2009

Finite sample sizes and phylogeny do not ACCOUNT FOR THE MUTUAL INFORMATION OBSERVED FOR MOST SITE-PAIRS IN MULTIPLE SEQUENCE ALIGNMENT

Christopher S. DeHaan

Finite sample sizes and phylogeny do not account for the Mutual Information observed for most site-pairs in multiple sequence alignments

(Spine Title: Mutual Information Components in Multiple Sequence Alignments)

(Thesis Format: Integrated Article)

by

Christopher S. DeHaan

Graduate Program in Applied Mathematics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Mutual information (MI) is a measure frequently used to find co-evolving sites in protein families. However, factors unrelated to protein structure and function, in particular sampling variance in amino acid counts and complex evolutionary relationships among sequences, contribute to MI. Understanding the contribution of these components is essential for isolating the MI associated with structural or functional co-evolution. To date, the contributions of these factors to mutual information have not been fully elucidated.

We find that stochastic variations in amino acid counts and shared phylogeny each contribute substantially to measured MI. Nonetheless, the mutual information observed in real-world protein families is consistently higher than the expected contribution of these two factors. In contrast, when using synthetic data with realistic substitution rates and phylogenies, but without structural or functional constraints, the observed levels of MI match those expected due to stochastic and phylogenetic background.

Our results suggest that either low levels of co-evolution are ubiquitous across positions in protein families, or some unknown factor exists beyond the currently hypothesized components of intra-protein mutual information: sampling variance, phylogenetics and structural/functional co-evolution.

Statement of Coauthorship

The work presented in Chapter 2 has been submitted for publication.

> Christopher S. DeHaan, Andrew Fernandes, and L.M. Wahl. (2009) Finite sample sizes and phylogeny do not account for the Mutual Information observed for most site-pairs in multiple sequence alignments. *Bioinformatics, submitted.*

The original draft for the above article was prepared by the author. Subsequent revisions were performed by the author, Dr. Lindi M. Wahl and Dr. Andrew Fernandes. Some software components were originally designed and written by Dr. Andrew Fernandes. Development of the final software package, as well as analysis using C++ and Matlab, was performed by the author under the supervision of Dr. Andrew Fernandes and Dr. Lindi M. Wahl.

Acknowledgements

There are many people I would like to thank for their support over the last few years. Without their help, this thesis would not have been possible.

One person has served as both my mentor and inspiration: my supervisor, Dr. Lindi Wahl. Since I first attended her class in undergrad, I could see Lindi had a deep wealth of knowledge along with a genuine love of teaching. Two summers working for her in undergrad was such a positive experience that it brought me back to undertake a Masters with her. In my time here, she has shown incredible insight, approachability and encouragement.

Dr. Andrew Fernandes has supported my thesis work from the beginning. His attention to detail and rigour has undoubtedly strengthened this work. In a wide range of topics including programming and statistics, he was constantly available to help with any sized problem, as well as contributing insight which helped form the direction of the paper.

I am indebted to G. B. Gloor for providing high-quality alignments, and invaluable insight.

I want to thank my parents who have always supported my schooling, including my Masters; they have always been willing to help in any way they can including the recent daunting task of proof-reading. My office mates provided excellent company, both in and outside the office while my close friend Amanda Mundt has provided continued encouragement over the years, including attending many

of my talks. The office staff of this department has assisted in many ways while always maintaining a friendly disposition.

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

Table of Contents

List of Figures

Chapter 1

Introduction

With the aim of being accessible to the largest possible audience, an extremely brief overview of biological and statistical concepts fundamental to the thesis are presented in this first section. Interested readers are referred to texts such as Alberts *et al.* (2009) and Reza (1961) for additional background. Readers familiar with these areas may prefer to begin with Section 1.1 for a review of recent literature.

In most organisms, genetic information is stored in long, intricate strands of deoxyribonucleic acid or DNA. Through complex intermediate processes, this information is used to create strands of amino acids. These strands take on particular 3-dimensional configurations, or "fold" to form proteins. The overall shape and structure of proteins are significant, and critical to correct function.

Not all parts of a protein, however, are equally important for its proper function. Some amino acids, for example, may serve to physically connect adjacent regions of the sequence, but do not directly affect protein function. Thus, the exact type of amino acid in these locations, or "sites", in the protein sequence is less critical.

Some sites, however, serve an important role in the structure of the protein. Some pairs of amino acids may, for example, form bonds thereby connecting locations

which were well-separated in the sequence. In this way, protein folding leads to close proximity of sections in the sequence which were originally distant. The amino acid sequences in these structurally significant areas often cannot change freely since a different amino acid may be unable to form the required bond. If both amino acids in such a pair change, however, it may be possible to maintain the critical structural characteristics of the protein.

As well as site pairs which influence protein structure, some groups of amino acids are directly involved in the function of the protein, and are equally important. If the function of the protein involves binding to another molecule, or facilitating a chemical reaction, for example, amino acids with specific characteristics may be required at certain sites in the sequence. A single change in one of these sites may render the protein useless. However, as in the structural case described above, simultaneous mutations may preserve the protein's function.

A group of similar proteins, which perform the same biochemical function, can often be found across a range of organisms. The resemblance between these homologous proteins exists because they arise from common genetic ancestors. The entire evolutionary history of such a group can be represented as a phylogenetic tree. This tree shows which organisms share a common ancestor and the likely evolutionary distance to that point. These organisms themselves shared more distant ancestors, eventually tracing to a single common root. This paradigm presumes a single historical protein served as the founder of all existing forms of that protein. Selection facilitated changes and divergence of paths, as represented by tree branches, resulting in the currently observable set of proteins, represented by tree leaves.

A multiple sequence alignment (MSA) is the collection of homologous proteins from different organisms, the leaf nodes of the phylogenetic tree. The MSA lists the amino acid sequence from each protein as a row in the alignment. Many

areas in a protein are similar throughout most organisms allowing sequences to be lined-up with each other. However, because of different lengths of strands, gaps are added to shorter strands to line-up similar regions. Once complete, the MSA can be thought of as a table. Each row represents one organism's amino acid sequence for the protein of interest. Each column represents a common location or site shared in the protein family.

Given a MSA, it is trivial to spot regions in the protein which do not change among organisms, called "conserved sites". However, it is difficult to identify instances where the amino acids are not strongly conserved, but are still constrained by other locations in the protein. We use the term co-evolution, in this context, to represent a pair (or group) of amino acids that must evolve simultaneously to maintain an aspect of protein structure or function.

Mutual information (MI) is a measure of the mutual dependence of two variables. It represents the amount of information conveyed in one random variable about another. If two random variables are independent then their level of mutual information is zero. As the dependency between two random variables increases, so does the MI.

Formally, the formula for mutual information is:

$$\mathrm{MI} = \sum_{x \in X} \sum_{y \in Y} p_j(x, y) \log \left(\frac{p_j(x, y)}{p_m(x)\, p_m(y)} \right). \tag{1.1}$$

where X and Y are random variables which can take on values x and y respectively.

Here, $p_m(x)$ and $p_m(y)$ represent the probabilities of x and y independently while $p_j(x, y)$ is the probability of x and y occurring simultaneously. This measure can be applied to two sites in a MSA. In this case, X and Y represent the unknown amino acid at sites i and j respectively. This means $p_m(x)$ and $p_m(y)$ represent the probability of particular amino acids occurring at each site; $p_j(x, y)$

represents the probability of a particular pair of amino acids occurring together, meaning amino acid x occurs at site i, and amino acid y occurs at site j, in the same protein. Although we cannot directly know the probability of amino acids occurring, a common assumption is that the frequency with which an amino acid is observed, divided by the total number of observations, gives a good estimate of its probability. However, as described in detail in the following Chapter, when there are only a small number of observations, this estimation becomes less accurate. When this potential inaccuracy is assumed negligible, and probabilities are estimated from observed frequencies, a single point value for MI can be calculated for each site pair in the sequence. However, if this uncertainty is considered, then a distribution of possible MI values results for each site pair, as described in Chapter 2.

1.1 Applying Mutual Information techniques

The insight that MI is a tool that can be applied to MSAs in order to detect co-evolution was first suggested by KORBER *et al.* (1993). Point estimates of MI were originally used in a straightforward way with no correction terms; we will call this statistic "uncorrected MI". KORBER *et al.* (1993) found sites with high MI values frequently corresponded to sites which independent experiments indicated had structural significance.

CLARKE (1995) also applied uncorrected MI to MSAs from proteins whose structure was already partially known. He selected 16 highly co-varying pairs in the sequence and then looked at the specific structures found at these pairs in the folded protein. Indeed, many of these areas had structural significance such as a salt bridge or other binding mechanisms.

Most protein structures, however, are unknown. Methods may be developed to

identify structurally or functionally significant pairs, but there is no certain way to evaluate their accuracy. To resolve this, proximity is used as a gold standard to evaluate if a method can identify sites likely to be co-evolving. If a method effectively identifies sites which are physically close in the folded structure of the protein, it is considered accurate.

