DECOMPOSITION OF PROTEINS INTO FUNCTIONALLY AND EVOLUTIONARILY INDEPENDENT COOPERATIVE UNITS

APPROVED BY SUPERVISORY COMMITTEE

Rama Ranganathan

Richard Auchus

Johann Deisenhofer

Kevin Gardner

dedicated to my Mom and Dad

DECOMPOSITION OF PROTEINS INTO FUNCTIONALLY AND EVOLUTIONARILY INDEPENDENT COOPERATIVE UNITS

by

NAJEEB MAROOF HALABI

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, TX

Dec. 16, 2008

DECOMPOSITION OF PROTEINS INTO FUNCTIONALLY AND EVOLUTIONARILY INDEPENDENT COOPERATIVE UNITS

NAJEEB MAROOF HALABI

The University of Texas Southwestern Medical Center at Dallas, 2008

RAMA RANGANATHAN, M.D. / PH.D.

Understanding cooperative interactions within proteins is an important goal because cooperativity underlies protein functions such as catalysis, allostery, ligand binding and folding. Cooperative units within proteins can be revealed via an evolution based method called statistical coupling analysis that quantifies the correlations between positions in a multiple sequence alignment of a protein family. In this work, coupling analysis and experimental studies were used to analyze two protein families – the TonB dependent receptors and the S1A serine proteases.

The TonB dependent receptor family members are membrane bound siderophore transporters. Ligand binding on the extracellular side of the transporter transmits a signal to the periplasmic side where another protein (TonB) provides the energy for transport. Coupling analysis, based on a diverse alignment of 541 family members revealed a network of physically contiguous positions extending close to 50 angstroms from the ligand binding pocket to the putative periplasmic interaction sites. A

mutational analysis of FecA, a representative member of this family, confirmed the functional significance of the coupled positions.

The S1A serine protease family members are involved in diverse functions such as digestion, coagulation, immunity and reproduction. Coupling analysis on 1470 serine proteases revealed at least three sets of independently evolving positions. Each set of positions is called a sector. Structural analysis revealed that each sector is physically contiguous suggesting mechanical independence. Extensive data in the literature on this protein family allowed the assignment of function to two of the sectors. One sector comprised positions making up the catalytic machinery, while another sector comprised positions important for substrate binding. To determine the function of the sector with unknown function, mutations were done on positions making up this sector and tested for catalytic and stability effects on proteins. The data showed that the positions in the unknown function sector affected stability but not catalysis. Each sector therefore performs a different and independent cooperative function: one sector for catalysis, one for substrate binding and one for fold stability.

Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

SCA: Statistical Coupling Analysis

PCA: Principal Component Analysis

ICA: Independent Component Analysis

Chapter 1: Protein cooperativity

This chapter is an introduction to the cooperative nature of proteins and the methods available to observe this cooperativity. It also explains why cooperative units of proteins are observed in evolution based methods, a subject that will be explored further in subsequent chapters.

Section 1: Cooperative protein behavior

Proteins do many amazing things: catalyze reactions in milliseconds that would otherwise not occur in geological time, bind ligands with atomic specificity and avidity, manipulate single atoms, harness single photons and adopt specific three dimensional shapes. These functions all occur while the protein is immersed in liquid water or in lipids or both and while constantly vibrating with thermal energy. Although understanding these protein properties has been the aim of decades of structural, biochemical and theoretical investigation, it is not yet possible to choose an arbitrary function and design an amino acid sequence that will carry out that function.

A reason why a deep understanding of proteins has been elusive is because it is not yet possible to easily observe or account for the cooperative interactions within a protein's atoms [1]. In other words, while the sequence and structure of proteins are often known, the functional units are not because the functional units are made up groups of amino acids that work together in as yet mysterious ways. As I will explain below, many protein functions depend on or exhibit cooperativity. Catalysis, ligand binding, allostery and folding all arise from cooperative interactions between multiple amino acids. In addition, the dynamic nature of proteins is a cooperative one. Because cooperativity plays such an important role in protein function, it is important that methods to measure, observe and understand it are developed.

Enzyme catalysis

Enzymes are biological catalysts. The degree to which enzymes increase the rate of chemical reactions can be estimated by calculating the rate acceleration which is the ratio of the rate of the catalyzed reaction to the rate of the uncatalyzed reaction. The rate acceleration for some enzymes is shown in Table 1 (adapted from two of Richard Wolfenden's papers [2, 3]). The rate acceleration ranges from 10^5 to 10^{17}. Some enzymes also exhibit diffusion-limited or 'perfect' second order rate kinetics; these enzyme rates are limited only by time it takes substrate to bind to the enzyme (on the order of 1-$10x10^9$ $M^{-1}s^{-1}$). The high second order rate constant of 'perfect' enzymes arises from an optimal arrangement of atoms to carry out the chemical reaction and the attraction of the substrate to the

active site via charged interactions [2]. Enzymes also display high specificity for a particular substrate. They do not, for example, catalyze the same reaction using the enantiomeric form of a substrate, which is a problem that synthetic chemists cannot easily solve.

The specificity, rate acceleration and chemical efficiency of enzymes have been the subject of much

Table 1: Rate acceleration (ratio of catalyzed to uncatalyzed rate) for some enzymes (adapted from Radzicka and Wolfenden, 1995).

ENZYME	RATE UNCATALYZED (s^{-1})	RATE CATALYZED (s^{-1})	RATE ACCELERATION
OMP decarboxylase	$2.8 * 10^{-16}$	39	$\mathbf{1.4x10^{17}}$
Staphylococcal nuclease	$1.7 * 10^{-13}$	95	$\mathbf{5.6x10^{14}}$
Adenosine deaminase	$1.8 *10^{-10}$	$3.7*10^{2}$	$\mathbf{2.1x10^{12}}$
AMP nucleosidase	$1.0 *10^{-11}$	60	$\mathbf{6.0x10^{12}}$
Phospotriesterase	$3.2 *10^{-10}$	$2.99*10^{2}$	$\mathbf{1.2x10^{12}}$
Cytidine deaminase	$7.5*10^{-9}$	$2.1*10^{2}$	$\mathbf{2.8x10^{11}}$
Carboxypeptidase A	$3.0*10^{-9}$	$5.78*10^{2}$	$\mathbf{1.9x10^{11}}$
Ketosteroid isomerase	$1.7*10^{-7}$	$6.6*10^{4}$	$\mathbf{3.9x10^{11}}$
Serine proteases	$0.2*10^{-9}$	50	$\mathbf{1x10^{10}}$
Triosephosphate isomerase	$4.3*10^{-6}$	$4.3*10^{4}$	$\mathbf{1.0x10^{9}}$
Chorismate mutase	$2.6*10^{-5}$	50	$\mathbf{1.9x10^{6}}$
Carbonic Anhydrase	$1.3*10^{-1}$	$1*10^{4}$	$\mathbf{7.7x10^{6}}$
Cyclophilin, human	$2.8*10^{-2}$	$1.3*10^{4}$	$\mathbf{4.6x10^{5}}$

investigation. It was thought at least since 1921 (reviewed in [3]) that an enzyme binds with strong affinity to the transition state of the reaction relative to that of the substrate. Figure 1 shows a schematic of how an enzyme catalyzes a reaction relative to the uncatalyzed reaction with the formation of the transition state being the highest energy point in the reaction scheme. The central role of the transition state has been largely accepted today [4-6]. Creation of catalytic antibodies by using as an antigen a molecule thought to mimic the transition state of a particular reaction is consistent with

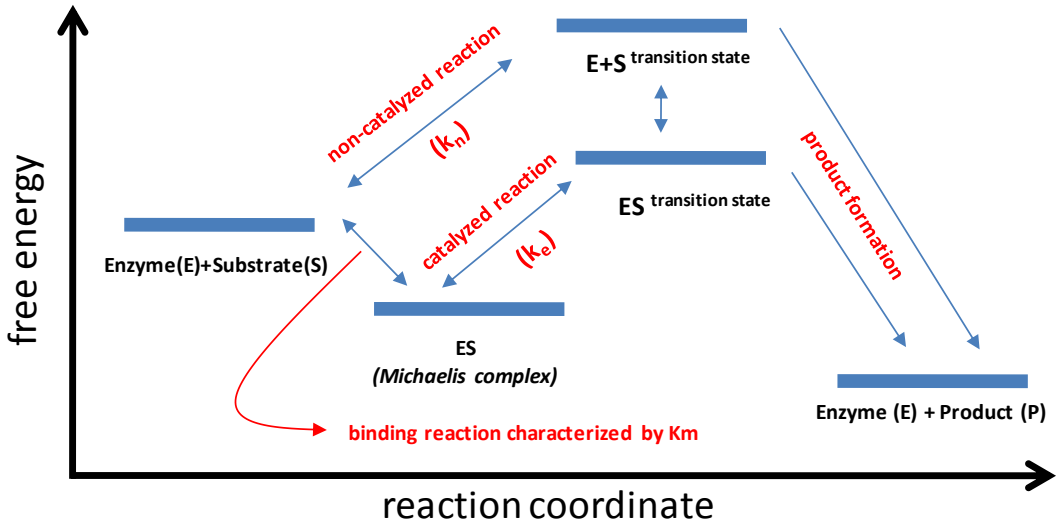
this model for enzyme function [4]. The role of the transition state is expressed in the following formula for the rate of a reaction (*k*) [5]:

***k*=(transmission coefficient) x (barrier crossing frequency) x (equilibrium constant of transition state formation)**

The transmission coefficient (tc) is a correction factor for tunneling, barrier recrossing and solvent friction that is usually between 0.1 and 1. The barrier crossing frequency (v) is the frequency of oscillation along the reaction coordinate and is often approximated by k_BT/h but could differ with different transition states. The equilibrium constant for transition state formation is the concentration of the transition state divided by the product of the concentrations of enzyme and substrate.

One consequence of transition state theory, first worked out by Wolfenden [6] is that the enzyme must bind the transition state with a binding affinity proportional to the enzymatic rate acceleration. To see this, a thermodynamic cycle can be written based on the scheme in Figure 1.

Figure 1: Scheme for an enzyme catalyzing a reaction going through a high energy transition state. The double tipped arrows represent an equilibrium reaction and the single tipped arrows represent an irreversible reaction.



To work out how the thermodynamics leads to an understanding of the rate acceleration, two assumptions of transition state theory are needed. The two assumptions are:

1. The limiting step is the rate at which the transition state decays to the product.
2. The enzyme-transition state complex is in equilibrium with the enzyme substrate complex.

One can now write an expression using the transition state equation for a catalyzed (e) and uncatalyzed reaction (n):

$$k_e = tc_e v_e K_e^{\text{transition state}}$$
$$k_n = tc_n v_n K_n^{\text{transition state}}$$

The ratio of rates can now be written as:

$$\frac{k_e}{k_n} = \frac{tc_e v_e K_e^{\text{transition state}}}{tc_n v_n K_n^{\text{transition state}}}$$

And because of the equilibrium assumption, a thermodynamic cycle can be written so that the ratio of the substrate binding equilibrium (K_m) to substrate transition state equilibrium (K_{ts}) equals the ratio of activated substrate equilibrium ($K_e^{\textit{transition-state}}$) to enzyme- substrate equilibrium ($K_n^{\textit{transition-state}}$). That is:

$$\frac{K_{ts}}{K_m} = \frac{K_n^{\text{transition state}}}{K_e^{\text{transition state}}}$$

Substitution of K_m and Kts back into the rate expression k results in:

$$\frac{k_e}{k_n} = \frac{tc_e v_e K_m}{tc_n v_n K_{ts}}$$

Now if the ratios of *tc* and *v* are close to one then this means that the ratio of catalyzed to uncatalyzed rates (the rate acceleration) must be close to the ratio of substrate binding to transition state binding. Since the rate acceleration data indicate that the rate acceleration is very high then that means that the enzyme binds the transition state more than the substrate by a factor of 10^5-10^{17} (depending on the enzyme).

The question now becomes one of trying to understand how an enzyme binds a transition state so tightly. The answer is that the enzyme binding to the transition state are a result of the cooperative behavior of many enzyme atoms or amino acid residues so that the transition state is precisely bound [3, 6-8]. In chapter five I will discuss a specific example of an enzyme utilizing five different residues to bind the transition state.

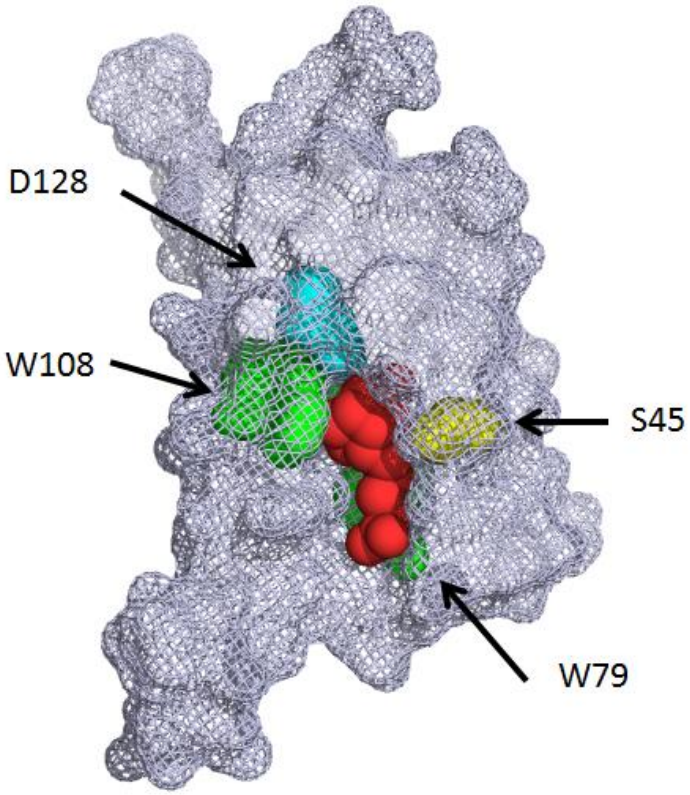
Ligand binding

Proteins are capable of binding other proteins as well as an extremely diverse array of non-protein molecules. Protein ligand binding is often characterized by high specificity and high affinity. These

properties are also mediated by cooperative protein interactions. As an illustrative example, I will consider two examples of proteins binding to ligands.

The first example is streptavidin binding to biotin which has one of the highest affinities found in nature (dissociation constant on the order of 10^{-15} M) [7, 8]. Streptavidin is a tetrameric protein with each subunit consisting of 133 residues. Biotin is a small molecule more commonly known as vitamin H or Vitamin B7. The structure of streptavidin [8] together with extensive mutagenesis and binding energy measurements [8-10] demonstrates that biotin is held tightly by interactions among at least six residues widely distributed in the structure and with different biochemical properties (four tryptophans, a serine and an aspartate) (see Figure 2). Thus, the biotin is held via a network of hydrogen bonding and van der Waals forces.

Figure 2: Structure (1STP) of streptavidin bound to biotin (in red). Four of six residues that have been shown to interact with biotin are in color. Note that this structure is only one monomer of the streptavidin native tetrameric structure.



Another example of a protein bound to its ligand is human growth hormone bound to its receptor. The structure of this complex reveals as many as 30 amino acids at the binding interface [9]. However, extensive mutagenesis revealed that just two residues contribute 75% of the binding energy [9], a phenomenon seen in other protein families as well [10]. What these works suggest is that although

structurally many positions are observed contacting the ligand, it is possible that only a subset of those make strong interactions with the ligand.

Allostery

The term allostery was originally used to refer to the interactions between two ligand binding sites within a protein [11]. The observation made in Jacques Monod and Francois Jacob's lab by their student Pierre Changeaux, that the product of an enzyme inhibited the enzyme by binding at another site on the same enzyme [12] prompted their investigation into allosteric effects in proteins. Subsequently, it was discovered that allostery was common in proteins. One of the most commonly studied examples is the binding of oxygen to hemoglobin. Hemoglobin is composed of two pairs of subunits and binds four oxygen molecules with the binding of one oxygen progressively increasing the affinity of the other subunits for oxygen [11].

In general allosteric systems can occur where one ligand affects the binding of the same ligand at a distant site (homotypic) or of a different ligand (heterotypic). There could be positive cooperativity where the binding of a ligand increases the affinity for another ligand and negative cooperativity where the binding of a ligand decreases the affinity for another ligand [11]. In hemoglobin the cooperativity is homotypic and positive.

Allosteric effects are typically detected by changes in the slope of binding curves (graphs of concentration of substrate against fraction bound). Scatchard plots or more commonly Hill plots are used. Two classical models have also been proposed to explain how cooperativity could occur. One model developed by Monod, Wyman and Changeaux [13] (MWC model) is called the concerted model and another developed by Koshland, Nemethy and Filmer [14] (KNF model) is called the sequential model. The models have implications about the molecular nature of interaction between sites. The sequential model postulates that one binding site directly contacts the other binding site whereas the concerted model postulates that one binding site changes the form of the subunit in which the second binding site exists. The MWC model implies that a system has two states (called the tense and relaxed states), whereas the KNF model suggests that there are many states along the path from one binding site to the other. Different models could apply to different proteins. In the case of hemoglobin two distinct states are observed structurally which provides support for the MWC model These two models, introduced in the 1960's, have been continuously revised as additional data are collected [10, 12].

Irrespective of the details of the specific model, both agree that an energy/information transfer pathway exists in allosteric proteins. By definition this involves the cooperation of several residues to

transfer this information [15]. Identifying these networks of sites is very difficult to do with current methods.

Folding

Folding is the process by which a long polypeptide chain adopts a specific tertiary structure. For many small proteins an idea known as the thermodynamic hypothesis applies to this process [16]. The thermodynamic hypothesis means that the native state is the free energy minimum of a protein. Experimentally the thermodynamic hypothesis is supported for many proteins by the observations that a protein can be denatured by temperature or chemical denaturants but then recover its structure and function when the denaturants are removed.

Though the thermodynamic hypothesis suggests that the protein fold is thermodynamically stable (exceptions to this have been discovered-for example in alpha lytic protease [17]), there are still unresolved problems with the understanding of protein folding. One problem is to understand the speed of folding. Another problem is to understand the forces in the protein-solvent system that lead to folding. These problems are not mutually exclusive.

The problem of explaining the speed of folding was reported by Cyrus Levinthal in 1968 [18] and is known as Levinthal's paradox. Levinthal calculated the number of conformations an unfolded polypeptide can adopt and estimated how long it would take the protein to adopt one conformation (the folded state conformation, for example). The number of conformations can be calculated by assuming that an amino acid can adopt one of three conformations in an unfolded state. Then if a protein has 100 amino acids in a polypeptide, the total number of conformations is 3^{100}. The maximal rate of folding would be limited to the bond vibration frequency which is on the order of 10^{13}/second. That means that a random search through all the possible conformations would take something on the order of 10^{27} years. Since proteins fold on time scales of microseconds to seconds [19], the hypothesis that a protein samples conformations randomly or has a large number of conformations to sample cannot be correct. Thus there must be pathways along a reaction coordinate the lead to folding or that the energy landscape of protein folding is shaped like a funnel [20, 21].

The problem of understanding the nature of the forces that lead to a folded protein has received much attention in the literature. Folded proteins adopt a very compact, dense, glass-like, almost uncompressible state [22-25]. In this compact state, there are many possible interactions that could promote folding (such as hydrogen bonding, ionic interactions, van der Waals forces, local interactions, distant interactions) and there are interactions that oppose folding (reviewed in [26]). It is estimated that the hydrophobic effect accounts for most of the energy of folding although other interactions can

contribute as well under certain circumstances. The hydrophobic effect can be described as the way that the hydrophobic residues in a protein partition away from water; the partitioning is maximized when these hydrophobic residues interact with each other forming the highly compact protein structure. The energetic interactions that oppose folding are entropic in nature: the folded state of a protein strictly reduces the conformations accessible to the protein. The balance between the hydrophobic effect and entropic effects results in the marginal stability of a protein [26].

The hydrophobic effect, while explaining the compact nature of a protein, does not explain the fine structure of a protein. A protein, for example, has alpha helices and beta sheets in specific orientations. It is thought that hydrophobic collapse is followed by the establishment of specific interactions between elements of the structure through hydrogen bonding, ionic interactions and van der Waals forces [26, 27].

Regardless of the exact nature of folding, the available evidence indicates that protein folding is extremely cooperative. Multiple interactions in the core of a protein are required for the hydrophobic effect to overcome the entropic penalty [28, 29]. Furthermore, specific aspects of structures such as helixes and sheets are defined as groups of residues. Evidence for the cooperative nature of folding is also seen by the temperature dependence of folding for small proteins which show a single sharp transition indicating that the protein folds and unfolds as a unit [30-32] . Thus an understanding of cooperativity in protein folding is important for a complete picture of the protein folding process.

Dynamics

The protein functions discussed above all occur through cooperative interaction although the molecular basis for these cooperative interactions is as yet unknown. One idea that may unify all these disparate functions comes from the low frequency dynamic nature of proteins. As Richard Feynman wrote , "...everything that living things do can be understood in terms of the jigglings and wigglings of atoms." [33]. The protein atoms, like all atoms above absolute zero, vibrate at fast time scales (0.1 ps) reflecting rapid bond fluctuations. However, there is another aspect of proteins which is that they also possess, at physiological temperature, slow (microsecond to millisecond) collective motions [34, 35] . These slow collective motions occur only at temperatures above 180-220 K which correlates with the temperature at which protein function first occurs [36, 37]. The correlation of these observed motions with function is striking. In one specific example in bacteriorhodopsin, the photocycle was examined below and above 220 K and it was observed that while light absorption could occur below 220 K the cycle could not be completed at that temperature. To paraphrase the authors, the enzyme was in effect frozen in the middle of the photocycle due to uncoupling of the protein from thermal motion [36]!

Once the temperature was raised the bacteriorhodopsin completed the photocycle and could be stimulated again. In another illuminating study from Gregory Petsko's lab [37] a protein crystal was studied with x-ray crystallography for more than one month while the temperature was varied over the 180-220 K transition while different aspects of function were tested. As in the bacteriorhodopsin case, at low temperatures an inhibitor could be trapped in the enzyme once the temperature was lowered below 180 K (the binding was done at higher temperature). It was also observed that catalysis did not occur with native substrate at low temperatures.

More evidence for the occurrence and importance of slow collective motions comes from work done using NMR to look at the motions of most residues in several protein systems [38-42]. NMR has an advantage over other methods in that the motions of individual atoms can be resolved. In one study from Peter Wright's lab [42], using the enzyme DHFR, the authors measured the conformational changes for many intermediates along the reaction pathway (including free enzyme, substrate bound and product bound). One result is that the enzyme without substrate shows conformational change in the substrate binding pocket, suggesting that enzyme samples conformations that would bind the substrate. The authors also find that while product is bound, the regions in the product binding site show motions as if the product binding site is empty. Another interesting result is that the rate of conformational change is correlated with the catalytic rate (specifically the rate of conformational change of the product binding region is close to the rate limiting step of the reaction which is product release) . Another study from Dorothee Kern's lab [39] on adenylate kinase presented evidence showing that positions that were moving on the ps-ns timescale were in the same regions that were involved in hinge motion thus providing a link between fast time scale motions and slow time scale motions.

There are two general points to make about the dynamic properties. First, that slow scale dynamics are cooperative as they require the synchronized movement of many atoms and residues. Second, this dynamic, cooperative behavior may contribute to or be the cause of other cooperative behavior in proteins such as catalysis, ligand binding, folding or allostery. For example, one notable study [43] measured using NMR the dynamics of a protein with and without ligand and related the dynamics to the conformational entropy and they found that the conformational entropy strongly contributes to the ligand binding energetics. Another important study [44] measured, using neutron inelastic scattering, the dynamics with and without the binding of a ligand to the dihydrofolate reductase enzyme and observed a change in dynamics to a softer, more flexible protein upon ligand binding; the dynamic change was estimated to contribute a factor of 1000 to the binding constant.

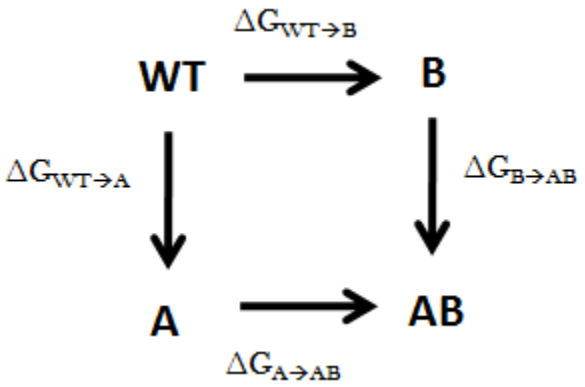
Studies of this type will lead to a greater understanding of the relationship between cooperative dynamics and other cooperative behavior in proteins.

One aspect of the previous methods is that while they are able to detect cooperative interactions, they cannot reveal the magnitude and nature of interactions between the cooperative units. That is, one cannot say how much a particular atom contributes energetically to the function. In the next section, I will discuss a method known as thermodynamic mutant cycles that can be used to examine the interaction between sites.

Section 2: Measuring the magnitude of cooperativity : thermodynamic mutant cycles

One powerful way to uncover the cooperativity of amino acids is through mutational analysis. Specifically, double (or more than double) thermodynamic mutant cycles [45-48] can reveal whether two mutated positions interact and if so by how much they interact. The simplest thermodynamic mutant cycle consisting of three sets of mutations is depicted in Figure 3. The basic idea is that one can

Figure 3: Thermodynamic cycle scheme. WT is the native protein. A refers to a mutation and B refers to a mutation different from A. AB is both mutations.



calculate whether two positions interact in a protein by making a double mutant and each single mutant. So, referring to Figure 3, one can make a mutation A and a mutation B and then measure the free energy change of the mutation. Then one makes the double mutation AB and measures the free energy change. The free energy change can be measured for any function-catalysis, ligand binding, folding or any other measurable function. Then one compares the sum of the free energies of the single mutants to the actual measured free energy of the double mutant. There are two possibilities to consider: the double mutation can be the sum of the two single mutants or the double mutation can be different from the sum of the two single mutants. (In real experiments, the error of the measurement must be considered when deciding whether the summed free energy is different from the measured double

mutant free energy). In other words, the interaction between positions A and B can be additive or non additive.

If the effect is additive then one likely conclusion is the sites act independently. This is because free energies for independent processes are additive. On the other hand, if the effect is non-additive then one can consider that the two positions in the protein interact together to carry out the measured function.

One assumption of thermodynamic mutant cycles is that the interaction of the mutations, which is what is being measured, is similar to that of the interaction of the amino acids in the native state. This assumption can easily be wrong as the following thought experiment shows. Imagine replacing positions A and B of a protein have two valines and what one would like to determine is if those valines are interacting. Now imagine if the valines are mutated to cysteines and that the valines are adjacent in space. When the double mutant is formed it will form a disulfide bond in reducing conditions. Now if the interaction being studied is in the context of fold stability of the proteins, one would conclude that those two positions interact to a large extent because the fold stability change of the double mutant is much greater than the sum of the stability changes of the single mutants. This conclusion would be reached regardless of the interaction of the valines in the native structure. Granted this thought experiment requires the residues to be adjacent and mutated to cysteines and the assay be fold stability but it is used to illustrate the assumption underlying thermodynamic mutant cycle conclusions. It is possible to perform control experiments to test whether the protein is drastically perturbed when the mutations are made but in general, without some independent way of verifying the interaction of the native amino acid at positions A and B there is no certainty that the interaction effect exists in the native state.

One controversy that I will briefly address here that has occurred in the literature that applies to thermodynamic mutant cycles is whether the individual components that give rise to energy can be parsed out from the energy term [1, 49]. That is, if one is given an entropy or enthalpy or free energy change, can one conclude that this change is due to a particular hydrogen bond interaction or van der Waals interaction or charge-charge interaction? The short answer is that one cannot because the free energy is a function of the state of the entire system and these states can only be considered as additive if they are shown to be independent [49]. But, as I have shown, for many protein functions, many functional aspects are cooperative and therefore not independent. Thus the free energy changes measured during thermodynamic mutant cycle experiments cannot be used to say whether specific interactions account for the energetic change. In some cases, one can make a reasonable case that the

entropy changes are minimal and therefore conclude that the enthalpy of a particular bond accounts for the energy change. However, it is important to keep in mind that the free energy, in general, is a property of a system and decomposing a system into subsets of interactions is possible only if the subsets are independent.

What has been learned about proteins from thermodynamic mutant cycles? The first report in 1984 of the use of this method from Alan Fersht's lab [46] reported how a mutation at one position of a tRNA synthase affected the strength of another position to the ATP substrate. Since then many studies reviewed in 1990 by Wells [50] and in 1995 by LiCata and Ackers [51] reveal an interesting protein energetic architecture. The earlier review by Wells found that most mutations were additive whether the interactions were protein binding to protein or protein binding to DNA or protein folding. Interactions that were not additive were due to positions that were in direct contact or change the structure or electromagnetic field or were cooperative in their function (as in enzyme active sites).

The later review disagreed somewhat with the 1990 review. LiCata and Ackers specifically looked for non-additivity between sites that were far enough away that they could not be contacting. They then observed that greater than 50% of interactions in several different systems showed small but significantly non-additive interactions. These small non-additivities were considered as additive by previous researchers. The authors also state that these small non-additivities were almost always positive indicating that they were real because if these small non-additivities were due to noise then there would be negative additivities as often as positive additivities. The authors conclude by stating that the mechanism of propagation could be due to long-range electrostatics, long-range structural perturbation or dynamical effects. The authors also warn that the implications of non-additivity should be considered by computational methods as those computations generally assume that distant sites are non-additive. Thus thermodynamic mutant cycles show that proteins have a highly heterogeneous structure with some positions being cooperative (non-additive) and others not cooperative (additive). Also, cooperativity can occur between physically non contacting sites through several possible, though generally unknown, mechanisms.

From the perspective of protein physics, the implications of thermodynamic mutant cycles is that they reveal and quantify the cooperativity between positions. They are one of the few readily accessible methods that can do so (one must keep in mind the assumptions stated before). However, to understand the interaction of one site with another requires making all possible pairs of mutations between all positions and assaying each one for function. Such a task has not been done due to the number of mutations and assays needed. Moreover such pairwise mutations reveal only pairwise

additivities or non-additivities; it is possible to have cooperative interactions between three or more positions and general approaches for elucidating those are not available (although Yifrach's group has studied higher order coupling for several sets of residues in the potassium channel [52, 53]) . It is therefore important to seek alternative ways to look for cooperativity between protein positions.

Conclusion

This chapter presented some of the available evidence for the role that cooperativity plays in proteins. Cooperative interactions are observed in catalysis, ligand binding, allostery and folding. Furthermore, the very nature of a protein seems to be a cooperative one with the occurrence of slow time scale cooperative motions.

One problem with most available methods that observe cooperativity is that most cannot measure the energy of the cooperative interaction. For example, an observation that an enzyme substrate is bound by several residues cannot determine what the relative strength of each interaction is. Experimentally, thermodynamic mutant cycles have shed light on the energetic nature of the interactions but these experiments are both difficult to conduct and make assumptions that can be difficult to justify. Therefore, the development of new methods to observe cooperative interactions and to estimate their strength is desirable.

One relatively new method that has been explored in the Ranganathan lab and others to detect cooperative units in proteins is based on the evolution of a protein family. The rationale for this evolution-based method is that residues that cooperate for function are likely to exhibit mutual conservation whereas residues that do not cooperate with any other residues will have little, if any, mutual conservation. Therefore, if a protein family with members that have the same function has a distribution of conservation at different positions it is likely that the residues with a similar conservation pattern cooperate together. In other words, an analysis of conservation patterns based on a multiple sequence alignment of a protein family can reveal cooperative units. The use of one evolution based method, called statistical coupling analysis (SCA), for the analysis of two protein families is the subject of this dissertation. I will show, in conjunction with experimental evidence, the use of SCA to identify a possible allosteric pathway in the TonB dependent receptor family and to identify three independent cooperative units of function in the serine protease family.

Chapter 2: Calculating Covariance Between Positions in Multiple Sequence Alignments

Cooperativity is an important determinant of protein function that is hard to reveal experimentally. Our lab and others have developed evolution based methods to detect cooperativity. These methods are discussed and compared in sections 1 and 2 of this chapter.

The underlying idea behind evolution based methods to detect cooperativity is that sites that interact together for a certain function will be less likely to tolerate mutations than sites that act independently. One could say that those positions in the protein are evolutionarily constrained. The reason why sites that act cooperatively are likely to mutate less is because if a mutation at a cooperative site occurs it is much more likely to be deleterious to the function because of its interactions with other sites. Therefore, the organisms that carry the protein with reduced functionality will likely be selected against (unless further mutations that restore function occur before the population carrying the deleterious mutation becomes extinct). Independent sites, on the other hand, are less likely to result in deleterious effects due to the fact that mutations in them are not likely to disrupt other positions in the protein.

Evolution based methods exploit the constraints that cooperativity imposes on protein evolution. There are two consequences of cooperative constraints that can be analyzed. The first is that cooperative constraints will lead to conservation in a multiple sequence alignment at positions whose amino acids mediate the cooperative interaction. The second consequence of cooperative constraints is that a correlation between the conservation at one position and the conservation at another position should exist if two positions are cooperative. Analyzing the conservation and correlation is possible using different statistics based methods.

One complicating factor with this analysis of conservation and correlation is that there is another source of correlation among positions. The source is the historical relationship between the sequences arising from the common ancestors that different sets of sequences have. This historic source of correlations is often called historical or phylogenetic noise and it can greatly affect the use of correlations to detect correlations coming from cooperative interactions. In practice this problem is minimized by using as phylogenetically diverse an alignment as possible and by using algorithms that take the effect of phylogeny into account as will be explained in the following sections.

Section 1: Statistical Coupling Analysis:

Our lab has developed three related methods to calculate coupling between amino acids in a multiple sequence alignment (MSA). These methods will be referred to as the weighted covariation method, the bootstrap method and the perturbation method. All three methods are related together through a common framework called statistical coupling analysis (SCA) which will be described.

There are two aspects of SCA that differentiate it from other methods. First, SCA uses the binomial distribution to calculate the probability of a particular amino acid frequency at a position given a multiple sequence alignment. Second, SCA implements methods to downweight correlations that could come from historical noise and upweight correlations that are thought to arise from cooperative constraints. These will be discussed in the following section.

The background frequency

One requirement for statistical coupling analysis is to be able to quantify how probable the frequency of an amino acid is. The question is if there is an amino acid *a* at a position *i* present at a frequency of 10%, then how does one calculate if that frequency is low, high or average? To answer this problem requires knowledge of how frequent a particular amino acid is in nature. The frequency of occurrence of an amino acid is called the background frequency.

The background frequencies are calculated as the average frequency of amino acids from all proteins in NCBI's non-redundant database. The background frequencies used in all of the statistical coupling work are shown in Table 2. The same background frequency is used to analyze every protein family based on the assumption that there are no amino acid frequency biases across different protein families. To investigate whether this assumption is reasonable, I analyzed the background frequencies of all 10340 protein families in the PFAM database [54]. The results of this analysis are shown in Figure 4. Note that the background frequency from the table above which is shown as a green dot in the following figure is close to the mean frequency of the distribution. Note also that while the variance of the distribution is small, there are some families that clearly have amino acids far from the mean. I calculated that 66% of families have at least one amino acid that occurs at 2 fold greater than the background and that 3% of families have at least one amino acid that occurs at 4 fold greater than the background. In a way that I will discuss later, differences from background frequency on the order of 2 fold or greater could potentially change the frequency at some positions.

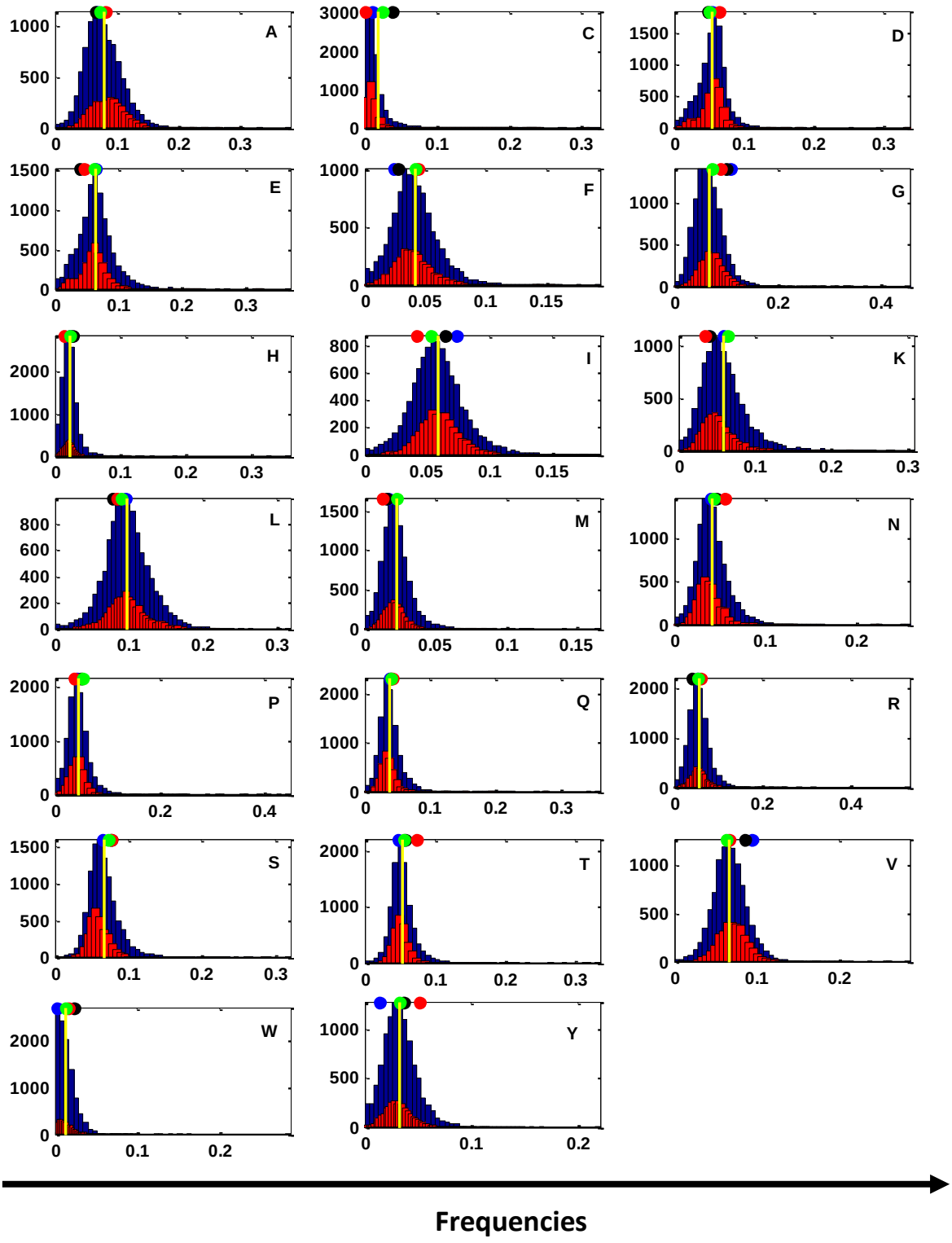
However, as shown in Figure 4, the families that I will use for coupling analysis are all close to the mean background frequency for most amino acids and hence results are not affected to any measurable extent.

Table 2: Background probabilities calculated using all sequences in the NR database (1998).

Amino acid	Background Probability (%)
Alanine (A)	7.3
Cysteine (C)	2.5
Aspartic acid (D)	5.0
Glutamic acid(E)	6.1
Phenylalanine (F)	4.2
Glycine (G)	7.2
Histidine (H)	2.3
Isoleucine (I)	5.3
Lysine (K)	6.4
Leucine (L)	8.9
Methionine (M)	2.3
Asparagine (N)	4.3
Proline (P)	5.2
Glutamine (Q)	4.0
Arginine (R)	5.2
Serine (S)	7.3
Threonine (T)	5.6
Valine (V)	6.3
Tryptophan (W)	1.3
Tyrosine (Y)	3.3

Examination of the background frequencies shown in Table 2 shows that there is a lot of variation. The most common amino acid is leucine occurring in 8.9% of all amino acids. The least frequent is tryptophan occurring in 1.3% of all amino acids. Based on this it is clear that the presence of tryptophan at 9% frequency would be highly unusual whereas observing the same frequency for leucine would not be unusual. A quantitative measure of unusualness that will assign a probability to the observed frequencies of amino acids given the background frequency is needed. A measure that I will discuss next which can be used for this purpose is called the binomial probability.

Figure 4: Background frequencies in different protein families. The blue histograms show frequencies of 20 amino acids in 10340 protein families in the PFAM-A database (Oct. 2008). The yellow line is the mean frequency. The red histograms are a subset of the blue ones where all protein families are represented by at least 100 sequences (to rule out biases of small sample size). The green dot shows the background probability used in the current work as in Table 2. Background frequencies of families used for SCA in this work are shown by blue (PDZ domain), black (serine proteases) and red (TonB dependent receptors) dots.



The binomial probability

The binomial probability is used to enable the estimation of the probability of an observed frequency of amino acids given a background frequency. It assumes two states for an event: an event either happens or does not happen. In the context of sequence data, the binomial probability considers whether an amino acid occurs or does not occur.

The binomial probability, in words, is the probability of getting *n* successes in a set of trials where every trial is independent. Calculating the binomial probability requires knowing the number of ways *n* can occur and the probability of success of *n*. Note that the probability of NOT success of *n* can be calculated from the probability of success as [1-probability of success].

The number of ways *n* can occur is given by the combination of the number of times an event can occur given a total number of trials (*N*). This is called (N choose n). The formula to calculate this is:

$$\text{N choose n}=\binom{N}{n}=\frac{N!}{n!(N-n)!}$$

The formula for binomial probability then becomes:

$$\text{binomial probability}=\binom{N}{n}\times(\text{probability of success})^{n}\times(\text{probability of not success})^{N-n}$$

If the probability of success is denoted as *q* and the probability of not success as *p* (note that *p*+*q*=1 and *p*=1-*q*) then the binomial probability can be written as:

$$\text{binomial probability}=\binom{N}{n}\times(\text{q})^{n}\times(1-q)^{N-n}$$

This is the general expression of the binomial probability. To calculate the amino acid frequency probability using the binomial probability is now straightforward. The goal is to calculate the binomial probability of the observed amino acid frequency at a position in a multiple sequence alignment.

The first step is to calculate given a multiple sequence alignment with *M* sequences the frequency, *f*, of an amino acid, *a*, at particular position, *i*. I will denote this observed frequency as f_i^a. In the binomial probability expression it is necessary to know the number of amino acids *a* that occur at position. This can be readily calculated by either counting or by multiplying the calculated frequency f_i^a by the total number of sequences, *M*. Thus one can substitute for *n* the product, Mf_i^a. *N* is the number of sequences, *M*. The only unknown quantity before the binomial probability is computed is the

probability of success. This is where the background probabilities that were discussed in the previous section become essential because the background probability of a particular amino acid is the probability of success (the background probability will be denoted as *q*).

This means that what is being done is to calculate the frequency of a particular amino acid at a position and comparing that to the average frequency of amino acids observed in all protein families.

Now that the background probabilities are considered to be the probability of success, the last variable is known and the binomial probability can be calculated. The expression for the binomial probability can now be written as:

$$\text{binomial probability}=\mathrm{P}(f_i^{(a)})=\begin{pmatrix} M \\ Mf_i^{(a)} \end{pmatrix}\times(\mathrm{q})^{Mf_i^{(a)}}\times(1-q)^{M(1-f_i^{(a)})}$$

This expression can be written in a simpler form which makes it easier to analyze. The simplification begins by expanding the combination expression (subscripts and superscripts will be dropped for clarity):

$$\mathrm{P}(f_i^{(a)})=\frac{M!}{Mf!(M-Mf)!}\times(\mathrm{q})^{Mf}\times(1-q)^{M(1-f)}$$

Now take the natural logarithm of both sides:

$$\ln(\mathrm{P}(f_i^{(a)}))=\ln[\frac{M!}{Mf!(M-Mf)!}\times(\mathrm{q})^{Mf}\times(1-q)^{M(1-f)}]$$

which expands to:

$$\ln(P)=\ln M!-\ln Mf!-\ln(M-Mf)!+\ln q^{Mf}+\ln(1-q)^{M(1-f)}$$

This expression can be further simplified by using the Stirling approximation for the factorials. The Stirling approximation to the first order is:

$$\ln N!=N\ln N-N$$

Therefore:

$$\ln(P)=M\ln M-M+Mf\ln q+M(1-f)\ln(1-q)-Mf\ln Mf+Mf-M(1-f)\ln M(1-f)+M(1-f)$$
$$\ln(P)=M\ln M-M+Mf\ln q+M(1-f)\ln(1-q)-Mf\ln Mf+Mf-M(1-f)\ln M(1-f)+M-Mf$$

The *M* and *Mf* terms cancel out and the equation reduces to:

$$\ln(P)=M\ln M+Mf\ln q+M(1-f)\ln(1-q)-Mf\ln Mf-M(1-f)\ln M(1-f)$$

'M' is now common to all expressions and can be factored out:

$$\ln(P) = M[\ln M + f \ln q + (1-f)\ln(1-q) - f \ln Mf - (1-f)\ln M(1-f)]$$

After some more rearrangement and factorization:

$$\ln(P) = M[(1-f-1+f)\ln M + f(\ln q - \ln f) + (1-f)(\ln(1-q) - \ln(1-f))]$$

$$\ln(P) = M[0 \ln M + f \ln \frac{q}{f} + (1-f)\ln \frac{1-q}{1-f}]$$

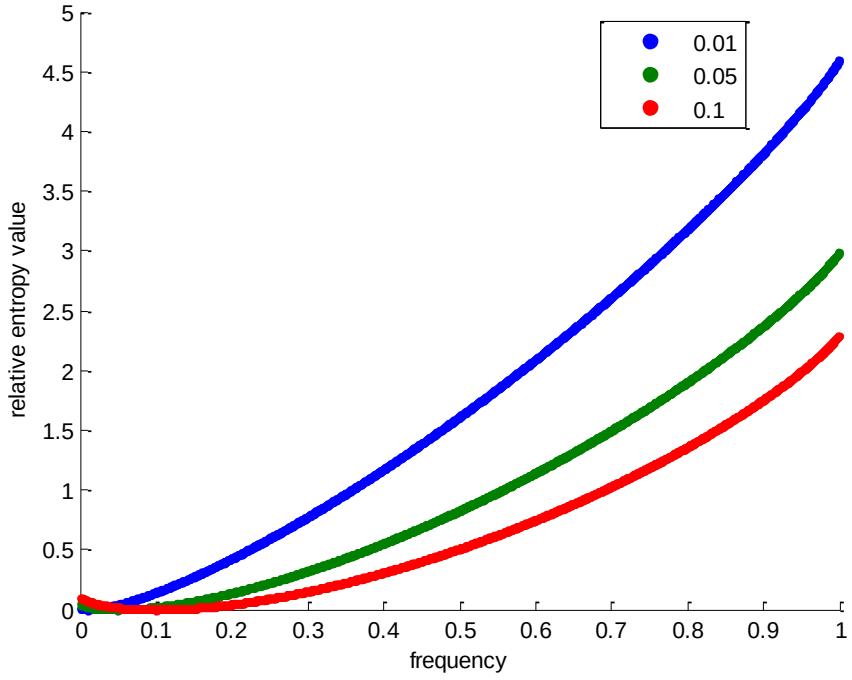
$$\ln(P) = -M[f \ln \frac{f}{q} + (1-f)\ln \frac{1-f}{1-q}]$$

The final expression above relates the probability of amino acid frequency to the number of sequences in the alignment multiplied by a function that only has observed frequencies (*f*) and background probabilities (*q*) in it. This function is called the relative entropy (denoted by *D*):

$$\text{relative entropy} = D = f \ln \frac{f}{q} + (1-f)\ln \frac{1-f}{1-q}$$

One thing to note about the relative entropy is that it depends only on the observed frequency and the background probability. With the relative entropy expression it is easy to see what will happen if the observed frequency of an amino acid is equal to the background probability. In that case, $q = f$ and $f \ln 1 + (1-f)(\ln(1))$ is zero (because ln1 = 0). The form of relative entropy with different background frequencies is shown in Figure 5.

Figure 5: The form of relative entropy plotted with three different background frequencies shown in the legend.



Calculation of the binomial probabilities of amino acids allows one to state how significant a particular frequency of amino acid at a position is. The goal of SCA however is not to calculate significance at one site but to calculate how the amino acid distribution at one site is similar to that at another site. One measure of similarity is the covariance which I will discuss next. Keep in mind however, that the covariance expression will take into account the binomial probabilities.

Covariance

Covariance is a commonly used concept in statistics. Given two vectors, one can calculate how close they are to each other using the covariance. The covariance (between two vectors x and y of dimension n, with means denoted with $\bar{x}$ and $\bar{y}$ respectively and components denoted by x_i and y_i) is defined as:

$$\text{covariance} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_i^n x_i y_i}{n} - \bar{x}\bar{y}$$

As the above equation shows the covariance depends on the product of x_i and y_i as well as the on the means of *x* and *y*. The dependence on the product means that if two vectors grow together, their product will be high and so the sum will be high and therefore the covariance. If one vector is negative while the other is positive, the covariance will have a high negative value. If the vectors have a random mixture of high and low values, the covariance will be close to zero. Thus the magnitude of the covariance indicates how close to each other two vectors are.

In the case of sequence data, one wants to calculate the covariance between two positions. However there is one complication in that the sequence data is not numerical but is composed of discrete amino acids. One cannot multiply or add amino acids together. To be able to perform calculations, the sequence data is binarized.

Binarization considers one type of amino acid at a time for a given position. If an amino acid '*a*' at a position '*i*' is present, then a value of 1 is assigned to that position; if not present, then a value of zero is assigned to the position. This process is carried out for all the other positions. It is then possible to calculate any covariation between the two vectors for a particular pair of amino acids. In a sense, what is done is to convert one alignment into 20 alignments, with each alignment containing only one type of amino acid.

