produce “Correct” Alignment Inputs

Correlated mutation algorithms obviously require a multiple sequence alignment as input. As such, collation and alignment of the sequence dataset are critical first steps in this process. Unfortunately, there is little consensus on dataset and alignment protocols. Since evolutionary correspondence is questionable, the most sensitive regions are those at the alignment ends and in gapped positions. In fact, we suspect that covariance is just due to noise when an arbitrarily gap threshold is invoked (e.g., it is common to include columns with less than 50% gaps). Recently, it was demonstrated that the introduction of even modest alignment errors could produce a substantial number of false positives when using $MI_p$ [27]. The poor performance of MI is caused in part by sensitivity to the background conservation of each column in the alignment [18], and the $MI_p$ correction only partially removes this sensitivity [28]. To correct for this, Little and Chen [28] introduced a further correction to $MI_p$ that regressed background and observed MI scores against one another and used residuals from the regression model normalized as *Z*-scores as the covariance score. Using a separate mathematical formulation of the algorithm that produces essentially identical results, Dickson *et al.* [27] demonstrated that this further improvement to $MI_p$ reduces sensitivity to alignment artifacts. Even for this algorithm, however, alignment errors can still substantially mask true covariance hampering the sensitivity of the correlated mutation approach [27]. A fully automated multiple sequence alignment procedure that produces alignments without errors is still beyond the reach of current bioinformatics [29], and as such this issue remains a potential problem that can hinder successful application of covariance techniques. Interestingly, covariance algorithms may, themselves, be an important tool for determining when an alignment has an error [27]. Clearly, there is much work left to be done in this area.

Another largely unsolved problem is how many sequences are required for successful application of covariance techniques. Gloor *et al.* have demonstrated that at least 125 sequences must be considered before the random MI signal is surpassed by the true correlated mutation signal [30]. Others go further. For example, it has been suggested by Hamacher *et al.* that 200-300 sequences are required [31], whereas Nielsen *et al.* suggest that at least 400 sequence 'clusters' are needed [32] where a cluster indicates groups of sequences that are related to one another by some similarity threshold, typically ~60% identity. Interestingly, application of the joint alignment background correction used by $MI_p$ to McBASC improves predictive power for alignments up to ~100 sequences [17], although $MI_p$ outperforms the corrected McBASC in larger sequence alignments. Guidelines for the minimal number of required sequences remain empirically derived and development of a more rigorous theory to guide algorithm choice for a given alignment may be helpful. Recent work looking at the impact of how different assumptions used in calculating a probability of a residue in a column from the number of times that residue is observed in a column may be a first step in this direction [33]. Also showing great promise is a bootstrap approach that randomly divides an alignment into subsets many times over, and asks how often the same set of covarying pairs is observed in each permutated subalignment [34]. Intriguingly, this approach found that some covariance algorithms were more accurate while others were more reproducible, although neither the accuracy nor reproducibility of any of the algorithms was perfect. Going forward, explicit consideration of the trade-offs in power vs. sensitivity to alignment artifacts should help guide future algorithm design.

### 1.6 Moving Beyond Pairwise Covariance

Even with the above improvements made to covariance algorithms, nearly all protein families contain a large number of covarying pairs that are not in close physical contact [35]. These “covarying but distal” pairs may reflect algorithmic errors or phylogenetic biases in the alignment that still remain uncorrected for by the covariance algorithms. Alternatively, a more intriguing hypothesis is that the covarying resiudes are functionally linked via allostery. A recent paper that rigorously demonstrates that most distal coevolving pairs are simply explained by coevolving contact chains provides considerable support for this view. Using a Bayesian network model, Burger and van Nimwegen [35] demonstrate that most covarying distal pairs are in fact connected by chains of residues that are also covarying. Because of the Bayesian formalism, this approach is generally computationally efficient, especially as compared to methods with similar intentions such as Weigt *et al* [36]. Instead of using the $MI_p$ score to rank residues, Burger and van Nimwegen propose using a posterior probability reflecting the strength of the $MI_p$ score between the two residues relative to the covariance scores of all possible residues that link the two residues in a chain. Remarkably, this procedure dramatically improves the performance of $MI_p$, reflecting the second major improvement to the performance of correlated mutation algorithms that has been described in the last three years [26, 28, 35]. Considering the fact that these algorithms have been actively studied since the early 1990s [37, 38], this progress is both unexpected and exciting.

### 1.7 A Critical View of the Underlying Concept of Conserved Allosteric Pathways

The stability of a protein, $\Delta G$, compares the free energy of the folded versus unfolded state, and the stability of a protein double mutant is described in Eq. 1.1. The $\mathbf{\Delta}_{ij}$ term quantifies the amount of nonadditivity within the cycle relative to the sum of the constituent single mutants, which identifies thermodynamic coupling between the pair of single mutants. It has been appreciated for over 25 years that nonadditivity within double mutant cycles is trivially expected within structural neighbors [39]. Conversely, when nonadditivity occurs within distal pairs, thermodynamic coupling is a convenient and commonly used reporter of allosteric coupling. However, while correlated mutations can be used to identify functionally important residues [40], there is little evidence to suggest that thermodynamically coupled pairs are limited to correlated mutations [4]. Rather, no correlation is observed between $\mathbf{\Delta}_{ij}$ and correlated mutation scores in three example protein families with good double mutant cycle coverage[4]. In fact, long-range thermodynamic coupling is, in itself, quite rare across the three datasets, which is consistent with the much larger double mutant dataset considered by Istomin *et al.* [41].

$$\Delta G_{ij} = \Delta G_i + \Delta G_j + \delta_{ij}$$

EQUATION 1.1: The amount of energy change for a combination of position *i* and position *j*, described here as the canonical delta G, would normally be expected to be a summation of individual energy changes for separate changes *i* and *j*. For a number of reasons however, what is actually measured is a phenomenon known as non-additivity, whereby a third term, here described by the lower case Greek letter delta is required to balance out the equation. The amplitude of the non-additivity is general a measure of the interaction between the two positions *i* and *j*, whether direct or indirect (allostery).

Nevertheless, the findings of Burger and van Nimwegen that consideration of covarying chains improves predictive power over considerations of residue pairs alone seems, at first glance, consistent with ideas put forward by the Ranganathan lab [3, 42, 43]. Perhaps these chains of correlated mutations that we can now rigorously identify with Bayesian statistics reflect the long-range allosteric couplings proposed by Lockless and Ranganathan [3]. Suel *et al.* [42] assert a “sparse network” of allosteric interactions, which may be related to the chains of covarying residues identified by Burger and van

Nimwegen [35]. We hope that investigators will use the new and improved tools for detecting covarying chains to test this relationship. While the veracity of this link will be ultimately decided by how well predictions derived from these new algorithms match mutagenesis experiments, our suspicion is that the information that can be gleaned from multiple sequence alignments is unlikely to reflect the plasticity of allosteric mechanisms [44], even with the improved covariance detection algorithms.

Our skepticism arises from our sense that the physical basis of the long-range intramolecular couplings that underlie allosteric response remains ambiguous [45]. Consistent with the idea of coevolving chains, *molecular wires* describe allostery as a cascade of local induced fit events that sequentially propagate over long distances [46-48], like a series of dominos falling in a line. Conversely, concerted *population shift* models describe pairwise couplings based on global changes in the free energy landscape [49, 50], which is akin to the conformer selection model of ligand binding [51]. Regardless of which model is "correct," both stress the importance of protein dynamics in allostery [52, 53]. That is, upon perturbation of an allosteric site, a signal is propagated to the effected site via a complex and dynamic change in structure. It is exactly this point that calls into question the notion that allosteric pathways are precisely conserved across a family. The literature includes nearly countless examples demonstrating that protein structures and their dynamics are highly sensitive to small perturbations, regardless of whether the perturbation is mutation [54], ligand binding [55], or simply changing the type of metal ion bound to the protein [56]. Related, several reports stress the diversity within dynamic signatures across protein families [57, 58]. If point mutations can perturb allostery in multiple ways, than it stands to reason that even closely related orthologs that have as much as 90 to 95% sequence identity may also have drastically different allosteric pathways. If this is true, it seems unlikely to us that the information required to predict these allosteric pathways could be contained in a multiple sequence alignment, at least using the methods in use today to create them..

An appreciable number of sites are frequently identified as being critical to intra-molecular communication within a given structure [59]. The large amount of variation within intra-molecular couplings makes it difficult to uncover general “traffic rules” regarding allosteric mechanisms across a protein family. For example, as discussed by del Sol *et al.* [14], the plasticity within allosteric response is such that nearby residues can easily functionally substitute for one another. Based on the ubiquitous diversity of allosteric mechanisms, Kuriyan and Eisenberg [44] have intriguingly suggested that allosteric diversity is responsible for the complexity of life. They argue for a “*rule of varied allosteric control*” where sensitivity in allosteric response is a fundamental evolutionary mechanism used to discover new pathways and functions. We are not arguing that such allosteric control, at least as it relates to non-evolving ligands, does not ultimately arise from sequence or that it is not subject to selection. Rather, we are asserting that allosteric control is so sensitive to context that it is unlikely to be recovered from sequence alignment information. The underlying assumption of a sequence alignment is that the sequences within that alignment share something in common evolutionarily. If sequences within an alignment are unlikely to share a common allosteric pathway, then that pathway cannot be reconstructed from the alignment.

As an instructive example, consider the case of cyclic-nucleotide gated ion channels. It has been demonstrated that the opening of these channels can be modeled as two independent steps: a ligand binding step, followed by a fully-liganded allosteric transition from the closed to the open state [60]. Consistent with our arguments of the general complexity of allostery in proteins, mutations throughout the channel sequence can alter the free energy of the allosteric transition without affecting ligand binding kinetics [60]. In the PFAM database, there are 11,189 sequences in the cNMP_binding family (PF00027) that contain a cyclic nucleotide binding domain. We can be reasonably confident that nearly all of these protein domains share the same fold and that most will bind cyclic nucleotides. One might expect a functional covariance analysis to find residues that are involved in, for example, discrimination of cAMP vs. cGMP binding, whereas the allosteric mechanisms found in these proteins will vary widely and depend on many factors not described by the alignment such as interactions with other protein domains and the microenvironment in which the protein is expressed. We assert that there is little reason to suppose a common allosteric mechanism shared by all, or even a significant fraction, of these eleven-thousand sequences.

On the other hand, allostery is likely conserved across very short evolutionary timescales. For example, one of us has recently demonstrated using computational modeling that, while residue-specific differences within CheY allostery are large and frequent, there is a general tendency for residues that initiate allostery to be structurally clustered [61]. Furthermore, we also observe stronger correlation within allostery across two closely related *E. coli and S. typhimurium* CheY orthologs, whereas there is greater diversity when compared to the more divergent *T. maritima* ortholog. In addition, quantitative differences in allosteric regulation have been used as a molecular basis of taxonomic assignments within the 3-deoxy-D-arabino-heptulosonate-7-phosphate synthetase family [7, 62, 63]. That is, while the extent of allosteric response is overall variable across the family, response is conserved within taxonomic groups, highlighting that conservation of allostery can occur over short evolutionary distances.

As such, a challenge to our assertion that allosteric information cannot be gleaned from sequence alignments will therefore require a departure from the current practice of building alignments based on every sequence that can be found with sufficient conservation to the experiential protein of interest. If there is any hope of detecting allostery in proteins, we assert that it will require restricting alignments to a subset of proteins for which there is reason to believe there is a common allosteric mechanism. However, as we have discussed above covariance algorithms tend to perform better on alignments containing large numbers of sequences. Unfortunately, there is no rigorous theory to guide the minimum number of sequences required for covariance analysis. It will be interesting to see in the next few years how the tension between a requirement for limiting alignment to proteins with well described functions balances against a requirement for sequence depth in covariance analysis.

## 1.8 Conclusions

In conclusion, the last few years have seen rapid development in the sophistication and power of algorithms for discovery of correlated mutations. Application of this new generation of algorithms should improve the utility of covariance methods in structure determination and the discovery of functionally and structurally constrained residues. The future of these techniques for discovery of prediction of allosteric pathways remains less clear, although we expect continued experimental effort focused on this question.

# CHAPTER 2: PHYLOGENETIC CORRECTIONS AND COVARIANCE.

## 2.1 Abstract

An established information theory algorithm known as mutual information (MI) has been used to study covariance with proteins, but with very limited success. A recent correction termed the Average Product Correction (APC) has increased both the specificity and sensitivity for MI in finding covarying protein portions. This increase has taken MI, which was one of the lowest performing algorithms in the study of protein covariance, to one the highest performing. In this chapter, we ask whether we can generalize the correction used successfully with MI to other classes of algorithms.

## 2.2 Introduction

Quantifying the amount of actual versus artifactual co-evolution in a multiple sequence alignment (MSA) is an area of active research [64]. First generation algorithms of protein covariance attempted to handle issues of phylogeny with what might be termed coarse-grained approaches such as removing sequences from an alignment based on sequence similarity. Confounding the difficulty of analysis of covariance are some key assumptions for algorithms such as MI that all of the sequences that make up an MSA are arrived at independently [65]. This key assumption ignores some of the largest influences of evolution such as gene duplication or accurately measuring the distance of phylogeny [66] to best prevent over representation in the MSA, which may bias results.

MI by itself has been shown to be a poor performer when compared to most other algorithm’s ability to recognize non-local covariance [18]. This is true for MI, even though this is one of the most widely used algorithms for correlation in engineering, statistics, and physics, has a well-understood formulation that several [26, 64] have exploited for sequence analysis.

A key finding in the study of MI and covariance was that the impact of phylogenetic proximity on the “tree” of evolution was a confounding factor responsible for false positive signals [19]. Empirical studies have indicated that if you included a bias or weighting scheme to differentiate what signal of MI was from evolutionary distance as opposed to structural or functional factors, MI scores would be more predictive of spatial proximity. The performance of MI in predicting spatial contact will tend to increase if corrections for phylogenetic trees are taken into account. Unfortunately, determining phylogenetic trees can be computationally difficult or inconsistent depending on the methods and seed data used. Worse yet, researchers have hypothesized the selective pressure placed on individual residues may be dynamic over evolutionary time [67].

Even though different scoring algorithms have drastically different approaches to detecting covariance, the underlying principle of coordinated change within the protein should be susceptible to the same corrections of assumptions. There is no *a priori* reason to believe that a correction based on phylogeny that assists in attenuating a sampling affect from an MSA should be successful with only one algorithm type no matter whether that algorithm is MI or based on perturbation, chemical similarity, or any other methodology. If our hypothesis is correct, it would seem that no one particular algorithm’s sensitivity or specificity would be particularly well suited for increase, implying that all algorithms should see a benefit. Our null hypothesis for this work then is to show that there is no measurable increase in performance using these corrections on other approaches. To test this hypothesis, in this chapter we report results where we measure specificity and sensitivity of algorithms both before and after an APC correction is applied. Our null hypothesis for this work is that there is no change to sensitivity and specificity for the algorithms tested before and after correction. By reviewing the relative performance of algorithms before and after such a correction we can gauge whether APC can be generalized or failing that at least leveraged by one ore more classes of covariance algorithms.

## 2.3 Methods

The November 2008 (version 23.0) of PFAM was downloaded from ftp.sanger.ac.uk. Columns with more than 50% gaps were removed from analysis. Sequences with more than 90% sequence similarity were removed from that particular family. For reasons of performance, families with more than 2,000 sequences were removed from analysis. For concerns raised by other authors over sampling bias problems, only families with at least 200 sequences were used. Each family must have a GF line that contained at least one PDB structure accession number that could be downloaded from rcsb.org. This left us with a dataset of 1186 families from which to make our computations. In order to match a reference sequence in MSA to a candidate structure, we took the best scoring Clustalw [68] version 2.1 ranking.

For receiver operating curve (ROC) curve analysis used in Fig. 2.1, Fig. 2.2, and Fig. 2.3, we defined true positive as those residues within 8 Å as measured by β-carbon to β-carbon. We completed a re-implementation of Gloor's original ANSI C code, which was graciously provided in source format. Our Java implementation was tested against the sample 3H PFAM family (PF02829), which provided a score per score calculation match to within 0.001%.

### 2.3.1 Mutual Information

Mutual information (MI) is a way to measure dependency between two random or seemingly random variables. MI is defined for proteins when using an MSA with columns *i* and *j* as:

$$MI = \sum_{i \in \theta} \sum_{j \in \theta} P_{i,j} \times \log\left(\frac{P_{i,j}}{P_i \times P_j}\right)$$

EQUATION 2.1: Here we have abbreviated Θ to represent all 20 amino acids. This reads as the sum of the products of the probability of each amino acid for each column *i* and *j* joint probability ($P_{i,j}$) times the log probability of that same joint probability divided by the probability of each amino acid in each of the columns *i* and *j*.