OLIVEIRA *et al.* (2002) also used uncorrected MI, subdividing MI from "structure and function" into three categories: "part of the main active site", "part of a modulator binding site" and "transducing a signal between those sites". These authors used a broad class of methods called Correlated Mutation Analysis (CMA) to categorize amino acids into pairs and groups which are likely related in terms of the function they perform. CMA looks at a combination of entropy, co-variation and physicochemical similarities between amino acids. The paper also touched on the very high conservation of functional location in a protein as compared to the conservation of the particular amino acids which perform that task.

Although MI can be modified with correction terms, some recent papers still use uncorrected MI values. One such recent application includes the discrimination of phylogenetically related organisms. In particular, WECKWERTH and SELBIG (2003) showed that applying mutual information assists in the discovery and identification of amino acid sites or motifs which are particular to different kinds of organisms. Regions in proteins from mammals, plants and bacteria may be consistent within their own group, but show marked differences across groups.

A second important application of uncorrected MI has been the identification of site pairs which are critical for drug resistance in HIV. HOFFMAN *et al.* (2003) created a MSA from HIV sequences and subsequently divided it into two groups. One contained sequences from people who had not been taking drugs to combat the virus; the other group had sequences from people who did. There was suffi-

cient variation in these sequences to pinpoint co-evolving sites which existed only in the sequences from people undergoing drug treatment which in turn indicated these pairs are critical to drug resistance.

1.2 Mutual Information components

Calculating the point MI for a site pair is straightforward; however the intuitive explanation for the resulting value is not. Unfortunately, diverse sources contribute to the total observed MI. These sources may be summarized as:

$$\mathrm{MI}_{(\mathrm{observed})} = \mathrm{MI}_{(\mathrm{stochastic})} + \mathrm{MI}_{(\mathrm{phylogeny})} + \mathrm{MI}_{(\mathrm{structure})} + \mathrm{MI}_{(\mathrm{function})} \qquad (1.2)$$

as hypothesized by ATCHLEY *et al.* (2000).

Observed MI represents the total MI, calculated directly from the MSA. Stochastic MI includes effects such as random variations and small sample sizes. In other words, one would expect that purely by chance, certain pairs would appear to undergo some level of co-evolution. Since there are a finite number of organisms in the MSA, small sample size becomes an unavoidable issue. MI is always a non-negative value meaning any such variation changing the value from zero will always produce a positive level of MI. A second factor, Phylogenetic MI, considers the effect of a shared evolutionary history. Similarities in a group of organisms' proteins may arise from the fact that the organisms have a common ancestor and may not be related to any structural or functional constraints. If by chance two sites mutate far back in the phylogenetic tree, then observed levels of MI would increase despite a lack of structural or functional significance in the pair. Similar to Stochastic MI, this effect will always contribute positively to the observed MI levels for a pair. The final factor we consider is the effect of structural and func-

tional constraints. This component reflects the relationship between sites that are critical to the operation of the protein.

Of these components, the aim of co-evolution analysis is to isolate the MI arising from structure or function. In recent years, a great deal of research has focused on ways to remove or account for both the stochastic and phylogenetic components of the measurement which we will call background MI. This includes correction terms which consider what levels of background MI are expected for a given site pair.

1.3 Modifications to basic Mutual Information methods

The uncorrected MI value calculated for a pair can be modified in various ways to improve co-evolution prediction accuracy. To this end, many authors have proposed methods which compare one pair's MI values in some way with the MI observed for all other site pairs. For example, TILLIER and LUI (2003) searched for pairs which co-varied differently to most others. This involved identifying groups of organisms in the MSA which tended to share mutations. Next, variations which occurred within groups of otherwise similar organisms were flagged. Since wide-spread co-variation typically arises from shared phylogeny, sites which did not follow the overall pattern likely co-varied for non-phylogenetic reasons. This paper indirectly accounted for phylogeny by assuming groups of proteins with similar patterns were closely related, but did not take phylogenetic history directly into account.

SÜEL *et al.* (2003) used a statistic called "statistical coupling energy", a measure that is closely related to MI, but which takes globally observed amino acid frequencies into account. With this measure, SÜEL *et al.* (2003) were able to identify groups of amino acids which represent disjoint functional areas on the

protein. These authors expanded the common conception that co-evolution primarily occurs between amino acids which are physically close, hypothesising that a network may link physically distant areas on the protein thereby making them functionally related.

Similarly, correlation coefficients computed between sites may be used to identify regions of interaction within a protein (SARAF *et al.* 2003). Correlation coefficients consider the overall similarity of a pair compared to the level of conservation of each site. This measure was likewise used to find regions in the protein which were functionally important.

A different modification of MI, normalizing MI by the joint entropy of the pair of positions, was shown to offer substantial improvements over uncorrected MI in predicting contacting pairs (MARTIN *et al.* 2005). Joint entropy is a measure of the variability of the amino acids at a particular pair of sites. If the amino acids are highly conserved across organisms then the entropy value will be low. A modified MI value is found by dividing point MI estimates by the joint entropy to create a statistic called MIr.

An extension of this work is to identify co-evolving groups of sites. GLOOR *et al.* (2005) used MIr to support the finding that amino acid sites that co-vary with relatively few other sites are more likely to be structurally important while those that co-vary in larger groups of related sites are usually critical for the function of the protein.

FARES and TRAVERS (2006) continued this work, which sought to isolate the notoriously difficult to separate MI components "structural" and "functional". Proteins with known 3D structures and areas of functional importance were analyzed. The results confirmed previous findings that sites with a small number of co-varying sites tended to be in close physically proximity and related to the

structure of the protein. Sites which were part of a larger network of co-variation tended to be important to the function of the protein and were not necessarily physically close in the folded protein.

More advanced normalizations can even better account for phylogenetic effects indirectly by assuming that the phylogenetic background will be largely the same across all sites in the protein family. When searching for unusually high levels of mutual information, DUNN *et al.* (2008) used all pairs to account for background levels of mutual information. This entails comparing the observed mutual information of one pair to the observed mutual information of all other possible pairs which include one of the original sites. That is, if sites 3 and 5 are of interest, then the observed mutual information is normalized by the average MI for all other site pairs that include either 3 or 5.

LITTLE and CHEN (2009) made a further refinement to the analysis of this normalized value by measuring the level of variability in MI values. This allowed for normalization of both MI levels and variability. With these values, a refined calculation of the discrepancy between observed MI point estimate and expected MI as calculated by DUNN *et al.* (2008) was possible. This in turn identified sites with unusually high MI levels and a high probability of co-evolution.

A recent development has been the use of weighted importance of sequences and low-count corrections to reduce the impact of small, redundant samples (BUSLJE *et al.* 2009). This stops similar, but not identical, sequences from being totally eliminated while still preventing them from biasing sequence calculations. This allows for better extraction of data from observed MSAs.

1.4 Including phylogenetic information

None of the methods described above directly consider the phylogenetic history of the protein family when computing co-evolution scores. However, a number of co-evolution analyses which include the underlying phylogeny, so-called "tree-aware" methods, have also been proposed.

POLLOCK *et al.* (1999) described an early method to include phylogenetic information which also reduced amino acids from the original 20 possibilities to a two-state system, considering properties of amino acids such as size or charge. Analysis of the results focused on physical proximity and simple structural constructs. The method identified sites which were touching in a helix loop, and frequently had opposing charges. In addition, many sites found to be coevolving on the surface of the protein were in close physical proximity.

TUFF and DARLU (2000) used phylogenetic history to reconstruct simulated protein families from 15 different sets of homologous proteins. Similar to POLLOCK *et al.* (1999), they considered substitutions which affected the physicochemical properties of the site. The authors investigated different methods to construct the phylogenetic tree as well as different possible reduced amino acid alphabets. The number and location of co-varying sites found under different tree-construction methods and amino acid alphabets were reported. This large sensitivity to the particular methods used highlighted the difficulty of obtaining robust predictions of co-evolution leaving the problem, admittedly, unresolved.

A method used in this thesis, bootstrapping, was described by WOLLENBERG and ATCHLEY (2000). In this method, the phylogenetic tree and MSA are used to find a probabilistic root ancestor based on tree branch length and position. With this, a new MSA is created by first instantiating one root ancestor given the probabilities of each amino acid at each site. Now, mutations are added

stochastically, depending on the branch lengths, until all leaf nodes contain a new protein sequence. MI calculated from this MSA will include the effects of phylogeny and stochastic variation but not structural or functional constraints, since they were not included in the simulated evolution. WOLLENBERG and ATCHLEY (2000), however, considered only point estimates of MI, as opposed to distributions considered in this thesis.