For example, consider that an amino acid '*a*' at position '*i*' is present in the first four sequences of an alignment. The vector for amino acid '*a*', position '*i*' is then (1 1 1 1 0). Similarly if at position *j*,

amino acid '*a*' is not present then the vector for that will be (0 0 0 0 0). These vectors of ones and zeros are termed binary vectors.

A binary vector can be represented as $x_{i,s}^{(a)}$, where *s* refers to sequence, *i* to the multiple sequence alignment position and *a* to the amino acid. The frequency of an amino acid at a position *i* can now easily be calculated as the average of the sum of binary vector. The frequency of binary vectors is denoted as $\left\langle x_{i,s}^{(a)} \right\rangle_s$.

A covariance score can now be calculated for any two pairs of amino acids at two positions using the binary vectors. The covariance is based on the above formula and can be written with the expression for binary vectors as:

$$\text{covariance} = C_{ij}^{(ab)} = \left\langle x_{i,s}^{(a)} x_{j,s}^{(b)} \right\rangle_s - \left\langle x_{i,s}^{(a)} \right\rangle_s \left\langle x_{j,s}^{(b)} \right\rangle_s = f_{ij}^{(ab)} - f_i^{(a)} f_j^b$$

A high covariance score says that the two vectors have the same pattern across sequences meaning that wherever there is a 1 in one vector, there is a 1 in another vector. A low covariance score implies that the vectors have no pattern.

Since twenty amino acids are possible at a given position *i* and twenty amino acids are possible at position *j*, there are four hundred values quantifying the degree of covariance between every possible pair of amino acids. These four hundred values can be visualized by thinking of a 20×20 matrix of couplings. The question that arises then is what pair of couplings or combination of pairs should be considered as representative of the coupling between two positions? If all 400 numbers are equally prevalent or large, then this would complicate the analysis. With real alignments, however, this 20×20 matrix of coupling values is sparse because most amino acids are not present in all positions. Furthermore, there is usually one pair that produces a high covariation score. Given these observed properties of the matrix of coupling values, two approaches are possible to obtain a single number representing the coupling between two positions of the alignment. The first approach is to consider only the pair that has the highest covariance. The second approach is to sum all the values in the 20x20 matrix which means that if two pairs have a high score then the total covariance at the position will reflect the two scores. In practice, there is not much numerical difference between these approaches so in this work the sum of all the covariance values will be used.

Weighted covariance

Now that binomial probabilities and covariances have been explained, the current SCA method can be derived. The SCA method combines both the binomial probabilities and the covariance calculation.

The idea is to weight the covariance by a function of the binomial probability. In other words, what is desired is a weighting function that would downweight the contribution of an amino acid to the covariation if the frequency of an amino acid is not significantly above background. On the other hand, if an amino acid frequency is significantly above background, then its covariance score should be weighted higher. The effect of this type of weighting is that it will reduce the covariation between two positions caused by a few related sequences that happen to exist in the alignment. The weighting equation can be written in the following way ($C_{ij}^{(ab)}$ is the covariance between amino acids *a* and *b* at positions *i* and *j*/ $D_i^{(a)}$ is the relative entropy of amino acid *a* at position *i*):

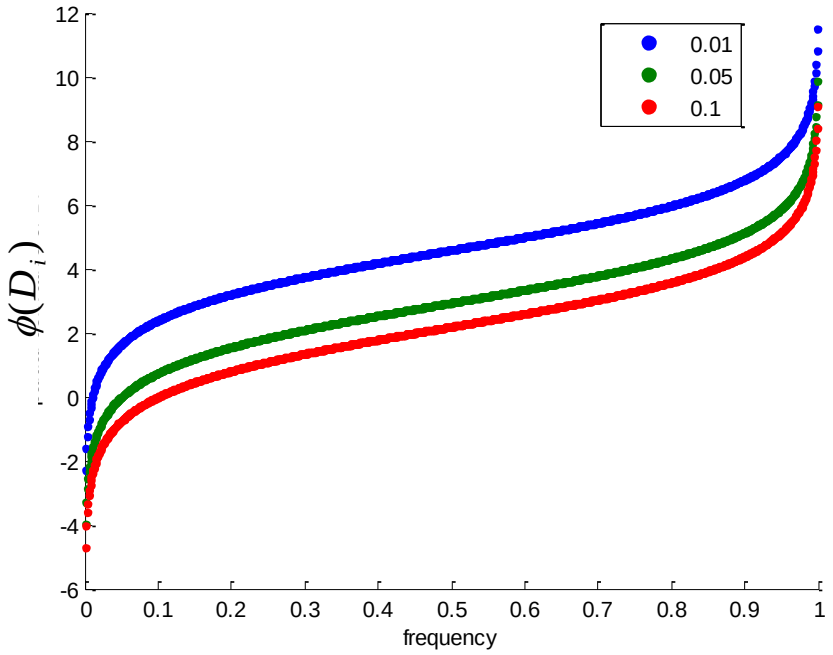
$$\text{weighted covariance}=\tilde{C}_{ij}^{(ab)} = \phi(D_i^{(a)})\phi(D_j^{(b)})C_{ij}^{(ab)}$$

$\phi(D_i^{(a)})$, is called a functional which is a function of a function and has the form:

$$\phi(D_i^{(a)}) = \frac{\partial D_i^{(a)}}{\partial f_i^{(a)}} = \ln\left[\frac{f_i^{(a)}(1-q^{(a)}}{(1-f_i^{(a)})q^{(a)}}\right]$$

The derivation of this will be given later, but for now, I will state that the weighting function is given by the first derivative of the relative entropy term with respect to frequency. Therefore if the covariation score and the frequency of an amino acid and the background frequency are known, the weighted covariance can be calculated. The form of $\phi(D_i^{(a)})$ is given in Figure 6. One can see by the

Figure 6: The form of the partial derivative of the relative entropy with respect to frequency. The different curves represent different background frequencies of 0.01, 0.05 and 0.1.



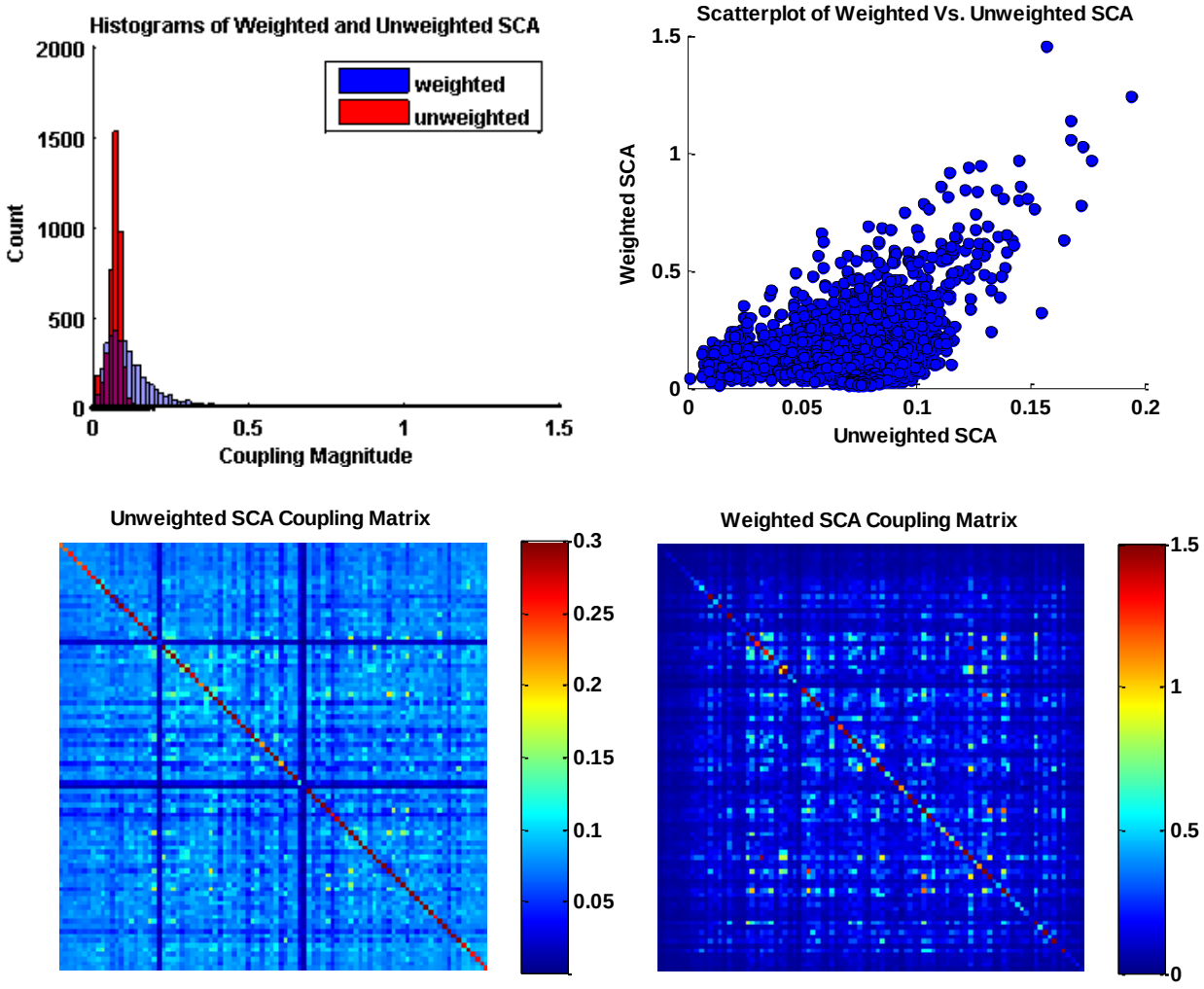
graph and by the equation that if the observed frequency is at the background frequency, the derivative of Di would be zero and therefore the weighted covariation including that site would to go to zero regardless of the actual covariation. On the other hand as the frequency increases above that of the background the weighting factor increases.

With this the reader can see that a high weighted covariation score is only possible if there is a high covariation score between two vectors and if the frequencies of amino acids in either vector are higher than the background frequency.

The impact of weighting on the covariance is shown in the Figure 7 on the following page. As is clear from the histograms of the coupling values, the weighting scheme decreases the values of some of the positions while dramatically increasing the magnitude of a few positions. In effect the weighting makes some positions stand out from the cluster of values observed in the unweighted covariance matrix. Note that mathematically, the matrix after weighting is not a true covariance matrix and so is referred to as coupling matrix. However, for certain purposes as will be explained later, the coupling matrix is used as if it is a covariance matrix because numerically it is close to being a true covariance matrix.

Taking the first derivative of the relative entropy may seem at this stage a somewhat arbitrary step. However, this weighting is actually the most natural one as will be shown next when another implementation of SCA called the bootstrap method is discussed; as will be shown the bootstrap method is equivalent to the weighted covariance.

Figure 7: Impact of weighting on the covariance values. The following histograms, matrices and scatter plot show that the weighting scheme increases the range of the covariance values so that some positions are more easily distinguished from the weak coupling. A PDZ domain alignment was used for this analysis. The coupling matrices are shown with values colored according to the color scale.



The bootstrap method

The bootstrap method [55] is a way to calculate the covariation using, not the positional frequencies (vectors of one and zero) but rather vectors of the relative frequency. The relative frequency is the change in the frequency of an amino acid at a position when one sequence is removed from the alignment. *Note that the relative frequency is not the same as the relative entropy.* The basic idea is to calculate coupling based on the correlation between relative frequencies at different sites as sequences are removed.

The bootstrap method removes a sequence *'s'* from an alignment and then calculates the new relative frequency (*D*) for every position. . The relative frequency changes because a sequence is removed, so the total number of sequences (denoted *M* in the binomial probability expression) decreases. Then, if every sequence of the alignment is removed and a new relative frequency calculated, then one obtains a vector of relative frequency (*D*) for every position and every pair. Next one calculates which positions show similar changes in relative entropy as the sequences are eliminated by calculating the covariation between the vectors of relative entropy obtained with single sequence elimination. Positions that covary would have the same variation in relative entropy and so will have a high covariance.

The covariation given by the bootstrap method when a sequence, s is removed, is written like this (the $\langle\ \rangle$ symbols represent calculation of a mean value of the expression in the brackets):

$$\text{relative frequency covariation}=\hat{C}_{ij}^{(ab)} = \left\langle D_{i,s}^{(a)} D_{j,s}^{(b)} \right\rangle - \left\langle D_{i,s}^{(a)} \right\rangle \left\langle D_{j,s}^{(b)} \right\rangle$$

This bootstrapping method can be shown to be equivalent, within a multiplicative factor to the weighted covariation method.

First the expression for the frequency of an amino acid (*f*) can be written as the number of sequences containing an amino acid '*a*' at position '*i*' divided by the total number of sequences, *M*:

$$f_i^a = \frac{M_i^a}{M}$$

Now if a sequence'*s*' is removed as in the bootstrap procedure the expression for the frequency changes. *M* is reduced by one and M_i^a could change depending on whether an amino acid '*a*' is present in that sequence. In any case M_i^a will either be reduced by one or not reduced at all depending on whether the removed sequence '*s'* contained amino acid *a* at position *i* (represented by $x_{i,x}^a$ in the following expression):

$$\text{frequency after eliminating one sequence= } f_{i,s}^{a} = \frac{M_i^a - x_{i,x}^a}{M-1}$$

If *M* is large so that the elimination of one sequence does not change *M* very much, then the removal of one sequence can be considered a slight perturbation. This slight perturbation allows the use of the Taylor series approximation.

The Taylor series approximation is a method used to approximate functions as a series of infinite polynomial terms. A Taylor series for a function f is written as:

$$\text{Taylor series=} f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

The important property of the Taylor series is that the Taylor series of order n has the same values when *x*=*a*. When *x* is close to *a* then the Taylor series of a particular order approximates a function. The farther x is from a, then the more terms need to be considered for a good approximation.

The Taylor series for the frequency function can be written as:

$$f_{i,s}^{a}(M-1) \approx f_{i,s}^{a}(M) + f_{i,s}'^{a}(M)\left[(M-1)-M\right] + \ldots = f_{i,s}^{a}(M) - f_{i,s}'^{a}(M) + \ldots.$$

To evaluate this, the function for the frequency and the derivative of the frequency are needed. The frequency was given before as $f_i^{(a)} = \frac{M_i^{(a)}}{M}$. The derivative of the frequency with respect to M is:

$$f_i'^{a} = \frac{-M_i^a}{M^2}$$

The derivative and the frequency after single sequence elimination are substituted into the Taylor series expression:

$$f_{i,s}^{a}(M-1) \approx \frac{M_i^a - x_{i,s}^a}{M} + \left[-\frac{(M_i^a - x_{i,s}^a)}{M^2}(-1)\right] \approx \frac{M_i^a}{M} - \frac{x_{i,s}^a}{M} + \frac{M_i^{a2}}{M} - \frac{x_{i,s}^a}{M^2}$$

The last term is divided by M^2 and if *M* is large will be small so it will be discounted. The expression for frequency will then be substituted and the M term grouped to give:

$$f_{i,s}^{a}(M-1) \approx f_i^a - \frac{x_{i,s}^a}{M} + f_i^a \frac{1}{M} \approx (1+\frac{1}{M}) f_i^a - \frac{x_{i,s}^a}{M}$$

The same approach can now be used for the relative entropy, *D* which is a function of the frequency.

$$D_{i,s}^a = D[\, f_{i,s}^a\,] \approx D[(1+\frac{1}{M})\, f_i^{\,a}\,] + \frac{\partial D_i^a}{\partial f_i^{\,a}}[(1+\frac{1}{M})\, f_i^{\,a} - \frac{x_{i,s}^a}{M} - (1+\frac{1}{M})\, f_i^{\,a}\,]$$

$$D_{i,s}^a = D[\, f_{i,s}^a\,] \approx D[(1+\frac{1}{M})\, f_i^{\,a}\,] + \frac{\partial D_i^a}{\partial f_i^{\,a}}[-\frac{x_{i,s}^a}{M}]$$

$$D_{i,s}^a = D[\, f_{i,s}^a\,] \approx \widehat{D}_i^a - \frac{x_{i.s}^a}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}}$$

Now the definition of covariation comes into play:

$$\text{relative entropy bootstrap covariation} = \widehat{C}_{ij}^{ab} = \left\langle D_{i,s}^a D_{j,s}^b \right\rangle_s - \left\langle D_{i,s}^a \right\rangle_s \left\langle D_{j,s}^b \right\rangle_s$$

The relative entropy can be substituted with the relative entropy approximation. This will give the following:

$$\hat{C}_{ij}^{ab} \approx \left\langle (\widehat{D}_i^a - \frac{x_{i.s}^a}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}})(\widehat{D}_j^b - \frac{x_{j.s}^b}{M}\frac{\partial D_j^b}{\partial f_j^{\,b}}) \right\rangle_s - \left\langle (\widehat{D}_i^a - \frac{x_{i.s}^a}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}}) \right\rangle_s \left\langle (\widehat{D}_j^b - \frac{x_{j.s}^b}{M}\frac{\partial D_j^b}{\partial f_j^{\,b}}) \right\rangle_s$$

This expression can be rewritten by multiplying the joint product and simplifying:

$$\hat{C}_{ij}^{ab} \approx \left\langle \widehat{D}_i^a \widehat{D}_j^b - \widehat{D}_j^b \frac{x_{i.s}^a}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}} - \widehat{D}_i^a \frac{x_{j.s}^b}{M}\frac{\partial D_j^b}{\partial f_j^{\,b}} + \frac{x_{i.s}^a x_{j.s}^b}{M^2}\frac{\partial D_i^a}{\partial f_i^{\,a}}\frac{\partial D_j^b}{\partial f_j^{\,b}} \right\rangle_s - (\widehat{D}_i^a - \frac{\left\langle x_{i.s}^a \right\rangle_s}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}})(\widehat{D}_j^b - \frac{\left\langle x_{j.s}^b \right\rangle_s}{M}\frac{\partial D_j^b}{\partial f_j^{\,b}})$$

$$\hat{C}_{ij}^{ab} \approx \left\langle \widehat{D}_i^a \widehat{D}_j^b - \widehat{D}_j^b \frac{x_{i.s}^a}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}} - \widehat{D}_i^a \frac{x_{j.s}^b}{M}\frac{\partial D_j^b}{\partial f_j^{\,b}} + \frac{x_{i.s}^a x_{j.s}^b}{M^2}\frac{\partial D_i^a}{\partial f_i^{\,a}}\frac{\partial D_j^b}{\partial f_j^{\,b}} \right\rangle_s$$

$$-\widehat{D}_i^a \widehat{D}_j^b - \widehat{D}_j^b \frac{\left\langle x_{i.s}^a \right\rangle_s}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}} - \widehat{D}_i^a \frac{\left\langle x_{j.s}^b \right\rangle_s}{M}\frac{\partial D_j^b}{\partial f_j^{\,b}} + \frac{\left\langle x_{i.s}^a \right\rangle_s \left\langle x_{j.s}^b \right\rangle_s}{M^2}\frac{\partial D_i^a}{\partial f_i^{\,a}}\frac{\partial D_j^b}{\partial f_j^{\,b}}$$

Now expand the first average:

$$\hat{C}_{ij}^{ab} \approx \widehat{D}_i^a \widehat{D}_j^b - \widehat{D}_j^b \frac{\left\langle x_{i.s}^a \right\rangle_s}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}} - \widehat{D}_i^a \frac{\left\langle x_{j.s}^b \right\rangle_s}{M}\frac{\partial D_j^b}{\partial f_j^{\,b}} + \frac{\left\langle x_{i.s}^a x_{j.s}^b \right\rangle_s}{M^2}\frac{\partial D_i^a}{\partial f_i^{\,a}}\frac{\partial D_j^b}{\partial f_j^{\,b}}$$

$$-(\widehat{D}_i^a \widehat{D}_j^b - \widehat{D}_j^b \frac{\left\langle x_{i.s}^a \right\rangle_s}{M}\frac{\partial D_i^a}{\partial f_i^{\,a}} - \widehat{D}_i^a \frac{\left\langle x_{j.s}^b \right\rangle_s}{M}\frac{\partial D_j^b}{\partial f_j^{\,b}} + \frac{\left\langle x_{i.s}^a \right\rangle_s \left\langle x_{j.s}^b \right\rangle_s}{M^2}\frac{\partial D_i^a}{\partial f_i^{\,a}}\frac{\partial D_j^b}{\partial f_j^{\,b}})$$

Many terms cancel now and the expression simplifies considerably:

$$\hat{C}_{ij}^{ab} \approx \frac{\left\langle x_{i.s}^a x_{j.s}^b \right\rangle_s}{M^2}\frac{\partial D_i^a}{\partial f_i^{\,a}}\frac{\partial D_j^b}{\partial f_j^{\,b}} - \frac{\left\langle x_{i.s}^a \right\rangle_s \left\langle x_{j.s}^b \right\rangle_s}{M^2}\frac{\partial D_i^a}{\partial f_i^{\,a}}\frac{\partial D_j^b}{\partial f_j^{\,b}}$$

Grouping terms yields:

$$\hat{C}_{ij}^{ab} \approx \frac{1}{M^2}\frac{\partial D_i^a}{\partial f_i^a}\frac{\partial D_j^b}{\partial f_j^b}(\left\langle x_{i.s}^a x_{j.s}^b \right\rangle_s - \left\langle x_{i.s}^a \right\rangle_s \left\langle x_{j.s}^b \right\rangle_s)$$

This is just the covariation matrix multiplied by the first derivative of the relative entropy and divided by the square of the number of sequences:

$$\text{weighted covariation}=\hat{C}_{ij}^{ab} \approx \frac{1}{M^2}\frac{\partial D_i^a}{\partial f_i^a}\frac{\partial D_j^b}{\partial f_j^b}(C_{ij}^{ab})$$

This shows that the bootstrap method for calculating covariance is the same as the relative entropy weighting up to a constant. Thus, there is no need to perform the bootstrap procedure as the weighted covariance can be calculated analytically if *M* is large enough (more than 100) that the assumption (that removing one sequence does not change the frequency very much) holds .

Perturbation method

This was the first method developed in the Ranganathan lab and used in several of the first studies [56-60]. This approach will be presented as it was first published and then it will be related to the framework presented previously of weighted covariances.

The coupling analysis method starts by defining a multiple sequence alignment as a thermodynamic ensemble. In this evolutionary ensemble, the different positions have different energy states. With this assumption, the difference in energy between positions can be calculated using the Boltzmann distribution. The Boltzmann distribution, for two populations, n_i and n_0, separated by energy (e_i-e_0) can be written as

$$\frac{n_i}{n_0} = e^{\beta(e_i - e_0)}, \beta = \frac{1}{kT}$$

In the 1999 Lockless and Ranganathan paper [58], the following was written down as the energy difference (ΔG) between mutating two positions (i and j) in an alignment:

$$\frac{P_i^x}{P_j^x} = e^{\frac{\Delta G_{i\to j}^x}{kT^*}}$$

This is considered as equivalent to the Boltzmann distribution expression with $\Delta G=\varepsilon_i - \varepsilon_{o,}$ and where the probabilities can be obtained from the numbers by dividing by the total number of amino acids at

that position. In addition $1/kT^*$ is denoted as '*' to note that this is not the same as that used in classical thermodynamic systems. (The ϐ value of $\frac{1}{kT}$ in classical thermodynamic systems is obtained from comparing the expression of internal energy of an ideal gas obtained using the equipartition expression with the expression of internal energy obtained using the translational partition function [61]. For sequence data, the value of ϐ is unknown because the internal energy of a system of sequences and the partition functions cannot as yet be calculated). Also since there are 20 different amino acids at a position, there is an 'x' for every amino acid. Using this expression, the observed probabilities for an amino acid can be converted into an energy score by taking the logarithm of both sides and rearranging:

$$\Delta G^{x}_{i\to j} = kT^{*} \ln \frac{P^{x}_{i}}{P^{x}_{j}}$$

Then, based on the above equation, a measure of conservation is defined in the following way:

$$\Delta G^{STAT}_{i} = kT^{*} \sqrt{\sum_{x} (\ln \frac{P^{x}_{i}}{P^{x}_{MSA}})^2}$$

Introduction of this measure is necessary because at any site there are multiple amino acids and this accounts for them by summing over all of the amino acids at a position. In addition, P_{MSA} refers to the probability of an amino acid in the total alignment.

This conservation measure is then used to calculate the coupling. The idea behind coupling is to measure the degree that an amino acid at one position correlates with an amino acid at another position. In this paper, this was implemented in two steps:

1. Choose a position *i* and calculate ΔG^{STAT} for position *i*.
2. Choose an amino acid *x* at position *j* and then calculate ΔG^{STAT} for position *i*. Subtract the ΔG^{STAT} after selection from the ΔG^{STAT} before selection. This defines $\Delta\Delta G^{STAT}$ between positions *i* and *j*. The actual formula incorporates the fact that different amino acids occur at a given position by summing them. The formula for this is:

$$\Delta\Delta G^{STAT}_{i} = kT^{*} \sqrt{\sum_{x} (\ln \frac{P^{x}_{i|\delta j}}{P^{x}_{MSA|\delta j}} - \ln \frac{P^{x}_{i}}{P^{x}_{MSA}})^2}$$

The one thing needed to solve this equation are the probabilities of amino acids at positions. In this paper the probabilities were defined as binomial probabilities. Recall that a binomial probability measures the probability that an event occurs relative to that event not occurring. In this approach the

probability of every amino acid is calculated separately so that a position in an alignment has a vector containing 20 elements (although any particular value could be 0). In the case of a multiple sequence alignment, the binomial probability is calculated as follows:

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

where,

p_x: occurrence of amino acid x in a protein database,

N: total number of sequences

n_x: the number of occurrences of amino acid x

This completes the description of the original SCA method. Its use in practice will be discussed now. One practical limitation is with the method of selecting a position *j* when calculating $\Delta\Delta G^{STAT}$. The limitation is that in some cases one can calculate the coupling at position i but not the coupling at position j. This resulted in asymmetric coupling patterns. This occurs because of the imposition of a threshold when selecting sequences. The threshold is that after selecting an amino acid a position, the distribution at several pre-selected non conserved sites should not change. This threshold is important to make sure that the selection is not globally changing the alignment (i.e. selecting for closely related sequences). If that number is not reached, then no calculations are made. In practice this is a sampling issue when there are few sequences present in the alignment. However, in cases where the coupling can be calculated between two positions, the coupling energies are similar. Although this method was used for several papers [56-60, 62], it has now been supplanted with different methods of calculating correlations.

This method also fits into the framework of weighted covariation. This can be seen by deriving an analytical equation for the perturbation method.

The starting point is to write down how the frequencies change in the subalignment vs. the full alignment. The subalignment frequencies are technically conditional frequencies; that is, they are the frequency of an amino acid given that the previous position has been fixed. This conditional frequency is denoted like this:

$$\text{conditional frequency} = f_{j|i}^{b|a_i}$$

In general, a conditional frequency can be written as the ratio of the joint frequency to the marginal frequency. That is, the joint frequency of '*a*' and '*b*' is the frequency of '*a*' times the conditional frequency of '*b*' given '*a*'.

$$\text{joint frequency}=f_{ij}^{a,b} = f_i^{a_i} f_{j|i}^{b|a_i}$$

Next, add and subtract f_j^b to the previous equation so that:

$$f_{j|i}^{b|a_i} = \frac{f_{ij}^{a,b}}{f_i^{a_i}} = f_j^b - \frac{f_{ij}^{a,b}}{f_i^{a_i}} - f_j^b$$

If $f_i^{a_i}$ is not taken as a common factor one obtains the following expression:

$$f_{j|i}^{b|a_i} = f_j^b - \frac{f_{ij}^{a,b} - f_i^{a_i} f_j^b}{f_i^{a_i}}$$

The numerator in the fraction above is a covariance so:

$$f_{j|i}^{b|a_i} = f_j^b - \frac{C_{ij}^{a_i b}}{f_i^{a_i}}$$

Now the expression for the conditional relative entropy can be written, using the Taylor series approximation discussed previously, as:

$$D(f_{j|i}^{b|a_i}) \approx D(f_j^b) + \frac{\partial D_j^b}{\partial f_j^b}[f_{j|i}^{b|a_i} - f_j^b]$$

The joint frequency expression can be substituted as above so that the relative entropy is given by:

$$D(f_{j|i}^{b|a_i}) \approx D(f_j^b) + \frac{\partial D_j^b}{\partial f_j^b}[\frac{C_{ij}^{a_i b}}{f_i^{a_i}}]$$

This expression now has a covariation term in it as well as a partial derivative. With it, the analytical solution to the perturbation is at hand. The $\Delta\Delta G_{j|i}^{STAT,b,a_i}$ is written as:

$$\Delta\Delta G_{j|i}^{stat,b,a_i} = -\frac{1}{M}[\ln P_M[f_j^b] - \ln P_M[f_{j|i}^{b|a_i}]]$$

To remind the reader, the expression $\frac{-1}{M}\ln P_M[f_j^b]$ is the relative entropy D_j^b so that $\Delta\Delta G_{j|i}^{STAT,b,a_i}$ is the difference between the marginal relative entropy and the conditional relative entropy like this:

$$\Delta\Delta G_{j|i}^{STAT,b,a_i} = D(f_j^b) - D(f_{j|i}^{b|a_i})$$

This statement can now be compared to the equation for the joint relative entropy. The reader can see that $\Delta\Delta G_{j|i}^{STAT,b,a_i}$ can be written like this:

$$\Delta\Delta G_{j|i}^{STAT,b,a_i} = \frac{-1}{f_i^{a_i}} \frac{\partial D_j^b}{\partial f_j^b} C_{ij}^{a_i b}$$

This can be considered as a weighted covariation with one weight being the partial derivative of the relative entropy and the other weight the inverse of the frequency. Note that this weighting scheme is not symmetrical. This completes the analytical solution of the perturbation experiment.

The perturbation method is presented here for completeness; it is not currently used to calculate coupling. The weighted covariation is used.

In the next section I consider the methods other labs have used to calculate coupling.

Section 2: Methods reported in the literature to calculate covariation:

Many approaches to calculate coupling between positions in a multiple sequence alignment have been published (see Table 3). In this section, I describe a subset of the published methods along with brief discussions of specific results or conclusions. I also compare the SCA method to selected methods. I have divided the different literature studies into the following categories: early studies, statistical algorithms to calculate coupling, comparison of selected methods with SCA, statistical methods to correlate coupling with physical contacts, experimental approaches, learning algorithms, inter-protein coupling, methods that attempt to correct for phylogeny, miscellaneous work, and a review of reviews.

Early studies

The observation of correlations between positions in sequence data predated the advent of fast and reliable DNA sequencing methods in 1975. The first report I know of was in 1972 when Crowson, at the zoology department in the University of Glasgow observed by visual inspection correlation between positions of 43 cytochrome c (a protein involved in apoptosis and the electron transport chain) sequences from several species [63]. The author noted that the correlations could be important and suggested that extracting significant meaning would have to await more sequences. Then in 1976 a paper from Wang's lab [64] at Carnegie Mellon employed statistical methods to calculate the significance of correlations on a set of 63 cytochrome c sequences concluding that correlated sites tended to be on the surface of a protein and suggesting that this is functionally important because other molecules bind to the surface of proteins. This paper was ahead of its time as it was not until more than 25 years later that methods of this level of sophistication are used.

The next paper to report coupling came in 1987 from Klug's lab [65] at Cambridge,11 years after the previous paper. In this work, the authors studied an alignment of seven viral sequences. The methodology they used was not based on statistical methods but on visual pattern recognition. The authors numbered the amino acids at a position. For example, if an amino acid at one position is absolutely conserved then the amino acids at all those positions would be given the value 1, and so for seven sequences the numbering would be a vector of 1's(1 1 1 1 1 1 1). If the position was not absolutely conserved but there were only two amino acids at a position, the vector of that position could be (1 2 1 1 1 1 1) if the second sequence had an amino acid different from the others. If there are three types of amino acids at a position, then the vector could be (1 2 3 1 1 1 1). This was done for all positions. The identity of the amino acid was not important for this study, only that the amino acid was different from others in a column.

Table 3: The following are published studies that analyze covariation between positions in a multiple sequence alignment. 'EXP?' refers to whether experiments were done and 'Purpose' refers to the whether the primary purpose of the study is to determine functional residues (F) or contact residues (C) or both (F/C).

#/ref	First Author	Last Author	Year	Coupling method	EXP?	Purpose
1 [63]	Crowson, RA		1972	visual pattern recognition	No	F
2 [64]	Wong, AK	Wang, CC	1976	mutual information based	No	F
3 [65]	Altschuh,D	Klug,A	1987	visual pattern recognition	No	F
4 [66]	Altschuh,D	Nagai, K	1988	visual pattern recognition	No	F
5 [67]	Vernet, T	Altschuh, D	1992	visual pattern recognition	Yes	F
6 [68]	Korber, BT	Lapedes, AS	1993	mutual information	No	F
7 [69]	Neher, E		1994	correlation, physiochemical	No	C
8 [70]	Gobel, U	Valencia, A	1994	correlation, distance metric	No	C
9 [71]	Taylor, WR	Hatrick, K	1994	correlation, physiochemical	No	C
10 [72]	Hatrick, K	Taylor, WR	1994	review	No	
11 [73]	Kovalenko, O	Horovitz, A	1994	not described	Yes	F
12 [74]	Clarke, ND		1995	mutual information, weighted	No	F
13 [75]	Lichtarge, O	Cohen, FE	1996	evolutionary trace	No	F
14 [76]	Thomas, DJ	Sander, C	1996	correlated, distance metric, learning	No	C
15 [77]	Pollock, DD	Taylor, WR	1997	review using simulated alignments	No	F/C
16 [78]	Chelvanayagam, G	Benner, SA	1997	phylogenetic, physiochemical corr.	No	F/C
17 [79]	Pazos, F	Valencia, A	1997	intraprotein physiochemical correlation	No	C
18 [80]	Olmea, O	Valencia, A	1997	correlation+conservation	No	C
19 [81]	Ortiz, AR	Skolnick, J	1997	correlation, physiochemical	No	C
20 [82]	Pollock, DD	Goldman, N	1999	maximum likelihood, physiochemical	No	C
21 [83]	Olmea, O	Valencia, A	1999	correlation+conservation	No	C
22 [84]	Ortiz, AR	Skolnick, J	1999	correlation, physiochemical	No	C
23 [85]	Farsiselli, P	Casadio, R	1999	correlation, neural network	No	C
24 [86]	Tuff, P	Darlu, P	2000	phylogenetic, physiochemical	No	C/F
25 [87]	Wollenberg, KR	Atchley, WR	2000	mutual information, phylogeny	No	F
26 [88]	Atchley, WR	Dress, AW	2000	mutual information, phylogeny	No	F
27 [89]	Fariselli, P	Casadio, R	2001	physiochemical, neural network	No	C
28 [90]	Pritchard, L	Dufton, M	2001	rule-based correlation	No	C
29 [91]	Filizola, M	Weinstein, H	2002	correlation, physiochemical	No	C
30 [92]	Oliveira, L	Vriend, G	2002	weighted covariance	No	F
31 [93]	Kass, I	Horovitz, A	2002	chi-square	No	F
32 [94]	Tillier, ER	Lui, TW	2003	mutual interdependency	No	F
33 [95]	Saraf, MC	Maranas, CD	2003	correlation, distance based	No	C/F
34 [96]	Fodor, AA	Aldrich, RW	2004	review-energetic coupling, contacts	No	C/F
35[97]	Fodor, AA	Aldrich, RW	2004	review-OMES, MI, SCA, McBasc	No	C/F
36 [98]	Dekker, JP	Yellen, G	2004	explicit likelihood of subset covariation	No	C
37 [99]	Fleishman, SJ	Ben-Tal, N	2004	correlation, phylogeny	No	F
38 [100]	Dutheil, J	Galtier, N	2005	correlation, phylogeny	No	C
39[101]	Gloor, GB	Dunn, SD	2005	mutual information	No	C/F
40 [102]	Noivirt, O	Horovitz, A	2005	chi-square, phylogeny	No	C/F
41 [103]	Fares, MA	Travers, SA	2006	correlation, phylogeny	No	C
42[104]	Halperin, I	Nussinov, R	2006	correlation	No	C
43 [105]	Kundrotas, PJ	Alexov, EG	2006	filter based correlation method	No	C
44 [106]	Fuchs, A	Frishman, D	2007	review of nine methods	No	C
45 [107]	Dunn, SD	Gloor, GB	2008	mutual information	No	C
46 [108]	Sayar, K	Onaran, O	2008	chi-square based	No	F
47 [109]	Skerker, JM	Laub, MT	2008	mutual information	Yes	C/F

To identify correlated positions the authors listed all the positions which had the same pattern. For example if position 1 had a vector (1 2 1 1 1 1 1) and position 2 had a vector (1 2 1 1 1 1 1) then they would be grouped together. The authors found several such groups. In the cases where the positions are not absolutely conserved then this would be grouping positions based on the pattern of substitutions at a position. The authors then mapped the correlated groups onto the known structure of the virus. They found that in most cases the grouped patterns were forming important structural contacts.

The visual pattern recognition process while conceptually sound did not have any statistical tests to ensure significance or to separate phylogenetic correlations from functional ones. In the follow-up paper in 1988, a contributor to the previous paper, Altschuh, now at Nagai's lab in France, addressed some of the limitations of the previous work [65]. The authors examine three protein families (serine proteases, hemoglobins, cysteine proteases) in the manner previously described. However, the authors realize certain complications. They note for instance, that the more number of sequences that one has in the alignment, the harder it will be to find exact pattern matches and so they introduced a method they called 'relaxed' to look for similar patterns rather than identical ones. The authors also introduce the idea of significance of positions as well as grouping amino acids based on chemical similarity. Note that these ideas are presented in a conceptual, not mathematical way. Later studies addressed coupling mathematically.

Statistical algorithms to calculate coupling

In this section I describe methods that have been used to calculate correlations. These methods use statistical approaches between positions to measure coupling and do not include any information about the physio-chemical nature of the amino acids for the statistical tests.

➔ **Mutual Information, Korber and Lapedes, 1993 [68]**

The paper by Korber and Lapedes in 1993 introduced a mathematical approach to quantify the coupling between positions in a protein multiple sequence alignment. The authors applied their method to the human immunodeficiency virus (HIV) (specifically, the V3 loop of the envelope protein).

The authors used a mutual information based method to detect coupling. Since mutual information is a widely used method I will explain it in some detail and then, in the next section, compare the values from mutual information to SCA values for the same alignment.

Mutual information is based on the idea of sequence entropy:

$$\text{sequence entropy}=-\sum_{x}(p_x(i)\ln p_x(i))$$ *$p_x(i)$ is the probability of occurrence of a state x at a position i. In the case of alignment data, x refers to the set of amino acids and i to the column of an alignment.*

For every column in the alignment, one can calculate a sequence entropy for a particular amino acid. Because there are 20 amino acids, one would have 20 sequence entropy values. For a second column in the alignment, one would then have another set of 20 entropy values.

If one now considers the occurrence of pairs of amino acids, one can calculate an entropy for that as well. Since there could be 20 amino acids at one site and 20 amino acids at another site, there could be 400 possible pairs of amino acids. The entropy of pairs is called joint entropy and it is calculated in the following way:

$$\text{joint sequence entropy}=-\sum_{xy}(p_{xy}(ij)\ln p_{xy}(ij))$$

With the single and joint entropies defined, a quantity called mutual information can be defined. Mutual information is calculating by summing the entropy of site i with the entropy of site j and then subtracting the joint entropy of sites i and j:

mutual information=entropy(i)+entropy(j) – joint entropy (i,j).

If two vectors are not independent then the mutual information will be high. Basically, mutual information is a measure of the reduction of uncertainty of one vector given another vector. The mutual information is always positive and has a maximum value when the covariation is one. If the mutual information is zero this could mean two things: either sites i and j vary independently or that there is no variation at all.

With real, finite data, the mutual information, even for independent, varying sites is almost never zero because the sample size is too small. This means that there is a slight bias towards positive mutual information in finite data sets. To account for this, the authors vertically shuffled the alignment 750,000 times and calculated a mutual information score. Then they counted the number of times that a mutual information score in the shuffled data sets was at least as large as the mutual information score calculated from the unshuffled data set. They divided this number by the total number of randomly vertically permuted alignments (750,000) to obtain a significance probability. The significance cutoff they chose was 0.1.

Their results are that only a subset of positions in the V3 loop were coupled based on the threshold that they chose. Note that the authors excluded columns with more than 95% identical positions and those with 'primarily' gaps.

The authors also calculated a term called the 'specific information'. This term describes the degree to which an amino acid at position i can predict an amino acid at position j. They listed for the columns with highest mutual information the amino acid pairs with high specific information.

The authors' primary purpose in developing this method is to suggest using the covarying residues to develop vaccines and to identify functional sites. They also acknowledge that sites could covary because of functional constraints and structural constraints.

➔ **Weighted mutual information, Clarke, 1995 [74]**

This 1995 paper by Clarke is one of the clearest and most explanatory papers dealing with coupling analysis. The author discusses coupling from a fundamental perspective and goes through all the factors that could affect the coupling signal. He discusses the assumption that all sequences in the analyzed set should undergo the same selective pressure and that sequence sets could be too small and that the sequences are not 'phylogenetically uniform'.

From an analytical perspective, the author also introduces a new metric based on mutual information. The author states that this metric is not rigorous but is based on the fact that it agrees better with functional and structural data. The expression used is:

$$\text{weighted mutual information} = \sum_{a_i, a_j} (P_{a_i, a_j})^2 \log\left(\frac{P_{a_i, a_j}}{P_{a_i} P_{a_j}}\right)$$

This differs from the usual mutual information in that the joint probabilities are squared. Also, the sum here is over residue pairs and not over sequences. The result of this is that amino acid pairs with high joint entropy (high information) that are present most often will contribute most to the sum. The author also states that many pairs of positions that would have been found to be highly covarying would be downweighted with his method because the covariation comes from low frequency pairs.

The author then applied this methodology to analyzing the coupling pattern and significance on the homeodomain family. He also identified a network of positions and depicted this with a network diagram. Covarying positions were also mapped onto structures to try to structurally rationalize the coupled pairs. Also, he analyzed the functional data available on the pairs and found that the covarying pairs are implicated in the function of the homeodomain.

Again, this paper is commendable for its scope and for the clarity of the writing and reasoning that went behind what the author did.

➔ **Weighted covariation, Vriend's lab, 2002 [92]**

Four protein families were analyzed using correlation. The authors differentiate between different sets of correlated positions and in fact find that proteins have different networks of correlated positions. This presages our work in the serine protease family by detecting four networks one of which is the active site residues, another is one of unknown function and two of which have to do with calcium binding.

The method used is quite similar in spirit to the one I described in section 1. It is a weighted covariation matrix. The difference is in the weights and the covariation calculation however:

$$\text{covariation(i,j)=W(i,j)}\sum_{p=1}^{n-1}\sum_{q=p|1}^{n}\partial(i_p i_q j_p j_q)$$

$$W=\frac{2}{n(n-1)}$$

The delta function can be 1 or 0 according to the following rules:
1: ip=jp and iq=jq OR ip~=jp and iq~=jq

2:ip=jp and iq~=jq OR ip~=jp and iq~jq

Another interesting part of this paper is that they make plots of entropy vs. variability and categorize rather arbitrarily the residues that fall in different regions of the plot. They make a model that says that for any given protein there is a set of highly conserved and invariable positions that make up the active site. This is surrounded by slightly less conserved positions. Then, on the surface are usually unconserved positions except at particular sites which then interact through unknown mechanisms through the conserved core of the protein to the active site. The model is interesting because it describes an information transfer pathway from one part of the protein to the other which is something that Lockless et al, 1999 [58] described as well.

➔ **Chi-square, Horovitz, 2002 [93]**

In this paper, Kass and Horovitz compute a measure of coupling based on the chi-square test. First, they calculated the expected frequency of number of sequences containing amino acid X and position i and amino acid Y at position j. This was calculated as the number of sequences*frequency

Xi*frequency Yj. An example if X at i occurs with a 0.2 frequency and Y at j occurs with 0.4 frequency, then assuming that they are independent, the expected frequency is 0.2*0.4=0.08. If 0.08 is multiplied by the number of sequences in the alignment, one gets the expected number of sequences.

This expected number is compared to the observed number of sequences. If this number is higher than the expected then one can consider that the residues are coupled. The significance of this is evaluated using the chi-square statistic.

One additional thing they did is to consider that multiple amino acids at positions i and j could occur. In that case they summed over all possible combinations of amino acids at sites i and j.

The formula they published is:

$$\chi^2(i,j) = \sum_n \frac{(N_{n,OBSERVED} - N_{n,EXPECTED})^2}{N_{n,EXPECTED}}$$

n=amino acids at position i×amino acids at position j, N=number of sequences

➔ **Mutual Interdependency, Tillier and Liu, 2003 [94]**

This paper follows up on the 1976 work [64] by using a mutual information based metric. The authors apply their method to many sets of proteins and find that it works better than mutual information because it predicts more close or contacting residues than mutual information. The weighting process includes several complex steps and would be interesting to investigate further.

➔ **Explicit Likelihood, Dekker, 2004 [98]**

This paper is a useful contribution to the literature as it introduces a new algorithm for coupling and compares that algorithm to the original SCA algorithm. In addition, it is a well written paper with a lot of explanations. The method they used to calculate covariation has very similar ideas to the original SCA in that one fixes a residue at one position in an alignment and looks at the change in conservation at another site in the alignment. However, the actual difference metric used by Dekker is different. Their metric is called Explicit Likelihood of Subset Co-variation (ELSC).

What they do is to calculate the number of ways that an observed distribution of amino acid frequencies in a subalignment can occur given the distributions of amino acids in the full alignment. This is done by calculating the number of times an amino acid *α* occurs in the subalignment given the total number of alanines in the full alignment (N_{ala} choose n_{ala}) multiplied by the combinations for all amino acids. So for example, if there are 10 alanines in the full alignment and 5 in the subalignment, then the total number of ways of getting 5 out of 10 is: 10!/5!(10-

5)!=10!/5!5!=10*9*8*7*6/5*4*3*2=9*28=252. This would be multiplied by the combinations for all 20 amino acids that could be present in the subalignment:

$$\Omega_j^{<i>} = \binom{N_{ala,j}}{n_{ala,j}} \bullet \binom{N_{asn,j}}{n_{asn,j}} \bullet \binom{N_{asp,j}}{n_{asp,j}} = \prod_r \binom{N_{r,j}}{n_{r,j}}$$

where the combination is given by the following:

$$\binom{N_{ala,j}}{n_{ala,j}} = \frac{N_{ala,j}!}{n_{ala,j}!(N_{ala,j} - n_{ala,j})!}$$

This total combination number is then divided by the total number of possible subsets which is (N_{total} choose n_{total}). This gives a probability, denoted by L:

$$L_j^{<i>} = \frac{\prod_r \binom{N_{r,j}}{n_{r,j}}}{\binom{N_{total}}{n_{total}}}$$

This L value is then further normalized to the likely frequency of an amino acid (denoted as m(r,j) at a position.

$$m_{r,j} \approx (\frac{N_{r,j}}{N_{total}}) * n_{total}\,,\ \sum_r m_{r,j} = \sum_r n_{r,j}$$

$$L_{j,\max}^{<i>} = \frac{\prod_r \binom{N_{r,j}}{m_{r,j}}}{\binom{N_{total}}{n_{total}}}$$

The normalization is then done by dividing L by Lmax which results in a score denoted by the authors as:

$$\wedge_j^{<i>} \equiv \prod_r \frac{\binom{N_{r,j}}{n_{r,j}}}{\binom{N_{r,j}}{m_{r,j}}}$$

Their results are that the algorithm performs with greater power than SCA at detecting contacting residues although the results by both SCA and ELSC were statistically significant.

➔ **Normalized mutual information, Dunn's lab, 2005 [101]**

This paper presents a new method for calculating covariation based on mutual information. Mutual information is based on the idea of sequence entropy provided before:

$$\text{sequence entropy}=-\sum_x (p_x(i) \ln p_x(i))$$

For every column in the alignment, one can calculate a sequence entropy for a particular amino acid. Because there are 20 amino acids, one would have 20 sequence entropy values. For a second column in the alignment, one would then have another set of 20 entropy values. If one now considers the occurrence of pairs of amino acids, one can calculate an entropy for that as well. In that case, it would be a joint entropy in the following way:

$$\text{joint sequence entropy}=-\sum_{xy} (p_{xy}(ij) \ln p_{xy}(ij))$$

For any given pair, there are potentially 400 possible pairs to consider when calculating joint entropies. Mutual information is calculating by summing the entropy of site i and amino acid x with the entropy of site j and amino acid y and then subtracting the joint entropy of sites i and j, amino acids x and y. If the joint entropy is high it means that the mutual information is high. Basically, mutual information is a measure of the reduction of uncertainty which one could see applies...if there is a high joint entropy then the uncertainty is less than it would be if positions i and j were evolving independently.

The authors noted some problems with mutual information. Those problems are that there is strong mutual information due to random pairings of amino acids in small sequence sets. They calculate that a minimum of 125 sequences be used to eliminate this. A second problem which exists is that low conserved positions have higher random pairings. To eliminate these problems the authors devised a

simple normalization where the mutual information is divided by joint entropy of the pair of positions. The authors also state that this ratio (mutual information/joint entropy) is a true distance measure.

A third problem that the authors discuss is the background mutual information coming from the fact that sequences are homologous. They have also done evolutionary simulations and found that if positions are highly conserved it is not easy to separate those that are co-evolving from those that are not. In this case, they impose a cutoff where they do not consider highly conserved positions. Also, they do not consider gapped positions.