Mutual information is at its maximum value when the amino acids always covary with one another. When the probability for column *i* ($P_i$) and the probability for column *j* ($P_j$) are the same a maximal value will be reached. The maximum $\log_{20}$ value for MI, with all 20 amino acids present in both columns is approximately 2.9957 [69].

Recently, a general-purpose phylogenetic correction was published called the average product correction (APC) [26], which functionally improves the predictive power of MI. Using APC on MI, produces what the authors term MIp. MIp is “general-purpose” in that there are no free parameters to rely upon, and the algorithm is computationally trivial. We can think of this APC correction as being a reasonable corollary of the ability of a given column to change. Formally, MIp is defined as:

$$\text{MIp} = \text{MI} - \text{APC}$$

$$\text{APC}(\text{i},\text{j}) = \frac{\text{MI}(\text{i},\bar{\text{x}}) \times \text{MI}(\text{j},\bar{\text{x}})}{\overline{\text{MI}}}$$

EQUATION 2.2: For a given correction for columns *i* and *j*, to determine the average MI for column *i*, represented by $MI(i,\bar{x})$ times the average MI for column *j*, represented by $MI(j,\bar{x})$, and normalize by the average MI, represented by $\overline{MI}$ , to signify the average of all MI scores for all columns.

### 2.3.2 Conservation

Conservation is most often measured using Shannon entropy [70] as described by Shenkin *et al.* [71]:

$$H = -\sum_{\Theta}\left[p_{\Theta}(i)\ln p_{\Theta}(i)\right]$$

EQUATION 2.3: Sequence (or Shannon) entropy, provides us a metric of conservation for column *i* across all possibly amino acids Θ. It is common to see this calculation carried out in protein sequence space using the natural log. Here for a given column *i* the probabilities times the log of those probabilities are summed for all amino acids Θ.

Using the above equation, Shannon entropy would be at its smallest for a column with complete conservation. Since conservation is truly a single column manifestation, to compare this to other algorithms, which score columns *i* and *j*, we took the average of these two values to report.

### 2.3.3 Statistical Coupling Analysis

In 1999, Rama Ranganathan and Steve Lockless [3] published a result that seemed to indicate a veritable breakthrough in the difficulties around determining quantitative thermodynamic answers from sequence data. They proposed an algorithm called Statistical Coupling Analysis (SCA) that was designed as a computationally tractable way of determining the energetic relationship between any two residues based on available sequences. If SCA was to be "the basis for efficient energy conduction within proteins", it would be a generalized formula that would easily show both trivial and non-trivial coupling. In this case according to the authors, trivial being spatially proximate from a known structure, and non-trivial (e.g. non-local) would "represent conduits along which energy distributes through a protein". SCA was defined as follows:

$$P_i^x = \frac{N!}{n_x!(N-n_x)!} \times p_x^{n_x}(1-p_x)^{N-n_x}$$

EQUATION 2.4: Probability distribution function for SCA. Here *i* indicates the particular column, *x* the residue, N is 100, $n_x$ is the % of residue x present in column *i* (expressed as a real number), and $p_x$ represents the probability of a given residue type based on Swiss-prot frequencies, originally taken in October of 1998.

One of the benefits of this formulation is that it scores relative to expected background frequencies of each residue. That is to say the lower a mean frequency of a residue is as given by Swiss-Prot, the more significant the contribution of that residue is to the final score. If we imagine that a residue has a frequency of 2%, and that column *i* has a 10% frequency and column *j* a 20% frequency a 'perturbation' that increases either column by 5% will be more significant for column *j*. This is related to energy by Ranganathan using a Boltzmann constant:

$$\Delta G_i^x = kT \times \ln P_i^x$$

EQUATION 2.5: The change in energy for a given column *i*, is a factor of "an arbitrary energy unit" *kT*, which is the equivalent of the Boltzmann's constant and the measure of mean transition between states. The right most portion of the equation (after the multiplication symbol) is the natural log of the probability of column *i* containing residue *x*. The probability is measured as the binomial probability of the observed number of amino acids *x*.

For SCA, conservation and perturbation are defined respectively on the first and second line of the following equation:

$$\Delta G_i^{stat} = kT * \sqrt{\sum\nolimits_x \left( ln \frac{P_i^x}{P_{MSA}^x} \right)^2}$$

$$\Delta\Delta G_{i,j}^{stat} = \sqrt{\sum\nolimits_x (ln \frac{P_{i|\delta j}^x}{P_{MSA|\delta j}^x} - ln \frac{P_i^x}{P_{MSA}^x})^2}$$

EQUATION 2.6: In the original paper the first line for "statistical" ΔG (abbreviated as "stat") is termed the "overall empirical evolutionary conservation parameter". Here, as previously x are the set of all possible amino acids, with *P* standing for probability and *i* for the column in question. On the second line the lower case delta (δ) indicates a subset of an alignment (by rows), which constitute highest frequency residue x for a given column. This sub setting is why this method and those related to this concept are referred to as "perturbation" methods. Note that on the second line, contrary to the original authors, we drop the kT, which amounts to scalar and is not useful for our calculations.

The original SCA paper made mention of excluding some columns from analysis, as they may not have come to "equilibrium" for analysis. For the purposes of this work, we will ignore this criterion, as none of the other algorithms make such an assertion, and this excludes a significant number of potential scores arbitrarily from those competing score generators.

### 2.3.4 Explicit Likelihood of Subset Co-variation

In 2003, statistical improvements to SCA showed promise, at least for probability of determining CASP like proximity significantly [72]. This improvement to SCA was named the explicit likelihood of subset co-variation (ELSC). SCA before or after improvement (ELSC) was originally described as being able to show the relative "coupling" between any arbitrary positions *i* and *j*, this new algorithm improved results as measured by proximity significantly.

The improvements were achieved by modifying the probability function to a more simplified sequence entropy computation, which was verified by comparing the SCA scores to sequence entropy via a Spearman rank-order correlation which showed an $r^2$ correlation >90%. The formulation was to replace what SCA used in Eq. 2.4 with the more traditionally statistically used equation for Shanon entropy as seen in Eq. 2.3. Then to determine the possible combinations of subsets (perturbations), and introduce a term $\Omega j^{<i>}$ relating the two columns *i* and *j*:

$$\Omega_j^{<i>} = \left( \frac{MSA_{x_1,j}}{MSA \mid \delta j_{x_1,j}} \right) \times \left( \frac{MSA_{x_2,j}}{MSA \mid \delta j_{x_2,j}} \right) \cdots \left( \frac{MSA_{x_{20},j}}{MSA \mid \delta j_{x_{20},j}} \right)$$

EQUATION 2.7: Note here that the convention for $x_1$ through $x_{20}$ is for all 20 amino acids alanine through Valine. "MSA" here refers to the multiple sequence alignments, with *i* and *j* referring to the respective columns within that MSA.

To determine the combination probability, we use a common statistical tool most often referred to as the "choose" function. If we make allowances for substitutions, in this case using *N* and *n*, we can more easily see how this common predictor for probability is being leveraged.

$$N = MSA_{x_1,j}$$

$$n = MSA \mid \delta j_{x_2,j}$$

$$\left( \frac{N}{n} \right) = \frac{N!}{n!(N-n)!}$$

EQUATION 2.8: This combination equation is done for each of the terms of the previous Eq. 2.6. The last line is the combinatorial factor which measures the subset (*n*) for the residue *x* for the permutations that are found in the complete MSA (N).

Using these more traditional statistical methods, ELSC was able to improve proximity CASP like calculations approximately twice as successfully when compared to SCA.

### 2.3.5 The McLachlan based Substitution Correlation

The McLachlan [73] based Substitution Correlation (McBASC), is dissimilar to a perturbation approach of ELSC or SCA, and in that sense more similar to mutual information. McBASC at its heart is a comparison of vectors, where the columns of an MSA represent those vectors. An item that is unique to McBASC is the use of a matrix similarity table [37]. One of the more common in the literature is the McLachlan substitution matrix from a 1971 work of the same name [74]. McLachlan reasoned that both the chemical nature of amino acids such as polarity, hydrophibicity etc. needed to be weighted in addition to the observed frequencies of amino acid replacements in homologous proteins. McBASC then uses a Pearson Correlation Coefficient (PCC) to determine similarity.

For ease of description, we will assume that a substitution matrix “COVARYBINARY” where we will use is "1" for identical residues (e.g. either no substitution occurred or a synonymous substitution occurred), and "0" otherwise. In a real substitution matrix, the comparisons would be much more varied, but for illustration this should suffice. If we use TABLE 2.1 as a reference, then the vectors can be created as:

$$V_I = \{AC, AH, AC, AA, AQ, CH, CC, CA, CQ, HC, HA, HQ, CA, CQ, AQ\}$$
$$V_J = \{AG, AL, AA, AG, AA, GL, GA, GG, GA, LA, LG, LA, AG, AA, GA\}$$

EQUATION 2.9: Using TABLE 2.1 as the reference, String representation for vector of column *i* and column *j*. Note how we compare pairs in a descending recursive fashion. For this to work, the columns must be of the same length, so for columns with gaps we can either omit those gaps for both vectors, or not score that column pair.

If we consider Eq. 2.9 but with the substitution matrix of COVARYBINARY, we can simplify the vectors, with cardinal integers. Then, as the penultimate step, we create vectors that we can use in our PCC calculation:

$$V_I = \{0,0,0,1,0,0,1,0,0,0,0,0,0,0,0\}$$
$$V_J = \{0,0,1,0,1,0,0,1,0,0,0,0,0,1,0\}$$

EQUATION 2.10: Vector from EQUATION 2.11 after transformation using our "1" and "0" substation matrix, for column *i* and *j*.

As the last step to produce a McBASC score, we calculate the Pearson Correlation Coefficient:

$$PCC = \frac{\sum IJ - \frac{\sum I \sum J}{N}}{\sqrt{\left(\sum I^2 - \frac{\left(\sum I\right)^2}{N}\right) - \sum J^2 - \frac{\left(\sum J\right)^2}{N}}}$$

EQUATION 2.11: N here is the number of rows, which for our example presented in Table 2.1 would be 6. ΣIJ is the sum of the products of the paired scores, in this case 0. ΣI is the sum scores for I, in this case 2. ΣJ here is 4.

Note that this formulation differs from SCA in that the results are symmetrical. That is to say that a score (*i*,*j*) and (*j*,*i*) would result in the same value.

## 2.4 Results

We compared the algorithms for MI, SCA, OMES, and McBASC with and without phylogenetic correction on a test set of 1186 protein families (see methods). For all algorithms, we ranked results by score and considered a true positive to be two residues in a corresponding protein structure of ≤ 8 Å. ROC curves for algorithms not corrected are shown in Fig. 2.1. Fig 2.2 shows the APC corrected version of the algorithms shown in Fig. 2.1. Lastly, Fig. 2.3 shows a combined view of both Fig. 2.1 and Fig. 2.2.

What we can see in Fig. 2.1 the generalized performance of these covariance algorithms is quite poor. What we notice in Fig. 2.2 and Fig. 2.3 however is that a phylogenetic correction has a large effect on the area under the curve for both the MI and McBASC algorithms. For the different approaches, we can see that McBASC responds favorably to the APC correct, but unfortunately is still performing much worse than correct MI (e.g. MIp). Ashkenzy and Kilger showed that APC can improve McBASC scores, but most dramatically when compared against other algorithms only at very shallow alignment depths [75] which we had excluded in this study.

## 2.5 Conclusions

Our work here was completed during the time that Ashkenzy and Kilger [75] were submitting their work for publishing. Because of the filtering for small alignments, we would not have found the correction for McBASC to be as compelling, as their work concentrated on sequence depths of 20-100. It is interesting nonetheless that it would appear that the improvement of APC on McBASC is consistent on all alignment sizes, but only at trivially small alignments does it outperform MI with APC. Using McBASC uncorrected no longer bears any benefit, and in that McBASC is much the same as MI. The underlying cause for why APC produces better results for some algorithm types but not others is not known.

What we have shown here is that some covariance algorithms are more susceptible to corrections in phylogeny than others. Possibly, those algorithms not responsive to APC would perform better if measured using a Mahalanobis distance as the metric [76], which might indicate that some algorithms are detecting covariance necessary during protein folding rather than resulting structure. It may also be the case that algorithms like OMES or SCA behave more like McBASC in that an APC correction would show measurable benefits of detecting proximity with very small sequence depths. Regardless of underlying principle, it would appear that in order to discriminate what the fundamental cause(s) are that account for this difference in algorithmic behavior will take additional research.

TABLE 2.1: A toy McBASC alignment example. The one-letter amino acid codes are used to form a set of vectors, which are named for the columns that they represent. These vectors can be seen in EQUATION 2.11

| **Column *i*** | **Column *j*** |
|---|---|
| A | A |
| C | G |
| H | L |
| C | A |
| A | G |
| Q | A |

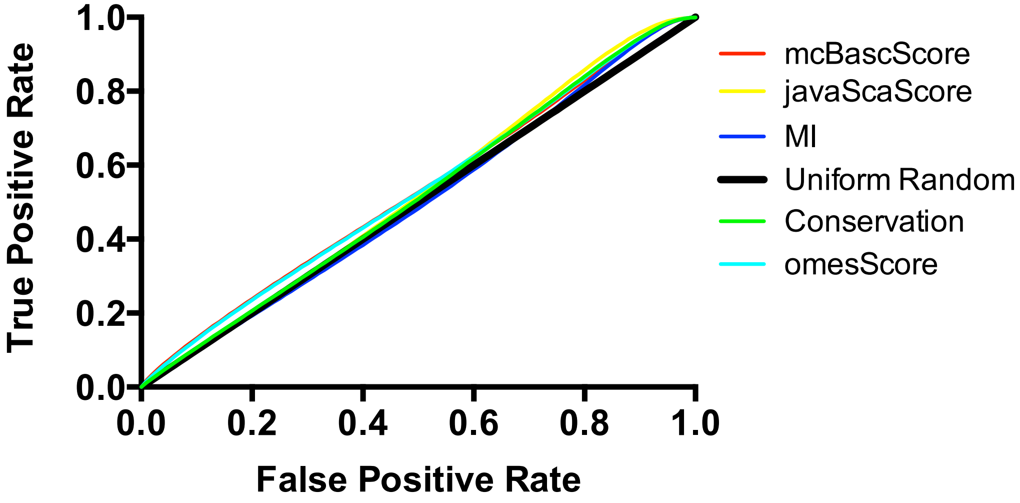

FIGURE 2.1: ROC curve showing the relative performance of each algorithm. Here the true positive was defined as a CASP like proximity of 8 Å as measured between β-carbons of the participating residues of the protein structure. These are the base algorithms without any phylogenetic correction.

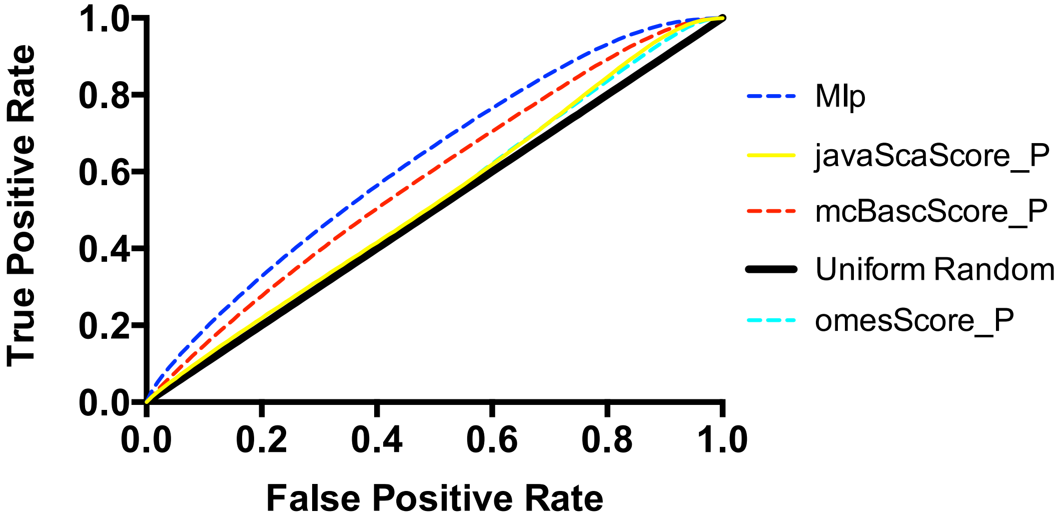

FIGURE 2.2: ROC curve showing the relative performance of each algorithm. The algorithms that are shown with a "_p" suffix indicate that the algorithm has had the APC correction applied. Here the true positive was defined as a CASP like proximity of 8 Å as measured between β-carbons of the participating residues of the protein structure. Here, each of the algorithms is corrected using the APC method to correct for sampling bias.