By comparing point estimates of the MI found in the bootstrapped MSAs with the MI point estimate from the original alignment, sites with high co-variation for structural or functional reasons were identified by ATCHLEY *et al.* (2000). While WOLLENBERG and ATCHLEY (2000) had focused on a known structure, the bootstrapping method was expanded in this contribution to investigate proteins with unknown structure. The idea of comparing bootstrap (expected) MI with observed MI is a cornerstone of this thesis.

This method of expected and observed MI comparison is not limited to site pairs and was extended by BUCK and ATCHLEY (2005) to find networks of co-evolving sites. The paper investigated the serpin proteins and found extensive networks of highly correlated sites corresponding to groups which played a significant functional role in the protein.

Mutations are typically considered to occur with equal probability over any length of branch in the phylogenetic tree. However, DIMMIC *et al.* (2005) considered on which branches of the tree mutations were most likely to occur for each site. Mutation likelihood was calculated for branches in the phylogenetic tree for all sites independently. To develop a null hypothesis of no co-evolution, bootstrap methods described by POLLOCK *et al.* (1999) were used. Test statistics then determined if the null hypothesis could be rejected. If so, the pair was considered to be co-evolving. DIMMIC *et al.* (2005) were able to identify a large number of co-evolving sites in real-world eukaryotic proteins, using amino acid proximity in

the folded structure to assess accuracy.

1.5 Comparison of methods

Despite this progress, it is as yet unclear which methods work best when used with real data and even whether including complete phylogenetic information in co-evolution analysis improves the accuracy or statistical power of the measure. FODOR and ALDRICH (2004) investigated 4 tree-unaware methods, including uncorrected MI, and focused on their ability to correctly detect covariance in both synthetic and real world data. Each method was found to have an optimal level of site conservation at which it performed best. This implied that different methods on different types of data, or an intelligent hybrid of multiple methods, would improve detection. CAPORASO *et al.* (2008) directly addressed the question of co-evolution detection by tree-aware and tree-ignorant methods. In the same paper, the effect of using reduced-state amino acid alphabets was investigated. Using 4 tree-aware and 5 tree-unaware metrics, CAPORASO *et al.* (2008) analysed a protein which contained a large, well-known helix structure. Since amino acids separated by 4 sites were known to interact and be critical to the helix structure, the paper investigated each method's ability to detect co-evolution in these site pairs. Tree-ignorant methods, it was found, were generally as powerful as tree-aware methods. The use of reduced alphabets was less clear, with some recodings offering improvements while others did not.

1.6 Conclusion

Explicitly or implicitly, all of the methods proposed for identifying co-evolving sites in protein families aim to minimize or eliminate the contribution of shared

phylogeny and stochastic variation to the observed co-evolution signal. As mentioned by Buslje *et al.* (2009) and Little and Chen (2009) however, the effect of the $MI_{(stochastic)}$ and $MI_{(phylogeny)}$ components has not been fully explored to date. In Chapter 2, we shed light on the nature of these contributing factors.

REFERENCES

ALBERTS, B., D. BRAY, K. HOPKIN, A. JOHNSON, J. LEWIS, M. RAFF, K. ROBERTS, and P. WALTER, 2009 *Essential Cell Biology.* Garland Publishing.

ATCHLEY, W. R., K. R. WOLLENBERG, W. M. FITCH, W. TERHALLE, and A. W. DRESS, 2000 Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol *17*(1): 164–178.

BUCK, M. J. and W. R. ATCHLEY, 2005 Networks of coevolving sites in structural and functional domains of serpin proteins. Mol Biol Evol *22*(7): 1627–1634.

BUSLJE, C. M., J. SANTOS, J. M. DELFINO, and M. NIELSEN, 2009 Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics *25*(9): 1125–1131.

CAPORASO, J. G., S. SMIT, B. C. EASTON, L. HUNTER, G. A. HUTTLEY, and R. KNIGHT, 2008 Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. BMC Evol Biol 8: 327.

CLARKE, N. D., 1995 Covariation of residues in the homeodomain sequence family. Protein Sci *4*(11): 2269–2278.

DIMMIC, M. W., M. J. HUBISZ, C. D. BUSTAMANTE, and R. NIELSEN, 2005 Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics **21 Suppl 1:** i126–i135.

DUNN, S. D., L. M. WAHL, and G. B. GLOOR, 2008 Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics *24*(3): 333–340..

FARES, M. A. and S. A. A. TRAVERS, 2006 A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. Genetics *173*(1): 9–23.

FODOR, A. and R. ALDRICH, 2004 Influence of Conservation on Calculations of Amino Acid Covariance in Multiple Sequence Alignments. Proteins: Struct Funct Bioinf **56:** 211–221.

GLOOR, G. B., L. C. MARTIN, L. M. WAHL, and S. D. DUNN, 2005 Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry *44*(19): 7156–7165.

HOFFMAN, N. G., C. A. SCHIFFER, and R. SWANSTROM, 2003 Covariation of amino acid positions in HIV-1 protease. Virology *314*(2): 536–548.

KORBER, B. T., R. M. FARBER, D. H. WOLPERT, and A. S. LAPEDES, 1993 Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A *90*(15): 7176–7180.

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

MARTIN, L. C., G. B. GLOOR, S. D. DUNN, and L. M. WAHL, 2005 Using information theory to search for co-evolving residues in proteins. Bioinformatics *21*(22): 4116–4124.

OLIVEIRA, L., A. C. M. PAIVA, and G. VRIEND, 2002 Correlated mutation analyses on very large sequence families. Chembiochem *3*(10): 1010–1017.

POLLOCK, D. D., W. R. TAYLOR, and N. GOLDMAN, 1999 Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol *287*(1): 187–198.

REZA, F. M., 1961 *An Introduction to Information Theory.* McGraw-Hill.

SARAF, M. C., G. L. MOORE, and C. D. MARANAS, 2003 Using multiple sequence correlation analysis to characterize functionally important protein regions. Protein Eng *16*(6): 397–406.

SÜEL, G. M., S. W. LOCKLESS, M. A. WALL, and R. RANGANATHAN, 2003 Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol *10*(1): 59–69.

TILLIER, E. R. M. and T. W. H. LUI, 2003 Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics *19*(6): 750–755.

TUFF, P. and P. DARLU, 2000 Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol Biol Evol *17*(11): 1753–1759.

WECKWERTH, W. and J. SELBIG, 2003 Scoring and identifying organism-specific functional patterns and putative phosphorylation sites in protein sequences using mutual information. Biochem Biophys Res Commun **307**: 516–521.

WOLLENBERG, K. R. and W. R. ATCHLEY, 2000 Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci U S A *97*(7): 3288–3291.

Chapter 2

Teasing apart components in observed Mutual Information from multiple sequence alignments

2.1 Introduction

It has long been understood that co-evolution, both among and within proteins, may indicate associations of structural or functional significance (Fitch and Markowitz 1970). A large number of techniques have thus been developed which quantify some signature of co-evolution, and use this pattern to identify protein-protein interactions (Marcotte *et al.* 1999; Goh *et al.* 2000; Goh and Cohen 2002; Pazos *et al.* 2005; Waddell *et al.* 2007), domains of interaction within and between proteins (Larson *et al.* 2000; Kim and Subramaniam 2006; Kim *et al.* 2006; Fares and Travers 2006), and co-evolving sites within protein families (for review see Codoñer and Fares (2008) and Pazos and Valencia (2008)).

For intraprotein interactions in particular, Mutual Information (MI) is a tool that can be applied to multiple sequence alignments (MSAs) in order to detect co-evolution (Korber *et al.* 1993; Clarke 1995; Oliveira *et al.* 2002; Fodor and Aldrich 2004). When applied to an MSA, the observed MI reflects the information content carried by one site in the protein about another

site. Since co-evolving sites are expected to share high MI, MI may be used to identify sites of structural or functional significance within the protein sequence. Recent applications include the discrimination of phylogenetically related organisms (WECKWERTH and SELBIG 2003) and the identification of site pairs which are critical for drug resistance in HIV (HOFFMAN *et al.* 2003).

Unfortunately, diverse sources are believed to contribute to the total observed MI. These sources may be summarized as: $MI_{(observed)} = MI_{(stochastic)} + MI_{(phylogeny)} + MI_{(structure)} + MI_{(function)}$, as hypothesized by ATCHLEY *et al.* (2000). Of these components, the aim of co-evolution analyses is to isolate the MI arising from structure or function. Intense research effort has thus been devoted in recent years toward developing an appropriately normalized or corrected measure of MI, which reduces or eliminates both the stochastic and phylogenetic components of the measurement.