The authors' main result is their identification of two classes of coevolving positions. The first class is composed of pairs that are near each other in the structure and therefore have to coevolve together. The second class, which the authors suggest is more significant, is composed of sites that are grouped together in ligand binding regions or active sites of 22 analyzed protein families. They also report that mutation of sites that are grouped are likely to cause mutations and compare their results with two protein families-homeodomain and E. coli ATP synthase.

Comparing SCA with selected algorithms

The development of different algorithms for calculating coupling raises the question of how different these algorithms are from each other and how different they are from statistical coupling analysis. In the next six figures I show comparisons of SCA and three other algorithms: mutual information, normalized mutual information and the chi-square test. These algorithms were chosen because they are widely used (especially mutual information). The comparison is presented for two alignments: a PDZ domain alignment (Figure 8 and Figure 9 and Figure 10) and a serine protease alignment (Figure 11 and Figure 12 and Figure 13). For each alignment, I show the symmetrical coupling matrix, histograms of the coupling values (excluding the diagonal) and a scatter plot of each coupling pattern against the other. The matrix and histograms give an overview of the data. The scatter plots should reveal if there is a relationship between the different coupling measures, although this relationship does not need to be a linear one.

As the scatter plots clearly show there is a strong relationship between mutual information, normalized mutual information and chi-square test for both alignments. On the other hand the relationship of with the other methods is poor, with normalized mutual information being the closest in the serine protease alignment. This difference of SCA with other methods means that SCA would make different predictions about what the cooperative constraints are. The correctness of these predictions can be discerned by either a much better understanding of evolution or by experimental work. The

focus of my work (and generally that of the Ranganathan lab) is experimental testing of predictions made using the SCA algorithm discussed in the previous chapter.

Figure 8: Coupling matrices for the PDZ domain family with values calculated using different algorithms. The coupling matrices are shown with values colored according to the color scale.

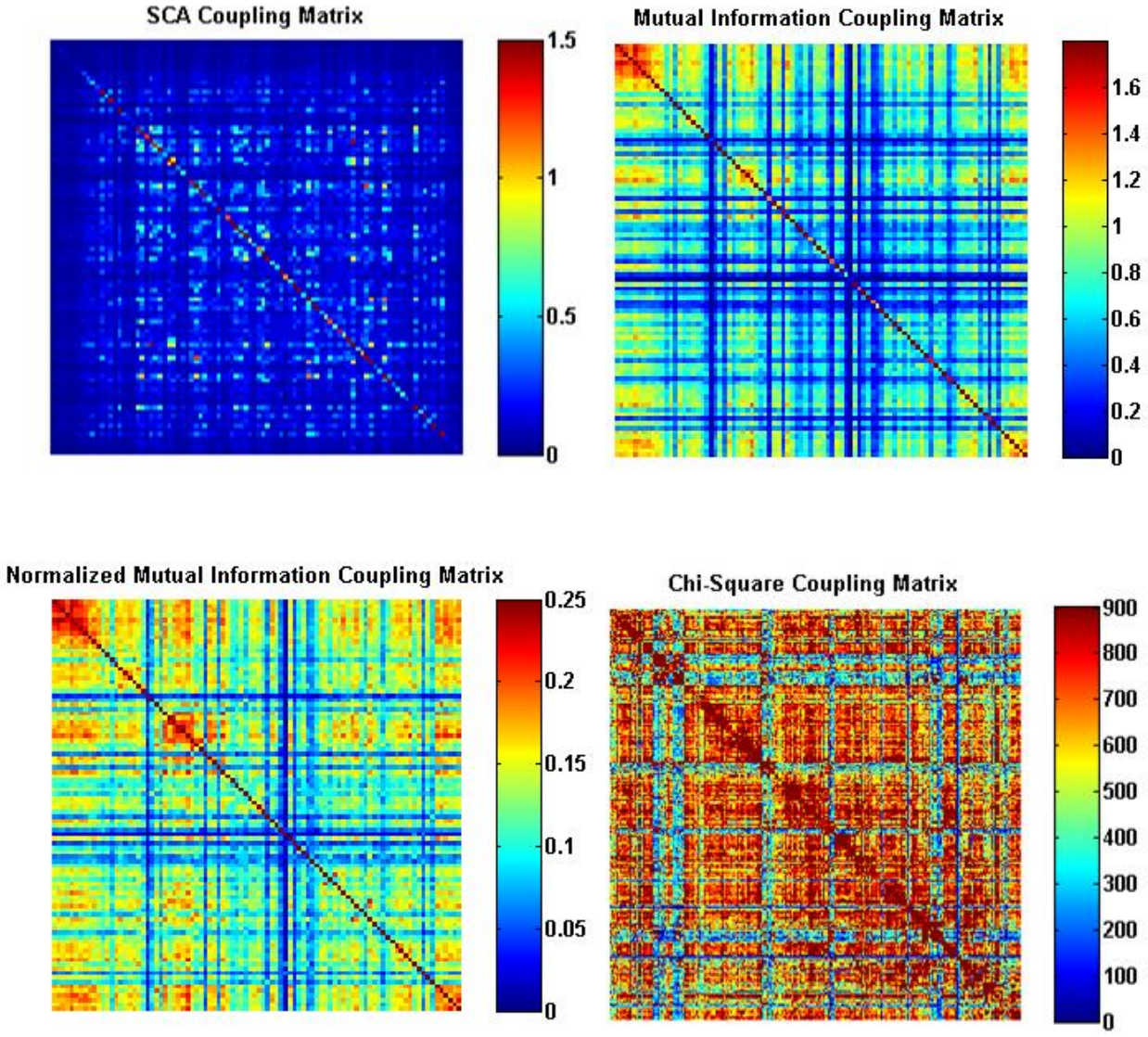


Figure 9 : Comparison of SCA with other methods for the PDZ domain. Histograms for different ways to calculate coupling are shown.

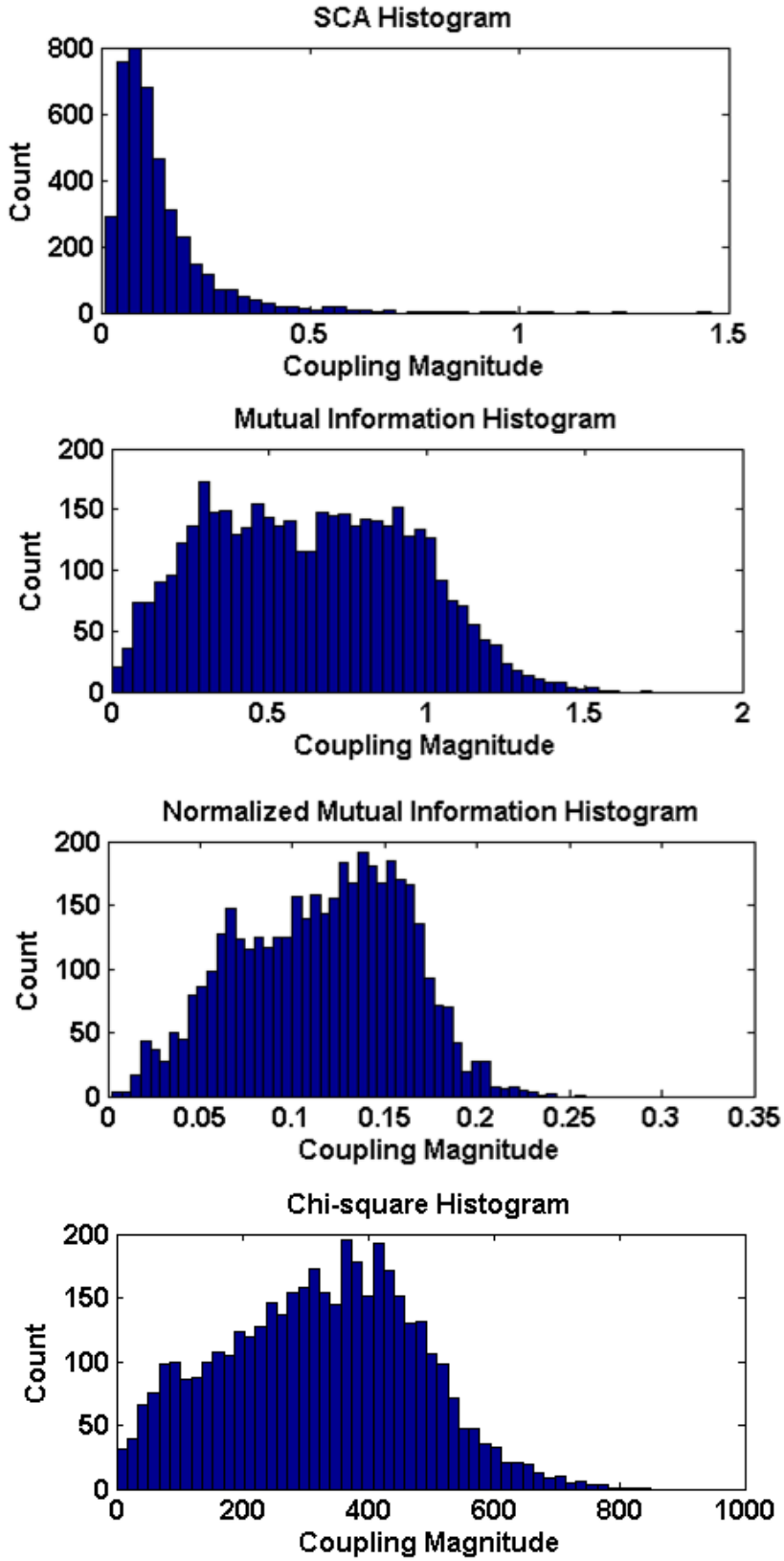


Figure 10: Comparison of SCA with other methods for the PDZ domain. Scatter plots of one coupling measure vs. another are shown.

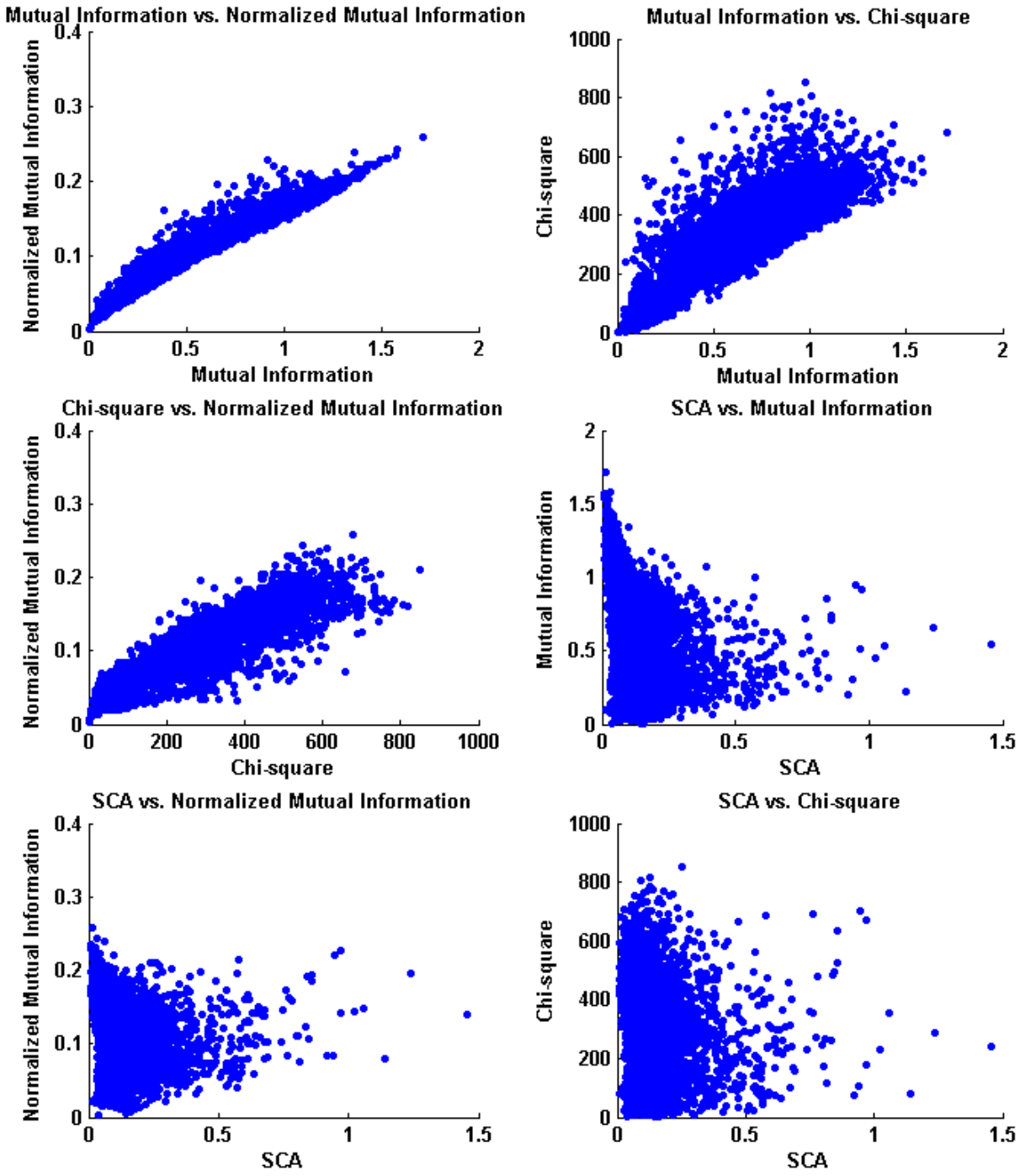


Figure 11: Coupling matrices for the serine protease family with values calculated using different algorithms. The coupling matrices are shown with values colored according to the color scale.

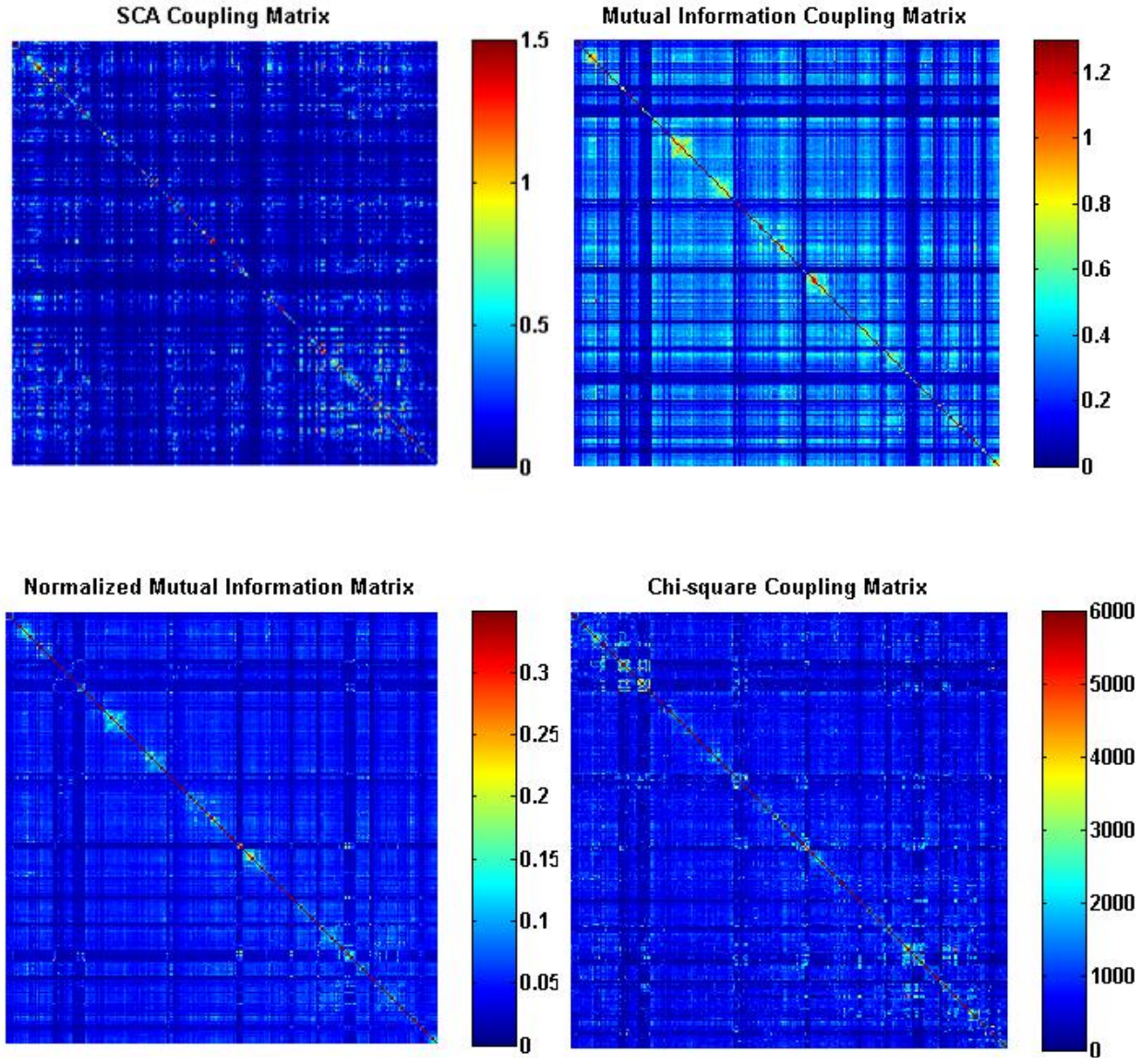


Figure 12: Comparison of SCA with other methods for the serine protease family. Histograms for different ways to calculate coupling are shown.

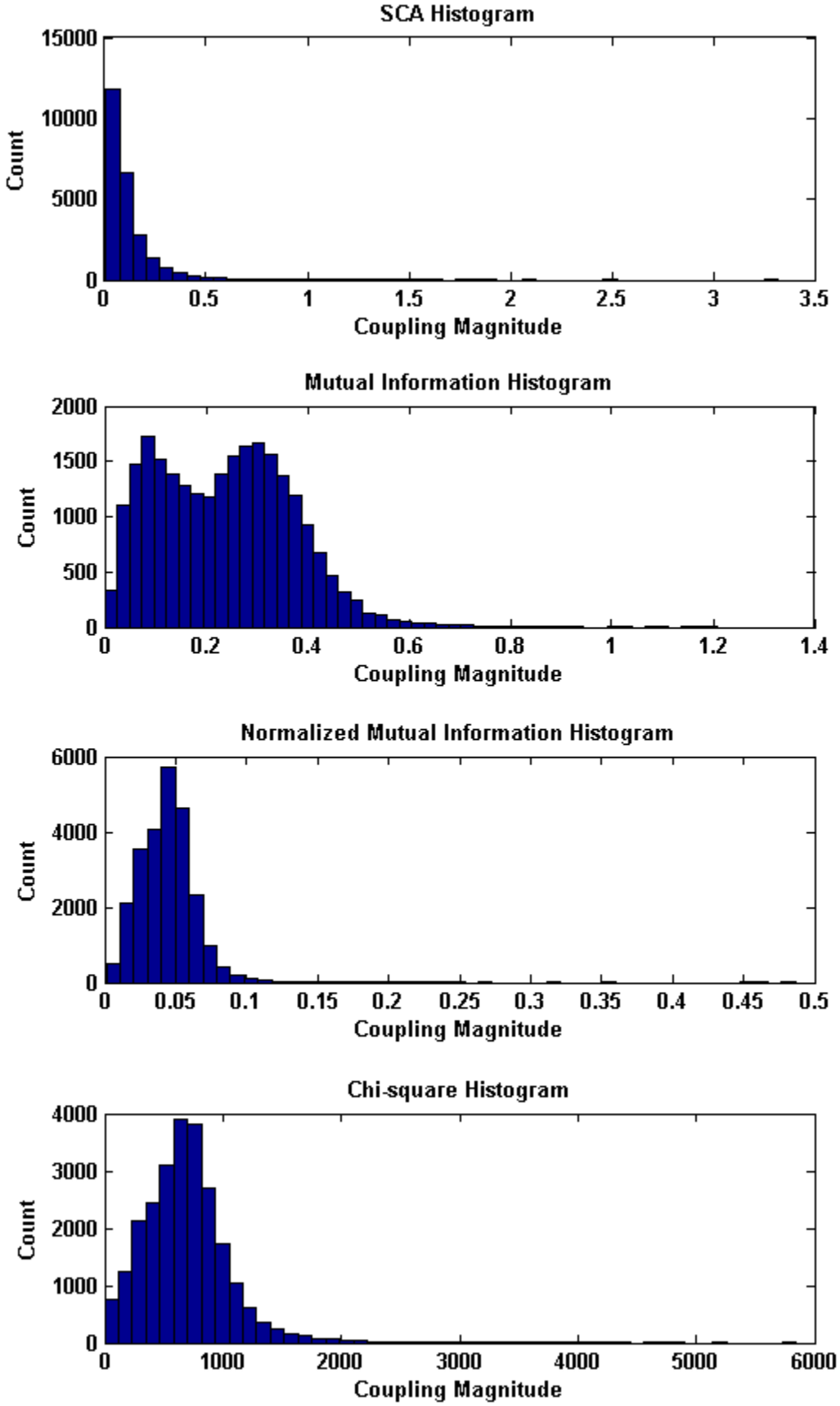
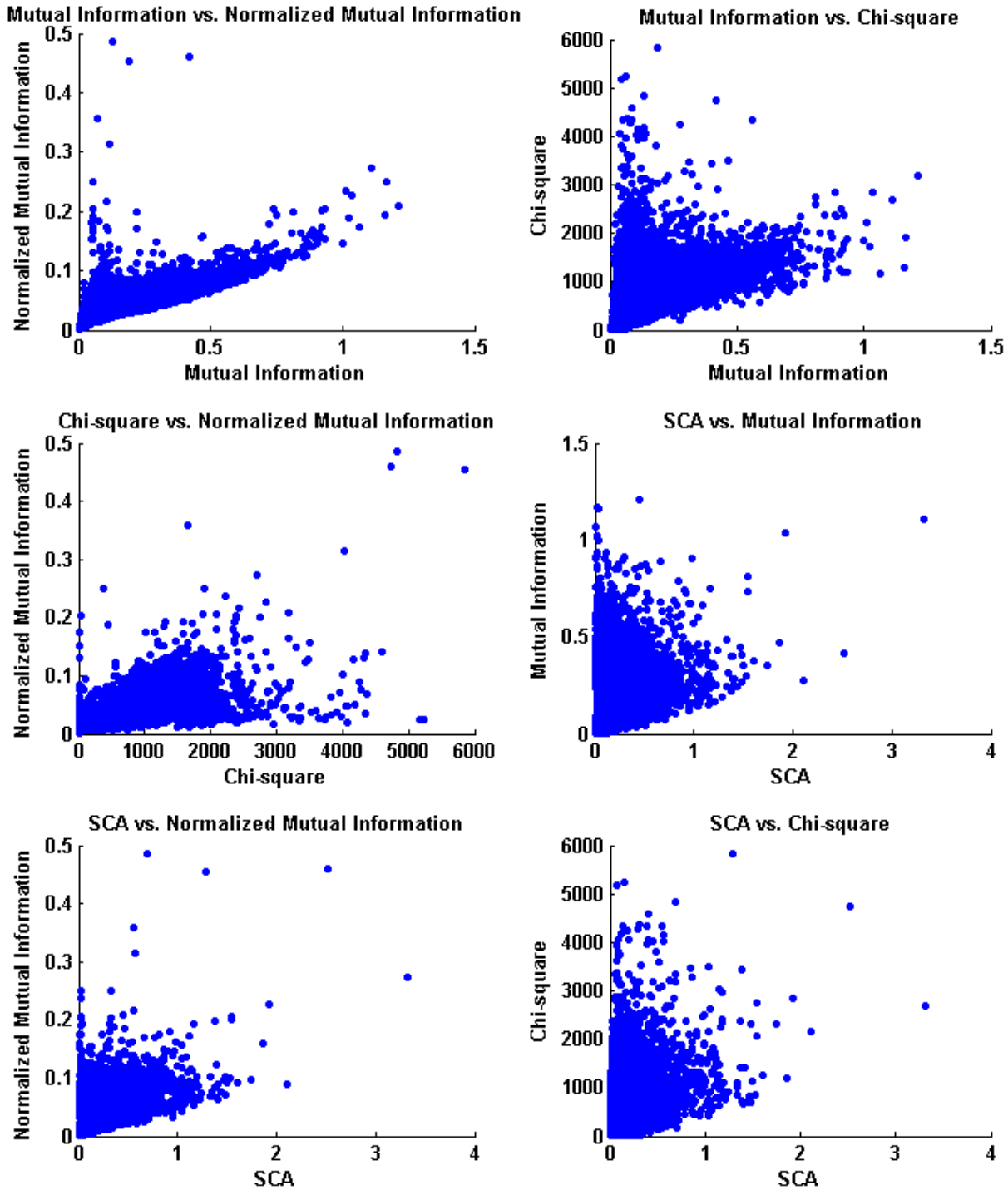


Figure 13: Comparison of SCA with other methods for the serine protease family. Scatter plots of one coupling measure vs. another are shown.



Statistical methods to correlate coupling with physical contacts:

These methods are listed separately from the methods of the previous section because they specifically look for relationships between coevolving residues and contact residues in proteins.

➔ **Neher, 1994 [69]**

Erwin Neher who had previously developed the patch clamp method, worked on coevolution too. Neher's approach will be described in some detail because later approaches are based at least conceptually on this work. The novelty of this work is that Neher not only introduced a new metric for correlation, but he also attempted to correlate the coupling to physical parameters of the protein. The main thrust of Neher's work was not the identification of correlated positions but to attribute correlation to some physical attribute of the amino acids making up the correlated pairs.

In this work, the frequency of amino acids is defined as a simple frequency although he also stated that in the general case, the frequencies should be considered as functions not only of the considered positions but also of all other positions. Neher stated that that if the positions are independent then the joint probability P of finding amino acid x at position i and amino acid y at position j is: P(amino acid i=x; amino acid j=y) equals the product of the individual probabilities.

To associate physical properties with the amino acid frequencies, Neher calculated the average charge or surface area over that position as:

mean (expected) value= $\sum_x \text{charge}^x \text{frequency}_i^x = \overline{Q}_i$

variance= $\sigma_i^2 = \sum_x (\text{charge}^x - \overline{Q}_i)^2 \, f_i^x$

The correlation was measured using Pearson's correlation coefficient:

correlation among positions i and j= $\sum_{x,y} \frac{(\text{charge}^x - \overline{Q}_i)(\text{charge}^y - \overline{Q}_j)}{\sigma_i \sigma_j} f_{i,j}^{x,y}$, where $f_{i,j}^{x,y}$ refers to the joint frequency of x and y at positions i and j.

In the next equation, Neher formulates a correlation between sequences assuming the sequences in a multiple sequence alignment are independent. The expression is similar to that above except the numerator of $f_{i,j}^{x,y}$ is 1 and numerator is just the total number of sequences and the sum is over all sequences. So what would be done is the correlation between some physical parameter (volume or charge of side chain) of one position at one sequence is compared to another position at the same

sequence. Then the total correlation for that pair of positions is obtained by summing. This works in this case because the correlations are between continuous variables (i.e. charge and volume) and not the discrete variables of amino acid identity. The expression is:

$$r_{i,j=}\frac{1}{N}\sum_{n}\frac{(\text{charge}_i^{\,n}-\overline{Q}_i)(\text{charge}_j^{\,n}-\overline{Q}_j)}{\sigma_i\sigma_j}$$

This measure of correlation would give a value of zero if the positions were independent and the sequences were independent. However, sequences are not independent as they are homologous and so any calculation of position independence could be obscured by the fact that the sequences are not independent. To get around this problem, Neher attempted to calculate the correlations among pairs of sequences by subtracting the charge at one position from the charge at another position (and not from the mean of the charges as shown in the above equation). The reader can see that if the amino acid is the same and therefore the charge, then the whole expression becomes zero. This expression is:

$$r_{i,j=}\frac{1}{N}\frac{1}{2}\sum_{n}\frac{(\text{charge}_i^{\,n1}-\text{charge}_i^{n2})(\text{charge}_j^{\,n1}-\text{charge}_j^{n2})}{\sigma_i\sigma_j}$$

One further modification is introduced to the above expression in order not to consider those positions which are zero is the sum. This is easily done by considering N as only the pairs that are not zero. So this calculation minimizes correlations due to homology by essentially removing those positions which have identical amino acids in two different sequences at a given position. This is the final correlation score given by Neher.

The last aspect of this considered by Neher is trying to calculate the variance of the data. Here, he used standard statistical analysis but modified to deal with the lack of independent sequences and the corrected expression. Without going into the details that Neher does not describe, the expression for variance used is:

$$\text{variance}_{r,i,j}=\frac{1}{\alpha N_{i,j}-4}$$

The coefficient α requires some explanation. It is obtained empirically by an iterative method from the distribution of $r_{i,j}$. It is a measure of the apparent degrees of freedom. As the number of sequences go up, the variance decreases subject to the α coefficient. With this definition of variance the noise of the data is calculated as the signal (the $r_{i,j}$) divided by the variance.

This completes the mathematical description of Neher's work. In the paper, Neher empirically determined the point at which he obtained maximal signal to noise ratio which was when the similarity between sequences was in the interval (60%-95%) for the set of 68 initial myoglobin sequences which resulted in a final analysis set of 42 myoglobins.

Briefly, the first result that Neher obtained is that the correlation between amino acid volumes was close to zero when summed over all positions and when considered over only a subset of positions known to be neighbors based on structure. This was unexpected because the assumption was that there would be correlation between amino acids due to size compensation. The second result is that when charge is considered, the correlation between all positions was zero but when structural neighbors were considered the correlation was significant.

Neher's work is very interesting because it attempts to do, from little previous body of work, a lot. He attempts to measure a correlation score and the noise of the data and then use this to see if protein physical properties are correlated. Neher also makes an intriguing statement where he says that he cannot predict the charge partner of a given position because the signal to noise ratio is too low. He calculates that if he did have 250 sequences, he would have a signal to noise ratio enough to make a prediction. It would be interesting to see if that analysis could be carried out with the large numbers of sequences and much more computing power present today.

➔ **Valencia lab, 1994, 1997, 1999 [70, 79, 80, 83]**

The next reports come from Valencia's lab. The authors introduce a rather novel way to calculate coupling. Their aim is to predict contacts in proteins. The method they use is to replace the sequence of amino acids at a position with a distance metric. The distance metric is calculated by obtaining a distance metric for each possible pair of amino acids at a position using the McLachlan distance metric (this is a matrix of chemical similarity scores). After encoding the alignment in this way, the authors calculate a correlation score which is similar to the covariation except that the denominator is the standard deviation of i*standard deviation of j instead of the total number of sequences.

Another study in 1997 also by Valencia's group continued the effort to predict intra-protein contacts. The coevolution method that is used combines correlation with conservation. With this combination the authors achieve a success rate of 30%; that is 30% of predicted contact sites based on their correlation based metric is correct based on the known structures of the analyzed proteins. The authors state that this is a promising start and plan on refining their method.

Valencia's group one year later, in 1999, published another paper with the aim of predicting contacts. Their main approach is to consider a few sequences which show covariation. They compare the contacts predicted with threading algorithm to that predicted from conservation and correlation and conclude that conservation and correlation or a linear combination of both can predict with low but significant accuracy contacts. They do not do the immediately obvious experiment which is to use the constraints obtained from conservation and correlation data in the threading algorithms and see if that improves the fit between the threaded structure and the crystallographic structure.

➔ **Hatrick, 1994 [71]**

This paper follows many of the previous papers in trying to devise a way to quantify covariations between positions with the idea that those covarying positions represent contacting positions and thus could be useful for predicting the structure of a protein. The authors in fact write the following, "The importance of correlations of this kind stems from the supposition that they might arise through the direct physical effect of one sequence position on the other. This implies that the side chains of the residues involved are close enough to interact."

The approach here is more sophisticated than previous approaches in that the authors specifically address the question of spurious correlations. Spurious correlations are those that may arise just due to historical correlations. To address this problem the authors formulated a method where in calculating correlation only non-identical sites are examined. This is similar to what Neher proposes but they do not explicitly mention his work. For example, if at a position i in a multiple sequence alignment, there are two amino acids, a and b and at another position j, there are two amino acids, c and d, correlations are not calculated when a=b or a=c or b=d or c=d.

Furthermore, the authors, to calculate correlation, assigned a 'spatial' model and a physiochemical scale to amino acids and in this way particular numbers could be assigned to the occurrence of amino acids. They then use a clustering method to cluster pairs based on their effect.

The author's conclusion is somewhat surprising for published work. They specifically write that their method has little predictive power in identifying contacting residues and that previous methods had more significant values.

➔ **Skolnic lab , 1998, 1999 [81, 84, 110, 111]**

The next series of studies from Skolnic's lab are again focused on using correlated mutations to predict contact sites in proteins but with a twist. The authors do not predict the contacts and stop there

but they simulate the folding of a protein using restraints derived from correlated mutations. The details of the method will not be described here but the authors report that their work is equivalent to that of threading algorithms (threading attempts to fold proteins by comparing sequences with unknown folds to homologous sequences with known structures).

The authors also report that they entered the CASP contest and found that they can predict with their method some unknown structures. They report that their method does well on small helical and alpha/beta proteins less than 110 residues, including those with some novel folds. Their method fails however, with beta proteins or large proteins.

➔ **Maranas Lab, 2003 [95]**

This group introduces a method called residue correlation analysis which is very similar to the method introduced by Gobel et al in 1994 [70] . The correlation between positions i and j is:

$$r_{i,j} = \frac{2}{N(N-1)} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \left(\frac{X_{ikl} - < X_i >}{\sigma_i}\right)\left(\frac{X_{jkl} - < X_j >}{\sigma_j}\right)$$

The X_{ikl} and X_{jkl} were obtained from the McLachlan scoring matrix.

As a practical matter, they excluded positions with more than 70% gaps. They chose a cutoff of a correlation of 0.4 to indicate significance based on the fact that random vertical or horizontal shuffling resulted in low correlation scores. They also considered only pairs that did not have the same sequences in both positions to minimize biasing the correlation scores to overrepresented sequences.

The author's goal however, is not to introduce this previously developed correlation score but to use it for the calculation of something they call a correlation tendency. The correlation tendency is a measure of how a 'contiguous string of residues' is correlated to several other residues. The authors are interested in regions of a protein that form contacts with another set of residues that is not contacting the original region (a cutoff of three residue away is chosen).

The authors state that the correlations are noisy and introduce a metric that aims to reduce the noise. This metric, for the correlation tendency is:

$$t_m = \frac{x_m}{I_m / L},\ x_m = \frac{\text{number of correlated pairs with at least one residue in region m}}{\text{total number of correlated pairs}}$$, I_m=length of region M, L=total sequence length

A value greater than one of the correlation tendencies is considered significant because shuffled alignments show values below one.

The authors also then state that conserved positions are important but do not show up in the correlated positions. To take advantage of this conservation they introduce a metric for conservation called site entropy:

$S_i = \sum_{x=1}^{20} p_x \log_2 p_x$, the probabilities are the probabilities of amino acid x at position i.

Using these metrics that authors attempt to understand whether covariation can capture any information that is not captured by conservation and whether both methods together can yield more information.

Their conclusions, briefly, are that correlations, when signal is high, identify functional sites better than contact maps and conservation and that the combination of correlation and conservation is best for capturing most of the functional sites.

➔ **Halperin, 2006 [104]**

This paper uses correlated mutation to measure contacts between two different proteins. They find that in one protein correlated mutation analysis works but in another protein family it doesn't and they therefore conclude that this method is not suitable for aiding in docking which is different from what other people have published.

➔ **Kundrotas and Alexov, 2006 [105]**

This is yet another paper where the author tried to use correlation mutations to predict physical contacts. The authors believe that they achieved an advance over previous methodologies by imposing a set of filters and optimizations on the multiple sequence alignment prior to the predictions. The filters are:

1) Removing non-homologous sequences (<20% identity) and removing highly similar sequences (>90% identity).
2) Ignoring absolutely conserved positions
3) Obtaining sub multiple sequence alignments
4) Incorporating a minimum degree of conservation for the calculation of coupling in the sub alignments.
5) Filtering the predicted positions by some physiochemical inclusion list. That is the authors chose allowed residues based on whether they form hydrophobic pairs, ion pairs, hydrogen bonds and disulfide bridges. Pairs of residues that did not make chemical sense were excluded.

Then the authors varied each of the variables to find which threshold maximized the predicted residues for several families which have high resolution structures. The net result of all this is that the true predicted residue pairs were about 0.09 on average which is on the low end of other methods. The authors claim that this can be improved with further work.

➔ **Dunn's lab, 2008 [107]**

The authors of this report adopt a coupling method that is similar to SCA in spirit. They develop a new method to remove background mutual information from actual mutual information values in order to better estimate correlations due to contacts.

The idea is quite simple. The authors take two unrelated protein families and make a joint alignment. Then they calculate the mutual information between positions in the two families. Then they make what they say is a surprising observation that the mutual information between positions in the unrelated families can be estimated by the average mutual information. Then the average mutual information is factored out of the calculated mutual information.

This paper is similar to SCA in spirit because with SCA the probability of every amino acid is based on a background probability. Here, the background mutual information comes from all the alignment and not the individual amino acid but the idea is the same: to account for a background level of correlations. The authors claim that their method increases by three to four fold the prediction of physically contacting residues.

Experimental approaches

Of close to 50 reported papers on coevolution that are by other labs, only 3 that I am aware of attempt to do an experiment to test the significance of coevolution. These three are briefly described here.

➔ **Altshcuh, 1992 [67]**

The first experiment I am aware to test the significance of correlated sites experimentally was made by Altschuh's lab in 1992. The authors took one pair of positions in the cysteine protease, papain, and investigated whether the 'complementary' pair identified by a multiple sequence alignment rescues function compared to the single mutants. They find that one single site mutation kills the enzyme which can be rescued by another mutation at another site. A single mutation at the other site has small effects.

This study is noteworthy as it represents a test of how a statistically coupled pair can actually function in a protein.

➔ **Horovitz, 1994 [73]**

Another early experimental work comes from Amnon Horovitz's laboratory in 1994. Horovitz was one of the first to work on thermodynamic mutant cycles along with Alan Fersht. In this short report , they investigate whether thermodynamic coupling is observed in co-evolved pairs. First, they identified a pair of correlated sites in GroEL (a chaperone protein). Then they made double mutant cycles on this pair and concluded that the effect of the mutants was non-additive suggesting that the sites were interacting cooperatively in the native protein. The authors do not describe the method for calculating covariation and instead only say that it is so and show a table with the claimed, correlated site. The value in this paper though is the experimental finding that the predicted cooperative sites were non additive. This paper was the first to make explicit the link between thermodynamic mutant cycles and statistical coupling.

➔ **Laub, 2008[109]**

This year Laub's group published a paper where they changed the interface between two proteins that contact each other based on the coupling between them. The system used was bacterial proteins that are in the two-component signal transduction system. These proteins consist of a histidine kinase and a response regulator. Both proteins take part in the same function and exist on the same operon. They used this fact to make an alignment of the two proteins and then they used mutual information to calculate the coupling between and within domains. The coupling data between domains identified clusters of residues that could be the binding interface between the two proteins.

The authors then switched the coupled region along with the connecting loops of one protein to another and they found that the switched protein exhibited the specificity of the protein from which the switch regions were obtained. This was done for several proteins and for each protein the authors were able to transfer the specificity just by transplanting the coupled positions and connecting loops. This is the first experimental demonstration of such a specificity switch using coupling data.

Learning algorithms

Several groups have attempted to use learning algorithms to see if they can develop rules that would enable the prediction of contacts. None of these studies was very successful.

➔ **Learning algorithm, Thomas and Sander, 1996 [76]**

The work from Thomas and Sander continues the trend of using coevolution to predict contact sites in proteins but without much success as acknowledged by the authors. This paper attempts to correlate contacts with correlated mutations using a learning algorithm. They take a set of alignments for which there are known structures. Then they bin the contacting pairs of sequences and the non-contacting pairs into different categories. They use this as a training set. They compare the predictive power on a set of unknowns. Their method achieves only a 1% improvement in predictive power over previous methods, which the authors say is,"...weaker that we might have hoped."

➔ **Neural Networks, Fariselli and Casadio, 1999 [85], Valencia's lab 2001 [89, 112]**

In this study, neural networks are designed that integrate chemical and evolutionary information with the aim of predicting contact maps. The authors report an improvement over previous methods in that their method is independent of protein size but still, to quote, 'far from useful' owing to the fact that the fraction of predicted contacts is never more than 26% accurate.

Another method published by Valencia's group in 2001 uses a similar idea as the Thomas and Sander paper. The new method tries to incorporate coupling information into a neural network with the goal being to improve the prediction of contacting sites. The details of this approach will not be discussed but the authors report that without the neural network they predicted 15% of all contacts but when incorporating neural networks they obtained an accuracy of 21% which represents a 6-fold increase over random guessing (random guessing would give an accuracy of ~3.5%).

➔ **Maximum likelihood, Pollock and Taylor, 1999 [82]**

Similar in spirit to neural networks, Pollock and Taylor who wrote a 1997 review critical of methods used for correlated mutation analysis developed in this paper a new method to deal with some of the uncertainty regarding what is the best way to calculate correlations between sites. The authors here use a maximum likelihood method to calculate coupling in an effort to reduce the effect of random correlations which they say affected previous methods. Their method modeled evolution using complementary states of amino acids. The complementarity was based on physical aspects of amino

acids such as the size of the amino acid or the charge. So inherent in their method is a certain assumption for what compensating mutations should be.

In general, maximum likelihood methods choose some model for the data and then calculate using the model, the parameters. Then, parameters are compared to experimental data to see which model most closely resembles the experimental data. Every model has some free parameters which are simulated during each calculation.

The authors used two models of evolution, a model where amino acids are independent and a model where amino acid mutations are dependent according to an evolutionary scheme. The details of the models are rather complex and are essentially an evolutionary model that will not be described here. The authors conclude that their method can detect coevolution in both simulated and real data. They also say that residues that are strongly co-evolving are more likely to be near each other. They also admit that they used size and charge arbitrarily and say they will work on other ways to group proteins.

Inter protein coupling

These studies attempt to calculate the contact sites of two proteins based on the co-evolution of the two proteins.

➔ **Pazos and Valencia, 1997 [79]**

The first study by Pazos and Valencia widens the scope of correlated mutation analysis by looking at the coevolution between two proteins that interact. The idea is that if two proteins interact together then the residues that mediate that interaction would co-evolve together even though they are on different peptides. The authors used the correlation metric introduced before (Gobel et al, 1994 [70]) but introduced a new method to compare distances between pairs in actual structures.

They test their method on known interacting pairs. They find that sequence correlation data can predict which of the docking solutions is best. This to them is strong evidence that their method can find contacting interfaces. Then they make a prediction for interacting pairs using a protein with an unknown structure to them. They find that their method could predict the interacting regions at the interface although they do not do any mutagenesis.

The value of this work is it is the first attempt to my knowledge at using correlated mutations to predict interactions between two different proteins.

➔ **Filizola, 2002 [91]**

This work is aimed at identifying interfaces between proteins. It is based on previous work discussed above, but in this work the authors claim to have developed a new method called subtractive correlated mutation (SCM). This method is easy to explain: the authors take two similar protein families and then compute four alignments. The first alignment is of protein A. The second alignment is of protein B. The third alignment is A appended to B so A and B seem like one protein(A+B). The fourth alignment is AB where A and B are aligned together because they have similar structures. Then each alignment has its coupling matrix calculated using previous methods. However, to obtain the intermolecular coupling the authors subtract the coupling values in the following manner: intermolecular coupling=Coupling(A+B)-Coupling(A)-coupling(B)-coupling(A,B). Then the set of positions identified by intermolecular coupling are filtered based on solvent accessibility (as the structures are known, this is possible to do reliably).

This method was applied specifically to the opiod receptor heterodimers, which are members of the GPCR family where specific interactions were found. A negative control where no interfaces were predicted was used also.

➔ **Laub's lab 2008 [109]**

This paper is a very interesting experimental paper where the authors measured coevolution using mutual information between two bacterial interacting proteins in the same operon. They were able to then redesign the interaction surfaces between the two proteins based on the coevolution pattern. This study was discussed further in the experimental studies section.

Considering phylogeny

Understanding the phylogenetic history of individual positions in alignments can allow for more accurate extraction of functional signals. The studies reported here attempt to model evolution. The evolutionary methods are not discussed in any detail but the results and conclusions of the authors are reported.

➔ **Benner's lab, 1997 [78]**

A paper in 1997 in Benner's lab partially addressed the concerns raised by Pollock and Taylor [77]. This paper has a combination of several novel approaches to calculate coupling. To begin with, the authors only considered interior positions (surface accessibility <40%). Second, they defined

'simultaneous variation' as two positions in aligned proteins that have both mutated. Third, they subdivided the 'simultaneous variation' into proximal and distal variation. The proximal variation is variation between residues that are close (<6 angstrom) in the structure whereas distal variation is variation between residues that are farther than 6 angstrom in the structure. They only considered proximal covariation in their analysis as they could hope to explain the covariation based on specific amino acid property compensation.

The calculation of 'simultaneous variation' is novel and I will explain it in more detail. Consider two positions in a sequence alignment at sequence a: say at position i there is amino acid A and at position j there is an amino acid B. Now consider another sequence: at position i there would be an amino acid C and at position j there would be amino acid D. The mutated pair would be: A->C|B->D. As there are 20 amino acids and every amino acid can be changed to any other one, there are a total of 400 pairs for a single site and 400 pairs for a second site, and to describe the changes at both sites requires a 400x400 matrix. The authors excluded changes that occur as (A->A|C->D) and so the final data set has 36290 elements rather than 160,000 (400*400). However, the authors did not use these covariances. Instead they categorized them based on the physical properties of the amino acids. Amino acids could be volume compensated, hydrogen bond compensated or charge compensated.

The authors also introduced a weighting scheme based on constructing phylogenetic trees from their sequence data. They looked at the effect of different evolutionary depths on the covariation. The main result is that at different evolutionary depths, different types of physical compensation become detectable. That is, when sequences are at intermediate distance, the covariation tends to account for hydrogen bonding. At low evolutionary distance, volume compensation is present. At all distances, charge compensation is present. The authors also see if their methodology can predict contacts but they cannot improve on previous methods.

The idea that at different evolutionary distances, different effects are possible is quite new and may be useful to think about although the specific compensations that the authors use may not be correct.

➔ **Tuffery and Darlu, 2000 [86]**

This paper attempts to model phylogeny as a way to get at the significant correlations. This was the most extensive model of evolution yet built to detect significant correlations in multiple sequence alignments. Only an outline of the methods and results is given here.

First, the authors constructed a tree of the sequences using maximum parsimony. Then using two different approaches (parsimony and maximum likelihood) they constructed ancestral sequences at the nodes of the tree. Then, they simulated the evolution from the ancestral sequences to the known sequences.

From the simulated evolution they were able to get at what they call 'cosubstitution', which is two different sites undergoing substitution in the same tree branch. The significance of the 'cosubstitutions' was evaluated via two evolution-based approaches the details of which will not be discussed here. For both approaches however the idea is that one can arrive at an expected substitution given that evolution occurs at sites independent than at others sites. These expected substitutions can be compared to the observed one and a significance ascribed.

One key finding they reported is that in order to get significant cosubstitution they need to categorize the amino acids into physically meaningful pairs. Their overall conclusion is that the coevolving pair detection is sensitive to the method and reliable analysis of co-evolution requires further work.

➔ **Wollenberg and Achtley, 2000 [87]**

This paper clearly describes the problem of detecting covariance between positions. The authors list that there are three sources that could result in coevolution between positions: 1) chance 2) phylogeny and 3) structural/functional constraints. The goal, as the authors state, is to understand what signals come from chance and phylogeny and what can come from functional constraints.

To achieve this goal, the authors constructed three sets of data. First is the coevolution of the natural alignment calculated using mutual information. The second set is the coevolution of an alignment reconstructed from a phylogenetic tree of the data and a substitution matrix of the data. The third set is made by computing the tree from the natural data but using a substitution matrix made from all protein alignments. Then the authors compared the data sets.

According to the authors in a way not fully explained, the third data set preserves the phylogenetic and chance properties of the alignment. The second data set preserves the functional data as well as the phylogenetic and chance data. Therefore, if one wants to extract functional data, all one needs to do is to take the threshold set by the third data set as things that are not functionally important.

To test this idea, they used the basic-loop-helix motif. They found that the mutual information of the phylogenetic data set are much less than that of the natural data set and the model of the natural data set. They therefore conclude that they can extract functional information with a particular

probability based on this method. The actual method used to make the reconstructed data sets is not described but is referenced. Other than that, the paper is quite well written and is a good attempt in trying to separate chance and phylogeny from function.

The same authors published another paper [87] that resembles very much the previous one but is more extensive it its explanation. They even use the same protein family. The only extra type of information they provide is the construction of network graphs of residues with mutual information above a certain threshold.

➔ **Dufton's lab, 2001 [90]**

This seems to be a set of rules for calculating covariation based on the detection of significant blocks. A block is a set of positions that covary together perfectly. The authors correctly consider that a block made up of one sequence only is not significant. However, the exact calculation of correlation between positions is not explicitly defined. The authors then simulate evolution and apply their correlation analysis to see if their method can detect true correlations. They were limited to 60 sequences. They calculated both the noise (defined as the correlations at positions that they did not choose to be correlated) and the signal (the number of pairs that were identified to be correlated at the sites which were designated to be correlated). The signal/noise ratio was then calculated which expresses the proportion of true correlated changes. This was done for different evolutionary models.

Conclusions: number of sequences should be high to minimize noise and get a high signal (>16), most reliable results are when there are large data sets with high sequence divergence and high rate of substitution. This paper described the effect on correlations of different evolutionary methods and as such it is useful.