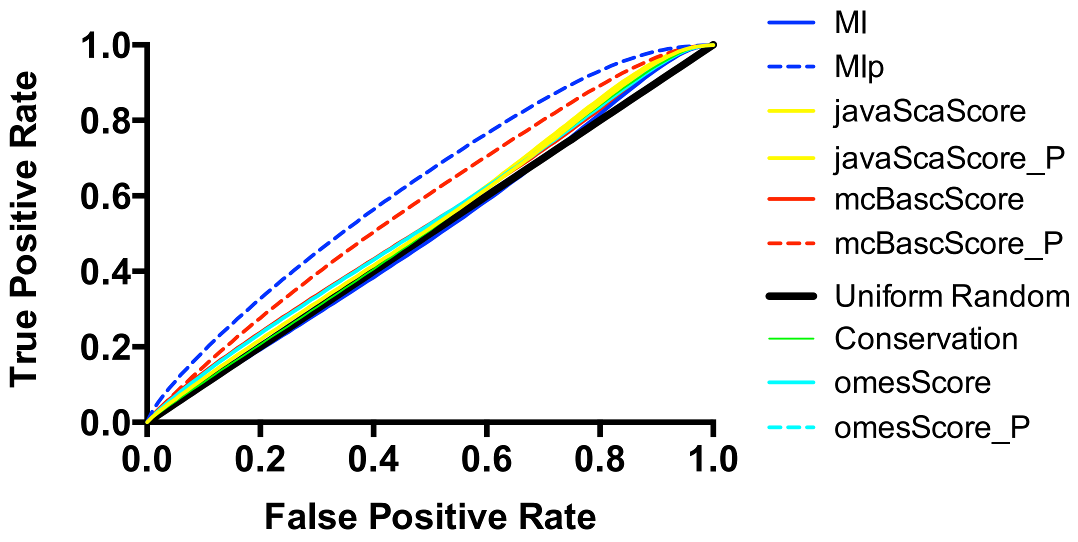

FIGURE 2.3: ROC curve showing the relative performance of each algorithm. Here the true positive was defined as a CASP like proximity of 8 Å as measured between β-carbons of the participating residues of the protein structure. The algorithms that are shown with a "_p" suffix indicate that the algorithm has had the APC correction applied. Here was can easily see the breakout performance for MIp.

# CHAPTER 3: EVIDENCE FOR COVARIANCE IN RIGID STRUCTURES

## 3.1 Abstract

It is clear from evidence of energetics and functionality that all sequence pairs of a protein are not equal participants. Certain sequence combinations are more critical than others, which we know from empirical evidence if not from first principles. Predicting which changes in sequence are maintained (e.g. covary with each other) has been only moderately successful (see Chapter 1).

Covariance algorithms have historically been measured by the distances of residues in the static structure coordinates represented by a PDB file resulting from NMR or crystallography. Recognizing that residues have varying freedoms of movement in Cartesian space may allow us to better understand the prevalence of false positives of residues that rank highly in covarying algorithms.

In this work we will use a methodology termed Floppy Inclusions and Rigid Substructure Topography (FIRST) that takes as input a protein structure and outputs information on the nature of which residues are fixed in relation to other residues in a given structure. When residues are determined to be fixed in relation to one another they are termed "rigid", which has been shown to have thermodynamic properties distinguished from residues that do not have this relationship. Then, we will scrutinize covariance for pairings that occur in the same rigid body versus pairings that are in

different rigid bodies. If the performance of covariance algorithms is significantly different based on these criteria, it could inform both methodologies.

3.2 Introduction

Rigidity can be a useful tool in describing a mechanical system. Rigidity has been used to study colloids [77], glasses [78], gels [79], and proteins to determine a variety of physically macroscopic phenomenon. Changes in rigidity have been linked to ligand binding, nonadditivity, deformation at stress, "gel point", just to mention a few well characterized empirical uses.

Rigidity theory has been evolving on mechanical systems for more than a hundred years. In the mid 19th century, Maxwell [80] determined how to easily compute the stability of an arbitrary system constructed of struts, which are only joined at their ends via determination of the elastic and deformation behaviors of sub-systems. The number of sub-systems in a given topology that would deform with a trivial energy input are referred to as *floppy-modes* [81]. For our purposes these floppy-modes would be found in the non-colored sections of Fig. 3.2.

If we consider a protein as a mechanical system, there are several items to model mechanically such as angular constraints and bond length/strength. Bond strength can vary from strong non-bending covalent to weak long-range electrostatic forces, making the basic problem of what constitutes a "constraint" all the more difficult. Rigidity can be, for our purposes, broken down into two separate modes. These modes herein we will refer to as "rigid" or "floppy". Floppy or rigid can refer to either an entire graph of vertices and edges or a portion (sub-graph). Because sub-graphs can themselves be rigid, but have floppy portions between them we call these areas between rigid regions "hinges". It is crucial to describe the nodes as belonging to named (or in our case numbered) sub-graphs, so that relative comparisons between nodes -- atoms in our case, are meaningful.

A major milestone in working on large systems such as proteins (all atom), came about in the 1990s with the work of Jacobs *et. al.* in an algorithm called the pebble game [82], which was later extended to a three dimensional models with software called Floppy Inclusions and Rigid Substructure Topography (FIRST). FIRST is capable of reading in standard descriptions of Cartesian coordinates of proteins such as a PDB flat file. The pebble game is capable of recognizing that certain constraints are redundant, which is to say that once a region is rigid, adding more constraints is futile with respect to achieving additional rigidity (being a binary value). This can be illustrated with a simple two-dimensional model as shown in Fig. 3.1. For the work here a numbered (e.g. labeled) graph equates to the rigid cluster identifier from "rigid" column as output from FIRST.

What is clear is that "not all sterically allowed conformations are equally populated", this is quite likely true of all conformational space, not just the calmodulin that was studied by Bertini [83]. Describing the portions of structure space that are more or less fixed (relative to some portion or cluster) is a major goal of rigidity theory as it relates to proteins. Studies have also shown that energies for double mutant cycles for residues in the same cluster to have significant probability of non-additive "thermodynamic coupling" [41].

The literature on the relationship between protein function and structure being related to various flexible and rigid criteria is well documented. Bertini noted that active sites tend to be at or near hinge regions [83], that is to say rigid bodies separated by relatively small flexible regions. Tsai *et. al.* noted that there is a link between the so-called protein folding "funnels" and flexibility [84]. When proteins are denatured, there is a concomitant loss of rigidity [85] that occurs precipitously to make the loss of functionality.

Mutual information (MI) has been used to help detect active sites [86]. Covariance algorithms have also been used, albeit indirectly through constraints and molecular dynamic simulation to have success at predicting proper protein folding [87]. A version of MI that uses a phylogenetic correction (Eq. 2.2), termed MIp has been shown to correlate well with residue distance in a protein structure (Fig. 2.3).

If both covariance signal (as measured by MIp) and rigid body analysis (as measured by FIRST) have overlap in measuring the characteristics of protein function and structure then our null hypothesis will be that the signals that these two approaches yield will have no correlation. If there is a correlation between measurements of covariance as measured by MIp and whether residues are in the same rigid cluster it could help to explain some of the false positives that are so common to covariance algorithms, because the covariance is mechanically linked even if not proximate in terms of a CASP cutoff value for interaction. Essentially, a link between covariance and rigidity could be shown to be modeling the same underlying constraints from two orthologous techniques.

## 3.3 Methods

Protein structures were all downloaded from *rcsb.org* unless otherwise noted. All PDB files had hydrogen atoms added prior to further analysis in FIRST using web services available at Duke University labs MolProbity [88]. These Hydrogen atoms themselves that are added, are used for rigid structure determination. MolProbity was used with all of the default values available with version 3.15. To produce Fig. 3.2 coloring was the default from Pymol [89] version 1.5.0.2 output of FIRST, with colors representing unique rigid bodies.

FIRST was compiled from version 6.0.1 that was graciously provided in source and binary format from Kirill Speranskiy at Arizona State University. The default values were used, with the exception of hydrophobic tethers, which were manually set to 3, as our compiled version would not default correctly. In the output from FIRST, most protein residues will have atoms that participate in several clusters, though most of these are trivially small. In our case we chose the simple majority of numbered clustered output for the entire residue. The last step in determining whether residues were part of the same rigid body was to compare the majority identifier provided in the FIRST output to one another, with a simple Boolean value of TRUE for identical majorities, and otherwise FALSE.

The output from FIRST has a bias towards sequence distances that are trivially close. This bias of the same cluster cohort can be peripherally viewed in Fig. 3.2 and Fig. 3.3, and is discretely shown in Fig 3.4. What was needed was a way to remove the difference in the population of sequence distance for the two cohorts.

To overcome the bias of proximity in sequence, a sampling methodology was instituted. To sample each cohort, we first binned the sequence distance as would be measured by the absolute value of the difference of column *i* and column *j* of the MSA. For all but the most distal sequence pairs, several hundred results were measured for each discreet bin. In the cases where there were fewer than 100 scores present for a given bin size, those *i,j* combinations were excluded from any further analysis. We then randomly sampled 100 data points from each bin (e.g. sequence pairing distance). A technical replicate of the sampling was performed 100 times for each MSA tested.

To produce a p-value, we took a measurement of the median amino acid distances for the top 100 MIp values. Distances between amino acids were measured as all atom. The choice of 100 residue pairs is arbitrary, though changing this threshold to 75 or 125 residues had almost no change in p-values. If we set a threshold of the 50th percentile we would expect that random pairings of all columns would find 50 pairings below this median distance. Because the pool of potential scores was several thousand, and 100 scores is a relatively small number of scores to choose from a population so large, we used a cumulative binomial distribution to produce p-values.

$$\sum_{1}^{N}\left(\frac{N!}{n!\,(N-n)!}\right) * 0.5^{n} * (1-.5)^{N-n}$$

EQUATION 3.1: N is the number of top scores that we measured, which was equal to 100. Since we chose the 50th percentile of distances for the entire population of pairings, the probability of random chance would be 50% (0.5) for a particular *i,j* combination to be within the median Å distance for the highest 100 MIp scores. *n* are the number of pairings that we counted below the 50th percentile in the top 100 scores. A random scoring algorithm would find 50 pairings below the median Å distance, with better than random algorithms finding more than 50 such column pairings.

The results of all pooled PFAM families for this test are given in Fig. 3.6.

Mutual Information with a phylogenic correction was computed as in Eq. 2.2. Distances were computed by β-carbon distances from the relevant protein structure, which was determined by a Clustalw [68] version 2.1 comparison. Clustalw was used to compare the PFAM sequences scored against the relevant PDB structure sequence. The best match was used to map between PFAM column and structure residue. If Glycine was the residue in question for distance measurements, an α-carbon was used instead.

## 3.4 Results

To determine if there was a correlation between the ability of MIp to identify proximate residues the results of MIp for the two cohorts, namely those in the same rigid body and those not in the same rigid body, were compared,. A set of 11 PFAM (see Tab. 3.1) families that were used from the original MIp paper that have documented correlation detectable by MIp of residues at sequence distance greater than 10 residues. The initial step after producing MIp scores (see methods) was to create the sets of rigid bodies, which was accomplished using the FIRST tool. A representative example of FIRST output is provided for the 2axn PDB (from the PFAM01591 alignment). The output for FIRST with 2ax resulted in a structure with 13 rigid bodies as depicted in Fig. 3.2. The structure shown in Fig 3.2 is after Hydrogen atoms were added by Molprobity. FIRST calculations were carried out as detailed in the methods section.

In Fig. 3.3 (a) two cohorts of columns *i* and *j* being in the same rigid body or in different rigid bodies the same PFAM family as Fig 3.2 is shown as a scatterplot of MIp score and residue to residue distance. In the two panels of Fig. 3.3 (a), we can see the line-of-fit for both cohorts. In panel (a) the more positive a slope (as seen in the right panel same rigid cluster) would indicate a higher correlation. The green colored dots in Fig 3.3 (a) are those *i,j* column pairings that are at a trivial distance of < 10 sequence separation. There is a higher concentration of green dots in the right panel correlate with shorter distances. This was the first indication that the improved slope seen in Fig. 3.3 (a) was an artifact rather than an affect.

In order to investigate the potential of bias from sequence proximity a histogram of scores between the two cohorts was created illustrated as Fig. 3.3 (b). Here we can see both the preponderance of trivially distant pairs (1-10 in particular) and the surreptitious drop of distal residues for the same rigid cluster population.

To investigate the hypothesis that MIp scores are more predictive of distance in the case of identical rigid bodies we first looked to remove the sequence distance bias seen in Fig. 3.3a. This was accomplished by sampling each cohort at each discreet sequence distance (for details see methods). To assign p-values of the ability of MIp to correctly measure proximity from these resulting sampled graphs, the cumulative binomial distribution was used (see methods). The results for the 11 families studied after the sequence bias was removed are shown in Fig. 3.3 (c). This figure reveals little to no efficacy of rigid body relationship in influencing MIp's ability to discern Cartesian distance of residues.

## 3.5 Conclusions

Unfortunately we were unable to detect a relationship between covariance and rigid bodies. Whether this is due to weakness of these algorithms to detect relationships as outlined in Chapter 1 or a problem of approach is unknown. Since completing this work, further advancements in both rigidity theory and covariance algorithmic design have been made. Advancements in knowledge of rigidity in proteins now allows for assigning continuous assignment of relative rigidity instead of the ordinal assignments used in this study [90]. There are also indications of rigidity "networks" involved in distant covariance [91] as well as indications that flexibility between orthologous proteins may vary significantly enough within a given MSA to confound the signal we are looking to detect across such large evolutionary time periods. Covariance algorithms have recently also evolved to include additional information from groups of columns (contiguous or not) that may be a closer approximation of the evolutionary constraints driving covariance (see Chapter 4 for additional information).

### 3.6 Acknowledgements

Kirill Speranskiy at Arizona State University for his assistance with what was believed to be a bug in FIRST 5.0. His help in determining that this was instead a result of constraints on PDB file layout was both insightful and critical in the correct parsing of the FIRST output.

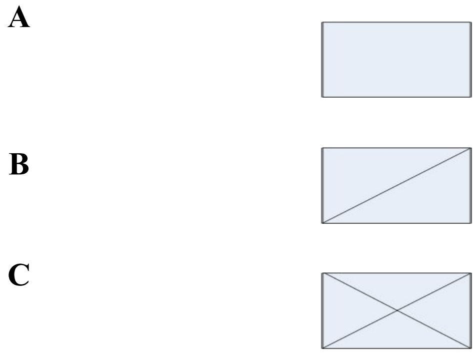

FIGURE 3.1: (**a**) A simple two-dimensional model indicating four constraints as lines of a rectangle. Note that this figure in two dimensions is flexible as any number of parallelograms are possible without breaking a constraint (line). (**b**) A rectangle with five constraints as indicated, allows us to illustrate a simple model as what we term in this manuscript as "rigid". In a chemical sense, this object has less enthalpy as a chemical bond has formed, and less entropy, as less motion is now possible as compared to our previous case. This particular bond network is sometimes termed "isostatic" because we have only just achieved a rigid state with this last bond being added. (**c**) If we create a six-constraint rectangle as indicated here, we will decrease the enthalpy as compared to a five-constraint rectangle. Note however, that a corresponding change in entropy has not occurred. The sixth constraint added has no impact on entropy because we are still "rigid". In this respect the additional bond that we have interjected is redundant.