To do this, many investigators have proposed methods which use simple normalizations or rapidly-computed comparisons over all site pairs. For example, multiple significant interdependency may be used to discriminate functional versus phylogenetic co-evolution (TILLIER and LUI 2003). Normalizing MI by the joint entropy of the pair of positions offers substantial improvements over unnormalized MI in predicting contacting pairs (MARTIN *et al.* 2005). An extension of this work is to identify co-evolving groups of sites (GLOOR *et al.* 2005), which may help distinguish important structural locations from those of functional significance (FARES and TRAVERS 2006). Statistical coupling energy, a measure closely related to MI, has also been used to indentify groups of amino acids which represent distant functional areas on the protein (SÜEL *et al.* 2003). Similarly, correlation coefficients computed between sites may be used to identify regions of interaction within a protein (SARAF *et al.* 2003). Normalizations can also account for phylogenetic effects indirectly by assuming that the phylogenetic background will be largely the same across all sites in the protein family (DUNN *et al.* 2008),

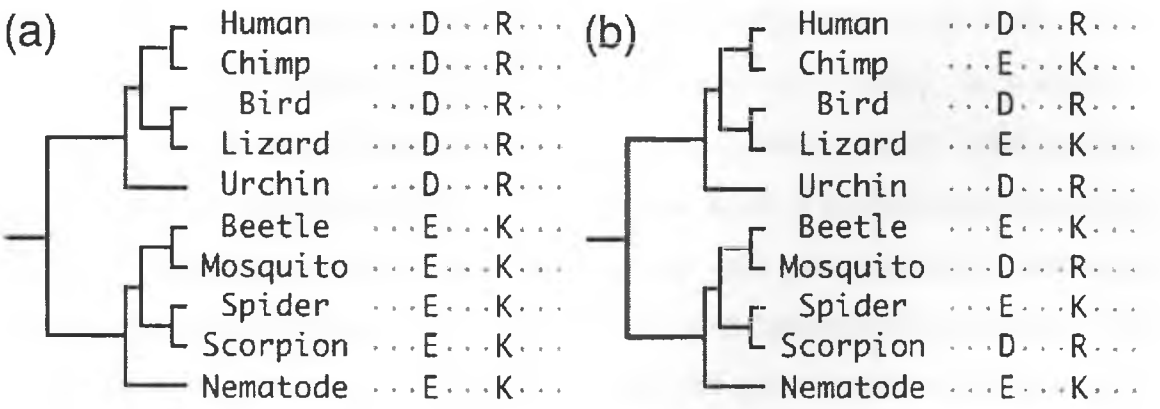

Figure 2.1: Two different pairs of sites in a protein. We consider one pair which is highly conserved with a single mutation in both sites near the tree root; this would result in the MSA depicted in (a). In contrast, panel (b) would result if the sites are strongly co-varying and not well conserved. These two very different situations produce identical MSAs and are indistinguishable if phylogenetic information is ignored.

and by manipulating residuals between observed and expected MI based on this assumption (Little and Chen 2009).

None of the methods described above considers the phylogenetic history of the protein family when computing co-evolution scores. The need for including phylogenetic information can be illustrated with a simple example, as shown in Figure 2.1. Suppose that one pair in this example, labeled (a), consists of highly conserved sites sharing a single coordinated mutation event near the root. This could for example indicate a distant selective sweep, or founder effect, and would not commonly be considered coevolution. In contrast, suppose that a second pair, labeled (b), consists of highly variable positions, in which substitution events are coordinated. These multiple, coordinated substitutions are commonly refered to as 'molecular coevolution'. Note that both panels of Figure 2.1 result in the same MSA, albeit with shuffled rows, despite describing very different biological events. Therefore without considering the underlying phylogeny we are unable to understand the evolutionary relationship between sites.

A number of co-evolution analyses which include the underlying phylogeny, so-

called "tree-aware" methods, have been proposed (POLLOCK *et al.* 1999; TUFF and DARLU 2000; DIMMIC *et al.* 2005; CAPORASO *et al.* 2008). In a study of nucleotide sequences, for example, (WOLLENBERG and ATCHLEY 2000) use the phylogenetic tree to recreate via bootstrap a new MSA. This new set of sequences captures co-variation from phylogenetic sources and also stochastic variations, but not structural or functional constraints. By comparing point estimates of the MI found in the bootstrapped MSAs with the MI point estimate from the original alignment, sites with high co-variation for structural or functional reasons (ATCHLEY *et al.* 2000), or networks of co-evolving sites (BUCK and ATCHLEY 2005), can be identified. Despite this progress, it is as yet unclear whether including complete phylogenetic information in co-evolution analyses improves the accuracy or statistical power of the measure (FODOR and ALDRICH 2004; CAPORASO *et al.* 2008).

In contrast with the phylogenetic component of MI, the component $MI_{(stochastic)}$ has received comparatively little attention. Estimates of the number of sequences required to obtain meaningful MI measures range from 30 (FARES and TRAVERS 2006) to 400 (BUSLJE *et al.* 2009). A recent development has been the use of weighted importance of sequences and low-count corrections to reduce the impact of small, redundant samples (BUSLJE *et al.* 2009).

Explicitly or implicitly, all of the methods proposed for identifying co-evolving sites in protein families aim to minimize or eliminate the contribution of shared phylogeny and stochastic variation to the observed co-evolution signal. To date, however, the contributions of $MI_{(stochastic)}$ and $MI_{(phylogeny)}$ have not been fully elucidated, as noted by several authors (WOLLENBERG and ATCHLEY 2000; BUSLJE *et al.* 2009; LITTLE and CHEN 2009). The purpose of this study is to identify and isolate the contribution of each of these factors to the observed MI. We find, strikingly, that these two components do not explain the levels of MI observed for most position pairs in MSAs of protein families.

2.2 Methods

Our analysis requires a MSA, site-specific substitution rates, and a corresponding phylogenetic tree for which both topology and branch lengths have been estimated. For this study, the following set of functionally and structurally diverse proteins were used: triosephosphate isomerase (TIM1) (454 sequences, 320 sites); dihydropteroate synthetase (215 sequences, 370 sites); phosphopyruvate dehydratase (162 sequences, 508 sites); methionine aminopeptidase-1 (223 sequences, 332 sites); phosphoglycerate kinase (326 sequences, 581 sites); the GroEL chaperonin (317 sequences, 591 sites), where the number of sites represents the number of amino acids in one sequence including gaps. All sequences were obtained from GenBank and had redundant sequences of more than 90% identity removed.

Protein families were aligned using a combination of the NCBI's Cn3D structural alignment system (Hogue 1997) and extensive manual curation. Midpoint-rooted phylogenetic trees were computed using PhyML under a Gamma model of rate heterogeneity (Guindon and Gascuel 2002; Guindon and Gascuel 2003) using the WAG substitution matrix (Whelan and Goldman 2001). Rates were estimated using both bRate (Fernandes and Atchley 2008) and Rate4Site (Pupko *et al.* 2002).

2.2.1 Algorithm

Estimating the Observed MI Distribution

Our algorithm proceeds as follows. For any two sites (columns in an MSA), we calculate the pair or "joint" counts, a 20-by-20 table of the number of times each amino acid pair occurs in the two columns, and the "marginal" counts, the number of times each amino acid occurs in each column. As described in greater

detail below, MI is not fundamentally a function of these amino acid *counts*, but of *frequencies*. Typically, this distinction is obscured by assuming that a reasonable estimate of the underlying frequency is given by the fraction: (counts)/(total observations). Under this assumption, MI can be computed in the usual way (MARTIN *et al.* 2005); we refer to this as the MI "point estimate".

Unfortunately, such point estimates are extremely sensitive to the number of sequences in the alignment, or the number of entries in the 20-by-20 contingency table. In any such table from real-world proteins, only a small fraction of the 400 possible amino acid pairs are typically non-zero. The MI point estimate implicitly assumes that if a pair is not observed, it would never be observed; its expected frequency is assumed to be zero. This is an extremely strong assumption, and acts as a large source of systematic bias (HUTTER and ZAFFALON 2005).

We therefore use the observed counts to estimate, in a statistically rigorous way, a *distribution* of the possible frequencies which might underly the observed data. We then generate repeated random samples from this distribution, calculating a value of MI each time. This yields a distribution of values of MI which might have been observed, given our best guesses about the underlying frequencies of amino acid pairs. In essence, this procedure accounts for the fact that we do not have every possible protein sequence in the alignment, but have observations based on only a subset of all possible organisms.

For brevity we call the distribution computed in this way the 'observed MI'. An example is given in the right-most histogram in Figure 2.2 and described further in the Results.

Estimating $MI_{(stochastic)}$

The procedure outlined above illustrates that stochastic variations in amino acid counts, based for example on which organisms are included in a specific alignment, produce variation in MI. We would like to quantify the contribution of this stochastic variation alone, in the absence of any structural/function or phylogenetic sources of MI. To do this, we assume that amino acid frequencies at the two sites in question are independent. In particular, we use the same statistically rigorous procedure to estimate the distribution of the marginal frequencies, the frequencies of each amino acid at each site, given the observed data. We then draw a random sample from this distribution. Since we have assumed independence, we compute each of the joint frequencies as the product of the appropriate marginal frequencies. A value of MI is then computed based on these joint and marginal frequencies, and the process is repeated to estimate the distribution of $MI_{(stochastic)}$, as illustrated by the leftmost histogram in Figure 2.2.