➔ **Fleishman, 2004 [99]**

This is another phylogenetic based correlated mutation analysis. An additional aspect of this work is that the authors used principal component analysis. The authors used their analysis on potassium channels and claim they need 50-100 sequences. The rationale behind phylogenetic approaches is to simulate the evolutionary path from hypothetical ancestral sequences to the extant sequences along a tree based on the extant sequences. The idea is to consider changes that occur in one branch of the tree. In addition to this, they weight the amino acid mutations differently based on the identity. As they say, a valine to isoleucine change is considered of smaller magnitude than a glycine for a tryptophan. These changes are accounted for by the Miyata substitution matrix.

The authors using this method computed basic correlation coefficients and calculated the significance of the coefficients via bootstrap sampling. After obtaining a set of significant correlation values (note that the correlation values were between pairs of amino acids at positions) the authors performed a principal components analysis to detect networks of coupled positions.

The results are the identification of networks of correlated amino acids that correspond in some cases to known functions of the potassium channel.

➔ **Dutheil 2005 [100]**

This work introduces yet another method for calculating coevolution. The basis of the work is to construct a detailed phylogenetic model of evolution of each site. This allows for the estimation of ancestral sequences. Correlations are then calculated between the ancestral sequences. According to the authors the use of this method predicted 95% of contact sites in a rRNA alignment. One issue with this work is that the authors state that the method could be used for protein and nucleic acid sequences but do not show a coevolution example for protein sequences.

➔ **Horovitz 2005 [102]**

In this paper, Horovitz's group modified the approach they used in their previous paper by incorporating by their own admission, ideas made by Wollenberd and Achtley (2000) [87]. They still used a chi-square test to rate significance but in this case they modified the calculation of observed sequences. In addition, they also tried to select sequences for analysis based on an evolutionary tree. According to them this method helps to reduce noise in the alignment.

The first difference they made is to shuffle a position i vertically many times. This has the effect of removing all correlations. For each shuffle they calculated a chi-square value. At the end of their shuffling trials they had a distribution of chi-square values. They then calculated a p-value to estimate the significance between the chi-square value calculated before shuffling and after shuffling. The reason for doing this, as the authors state, is that in certain cases the chi-square test results must be discarded, whereas in this case the results do not have to be discarded.

The second modification that they did is to select 20% of the sequences randomly and then 'permute' each sequence of the selected 20% with another sequence in the alignment. However, this 'permutation' was not done randomly but the probability that any given sequence is permuted with another depends on the evolutionary distance between the sequences. It is not clear from the mathematical description how the evolutionary distance is calculated but I presume that a sequence has

a higher chance of being permuted with a sequence that is like it. In this case, permutation I think means switched. Finally, the distribution of shuffled positions was compared with the non-shuffled position. In this case, this means that the correlations obtained with this method reflect ancestry or noise.

In order to estimate evolutionary noise what the authors did is to obtain sequences of non-interacting proteins from the same species and combine them. Then from another species they picked the same pair and they made a multiple sequence alignment of all of these non-interacting pairs. They then calculated the degree of coupling between positions. What they found is even though the two proteins did not interact they detected a large amount of coupling. However, when they vertically shuffled one of the proteins these correlations went away. Their conclusions are that the coupling is easily observable and reflects noise. One thing to note about this is that the authors did not distinguish the magnitude of coupling. They just chose a cutoff p-value.

When the authors applied their evolutionary weighting method they observed that both intra-residue and inter-residue correlations decreased. They then calculated a signal to noise ratio and found that for most of the protein pairs the signal to noise ratio increased.

My criticism of this paper is that due to incomplete explanation I cannot follow why shuffling based on evolutionary distance decreases the p-value relative to not shuffling and second, why the calculation of evolutionary distance depends on the correlated positions.

Finally they compare their correlation analysis to distance metrics of proteins and find that their method does better at revealing correlations that appear due to distance measures.

➔ **Fares and Travers, 2006 [103]**

This study aims to correct for phylogeny, 'replacement propensity', 'background sequence divergence' and 'three-dimensional' information, using a novel method that they describe.

They call their method CAPS for coevolution analysis using protein sequences. The idea is not to use the correlation between amino acids at a position as in other work but the correlated variance of the evolution of particular sites. An additional correction is applied which based on the estimated time of divergence between sequences. This is therefore a more complicated approach than used in other work. The authors also compare their work with the work of Pollock (parametric method), Tillier (dependency) and Korber (mutual information). The main advantage of this method is that it is highly sensitive and does not require large alignment sizes. It would be interesting to examine the methods in this work more closely.

Miscellaneous work

➔ **Evolutionary trace, Cohen's lab, 1996 [75]**

The next work in 1996 by Cohen's lab is not a coupling method in the sense of calculating the covariance between positions in a multiple sequence alignment. Instead of just relying on the multiple sequence alignment, phylogenetic trees of the sequences are used to select sequences to examine. The authors are interested in identifying functional sites with this method that they call an evolutionary trace.

The starting point is to make a multiple sequence alignment. Then a dendrogram is made of the sequences. This dendrogram groups sequences based on function. At a particular branch point of the dendrogram, groups of sequences can be defined. Subalignments are extracted from each group and then a consensus sequence is made of each subalignment. Then the subalignment consensus sequences are aligned together and a consensus sequence of the consensus sequences is made.

In this scheme, all positions can be classed into three categories: conserved, neutral, or class-specific. Conserved means that a position is 100% conserved. Neutral means that a position has some variation in the original consensus sequence. Class specific means that a position has some variation in the consensus of the consensus sequences.

These steps are repeated for different partitions of the dendrogram and then, the categorized positions are mapped onto the structure. The main result observed for two protein families (SH2, nuclear receptors) is that conserved and class-specific residues often form clusters that are regions where a peptide bonds. They also observe non-clustered positions but they state that they cannot interpret the functional significance of those sites.

Review of reviews

➔ **Hatrick and Taylor, 1994 review [72]**

The emphasis of this paper is a review of using conservation and correlation to estimate contacts in proteins. With respect to correlation, the authors review four methods including their own discussed above. They state that the Altschuh method which looks at pattern recognition is faulty because the patterns could be artifacts of the conservation pattern. They also state that that method is

hard to implement for many sequences. They also consider another example from the literature where Benner and Gerloff in 1991 [113, 114] identified covarying residues but they did not ascribe much validity to the method. The third method criticized is that of Oliveira et al [1993] where some positions where identified but without a structure are impossible to verify. For their own method the authors criticize their work by saying that while they wanted to solve the problem of conservation they were unable too because when conservation is taken into account the correlation signal is quite weak.

➔ **Pollock and Taylor, 1997 review [77]**

The authors of this 1997 review, Pollock and Taylor tested different correlation methods (Neher, 1994 [69]/Gobel et al, 1994 [70] /Taylor and Hatrick, 1994 [71]) on simulated evolutionary trees. Their conclusion is that no method can detect a significant number of true correlations without also including correlations due to phylogenetic structure of alignments. They go further and write that this may be impossible to address properly owing to the fact that it is not possible to make alignments in the first place that contain no phylogenetic information as alignments are built using homologous sequences that by definition are phylogenetically related. They suggest but do not provide evidence that the best approach to analyzing sequence information is to model the alignments phylogenetically to extract true correlations.

➔ **Fodor and Aldrich, 2004 review [96]**

In this paper, Fodor and Aldrich do not introduce new algorithms or data for coupling but attempt to see whether coupling by two different algorithms is correlated with distance in a protein and with thermodynamic mutant cycles. They write that they observed that statistical coupling with two different algorithms (Gobel score and SCA score) correlates with distance. This is true to some extent but the correlation is rather weak. The authors also compare thermodynamic mutant cycles in the PDZ domain (for binding), staphylococcal nuclease (folding) and shaker channel (for conductance) with distance and the Gobel score. They find that the correlation is poor for these systems. However, the authors only considered linear correlations but the relationship between sequence covariation and thermodynamics does not need to be linear. Finally the authors correlate thermodynamic coupling with distance and report that the correlation is strong although this varies among different data sets and it could be non-linear as well.

This paper is an attempt to understand the meaning of statistical coupling in terms of physical properties of the protein, which is what other groups tried to do. The authors also use the paper to

criticize the claims that they believe were made by proponents of statistical coupling analysis. Specifically they write that it is possible to observe thermodynamic coupling that is not predicted by statistical coupling analysis. They also imply that, if a structure is present, it is possible to correlate distance to statistical coupling and it is therefore not very useful to determine statistical coupling because distance is also correlated with thermodynamic coupling.

My viewpoint is that distance metrics, thermodynamic coupling and statistical coupling measure related but not identical constraints. There are problems with each of these metrics and thus the comparison between them to find out which is better than the other has many problems not acknowledged by Fodor and Aldrich. For example, thermodynamic coupling data is not readily available and is also fraught with problems because mutants may change other parameters of the protein and the system itself may change. In addition, there is a problem of what to mutate into. Moreover, the relationship between the thermodynamic coupling and statistical coupling need not be linear, a point not discussed by the authors.

➔ **Fodor and Aldrich, 2004 review [97]**

In this paper a comparison of different algorithms for calculating covariance is made. The authors compare OMES (Kass and Horovitz [93]), mutual information (Atchley et al [88]), SCA (Lockless et al [58] , McBasc (Olmea et al [83]).

The authors also introduce a measure of conservation at a position as the absolute sequence entropy:

$$\text{sequence entropy}=-\sum_{x}(p_x(i)\ln p_x(i))$$

The authors point out the obvious that correlation only works when there are intermediate values of conservation. This is because when a column is 100% conserved there is no correlation. If the conservation is very low then the correlation is also zero or close to zero.

The authors construct an artificial alignment and compare the behavior of the different methods on it. They find that all methods had the highest correlation between positions in which there is 50% MM sequences and 50% YY sequences (the alignment had only two columns). What differed, however, was the distribution about this maximal value. All methods had different distributions. The authors also compared the performance of the algorithms to random positions and found differences as well. Then they compared the behavior of the algorithms to real alignments and found that the different

algorithms had different shapes with some algorithms giving higher scores to less conserved positions. The third aspect that the authors looked at is the correlation between correlation scores and distance between positions. They found, again, a relationship between correlation score and distance that is slightly different for different positions. The authors also examine the predictive power of the algorithms for residue contacts and find that the algorithms generally have low accuracy. The final figure of the paper shows the pair distance prediction for multiple protein families. Again the algorithms perform differently.

The authors conclude the article with an interesting comparison of their analysis of the correlation algorithms with the claims made for the correlation algorithms. They single out SCA for special treatment. One criticism is that SCA has low power and hence the conclusions that the energetic interactions in a protein are sparse may be due to its low power and not due to actual sparse interactions. The authors also have a problem with all methods in that only the highly covarying positions are significant. They question whether the information in the highly covarying positions is actually very useful in describing thermodynamic coupling and residue distances in proteins. The authors end with the idea that different methods filter conservation in different ways and that some methods may be better suited for alignments of particular conservation.

Generally, the authors did a lot of analytical work in comparing the different methods. They also point out that what SCA attempts to measure is actual thermodynamic coupling and not distance (although they are skeptical of the idea that most adjacent residues in proteins are not thermodynamically in contact).

➔ **Fuchs and Frishman, 2007 review [106]**

This is a comprehensive study on the general applicability of correlation analysis to membrane proteins. The authors used nine different correlation algorithms (all the ones discussed previously) to analyze coupling in several membrane protein families. The goal again was to predict contacts in membrane proteins. In this case, though it's acknowledged that only about 10% of correlated residues are contacting so the authors used a previously used measure based on proximity. That is, they looked whether correlated positions are near each other. They find that, with this measure, close to 50% of correlated residues are in 'close vicinity to interhelical contacts'.

This paper is laudable in that they compared nine prediction algorithms and pooled the information from each. They also state the both the SCA method and mutual information performed poorly in calculating coupling and so they excluded them from further analysis although they also state that for

one protein family mutual information was the best measure at predicting contacting residues. The authors claim that their approach would be useful to predict the structure of contacting domains.

Conclusion

Extracting correlations from a multiple sequence alignment that reflect cooperative interactions within a protein is complicated by the phylogenetic history of the sequences. Deconvoluting cooperativity from phylogeny has been the key difficulty that has been addressed both in SCA and in other methods.

It is clear from the comparisons between SCA and mutual information and chi-square based methods that there are fundamental differences in the methods reflecting differences in an assumed evolutionary model. In the absence of true phylogenetic histories and true cooperative interactions, the correctness of each method can only be established by experimental evidence; i.e. methods should make testable predictions that can then be compared with experiments which is the approach taken in my work.

One specific comparison widely made in the literature is the comparison of correlation data to the contact sites in proteins. The reason to predict contacts is that the contact constraints can greatly aid in predicting the fold of the protein. However, a consistent them of these contact based methods has been that they do not predict contacts very accurately (on average only 20% of all predictions are real contacts). The modest success could mean that phylogenetic information has not been removed properly or it could also mean that contacts in proteins do not, in themselves, provide strong constraints on evolution. Further work will establish whether this is the case.

The most promising methods to separate phylogeny from function are phylogenetic reconstruction methods. These methods attempt to reconstruct the phylogenetic history of the sequences comprising an alignment allowing one to know whether observed correlations come from functional constraints or evolutionary history. Phylogenetic methods can be quite complex but they are worth examining in detail as they directly address the problem confounding coupling analysis.

Chapter 3: Analysis of statistical coupling matrices

A statistical coupling matrix, introduced in Chapter 2 (also in Figure 14), shows the coupling of every position in an alignment with every other position. In some cases, there is strong coupling between two positions and in others there is weak or no coupling. One can attempt to visually identify patterns in the data by looking for positions that have similar coupling patterns. However, the complexity of the patterns requires methods other than visual inspection. The goal of this chapter is to explain four methods to analyze the coupling matrix: network graphs, hierarchical clustering, principal components analysis and independent component analysis. Each method has its advantages and disadvantages and these will be discussed as well.

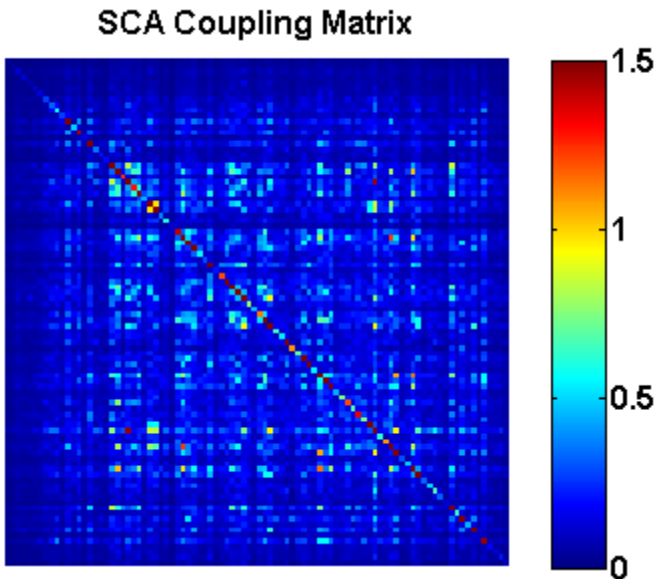


Figure 14: Example of a coupling matrix. This matrix comes from an alignment of 240 PDZ domains. The coupling matrix is shown with values colored according to the color scale.

Section 1: Network graphs

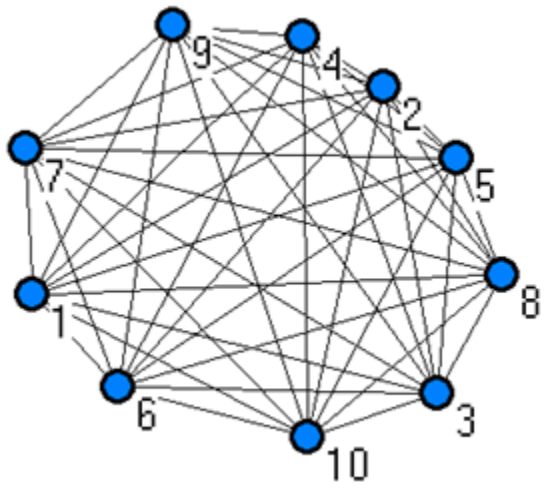
A useful tool to visualize coupling between positions is to draw a network graph. The network graph is a graph where every amino acid is represented as a point, also called a node, and the coupling between positions is represented by a line.

To illustrate the idea of network graphs I will first show a simple example consisting of only ten nodes (whereas a typical protein has a node for every position in the alignment (92 positions in the case of the PDZ domain alignment). The covariance matrix that will be represented as a network graph is shown in Figure 15. The network representation of that covariance matrix is shown in Figure 16.

Figure 15: Covariance matrix consisting of 10 positions. The shaded region is the diagonal which is the variance and is not required for this analysis. The matrix is symmetric about the diagonal.

	1	2	3	4	5	6	7	8	9	10
1	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02
2	0.01	0.05	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03
3	0.01	0.02	0.12	0.03	0.03	0.03	0.03	0.04	0.05	0.05
4	0.01	0.02	0.03	0.11	0.02	0.03	0.02	0.04	0.03	0.04
5	0.01	0.02	0.03	0.02	0.06	0.02	0.02	0.04	0.03	0.04
6	0.01	0.02	0.03	0.03	0.02	0.12	0.03	0.05	0.03	0.04
7	0.01	0.02	0.03	0.02	0.02	0.03	0.13	0.04	0.03	0.04
8	0.02	0.02	0.04	0.04	0.04	0.05	0.04	0.27	0.06	0.06
9	0.02	0.02	0.05	0.03	0.03	0.03	0.03	0.06	0.32	0.19
10	0.02	0.03	0.05	0.04	0.04	0.04	0.04	0.06	0.19	0.57

Figure 16: Network graph of the covariance matrix. Every position is a node shown by the blue circle. If a position is coupled to another then a line is drawn between the nodes. This representation cannot resolve the value of the coupling.



The problem with the network graph shown in Figure 16 is that it does not show the value of the covariation between two positions. In order to do show this it is necessary to either depict the lines with different colors and/or widths or to only draw lines above some covariance value threshold.

In the next figure(Figure 17) I show a network graph with the lines colored depending on the coupling score. The strongest coupling between amino acids 9 and 10 (covariance score=0.19) is shown as a dark black line whereas the other lines are fainter as they have lower covariance scores. In Figure 18 the lines with a covariation score less than 0.05 are removed leaving a much simpler network due to the lack of lines. This representation can be simplified even more by removing nodes that are not connected (Figure 19). Finally the nodes can be rearranged to more easily visualize the relationships among nodes (Figure 20).

Figure 17: In this network representation, the color of the lines depends on the covariation score with the higher the score, the darker the line color.

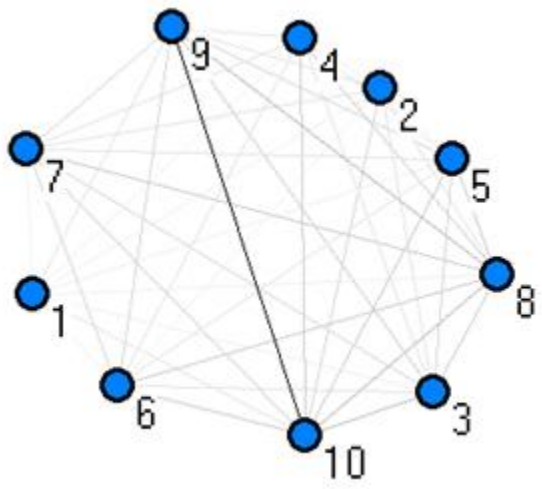


Figure 18: In this network all lines with a covariation score less than 0.05 are removed, resulting in a simpler network to analyze.

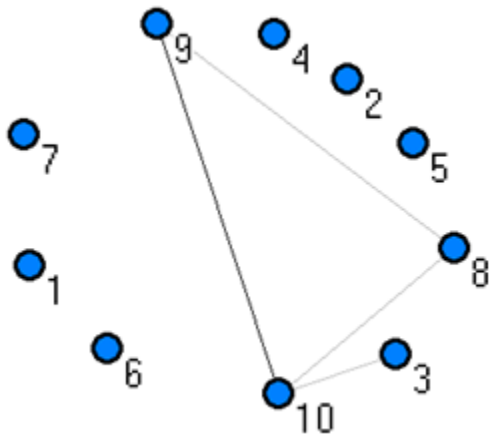


Figure 19: In this network, only nodes connected to other nodes are shown for clarity.

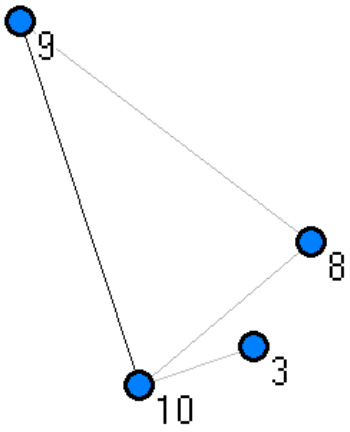
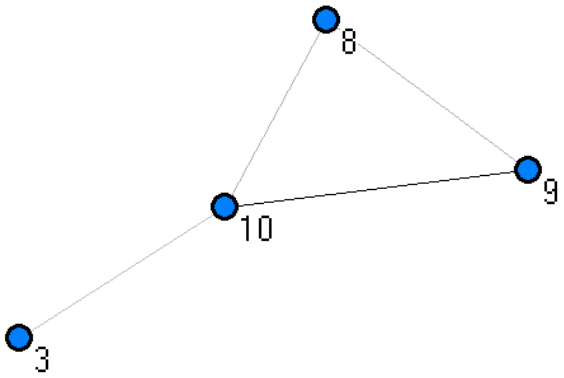
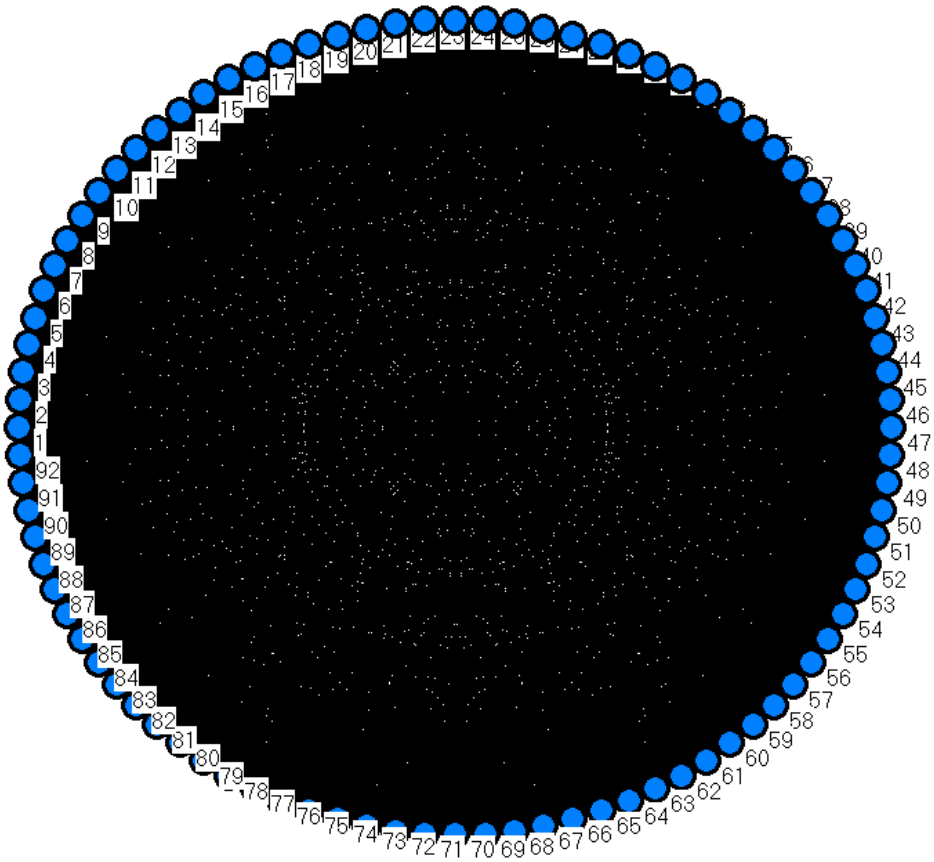


Figure 20: The nodes are arranged to show visually the relationships between the nodes better. In this case it is clear that amino acid 10 is connected to all other nodes.



The previous example showed only a small network of 10 nodes. Most protein covariance networks are larger. As an example, I will show the network graphs for the PDZ domain alignment consisting of 92 amino acids. As Figure 21 shows, it is not possible to show dark lines for 92 nodes all connected to each other. The 4186 lines that can be drawn between 92 nodes cannot be distinguished.

Figure 21: Network graph showing all lines to all nodes when there are 92 nodes.



Even if the lines are colored based on the covariance value as in Figure 22, it is still not possible to make sense out of all of this. It is therefore necessary to filter out the weak correlations. To do this, I shall use the same threshold as before which is to remove all correlations less than 0.5. This results in a much simplified network (Figure 23). If the threshold is set at 0.75 the network is further simplified (Figure 24), thus allowing connections to be visualized in an easier way.

It is now clear to see which residues are coupled to one another. This method of viewing different significance thresholds works but it would be more useful if one can see several layers of significance on one graph. This is possible if different colors are used (rather than grayscale) for the different thresholds. In the next figure (Figure 25) I show such a graph which makes it easy to see the networks and the relative strengths of the connections.

Figure 22: Network graph with 92 nodes where the lines between nodes are colored according to covariance value. Although some relationships are discernable (such as the linkage between 58, 71 and 75), the graph is difficult to analyze further.

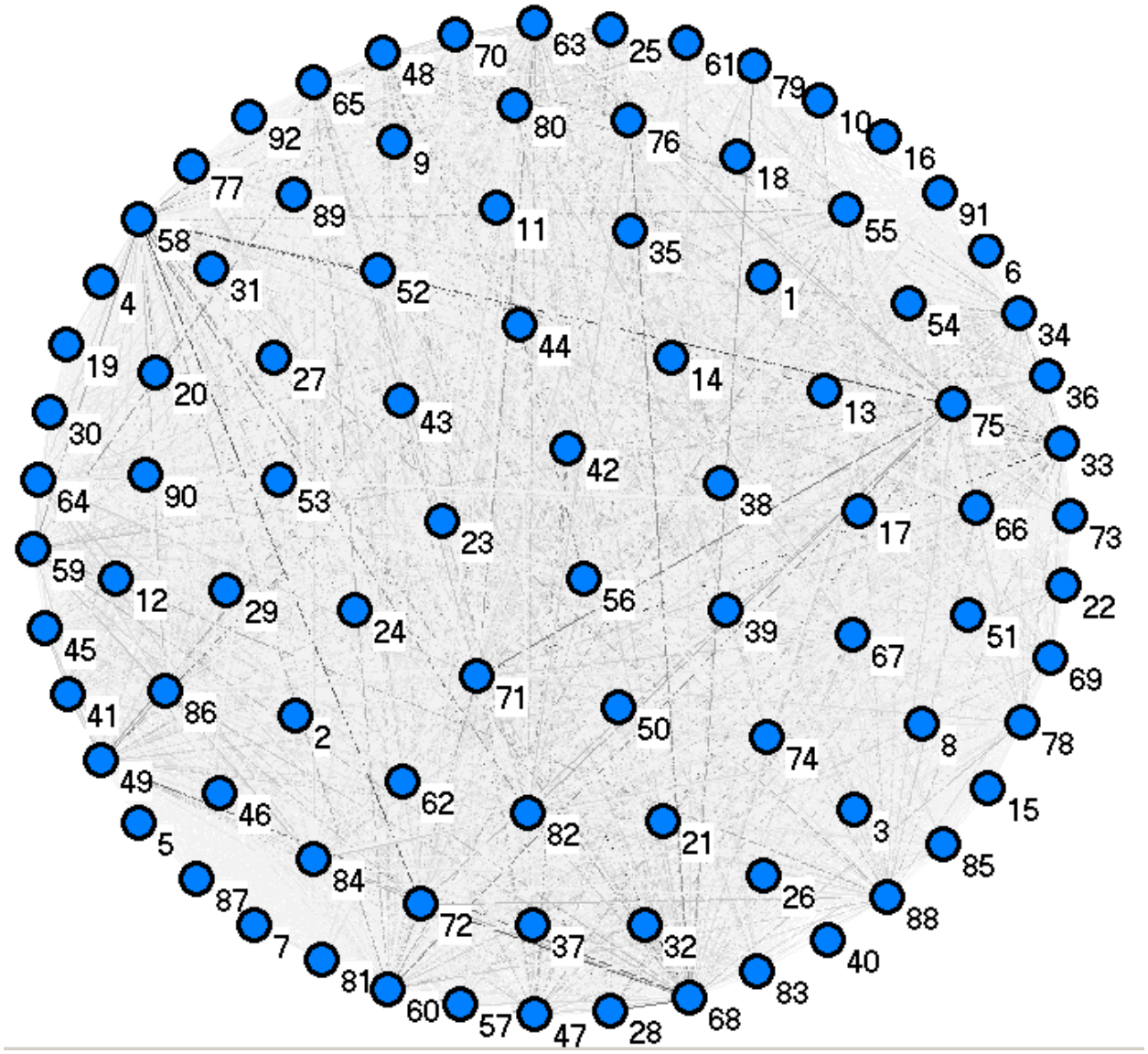


Figure 23: Network diagram with lines drawn only if there is a covariance value above 0.5. Note that nodes are arranged to show relationships clearly with highly connected nodes in the center.

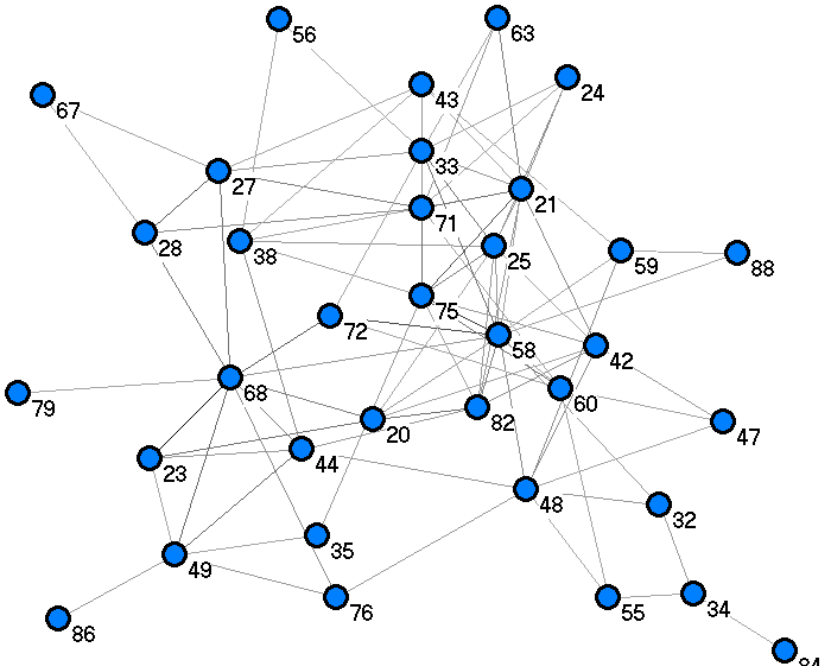


Figure 24: Network graph with lines drawn only if the covariance value between nodes is higher than 0.75. Note that nodes are arranged to show relationships clearly with highly connected nodes in the center.

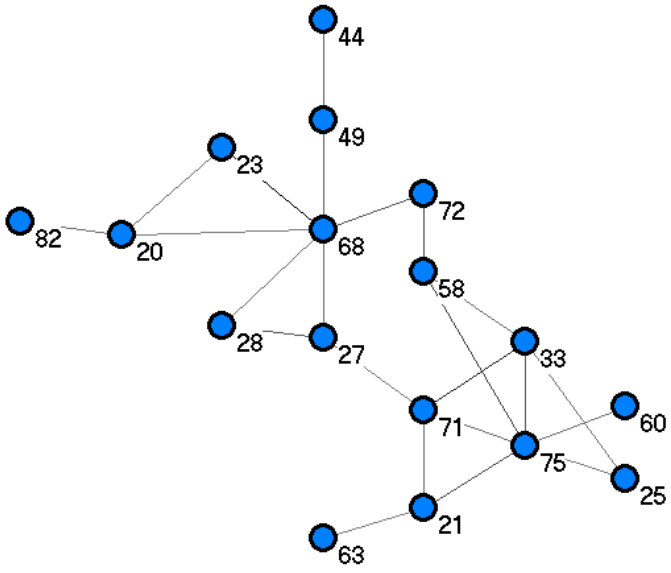
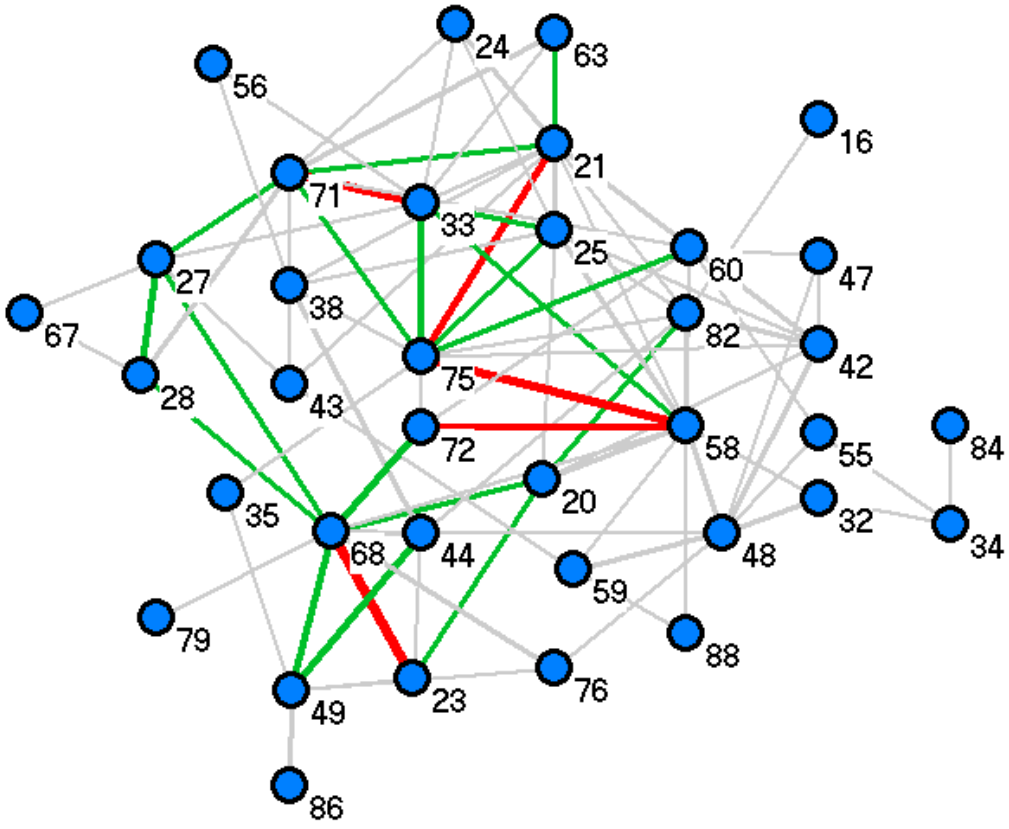


Figure 25: A network graph showing all nodes with lines greater than 0.5. The colors represent different coupling values. Red is coupling greater than 1. Green is coupling between 0.75 and 1. Grey is coupling between 0.5 and 0.75. With this type of visualization one can see clearly that residues 21-58-72-75 are linked by the highest coupling values.



This completes the network graph visualization methods section. These graphs are useful as they involve no assumptions or model for the analysis and as will be shown later, can reveal networks of coupled positions. However, one needs to choose particular, and what may seem to be arbitrary thresholds (although one may choose thresholds based on some statistical model) in order to visualize relationships as there are too many possible lines between points to show all the lines. What I will show in the detailed analysis of specific protein families in chapters 4 and 5 is thresholds from the highest down to coupling values of 0.5. In general, the interactions when considering lower thresholds are too numerous to visualize and different methods need to be used to include the contribution of weak couplings to the overall coupling pattern.

Section 2: Hierarchical Clustering

Hierarchical clustering (referred to as clustering) is a simple concept although it can be difficult to implement in practice. The idea behind clustering is to calculate for a set of data the differences between all the data points and then based on those distances calculate a tree of the data with the leaves of the tree reflecting the distances. The goal of clustering is to group similar data together. The problem with clustering is that both the distances and the tree building can be done in different ways which results in different groups being formed.

I will now give a simple example of clustering before moving on to an example for a covariance matrix. The data consists of these numbers: [1-2-3-8-10-11-12-31]. The goal is to group these data into clusters. The first step is to calculate the distance between individual elements in the data set. There are many different methods of calculating distances and it is not always clear which method to choose. The methods that appear in this text are explained in Appendix A.

Once a distance method is chosen, a distance matrix of the data is made using the metric as shown in Figure 26 using Euclidean distances. Another way to calculate distance is shown in Figure 27 where standardized Euclidean distances ((the data points are divided by the variance of the set) are shown.

Figure 26: Distance matrix of the example data set (highlighted in blue) calculated as euclidean distances. The diagonal is shown in grey.

	1	2	3	8	10	11	12	31
1	0	1	2	7	9	10	11	30
2	1	0	1	6	8	9	10	29
3	2	1	0	5	7	8	9	28
8	7	6	5	0	2	3	4	23
10	9	8	7	2	0	1	2	21
11	10	9	8	3	1	0	1	20
12	11	10	9	4	2	1	0	19
31	30	29	28	23	21	20	19	0

Figure 27: Distance matrix of the example data set (highlighted in blue) calculated as standardized euclidean distances. The diagonal is shown in grey.

	1	2	3	8	10	11	12	31
1	0	0.1	0.2	0.7	0.9	1	1.1	3.1
2	0.1	0	0.1	0.6	0.8	0.9	1	3
3	0.2	0.1	0	0.5	0.7	0.8	0.9	2.9
8	0.7	0.6	0.5	0	0.2	0.3	0.4	2.4
10	0.9	0.8	0.7	0.2	0	0.1	0.2	2.2
11	1	0.9	0.8	0.3	0.1	0	0.1	2.1
12	1.1	1	0.9	0.4	0.2	0.1	0	2
31	3.1	3	2.9	2.4	2.2	2.1	2	0

The next step is to calculate the linkage between the data using the calculated distances. This is done by first finding the data point with the smallest distance. In this example, the difference between 1 and 2 is one which is the smallest. Now a group consisting of 1 and 2 is made, called Group 1. Group 1

can either be added to or a new Group is created. To see whether additional elements should be added to Group 1, a second element, say element 3, is compared to Group 1. And here lies the second problem with clustering: how does one measure the distance from one element to a Group which consists of several elements? One method is to take the average of all possible distances between Group 1 and another element (this is called average linkage). In this example, the distance would be the average of [2-1]+[3-1] which is (1+2/2)=1.5 Another method is to consider the shortest distance (called single linkage) between Group 1 and element 3 (which would be to compare 3 with 2) or the longest distance (called complete linkage) (which would be to compare 3 with 1). Other linkage metrics are also possible and shown in appendix B. If the number 3 is of equal distance as 2 and 1 to each other then 3 is placed in Group 1. Otherwise a new group is made, called Group 2. Other elements are then compared with Group 1 and Group 2 to see where the new element would fit or whether an additional group is created.

It is easy to visualize these relationships for this example data set by way of a tree. I shall use the data using Euclidean distance to illustrate various tree formations using different linkage algorithms. The first linkage algorithm used is called 'single' and it calculates the shortest distance between elements and/or groups. This is illustrated in Figure 28. Close analysis of this tree shows that the numbers [1 2 3] are grouped together while [10 11 12] are grouped separately. The number 8 is grouped closer to [10 11 12] and 31 is in its own group. Another algorithm called 'complete', calculates linkage using the largest distance (Figure 29). There are two things to note when comparing the two

Figure 28: Dendrogram made using euclidean distances and single linkage

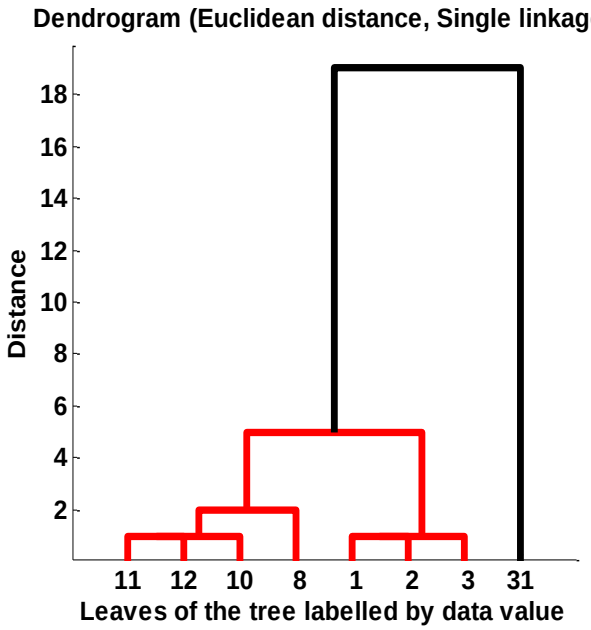
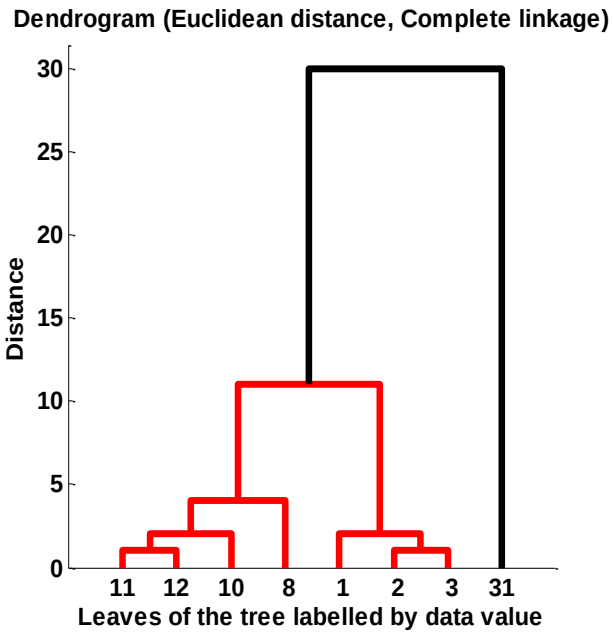


Figure 29: Dendrogram made using euclidean distances and complete linkage



figures. Firstly, while the overall structure of the tree is similar the fine structure is different. For example, [1 2 3] are not grouped in exactly the same way. Secondly, the distance on the y-axis is different. This arises because with single linkage the distances being considered are the shortest distance whereas in complete linkage the distances being considered are the largest distances. In single linkage, 31 is being compared to 12 which is the closest to 31 with a distance of 19. In complete linkage, 31 is being compared to 1 which is the furthest from 31 with a distance of 30. It is unclear which tree is more accurate or which groups cluster the elements better. This uncertainty becomes more problematic with more complex data sets.

One way to tell which tree is better is by realizing that the distances calculated with the tree algorithm should correlate well with the distances calculated from the original data. There is a measure that can quantify this which is to find the correlation coefficient of the data before and after clustering. In the example above, the correlation for the single linkage calculation is 0.958 and for the complete linkage calculation is 0.962. These are both very close to one and so both can be considered excellent solutions although the complete linkage is closer to one. Therefore when performing a clustering analysis one can create trees using different metrics and then find the one that is most consistent with the data as shown in Table 4. However, as will be shown later, this cannot be the sole criterion for

Table 4: Correlations for example data set between the tree distances and the actual data sets calculated with combinations of distances and linkages. All methods yield close values for this simple data set.
NaN means that the particular metric is not suitable for the data.

		linkage methods						
		'average'	'centroid'	'complete'	'median'	'single'	'ward'	'weighted'
distance methods	'euclidean'	0.962	0.962	0.962	0.961	0.958	0.943	0.961
	'seuclidean'	0.962	NaN	0.962	NaN	0.958	NaN	0.961
	'mahalanobis'	0.962	NaN	0.962	NaN	0.958	NaN	0.961
	'cityblock'	0.962	NaN	0.962	NaN	0.958	NaN	0.961
	'cosine'	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	'correlation'	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	'spearman'	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	'chebychev'	0.962	NaN	0.962	NaN	0.958	NaN	0.961

deciding what is the best tree.

The basic ideas and methodologies of hierarchical clustering are apparent with this simple example data set. Applying this to the covariances of sequence data is more complicated. With covariance data

there are many more pairs to consider as every position in a protein covaries with every other position. For example, the PDZ domain which has 92 positions in the alignment, has 4186 pairs that need to be considered (91*92/2). Because of this large number it is not a simple manner to visually inspect the resulting tree and see if it makes "sense".

In the next table (Table 5), I show a hierarchical analysis of the PDZ domain family done with different methods. It is clear in this example that different methods have a big impact on the quality of the tree. The tree shapes and the colored matrices for three different algorithms are shown in three figures (Figure 30, Figure 31, Figure 32). These show different trees with tree/data distance correlations of (0.442, 0.989 and 0.935). While all the trees contain groupings, it is difficult to conclude which is the best. Thus other methods to aid in the identification of covarying groups will be discussed next.

Table 5: Correlation of tree distances with actual distances for the PDZ domain.

		linkage methods						
		'average'	'centroid'	'complete'	'median'	'single'	'ward'	'weighted'
distance methods	'euclidean'	0.968	0.965	0.935	0.927	0.954	0.694	0.946
	'seuclidean'	0.892	NaN	0.681	NaN	0.810	NaN	0.844
	'mahalanobis'	0.433	NaN	0.301	NaN	0.154	NaN	0.382
	'cityblock'	0.914	NaN	0.442	NaN	0.908	NaN	0.888
	'cosine'	0.885	NaN	0.757	NaN	0.854	NaN	0.789
	'correlation'	0.723	NaN	0.427	NaN	0.670	NaN	0.629
	'spearman'	0.697	NaN	0.479	NaN	0.700	NaN	0.591
	'chebychev'	0.989	NaN	0.978	NaN	0.966	NaN	0.987

Figure 30: Tree of PDZ domain covariances made using city block metric and complete linkage. The numbers on the x-axis of the tree are distances. The coupling matrices are shown with values colored according to the color scale in figure 14.

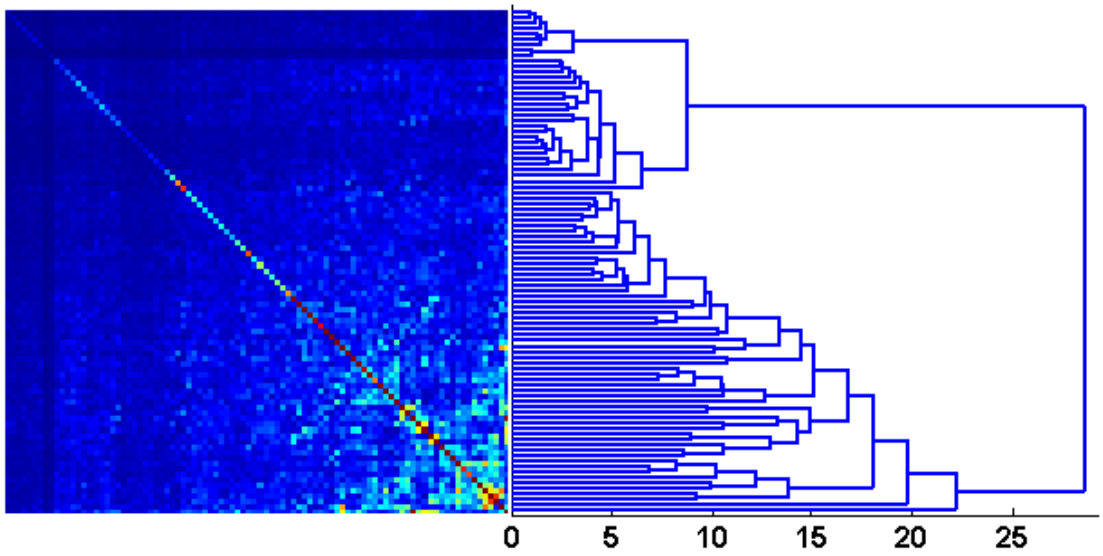


Figure 31: Tree of PDZ domain covariances made using euclidean metric and complete linkage. The numbers on the x-axis of the tree are distances. The coupling matrices are shown with values colored according to the color scale in figure 14.