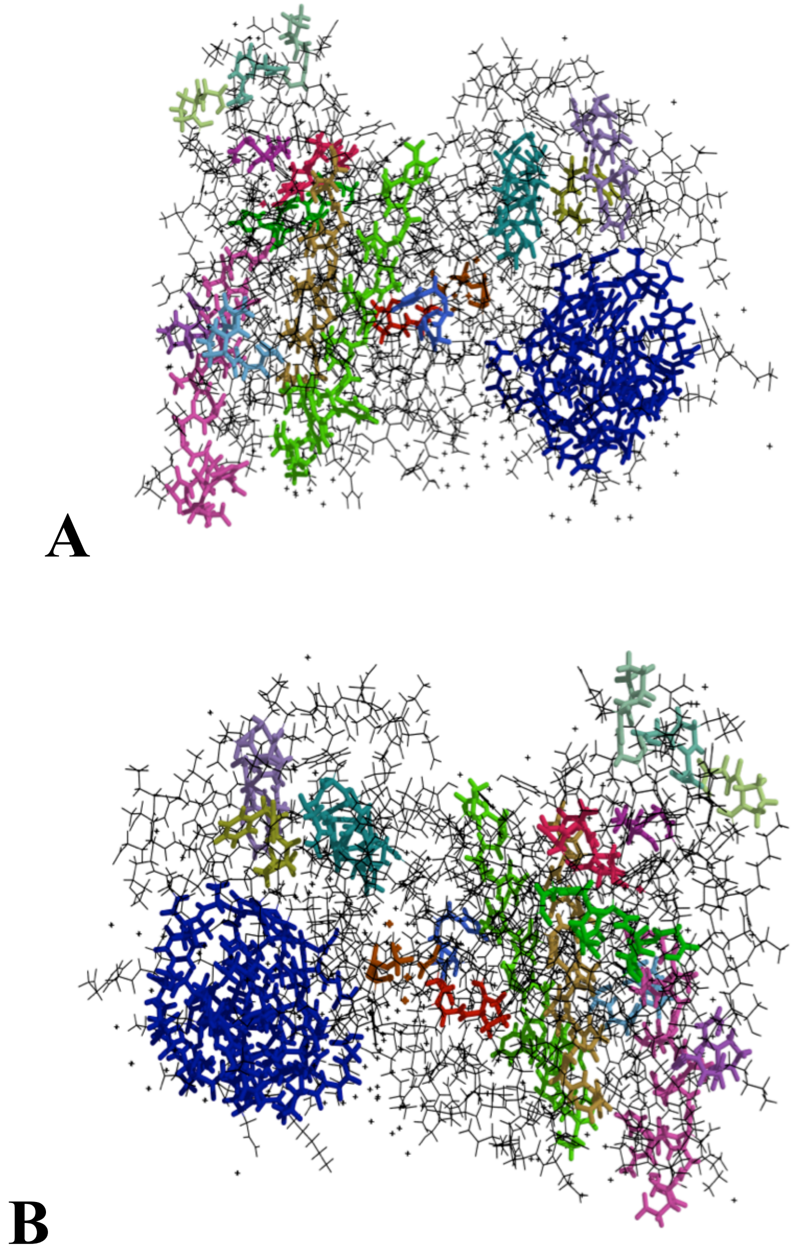

FIGURE 3.2: The 2axn protein structure as updated with Molprobity to include Hydrogen atoms. Once hydrogen atoms have been added, the FIRST program will generate Pymol compatible coloring, which can be output for viewing. Here we see 18 separate non-trivial rigid bodies, which are determined based on degrees of freedom as indicated in Figures 2.1, 2.2, and 2.3. Rigid bodies of any significant size tend to have coiled regions interspersed, so the likelihood of non-continuous sequences is likely. Panel "A" and "B" are 180 pivoted views of the same structure.

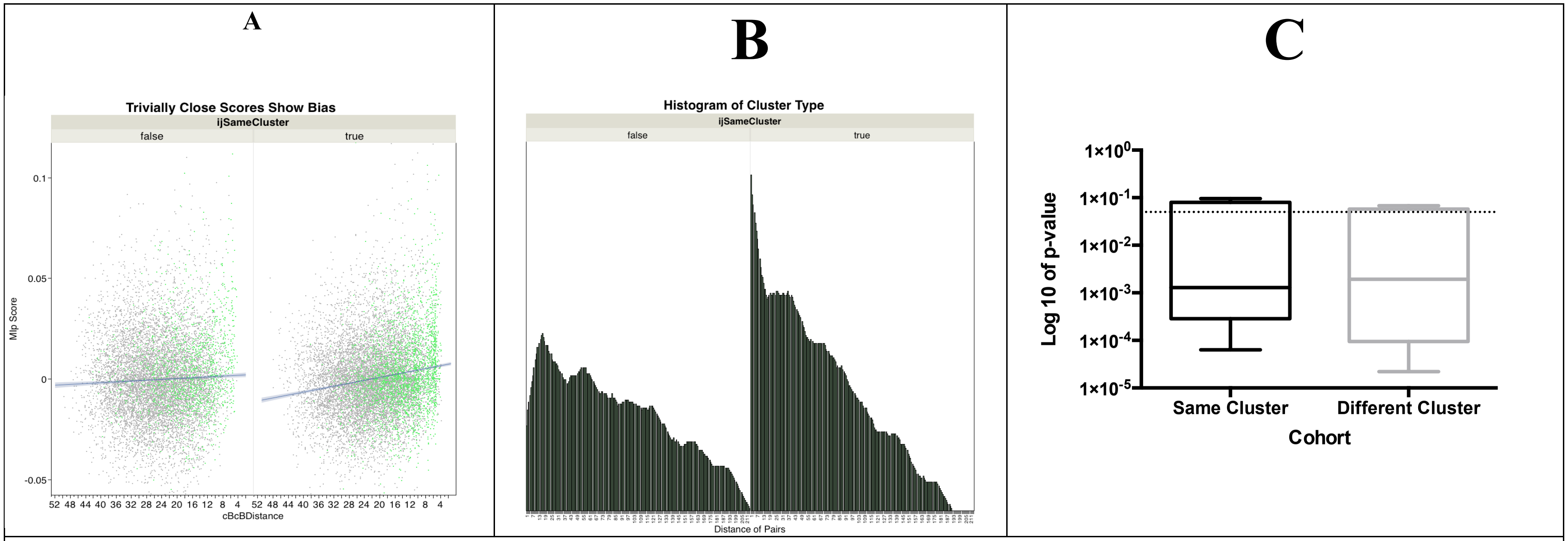

FIGURE 3.3: Shown in all panels are results for the 2axn protein, which was the best sequence match for the PFAM accession number pfam01591. (**A**) Shown here is the performance of MIp, measured the MIp score on the y-axis and the x-axis being a measurement of β-carbon distances in a descending scale. The green dots on the plot represent sequence separation for a given pairing of 10 residues or closer. Panels for the case of residues occurring in the same rigid cluster (right panel) are contrasted with those in a different cluster (left panel). The density of green is more concentrated on the rigid cluster panel at the closer distance. The line is a fit of all data points, with the shaded region indicating confidence. (**B**) A histogram of the relative abundance of sequence pairs for cases where the pairs are either in different rigid bodies (left panel) or in the same rigid bodies (right panel). What we notice here is that when comparing the parings of given columns *i,j* there is a strong bias for lower sequence distance when the sequences are found to be in the same cluster. (C) After removing the sequence distance bias with sampling, p-values were measured for all the PFAM families tested. Probabilities were generated from the binomial probability of picking better than the mean distance values for the top 100 MIp scores. Once sequence distance bias is removed we can see no discernable performance delta between the residues in the same rigid body cohort as those residue pairings in different rigid bodies. The dotted line represents a probability of 0.05. The centerline in each box represents the median value. The outer lines of each box plot represent the 25th and 75th percentile, with the whiskers representing the 10th and 90th percentiles.

# CHAPTER 4: COVARIANCE IN ALIGNMENTS USING COLUMN GROUPS

## 4.1 Abstract

Algorithms that detect covariance between pairs of columns in multiple sequence alignments are commonly employed to predict functionally important residues and structural contacts. However, the assumption that co-variance only occurs between individual residues in the protein is more driven by computational convenience rather than fundamental protein architecture. Here we develop a novel algorithm that defines a covariance score across two groups of columns where each group represents a stretch of contiguous alignment columns in the alignment.

We define a test set that consists of secondary structure elements (α-helixes and β-strands) across more than 800 PFAM families. Using these alignments to predict segments that are physically close in structure, we show that our method substantially out-performs approaches that aggregate the results of algorithms that operate on individual column pairs. Our approach demonstrates that considering units of proteins beyond pairs of columns can improve the power and utility of covariance algorithms.

## 4.2 Introduction

One of the “grand challenges” [92] of structural genomics is to elicit structural information from sequence alone. The relationship between compensatory changes (e.g. mutations) of amino acids within structurally con-strained regions of homologous proteins has been an active area of research since the pioneering work of Altschuh [93].

To date, most algorithms use pairs of single columns as the unit of covariation. Covariance between pairs of columns has been used to find errors in alignments [94], locate point of inter-protein docking [28], and to search for packing specific to α-helixes to α-helixes distances [95]. Approaches for these algorithms have varied with scoring based on substitution matrices [37, 75], chi-squared tests [96], perturbation [3, 72], and more recently for large multiple sequence alignments (MSA) the inverse of sparse covariance estimations [97]. Recent improvements in these algorithms have been substantial [1], though the basis for improvements have varied significantly including machine learning [97, 98], tangential information such as solvent accessibility [99], and phylogeny based corrections [26].

There is no *a priori* reason to think that covariance is limited to individual pairs of residues. A number of researchers therefore have explored methods beyond simple pairs of columns. These methodologies often work with groups of columns (GOC) that are not contiguous within sequence. In one example of this approach, Halabi *et al.* [43] utilized what they termed “Sectors” in which information from the SCA algorithm is expanded to multiple sets of columns. In another example, Burger *et al.* [100] utilized a graph based model that relies on Bayesian statistics to score sets of inter-related columns.
Stretches of residues that are continuous within the protein sequence around structural and functional sites are often conserved [101-104]. It is therefore reasonable to believe that algorithms that work on these continuous GOCs could provide insights into proteins that would be missed by algorithms that work on pairs of columns or on discontinuous sets of columns. With this in mind, we developed an algorithm that detects covariance in these continuous stretches of sequence. Since the number of permutations for arbitrary

non-overlapping GOC for even a modestly sized protein is exponentially large, we developed a test set for our algorithm that focuses on secondary structure elements (SS), specifically α-helixes and β-strands. This approach is attractive because secondary structure elements are predefined, obviously relevant to structure, non-overlapping, and modest in number. As demonstrated below, our approach significantly outperforms methods that aggregate covariance results from pairs of columns.

## 4.3 Methods

Covariance algorithms applied to individual pairs of columns and groups of columns. The goal of this paper is to compare algorithms that calculate covariance on a pair of contiguous "groups of columns" (GOC) within a protein multiple sequence alignment. These algorithms extend algorithms that calculate covariance on a pair of columns. The following algorithms are evaluated in this paper.

### 4.3.1 Average McBASC

The McLachlan [73] based Substitution Correlation (McBASC) algorithm works on a single pair of columns and has been previous described [18]. Briefly, if N is the number of sequences in the alignment, to calculate a covariance score for columns *i* and *j*, we create a vector of length $\binom{N}{2}$ for each column. With *k* and *l* defined as indexes of each sequence within the alignment, each vector is populated with the values of scores from the McLachlan substitution matrix that result from comparing the residues within each column for all possible comparisons of sequences *k* and *l* (with *k* != *l*). The McBASC score *r*, for a given *i,j* column combination is given by:

$$r_{i,j} = \frac{1}{N^2} \cdot \frac{\sum_{kl}(s_{ikl} - \langle s_i \rangle)(s_{jkl} - \langle s_j \rangle)}{\sigma_i \sigma_j}$$

EQUATION 4.1: $\langle S \rangle$ is the average and $\sigma$ is the standard deviation for all the entries in each of the two vectors (Fig 4.1A).

*r* can range from -1 to 1 inclusive with the highest score indicating the highest level of covariance. A score of 1 was assigned for any *i,j* column pair where there was a gap at sequence for that particular sequence *k* or *l*. For performance reasons, our implementation (https://github.com/afodor/cobs) produces values of r that are approximate to Eq. 1 (with differences for alignments of >50 sequences of less than 1% from the value of *r* defined in Eq. 4.1).

The "McBASC Average" as indicated by the name is the result of taking the average McBASC score for each pair of column within two GOCs. That is, if there are $g_1$ columns in one GOC and $g_2$ columns in another GOC, the "Average McBASC" score is defined as the mean of the $g_1 \cdot g_2$ McBASC scores produced by calculating McBASC for all $g_1$ versus $g_2$ columns.

### 4.3.2 COBS

As a simple alternative to averaging all possible McBASC scores within two GOCs, we propose COBS (COvariance By Sections), a straight-forward extension of McBASC to groups of contiguous columns (Fig. 4.1B). As in McBASC, we end up with a pair of vectors which are compared by Pearson correlation to give a final score. If *i* represents a GOCs of length *m* within the alignment, and *k* and *l* are indexes of each sequence within the alignment, then the value placed within the vector for *i* is given by:

$$\mathrm{S}_{ikl} = \sum_{m} \mathrm{McLachlan}(k_m, l_m)$$

EQUATION 4.2: The McLachlan function returns the substitution matrix value comparing the residues at position *m* within the GOC for sequences *k* and *l* (Fig. 4.1B)

To generate a COBS score for GOC *i* vs. GOC *j*, vectors of length $\binom{N}{2}$ are generated for each GOC and over all possible comparisons of *k* and *l* (with *k* != *l*), the two vectors are populated with the values generated by Eq. 4.2. The vectors are scored by the Pearson correlation as indicated by Eq. 4.1 to generate a final COBS score. GOCs that are perfectly conserved are given a score of 1.

### 4.3.3 Average Conservation

We calculated Shannon Entropy as canonically defined [71]:

$$-\sum_{x_1}^{x_{20}} (p_x(i) \ln p_x(i))$$

EQUATION 4.3: With *x* being indexed across all 20 amino acids, $p_x$ representing the frequency of the particular amino acid at the $i^{th}$ column.

The “Conservation Average” is the mean value for this value across all the columns in the pair of GOCs.

### 4.3.4 Mutual Information

Mutual information was implemented as previously described [18]. As was the case for average McBASC, we define average MI as the mean of the $g_1 \cdot g_2$ MI scores from 2 groups of columns (with $g_1$ columns in the first group and $g_2$ columns in the second group).

### 4.3.5 Phylogenetic Correction

MI has been shown to be an ineffective measure of covariance within protein alignments [18] with a high sensitivity to phylogenetic artifacts in the alignment. A procedure to correct for these artifacts has been introduced (Dunn, et al., 2008) and been shown to substantially improve the performance of MI. If MI scores have, as indicated above, been calculated for all pairs of columns *i* and *j* in the alignment, then MI with a phylogenetic correction [26] termed MIp is calculated as:

$$MIp(i,j) = MI(i,j) - \frac{MI(i,\bar{x}) \cdot MI(j,\bar{x})}{\overline{MI}}$$

EQUATION 4.4: $MI(i,\bar{x})$ is the average MI score of column *i* with all other columns in the alignment, $MI(j,\bar{x})$ is the average MI score of column *j* with all other columns in the alignment and $\overline{MI}$ is the average of all MI scores from all pairs of columns in the alignment.

The APC correction has been shown to work with McBASC, at least for small alignments [32] previously. We use the same correction on McBASC as we defined for MI:

$$McBASCp(i,j) = McBASCp(i,j) - \frac{McBASC(i,\bar{x}) \cdot McBASC(j,\bar{x})}{\overline{McBASC}}$$

EQUATION 4.5: $McBASC(i,\bar{x})$ is the average McBASC score of column *i* with all other columns in the alignment, $McBASC(j,\bar{x})$ is the average McBASC score of column *j* with all other columns in the alignment and $\overline{McBASC}$ is the average of all MI scores from all pairs of columns in the alignment.

As with “Average McBASC” and “Average MI”, we define “Average MIp” and “Average McBASCp” as the mean of the $g_1 \cdot g_2$ MIp or McBASCp scores respectively from 2 groups of columns (with $g_1$ columns in the first group and $g_2$ columns in the second group). Although implemented originally for use with algorithms that work on pairs of columns, the phylogenetic correction algorithm can be applied to any other covariance score [100, 105]. If we have Y GOCs in our dataset, we will generate a total of $\binom{y}{2}$ COBS scores. For each pair of GOCs, $i$ and $j$, the phylogenetic corrected COBS score (which we call COBSp) is given by:

$$COBSp(SS_i, SS_j) = COBS(SS_i, SS_j) - \frac{COBS(SS_i, \bar{x}) \cdot COBS(SS_j, \bar{x})}{\overline{COBS}}$$

EQUATION 4.6: $COBS(i, \bar{x})$ is the average COBSs score between GOC $i$ and all other GOCs in the alignment, $COBS(j, \bar{x})$ is the average COBS score for GOC $j$ and all other GOCs in the alignment and $\overline{COBS}$ represents the average of all COBS scores for the alignment in question.

As an additional permutation, the APC correction for MI and McBASC can be applied on a per column basis as defined above and was originally done for Mutual Information, or can be done on GOCs, which is what COBS must use since it only works at the GOC level. In our implementation (https://github.com/afodor/cobs), we output phylogenetic correction for McBASC and MI at the pair of column level (before pairs of columns are averaged), at the group of column level (after pairs of columns are averaged) and applied twice: initially at the pair of columns and then again at the group of columns level. None of these normalization schemes consistently out-perform any of the others for McBASC and MIp (data not shown). In this paper, therefore, we only report normalization at the single pair of column level (before pairs of columns are averaged) which is most consistent with how phylogenetic corrections have been previously utilized in the literature.