Estimating the expected MI distribution

To generate the expected MI distribution, we would like to generate possible values of MI which might occur given both stochastic variations in amino acid counts, and shared phylogenetic history, but without assuming any structural or functional relationship between the two sites. To do this, we bootstrap an evolutionary history, but assume that mutations occur independently at the two sites.

As stated above, our method requires a phylogenetic tree and site-specific substitution rates as inputs. We use these, along with the observed MSA, to generate a probabilistic ancestor at the root of the tree. The ancestor is probabilistic in the sense that we estimate a probability that the residue at a specific site in the

ancestor was a specific amino acid. We can then sample with these probabilities to generate a particular possible ancestor. Starting from this ancestor, we assume that the evolutionary history at each site can be modeled by a standard evolutionary Markov process at the estimated site-specific substitution rates, using the WAG substitution matrix (WHELAN and GOLDMAN 2001), and following the given phylogenetic tree. Again we note that substitutions occur independently at each site throughout this bootstrap procedure. At the end of this procedure, we have a single bootstrapped MSA. Repeating this procedure many times yields a distribution of MI, as illustrated in the central histogram of Figure 2.2. We note that this 'expected MI', or $MI_{(stochastic)} + MI_{(phylogenetic)}$, accounts explicitly for sampling variance, rate heterogeneity, amino acid substitution similarity and shared phylogeny for every site pair.

Comparing observed and expected MI

The expected MI distribution computed above contains both the stochastic and phylogenetic components of MI, but no structural or functional information. In contrast, when we use the original MSA to estimate the observed MI distribution, structural or functional constraints on the pair frequencies are preserved. Thus our null hypothesis is that the distribution of observed MI for most site pairs will not differ significantly from the expected distribution.

Determining the extent to which alternate probability distributions are distinguishable, via their samples, is a nontrivial task. The most popular approaches are based on the work of NEYMAN and PEARSON (1933) and are related to the parametric analysis of Receiver Operating Characteristic curves (LASKO *et al.* 2005). For this work we instead used a method similar to the Mann-Whitney U statistic (MANN and WHITNEY 1947) and the Area Under the Curve (AUC), as described by FAWCETT (2006). To verify robustness, numerous variations of this measure

were used. In all cases, differences between these variants were negligible, showing that our measure of separability was invariant to minor changes in the analytical-discrimination framework.

We scale our separability measure such that if the expected distribution lies to the left of the observed, and has very little overlap, the separability is close to negative one. If the two distributions overlap perfectly, the measure will return a value of zero separability. It is uncommon that the expected MI level is larger than the observed, but is possible with stochastic variations; in this case, if the expected distribution lies entirely to the right of the observed, the separability is one.

An internal control for the analysis

The integrity of our analysis was tested by using bootstrapped MSAs, in which site pairs do not have structural or functional constraints, as internal controls. In brief, we use randomly chosen MSAs created by the bootstrapping procedure described above as the input to our algorithm, along with the rates and phylogenetic tree of the original MSA. We then compute a control case of the "observed" MI distribution, a control case of the probabilistic root ancestor and a control set of bootstrapped MSAs. These allow us to estimate an expected MI distribution known to be free of structural and functional constraints. In principle, this distribution should be indistinguishable from the "observed" MI distribution for the respective bootstrapped MSA.

2.2.2 Implementation

Observed MI Distribution

As stated above, rare events (low or zero contingency table counts) have a large impact on the 'information' of the data (SHANNON 1948a; SHANNON 1948b), and can systematically bias subsequent computations, such as MI (HUTTER and ZAFFALON 2005). Attempts have been made to derive a closed-form expression for the sampling distribution of the MI estimator (HUTTER and ZAFFALON 2005), but only asymptotic moment-expansions have been found. Therefore we follow a standard Monte Carlo approach where a Bayesian posterior frequency distribution is derived from the presumed-multinomial contingency counts (ROBERT 2001). Specifically, a Dirichlet prior is combined with a multinomial likelihood to estimate posterior contingency frequencies p from counts n, both joint and marginal, such that

$$\Pr(p|n) \propto \Pr(n|p) \Pr(p|\alpha), \tag{2.1}$$

where $\Pr(n|p)$ is multinomial likelihood and $\Pr(p|\alpha)$ is Dirichlet. Since the multinomial distribution is conjugate to the Dirichlet, the posterior is also Dirichlet-distributed. Following BERGER and BERNARDO (1992), each component of hyperparameter α is set to $1/2$ such that the prior is minimally informative both in the sense of BERGER and BERNARDO (1992) and Jeffreys (ROBERT 2001).

Let n_{ij} represent the observed joint count of amino acid pair (i, j), while $n_{i\bullet}$ and $n_{\bullet j}$ denote the counts of the respective marginals. Then the joint frequencies p_{ij} and marginal frequencies $p_{i\bullet}$ and $p_{\bullet j}$ are Dirichlet distributed with parameters $n_{ij} + \alpha$, $n_{i\bullet} + \alpha$, and $n_{\bullet j} + \alpha$ respectively. The sampling distribution of the MI estimator is computed by repeatedly drawing values of p_{ij}, $p_{i\bullet}$, and $p_{\bullet j}$ and

computing

$$\mathrm{MI} = \sum_{i=1}^{20} \sum_{j=1}^{20} p_{ij} \log \left(\frac{p_{ij}}{p_{i\bullet}\, p_{\bullet j}} \right). \tag{2.2}$$

Inspection showed that 1000 replicates were required to reasonably estimate the posterior MI distribution from our typical amino acid pair count contingency tables.

$\mathrm{MI}_{(\mathrm{stochastic})}$

To estimate $\mathrm{MI}_{(\mathrm{stochastic})}$, we generate joint frequencies as the product of the appropriate marginal frequencies as estimated via the Dirichlet posteriors. New joint counts are then generated via multinomial sampling of joint frequencies. An MI value is computed based on this new joint table, and the process is repeated.

Expected MI distribution

Under the hypothesis of site-independence, parametric bootstrapping was used as per WOLLENBERG and ATCHLEY (2000). Specifically, the distribution of amino acids for each site of the root ancestor was estimated using standard phylogenetic likelihood techniques (FELSENSTEIN 2004). The probabilistic ancestor was repeatedly sampled and subject to tree-guided Markovian evolution with site-specific rates to realize hypothetical MSAs under the null model of site-independence. For each MSA realization an 'observed MI' distribution was generated via multinomial sampling (as described for the original MSA), and then a single MI value was drawn from this distribution. The distribution of an equal-weighted mixture of such bootstrap realizations is the 'expected MI'; we note that the expected MI distribution conditions only on site independence.

2.3 Results

2.3.1 Distribution comparison

Figure 2.2 shows the resulting distributions of mutual information for a typical pair of positions in a multiple sequence alignment. For illustrative purposes we have used a pair of positions from the alignment TIM1, as described in the Methods. The histogram on the left shows the mutual information attributable to finite sampling alone, $MI_{(stochastic)}$. The centre distribution shows the mutual information which is attributable to the effects of both finite sampling and shared phylogeny, $MI_{(stochastic)} + MI_{(phylogenetic)}$. The rightmost histogram shows the distribution of the observed MI shared by that pair, given the original data in the multiple sequence alignment.

In this example, we note that the observed MI is clearly well-separated from our expectations based on finite sampling and shared phylogeny. We measured the separability between the phylogenetic and observed MI distributions for all ungapped position pairs in TIM1 (over 10,000 pairs); the resulting histogram is shown in Figure 2.3. The surprising result here is that for almost all position pairs in this alignment, the observed and expected histograms are extremely separable.