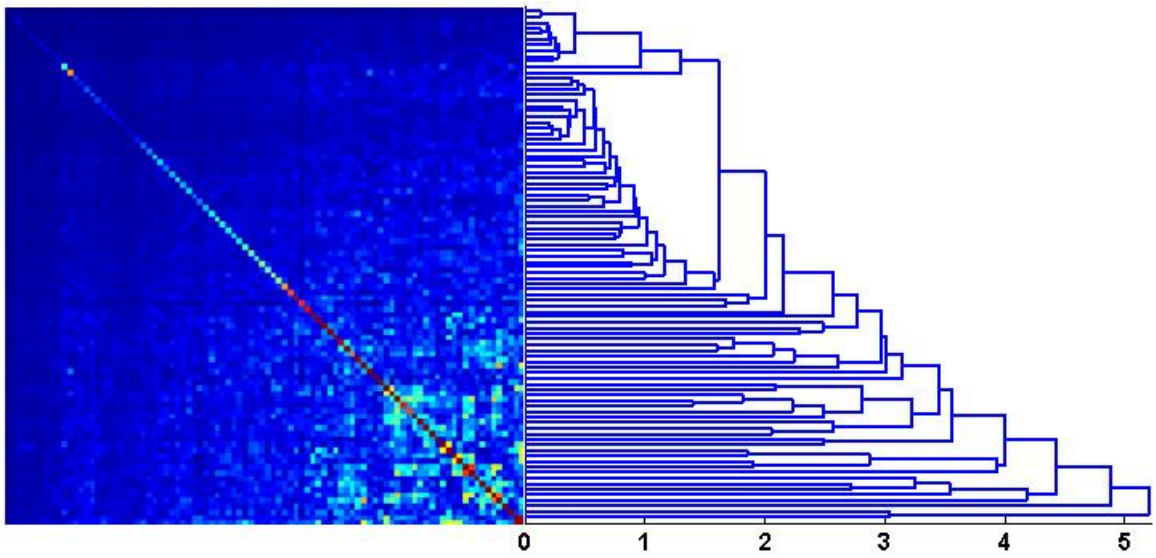
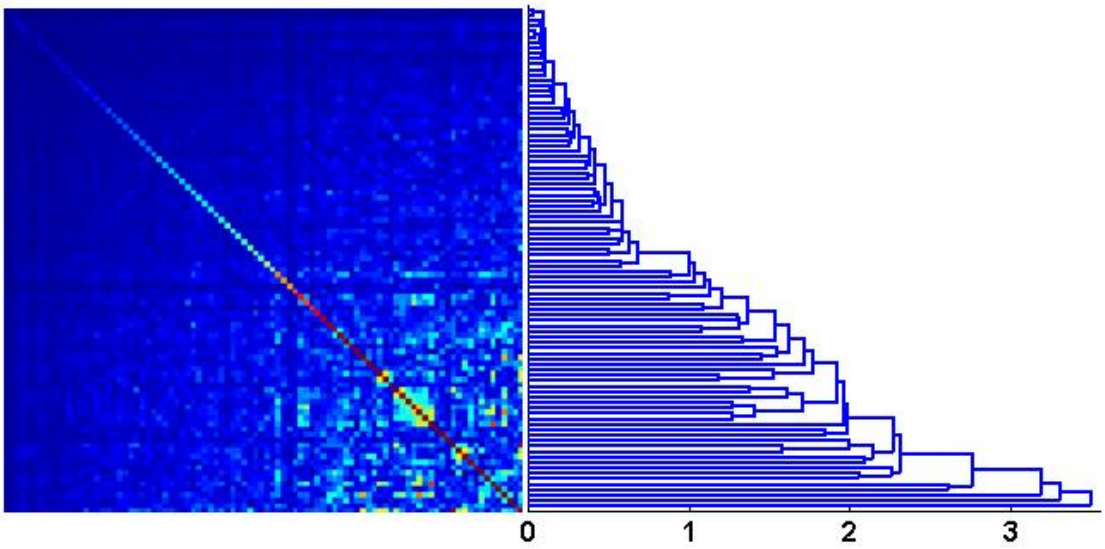


Figure 32: Tree of PDZ domain covariances made using chebychev metric and average linkage. The numbers on the x-axis of the tree are distances. The coupling matrices are shown with values colored according to the color scale in figure 14.



Hierarchical analysis is dependent on the distance metric and the linkage method. The resulting tree needs to be analyzed further to see if the resulting clusters are consistent with the original distance data and with possibly experimental data or other features of the data set that are independent of the actual distance data.

Section 3: Principal Component Analysis (PCA)

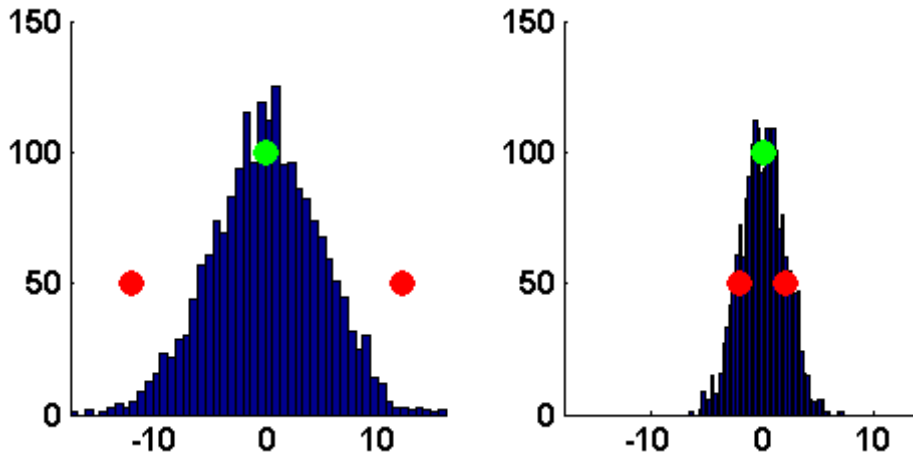
A natural starting point when discussing principal components analysis (and the similarly named independent component analysis (ICA) is to think about how one can visualize and express data and what parameters can describe a data set.

Consider for example a single variable data set. The data can be characterized by the mean and the variance. The variance describes the variability of the data. The formula for the variance of a vector of dimension *n* containing *x* elements and with mean $\bar{x}$ is:

$$\text{variance}=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n}$$

The variance describes the spread of the data about the mean because each member of the data set is subtracted from the mean. The sign does not matter because the difference is squared. The standard deviation is the square root of the variance and is used to show the variance because it has the same units as the data. Different data sets are characterized by different means and variances. A data set with the same mean, if it has a substantially different variance, can be said to come from another population. The mean and standard deviation for two different data sets is shown in Figure 33.

Figure 33: Histograms for two sets of data showing the mean of each data set (green dot) and one standard deviation away from the mean in either direction (red dots).



Is the variance a meaningful measure when comparing data sets? In general it is because it is unlikely that two variables which have the same variance are coming from different populations. Also, it is even more unlikely that data with different variances are coming from the same population. Though the variance is widely used to describe data sets, the suitability for a given data set would have to be determined after investigation of the data set. For example, a data set with several large outliers could generate a high variance even though most of the data could be characterized by a small variance.

Another parameter that can be used to describe single variables is the shape of the distribution. For example variables can be normally distributed, skewed to one side or have a sharp peak, be sinusoidally shaped, be uniform or have some other shape. The shapes of the distribution are important when comparing data sets as comparisons are often valid only when comparing data sets of some known shape. Also data sets with different distributions often mean that they are measuring different and possibly independent phenomenon. Independent component analysis makes explicit use of the distribution shape and will be discussed further. PCA does not use the distribution information but acts only on the variance as will be explained.

For single dimension data, the variance is a useful parameter. For a data set with two variables x and y a related quantity called covariance can measure the relationship between the two variables (as described in Chapter 2 and repeated here). The formula for covariance between vectors ***x*** and ***y*** is:

$$\text{covariance}=\sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{n}$$

To see how the covariance measures relationships between two vectors consider two sets of data, x and y shown as a scatter plot in Figure 34 and Figure 35.

Figure 34: Scatter plot showing a linear trend.

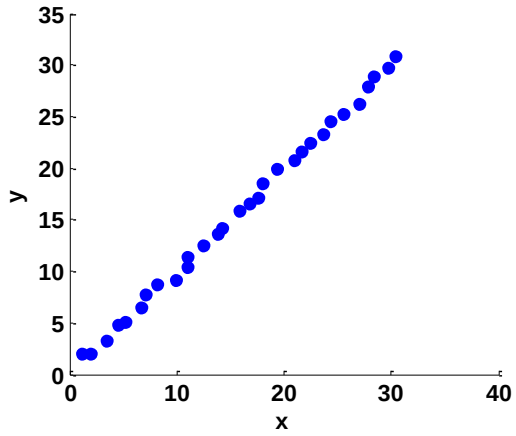
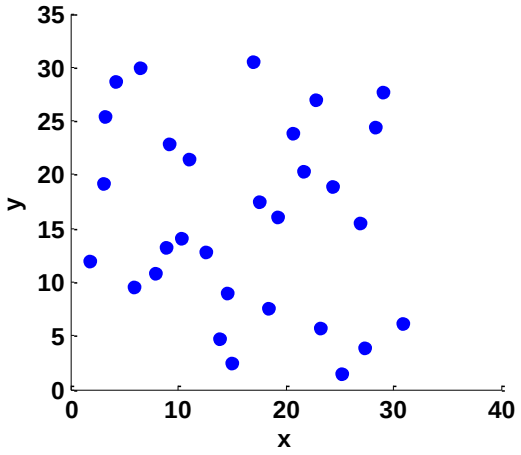


Figure 35: Scatter plot showing no trends.



Clearly, the data in Figure 34 show a linear trend and the data in Figure 35 shows no discernable trend. If two variables are linearly related then they will have a high covariance. If they are not linearly related, they will have a low covariance. In the above example, the covariance for Figure 34 is 77.7 whereas the covariance for Figure 35 is -12.0. Thus if one only had the covariance information one could come to the conclusion that the variables in the linear data set are much more related than the variables in the non-linear data set.

If there are more than two vectors in a data set, a covariance can be calculated between each vector. A covariance can be calculated between a vector and itself as well but then it is the variance as the covariance definition shows. The set of all covariances and variances in a data set is called the covariance matrix. The diagonals of a covariance matrix are the variances while the off-diagonals are the covariances. If the mean of each data set is zero (easily done for continuous data) then, using matrix algebra, the covariance can be calculated as the product of the data matrix (the data expressed in matrix form with the different variables in rows and different observations in the matrix columns) times the data matrix transposed:

$$\text{mean centered covariance matrix} = \mathbf{DataMatrix} \times \mathbf{DataMatrix}^T$$

The problem with analyzing multi-dimensional data sets is that one cannot easily see which variables are related to each other. Furthermore, it is possible that several sets of residues each have similar trends and it is not possible to discern these in large data sets. This is where principal components analysis is useful.

Principal components analysis developed In 1901 by Pearson [115] transforms the data in such a way as to more clearly show the relationships among the variables. PCA has its roots in linear algebra and so the linear algebra operations will be used to describe it.

A data set can be represented as a matrix **X**. It is possible to represent the data in **X** in another *basis*. A basis is a set of vectors by which one can express every other vector in the vector space. For example, consider the vector: [1,0] AND [0,1]. The set of these two vectors can be used to form any vector in a two-dimensional vector space. For example: the vector [5,9] can be expressed as 5×[1,0] + 9×[0,1]. Similarly any vector [a,b] can be expressed as: a×[1,0] + b×[0,1]. The basis that was listed above |[1,0],[0,1]| is defined as the standard basis for a two-dimensional space. Similarly, the standard basis for three dimensional space is |(1,0,0),(0,1,0),(0,0,1)|. However this is not the only basis there is: there are an infinite numbers of basis by which one can express a vector. The problem then is to find a basis that can express the data in a better way than the standard basis (or whatever basis the data happen to be in).

Given the coordinates of a vector in an old basis, it is possible to find the coordinates of a vector in a new basis by the following matrix multiplication:

Y (new coordinates) = **P** (new basis)× **X**(old coordinates).

For PCA purposes, the question is what is **P** that when multiplied by **X** leads to *a more meaningful* representation of **X**.

To answer this question it is necessary to know what a *more meaningful* representation of **X** means . In this case a more meaningful representation means that the data in the data set are grouped together so that the variance is maximized. For example, consider a data set in which there is some noise and in which there is a linear trend. Generally the linear trend data would exhibit large variances while the noise will have a small variance. In fact, in order to make any sense of data at all, the variance of the interesting relationships in the data must be greater than the noise. And the noise vs. data can be characterized as one of small variance vs. large variance. Therefore a meaningful representation of the data is one that maximizes the variance.

How is the variance maximized? The variance is maximized if the covariance between the vectors becomes zero. The covariance can be made zero by finding the correct basis to express the data as illustrated in Figure 36. The steps to find the basis are described next.

Figure 36: This figure illustrates how an axis can be rotated (dashed lines) so that the points lie only on one axis whereas it took two axes to show the data previously. The covariance that existed between the variables becomes zero after rotation, while the variance is maximized. This can be considered a reduction in dimension from two to one.

To restate, the goal is to find a basis such that a covariance matrix is obtained in which all the covariances are zero. The covariance matrix with zero covariance is called the transformed covariance matrix. The transformed covariance matrix can also be written in terms of the data which in this case is not the original data but the data transformed in an as yet unknown way. The expression for the transformed covariance matrix in terms of the transformed data can be written as:

Transformed covariance matrix $= (\boldsymbol{transformed\ matrix})(\boldsymbol{transformed\ matrix})^T$

Next, substitute Y for the transformed matrix:

$$= (\boldsymbol{Y})(\boldsymbol{Y})^T$$

The transformed data matrix, Y, is formed by multiplication of the original data matrix, **X**, by the as yet unknown basis **P**.

$$= (\boldsymbol{PX})(\boldsymbol{PX})^T$$

The transpose of two matrices can be written as the transpose of the innermost matrix times the transpose of the outermost matrix to give:

$$= (\boldsymbol{PX})(\boldsymbol{X}^T\boldsymbol{P}^T)$$

The terms can then be grouped together to give:

$$= \boldsymbol{P}(\boldsymbol{XX}^T)\boldsymbol{P}^T$$

A matrix times its transpose forms a symmetric matrix. Call the symmetric matrix **A**.

$$= \boldsymbol{PAP}^T$$

Now two theorems from linear algebra (described in any linear algebra textbook-Gilbert Strang's book [116] is popular) are used. One theorem, called the Real Spectral Theorem, states that if **A** is a symmetric matrix, then **A** is diagonalizable. A matrix that is diagonalizable is one that can be written as follows: $\mathbf{A=PDP^T}$, where **D** is a diagonal matrix. The second theorem is called the Fundamental Theorem of Symmetric Matrices and it says that if **A** is symmetric then it can be written as $\mathbf{A=EDE^T}$. This differs from the previous expression in that the matrix written as **P** is now the matrix of eigenvectors, **E**. Given those theorems this means that **A** can be written as a diagonal matrix multiplied on each side with the eigenvectors of **A**.

Now the goal is to find what **P** is. Guess that **P** is equal to $\mathbf{E^T}$ and see where that leads by substituting into the expression for **A**. This leads to $\mathbf{A=P^TDP}$. Then substitute **A** into the expression above to give:

$$= \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{DP})\boldsymbol{P}^T$$

Rearrangement yields:

$$= \boldsymbol{PP}^T\boldsymbol{DPP}^T$$

Another theorem can be used here. The eigenvectors of a symmetric matrix have a property that their transpose is equal to the inverse matrix. This is a highly unusual property but applies here. The solution to the problem of finding the right basis is at hand because using the property gives the following:

$$= \boldsymbol{PP}^{-1}\boldsymbol{DPP}^{-1}$$

A matrix times its inverse is the definition of an identity matrix. So:

$$= \boldsymbol{IDI}$$

Again, by definition, an identity matrix times a matrix, **X**, is **X**:

$$= \boldsymbol{D}$$

The problem of finding the basis that when multiplied by the original data results in a diagonalized matrix is solved. The basis is: the eigenvectors transposed of the original data matrix. This means that if one has a data set, and one wants to diagonalize it, i.e. express the data in a different coordinate system so that the covariances between all the vectors are zero, all one has to do is find the eigenvectors, and then multiply the original data matrix by the transpose of the eigenvectors. With modern day computers this process is easily done although when Pearson wrote about it in 1901 he said that it is not hard to envision doing this for 4 or 5 dimensions.

One other property is at the core of PCA and that has to do with what is on the diagonals of the diagonalized matrices. That property is that the diagonals of the diagonal matrix are the eigenvalues associated with the eigenvectors. However, these eigenvalues are actually equivalent to *something else*: because the diagonalized matrix is formed as the transformed data matrix times the transposed data matrix, *the diagonalized matrix is also a covariance matrix*. That means that the diagonals which are the eigenvalues of the transformed matrix are also variances of the transformed data matrix.

Think about this for a minute with respect to what variances represent. As I stated earlier in the chapter, the variances of a data set is an important property of the data set as the larger the variance the greater the spread of the data. If the data set is composed of low noise relative to the signal, then the largest variances consist of the most interesting aspects of the data. Therefore to find the most interesting behavior of a low noise data set, one powerful approach is to calculate the eigenvectors and eigenvalues and then sort the eigenvalues which are the variances of the transformed data set from high to low. The eigenvector corresponding to the highest eigenvalue is called the principal component.

The original data, if transformed with the transpose of the principal component, will place on one axis the contribution of all points towards that variance. Typically however, one plots several of the top components against each other to see which ones yield interesting patterns or groupings of the data. A simple example of PCA is shown in the next page as Figure 37.

Figure 37: An example of principal components analysis for a simple data set. The steps are shown in sequence.

1. Collect data

-9.5	-8.5	-7.5	-6.5	-5.5	-4.5	-3.5	-2.5	-1.5	-0.5	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
-7.9	-7.0	-6.4	-5.8	-7.0	-5.3	-5.0	-4.7	-2.4	0.1	0.8	1.5	2.0	2.4	4.3	6.3	7.3	7.6	9.3	9.9

2. Calculate a covariance matrix

XX^T

cov ariance matrix:

35.0000 34.8843
34.8843 35.8889

3. Calculate eigenvectors and eigenvalues of the covariance matrix

eigenvectors :

-0.7116 0.7026
0.7026 0.7116

eigenvalues :

0.5573 0
0 70.3316

4. Sort the eigenvalues

sorted eigenvalues:

70.3
0.5

5. Choose a subset of the eigenvalues based on the variances they capture

percent of total variance:

eigenvalue 1: 99.2
eigenvalue 2: 0.8

6. Select the corresponding eigenvectors

sorted vector:

0.7026 -0.7116
0.7116 0.7026

7. Take the transpose of the selected eigenvectors

transpose the sorted vector:

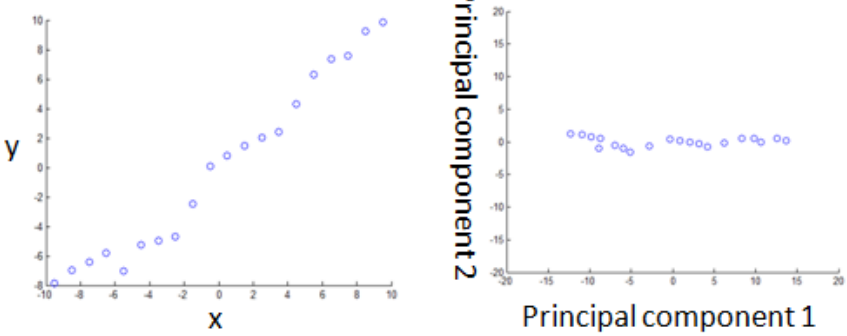
0.7026 0.7116
-0.7116 0.7026

8. Multiply the data by the transpose of the selected eigenvectors

Y

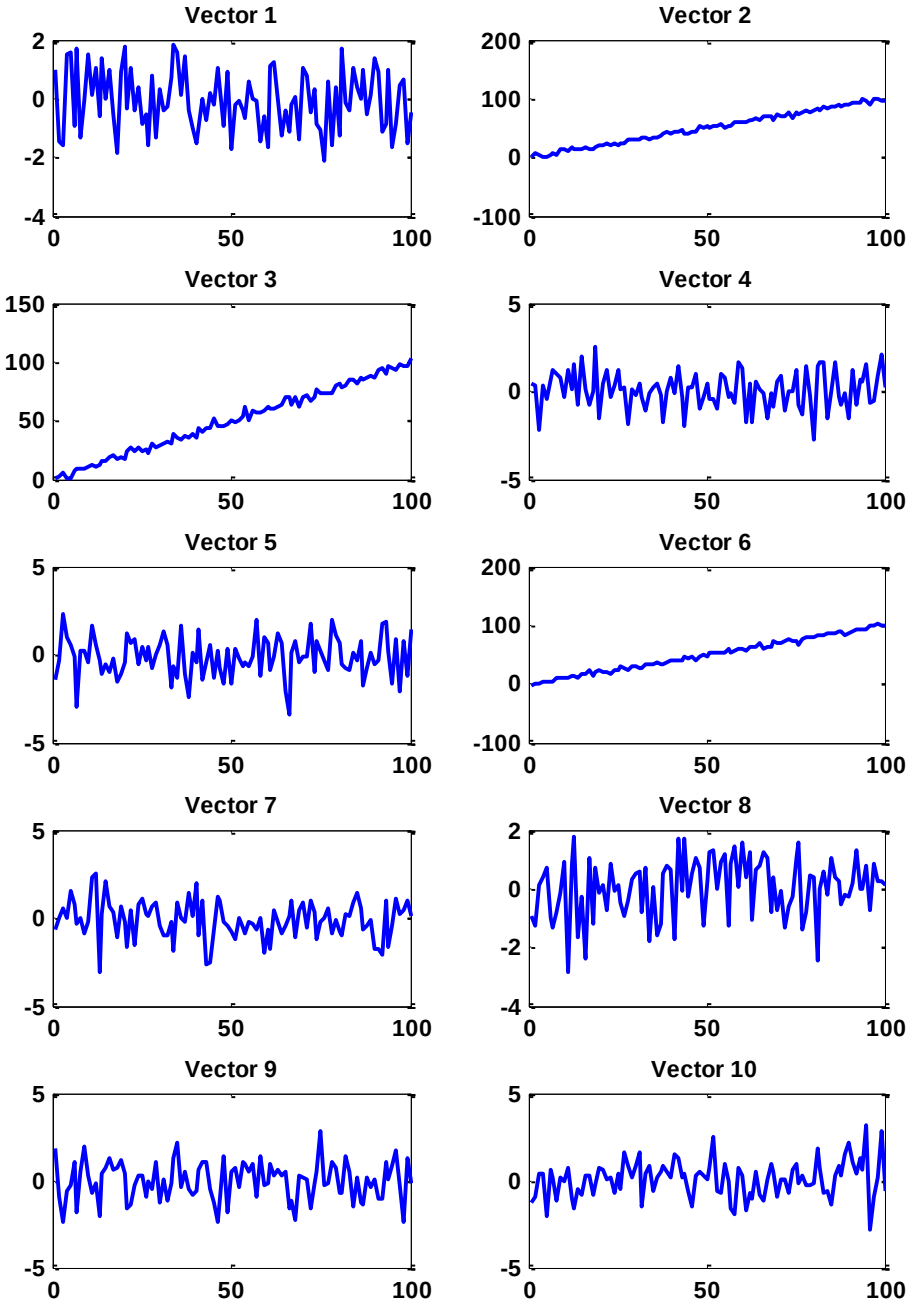
-12.3	-10.9	-9.9	-8.7	-8.9	-6.9	-6.0	-5.1	-2.8	-0.3	0.9	2.1	3.2	4.2	6.2	8.4	9.8	10.7	12.6	13.7
1.2	1.2	0.8	0.5	-1.0	-0.5	-1.0	-1.5	-0.7	0.4	0.2	0.0	-0.3	-0.8	-0.2	0.5	0.5	0.0	0.5	0.2

9. Plot the original and transformed data



That example shows how PCA can rotate the axis so that the important variation in the data lies on one axis. However, another use of PCA is not to rotate the axis but to group vectors based on their contribution to the variance. This form of PCA is known as spectral clustering. In the next example, I will show how PCA and spectral clustering works with data consisting of ten vectors. The vectors are shown in Figure 38.

Figure 38: Plots of ten data vectors



Note that this is example is chosen so that one can see by eye that vectors 2, 3 and 6 are all linearly related while the other vectors are noise. Next, compute the covariance matrix and the eigenvalues and eigenvectors.

The covariance matrix is shown in Figure 39. Note that the covariance between vectors 2, 3 and 6 is high relative to the covariance of other vectors.

Figure 39: Covariance matrix of 10 vectors.

		Vectors									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Vectors	**1**	1.03	-3.67	-3.69	-0.06	-0.19	-3.67	0.06	-0.22	-0.12	-0.07
	2	-3.67	852.19	830.85	-0.73	0.89	850.35	-3.76	5.11	-0.24	5.37
	3	-3.69	830.85	821.90	-0.61	1.08	836.17	-4.09	4.74	-0.10	4.63
	4	-0.06	-0.73	-0.61	1.08	-0.04	-0.54	0.09	-0.02	0.16	0.13
	5	-0.19	0.89	1.08	-0.04	1.25	1.57	-0.08	-0.11	-0.22	0.04
	6	-3.67	850.35	836.17	-0.54	1.57	864.83	-3.15	4.95	0.14	4.84
	7	0.06	-3.76	-4.09	0.09	-0.08	-3.15	1.13	-0.07	0.03	-0.15
	8	-0.22	5.11	4.74	-0.02	-0.11	4.95	-0.07	0.97	-0.01	-0.06
	9	-0.12	-0.24	-0.10	0.16	-0.22	0.14	0.03	-0.01	1.20	0.02
	10	-0.07	5.37	4.63	0.13	0.04	4.84	-0.15	-0.06	0.02	1.08

The eignevalues of the covariance matrix are shown next in Figure 40. Remember that the diagonals of the eigenvalues correspond to variances. Also note that this matrix has zero for the value of the covariances between vectors indicating that the variances have been maximized. A more useful way to look at the eigevalues is to show a histogram as in Figure 41. Clearly one eigenvalue, corresponding to the tenth vector stands out.

The eigenvectors along with their values are shown in Figure 42.

Figure 40: Eigenvalue matrix. The eigenvalues correspond to variances.

		Eigenvalues									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Eigenvalues	**1**	0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.00	0.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	1.07	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00	1.19	0.00	0.00	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.00	0.00	1.35	0.00	0.00	0.00	0.00
	7	0.00	0.00	0.00	0.00	0.00	0.00	1.52	0.00	0.00	0.00
	8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.92	0.00	0.00
	9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.36	0.00
	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2524.92

Figure 41: Eigenvalue histogram

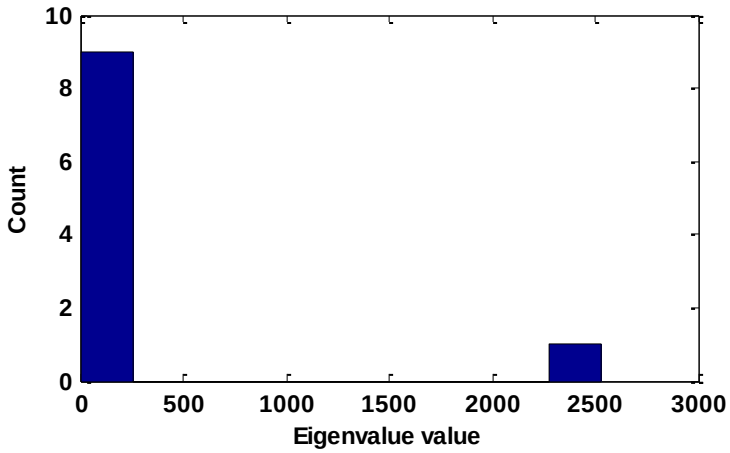
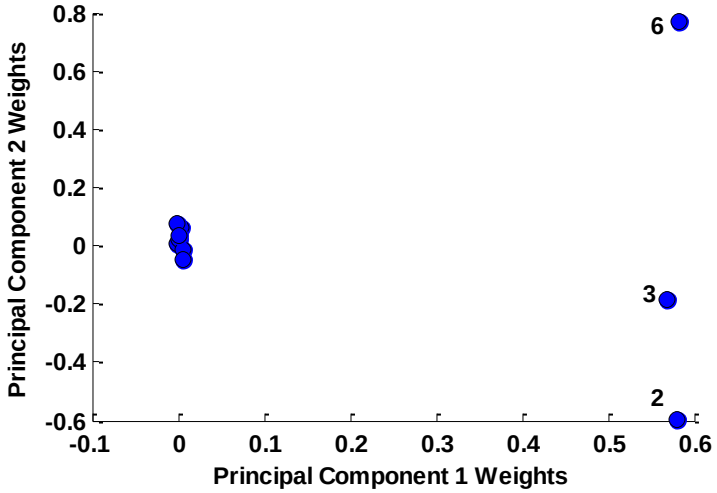


Figure 42: Eigenvector matrix. The vector is defined as a column. Note that this is not a symmetric matrix.

		Eigenvectors									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Eigenvector weights	**1**	0.54	0.34	-0.06	0.21	0.48	-0.56	-0.05	-0.01	0.00	0.00
	2	-0.02	0.06	0.05	-0.03	-0.01	0.00	0.01	-0.55	-0.60	0.58
	3	0.06	-0.06	-0.01	-0.05	0.02	-0.01	-0.02	0.79	-0.18	0.57
	4	-0.09	0.49	-0.45	-0.42	0.33	0.40	-0.31	0.00	0.02	0.00
	5	0.38	0.23	0.34	-0.29	0.06	0.36	0.68	0.00	0.06	0.00
	6	-0.04	0.01	-0.04	0.08	-0.01	0.00	0.00	-0.23	0.77	0.58
	7	0.23	-0.46	0.09	-0.77	0.05	-0.29	-0.21	-0.09	0.08	0.00
	8	0.58	0.03	-0.46	0.05	-0.65	0.10	-0.08	-0.03	-0.01	0.00
	9	0.28	0.11	0.62	0.14	-0.05	0.34	-0.62	0.00	0.04	0.00
	10	0.27	-0.60	-0.25	0.28	0.48	0.44	0.02	-0.07	-0.04	0.00

The eigenvalue matrix shows that eigenvector 10 has the highest variance. Eigenvector 10 is also known as the first principal component. Observation of eigenvector 10 shows that it has three values of about 0.6 while all the other values are zero. The weights of the eigenvectors reflect the contribution of each original vector to the variance. Hence, in eigenvector 10, the original vectors 2, 3 and 6 contributed most to that eigenvector. If a scatterplot is made of eigenvector 10 vs. eigenvector 9 (the eigenvector with the second highest variance) as shown in Figure 43, it is clear that vectors 2, 3 and 6, are very far away from the remaining vectors. Thus one could consider vectors 2, 3 and 6 to form a group. In the next chapters I will perform the same analysis with sequence data where there are many more vectors, but the procedure is exactly as outlined here.

Figure 43: Scatter plot of the highest eigenvector with the second highest eigenvector. The numbers refer to the weights that contributed to the vectors.



Note that in the analysis of sequence data what is done with PCA is to look at the degree to which each position contributes to any given significant component. Some positions will contribute more than others. Positions that contribute similarly to a particular vector can be considered to be clustered as will be shown in the next chapter. Also note is that the coupling matrix is considered a covariance matrix and so the steps for calculating eigenvectors and eigenvalues are done directly on the coupling matrix.

One additional note about PCA is that it is an exact solution to the variance maximizing problem. A disadvantage is that it assumes the relationship between variables is linear. Another shortcoming is that PCA merely decorrelates variables. In some cases, what is interesting, are independent variables. While independence implies decorrelation, decorrelation does not necessarily mean independence. In the next section, I discuss a method called independent component analysis that finds vectors in the data that are independent.

Section 4: Independent component analysis (ICA)

As stated in the introduction to this chapter, ICA has different assumptions from PCA and is trying to solve a different problem. ICA tries to separate components that are independent whereas PCA only separates components that are uncorrelated. For this section I draw on the works of James Stone [117, 118] and of Aapo Hyvarinen and Erkki Oja [119].

A starting point for thinking about separating independent components is to think about the ways that samples can be mixed together and the ways those signals can be detected. For example, if one was in a room with many people talking, all the voices, unless everybody was singing in concert, would be mixed together. If one moved around the room one would hear different combinations of voices. Another example is if one is recording electrical signals from the brain. The signals are typically complex wave forms coming from a multitude of independent neural processes and one can record from different areas of the brain to obtain different combinations of those waveforms. So two common aspects of voices and electrical signals is that they are mixtures of independent events and that different recordings yield different combinations of the independent events.

The mixing and recording of independent signals can be expressed mathematically. If there are only two independent signals (denoted **signal 1** and **signal 2**) then one can write the mixing process by multiplying **signal 1** by a certain factor and **signal 2** by another factor to give rise to the mixed signal:

signal 1 × factor 1 + signal 2 × factor 2 = mixed signal 1

A second mixed signal can be obtained by changing the recorded position. The new mixed signal can be written as:

signal 1 × factor 3 + signal 2 × factor 4 = mixed signal 2.

It is possible to collect more mixed signals which is advisable if one does not know how many independent signals there are as for ICA to work the number of signals to be extracted cannot exceed the number of mixtures recorded. That is for every different voice that one would like to record, there should be a different microphone recording. In general, this means that one should collect as much data as possible if one does not know beforehand the nature of the recorded signals.

Inspection of the above two equations shows that there are two known variables consisting of the two mixed signals, but four unknown variables. There is no information about the factors (to be called mixing factors) or the signals (to be called source signals). To solve for the source signals, knowledge of

the factors is needed. However, though there is no specific knowledge of these factors or of the signals, there is some general knowledge about their form.

This general knowledge is that the form of the mixed signals is different from that of the source signals. Specifically the source signals are less Gaussian than the mixed signals. The central limit theorem from statistics ensures this, even if the signals are non-Gaussian.

For ICA this implies that the source signals are less Gaussian than the mixed signals. The feature that ICA can extract non-Gaussian data from Gaussian mixtures -while central for the success of ICA- introduces a limitation. The limitation is that if both source signals are themselves Gaussian, then they cannot be resolved from the Gaussian mixture. However, if the source signals are slightly non-Gaussian then they can be resolved from the mixed signal.

The next step in ICA is to try to search computationally for signals and factors that when combined together yield the observed mixed signal. In effect, this is a computational trial and error approach. One method to implement this is called maximum likelihood estimation (MLE).

Maximum likelihood estimation is a widely used method to estimate the probability that parameters of a set of data describe the data given some model of the data. In MLE, the data are known but the parameters used to describe the data are unknown. In a trivial case of the probability of coin flipping, the MLE would be used to estimate the probability that the mean value of a coin flip given the actual observed numbers of heads and tails. The model of coin flipping could be the binomial probability. In the case of MLE of ICA, the known data are the mixed signals whereas the unknown parameters are an unmixing matrix that results in independent and non-Gaussian parameters.

With MLE, a guess is made of **signal 1** and **signal 2** and then another guess of the factors and then the mixture can be calculated. Then the guessed mixed signals can be compared to the observed mixed signals. To reduce the search space, the signal guesses are based on non-Gaussian functions. The process iterates until an acceptable level of agreement between the observed data and guessed signals is reached.

To repeat the requirements of ICA: the number of mixtures should be greater than or equal to the number of signals, the signals must be non-Gaussian and the signals should be independent. In vector form a signal can be written as a row vector like this:

$$s_1 = (s_1, s_2, s_3, \cdots, s_n)$$

Two source signals can be written as:

$$s = \begin{bmatrix} s_1^1 & s_1^2 & \cdots & s_1^n \\ s_2^1 & s_2^2 & \cdots & s_2^n \end{bmatrix}$$

The mixing process can be written as:

$$\mathbf{x} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} s_1^1 & s_1^2 & \cdots & s_1^n \\ s_2^1 & s_2^2 & \cdots & s_2^n \end{bmatrix} = \mathbf{As}$$

On the other hand what is observed is **x**. What is needed is **s**. In that case one could write:

$\mathbf{s} = \begin{bmatrix} \alpha & \beta \\ \lambda & \delta \end{bmatrix} \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ x_2^1 & x_2^2 & \cdots & x_2^n \end{bmatrix} = \mathbf{Wx}$, where **W** is the called the unmixing matrix

In vector form, the search is to find the unmixing matrix coefficients in order to compute **s**.

As mentioned above, one can guess the unmixing coefficients, multiply that by the observed data, and compute a preliminary source signal, which I will call the guess signal. How is one to know how close the source signal is to the guess signal?

This is where the non-Gaussian nature and independence of the signals comes it to play. One can calculate the non-Gaussian properties and independence of the guess signals when different unmixing matrices are tried. When the maximum non-Gaussian transform and independence is reached, or is close to being reached, one can conclude that the guess signal is close to the source signals.

To explain further before the mathematical details are given, the unmixing matrix is the inverse of the mixing matrix; the knowledge of one matrix can be used to calculate the other. That implies that one can express MLE in terms of the mixing matrix, given a set of source signals, which is the implementation given here.

The probability density function (pdf) is what describes the non-Gaussianity or Gaussianity of a set of signals. The pdf is a histogram with infinitesimally small bins. Given a set of signals (s) one can calculate the probability that any given value of the signals occurs. This probability, that a signal has a particular value, t, is denoted as ps(st). Different types of non-Gaussian signals have different probability density functions. One would have to guess what type of probability function the signals are described by. For example, if one is dealing with speech signals, they are characterized by a lot of silence and so have a flat signal with a sharp spike.

Independence has a very specific statistical meaning. If signals are independent then the pdf function of a pair of signals (the joint pdf) is the product of the individual pdfs (the marginal pdfs). Mathematically, $p_s = p_{s_1} \times p_{s_2}$, if independent.

To repeat, maximum likelihood estimation will attempt to find an unmixing matrix where the guess joint pdf is as similar as possible to the model joint pdf. For the speech signal example, one would specify a model for the joint pdf of two independent speech signals where each speech signal would be non-Gaussian and then one would compare this model joint pdf to the pdf obtained with the guessed unmixing matrix.

The relationship between the pdfs of the source signals and mixtures assuming the correct matrix is:

$$p_x(\mathbf{x}) = p_s(\mathbf{s})\left|\mathbf{W}^{correct}\right|, \mathbf{W}^{correct} = \frac{\partial \mathbf{s}}{\partial \mathbf{x}}$$

The previous equation defines the likelihood (probability) of the observed mixtures given the source signals and the jacobian (the jacobian is a matrix of partial derivatives) of the unmixing matrix.

The source signal is the unmixing matrix, **W**, times the mixtures, **x**:

$$p_x(\mathbf{x}|\mathbf{W}) = p_s(\mathbf{W}^{guess}\mathbf{x})\left|\mathbf{W}^{guess}\right|$$

This is the likelihood given a particular unmixing matrix. The next two steps are to incorporate the independence and non-Gaussianity models.

If the source signals are independent then the joint pdf of the source signals, p_s, is the product of its marginal pdfs. Then one could write p_s as:

$$p_{\mathbf{S}} = p_{s_1} \times p_{s_2} = p(\mathbf{w}_1\mathbf{x}) \times p(\mathbf{w}_2\mathbf{x})$$

Then if one substitutes that expression into the likelihood expression and takes the natural logarithm (the natural logarithm is used because addition is computationally more stable than multiplication), one gets:

$$\text{ln likelihood=ln L}(\mathbf{W}) = \sum_g^m \sum_i^n \ln(p_s(\mathbf{w}_g^T\mathbf{x}_i)^2) + N \ln\left|\mathbf{W}\right|$$

The last thing to do is to put a non-Gaussian model for p_x. One such model that can be used is one that is leptokurtic, which models speech signals. In that case, the function 1-tanh(x)^2 can be used so the completed likelihood expression will be:

$$\text{ln likelihood=ln L}(\mathbf{W}) = \sum_g^m \sum_i^n \ln(1-\tanh(\mathbf{w}_g^T\mathbf{x}_i)^2) + N \ln\left|\mathbf{W}\right|$$

Then using optimization techniques, a computer iterates through various unmixing matrixes to obtain a point at which the above function reaches a maximum value. In the example of having two independent source signals, there are four parameters of the unmixing matrix to guess. If there are

three independent source signals then there are nine parameters. The number of parameters goes up n^n which is extremely quickly. Therefore, it is not possible to sample exhaustively all possibilities of coefficients. It is also important to choose algorithms that converge quickly. As with all optimization techniques it is possible to get stuck in local minima or not to converge at any minimum.

In practice, one can simplify the ICA data set first by doing PCA on the data. The use of PCA is called whitening. If the reader will recall, PCA can express the data so that the components with maximum variance are grouped together. One can then do ICA on a certain subset of the data that PCA has identified to be correlated, which is the approach that I will take in the next two chapters.

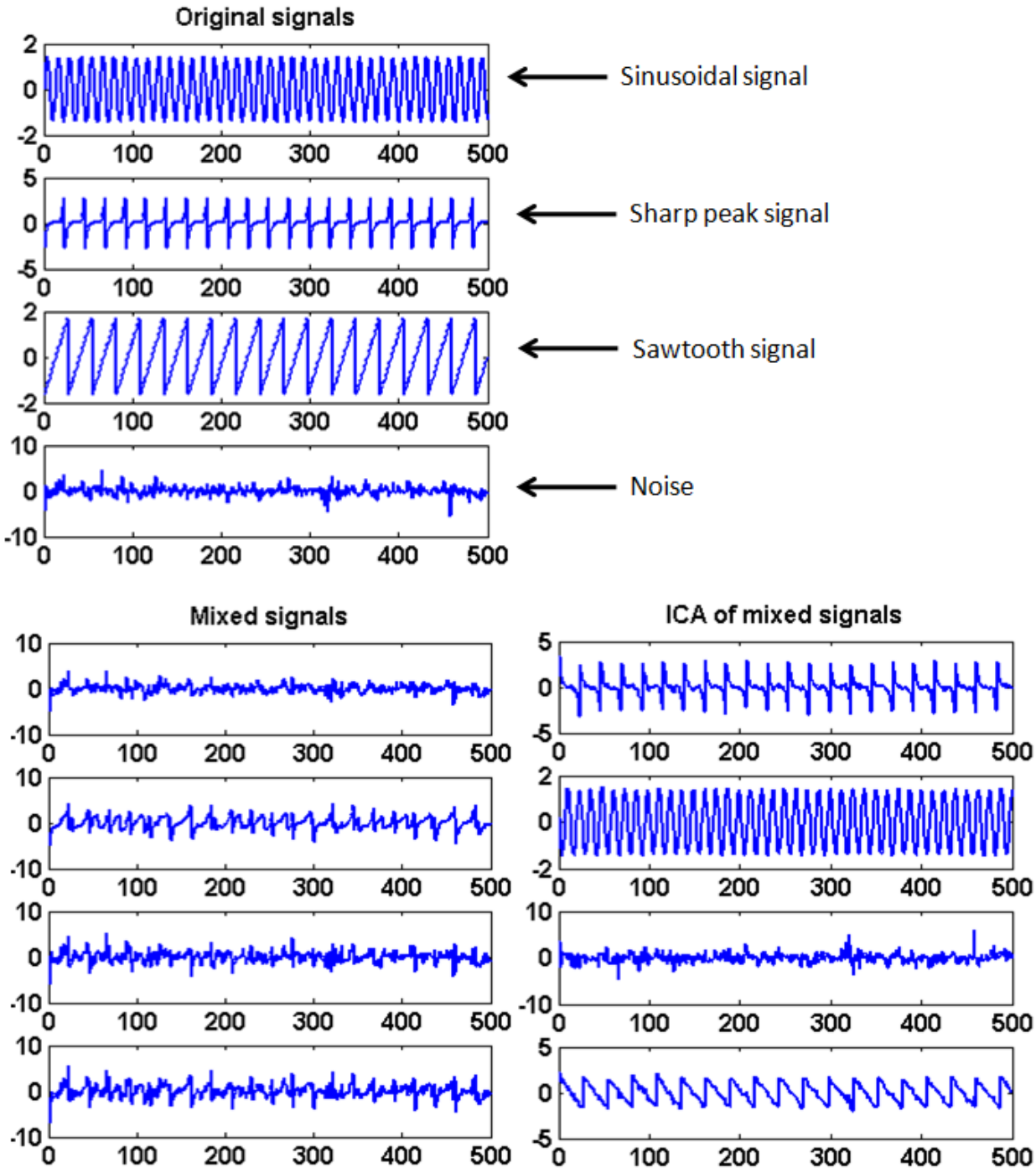
For the work that I will present, I have used freely available software called FastICA, developed by Aapo Hyvarinen and Erkki Oja. The software developers use a method similar to maximum likelihood; the software implements four different non-Gaussian models: cubic called 'pow3' , leptokuritc called 'tanh', super gaussian called 'gauss', and a skewed distribution called 'skew'.

One problem with ICA is that there is no mathematical knowledge of how many independent signals there are. It is up to the user to specify how many independent signals there are. If the user has two mixtures then the maximum search can be for two signals. If the user has three mixtures, then the maximum search can be for three signals (although one can search for two components as well). Also, similar to PCA, the sign of the signals is arbitrary. ICA will also fail if more than one of the signals is Gaussian, as ICA models looks for sources that are more non-Gaussian than the mixtures. Another way ICA can fail to converge is if the assumption of independent components in the data does not hold. One other problem with ICA is that since it is based on random seeding (guessing), every time it is run slightly different values are obtained. This should be kept in mind when generating data as it is very unlikely that one will obtain the same results on subsequent runs. Generally one can set the number of iterations to run as well as the threshold for convergence; time and computer constraints as well as the nature of the data should be considered in deciding the iteration and convergence thresholds.

Figure 44 demonstrates how ICA can find independent components. Four signals are mixed up and then separated by ICA. Note that it is not possible to determine the original order of source signals from the mixtures.

In the next two chapters I will use ICA on the covariance matrix of sequence data. The covariance of one position to another position is considered as a vector and ICA will search for combinations of weights that will result in independence relationships between the vectors. The contribution of each position to the independent vectors can then be examined. This will be shown in detail in the next two chapters.

Figure 44: Demonstration of how ICA can separate mixtures of independent signals. The original signals are mixed together resulting in the mixed signals. ICA procedure is then done to recover the form of the original signals with a multiplicative factor.



Conclusion:

In this chapter I covered several methods to extract information about patterns of coupling from a coupling matrix calculated with SCA. Network graphs depict the amino acids as nodes and the coupling between them as lines. It is necessary to choose thresholds of coupling score to visualize the data but it is possible to easily see patterns of coupling if the number of nodes is not too high. Hierarchical clustering calculates distances between the covariance vectors. Distances can be calculated with several methods. The distances are then used to construct clusters. Clusters, like distances, can be defined in different ways. The advantage of clustering over network graphs is that all positions are compared when calculating the clusters. The disadvantage is that it is not clear what the best distance/linkage method to use is. Principal components analysis is similar to clustering in that it considers the magnitude of correlations between positions; it is different in that it re-expresses the data such that the variance between vectors is maximized and covariance between different vectors is zero. The variance can be used to define groups of positions because one can easily observe how much a particular position contributes to that component's variance; positions that contribute similar amounts can be grouped together. Finally I also considered independent component analysis, which like PCA can also be used to obtain vectors of transformed data but in ICA the vectors are considered to be independent. One can, like in PCA, examine the independent vectors to see which positions contribute to the independence. Positions with similar contributions can be thought of as comprising independent units.

Typically all the methods are used for analyzing protein families. The methods agree with each other. Also under investigation in the lab are additional methods that can more robustly detect patterns of coupling in data.

Chapter 4: Statistical coupling analysis of The TonB dependent receptor family

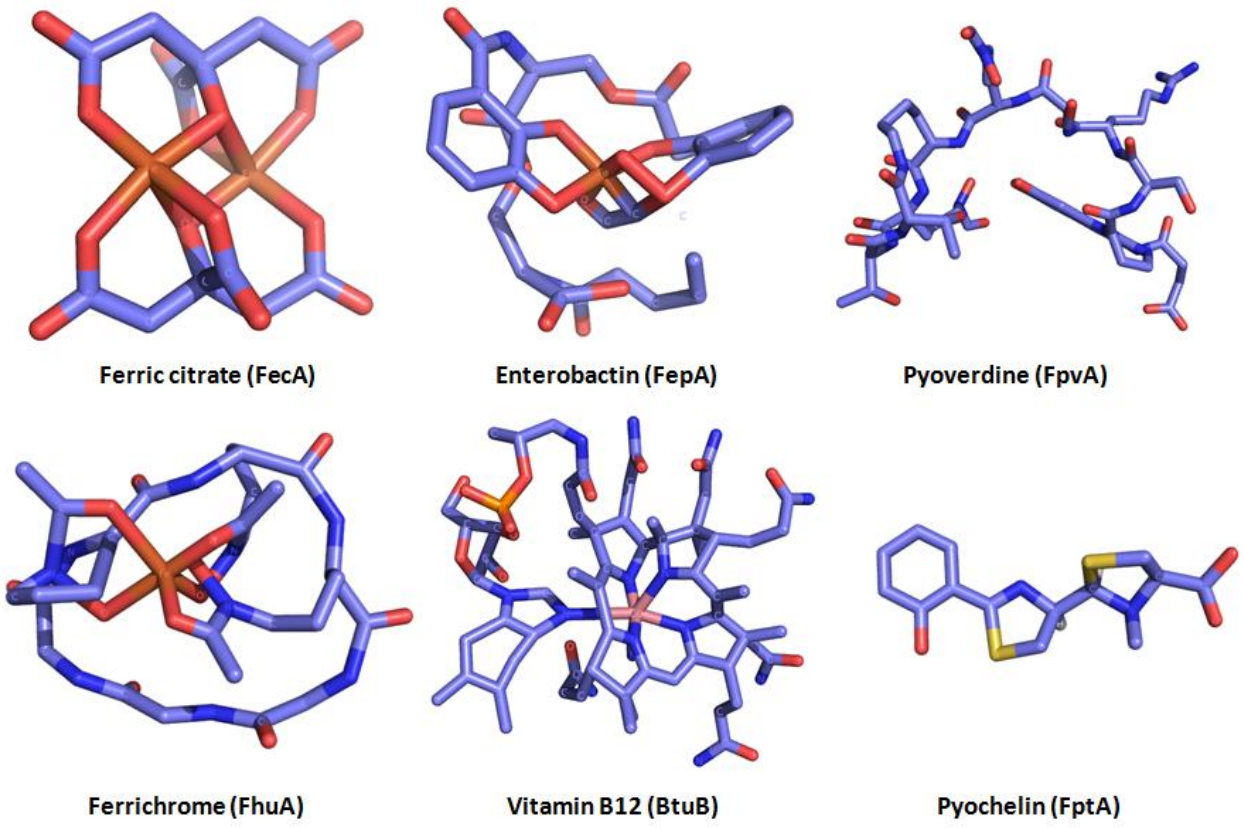
This chapter is divided into three sections. The first section is a brief review of the TonB dependent receptors. The second section describes the statistical coupling analysis (SCA) of the TonB dependent receptors. The third section is experimental work done based on predictions of SCA.

This work was done in close collaboration with Andrew Ferguson in Johann Deisenhofer's laboratory at UT Southwestern. These results were published [55] as part of a larger study by Andrew Ferguson on this family.