4.4 Source Data and Distance Computation

Version 26 (November 2011) of PFAM [106] was downloaded from ftp.sanger.ac.uk. Protein families were chosen that had at least one protein referenced in the GF DR line. PDB structures were assigned to PFAM families based on Sanger mappings (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/mappings/pdb_pfam_mapping.txt). SS elements were determined using the “HELIX” or “SHEET” indicators in the remarks section of the selected PDB file, which was downloaded from rcsb.org. Distances between all β-carbons for a given SS were measured against all the β-carbons of the other SS being compared. In the cases where Glycine was part of the measurement α-carbons were used.

To eliminate the possibility that quality of alignment would help explain the variation of the best performing three algorithms, namely McBasc, MIp and COBS we looked to measure this effect if any. First, we generated a graph for each algorithm, like what you would find in Figure 4.2. Next, we performed a Kendall rank correlation of the resulting graph for each family to produce a p-value. We then plotted that p-value versus the alignment quality, as judged by the Mumsa score [107]. As shown in Supplemental Fig. S4.1, the line of fit for such a plot would indicate no such correlation exists.

In total 1,116 PFAM families (Supplementary Table 1) were found that had PDB files that had a minimum of 7 secondary structure elements. Mapping of PDB files to PFAM was compiled using a BLAST search of the PFAM accession IDs and PDB sequence information, taking the best result as our reference sequence (RS).

For sampling size considerations, only families that had at least 200 sequences were considered. For performance considerations families with > 2000 sequences were removed from the data set. To ensure that we would have enough columns for analysis, MSA “width” was set to a floor of 80 columns. Finally, a minimum percentage identity score of the RS was set for 90% as compared to the sequence from the identified PDB.

In total we had 1,116 families in our final dataset with a total of 18,162 unique secondary structure elements. All scripts used to create figures in this paper can be found at https://github.com/afodor/cobs/.

In generating ROC curves, we simply took the scores for each PFAM family and aggregated them into one large spreadsheet sorted by score. This method has the disadvantage that the top hits that represent the initial set of predictions from the ROC curve may come from a disproportionally small number of PFAM families. An alternative would be to generate a separate ROC curve for each PFAM family and then produce an average ROC curve made up of each individual ROC curve. However, this procedure generated nearly identical ROC curves as simply taking absolute score (data not shown). In figures for this paper, therefore, we report ROCs based on absolute scores. Results for the alternative method can be generated by following the final step in the "Readme" instructions for the source code (https://github.com/afodor/cobs/).

4.5 Results

We defined a novel covariance method called COBS that works on contiguous groups of columns within a protein multiple sequence alignment (see methods). We evaluated the COBS algorithm on the proximity of secondary structure (α-helixes and β-strands) in 1,116 PFAM families (Supplementary Table 4.2). For each PFAM alignment, we asked in the corresponding structure how well the COBS algorithm could predict secondary structure elements that were in physical proximity. We compared the COBS algorithm to averaging results from the canonical covariance.

Fig. 4.2 shows the results of this comparison for the Bac_rhamnosid PFAM family (*PF05592*). There are 18 α-helixes and 17 β-strands in this family for a total for 35 secondary structure elements. For each of the 595 (or $\binom{35}{2}$) possible comparisons, we asked how well the scores from the variance covariance algorithms predict the average distance between all residues in these structures. As controls, we included the average conservation score for both columns as well as simply assigning a score from a random (uniform) distribution. Just by visual inspection, for this protein family the highest scoring COBS pair of GOCs (to the right on the x-axis) appear to have an average distance that is closer (to the bottom of the y-axis) than the highest scoring pair of GOCs chosen by average McBASC and average MI.

In order to gauge the performance of the algorithms across multiple PFAM families, we aggregated all predictions across 180,851 α-helixes and β-strands combinations from 1,116 PFAM families that met the criteria for inclusion in our study (see methods). We arbitrarily defined a success as a prediction in which the average distance between two secondary structure elements is less than the median distance of all secondary structure elements within the protein structure. We then ranked the predictions with the highest scoring prediction first. ROC curves based on these ranks are shown in Fig. 4.3A. As expected, an algorithm that chooses pairs of secondary structures at random falls on the identity line on the ROC curve (Fig. 4.3A black line). Average conservation (Fig. 4.3A, blue line) does little better than random, demonstrating that, unlike for pairs of columns [18], background conservation does not predict physically close secondary structures. Average MI (Fig. 4.3A, green line) and average McBASC (Fig 4.3A, yellow line) are also not much better than random but COBS (Fig. 4.3A, red line) displays a substantially improved performance.

### 4.6 Improved Prediction Accuracy Using Phylogeny Correction

We applied the phylogenetic correction term APC, introduced by Dunn [26] to the MI, McBASC, and COBS algorithms to produce algorithms called MIp, McBASCp, and COBSp (Fig 4.3B; see methods). The phylogenetic correction yielded a significant improvement in the performance of MIp (Fig 4.3B, green dashed line), McBASCp (Fig 4.3B, yellow dashed line) and COBSp (Fig. 4.3B, red dashed line) in predicting physically close secondary structures. When all corrected and uncorrected algorithms are compared, COBSp (Fig. 4.3C, red dashed line) clearly demonstrated the best performance among all the algorithms we tested.

The phylogenetic correction term is designed to eliminate "background" covariance due to non-random sampling across phylogenetic space in the multiple sequence alignment. Since we expect that most secondary structure elements within a protein will not covary, we would expect that after phylogenetic correction, the average covariance score for COBSp would be centered on zero, which would result from a Pearson correlation of unrelated vectors. Fig. 4.4 demonstrates that this expectation was realized for COBSp score providing further evidence that the simple phylogenetic correction terms is effective in reducing covariance introduced by phylogenetic artifacts.

## 4.7 Discussion

Using an algorithm based in part on average MI scores, Xu and Tillier [108] found that residues close to highly covarying residues also tended to be highly covarying. In their work, Xu and Tiller suggest a scoring scheme for a group of residues (what they term a "patch" and what we here call a GOC) based on the MI score for the pair of residues within the patch with the highest covariance score (what they term the "focal pair") divided by the average MI for the entire patch of continuous residues. Here, we suggest an alternative that computes covariance directly at the "patch" or GOC level without relying on average paired covariance. On a test set of α-helixes and β-strands derived from the PFAM database, our approach appears to have more power at detecting physically close sets of residues than methods that average over covariance scores derived from pairs of columns.

The dataset we used to test our algorithm, like recent work by Hopf *et al.* [95], focused solely on secondary structure covariance. It is easy to imagine future permutations that would extend COBS past α-helixes and β-strands. For example, a "greedy" algorithm could start with "focal pairs" of highly covarying columns and attempt to extend the region of significant covariance. Likewise, since Eq. 4.2 can be defined over any set of contiguous or non-contiguous columns, one can also imagine possible extensions that could apply COBS to non-contiguous columns to attempt to find a global network of covariance within each protein family. Such extensions, however, would require additional parameters to determine appropriate threshold cutoffs for when groups of covarying columns should be considered distinct clusters. Fitting these additional parameters would presumably require separating part of our data into a training set to estimate the parameters and a separate test set to evaluate performance. By pre-defining our GOCs as secondary structure elements whose composition is defined independent of any action of the algorithms, we have avoided the need for training and test sets, simplifying the interpretation of the relative power of the different algorithms that we tested.

While still modest in overall accuracy, our approach would appear to reveal regional patterns of covariance that are relatively unexplored by algorithms that focus on pairs of columns.  This approach may in the future have utility in assisting computational methods that discriminate likely and unlikely folds as well as methods that use sequence alignments to find functionally and structurally important regions in proteins [97, 109, 110].

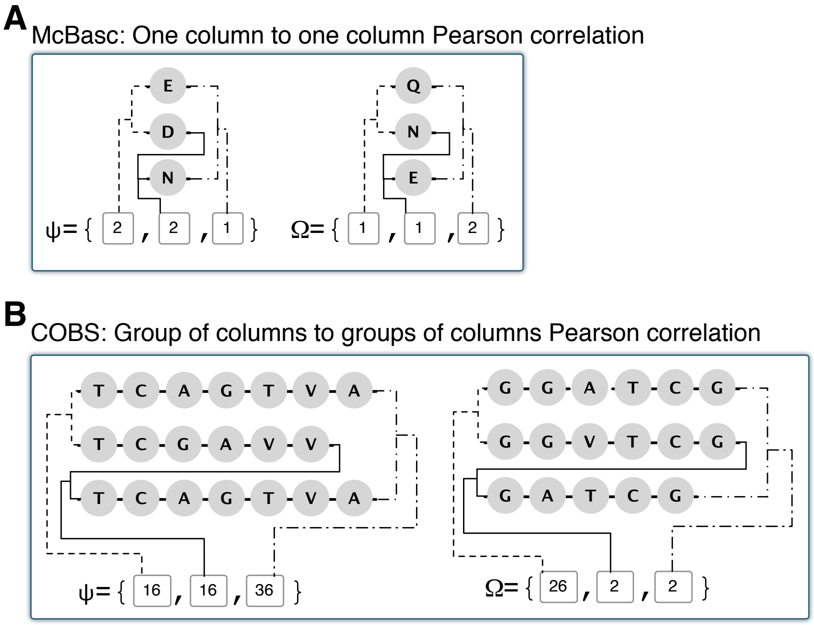

FIGURE 4.1: McBASC and COBS applied to simple alignments. (a) The McBASC algorithm applied to two columns from a multiple sequence alignment. The similarity of each pair of amino acids in each column is recorded using a McLachlan matrix. Each score of similarity from the McLachlan matrix is then added to a vector ψ (for the first column) and Ω (for the second column). (b) The COBS algorithm applied to two contiguous groups of columns within an alignment. The scores added to the vectors for each pair of sequences in the alignment is the sum of all substitutions from the McLachlan matrix.

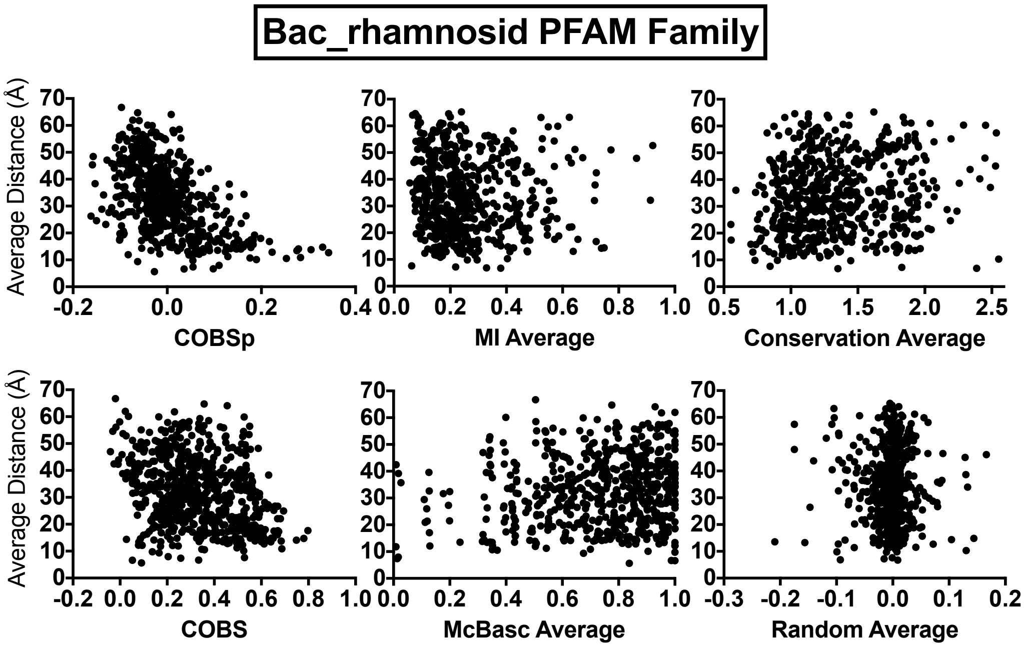

FIGURE 4.2: The performance of COBS on a single PFAM family. The bac_rhamnosid family has 943 sequences in the alignment. In each panel, the y-axis is the average distance between each pair of secondary structures in the corresponding pdb file (3cih). The x-axis is the score for the indicated algorithm. Random average is a score assigned at random from the uniform distribution. Conservation average refers to sequence entropy averages for the columns tested.

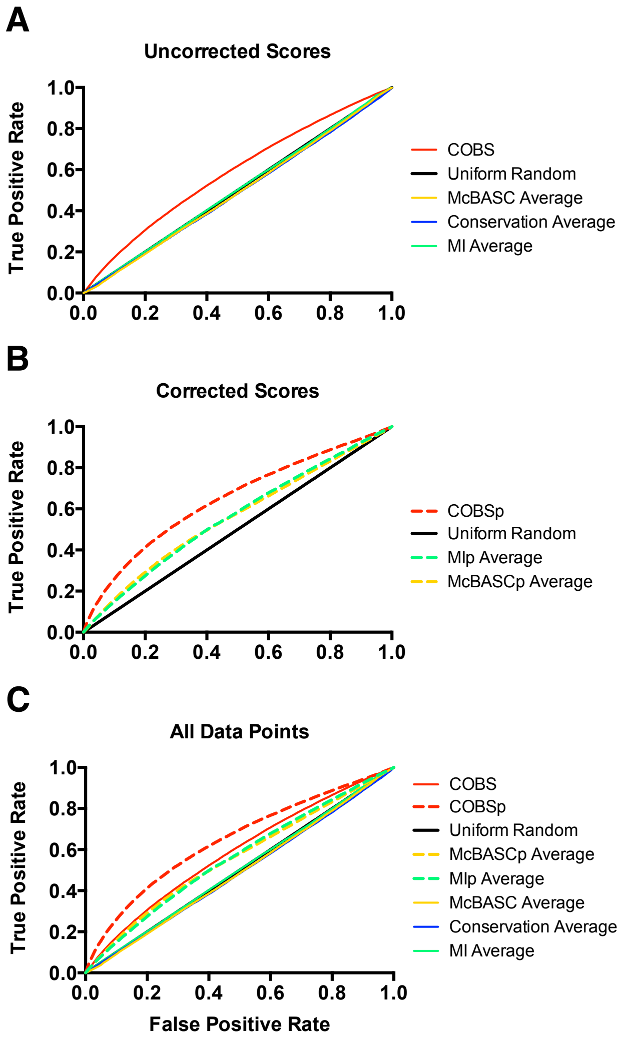

FIGURE 4.3: Receiver operating characteristic curves showing the relative performance for all algorithms. A true positive was defined as any distance that was less than the 50th percentile of the average distances of the secondary structures from each alignment. (A): Algorithms uncorrected for phylogenetic artifacts; (B): Algorithms with phylogenetic correction applied; (C): Superimposed receiver operating characteristic curves from corrected (dashed lines) and uncorrected (solid lines) algorithms.

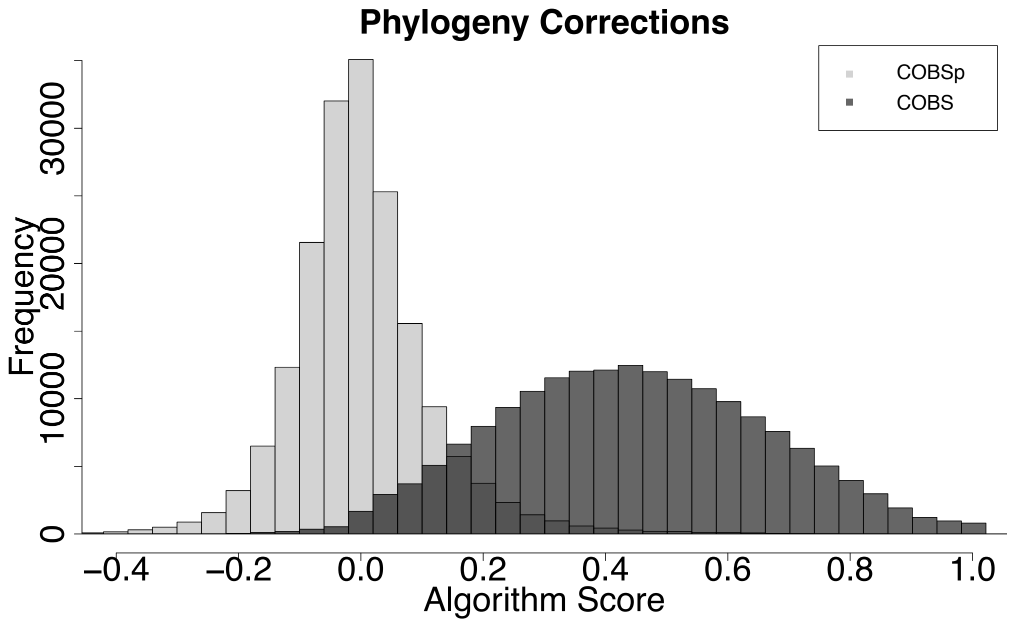

FIGURE 4.4: A histogram of the relative abundance of scores for COBS and COBSp, which is COBS with a phylogenetic correction. Prior to any phylogenic correction the distribution of scores has an apex at roughly 0.4. After applying the phylogenetic correction the distribution has reduced variance and a peak closer to zero indicating the success of the correction for background covariance introduced by alignment artifacts.