We repeated this analysis on each of the six structurally and functionally diverse protein families described in the Methods. For each protein, fifty sites were chosen uniformly at random, yielding approximately 1000 site-pairs per protein (with minor variability due to gap removal). The resulting separability histograms for each protein family are provided in Figure 2.4. For the majority of site pairs in every protein, the observed MI distribution was separable from the expected distribution, as indicated by the preponderance of large negative separability values. We also observed that as the length of the protein increased, a larger fraction of site pairs had separability at or near zero (compare top left panel with

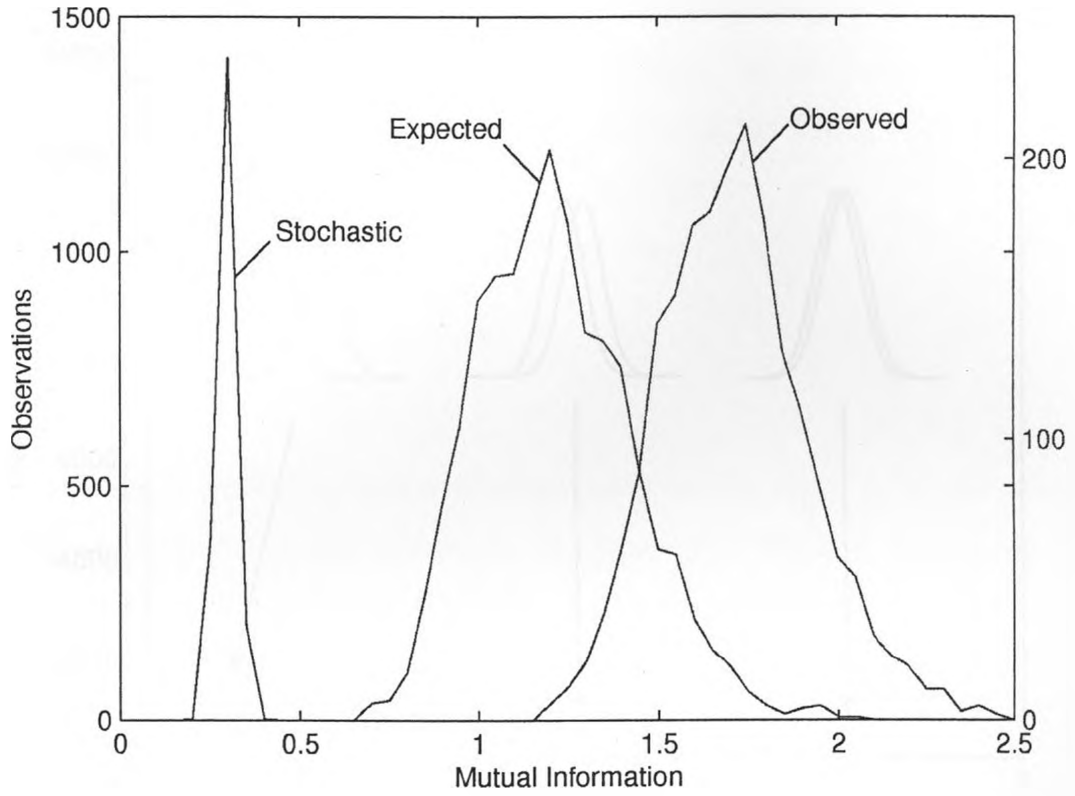

Figure 2.2: Typical distributions of mutual information from a pair of sites in a real world protein. The left distribution shows the expected mutual information when including stochastic effects, but ignoring phylogenetic information. The center distribution shows expected mutual information when phylogenetic information is also considered. The right distribution is the observed mutual information for the pair. Each distribution represents 2000 realizations for sites 12 and 69 in protein TIM1. The point estimate of MI for these two sites is 0.19. The left distribution uses the left axis while the center and right distribution both use the right axis.

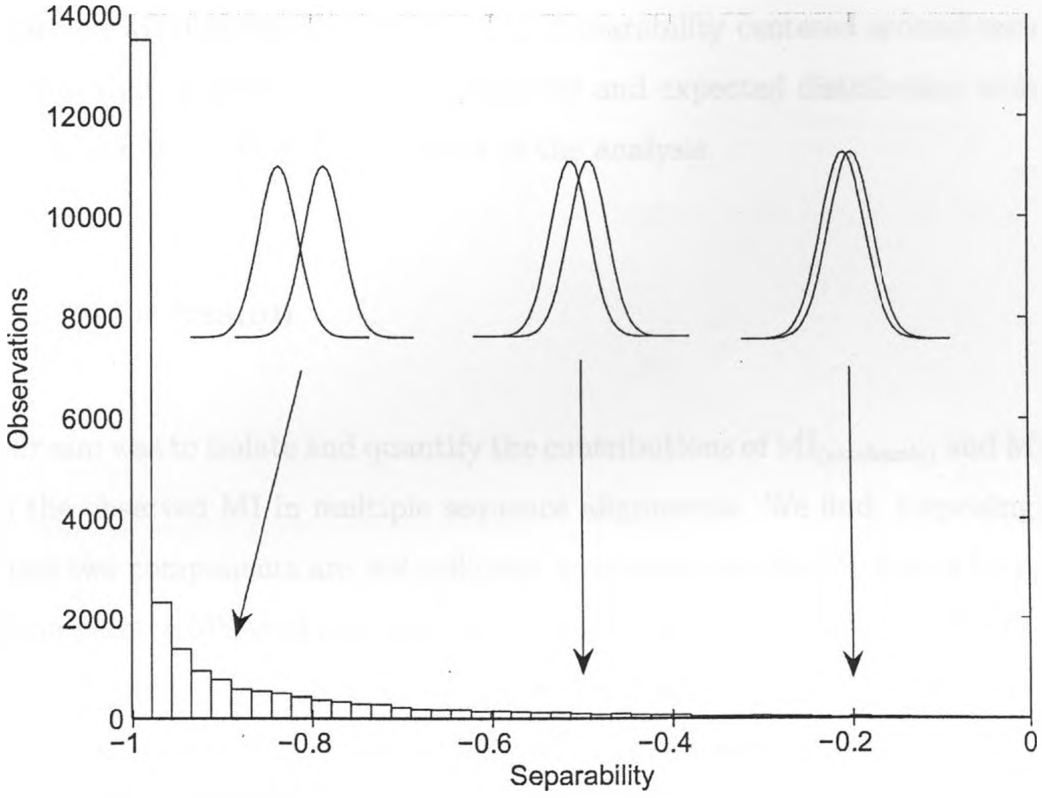

Figure 2.3: A histogram of the separability of expected and observed mutual information in TIM1. If the expected mutual information distribution is lower than the observed, this is represented by negative separability. Distributions with a great deal of overlap are represented near zero. Each of $\geq$10,000 ungapped site pairs undergoes 100 realizations of both observed and expected MI to calculate separability. Insets show theoretical examples of distributions with separability of -0.9, -0.5 and -0.2.

bottom right).

In contrast, the histogram in Figure 2.5 shows the separability when a MSA bootstrapped from TIM1 was used as a control (see Methods). Since the bootstrapped MSA contains no structural or functional constraints on pair frequencies, we expect, as shown in Figure 2.5, that the degree of overlap between the observed and expected MI distributions is very high (separability centered around zero). This verifies that differences between observed and expected distribution when using real-world MSAs were not artifacts of the analysis.

2.4 Discussion

Our aim was to isolate and quantify the contributions of $MI_{(stochastic)}$ and $MI_{(phylogeny)}$ to the observed MI in multiple sequence alignments. We find, surprisingly, that these two components are *not* sufficient to account for the MI shared by most position pairs in MSAs of real world protein families. In contrast with the prevailing wisdom that only a small fraction of position-pairs in a protein share functional or structural constraints (DUNN *et al.* 2008; LITTLE and CHEN 2009), this striking result suggests that almost all sites are co-varying, to some extent, for reasons beyond phylogenetic history. To relate this to the hypothesis of ATCHLEY *et al.* (2000) our results imply that either $MI_{(structure)}$ or $MI_{(function)}$ are shared by *most* position pairs in protein families, or that an additional component contributing to mutual information has not yet been described.

We estimated the distribution of expected Mutual Information via parametric bootstrap and multinomial sampling, accounting for shared phylogenetic history, amino acid similarity, rate heterogeneity and sampling variance. As an internal control, we also tested MSAs created by an evolutionary Markov process using the respective phylogenetic trees, at the site-specific rates inferred from the original