Section 1: Review of TonB dependent receptors:

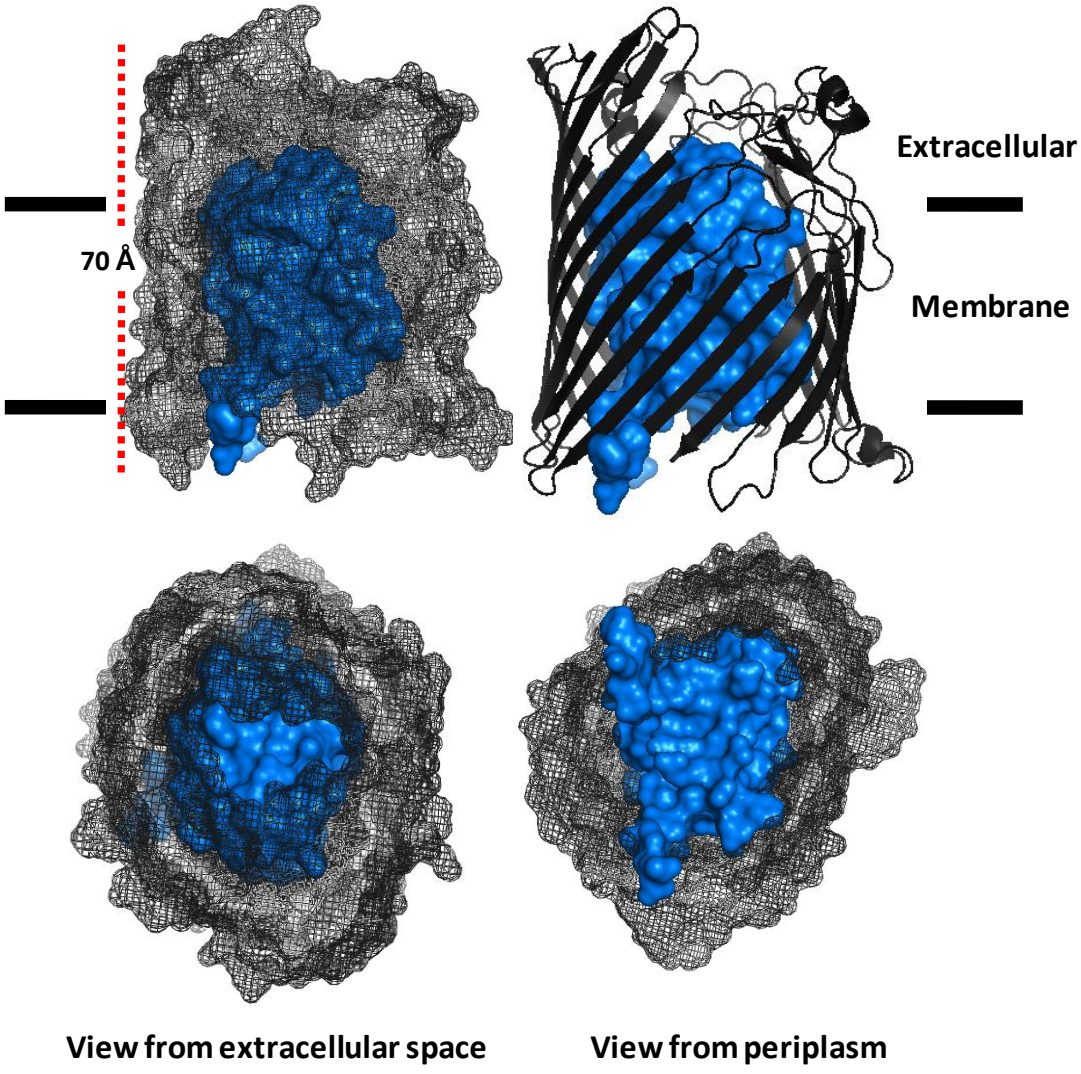
The TonB dependent family of receptors are membrane bound energy coupled transporters that exist in gram negative bacteria and transport metal binding compounds (siderophores), vitamin B and antibiotics among other substances [120]. The wide variation in the structure and size of the receptor ligands (see Figure 45) has resulted in a diverse family of receptors.

Figure 45: Different molecules that bind to TonB dependent receptors. The receptor names are in parenthesis.



In contrast to the diversity of the ligands, the receptors are all structurally similar. The receptors are 600 or more amino acids in size with two domains shared by most family members and an optional third domain by some family members. The two core domains are called the barrel and the plug. These domains are so named because structural analysis reveals a cylindrically shaped 22-strand beta sheet structure called the barrel that in the center has a structure composed of beta-sheets and helixes, called the plug [121]. The plug seals the interior of the barrel with no apparent way that ligands can enter. This is shown in Figure 46 with the iron transporter, FecA which binds ferric citrate.

Figure 46: FecA structure (pdb 1KMO, Ferguson et al, 2002). Different views of the same protein are shown. In this structure there is no ligand. The black mesh and cartoon are the barrel and the blue surface is the plug.



The third domain found in some proteins such as FecA is called the signaling domain. This domain is found N-terminal to the plug and is in the periplasm. Its role is not in transport but in signaling: it triggers a signaling cascade that activates the transcription of some iron-transport related genes upon binding of the ligand.

The protein has a beautiful structure and an important function as iron and other metals are vital for the survival of organisms. However, the way this protein functions is also intriguing. Transport is an active process requiring energy. The energy that drives transport comes from a protein complex (made up of TonB-ExbB -ExbD proteins) in which one member called TonB binds the receptor. Upon binding of TonB, transport commences. The signaling function that occurs in some proteins in this family also requires TonB.

Thus, these proteins have a fascinating energetic cycle:

1. Upon binding of ligand, they communicate to TonB that they have a cargo to transport.
2. Also, upon binding of ligand, they communicate to the signaling domain to activate the signaling process.
3. The TonB protein then binds to the receptor and triggers the opening of the receptor so that ligand transport can occur.
4. After transport, the receptor must then close and reset to be capable of binding another ligand.

Certain structural changes observed in liganded and unliganded structures suggested how the protein functions [121]. As shown in Figure 47 and Figure 48 the ligand binds on the extracellular side, triggering conformational changes that result in the unfolding of a helix in the plug domain. However, there is no hint from the structure as to what the allosteric pathway is nor does there exist feasible experimental assays that could determine this allosteric pathway.

This is where statistical coupling analysis could be useful. Allosteric pathways in proteins require that groups of residues work together to accomplish this function. Thus, a statistical coupling analysis of this protein family could perhaps find this allosteric pathway or provide some clues as to what residues may constitute it. The coupling analysis and part of the subsequent experimental assays was my contribution to this project.

Figure 47: Changes in the barrel of FecA upon ligand binding. The unbound structure is in blue while the ferric dicitrate bound structure is in grey. The ferric dicitrate is shown in red (note that the unbound structure does not have ferric dicitrate and is present there only for reference). There is a helix that unwinds in the bound structure and a loop that changes positions. These changes appear to seal the ferric dicitrate in the barrel prior to transport.

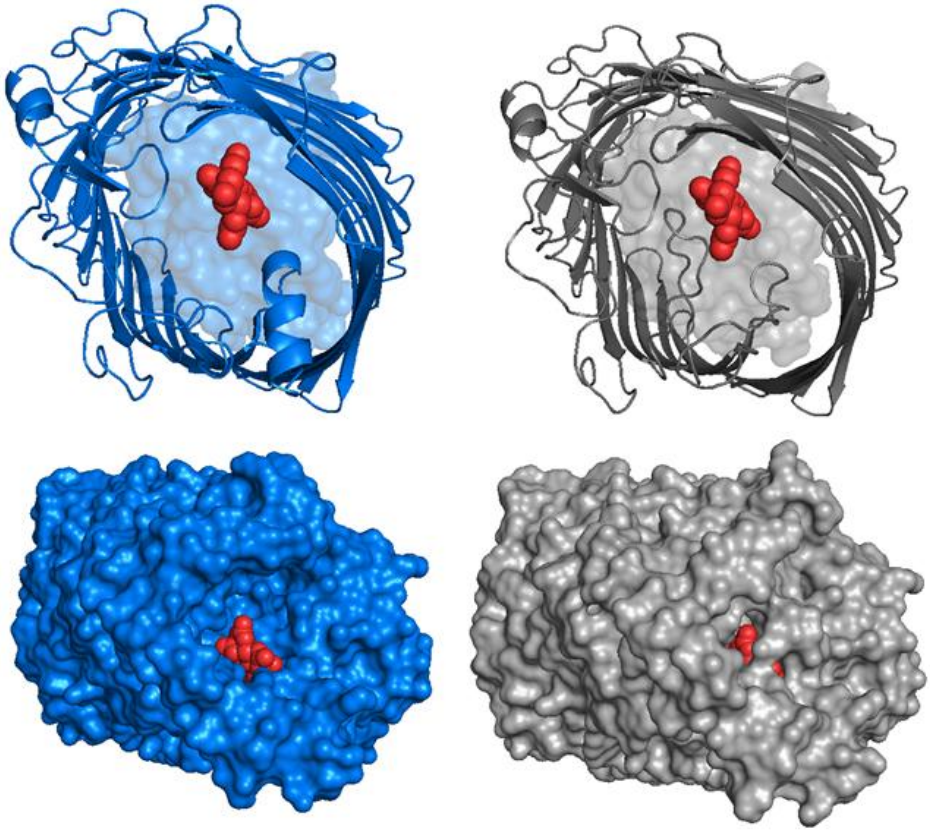
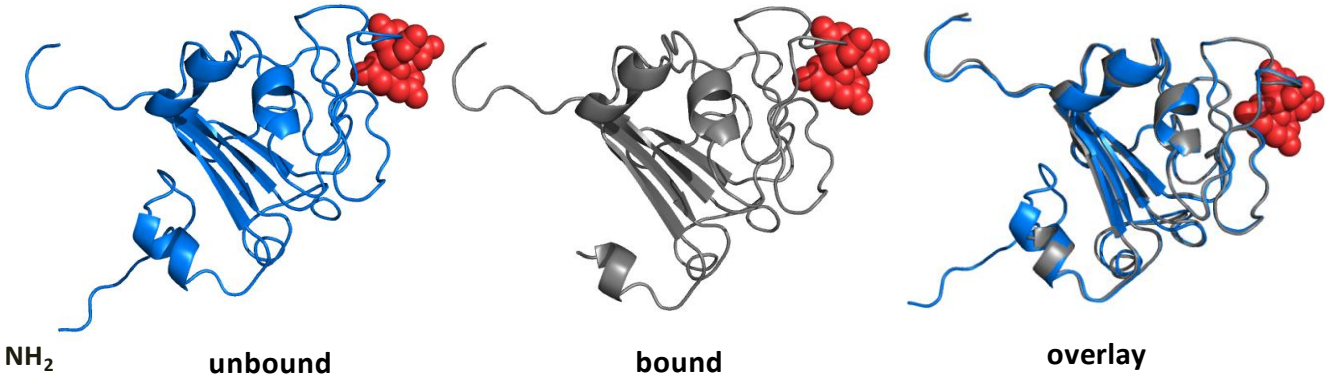


Figure 48: Changes in the plug of FecA upon ligand binding. The blue structure is unbound FecA while the grey is the bound FecA. The red surface is the ferric dicitrate (shown only as a reference in the unbound structure). Visible changes are present only in the N-terminal part of the plug; a helix unwinds upon binding of ligand. The overlay shows that no major changes happen anywhere else in the plug in this structure.



The general approach taken for this protein family, which I will describe in detail in the next sections, was to obtain sequences of four structures, perform a psi-blast on the sequences, do individual and group alignments and compare the features, then make one large alignment with which SCA was done to identify coupled (important) residues. This, as will be described, proved to be a challenging process because these proteins have many transmembrane sheets which share little sequence similarity and it took many approaches to have some confidence in the final sequence alignment. In addition to the barrel and plug alignments, some proteins have the extra signaling domain and a separate alignment was made for the signaling domain to identify functional sites there.

SCA analysis resulted in the identification of a coupled network. To test whether the coupled residues were functionally important, a mutational investigation was done. Andrew Ferguson and I adapted an assay already described and worked out a protocol for its use. I then performed the assay and analyzed the data. The data confirmed that many of the predicted mutants showed a large effect on protein function. A final interesting note on this work came with the publication by two other groups of the partial structure of TonB with the TonB dependent receptor [122, 123] which served to confirm our work as the binding surface predicted by SCA was the same as that observed by the structures.

Section 2: Statistical coupling analysis of Ton B dependent receptors:

The starting point of SCA is to make a diverse alignment of homologous sequences. The strategy used to build the alignment is to take as the starting point the four protein sequences for which structures were solved. This proved to be a very important step because the sequences themselves, without structures, were unalignable due to the low sequence identity in the transmembrane region. The sequences were those of FecA, Btub, FhuA and FepA. The following table (Table 6) summarizes the characteristics of the sequences used to make the seed alignment.

Table 6: TonB dependent receptors used as a basis for alignment

Protein	Transports	gi number	Pdb	Liganded and Unliganded	Signaling?	Length	Structure Reference
FecA	ferric dictitrate	20150926	1KMP	Yes	Yes	774	[121]
FhuA	ferrichrome	4389223	1BY3	Yes	No	714	[124]
FepA	ferric enterobactin	6730010	1FEP	No	No	724	[125]
Btub	Vitamin B12	30749694	1NQH	Yes	No	594	[126]

The challenge in building this alignment is due to the length, diversity, structure and number of these proteins in this family. The length of these proteins is about 600-800 amino acids which is challenging when one is attempting to align hundreds of them. The diversity of the sequences is also large. The pairwise identity between the proteins above is shown in Table 7. Identities below 15% are what is expected from two random sequences and these sequences have pairwise identities ranging from 11.2% to 17.2%. Specifically, the barrel regions where there is a 500 amino acid residue stretch of low identity region is most difficult. Many of the strands have very little conservation so that standard alignment programs that only use sequence information fail completely. Aligning this family required structure and profile based methods. After trying different software, I found that the NCBI program, Cn3D [127] remained stable and gave meaningful alignments. Cn3D works because it incorporates structure information into making a profile based alignment. Sequences can be aligned to this profile by sequential addition until all have been added. Even after Cn3D alignment, extensive manual adjustment was necessary.

Table 7: Pairwise identities between four sequences with structures

	FecA	FhuA	FepA	BtuB
FecA		14.8	13.8	14.9
FhuA	14.8		11.2	14.6
FepA	13.8	11.2		17.2
BtuB	14.9	14.6	17.2	

To obtain additional sequences, I used NCBI's psi-blast search program [128]. Psi-blast works by constructing an amino acid profile of sequences after each iteration and then finding additional sequences that match the profile after multiple iterations. This method has greater power to find distant homologues than the standard blast algorithm.

Each of the four proteins (FecA, FhuA, FepA, Btub) was used for a separate psi-blast search. Then all the identified sequences were compared together. The vast majority of found sequences were common to all four input sequences. Non redundant sequences were pooled together to yield a set of ~2300 sequences. This large set of sequences was further filtered by removing sequences that are larger than 800 amino acids, removing homologues that were identified as putative or hypothetical or heme binding or transferrin binding and removing sequences with differently sized secondary structure elements. The final alignment consisted of 541 sequences.

As an additional check on the alignment quality, the identified sequences from all four proteins were individually aligned to obtain four alignments. Inspection of these four alignments showed that all alignments contained the same motifs; or I should say there was no feature in one set of proteins that I could distinguish as coming only from that family. Therefore, the structure and sequence similarity is sufficient grounds to include homologues of all proteins in the final alignment.

During and after alignment construction two structures came out of the pyoverdine [129] and pyochelin [130] transporters from *Pseudomonas aerugionosa*. These sequences were in the final alignment I made allowing for an independent check of the alignment quality. The alignment was excellent agreement with these two structures.

As previously discussed, some transporters have a signaling domain. The signaling domain is typically a short (~80 amino acid) sequence N-terminal to the plug. To make this alignment I used the NMR structures obtained by Andrew Ferguson and colleagues [55] of the signaling domain of FecA and PupA (a TonB dependent receptor from *Pseudomonas putida)*. That alignment contains 296 sequences.

With the alignment done, SCA analysis could proceed. For this analysis, positions which contained more than 20% gaps were not considered; this has the effect of removing from the analysis poorly aligned positions. I will show the SCA results using network graphs, hierarchical clustering, principal components analysis and independent component analysis.

Covariance analysis shows little coupling between positions making it difficult to make sense of the coupling pattern (Figure 49). This is quantitatively shown as histograms of coupling values in Figure 50.

Figure 49: The weighted covariance matrix (SCA matrix) for 588 positions in the TonB dependent receptor family. Clearly most of the couplings are weak.

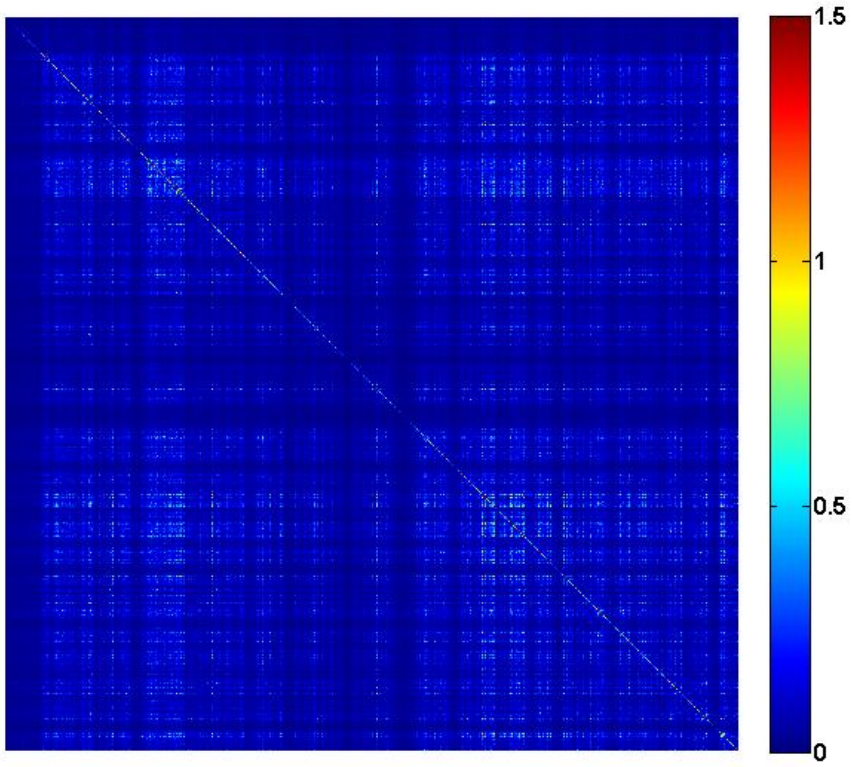
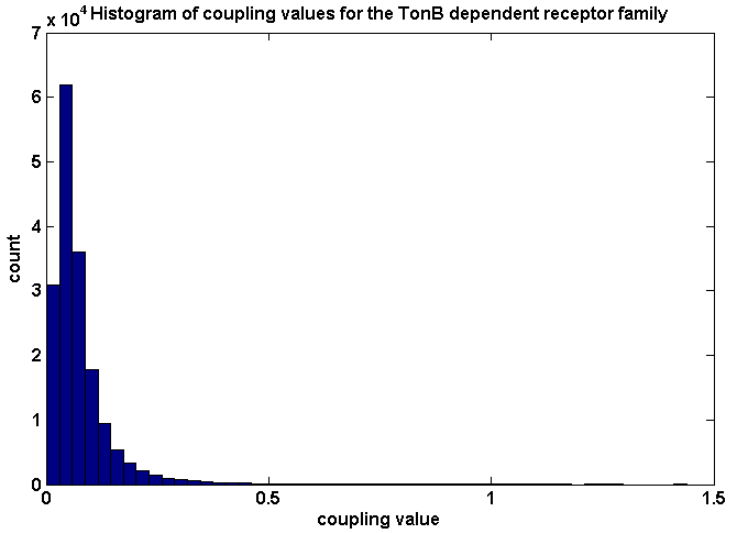


Figure 50: Histogram of coupling values for the TonB dependent receptor family. Most values are very small and there are few high value couplings.

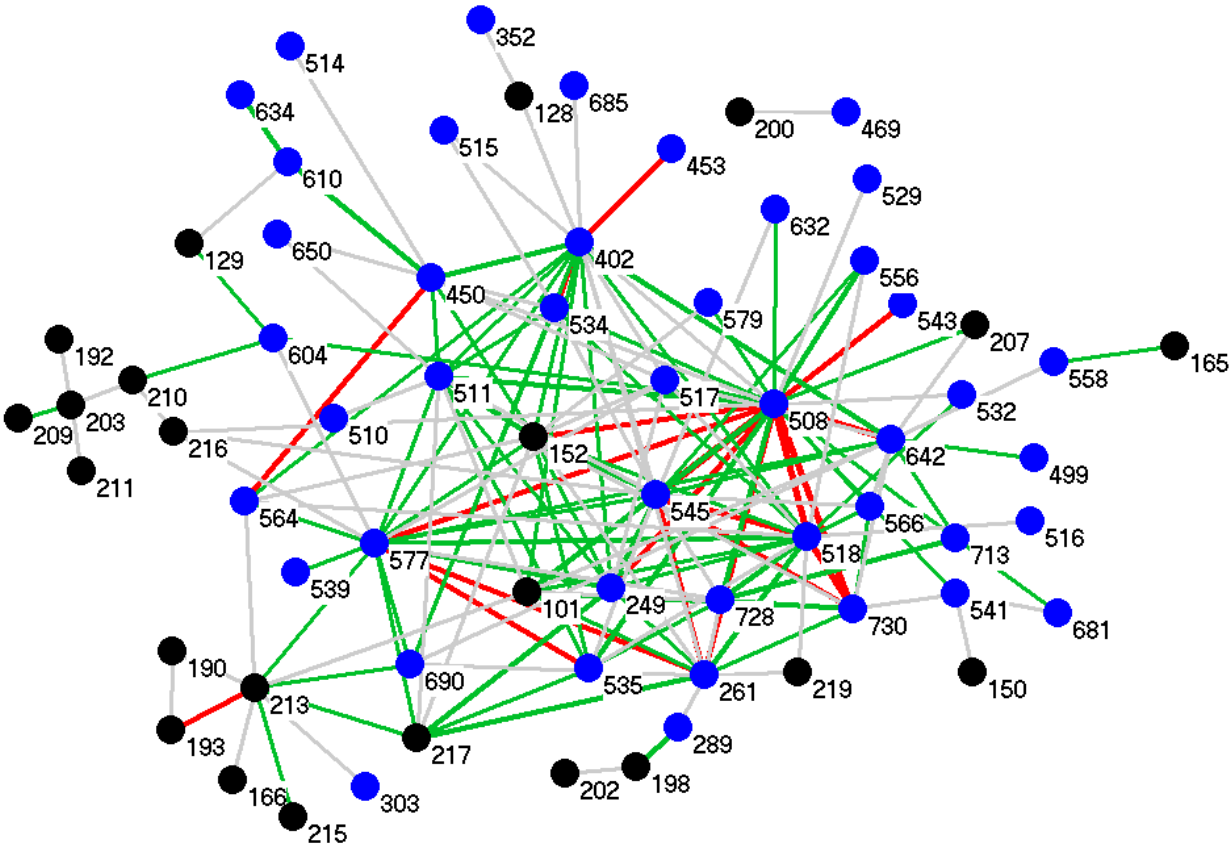


To see if any relationships can be discerned a network analysis is done (Figure 51). Analysis of Figure 51 shows some interesting relationships. First are two sets of positions in the plug: (209, 203, 210, 211,192, 216) and (190, 193, 213, 166, 215, 217). Then there is the large set of coupled positions in the

center (152, 511, 450, 402, 534, 517, 545, 249, 101, 728, 518, 535, 508, 566) that forms associations with other positions in the plug and the barrel. What is striking about the network of coupled positions is that positions distant in sequence space are nonetheless coupled strongly.

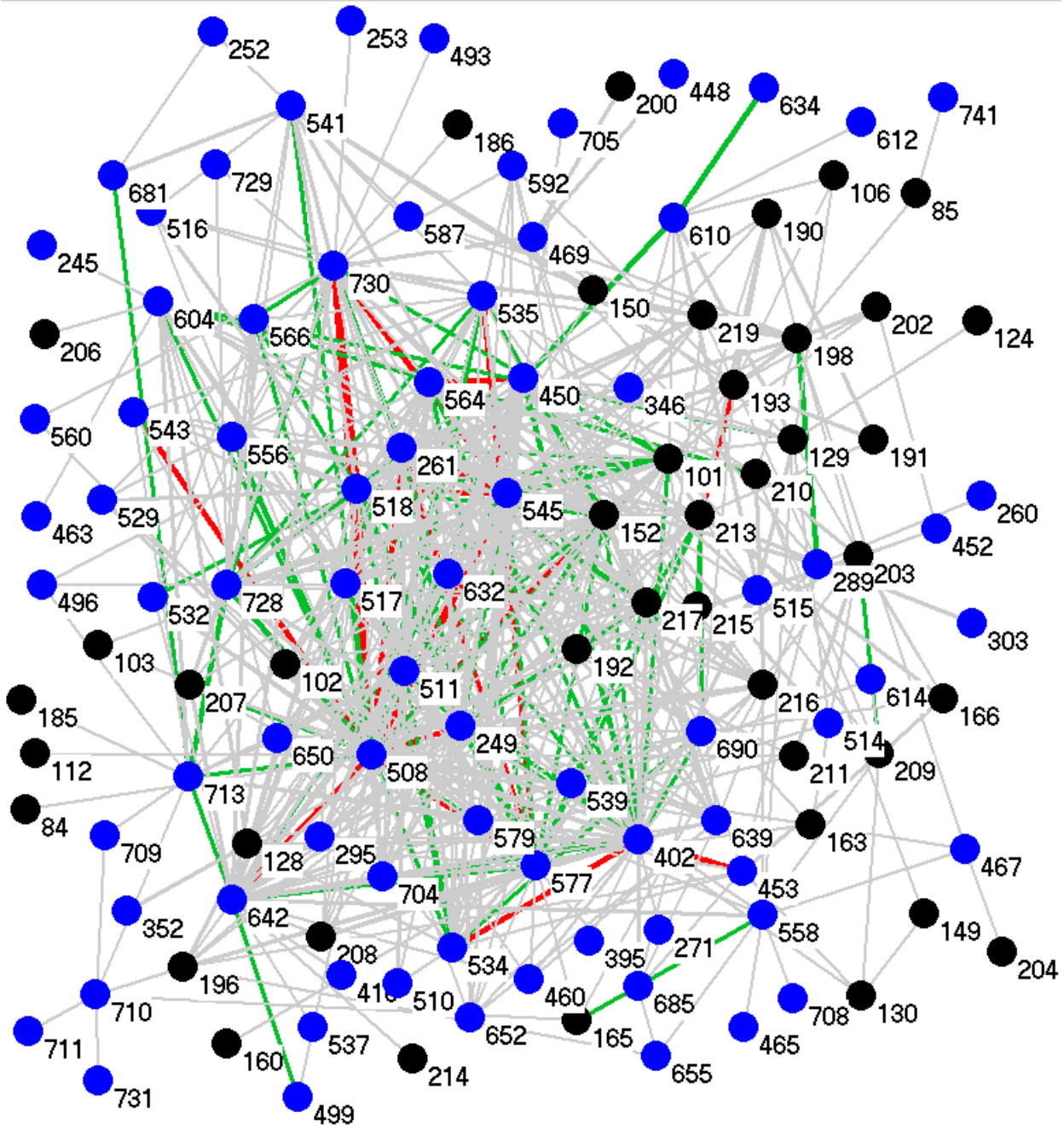
A more extensive network consisting of 120 nodes is shown on the next page (Figure 52) with the cutoff being coupling scores of greater than 0.5. Clearly it is difficult to distinguish relationships with this type of graph when the number of nodes is high. Note that everything seems to be coupled together in this figure. However, while this network contains 121 positions (20% of the aligned protein), they only comprise 0.3% of all possible couplings.

Figure 51: TonB dependent receptor network analysis. 67 Positions that have a coupling score greater than or equal to 0.67 are shown in the figure below. Red lines show coupling values higher than 1. Green lines show coupling values between 0.75 and 1 and grey lines show coupling values between 0.67 and 0.75. The black circles are positions that are in the plug and the blue circles are positions in the barrel.



Greater insight into the pattern of couplings can be appreciated by plotting these positions onto the structure but this will be shown after analyzing the coupling pattern with other methods.

Figure 52: TonB dependent network graph showing all 121 positions with coupling value greater than 0.5. Color scheme is as in the previous figure.



The second approach to use is for analyzing the coupling matrix is hierarchical clustering method. The purpose of this, as previously mentioned is to construct groupings of the data. There may be one or more groupings. The first thing I did with this method is to calculate which method gives the best trees. As shown in Table 8, the Chebychev method with average linkage has the 'best' tree (Figure 53).

Table 8: Hierarchical coupling analysis of the TonB dependent receptor family plug and barrel region. Centroid, median and ward calculations can only be done on Euclidean distances.

		linkage methods						
		'average'	'centroid'	'complete'	'median'	'single'	'ward'	'weighted'
distance methods	'euclidean'	0.965	0.949	0.902	0.919	0.901	0.781	0.945
	'seuclidean'	0.893	NaN	0.734	NaN	0.832	NaN	0.354
	'mahalanobis'	0.220	NaN	0.027	NaN	0.063	NaN	0.195
	'cityblock'	0.891	NaN	0.735	NaN	0.851	NaN	0.683
	'cosine'	0.916	NaN	0.547	NaN	0.890	NaN	0.838
	'correlation'	0.870	NaN	0.614	NaN	0.844	NaN	0.830
	'spearman'	0.844	NaN	0.488	NaN	0.811	NaN	0.671
	'chebychev'	0.996	NaN	0.992	NaN	0.984	NaN	0.989

The Euclidean distance with different linkage methods also works very well as the correlation scores range from 0.78 to 0.97. The cityblock distances with complete linkage does not work as well but is still shows a correlation of 0.74 and in fact visual inspection shows a more compact grouping of positions (Figure 54) with the complete linkage/cityblock distance tree. To obtain a list of the coupled positions, it is possible to zoom into the leaves of the trees and the matrix. This is done for the cityblock/complete linkage scheme as the leaves are clustered together more (shown in Figure 55). While the highly coupled positions are easily read from the zoomed graph, it is still not easy to see what is coupled to what. To further examine this, more analytical methods are presented next.

Figure 53: Hierarchical clustering of TonB dependent receptor family using chebychev distances and average linkage. The clustered matrix shows a regular pattern of small sets of clustered positions. The tree reveals no major groupings.

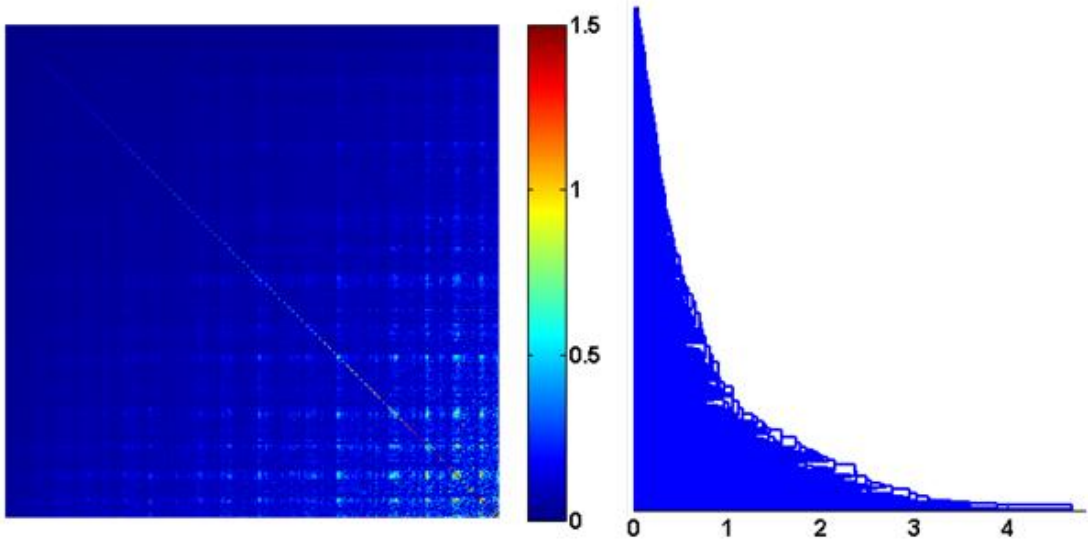


Figure 54: Hierarchical clustering of TonB dependent receptors using cityblock distance and complete linkage. There is one coupled set in the lower right.

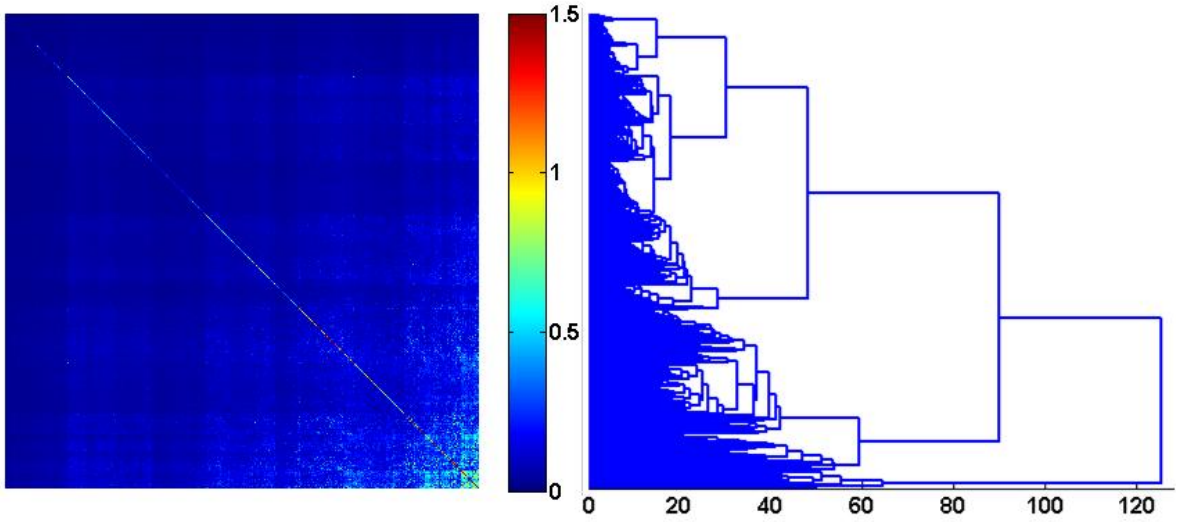
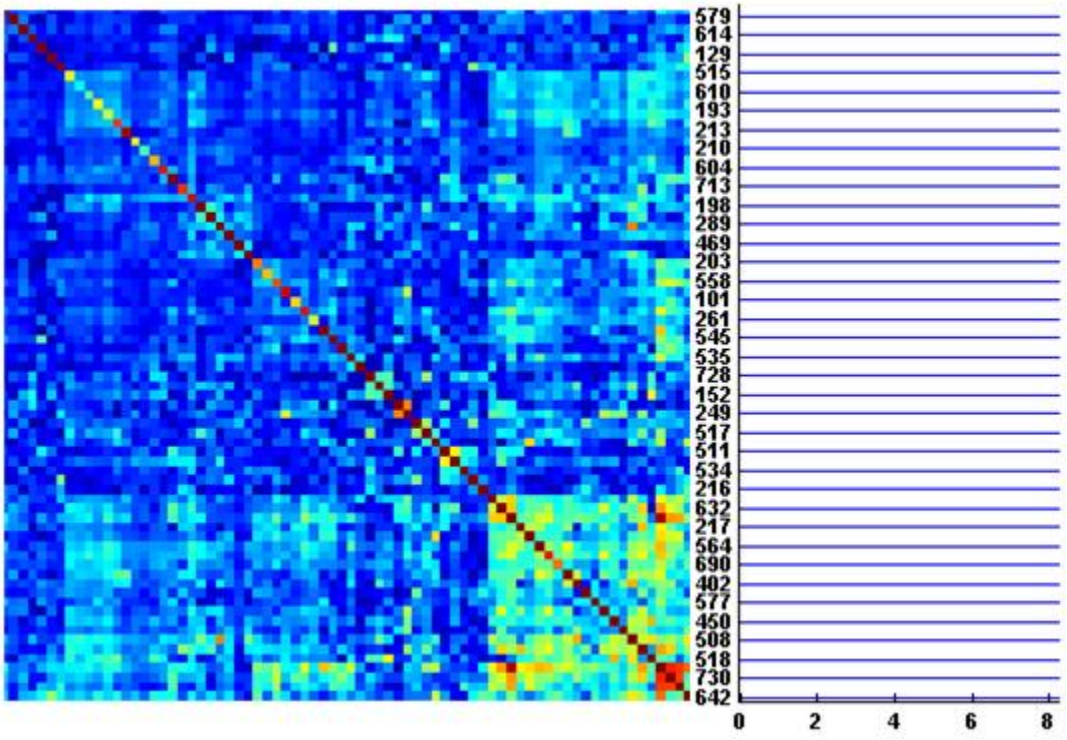


Figure 55: Close up of hierarchical clustering (cityblock/complete). The numbers on the left of the tree are the structure numbers of FecA.



Principal components analysis as described in chapter 3 can be used to analyze covariance matrices. In order to obtain some confidence of what associations are due to random noise, it is necessary to look at the distribution of the eigenvalues of randomized matrices. The eigenvalues of the original matrix that are of high magnitude are then considered significant. This use of PCA is known as spectral clustering.

The first thing to do is to calculate the eigenvalues of the SCA matrix for both the original alignment and a vertically scrambled alignment (Figure 57). As is apparent, both the random matrix and the normal matrix have a single high value eigenvalue (at 57 for the normal matrix and 31 for the random matrix). These values are expected because the covariance matrices are made from a set of non-zero mean data and with PCA, the first dimension of such a non-zero mean set of data is positive and the first eigenvalue reflects that. This is demonstrated in Figure 56.

Figure 57: Eigenvalue histograms of the coupling matrix and the random matrix. The left panel shows the whole histogram and the right panel zooms onto the y-axis so that additional eigenvalues can be resolved.

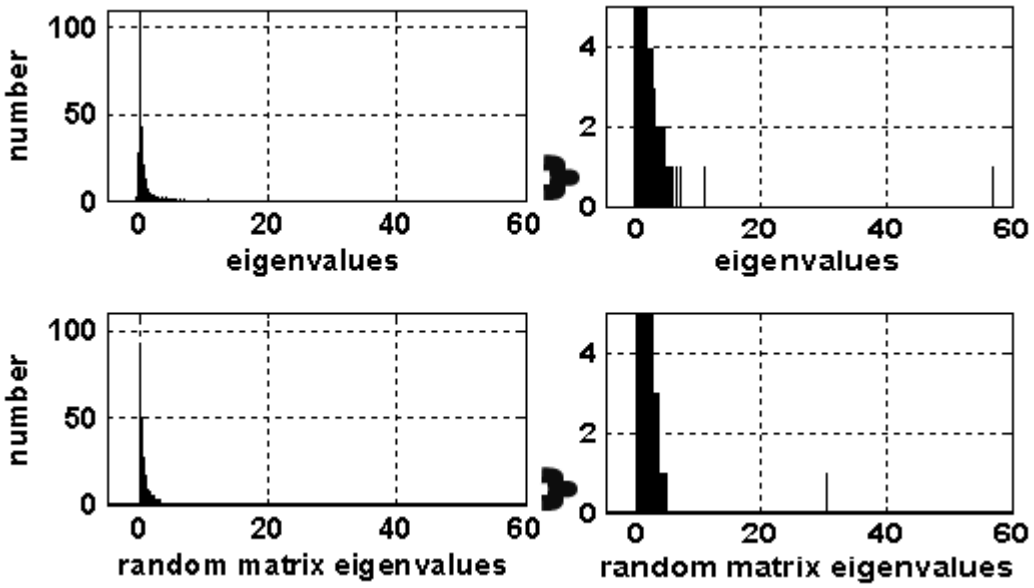
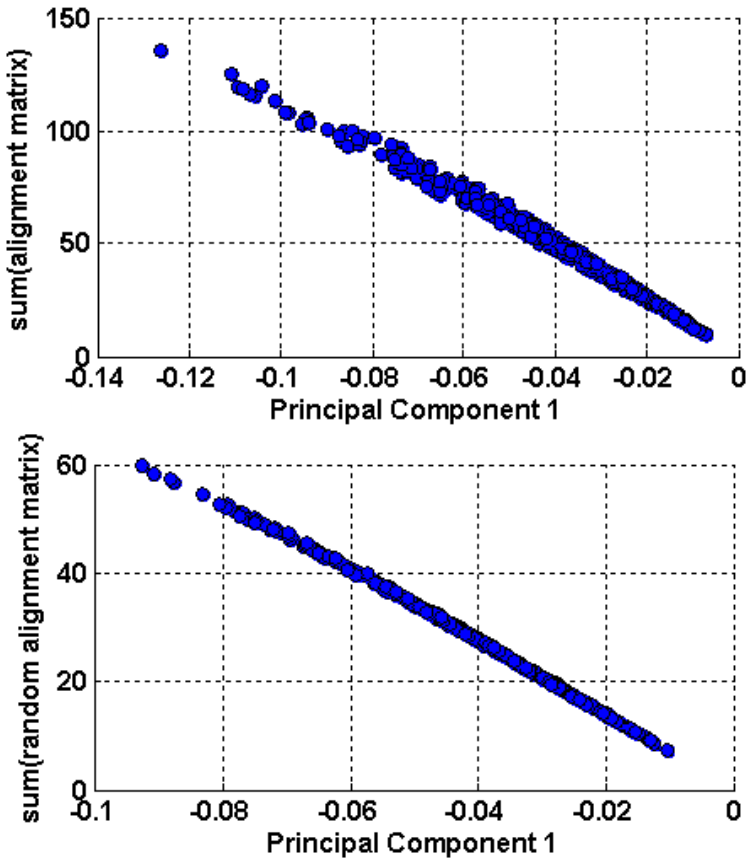


Figure 56: The first eigenvalue correlates well with the magnitude of the coupling for both the random matrix and the normal matrix.



Since the first eigenvalue and hence the first eigenvector does not carry the coupling information, it can be discarded. The eigenvalue histogram with the highest eigenvalue removed is shown in Figure 58. Now it is clear that the random matrix has a high eigenvalue of magnitude 4.92 whereas the normal matrix has eight eigenvalues higher than that. I will consider all eigenvalues higher than 5.0 to be significant and indicative of coupling. This results in eight significant eigenvectors (excluding the first eigenvector). An eigenvector in this case is a list of 588 numbers (also known as weights or factor loadings) that specify to what degree a particular position in the alignment contributes to the magnitude of the whole eigenvalue. The farther a weight is from zero, the more the position contributes (see Figure 59 for distributions of the weights and Figure 60 for plots of the weights against each other). If one takes weights that are 0.1 units away from zero in either direction (0.1 is at least two standard deviations away from the mean, see Figure 59), one obtains a list of positions corresponding to particular residues that are considered to be highly coupled (Table 9). In the TonB family, this list has 72 unique amino acids, comprising 12% of the alignment positions. These positions have been mapped onto the structure of the bound FecA in Figure 62 and Figure 61.

Figure 58: Eigenvalue spectrum discarding the top eigenvector for both the random and normal matrix.

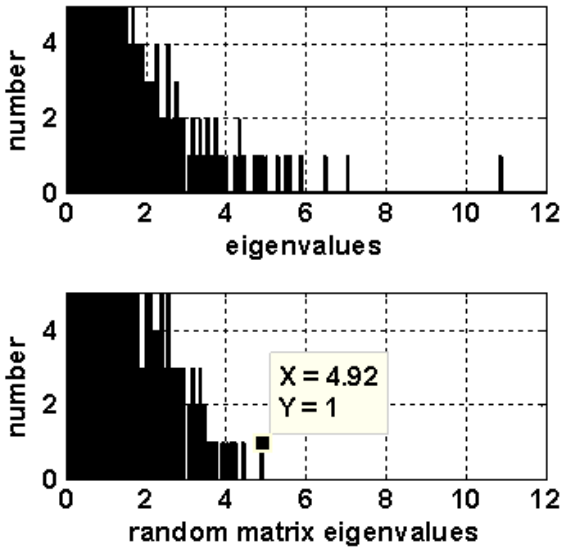


Table 9: The positions with weights more than 0.1 units from zero along the significant eigenvectors.

	eigenvectors							
	2	3	4	5	6	7	8	9
Positions (FecA numbering)	101	129	128	193	128	165	165	101
	152	192	190	198	193	166	166	106
	203	261	191	203	198	192	198	130
	209	303	193	213	210	203	203	165
	212	402	198	219	216	209	210	190
	217	450	213	289	450	210	213	191
	249	453	217	303	467	211	289	193
	261	465	261	453	516	213	452	198
	402	508	289	469	541	215	469	202
	508	511	450	558	558	289	541	209
	511	515	499	604	579	511	558	210
	517	518	508	642	604	539	604	219
	518	532	558	681	610	558	642	261
	535	534	610	709	614	639	650	303
	545	541	642	710	690	642	681	469
	556	543	685	711	710	681	690	515
	558	545	690	713	713	690		534
	564	558	713	730	728	709		535
	577	564				710		539
	709	566				713		543
	710	610				726		577
	728	650						579
	730	690						642
		730						685
								728
								741

Figure 59: Histograms of principal components of the TonB dependent receptor family. The distribution of weights in each significant principal component is shown. The green line is the mean of the distribution and the red dashed line is two standard deviations from the mean. The black dashed line represents the cutoff used (0.1 units away from zero).

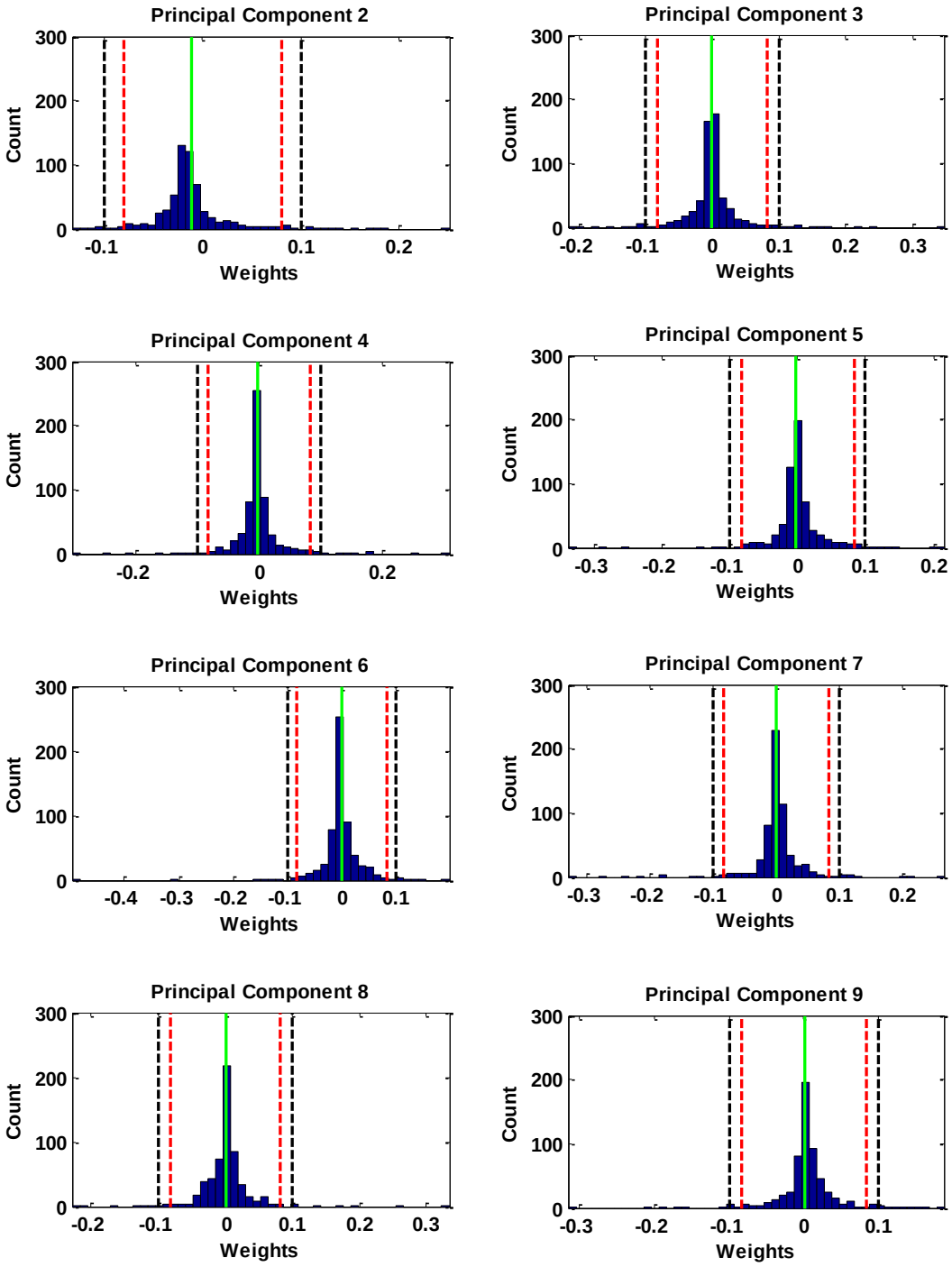


Figure 60: Weights of principal components 2-4 plotted against each other. The number next to the points corresponds to the amino acids of FecA. The blue dots are residues that come from the barrel and the black dots are residues that come from the plug. Dashed lines represent the 0.1 threshold.