# CHAPTER 5: CONCLUSION

## 5.1 Summary of Results

The overall power of covariance algorithms is quite weak. From the amount of data that we have currently, it would appear that we lack the ability to leverage covariance for many desirable tasks such as folding or prediction of allostery. As outlined in Chapter One, allostery appears to be a particularly thorny issue as this phenomenon can exhibit itself with a wide variety of mechanisms even for similar homologs. Until such a time as algorithms are more powerful, or the sequence alignments can be pre-screened or filtered to map only one mechanism for allostery, there would appear to be no short term solution.

We can gain significant power both by accounting for phylogenetic noise and by grouping columns for analysis. Secondary structures were the easiest way to approach determining which groups of columns to use for analysis, and because of their known significance with regard to function and structure they were easy to justify using in this manner.

## 5.2 Future Direction

Grouping columns by secondary structure was a matter of convenience. Other ways of determining how to group columns have been explored, usually with an ‘optimal’ window size of grouping columns together being searched for or modeled separately. The constraints placed on protein evolution differ dramatically for proteins that bind to other proteins versus globular versus membrane bound proteins. Analyzing the constraints and attempting to maximize signal from three large groupings of these protein types may inform future approaches. As recently as this year, additional power has been leveraged with respect to decoy elimination. This feat was accomplished by specializing the problem to consider only membrane bound proteins.

Another direction to take this work is to revisit the question of rigid body analysis but with COBSp instead of MIp, effectively revisiting Chapter 3 with a more powerful covariance algorithm. Statistically, the correlation between covariance and rigidity was close to statistically significance using MIp. If COBSp has additional power to discriminate true positives of rigid body covariance, it may be that we can find a significant correlation. To that end, I have started a collaboration with K.C. Dukka at North Carolina A&T University to answer that question.

# REFERENCES

1. Livesay, D.R., K.E. Kreth, and A.A. Fodor, *A critical evaluation of correlated mutation algorithms and coevolution within allosteric mechanisms.* Methods Mol Biol, 2012. **796**: p. 385-98.

2. Horovitz, A., et al., *Prediction of an inter-residue interaction in the chaperonin GroEL from multiple sequence alignment is confirmed by double-mutant cycle analysis.* J Mol Biol, 1994. **238**(2): p. 133-8.

3. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-9.

4. Fodor, A.A. and R.W. Aldrich, *On evolutionary conservation of thermodynamic coupling in proteins.* J Biol Chem, 2004. **279**(18): p. 19046-50.

5. Chi, C.N., et al., *Reassessing a sparse energetic network within a single protein domain.* Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4679-84.

6. Liu, Z., J. Chen, and D. Thirumalai, *On the accuracy of inferring energetic coupling between distant sites in protein families from evolutionary imprints: illustrations using lattice model.* Proteins, 2009. **77**(4): p. 823-31.

7. Jensen, R.A. and S.L. Stenmark, *Comparative allostery of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthetase as a molecular basis for classification.* J Bacteriol, 1970. **101**(3): p. 763-9.

8. Jensen, A.A. and T.A. Spalding, *Allosteric modulation of G-protein coupled receptors.* Eur J Pharm Sci, 2004. **21**(4): p. 407-20.

9. May, L.T., et al., *Allosteric modulation of G protein-coupled receptors.* Curr Pharm Des, 2004. **10**(17): p. 2003-13.

10. Hudson, J.W., G.B. Golding, and M.M. Crerar, *Evolution of allosteric control in glycogen phosphorylase.* J Mol Biol, 1993. **234**(3): p. 700-21.

11. Royer, W.E., Jr., et al., *Cooperative hemoglobins: conserved fold, diverse quaternary assemblies and allosteric mechanisms.* Trends Biochem Sci, 2001. **26**(5): p. 297-304.

12. Royer, W.E., Jr., et al., *Allosteric hemoglobin assembly: diversity and similarity.* J Biol Chem, 2005. **280**(30): p. 27477-80.

13. Chakrabarti, S. and A.R. Panchenko, *Coevolution in defining the functional specificity.* Proteins, 2009. **75**(1): p. 231-40.

14. del Sol, A., et al., *The origin of allosteric functional modulation: multiple pre-existing pathways.* Structure, 2009. **17**(8): p. 1042-50.

15. Cui, Q. and M. Karplus, *Allostery and cooperativity revisited.* Protein Sci, 2008. **17**(8): p. 1295-307.

16. Formaneck, M.S., L. Ma, and Q. Cui, *Reconciling the "old" and "new" views of protein allostery: a molecular simulation study of chemotaxis Y protein (CheY).* Proteins, 2006. **63**(4): p. 846-67.

17. Ashkenazy, H. and Y. Kliger, *Reducing phylogenetic bias in correlated mutation analysis.* Protein Eng Des Sel. **23**(5): p. 321-6.

18. Fodor, A.A. and R.W. Aldrich, *Influence of conservation on calculations of amino acid covariance in multiple sequence alignments.* Proteins, 2004. **56**(2): p. 211-21.

19. Wollenberg, K.R. and W.R. Atchley, *Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap.* Proc Natl Acad Sci U S A, 2000. **97**(7): p. 3288-91.

20. Noivirt, O., M. Eisenstein, and A. Horovitz, *Detection and reduction of evolutionary noise in correlated mutation analysis.* Protein Eng Des Sel, 2005. **18**(5): p. 247-53.

21. Dimmic, M.W., et al., *Detecting coevolving amino acid sites using Bayesian mutational mapping.* Bioinformatics, 2005. **21 Suppl 1**: p. i126-35.

22. Dutheil, J., et al., *A model-based approach for detecting coevolving positions in a molecule.* Mol Biol Evol, 2005. **22**(9): p. 1919-28.

23. Ashkenazy, H., R. Unger, and Y. Kliger, *Optimal data collection for correlated mutation analysis.* Proteins, 2009. **74**(3): p. 545-55.

24. Vicatos, S., B.V. Reddy, and Y. Kaznessis, *Prediction of distant residue contacts with the use of evolutionary information.* Proteins, 2005. **58**(4): p. 935-49.

25. Kundrotas, P.J. and E.G. Alexov, *Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives.* BMC Bioinformatics, 2006. **7**: p. 503.

26. Dunn, S.D., L.M. Wahl, and G.B. Gloor, *Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.* Bioinformatics, 2008. **24**(3): p. 333-40.

27. Dickson, R.J., et al., *Identifying and Seeing beyond Multiple Sequence Alignment Errors Using Intra-Molecular Protein Covariation.* PLoS One, 2010. **5**(6): p. e11082.

28. Little, D.Y. and L. Chen, *Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution.* PLoS One, 2009. **4**(3): p. e4762.

29. Edgar, R.C. and S. Batzoglou, *Multiple sequence alignment.* Curr Opin Struct Biol, 2006. **16**(3): p. 368-73.

30. Martin, L.C., et al., *Using information theory to search for co-evolving residues in proteins.* Bioinformatics, 2005. **21**(22): p. 4116-24.

31. Weil, P., F. Hoffgaard, and K. Hamacher, *Estimating sufficient statistics in co-evolutionary analysis by mutual information.* Comput Biol Chem, 2009. **33**(6): p. 440-4.

32. Buslje, C.M., et al., *Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information.* Bioinformatics, 2009. **25**(9): p. 1125-31.

33. Fernandes, A.D. and G.B. Gloor, *Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself?* Bioinformatics. **26**(9): p. 1135-9.

34. Brown, C.A. and K.S. Brown, *Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my!* PLoS One. **5**(6): p. e10779.

35. Burger, L. and E. van Nimwegen, *Disentangling direct from indirect co-evolution of residues in protein alignments.* PLoS Comput Biol. **6**(1): p. e1000633.

36. Weigt, M., et al., *Identification of direct residue contacts in protein-protein interaction by message passing.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 67-72.

37. Gobel, U., et al., *Correlated mutations and residue contacts in proteins.* Proteins, 1994. **18**(4): p. 309-17.

38. Clarke, N.D., *Covariation of residues in the homeodomain sequence family.* Protein Sci, 1995. **4**(11): p. 2269-78.

39. Mildvan, A.S., D.J. Weber, and A. Kuliopulos, *Quantitative interpretations of double mutations of enzymes.* Arch Biochem Biophys, 1992. **294**(2): p. 327-40.

40. Gloor, G.B., et al., *Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions.* Biochemistry, 2005. **44**(19): p. 7156-65.

41. Istomin, A.Y., et al., *New insight into long-range nonadditivity within protein double-mutant cycles.* Proteins, 2008. **70**(3): p. 915-24.

42. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Biol, 2003. **10**(1): p. 59-69.

43. Halabi, N., et al., *Protein sectors: evolutionary units of three-dimensional structure.* Cell, 2009. **138**(4): p. 774-86.

44. Kuriyan, J. and D. Eisenberg, *The origin of protein interactions and allostery in colocalization.* Nature, 2007. **450**(7172): p. 983-90.

45. Fenton, A.W., *Allostery: an illustrated definition for the 'second secret of life'.* Trends Biochem Sci, 2008. **33**(9): p. 420-5.

46. Koshland, D.E., *Application of a Theory of Enzyme Specificity to Protein Synthesis.* Proc Natl Acad Sci U S A, 1958. **44**(2): p. 98-104.

47. Yu, E.W. and D.E. Koshland, Jr., *Propagating conformational changes over long (and short) distances in proteins.* Proc Natl Acad Sci U S A, 2001. **98**(17): p. 9517-20.

48. Ottemann, K.M., et al., *A piston model for transmembrane signaling of the aspartate receptor.* Science, 1999. **285**(5434): p. 1751-4.

49. Swain, J.F. and L.M. Gierasch, *The changing landscape of protein allostery.* Curr Opin Struct Biol, 2006. **16**(1): p. 102-8.

50. Kumar, S., et al., *Folding and binding cascades: dynamic landscapes and population shifts.* Protein Sci, 2000. **9**(1): p. 10-9.

51. Monod, J., J. Wyman, and J.P. Changeux, *On the Nature of Allosteric Transitions: a Plausible Model.* J Mol Biol, 1965. **12**: p. 88-118.

52. Gunasekaran, K., B. Ma, and R. Nussinov, *Is allostery an intrinsic property of all dynamic proteins?* Proteins, 2004. **57**(3): p. 433-43.

53. Bruschweiler, S., et al., *Direct observation of the dynamic process underlying allosteric signal transmission.* J Am Chem Soc, 2009. **131**(8): p. 3063-8.

54. Schlegel, J., et al., *Characterizing and controlling the inherent dynamics of cyclophilin-A.* Protein Sci, 2009. **18**(4): p. 811-24.

55. Lee, A.L., S.A. Kinnear, and A.J. Wand, *Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex.* Nat Struct Biol, 2000. **7**(1): p. 72-7.

56. Mau, T., J.D. Baleja, and G. Wagner, *Effects of DNA binding and metal substitution on the dynamics of the GAL4 DNA-binding domain as studied by amide proton exchange.* Protein Sci, 1992. **1**(11): p. 1403-12.

57. Forman, B.M., et al., *Unique response pathways are established by allosteric interactions among nuclear hormone receptors.* Cell, 1995. **81**(4): p. 541-50.

58. Conigrave, A.D. and A.H. Franks, *Allosteric activation of plasma membrane receptors--physiological implications and structural origins.* Prog Biophys Mol Biol, 2003. **81**(3): p. 219-40.

59. Hardy, J.A. and J.A. Wells, *Searching for new allosteric sites in enzymes.* Curr Opin Struct Biol, 2004. **14**(6): p. 706-15.

60. Fodor, A.A., K.D. Black, and W.N. Zagotta, *Tetracaine reports a conformational change in the pore of cyclic nucleotide-gated channels.* J Gen Physiol, 1997. **110**(5): p. 591-600.

61. Mottonen, J.M., D.J. Jacobs, and D.R. Livesay, *Allosteric Response is Both Conserved and Variable Across Three CheY Orthologs.* p. In revision.

62. Whitaker, R.J., et al., *Comparative allostery of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase as an indicator of taxonomic relatedness in pseudomonad genera.* J Bacteriol, 1981. **145**(2): p. 752-9.

63. Jensen, R.A. and R. Twarog, *Allostery of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase in Clostridium: another conserved generic characteristic.* J Bacteriol, 1972. **111**(3): p. 641-8.

64. Fuchs, A., et al., *Co-evolving residues in membrane proteins.* Bioinformatics, 2007. **23**(24): p. 3312-9.

65. Gouveia-Oliveira, R. and A.G. Pedersen, *Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation.* Algorithms Mol Biol, 2007. **2**: p. 12.

66. Suarez-Diaz, E. and V.H. Anaya-Munoz, *History, objectivity, and the construction of molecular phylogenies.* Stud Hist Philos Biol Biomed Sci, 2008. **39**(4): p. 451-68.

67. Wang, H.C., et al., *Testing for covarion-like evolution in protein sequences.* Mol Biol Evol, 2007. **24**(1): p. 294-305.

68. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.

69. Durbin, R., *Biological sequence analysis : probabalistic models of proteins and nucleic acids*. 1998, Cambridge, UK New York: Cambridge University Press. xi, 356 p.

70. Cover, T.M. and J.A. Thomas, *Elements of information theory*. 2nd ed. 2006, Hoboken, N.J.: Wiley-Interscience. xxiii, 748 p.

71. Shenkin, P.S., B. Erman, and L.D. Mastrandrea, *Information-theoretical entropy as a measure of sequence variability.* Proteins, 1991. **11**(4): p. 297-313.

72. Dekker, J.P., et al., *A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments.* Bioinformatics, 2004. **20**(10): p. 1565-72.

73. Olmea, O., B. Rost, and A. Valencia, *Effective use of sequence correlation and conservation in fold recognition.* J Mol Biol, 1999. **293**(5): p. 1221-39.

74. McLachlan, A.D., *Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551.* J Mol Biol, 1971. **61**(2): p. 409-24.

75. Ashkenazy, H. and Y. Kliger, *Reducing phylogenetic bias in correlated mutation analysis.* Protein Eng Des Sel, 2010. **23**(5): p. 321-6.

76. Chou, K.C. and C.T. Zhang, *Predicting protein folding types by distance functions that make allowances for amino acid interactions.* J Biol Chem, 1994. **269**(35): p. 22014-20.

77. Miller, M.A. and D. Frenkel, *Simulating colloids with Baxter's adhesive hard sphere model.* Journal of Physics: Condensed Matter, 2004. **16**(42): p. S4901.

78. Dumitraş, M. and C. Friedrich, *Network formation and elasticity evolution in dibenzylidene sorbitol/poly (propylene oxide) physical gels.* Journal of Rheology, 2004. **48**: p. 1135.

79. Gohlke, H., L.A. Kuhn, and D.A. Case, *Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach.* Proteins, 2004. **56**(2): p. 322-37.

80. Maxwell, J.C., *L. on the calculation of the equilibrium and stiffness of frames.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1864. **27**(182): p. 294-299.

81. Phillips, J.C. and M.F. Thorpe, *Phase transitions and self-organization in electronic and molecular networks*. 2001: Springer.

82. Jacobs, D.J. and M.F. Thorpe, *Generic rigidity percolation: The pebble game.* Phys Rev Lett, 1995. **75**(22): p. 4051-4054.

83. Bertini, I., et al., *Experimentally exploring the conformational space sampled by domain reorientation in calmodulin.* Proc Natl Acad Sci U S A, 2004. **101**(18): p. 6841-6.

84. Tsai, C.J., et al., *Folding funnels, binding funnels, and protein function.* Protein Sci, 1999. **8**(6): p. 1181-90.

85. Rader, A.J., et al., *Protein unfolding: rigidity lost.* Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3540-5.

86. Gloor, G.B., et al., *Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions&#x2020.* Biochemistry, 2005. **44**(19): p. 7156-7165.

87. Ortiz, A.R., A. Kolinski, and J. Skolnick, *Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations.* Proc Natl Acad Sci U S A, 1998. **95**(3): p. 1020-5.

88. Davis, I.W., et al., *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W375-83.

89. Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 1.3r1*, 2010.

90. Verma, D., D.J. Jacobs, and D.R. Livesay, *Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged.* PLoS Comput Biol, 2012. **8**(3): p. e1002409.

91. Daily, M.D. and J.J. Gray, *Allosteric communication occurs via networks of tertiary and quaternary motions in proteins.* PLoS Comput Biol, 2009. **5**(2): p. e1000293.

92. Bradley, P. and D. Baker, *Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation.* Proteins, 2006. **65**(4): p. 922-9.