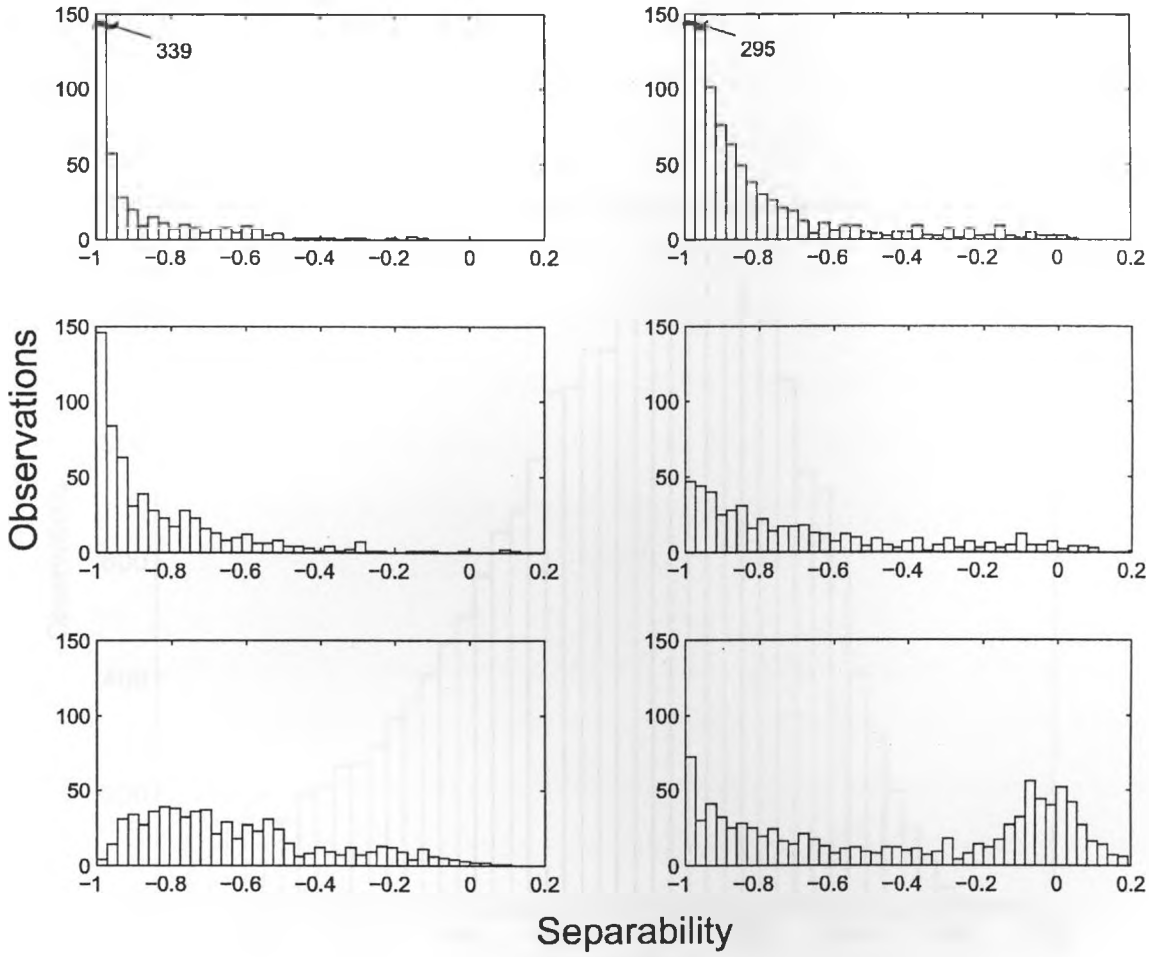

Figure 2.4: Separability histograms for six protein families. For each of $\geq$ 1000 site pairs per protein, 100 observed MI and 100 expected MI realizations were used to create distributions, as described in the text. The sequences are shown in order of increasing length with the shortest in the top left moving right and down to the longest in the bottom right. In that order, the protein names are: triosephosphate isomerase (TIM1); methionine aminopeptidase-1; dihydropteroate synthetase; phosphopyruvate dehydratase; phosphoglycerate kinase and; the GroEL chaperonin.

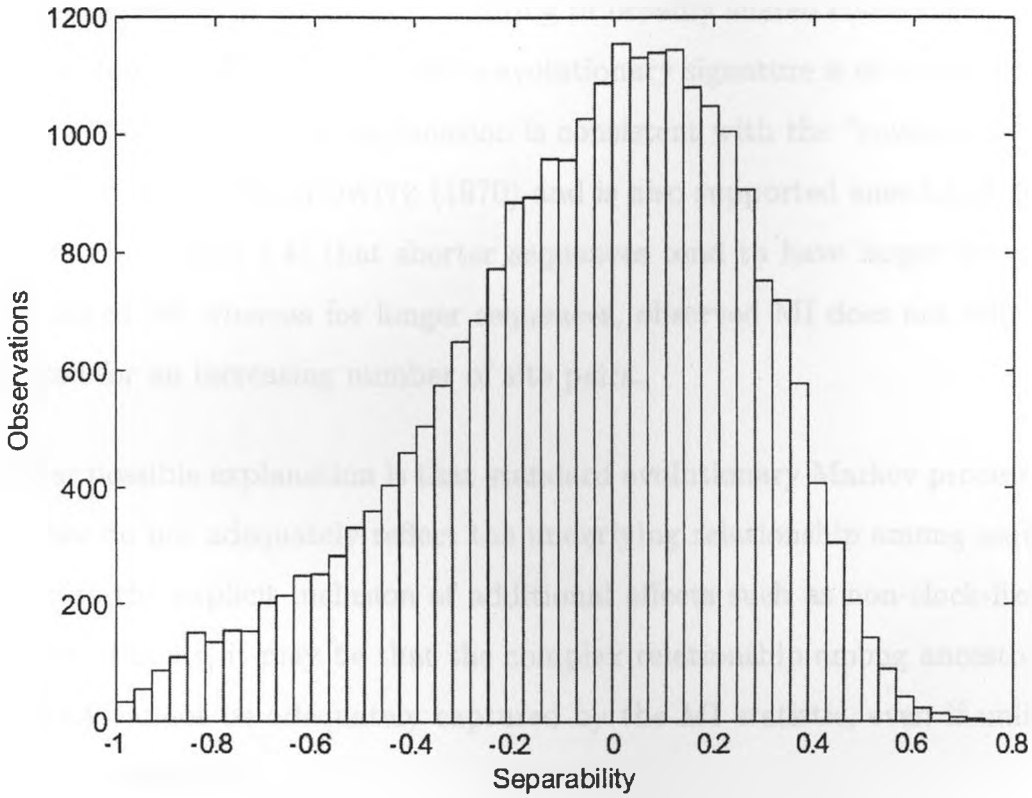

Figure 2.5: A histogram of the separability of expected and observed MI when bootstrapped data are used as input. Each of $\geq$ 10,000 ungapped site pairs undergoes 100 realizations of both observed and expected MI to calculate separability.

data. In these control cases we found that $MI_{(stochastic)}$ and $MI_{(phylogeny)}$ completely account for the measured MI. This disparity between control and real-world MSAs reinforces the conclusion that protein families contain broadly-shared contributions to MI that are not captured by standard, complex evolutionary models.

There are several possible explanations for this counter-intuitive result. The most straight-forward hypothesis is that local structural constraints "percolate" through sequences or structures, resulting in broadly shared co-evolution in real-world protein families. This broad co-evolutionary signature is of course absent in bootstrapped MSAs. This explanation is consistent with the "covarion hypothesis" of FITCH and MARKOWITZ (1970) and is also supported anecdotally by our observation (Figure 2.4) that shorter sequences tend to have larger amounts of unexplained MI whereas for longer sequences, observed MI does not differ from expected for an increasing number of site pairs.

Another possible explanation is that standard evolutionary Markov process models either do not adequately reflect the underlying relationship among sequences or require the explicit inclusion of additional effects such as non-clock-like substitution. Finally, it may be that the complex relationship among ancestor, tree and MSA cannot be adequately captured by the MI statistic, even if unlimited data were available.

A possibility for future investigation is to consider the effects of root location in the phylogenetic tree. In calculating MI from a MSA, each leaf is treated as equal; tree-unaware models share this feature. However, when incorporating evolutionary history, leaf nodes closer to the root have a larger effect on the probabilistic ancestor and thus a larger effect on the outcome of the expected mutual information. Although the rooting of phylogenetic trees is a notoriously complex question, future work involving sequence-weighting (BUSLJE *et al.* 2009) based on distance from the root might shed some light on the results described

here.

Although our work suggests that structural and functional MI may be more broadly shared in protein families than previously predicted, this by no means negates the use of sophisticated tools for identifying site-pairs which co-evolve most strongly. Ultimately, the characterization of factors contributing to MI should in fact facilitate the further development of these important methods.

REFERENCES

ATCHLEY, W. R., K. R. WOLLENBERG, W. M. FITCH, W. TERHALLE, and A. W. DRESS, 2000 Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol *17*(1): 164–178.

BERGER, J. O. and J. M. BERNARDO, 1992 Ordered Group Reference Priors with Application to the Multinomial Problem. Biometrika *79*(1): 25–37.

BUCK, M. J. and W. R. ATCHLEY, 2005 Networks of coevolving sites in structural and functional domains of serpin proteins. Mol Biol Evol *22*(7): 1627–1634.

BUSLJE, C. M., J. SANTOS, J. M. DELFINO, and M. NIELSEN, 2009 Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics *25*(9): 1125–1131.

CAPORASO, J. G., S. SMIT, B. C. EASTON, L. HUNTER, G. A. HUTTLEY, and R. KNIGHT, 2008 Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. BMC Evol Biol **8**: 327.

CLARKE, N. D., 1995 Covariation of residues in the homeodomain sequence family. Protein Sci *4*(11): 2269–2278.

CODOÑER, F. M. and M. A. FARES, 2008 Why should we care about molecular coevolution? Evol Bioinform Online **4**: 29–38.

DIMMIC, M. W., M. J. HUBISZ, C. D. BUSTAMANTE, and R. NIELSEN, 2005 Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics **21 Suppl 1:** i126–i135.

DUNN, S. D., L. M. WAHL, and G. B. GLOOR, 2008 Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics *24*(3): 333–340.

FARES, M. A. and S. A. A. TRAVERS, 2006 A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. Genetics *173*(1): 9–23.

FAWCETT, T., 2006 An Introduction to ROC Analysis. Pattern Recognition Letters *27*(8): 861–874. ROC Analysis in Pattern Recognition.