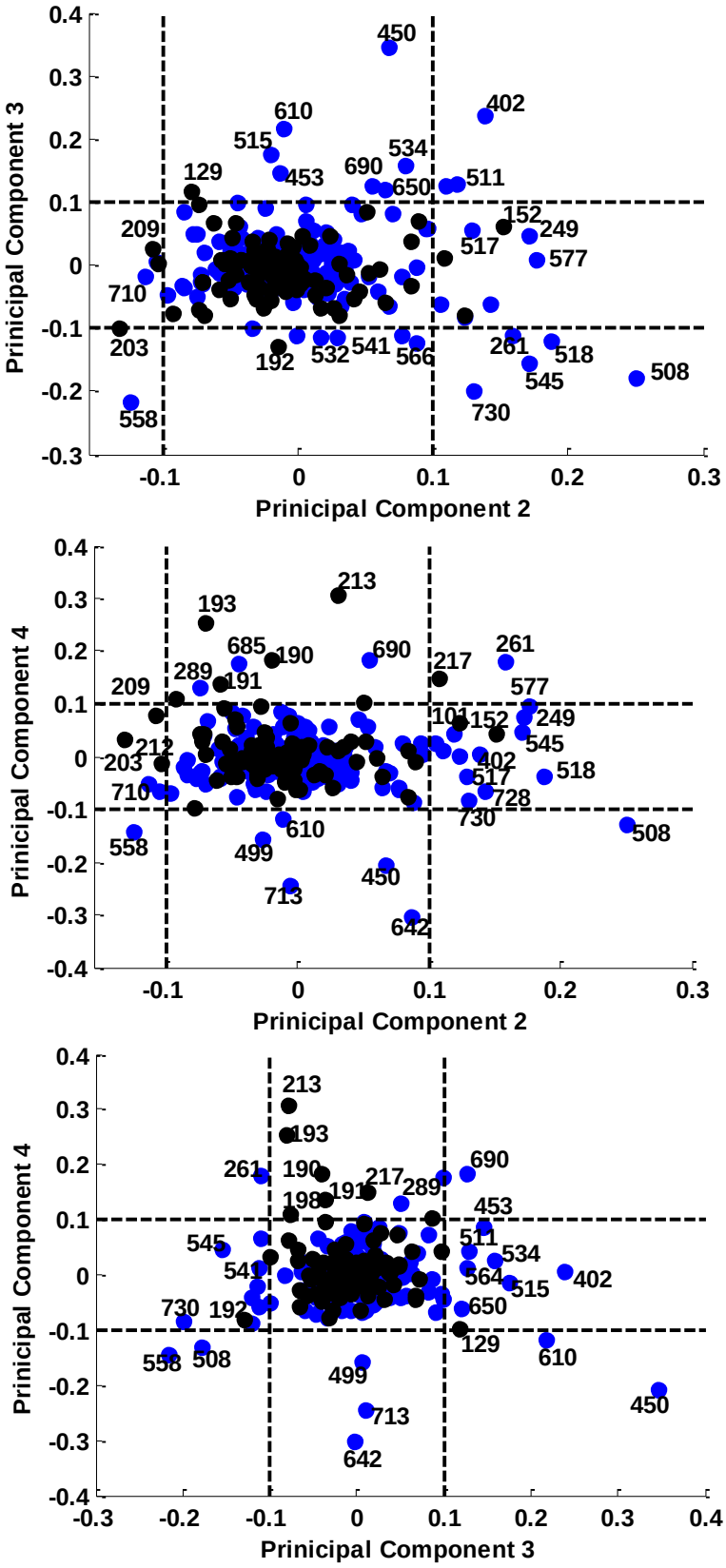


Figure 62: Mapping of coupled positions onto the structure of bound FecA. The black cartoon is the barrel, the blue cartoon is the plug, the orange spheres is the ferric dicitrate and the red surface is the network of 72 coupled amino acids.

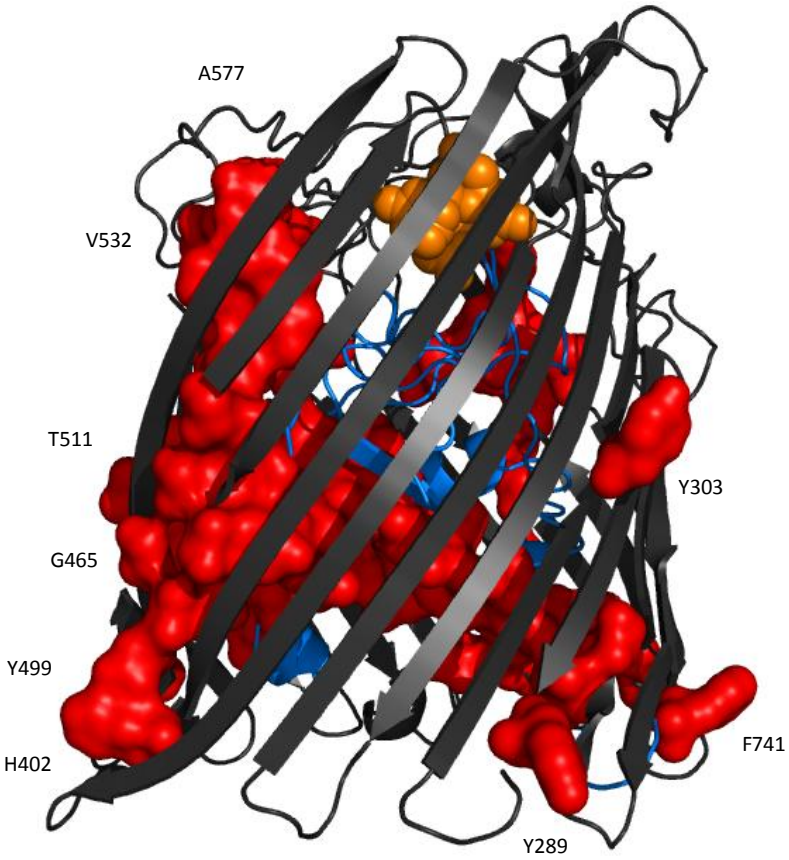
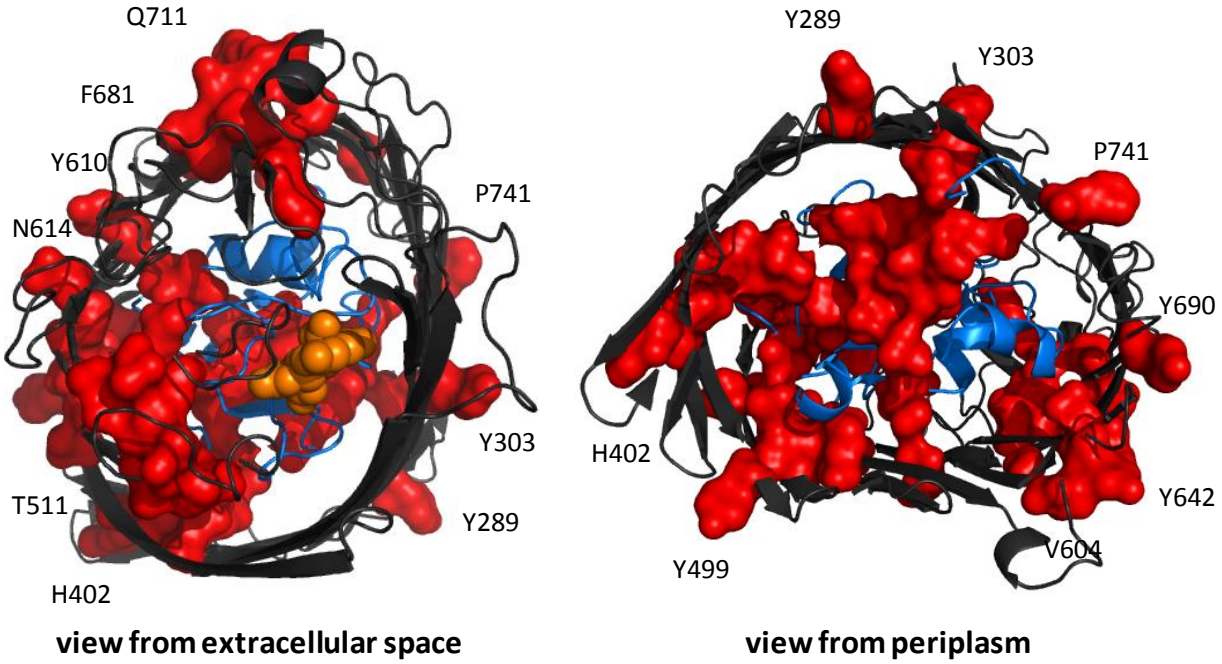
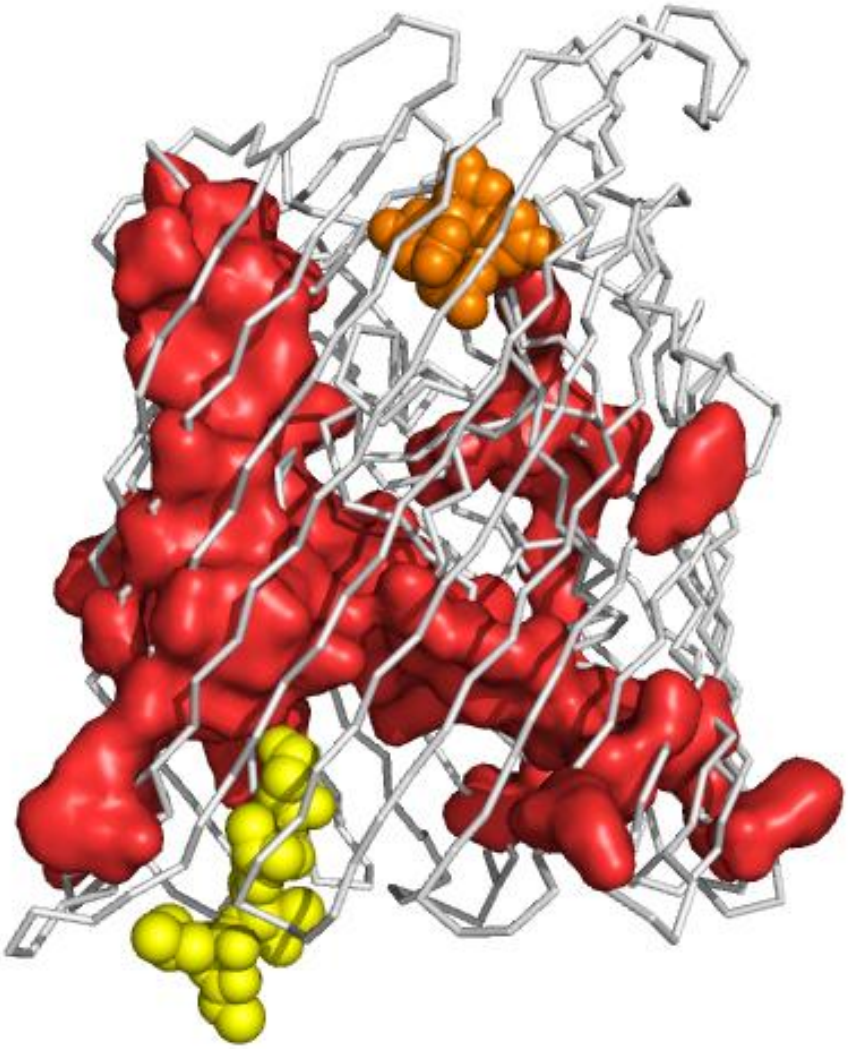


Figure 61: Additional views of the coupling network mapped onto the FecA citrate bound structure. The black cartoon is the barrel, the blue cartoon is the plug, the orange spheres are the atoms of the ferric dicitrate and the red surface is the network of 72 coupled amino acids. Selected positions are labeled.



What is striking about this network is how it extends in a physically continuous way from the region surrounding the binding pocket all the way down to what is known as the switch helix. This is better appreciated with simplified view as in Figure 63. There are also additional couplings whose significance was not apparent at that time. One could imagine that some of these regions could be important interactions with the membrane that the receptor or with the TonB protein that provides the energy for transport.

Figure 63: The coupled network with respect to the TonB box (yellow) and the ferric dicitrate (orange). Note that this is a combination of two structures: FecA bound and FecA unbound. The TonB box is disordered in the FecA bound structure and the dicitrate is not present in the FecA unbound structure. The grey alpha carbon trace is that of bound FecA.



The goal of ICA as explained previously is to identify independent components of the data. This differs from principal components analysis in several important ways. First, ICA finds statistically independent components whereas PCA finds decorrelated components. Second, ICA is non-deterministic owing to the fact that it uses maximum-likelihood methods in which one guesses the form of the data and then through iteration attempts from the guesses to get at the characteristics of the data. Third, ICA can find at most one non-Gaussian independent component. ICA and PCA are similar in one respect however, using both methods one cannot tell which are the important components. This ambiguity is compounded with ICA because one often cannot tell which model of the data to use AND one must give the algorithm the number of independent components to use. That is, I as the user would have to guess that there is some number of independent components and then analyze the output of the program and see if it makes sense.

The way I analyze ICA data is to plot the source signals. The source signals are weights for each residue in the alignment so if one source signal is plotted against another (as in a scatter plot), the pattern of scattered points can reveal whether there is independence. Specifically, if positions are orthogonal to each other in the scatter plot then they are considered to be statistically independent. The goal then of ICA analysis for sequence data is to see if there is any evidence that sets of positions could be independent.

The results of ICA analysis on the TonB dependent family does not reveal any independent components. As the network diagrams show, all coupled positions are about equally coupled to most other positions. This is revealed in the ICA analysis as, whatever model or parameters are chosen, the vast majority of points lie on one axis. This is shown in Figure 64 for the case where there are only two components. The bulk of positions lie along one axis while only one position (558) appears to lie along a separate axis. The same result holds if more components are added or if different eigenvalues are selected as shown in Figure 65 and Figure 66.

From this I can conclude that the coupled sets of positions in the TonB dependent family do not show any evidence of consisting of independent sets of positions.

Figure 64: Independent two component analysis of the TonB dependent receptor family reveals no independent sets of positions. The parameters used for ICA are shown in each graph. In this example, different models are tried, while two components are searched for using the eigenvalues from 2 to 10. The colors refer to the following:
red: residues identified as coupled using the PCA analysis
green: residues have at least one coupling score greater than 0.5. Note that this includes all the red colored positions.
blue: the remaining alignment positions. Selected residues are marked using FecA structure numbering.

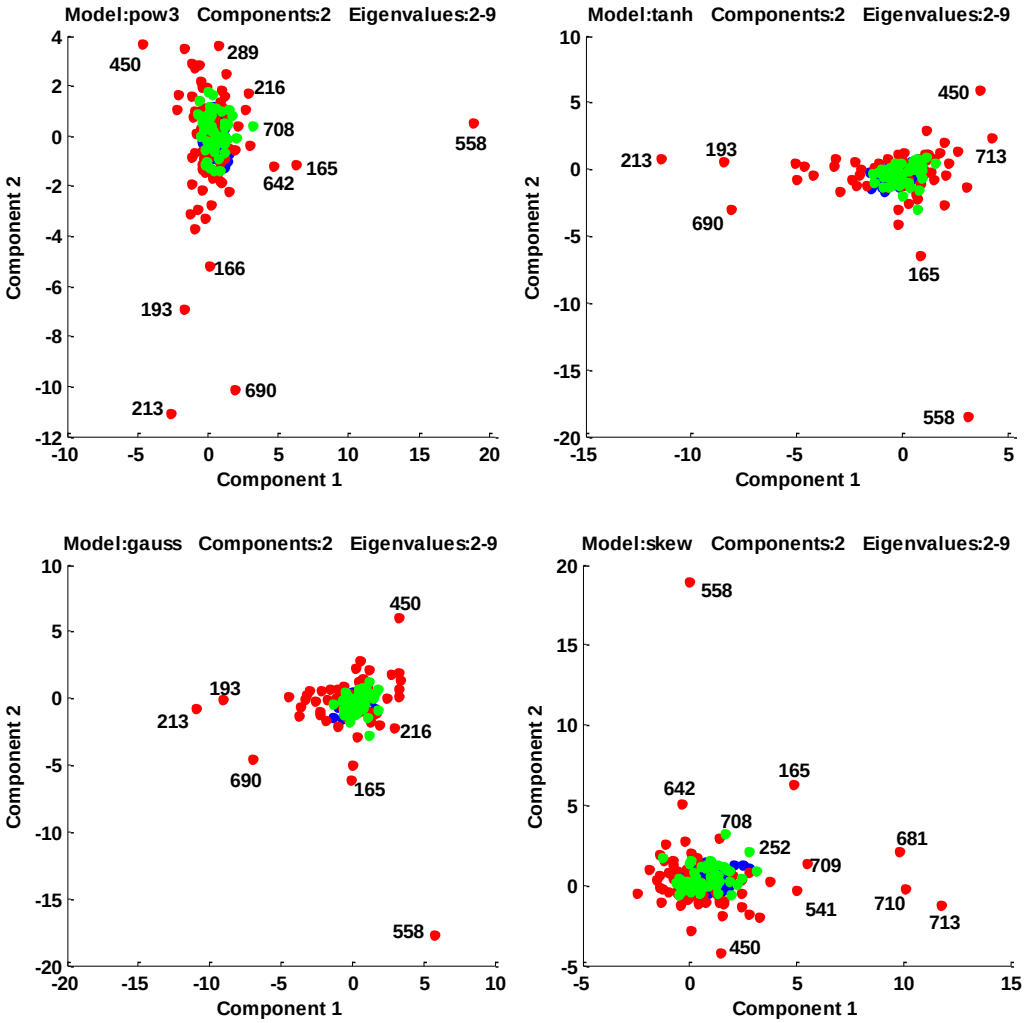


Figure 65: Independent three component analysis of the TonB dependent receptor family reveals no independent sets of positions. The parameters used for ICA are shown in each graph. In this example, different models are tried, while three components are searched for using the eigenvalues from 2 to 10. The colors refer to the following:
red: residues identified as coupled using the PCA analysis
green: residues have at least one coupling score greater than 0.5. Note that this includes all the red colored positions.
blue: the remaining alignment positions. Selected residues are marked using FecA structure numbering.

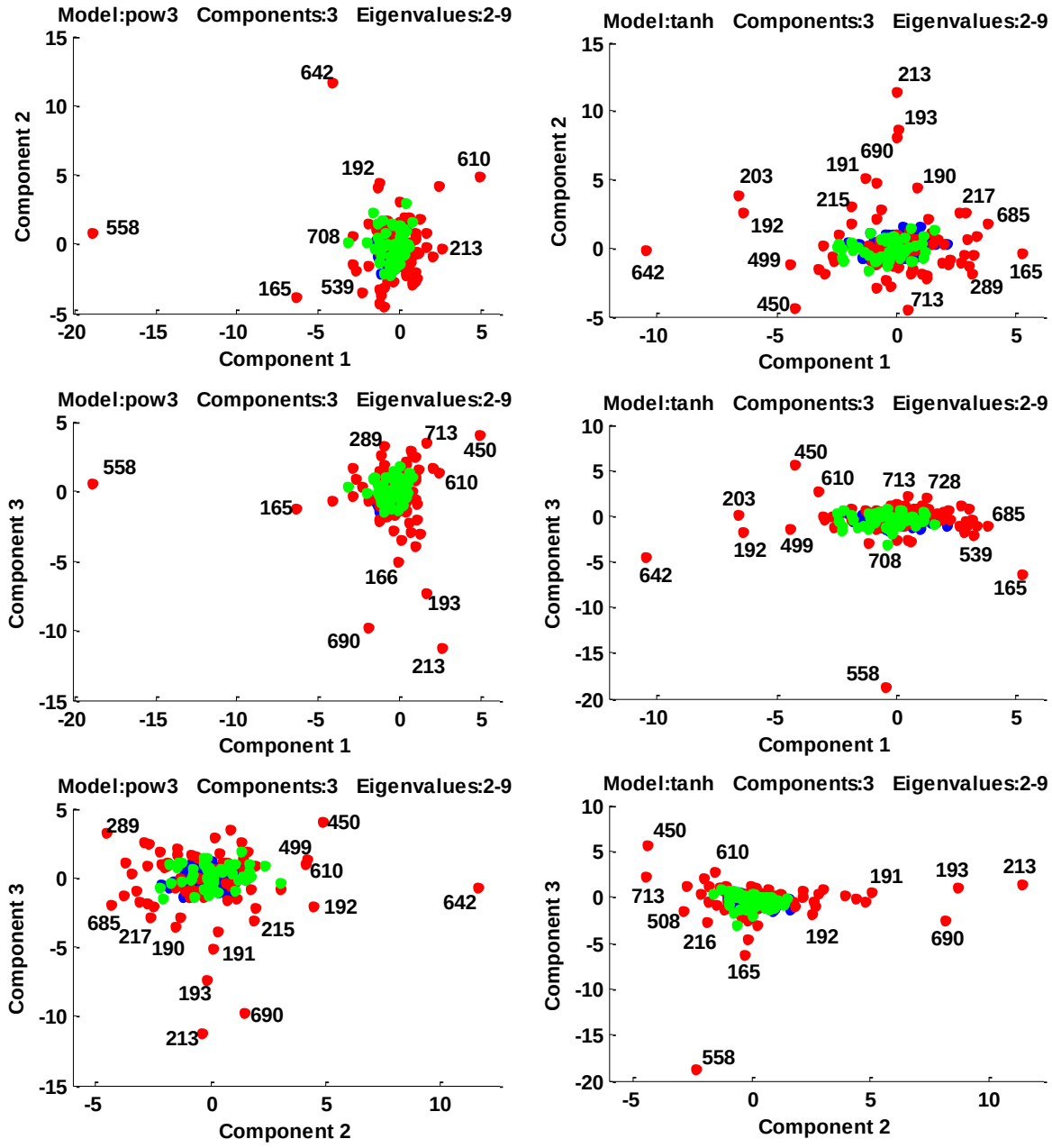
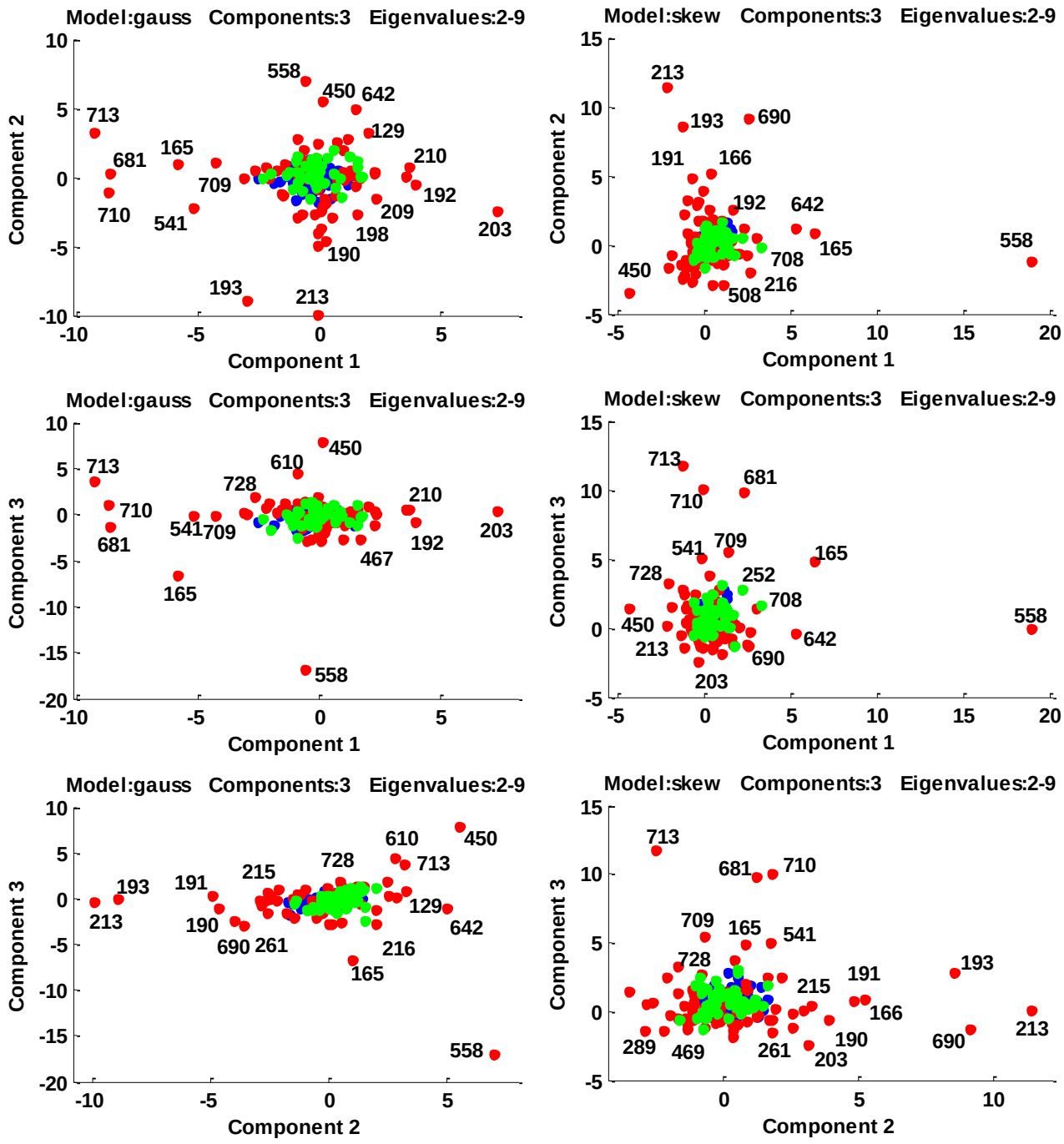


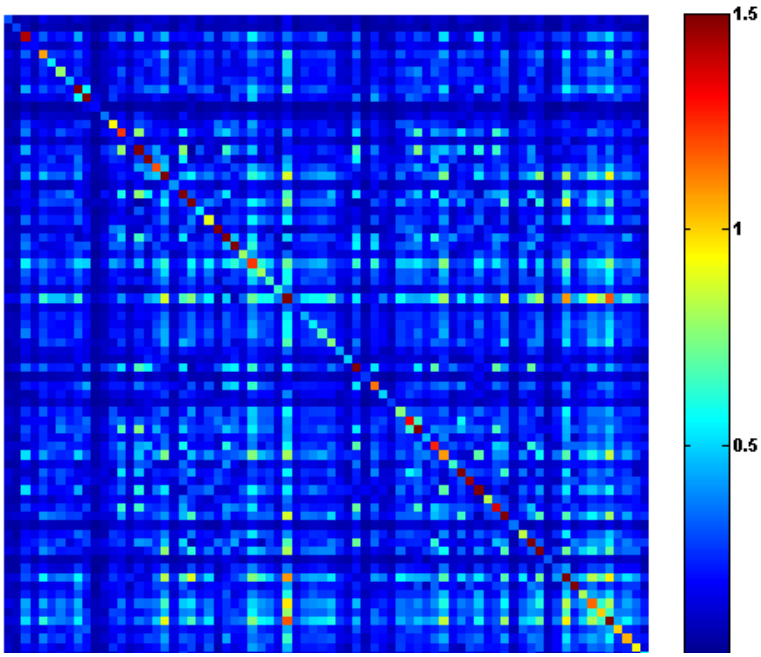
Figure 66: Independent three component analysis of the TonB dependent receptor family reveals no independent sets of positions. The parameters used for ICA are shown in each graph. In this example, different models are tried, while three components are searched for using the eigenvalues from 2 to 9. The colors refer to the following:
red: residues identified as coupled using the PCA analysis
green: residues have at least one coupling score greater than 0.5. Note that this includes all the red colored positions.
blue: the remaining alignment positions. Selected residues are marked using FecA structure numbering.



Coupling analysis of the signaling domain was also performed based on a structure based alignment of 296 signaling domains. The alignment was done separately from the barrel and plug domain and so is presented separately.

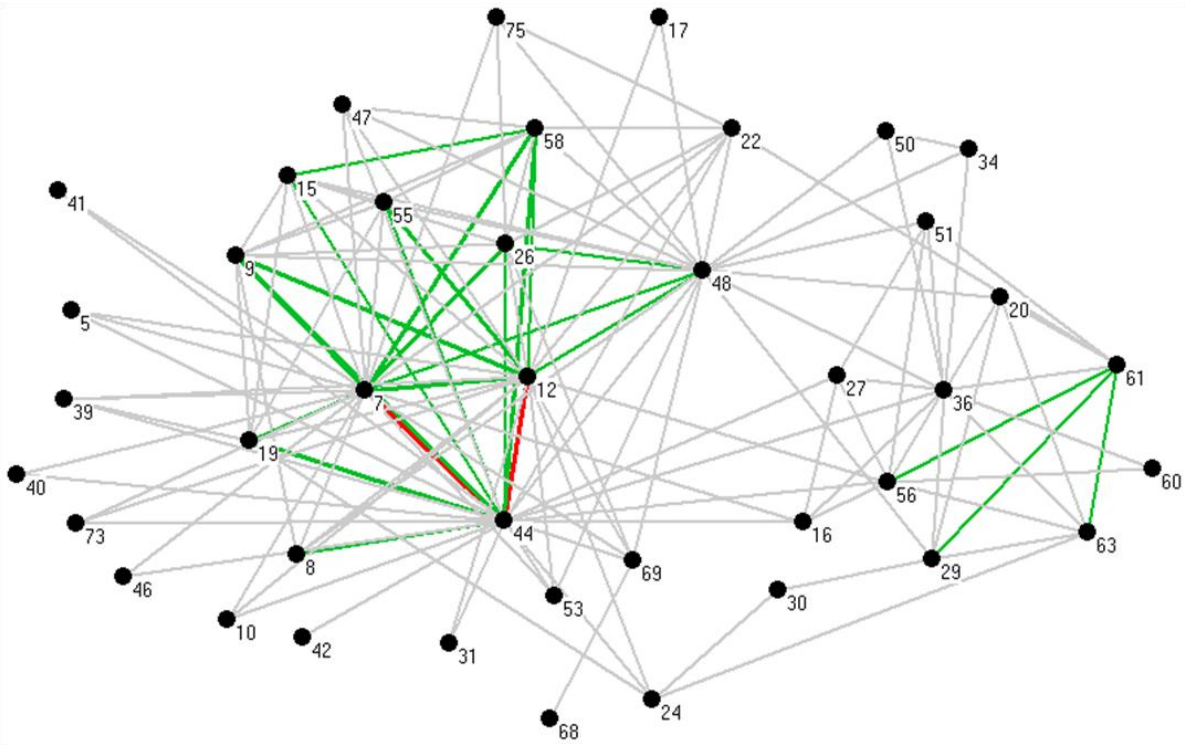
The coupling matrix of 74 aligned positions is shown in Figure 67. Clearly there are some coupled positions but this coupling matrix is not as 'blue' as the barrel and plug alignment. This coupling matrix will be analyzed with a network graph, hierarchical clustering, principal components analysis and independent component analysis as previously described.

Figure 67: Statistical coupling matrix of the signaling domain. Matrix values are colored according to the color scale.



The network graph shown in Figure 68 reveals a network of 41 coupled positions. This represents most of the coupling of this domain (made up of 80 amino acids or 74 positions in the alignment). The whole domain seems to be quite connected. There may be two clusters of positions where one set of

Figure 68: Network graph of the signaling domain of FecA. Red lines are coupling values greater than or equal to one, green lines are coupling values between 0.75 and 1 and grey lines are coupling values between 0.5 and 0.75. The numbers on the nodes are the signaling domain structure numbers.



residues forms a dense network of high value contacts with each other. The first cluster is made up of positions 7, 12 and 44 and the surrounding residues connected by green lines. The second cluster could be made of positions 29, 56, 61 and 63. Note though that the independence of these clusters is not absolute as position 12 of cluster 1 interacts weakly with position 56 of cluster 2.

After network analysis, hierarchical clustering is performed. The clustering was done with different combinations of distance and linkage methods as shown in Table 10. As was the case previously, the Chebychev distance with average linkage gave the best tree-data correlation score. The cityblock distance with complete linkage gave a score of 0.71 which is still pretty good. The tree diagrams are shown in Figure 69 and Figure 70 . One can see that the trees are actually very similar as the groupings are the same except they are ordered differently.

Table 10: Clustering of the signaling domain of the TonB dependent receptor. Centroid, median and ward calculations can only be done on Euclidean distances.

		linkage methods						
		'average'	'centroid'	'complete'	'median'	'single'	'ward'	'weighted'
distance methods	'euclidean'	0.85	0.82	0.80	0.82	0.77	0.68	0.79
	'seuclidean'	0.72	NaN	0.61	NaN	0.39	NaN	0.64
	'mahalanobis'	0.49	NaN	0.22	NaN	0.16	NaN	0.43
	'cityblock'	0.76	NaN	0.71	NaN	0.62	NaN	0.76
	'cosine'	0.89	NaN	0.68	NaN	0.87	NaN	0.83
	'correlation'	0.86	NaN	0.74	NaN	0.78	NaN	0.81
	'spearman'	0.71	NaN	0.62	NaN	0.58	NaN	0.69
	'chebychev'	0.97	NaN	0.94	NaN	0.94	NaN	0.95

Figure 69: Hierarchical clustering of signaling domain using chebychev distances and average linkage.

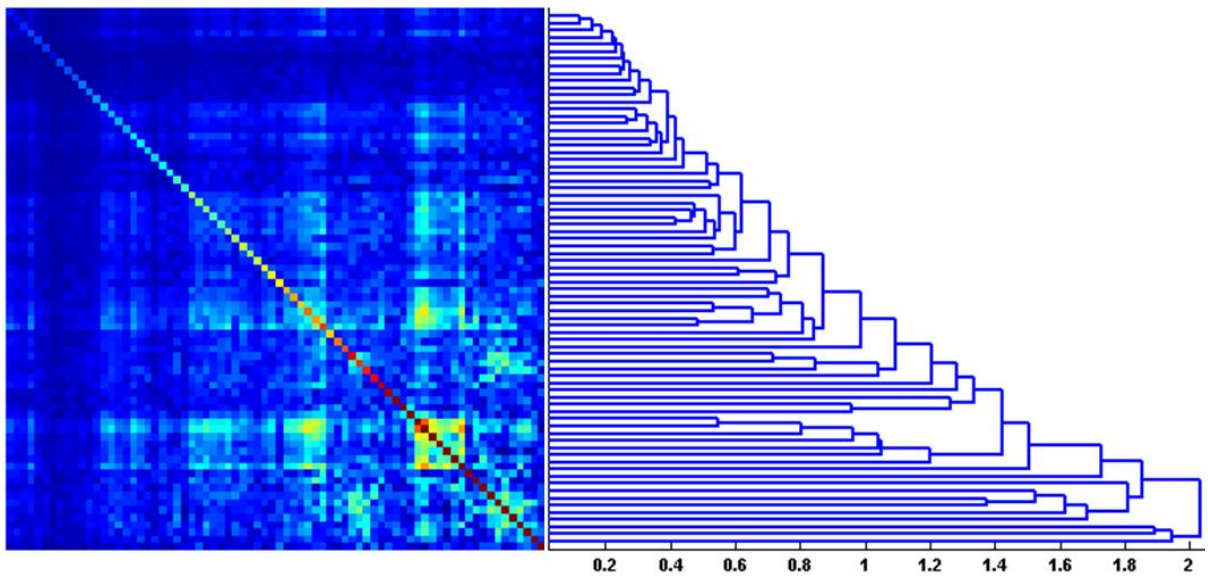
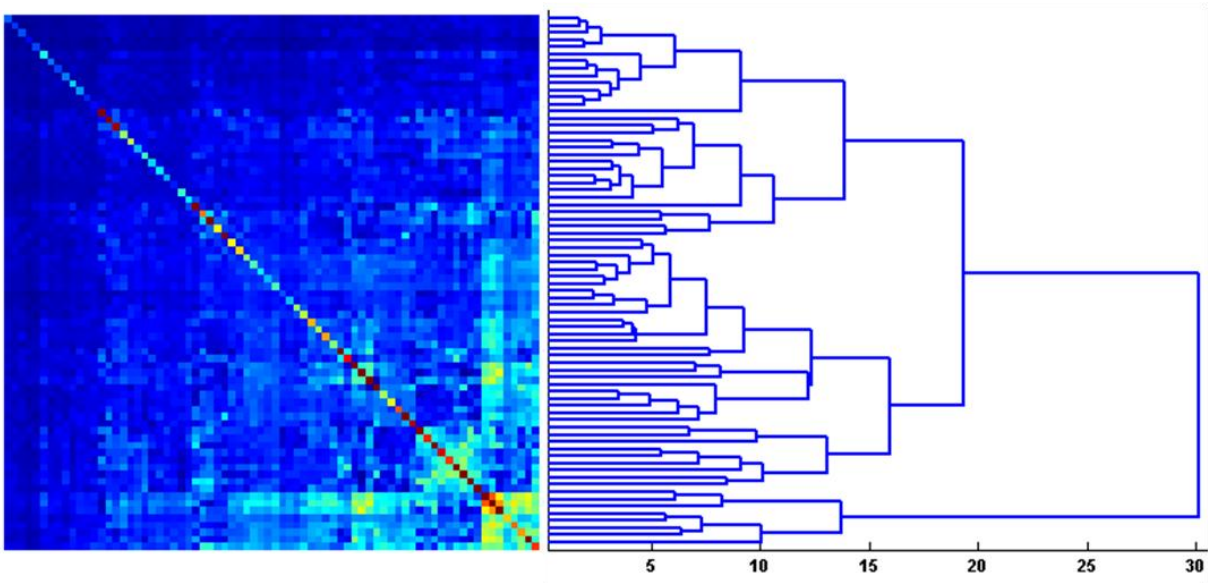
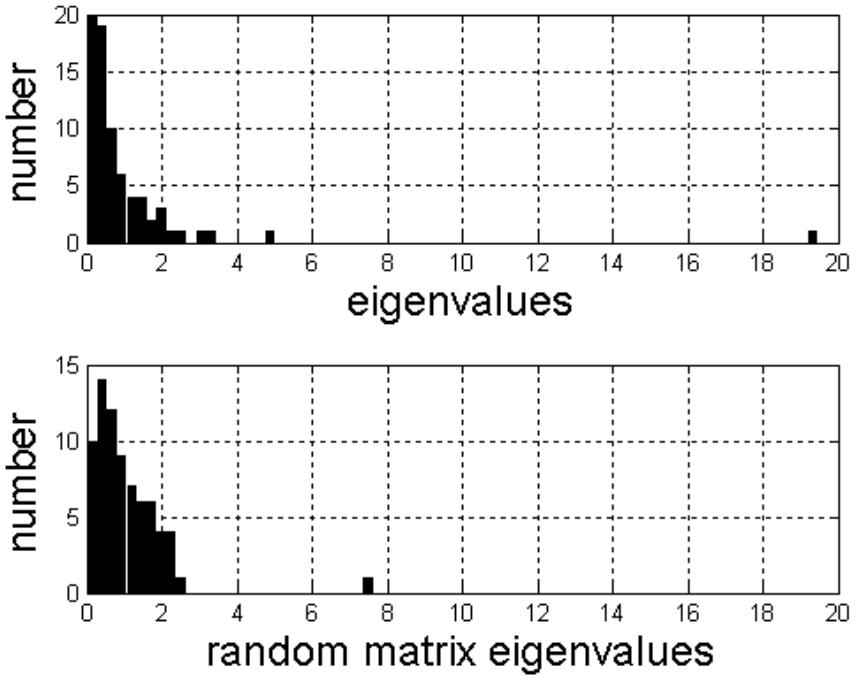


Figure 70: Hierarchical clustering of signaling domain using cityblock distances and complete linkage.



Next, principal components analysis is performed on the coupling matrix. A randomized alignment is made and the coupling matrix calculated for the random alignment. The eigenvalues of the original and random coupling matrices are plotted in a histogram form (Figure 71).

Figure 71: Eigenvalue histograms for PCA of TonB dependent receptor signaling domain.



Again, the first eigenvalue reflects the fact that the matrix is composed of positive values (Figure 72) and means that the first eigenvector can be discarded. If the first eigenvector of each matrix is discarded, the eigenvalue histograms reveal three eigenvalues with high values in the normal coupling matrix but not in the random alignment coupling matrix (Figure 73). The histograms of the weights of the significant eigenvectors are shown in Figure 74.

Figure 72: The first eigenvalue of the coupling matrix and the random coupling matrix are related to the magnitude of the coupling.

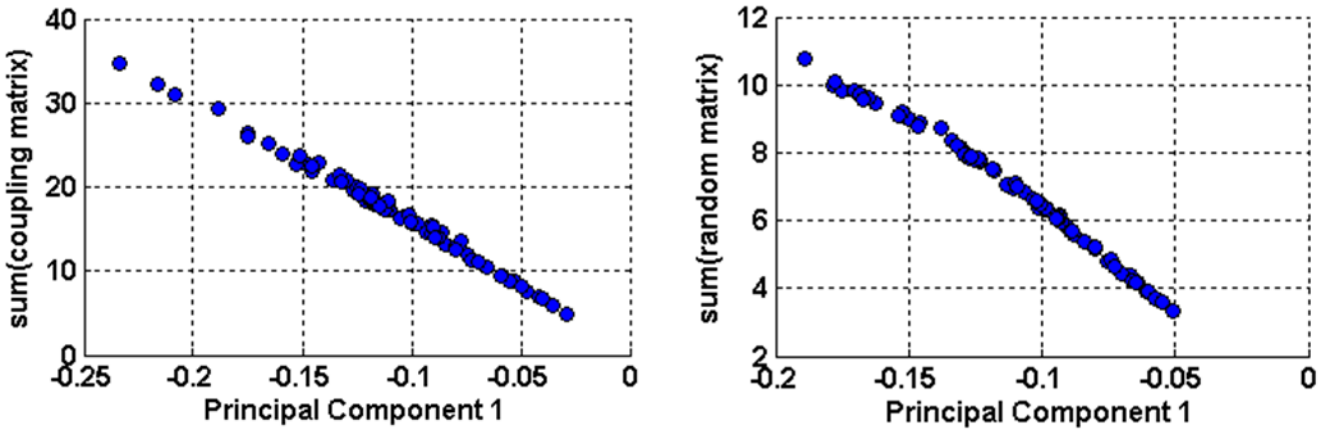


Figure 73: Eigenvalue histograms when the first eigenvalue is discarded for the PCA of signaling domain analysis

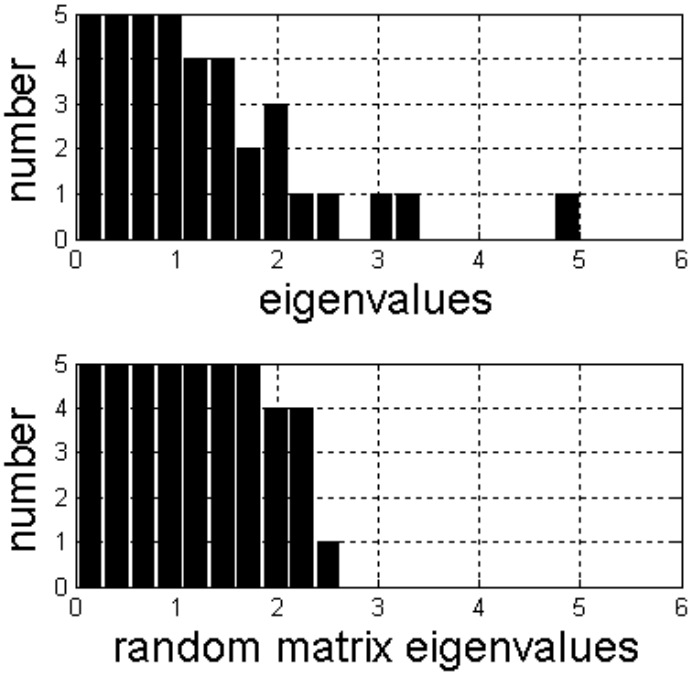
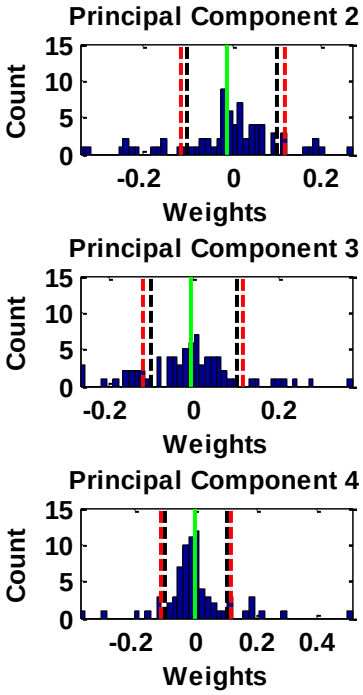


Figure 74: Histograms of the weights of the three significant eigenvectors. The green line is the mean; the red dashed line is one standard deviation and the black dashed line is at 0.1.



It is possible to visualize all three significant eigenvectors with either a 3D plot or three 2D plots. These three 2D plots are shown in Figure 75. The residues with weights higher than 0.1 are shown in Table 11. In the plots one can clearly see that the positions that could represent different networks in the network diagram are in different regions in the PCA plot. In the PC2 vs. PC3 graph one can see that positions 7 and 12 and 55 contribute along one axis differently than positions 60, 61 and 20.

Figure 75: Weights of principal component 2-4 plotted against each other for signaling domain of TonB dependent receptors. Dashed lines represent the 0.1 threshold.

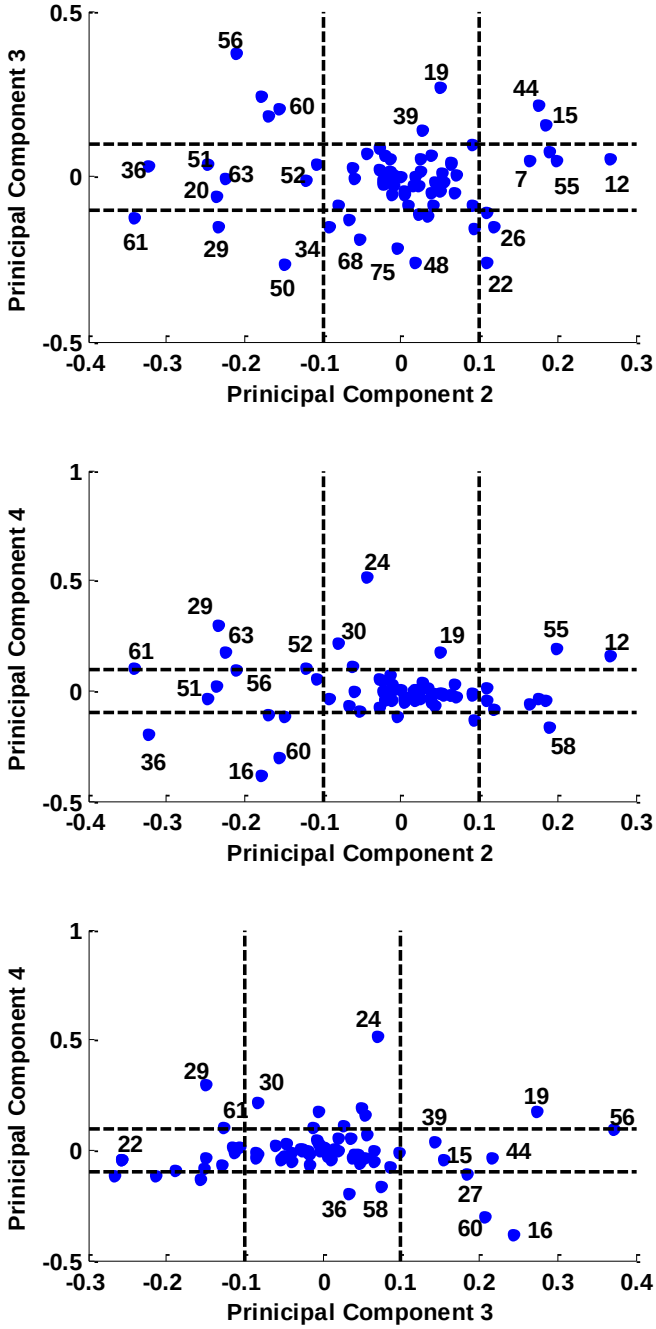
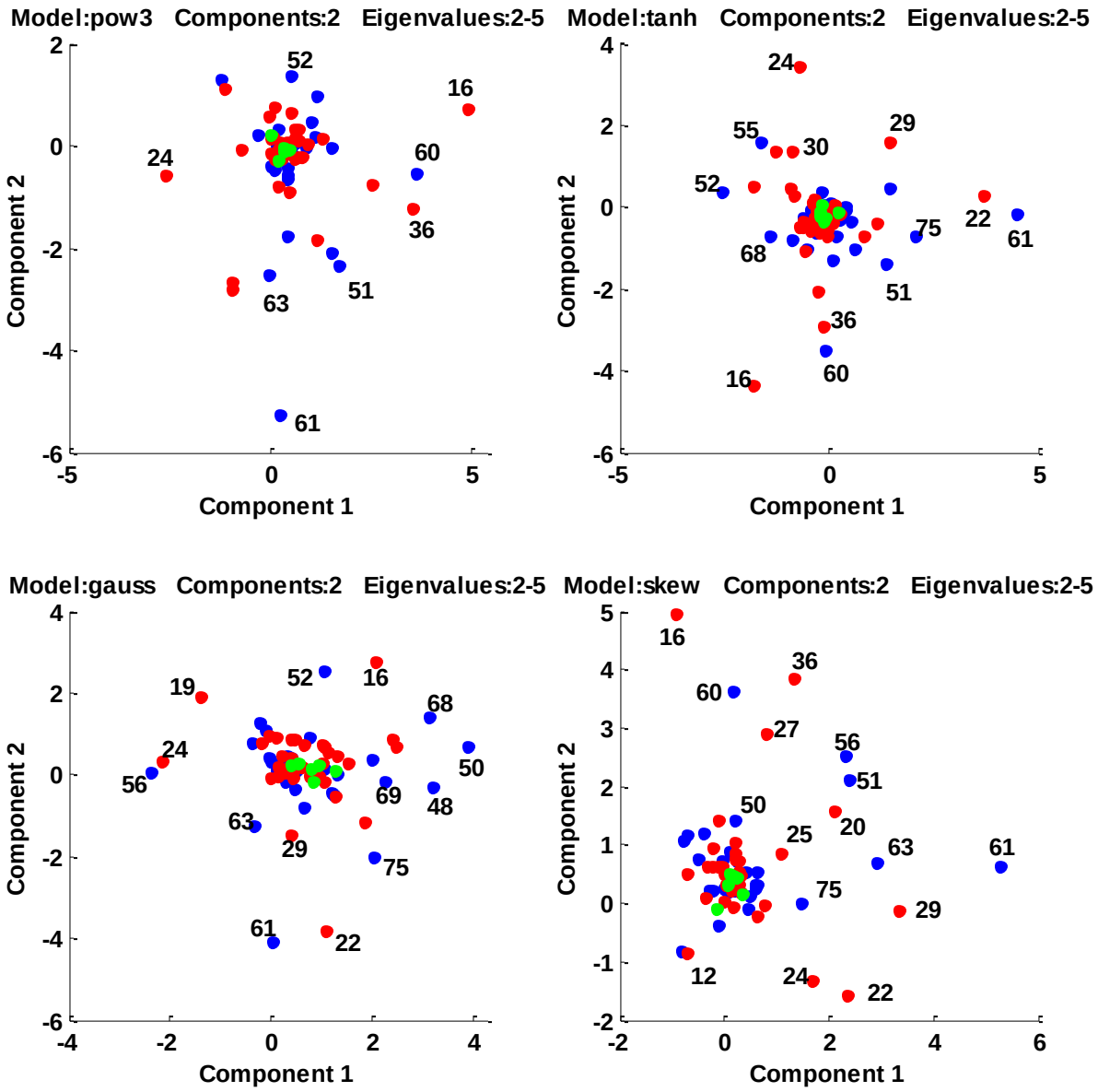


Table 11: Positions along the listed eigenvectors with weights higher than 0.1

	eigenvectors		
	2	3	4
Positions (FecA numbering)	7	15	12
	12	16	16
	15	17	19
	16	19	23
	20	22	24
	22	26	27
	25	27	29
	26	29	30
	27	34	36
	29	39	50
	36	44	52
	44	47	55
	47	48	58
	50	49	60
	51	50	61
	52	56	63
	55	60	69
	56	61	75
	58	64	
	60	68	
	61	69	
	63	75	

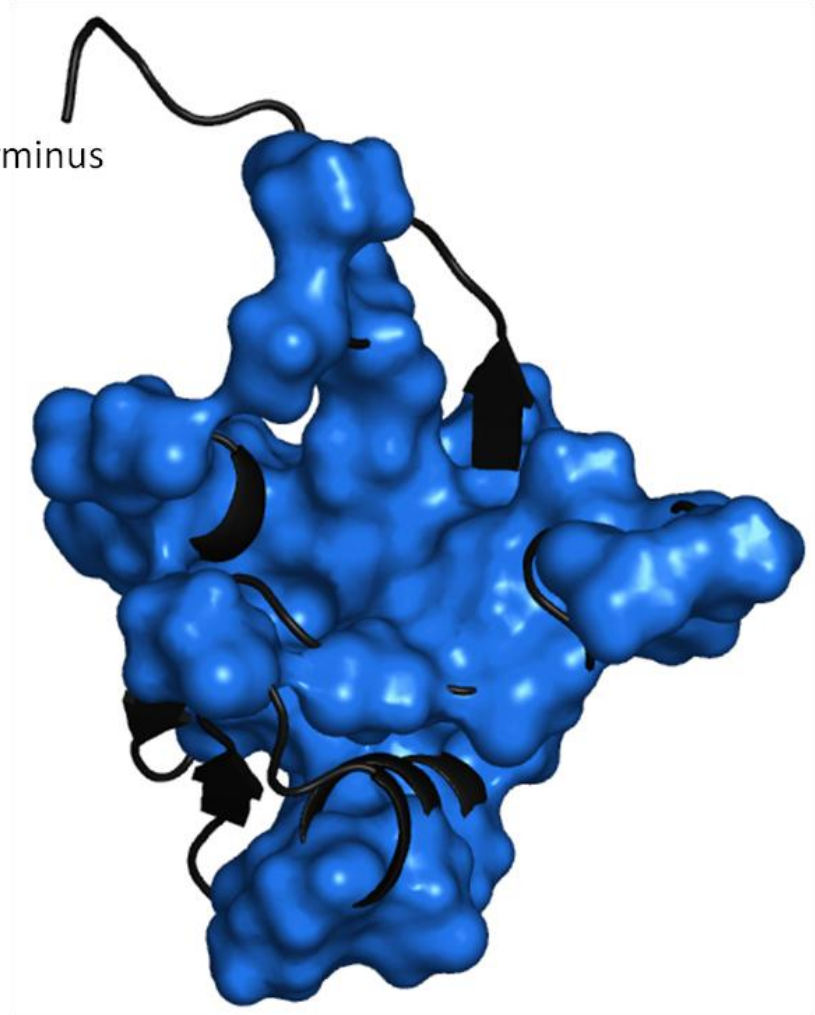
To see if these sets could be statistically independent, independent component analysis is done. Here I would like to determine whether there are two independent components to the data. As the next figure (Figure 76) shows, there are no sets of positions that clearly lie on two different axis indicating that these data are probably not composed of independent components.