93. Altschuh, D., et al., *Coordinated amino acid changes in homologous protein families.* Protein Eng, 1988. **2**(3): p. 193-9.

94. Dickson, R.J. and G.B. Gloor, *Protein sequence alignment analysis by local covariation: coevolution statistics detect benchmark alignment errors.* PLoS One, 2012. **7**(6): p. e37645.

95. Hopf, T.A., et al., *Three-dimensional structures of membrane proteins from genomic sequencing.* Cell, 2012. **149**(7): p. 1607-21.

96. Kass, I. and A. Horovitz, *Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations.* Proteins, 2002. **48**(4): p. 611-7.

97. Jones, D.T., et al., *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.* Bioinformatics, 2012. **28**(2): p. 184-90.

98. Cheng, J. and P. Baldi, *Improved residue contact prediction using support vector machines and a large feature set.* BMC Bioinformatics, 2007. **8**: p. 113.

99. Eyal, E., et al., *A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction.* Proteins, 2007. **67**(1): p. 142-53.

100. Burger, L. and E. van Nimwegen, *Disentangling direct from indirect co-evolution of residues in protein alignments.* PLoS Comput Biol, 2010. **6**(1): p. e1000633.

101. Ouzounis, C., et al., *Are binding residues conserved?* Pac Symp Biocomput, 1998: p. 401-12.

102. Hu, Z., et al., *Conservation of polar residues as hot spots at protein interfaces.* Proteins, 2000. **39**(4): p. 331-42.

103. Aloy, P., et al., *Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking.* J Mol Biol, 2001. **311**(2): p. 395-408.

104. Hoberman, R., J. Klein-Seetharaman, and R. Rosenfeld, *Inferring property selection pressure from positional residue conservation.* Appl Bioinformatics, 2004. **3**(2-3): p. 167-79.

105. Savojardo, C., et al., *Prediction of disulfide connectivity in proteins with machine-learning methods and correlated mutations.* BMC Bioinformatics, 2013. **14 Suppl 1**: p. S10.

106. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.

107. Lassmann, T. and E.L. Sonnhammer, *Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W596-9.

108. Xu, Y. and E.R. Tillier, *Regional covariation and its application for predicting protein contact patches.* Proteins, 2010. **78**(3): p. 548-58.

109. Aguilar, D., B. Oliva, and C. Marino Buslje, *Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features.* PLoS One, 2012. **7**(7): p. e41430.

110. Othersen, O.G., et al., *Application of information theory to feature selection in protein docking.* J Mol Model, 2012. **18**(4): p. 1285-97.

# APPENDIX A: SUPPLEMENTAL MATERIALS

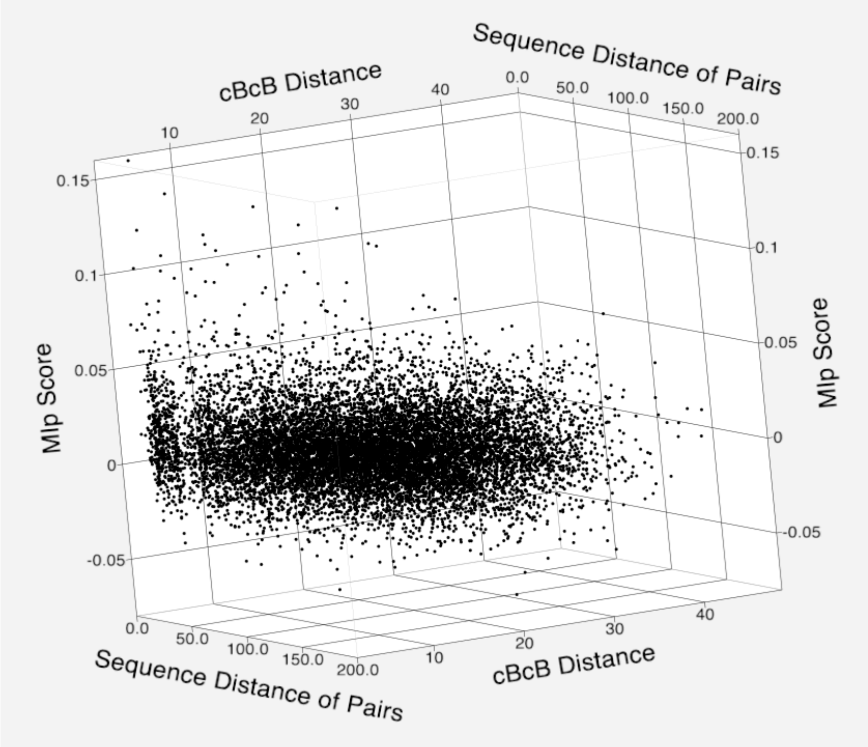

FIGURE S3.1: Represented here is the 2axn protein, which was the best sequence match for the PFAM accession number pfam01591. This view is only for those sequence pairs where the pairs of columns being analyzed are in the same rigid body. You can notice the over abundance in the situation of very low "sequence distance of pairs", which will also unsurprisingly also have very low β-carbon to β-carbon Cartesian distances. In this view the outlier top and bottom 5% of MIp scores are removed to better view the data in this format.

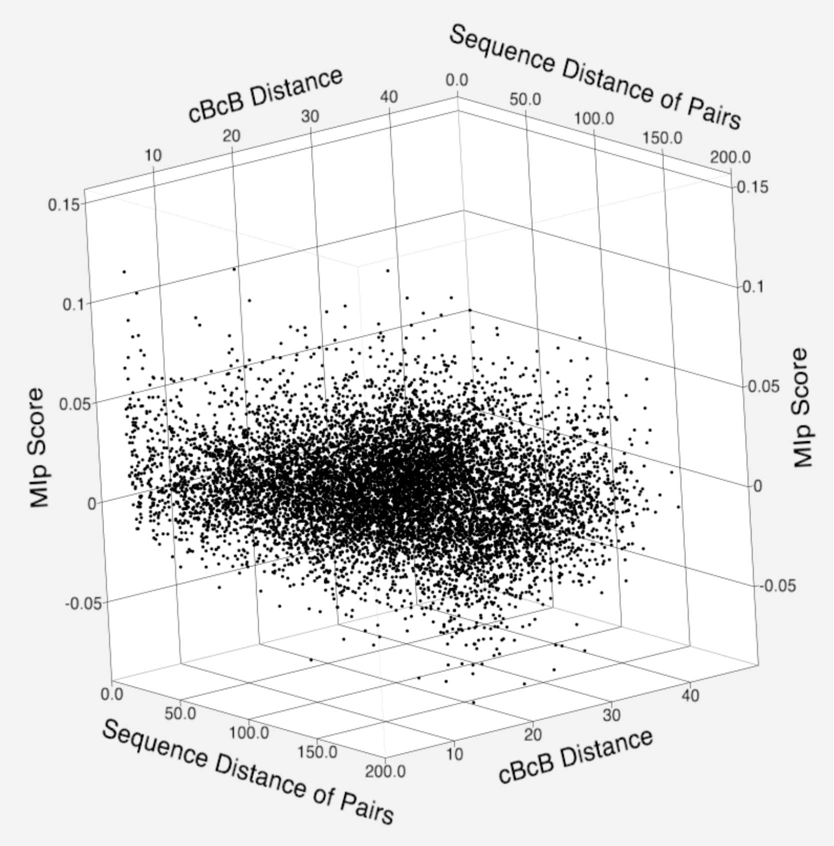

FIGURE S3.2: Represented here is the 2axn protein, which was the best sequence match for the PFAM accession number pfam01591. This view is only for those sequence pairs where the pairs of columns being analyzed are in separate rigid bodies. Here can note the most typical result for MIp where (excepting the very best scores), the overall correlation between distance and MIp score is weak. In this view the outlier top and bottom 5% of MIp scores are removed to better view the data in this format.

APPENDIX A: (CONTINUED)

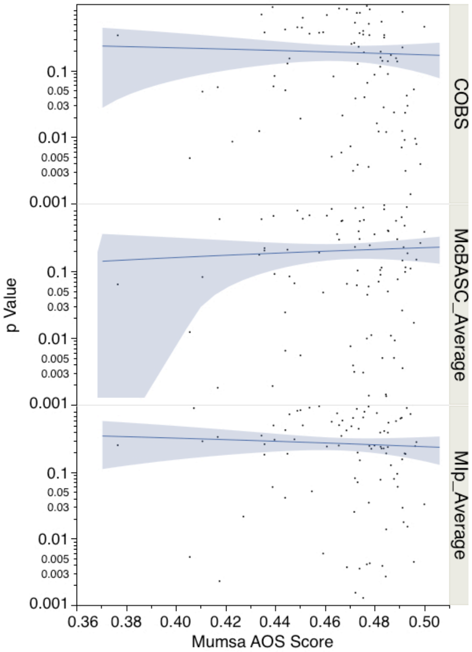

FIGURE S4.1: This graph was generated to view the potential conflation of alignment quality and algorithm success. Here, the p-value is expressed in log space as the Kendall rank correlation of a plot of algorithm score and distance of all atom to all atom residue distances. Shown on the right Y-axis are results for the three best performing algorithms. PFAM families have a relatively narrow band of Mumsa average overlap scores (AOS). PFAM families were re-aligned with Muscle 8.31, but the results were nearly identical (data not shown). The first 105 families (alphabetically) from Table Supplemental 1 were used to generate the p-values and Mumsa scores.

APPENDIX A: (CONTINUED)

| PFAM Family List A | PFAM Family List B |
|---|---|
| 14-3-3 | GshA |
| 2-ph_phosp | Gtr1_RagA |
| 2OG-FeII_Oxy_5 | HD_2 |
| 2_5_RNA_ligase2 | HD_3 |
| 3-HAO | HGD-D |
| 3-dmu-9_3-mt | HK |
| 4HBT_3 | HMG_CoA_synt_C |
| 5_nucleotid | HMG_CoA_synt_N |
| 6PF2K | HNOB |
| 7TMR-DISMED2 | HOOK |
| A2M_N_2 | HORMA |
| A2M_comp | HSF_DNA-bind |
| AAA_18 | HTH_WhiA |
| AAA_28 | HTS |
| AAA_32 | H_PPase |
| AAA_4 | Haemagg_act |
| AAA_8 | Hat1_N |
| AAL_decarboxy | Helicase_RecD |
| AAR2 | HemS |
| AAT | Hema_HEFG |
| ABC_sub_bind | Hema_esterase |
| ACOX | Heme_oxygenase |
| ADC | Hemocyanin_C |
| ADP_ribosyl_GH | Hemocyanin_M |
| AFOR_C | Hemocyanin_N |
| AFOR_N | Hepar_II_III |
| AHS1 | Herpes_BLLF1 |
| AHS2 | Herpes_TK |
| AIG1 | Herpes_glycop_D |
| AIG2 | Herpes_glycop_H |
| AKAP7_NLS | Hexokinase_1 |
| ALO | Hexokinase_2 |
| AMMECR1 | HgmA |
| AMNp_N | Hom_end_hint |
| AMP_N | HpaB |
| ANTH | HpaB_N |
| APC10 | Hpr_kinase_N |

| | |
|---|---|
| APG5 | HrcA |
| APOBEC_N | Hus1 |
| ARD | HutD |
| ARPC4 | HutP |
| ART | HyaE |
| ASC | HycI |
| ASF1_hist_chap | Hydantoinase_A |
| ASL_C | HypA |
| ATP-grasp_3 | HypD |
| ATP-grasp_5 | IDH |
| ATP-sulfurylase | IDO |
| ATP_bind_1 | IF4E |
| ATP_bind_4 | IIGP |
| AXE1 | IL1 |
| AceK | IP_trans |
| AcetDehyd-dimer | IalB |
| AcetylCoA_hyd_C | ImpE |
| AcetylCoA_hydro | Indigoidine_A |
| Acetyltransf_2 | Ins134_P3_kin |
| Acid_PPase | Ins_P5_2-kin |
| Acid_phosphat_B | Integrin_alpha2 |
| Aconitase_2_N | Integrin_beta |
| Aconitase_B_N | Intimin_C |
| Acyl-ACP_TE | Iron_transport |
| Acyl_CoA_thio | IucA_IucC |
| Adap_comp_sub | Ivy |
| Adenine_deam_C | JAB |
| Adeno_hexon_C | JHBP |
| Adenosine_kin | JmjC |
| AdoHcyase | Josephin |
| AdoHcyase_NAD | KAT11 |
| AdoMet_dc | K_oxygenase |
| Aerolysin | KaiC |
| Aha1_N | KdgM |
| Alginate_lyase2 | KdpD |
| Alginate_lyase | KduI |
| AlkA_N | Kin17_mid |
| Allantoicase | Kunitz_legume |
| Alpha-L-AF_C | LANC_like |
| Alpha-amylase_N | LBP_BPI_CETP_C |
| Alpha_E1_glycop | LBP_BPI_CETP |

| | |
|---|---|
| Alpha_E2_glycop | LIP |
| Alpha_L_fucos | LRAT |
| Alum_res | LacY_symp |
| Aminotran_MocR | Lact-deh-memb |
| An_peroxidase | Lact_bio_phlase |
| Antibiotic_NAT | LamB |
| Arabinose_Iso_C | LamB_YcsF |
| Arabinose_Isome | Ldh_2 |
| Arch_ATPase | Lectin_leg-like |
| Archease | Lectin_legB |
| ArdA | Leu_Phe_trans |
| Arena_RNA_pol | Leuk-A4-hydro_C |
| Arena_nucleocap | Leukocidin |
| ArfGap | Linocin_M18 |
| ArgJ | Lipase_2 |
| ArgK | Lipase |
| Arrestin_C | Lipase_chap |
| Arrestin_N | Lipocalin_2 |
| Arylsulfotrans | Lipocalin |
| AsnA | Lipoprot_C |
| Asp_decarbox | Lipoprotein_1 |
| AstA | Lipoprotein_2 |
| AstB | Lipoxygenase |
| AstE_AspA | LolA |
| Astacin | LolB |
| Atg8 | LptC |
| AurF | LpxC |
| Autoind_bind | Lumazine_bd_2 |
| Autoind_synth | Lyase_8 |
| B56 | Lyase_8_N |
| BAAT_C | Lys |
| BAR_3_WASP_bdg | M16C_assoc |
| BCDHK_Adom3 | M60-like |
| BNR_2 | MAGE |
| BNR_3 | MAM |
| BRO1 | MCR_alpha_N |
| BTAD | MDMPI_N |
| BTG | META |
| BTLCP | MIF4G_like |
| Bac_globin | MIF |
| Bac_rhamnosid | MOFRL |

| | |
|---|---|
| Bac_rhamnosid_N | MOSC_N |
| Band_3_cyto | MRJP |
| Bet_v_1 | MTS_N |
| Beta-Casp | MTTB |
| BetaGal_dom4_5 | MT |
| Bgal_small_N | Malate_synthase |
| Bile_Hydr_Trans | Malectin |
| Biopterin_H | MdoG |
| Birna_RdRp | Memo |
| Birna_VP2 | Meth_synt_1 |
| Branch | Methyltransf_10 |
| Brix | Methyltransf_14 |
| BsmA | Methyltransf_19 |
| C4 | Methyltransf_28 |
| C4dic_mal_tran | Methyltransf_30 |
| CAF1 | Methyltransf_7 |
| CARDB | Methyltransf_PK |
| CAS_CSE1 | Methyltrn_RNA_3 |
| CAT | Mfa2 |
| CBAH | MinC_C |
| CBM_21 | MipZ |
| CBM_4_9 | MlrC_C |
| CCP_MauG | MltA |
| CDC27 | MmgE_PrpD |
| CDH | Mn_catalase |
| CDO_I | Mo25 |
| CGI-121 | Mob1_phocein |
| CHB_HEX | MobA_MobL |
| CHMI | MobB |
| CHU_C | Motile_Sperm |
| CK_II_beta | MtfA |
| CM_1 | MtlR |
| COX4 | MucB_RseB |
| COXG | MukB |
| CP_ATPgrasp_1 | Multi-haem_cyto |
| CRISPR_Cse1 | Myotub-related |
| CRISPR_Cse2 | NAD_binding_5 |
| CRM1_C | NAGLU_C |
| CaMKII_AD | NAGLU |
| Calreticulin | NAGidase |
| CamS | NAM |