FELSENSTEIN, J., 2004 *Inferring Phylogenies.* Sunderland, Massachusetts: Sinauer Associates.

FERNANDES, A. D. and W. R. ATCHLEY, 2008 Site-specific evolutionary rates in proteins are better modeled as non-independent and strictly relative. Bioinformatics *24*(19): 2177–83.

FITCH, W. M. and E. MARKOWITZ, 1970 An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet *4*(5): 579–593.

FODOR, A. and R. ALDRICH, 2004 Influence of Conservation on Calculations of Amino Acid Covariance in Multiple Sequence Alignments. Proteins: Struct Funct Bioinf **56:** 211–221.

GLOOR, G. B., L. C. MARTIN, L. M. WAHL, and S. D. DUNN, 2005 Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry *44*(19): 7156–7165.

Goh, C. S., A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen, 2000 Co-evolution of proteins with their interaction partners. J Mol Biol *299*(2): 283–293.

Goh, C.-S. and F. E. Cohen, 2002 Co-evolutionary analysis reveals insights into protein-protein interactions. J Mol Biol *324*(1): 177–192.

Guindon, S. and O. Gascuel, 2002 Efficient biased estimation of evolutionary distances when substitution rates vary across sites. Mol Biol Evol *19*(4): 534–543.

Guindon, S. and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol *52*(5): 696–704.

Hoffman, N. G., C. A. Schiffer, and R. Swanstrom, 2003 Covariation of amino acid positions in HIV-1 protease. Virology *314*(2): 536–548.

Hogue, C. W., 1997 Cn3D: a new generation of three-dimensional molecular structure viewer. Trends Biochem Sci *22*(8): 314–316.

Hutter, M. and M. Zaffalon, 2005 Distribution of mutual information from complete and incomplete data. Comput Statist Data Anal *48*(3): 633–657.

Kim, Y., M. Koyutürk, U. Topkara, A. Grama, and S. Subramaniam, 2006 Inferring functional information from domain co-evolution. Bioinformatics *22*(1): 40–49.

Kim, Y. and S. Subramaniam, 2006 Locally Defined Protein Phylogenetic Profiles Reveal Previously Missed Protein Interactions and Functional Relationships. Proteins: Struct Funct Bioinf **62:** 1115–1124.

Korber, B. T., R. M. Farber, D. H. Wolpert, and A. S. Lapedes, 1993 Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A *90*(15): 7176–7180.

LARSON, S. M., A. A. D. NARDO, and A. R. DAVIDSON, 2000 Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol *303*(3): 433–446.

LASKO, T. A., J. G. BHAGWAT, K. H. ZOU, and L. OHNO-MACHADO, 2005 The use of receiver operating characteristic curves in biomedical informatics. JBI *38*(5): 404–415.

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

MANN, H. B. and D. R. WHITNEY, 1947 On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat *18*(1): 50–60.

MARCOTTE, E. M., M. PELLEGRINI, H. L. NG, D. W. RICE, T. O. YEATES, and D. EISENBERG, 1999 Detecting protein function and protein-protein interactions from genome sequences. Science *285*(5428): 751–753.

MARTIN, L. C., G. B. GLOOR, S. D. DUNN, and L. M. WAHL, 2005 Using information theory to search for co-evolving residues in proteins. Bioinformatics *21*(22): 4116–4124.

NEYMAN, J. and E. S. PEARSON, 1933 On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philos Trans R Soc Lond A *231*(694–706): 289–337.

OLIVEIRA, L., A. C. M. PAIVA, and G. VRIEND, 2002 Correlated mutation analyses on very large sequence families. Chembiochem *3*(10): 1010–1017.

Pazos, F., J. A. G. Ranea, D. Juan, and M. J. E. Sternberg, 2005 Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol *352*(4): 1002–1015.

Pazos, F. and A. Valencia, 2008 Protein co-evolution, co-adaptation and interactions. EMBO J *27*(20): 2648–2655.

Pollock, D. D., W. R. Taylor, and N. Goldman, 1999 Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol *287*(1): 187–198.

Pupko, T., R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal, 2002 Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics *18*(Suppl 1): S71–S77.

Robert, C. P., 2001 *The Bayesian choice: from decision-theoretic foundations to computational implementation* (2nd ed.). Springer texts in statistics. New York: Springer.

Saraf, M. C., G. L. Moore, and C. D. Maranas, 2003 Using multiple sequence correlation analysis to characterize functionally important protein regions. Protein Eng *16*(6): 397–406.

Shannon, C., 1948a A Mathematical Theory of Communication. AT&T Tech J **27:** 379–423.

Shannon, C., 1948b A Mathematical Theory of Communication. AT&T Tech J **27:** 623–656.

Süel, G. M., S. W. Lockless, M. A. Wall, and R. Ranganathan, 2003 Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol *10*(1): 59–69.

TILLIER, E. R. M. and T. W. H. LUI, 2003 Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics *19*(6): 750–755.

TUFF, P. and P. DARLU, 2000 Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol Biol Evol *17*(11): 1753–1759.

WADDELL, P. J., H. KISHINO, and R. OTA, 2007 Phylogenetic methodology for detecting protein interactions. Mol Biol Evol *24*(3): 650–659.

WECKWERTH, W. and J. SELBIG, 2003 Scoring and identifying organism-specific functional patterns and putative phosphorylation sites in protein sequences using mutual information. Biochem Biophys Res Commun **307**: 516–521.

WHELAN, S. and N. GOLDMAN, 2001 A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Mol Biol Evol *18*(5): 691–699.

WOLLENBERG, K. R. and W. R. ATCHLEY, 2000 Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci U S A *97*(7): 3288–3291.

Chapter 3

Summary and Future Work

This thesis assumes that the bootstrap procedure, described by Wollenberg and Atchley (2000), accurately reflects the impact of phylogenetic history on MI levels in MSA pairs. Additional research could be done into the effects of assumptions included in this model. One such consideration is that relative mutation rates are assumed constant. In the current model, mutation rates are averaged over the entire tree for each site. For example, if site 5 was determined to be twice as likely as site 3 to mutate, this would be true throughout the entire tree. It is possible, however, that for real world proteins, certain sites may have been more likely to mutate at certain times in evolutionary history.

In addition to fixed mutation rates, the creation of the probabilistic ancestor may require some reconsideration. When calculating the site probabilities for internal nodes, including the root, branch length determines how much the child node affects its parent. Leaf nodes close to the root, that is those with a short total branch length, will more strongly affect the root probabilistic ancestor. Further investigation into the effect of short-distance leaf nodes could be carried out, and possible weighting techniques could be developed to reduce this effect if needed.

A final bootstrapping consideration is the nature of internal nodes. Currently, all nodes except the root are totally cleared, then generated with data from the

root. By considering alternate methods which do not fully "erase" the probable amino acid identities at internal nodes, better estimations of historical internal nodes might be possible. This in turn may be used for better MI corrections related to phylogeny. Any such method, however, must be sure not to include any structural or functional constraints.

An additional consideration, unrelated to the bootstrapping already mentioned, is the removal of similar proteins in a MSA. Currently, if two sequences are more than 90% similar, one is removed. The weighting, instead of deletion, of these similar proteins might allow for more powerful statistics as the problems associated with small sample sizes could be reduced, as proposed by BUSLJE *et al.* (2009). Research into the full benefit of this idea, as it applies to this thesis, is needed. Sequence weighting would have at least two clear benefits. It would eliminate the need for an arbitrary similarity cut-off, currently set at 90%. Also, it would remove the random selection of which protein to keep when two or more are found to be similar.

In the area of software, the simulation software is currently not intuitive to use. Changing parameters, such as the number of MI points in each distribution, is done by changing the code and recompiling. The data output has minimal formatting or organization, requiring additional programming, currently done in Matlab, to make it useable. The code is relatively modular, and uses pre-existing optimization libraries; however, additional comments and an intuitive user interface could greatly improve accessibility.

Migration onto SharcNet would also be beneficial since analysing a protein with 320 sites, bootstrapping 100 times for each pair, takes upwards of 3 hours on a dual core, 1.6GHz processor. As proteins get larger, the number of pairs to analyse increases by $O(n^2)$, which has already limited the length of strand we can fully analyse.

Overall, an improved understanding of MI contributions, and possible subsequent improvements to background MI corrections, may lead to better identification of co-evolving site in protein families. Assisting in this eventual improvement of such methods is, in a way, the long-term goal of this entire work. In particular, a better understanding of MI component contributions may assist in the isolation of pairs, and groups, which share MI for structural or functional reasons. In the long term, the ability to better identify these locations in sequences of unknown function will assist in our understanding of the complex mechanisms that exist in proteins.

REFERENCES

Buslje, C. M., J. Santos, J. M. Delfino, and M. Nielsen, 2009 Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics *25*(9): 1125–1131.

Wollenberg, K. R. and W. R. Atchley, 2000 Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci U S A *97*(7): 3288–3291.