| | |
|---|---|
| Caps_synth_GfcC | NEAT |
| Capsid_NCLDV | NHase_alpha |
| Carb_anhydrase | NHase_beta |
| Carn_acyltransf | NIT |
| Cas_Cas6 | NMT_C |
| Cas_DxTHG | NO_synthase |
| CbiC | NPP1 |
| CbiD | NSP10 |
| CbiG_C | NSP13 |
| CbiK | NT5C |
| CdhD | NTF2 |
| Ceramidase_alk | NTPase_1 |
| Chalcone | NTPase_I-T |
| Channel_Tsx | NUDIX_2 |
| CheD | NYN |
| Chitin_bind_3 | Na_H_antiport_1 |
| ChitinaseA_N | Na_K-ATPase |
| Chlor_dismutase | NadA |
| Chor_lyase | Nairo_nucleo |
| ChuX_HutX | NanE |
| Circo_capsid | Ndr |
| CitF | Nepo_coat_C |
| CitG | Nepo_coat |
| Clp1 | NeuB |
| CmcH_NodU | NigD |
| CmcI | NinB |
| Coagulase | Nitrate_red_gam |
| Coatomer_E | Nol1_Nop2_Fmu_2 |
| Coatomer_WDAD | NosL |
| CobA_CobO_BtuR | Nuc_deoxyrib_tr |
| CobU | Nucleoplasmin |
| CodY | Nucleopor_Nup85 |
| Cofilin_ADF | Nucleoporin_N |
| Colicin-DNase | Nuf2 |
| CopC | Nup160 |
| Coq4 | Nup84_Nup100 |
| Creatininase | NurA |
| Crl | O-FucT |
| CrtC | OAS1_C |
| Cse1 | OCD_Mu_crystall |
| CsiD | OHCU_decarbox |

| | |
|---|---|
| CtaG_Cox11 | OKR_DC_1_N |
| Cu2_monoox_C | Octopine_DH |
| Cu2_monooxygen | OmpA_membrane |
| Cu_amine_oxidN3 | OmpH |
| Cu_amine_oxid | OmpW |
| Cucumo_coat | Omptin |
| Cullin | Opacity |
| Cullin_binding | Oxysterol_BP |
| Cupin_5 | P16-Arc |
| CutC | P21-Arc |
| Cutinase | P2X_receptor |
| Cyclase | P34-Arc |
| Cytochrom_C552 | PA14 |
| Cytochrome_C554 | PAC2 |
| DAHP_synth_2 | PAD_porph |
| DBI_PRT | PAF-AH_p_II |
| DCD | PAP_central |
| DCP1 | PAS_2 |
| DDR | PBP_like |
| DENN | PCI_Csn8 |
| DGOK | PCNA_C |
| DHHA2 | PCNA_N |
| DHquinase_I | PDEase_I |
| DIM1 | PEPCK_ATP |
| DNA_PPF | PEPCK |
| DNA_ligase_A_C | PHF5 |
| DNA_ligase_A_N | PI-PLC-X |
| DNA_pol3_chi | PI31_Prot_N |
| DNA_pol3_tau_5 | PI3K_C2 |
| DNA_pol_E_B | PI3K_rbd |
| DNA_primase_S | PI3Ka |
| DNase-RNase | PID |
| DOPA_dioxygen | PITH |
| DOT1 | PLA1 |
| DPPIV_N | PLAT |
| DPRP | PMEI |
| DSHCT | PMM |
| DS | PPV_E2_N |
| DUF1028 | PRiA4_ORF3 |
| DUF1034 | PTB |
| DUF1054 | PTE |

| | |
|---|---|
| DUF1080 | PTH2 |
| DUF1100 | PTPA |
| DUF1116 | PTSIIA_gutA |
| DUF111 | PTS_2-RNA |
| DUF1131 | PUA_2 |
| DUF1149 | P_gingi_FimA |
| DUF1217 | PaaA_PaaC |
| DUF1223 | PaaX_C |
| DUF1237 | PagL |
| DUF1273 | PagP |
| DUF1285 | Palm_thioest |
| DUF1307 | Pan_kinase |
| DUF1338 | ParBc_2 |
| DUF1341 | Parvo_NS1 |
| DUF1342 | PdxA |
| DUF1348 | PdxJ |
| DUF1349 | Pec_lyase_C |
| DUF1355 | Pectate_lyase |
| DUF1396 | Pectinesterase |
| DUF1398 | Pencillinase_R |
| DUF1439 | Penicil_amidase |
| DUF1445 | Pentaxin |
| DUF1460 | PepX_C |
| DUF1470 | PepX_N |
| DUF1479 | Peptidase_C12 |
| DUF1485 | Peptidase_C15 |
| DUF1498 | Peptidase_C28 |
| DUF1537 | Peptidase_C2 |
| DUF159 | Peptidase_C30 |
| DUF1611 | Peptidase_C39_2 |
| DUF1681 | Peptidase_C4 |
| DUF1684 | Peptidase_C65 |
| DUF1694 | Peptidase_C78 |
| DUF1696 | Peptidase_C80 |
| DUF1697 | Peptidase_M10_C |
| DUF1768 | Peptidase_M15_3 |
| DUF1775 | Peptidase_M15 |
| DUF1794 | Peptidase_M19 |
| DUF1795 | Peptidase_M27 |
| DUF179 | Peptidase_M29 |
| DUF1810 | Peptidase_M2 |

| | |
|---|---|
| DUF1811 | Peptidase_M32 |
| DUF1831 | Peptidase_M35 |
| DUF1846 | Peptidase_M43 |
| DUF1870 | Peptidase_M49 |
| DUF1877 | Peptidase_M4_C |
| DUF1896 | Peptidase_M4 |
| DUF1900 | Peptidase_M55 |
| DUF1906 | Peptidase_M75 |
| DUF1934 | Peptidase_M8 |
| DUF1935 | Peptidase_M9 |
| DUF1963 | Peptidase_S13 |
| DUF1969 | Peptidase_S15 |
| DUF1989 | Peptidase_S28 |
| DUF1993 | Peptidase_S32 |
| DUF2000 | Peptidase_S51 |
| DUF2002 | Peptidase_S58 |
| DUF2064 | Peptidase_S66 |
| DUF2077 | Peptidase_S6 |
| DUF2088 | Peptidase_S9_N |
| DUF2094 | Peripla_BP_5 |
| DUF2156 | Peroxidase_2 |
| DUF2200 | Pertactin |
| DUF2219 | PgpA |
| DUF2233 | Phage_cap_E |
| DUF2237 | Phage_lysozyme |
| DUF2263 | Phage_sheath_1 |
| DUF2267 | Phage_tail_U |
| DUF2479 | Phe_hydrox_dim |
| DUF2507 | Phenol_Hydrox |
| DUF2520 | PhnH |
| DUF2529 | PhoD |
| DUF255 | PhoQ_Sensor |
| DUF2886 | PhoU_div |
| DUF297 | Phosducin |
| DUF3013 | PhosphMutase |
| DUF303 | Phosphoesterase |
| DUF3048 | Phospholip_A2_2 |
| DUF3108 | Phospholip_B |
| DUF3168 | Phytase |
| DUF3231 | Phytochelatin |
| DUF3237 | Pico_P2A |

| | |
|---|---|
| DUF3251 | PilS |
| DUF330 | Pilin |
| DUF3327 | Pirin_C |
| DUF336 | Polysacc_deac_2 |
| DUF3372 | Porin_3 |
| DUF3416 | Porin_O_P |
| DUF3453 | Prim-Pol |
| DUF3457 | Pro-kuma_activ |
| DUF3458 | ProQ |
| DUF3478 | Pro_racemase |
| DUF364 | Profilin |
| DUF3749 | PrpF |
| DUF377 | PrpR_N |
| DUF3799 | Pyrid_oxidase_2 |
| DUF3829 | Pyridox_ox_2 |
| DUF3857 | RINT1_TIP1 |
| DUF385 | RIP |
| DUF3862 | RNA_lig_T4_1 |
| DUF386 | RNA_ligase |
| DUF3872 | RNA_pol_Rpb1_7 |
| DUF3988 | RNA_replicase_B |
| DUF399 | RNase_P_p30 |
| DUF4038 | RPE65 |
| DUF410 | RTC |
| DUF4136 | RTC_insert |
| DUF4147 | Rad1 |
| DUF416 | Rad4 |
| DUF436 | Rad51 |
| DUF442 | Rad52_Rad22 |
| DUF461 | Rad9 |
| DUF480 | RapA_C |
| DUF489 | Rap_GAP |
| DUF498 | RasGEF |
| DUF519 | RbsD_FucU |
| DUF520 | Rcd1 |
| DUF538 | RdRP_4 |
| DUF54 | RdRP |
| DUF576 | RdgC |
| DUF600 | Recep_L_domain |
| DUF615 | Regulator_TrmB |
| DUF619 | Rep_fac-A_3 |

| | |
|---|---|
| DUF633 | RhaA |
| DUF706 | Rieske_2 |
| DUF718 | Ring_hydroxyl_B |
| DUF72 | RnaseA |
| DUF830 | Rota_Capsid_VP6 |
| DUF849 | Rota_NS35 |
| DUF862 | Rota_VP2 |
| DUF866 | RusA |
| DUF86 | S1-P1_nuclease |
| DUF871 | S6PP |
| DUF885 | SAF_2 |
| DUF89 | SAG |
| DUF915 | SAM_MT |
| DUF917 | SAM_adeno_trans |
| DUF91 | SAM_decarbox |
| DUF924 | SBP56 |
| DUF925 | SCP2_2 |
| DUF961 | SIP |
| Dak1 | SMI1_KNR4 |
| DcpS_C | SNO |
| Ded_cyto | SOR_SNZ |
| Dehydratase_LU | SOUL |
| Dehydratase_MU | START |
| Dehydratase_SU | STAT_bind |
| Desulfoferrodox | STAT_int |
| DevR | STT3 |
| DinB | SUFU |
| Diphthamide_syn | SdiA-regulated |
| DisA-linker | Sec15 |
| DisA_N | Sec1 |
| Disulph_isomer | Sec23_trunk |
| DltD_C | Sec7 |
| DltD_M | Sedlin_N |
| Drf_FH3 | Sema |
| DsbC | Septin |
| DsrC | Ser_hydrolase |
| Dynamin_M | ShlB |
| Dyp_perox | Sif |
| E1_DerP2_DerF2 | Sina |
| E6 | Sipho_tail |
| EAP30 | SnoaL_4 |

| | |
|---|---|
| EHN | SnoaL |
| EIF_2_alpha | SoxG |
| EMG1 | SoxY |
| ENTH | SpoVAD |
| ERM | Spond_N |
| ERO1 | Spore_GerAC |
| EST1_DNA_bind | SsgA |
| ETF_QO | SspB |
| EVE | Ssu72 |
| Ecotin | Stap_Strp_tox_C |
| Endonuclease_1 | StbA |
| Endonuclease_5 | Stress-antifung |
| Endonuclease_NS | Sucrose_synth |
| Endotoxin_C | SufE |
| Endotoxin_M | Sulfotransfer_3 |
| Endotoxin_N | SusD-like |
| Enoyl_reductase | Sybindin |
| Ephrin | Syd |
| Ephrin_lbd | Syja_N |
| EpoR_lig-bind | T2SJ |
| Erythro_esteras | T2SK |
| EutB | T2SL |
| EutC | T4SS |
| Exo70 | T6SS-SciN |
| Exonuc_X-T_C | TCTP |
| F-actin_cap_A | TFIIF_alpha |
| F420_ligase | TGFb_propeptide |
| FA_desaturase_2 | TIM-br_sig_trns |
| FBP | TIMP |
| FBPase | TIP120 |
| FBPase_glpX | TNF |
| FGF | TPK_catalytic |
| FGase | TPMT |
| FLgD_tudor | TRM |
| FSH1 | TYW3 |
| FTR_C | Tagatose_6_P_K |
| FTR | TaqI_C |
| FadR_C | TehB |
| Fae | Tenui_N |
| Fascin | TerB |
| FbpA | TerD |

| | |
|---|---|
| FdhD-NarQ | Terminase_2 |
| FdhE | Terpene_synth_C |
| FdtA | Terpene_synth |
| Fe-ADH_2 | Thaumatin |
| Fe_hyd_lg_C | Thg1 |
| FemAB | Thi4 |
| FhuF | ThiC |
| FimH_man-bind | ThiI |
| Flavodoxin_4 | Thia_YuaJ |
| Flavodoxin_NdrI | Thioredoxin_5 |
| Flexi_CP | Thioredoxin_6 |
| FlhC | ThuA |
| FliG_C | TolB_N |
| FliM | Tol_Tol_Ttg2 |
| FliW | Toluene_X |
| FmdA_AmdA | TpcC |
| FmdE | Transglut_N |
| Frataxin_Cyay | Translin |
| Fructosamin_kin | Transthyretin |
| Fucose_iso_C | Trehalase |
| Fucose_iso_N1 | Trehalose_PPase |
| Fucose_iso_N2 | Triabin |
| Fumble | Trm112p |
| Furin-like | Trp_DMAT |
| G-alpha | Trp_halogenase |
| G3P_antiterm | TruD |
| GATase_4 | TrwB_AAD_bind |
| GBP_C | TrwC |
| GBP | Trypan_glycop |
| GCD14 | TylF |
| GCHY-1 | UBA_e1_C |
| GCS2 | UDPGP |
| GCS | UEV |
| GDA1_CD39 | UFD1 |
| GDI | UPF0027 |
| GFP | UPF0047 |
| GH3 | UPF0066 |
| GLF | UPF0075 |
| GLTP | UPF0113 |
| GNAT_acetyltr_2 | UPF0149 |
| GNAT_acetyltran | UPF0157 |

| | |
|---|---|
| GNT-I | UPF0302 |
| GPP34 | UPF1_Zn_bind |
| GRASP55_65 | UT |
| GSH-S_N | Ufd2P_core |
| GSH_synth_ATP | UreD |
| GSH_synthase | UreF |
| GSP_synth | Urease_alpha |
| GalP_UDP_tr_C | Ureidogly_hydro |
| GalP_UDP_transf | Uricase |
| GlcNAc_2-epim | Urocanase |
| Glu_cyclase_2 | UvdE |
| Glu_cys_ligase | UxaC |
| Glucokinase | UxuA |
| Glutaminase | V-ATPase_C |
| Glutaredoxin2_C | V-ATPase_H_C |
| Glyco_hydro_101 | V-ATPase_H_N |
| Glyco_hydro_10 | VHS |
| Glyco_hydro_11 | VanY |
| Glyco_hydro_12 | Viral_coat |
| Glyco_hydro_14 | Viral_protease |
| Glyco_hydro_15 | VitK2_biosynth |
| Glyco_hydro_17 | Vitellogenin_N |
| Glyco_hydro_19 | Vps26 |
| Glyco_hydro_26 | XFP_C |
| Glyco_hydro_30 | XFP |
| Glyco_hydro_35 | XFP_N |
| Glyco_hydro_38C | XdhC_C |
| Glyco_hydro_38 | Xpo1 |
| Glyco_hydro_39 | Xylanase |
| Glyco_hydro_4 | YTH |
| Glyco_hydro_53 | Y_phosphatase2 |
| Glyco_hydro_56 | YaeQ |
| Glyco_hydro_57 | YcgR |
| Glyco_hydro_65N | YdjC |
| Glyco_hydro_65m | YecM |
| Glyco_hydro_67C | YfbU |
| Glyco_hydro_67M | YfdX |
| Glyco_hydro_67N | YfiO |
| Glyco_hydro_68 | YhjQ |
| Glyco_hydro_6 | YiiD_Cterm |
| Glyco_hydro_72 | YkuI_C |

| | |
|---|---|
| Glyco_hydro_76 | YmdB |
| Glyco_hydro_7 | YodA |
| Glyco_hydro_85 | YopX |
| Glyco_hydro_88 | YqeY |
| Glyco_hydro_8 | YugN |
| Glyco_hydro_92 | YukC |
| Glyco_hydro_97 | YycH |
| Glyco_hydro_9 | YycI |
| Glyco_tranf_2_5 | Zeta_toxin |
| Glyco_trans_4_2 | ZipA_C |
| Glyco_transf_10 | Zn_dep_PLPC |
| Glyco_transf_15 | Zona_pellucida |
| Glyco_transf_20 | Zot |
| Glyco_transf_29 | bact-PGI_C |
| Glyco_transf_36 | dCMP_cyt_deam_2 |
| Glyco_transf_41 | eIF-5_eIF-2B |
| Glyco_transf_43 | eIF-6 |
| Glyco_transf_52 | efhand_1 |
| Glyco_transf_64 | iPGM_N |
| Glyco_transf_6 | mRNA_cap_enzyme |
| Glycoamylase | nsp8 |
| Glycogen_syn | nsp9 |
| Glycolytic | rRNA_methylase |
| Glycoprotein_B | s48_45 |
| Glycos_transf_N | tRNA-synt_1f |
| Glyoxalase_3 | tRNA-synt_2e |
| Glyoxalase_4 | tRNA_NucTran2_2 |
| Glyphos_transf | vATP-synt_AC39 |
| | zf-MaoC |
| | zf-ZPR1 |

Table S4.1: The exhaustive list of PFAM families studied in Chapter 4. This is presented in two-column format for brevity.