Figure 76: Independent component analysis of the TonB dependent signaling domain. Different methods do not reveal any points that are clearly on different axes. The colors refer to the following:
red: residues identified as coupled using the PCA analysis
green: residues that have at least one coupling score greater than 0.5. Note that this set includes all the red colored positions.
blue: the remaining alignment positions. Selected residues are marked using FecA structure numbering.



Now that the analysis of the signaling domain is complete, the identified positions are mapped onto the structure as shown in Figure 77. As many positions (35/80) are significantly coupled, the connectedness of this structure is expected. However it is interesting that there is a coupling to the strand that is linked to residue 80 of the protein. Residue 80-85 makes up the TonB box which is the set of residues that contact TonB. Selected mutations were made in the positions of the signaling domain (as well as the rest of the protein) and assayed for function. These experimental tests are described next.

Figure 77: Coupled positions identified by PCA of the FecA signaling domain.



Section 3: Experimental analysis of the coupled positions

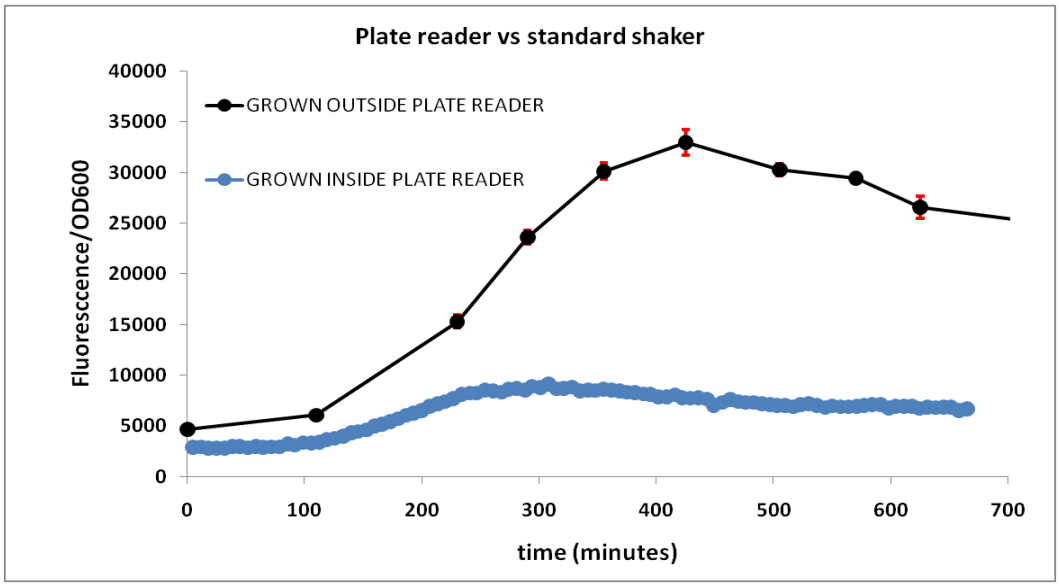
Residues identified by SCA are interpreted to be important in the proper functioning of the protein. Experimental tests are required to support this point of view. To test the effects of the coupled positions, several alanine mutants were studied of positions in the plug and barrel and signaling domain of FecA. These mutants were then tested for their ability to affect the function of the receptor.

The particular function tested in this assay was the proper signaling of the receptor. Remember that FecA has a signaling domain and the activation of signaling is dependent on the binding of ligand and the subsequent TonB activation of the receptor. On the other hand the transport function and signaling function are independent as the signaling domain can be experimentally removed resulting in no loss of transport function. This assay was used rather than a direct transport assay because the available transport assays are not very good due to the fact that there are a variety of transporters available and the difficulty of labeling the ferric dicitrate.

To observe whether signaling occurred, a GFP based assay was developed. Expression of GFP was under the control of a promoter that becomes activated via the signaling domain. Therefore, in principle the cells become fluorescent with increased signaling. The plan was to incubate cells harboring a GFP reporter plasmid with 1 mM sodium citrate and monitor over time the expression of GFP in whole cells.

Two conditions were tried to measure GFP fluorescence. First, I tried measuring GFP fluorescence while growing the cells in the plate reader. This method allows one to follow any change in GFP fluorescence over time as the cells grow. The second method was to grow the cells in standard culture tubes and then periodically remove aliquots and measure GFP fluorescence in a plate reader. This method, though more labor intensive, was chosen to collect all the data as the dynamic range is at least three fold higher than the first method (see Figure 78). This is probably because growing cells in a plate reader is not optimal. The data reported represent the fluorescence signal divided by the optical density at 600 nm (OD600) for each time point.

Figure 78: Comparison of GFP fluorescence between cells grown in plate reader and cells grown in a laboratory shaker. Error bars represent the standard deviation from three trials.



Several control experiments were conducted to verify that the assay can accurately and reproducibly measure signaling in these cells (see Figure 79). Inspection of the various panels reveals that cell growth is not inhibited by either a lack of citrate or absence of FecA.

To account for differences in growth, as previously stated, the fluorescence was divided by the OD600 at each time point. This results in a normalized graph as shown in Figure 80. Note that the profile is complex: there is a small lag phase, an exponential phase, a peak, and then a decrease in normalized fluorescence.

This experiment was then repeated for a set of mutants in the plug, barrel and signaling domain, with the normalized fluorescence data shown in Figure 81. This data for the mutations in the coupled network along with other control mutations have been reported in Ferguson et al. 2007 [55] and are shown in Table 12 taken from that paper. Mutations in the coupled positions affected the function of the protein in some cases completely disrupting the signaling activity. This mutational data supports the idea that the coupled network mediates the allosteric function of this protein.

Figure 79: Positive and negative control experiments for measuring signaling of the FecA receptor. Error bars are standard deviations of triplicate data. Growth is similar for all conditions, while the cells become fluorescent only with the presence of FecA and addition of citrate.

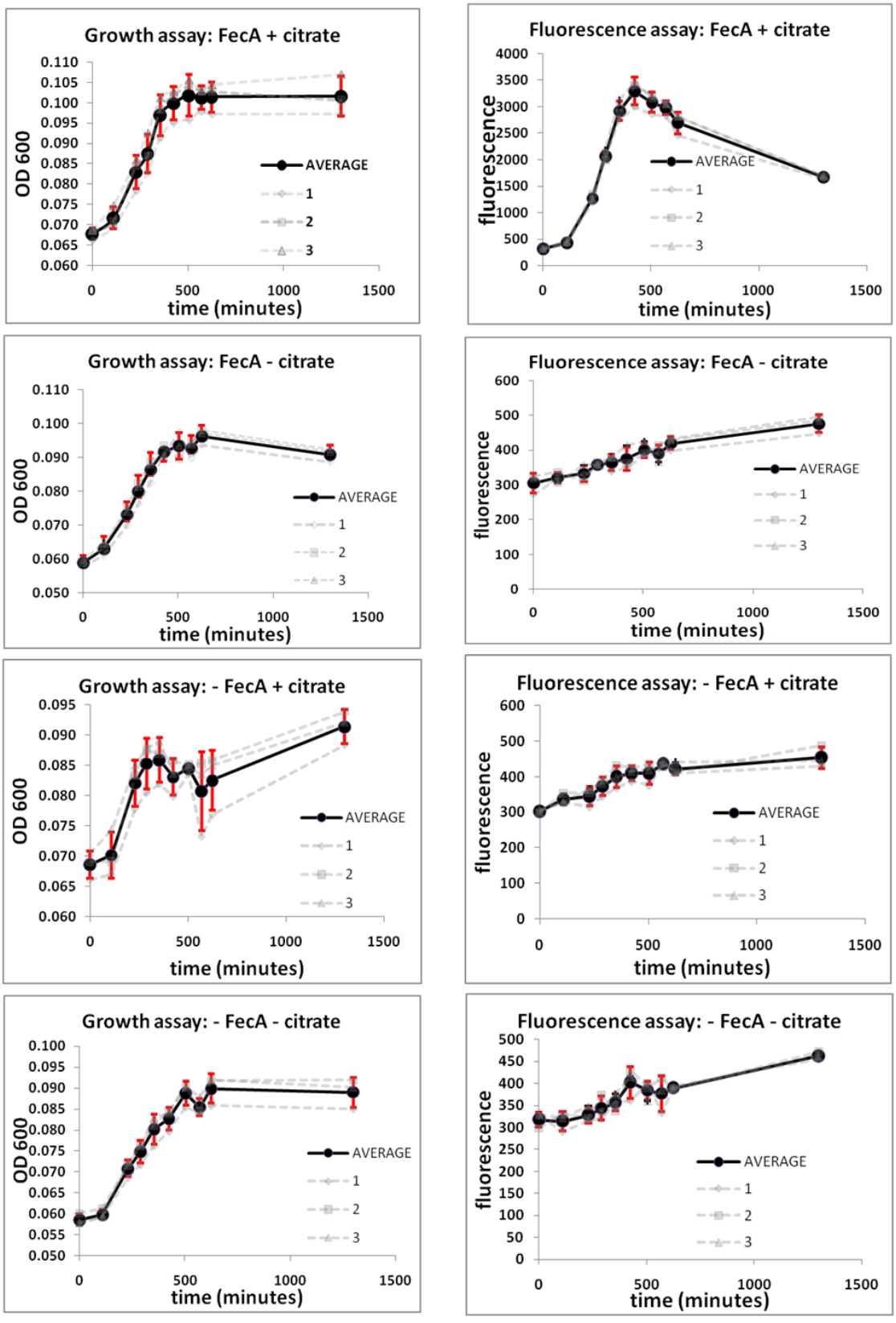


Figure 80: Fluorescence data normalized for cell growth. Fec refers to FecA, CIT refers to sodium citrate, the ligand of FecA, and + or – refers to the presence or absence of FecA or sodium citrate.

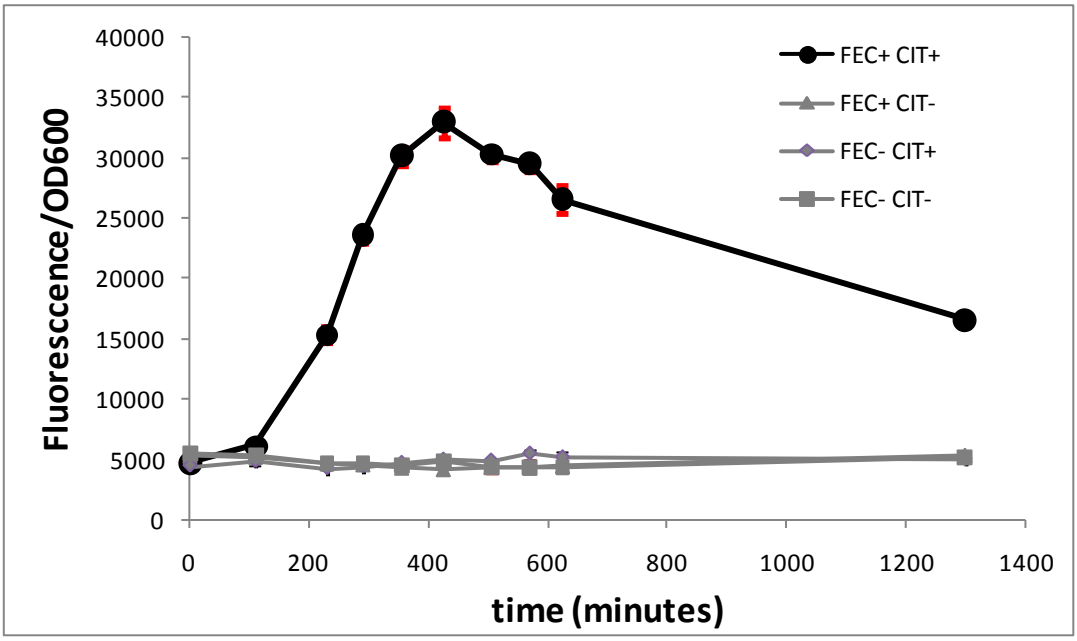


Figure 81: The effect of FecA mutants on signaling. The mutants were chosen based on having high coupling scores. Note that the error bars (+/- standard deviation) are present and within the marker for each data point. Note that D45 in the signaling domain seems to be much slower at activating transcription.

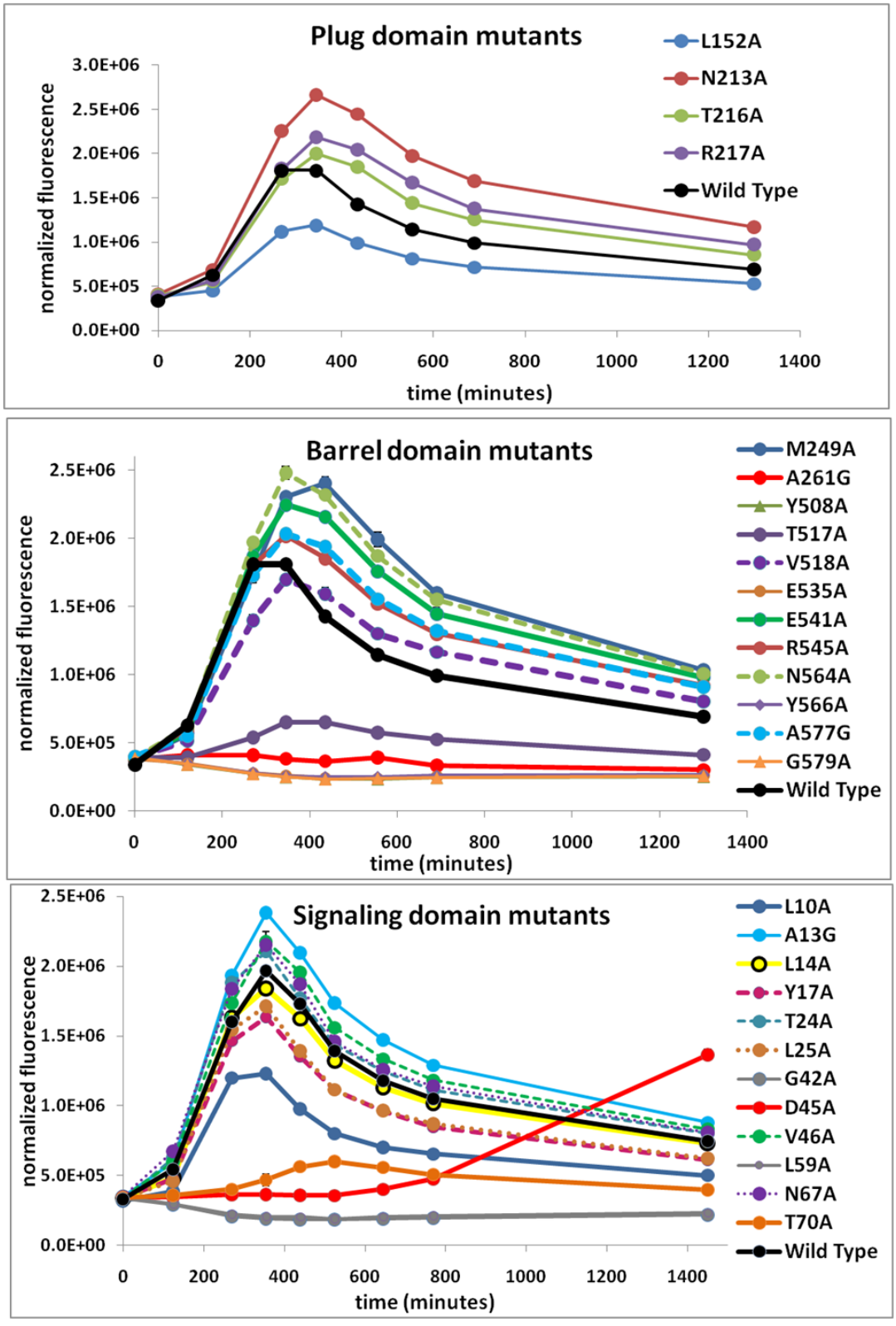


Table 12: GFP fluorescence assay as described in the text. The data is presented normalized to the expression of wild-type FecA (which is 1). The range of wild-type is 1 +/- 0.2 so any effect above 1.2 or below 0.8 is considered significant. FecIR is the negative control containing only the GFP plasmid while FecIRA is the positive control with wild type FecA. The residues in bold are mutations in the coupled network.

Mutation	Position	Fluorescence
FecIR	—	0
FecIRA	—	1
L10A	Signaling, α1	0.71
A13G	Signaling, α1	1.24
L14A	Signaling, α1	0.95
Y17A	Signaling, α1	0.85
T24A	Signaling, β2	1.06
L25A	Signaling, β2	0.92
G42A	Signaling, β3	0
D43A	Signaling, β3	1.09
D45A	Signaling, L5	0
V46A	Signaling, α3	1.13
L59A	Signaling,β4	0
N67A	Signaling, L7	1.13
T70A	Signaling, β5	0.06
Δ28–30	Signaling, α2	0
Δ80–86	TonB-box	0
T138A	Binding site	0.15
R365A	Binding site	0
R380A	Binding site	0
R438A	Binding site	0
Q570A	Binding site	0
L152A	Plug	0.46
N213A	Plug	1.35
T216A	Plug	1
R217A	Plug	1.14
M249A	Barrel, β2	1.29
A261G	Barrel, β3	0.08
Y508A	Barrel, β13	0
T517A	Barrel, β13	0.18
V518A	Barrel, L7	0.80
E535A	Barrel, β14	0
E541A	Barrel, β14	0.87
R545A	Barrel, β14	0.96
N564A	Barrel, β15	1.26
Y566A	Barrel, L8	0
A577G	Barrel, L9	1.02
G579A	Barrel, β16	0
E587A	Barrel, β16	0
Δ519–533	Barrel, L7	0
Δ568–577	Barrel, L8	0

Conclusion

This chapter discussed the identification of an allosteric pathway in the TonB dependent receptor family. Constructing the MSA for this protein family and the subsequent SCA were the most technically challenging parts of this study. The experimental assays and analysis were relatively straightforward and facilitated by Andrew Ferguson who was responsible for this project.

The results are quite interesting. Statistical coupling analysis was able to identify a set of positions encompassing the binding pocket in the extracellular space all the way through the protein to the interaction site where the periplasmic protein TonB binds. In addition, the signaling domain also contained a network of coupled residues. Experiments of point mutants of these coupled positions validated the functional significance of these positions as many affected the signaling activity.

Furthermore, after this work was complete, two groups [122, 123] published the structure of a truncated form of TonB with the TonB dependent receptor revealing the binding site where TonB binds to the receptor. This binding site position was identified by SCA as being coupled although the functional significance of the coupling was not apparent at the time.

The identification of a coupled network of positions highlights the fascinating energetic nature of proteins and how much there is yet to learn about the detailed interactions within proteins. The answers to the following would make for challenging projects:

1. Physical mechanism of allostery from the ligand site to the periplasm
2. Conformation of the protein in the open state
3. Mechanism for TonB energizing the receptor
4. Mechanism for signaling domain binding to the receptor
5. Energetics of releasing the iron which is closely associated with the siderophore
6. Examining the coupling patterns within other homologous proteins
7. The actual transportation mechanism

The network of positions identified with SCA could serve as a starting point to uncovering the complex cooperative interactions in the TonB dependent receptors.

Chapter 5: Serine Protease Family

This chapter is divided into three sections. The first section is a brief review of the chymotrypsin class serine protease family (family S1A). The second section describes the statistical coupling analysis (SCA) of the serine protease family. The third section describes experimental work done based on predictions of SCA and further evolutionary analysis.

Section 1: Review of the S1A family

The chymotrypsin class serine proteases (Family S1A according to Merops classification [131]) are a large and ancient family [131] of hydrolytic enzymes which are characterized by a catalytic triad consisting of serine, aspartate and histidine [132]. This family has been the subject of much research as different enzymes with very similar tertiary structures hydrolyze different substrates while utilizing the same catalytic machinery [132]. The basis of this substrate specificity is not well understood and attempts to switch the specificity from one enzyme to another has either required many mutations or resulted in enzymes with slow rates [133-136]. Allosteric regulation has also been observed in certain types of these proteases [137-139]. These proteases also bind calcium ions [132], are expressed as inactive zymogens and are stored in acidic endosomes until secreted into the extracellular space [140, 141]. The proteases of this family are involved in digestion of food, coagulation cascades, reproduction and immunity [142]. Their functional diversity and physiological importance makes them an interesting system to study.

Families of proteases and family S1A

Proteases are enzymes that hydrolyze other proteins. They comprise an extremely diverse and numerous set of proteins [143]. There are also at least six possible mechanisms by which proteases hydrolyze peptides giving rise to six functional classifications of proteases based on the nature of the catalytic residue: aspartic, cysteine, glutamic, metallo, serine and threonine [143]. These are not evolutionarily based classes however, because, for example, the serine proteases contain two non-homologous families (subtilisin family and chymotrypsin family that both use the exact same catalytic mechanism (constituting an example of convergent evolution) [[144, 145]].

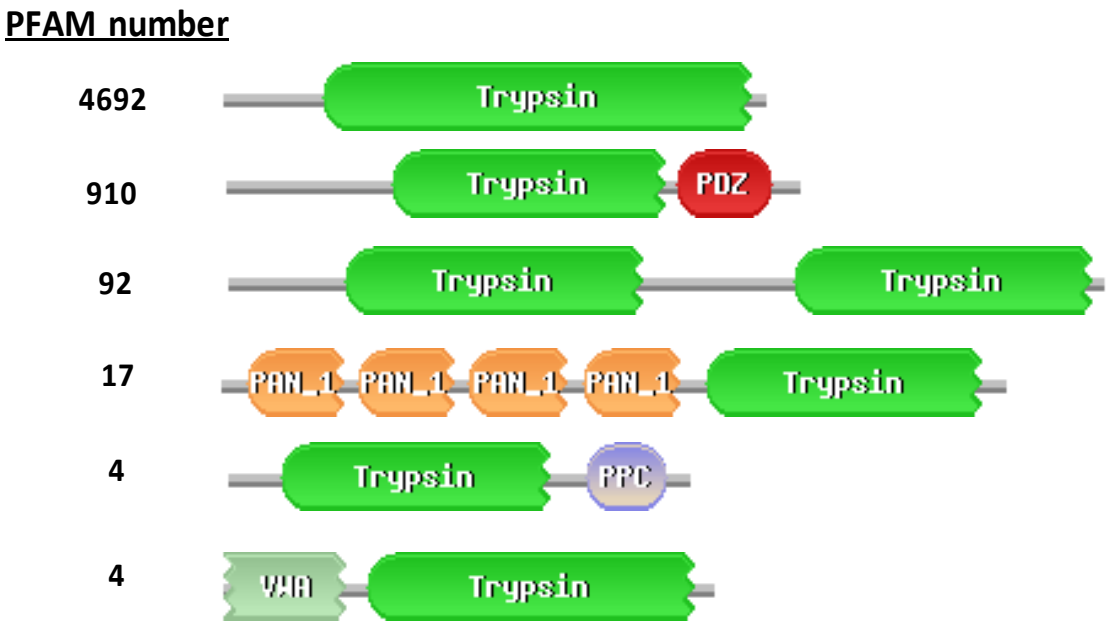
The family that is relevant to this project is the chymotrypsin class serine protease family also known as family S1A. This family has representatives in all eukaryotic species. The structurally related family S1B is also present in prokaryotes; however the level of sequence homology between families S1A and S1B is very low and in fact, running sequence similarity searches with family S1A does not result in identification of any sequences from family S1B). Some examples of widely known S1A proteases are

trypsin, chymotrypsin, elastase, thrombin and enterokinase (a protease-misnamed as a kinase). Note that these names do not refer to a single protein sequence and that each could have multiple isoforms.

While the protease function defines the vast majority of homologs of this family, some family members have lost the catalytic active site residues and carry out binding functions. One example of this is haptoglobin which has evolved to bind heme [146, 147]. Other examples will be discussed later.

Most commonly, family S1A members are part of one domain. However, some family members exist as multi-domain proteins as shown in Figure 82. Furthermore physiologically important members such as some coagulation and complement proteases are part of multidomain or membrane bound proteins.

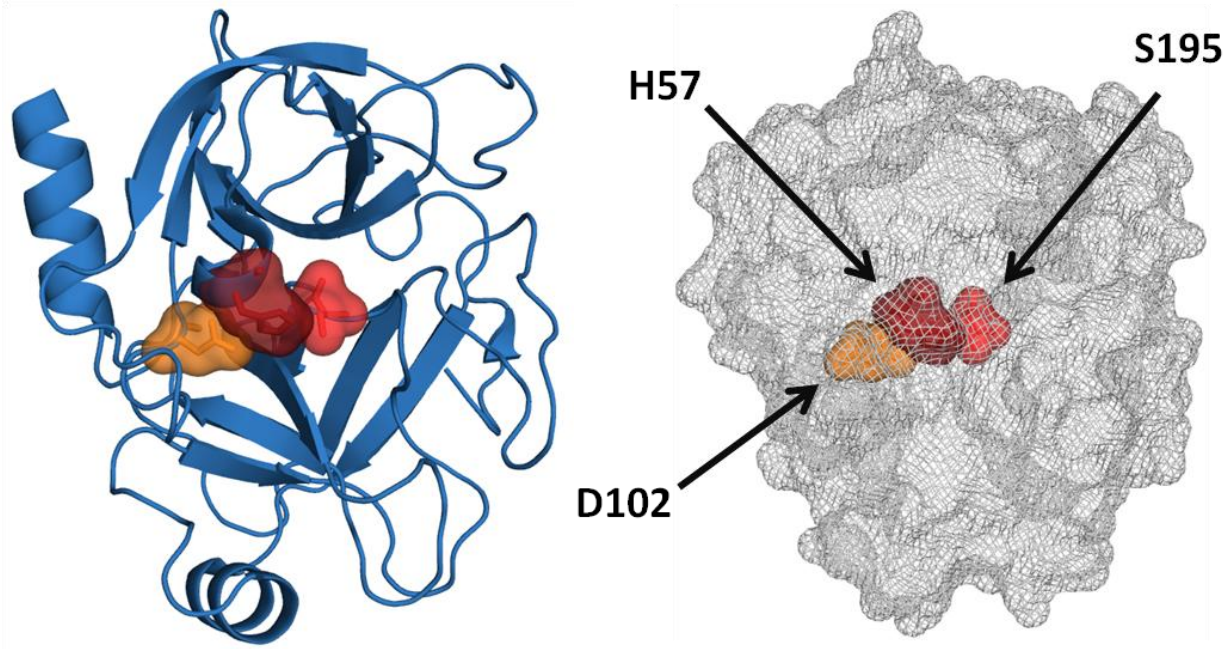
Figure 82: Examples of trypsins with different domain architectures. PDZ: a c-terminal peptide/protein binding domain. Pan-1: A protein or carbohydrate binding domain. PPC: domain of unknown function. VWA: domain that binds to von Willebrand factor. The PFAM number is the number of sequences in the PFAM database (Dec. 2008) that contains the displayed domain structure.



Proteolysis mechanism

The mechanism of proteolysis of trypsin involves three important residues (serine, histidine and aspartate) at the catalytic core. Studies have elucidated the general mechanism of trypsin proteolysis; however, additional studies continue with the goal of determining the molecular details. The numbering of these amino acids is usually based on their occurrence in the sequence of chymotrypsin and so I will refer to the catalytic residues as S195, H57 and D102. Figure 83 shows the catalytic triad on the structure of rat trypsin.

Figure 83: Catalytic triad of serine proteases mapped on the structure of rat trypsin (pdb ID 2AGE).



The current model of the chemical mechanism is described below:

Histidine 57 is thought to abstract a proton from the gamma oxygen of serine 195 resulting in the creation of a positive charge on histidine and a negatively charged oxygen on the serine. The positively charged H57 is stabilized by D102. The negatively charged oxygen atom acts as a nucleophile and attacks the carbonyl carbon of the peptide bond. This results in a carbon with four bonds on it in a tetrahedral form (known as the tetrahedral intermediate). The tetrahedral intermediate is stabilized via interactions with residues such as S214 and G193 at a site known as the oxyanion hole. The tetrahedral intermediate, assisted by the positive charge of histidine 57, rapidly breaks down to give the first product which is the first peptide. The enzyme is still covalently attached to the second peptide. This intermediate with bound peptide on it is called the acylenzyme intermediate. The second half of the reaction is known as the deacylation reaction. Water, thought to be activated via H57 now serves as

the nucleophile to attack the ester bond releasing the second product and regenerating the enzyme ([132] and reviews within).

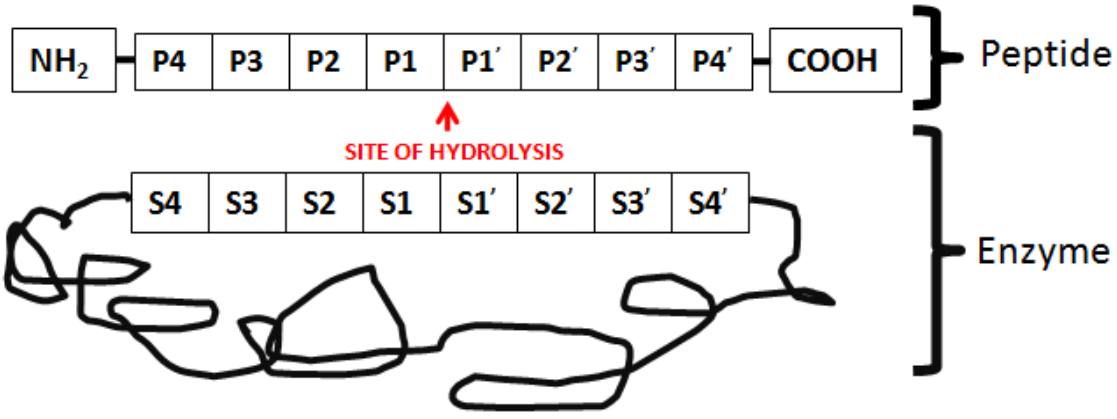
While this mechanism is generally accepted additional details have continued to emerge. For example, a 2006 paper from Daniel Koshland's group [148] captured the structure of real substrates in the active site of rat trypsin by quickly freezing the crystal after immersing it in substrate. The structures they obtained were the first of native substrate and native enzyme. The enzyme had reacted with the substrate forming an acyl intermediate. The structures were of high resolutions (1.15 angstroms) and the authors obtained strong evidence that one particular water resolvable in the structures is the nucleophile in the deacylation reaction.

Although understanding the chemical mechanism is important to uncover fully especially given that the rate acceleration is of the order of 10^{10}, understanding the substrate specificity is the primary goal within the field.

Substrate Specificity

An important feature of family S1A is that while the same catalytic mechanism is used, different enzymes bind and hydrolyze different substrates. For example the trypsin protease cleaves most efficiently after lysine or arginine residues whereas chymotrypsin cleaves most efficiently after phenylalanine. This selectively translates into a k_{cat}/K_m difference of 100,000 for peptide substrates [136, 149]. The molecular determinants of this specificity have been the subject of much investigation [133, 135, 136, 149-153]. The nomenclature used historically to describe the determinants of specificity [154] is based on assuming that a site on the enzyme (S) binds a particular site on the peptide (P). The numbering of sites begins with one at the site of cleavage. So site P1 of the peptide is bound by site S1 of the enzyme. Sites on the peptide's N-terminus are incremented, while sites C-terminal are incremented with a P'. This is shown schematically in Figure 84. This scheme has to be considered

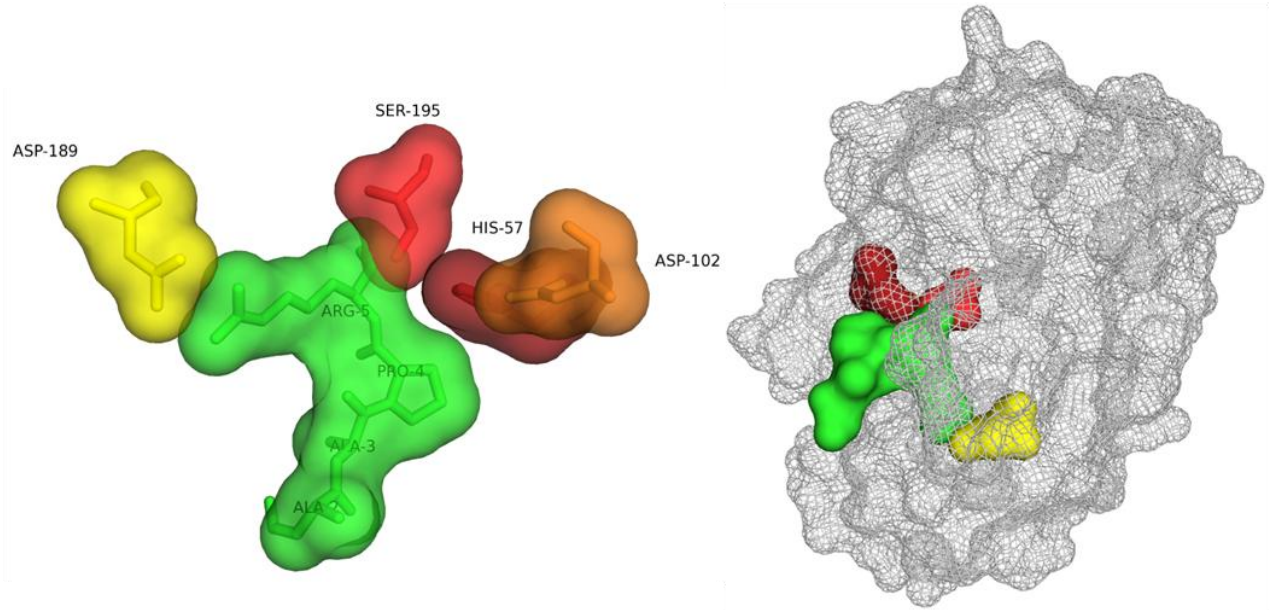
Figure 84: S and P nomenclature of proteases.



approximate as multiple sites on the enzyme could bind the same peptide site and multiple sites on the peptide could bind to the same site on the enzyme when interactions are considered on a residue level and possibly at atomic level too (as atoms can form multiple interactions). Nonetheless this nomenclature is used extensively in the literature.

When the first structures of trypsin and chymotrypsin were solved it was thought that specificity could arise due to residue 189 which formed the S1 binding site for residue P1 of a substrate; the side chains of lysine/arginine for trypsin substrates and phenylalanine or tyrosine for chymotrypsin substrates interact with this residue (see Figure 85). Most importantly, residue 189 is different in proteases with different specificities. In trypsin residue 189 is aspartate and in chymotrypsin residue

Figure 85: The association of a peptide substrate (Suc-Ala-Ala-Pro-Arg-PNA), shown in green, with residues from rat trypsin, D189. The other residues present are the active site of rat trypsin. This is based on structure 2AGE. The mesh diagram shows how part of the substrate is embedded within the active site pocket (also known as the S1 pocket) of rat trypsin.



189 is serine. The chemical rationale behind the specificity difference is that in trypsin aspartate binds the positively charged lysine and arginine whereas in chymotrypsin the serine with hydrophobic character can accommodate the large phenylalanine or tyrosine side chains. This view however proved to be simplistic as the next experiment shows.

The logical experiment to test this specificity model was to mutate residue 189 to either aspartate in chymotrypsin or serine in trypsin and measure catalytic rates to determine if specificity changed. These studies was done in 1988 in William Rutter's group [151] and they found that the D189S mutant in trypsin killed both the tryptic and chymotryptic activities (the tryptic activity k_{cat}/K_m was reduced 57800 fold and the chymotryptic k_{cat}/K_m was increased 8 fold but this 8 fold increase was still low as the

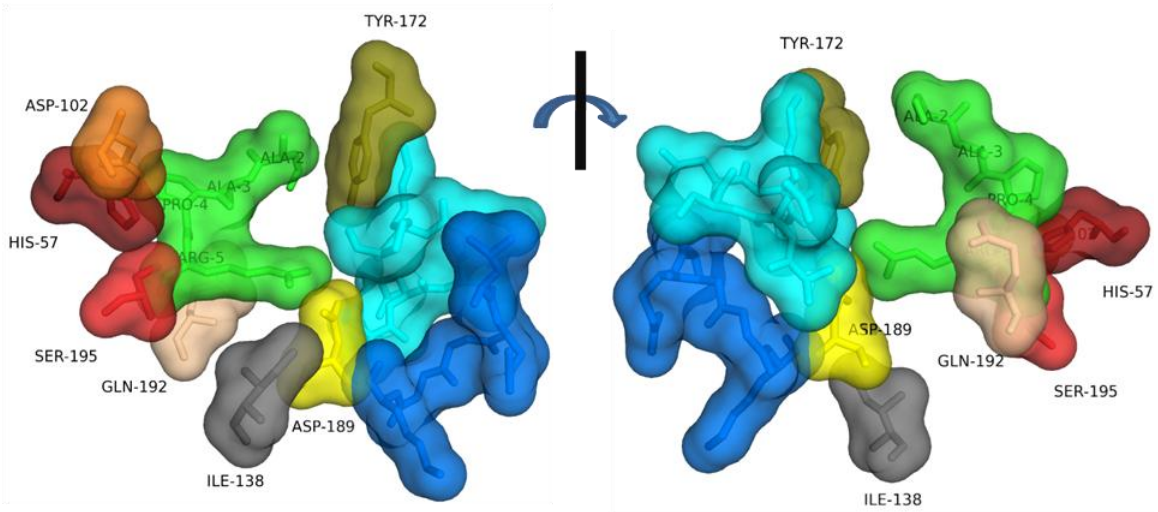
tryptic substrate with native enzyme was still 21000 fold less in k_{cat}/K_m than the chymotryptic substrate with S189). As it turns out to transfer chymotryptic specificity to trypsin requires many more mutations. This work was done by Lizbeth Hedstrom again in William Rutter's lab reported in 1992 [136] and 1994 [135, 149]. To get a trypsin swap that is within one percent of the activity of chymotrypsin on a PHE-substrate it is necessary to mutate what is known as the S1 pocket (D189S, Q192M,I138T, +T219) and two surface loops termed Loop 1 or the 80's loop(residues 185-188) and Loop 2 or the 20's loop (residues 221-225). To get within 15% of the activity of chymotrypsin required the addition of one more mutation, Y172W. The result is that it is necessary to make 14 mutations to switch the specificity of trypsin to chymotrypsin (up to 15%). The following Figure 86 shows these positions.

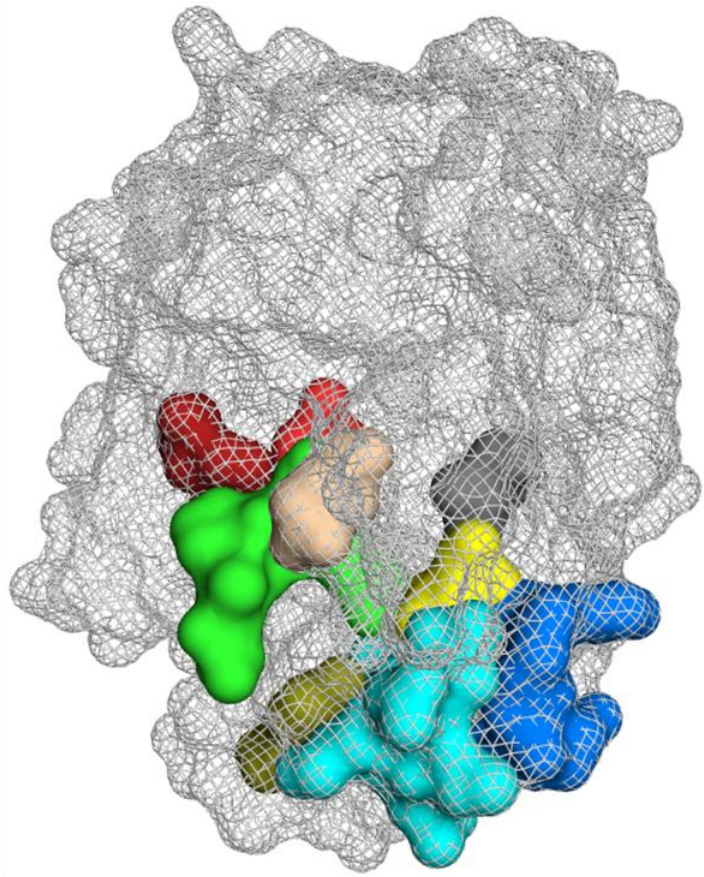
This work showed that switching the specificity of a serine protease is difficult to do and the mutations required not obvious from the structure. Furthermore equivalent attempts to change chymotrypsin to trypsin [134, 150] or trypsin to elastase [133] were much less successful than that of the trypsin to chymotrypsin switch.

One question that was addressed by Lizbeth Hedstrom and colleagues is what chemical step was improved in the mutant trypsin switch. They addressed whether it was binding, acylation or deacylation. The data published in [135, 149] showed that it was improvement in the acylation step that most contributed to the switch with the Y172W mutation. Furthermore, there is biochemical [149] and structural [155] evidence that the efficiency of good trypsin-chymotrypsin proteases is due to increased stabilization of the S1 pocket . These studies have helped elucidate the cooperative nature of substrate binding in this family.

Though trypsin and chymotrypsin have been the most extensively studied serine proteases, other family members with greater specificity to substrates have also been examined [156-158] based on crystal structures and mutations. In general the more specific the substrate the more sites are used for binding to the protein.

Figure 86: Positions of rat trypsin that had to be mutated to the equivalent ones (based on alignments) to switch trypsin to chymotrypsin. The structure used was 2AGE. The green positions are from the covalently bound substrate (Suc-Ala-Ala-Pro-Arg). while all other positions are from rat trypsin.





Proteases as Zymogens

Another feature common to proteases is that they are first expressed as inactive zymogens and stored in the cell endosomes at low pH and then are activated by proteases outside the cell. Proteolytic cleavage of an N-terminal region is necessary to convert the zymogen to an active protein. The fragment removed is small; in the case of trypsin it is 18 residues in length whereas in chymotrypsin the cleaved fragment is 15 residues. The zymogen is generally denoted by the addition of the suffix –ogen to the protease. For example, the zymogen form of trypsin is called trypsinogen. The zymogen form of chymotrypsin is called chymotrypsinogen.

Structurally, cleavage of the N-terminal pro-form of the protease results in changes that include an ordering of the S1 pocket and the oxyanion hole. It is thought that interaction of residue 16 (based on chymotrypsin numbering) to D194 (adjacent to the catalytic S195) is responsible for turning the protease into the active form [132].

Allosteric effects in Proteases

There are two documented types of allosteric control in this family of proteases. Both involve the modulation of function via ions: one of the ions is sodium and the other is calcium.

Sodium binding to the protease thrombin is important for its regulation. Thrombin is the final protease in the coagulation cascade. Binding of sodium to thrombin has been shown to increase the rate of proteolysis by thirty fold and to increase the range of substrates hydrolyzed by thrombin [137, 159, 160]. Some of the molecular determinants of this have been identified [138, 161-163] and recently the allosteric control has been transplanted into trypsin via 19 mutations [164].

Calcium binds to all known proteases. It has been shown that calcium stabilizes the trypsinogen form of the protease preventing autoactivation of the protease [165]. Preventing autoactivation is physiologically very important as life threatening conditions such as pancreatitis (where the pancreas digests itself because the proteases are prematurely activated) would occur if proteases become activated at the wrong time or location [140, 141].

Section 2: Statistical coupling analysis of Family S1A

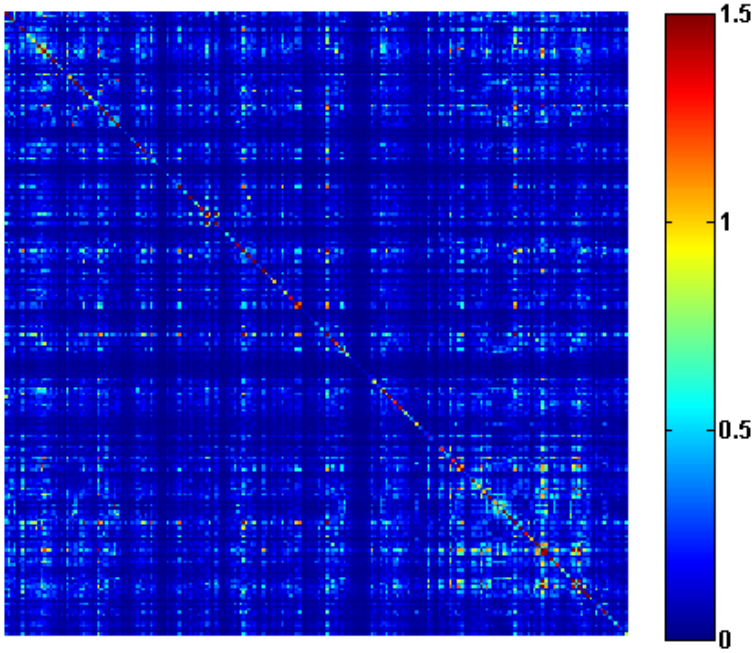
The important and complex functions of this protein family made it interesting to study via statistical coupling analysis (SCA). The starting point of my analysis is to create an alignment of this family. The alignment consists of 1470 sequences of family S1A members. The alignment includes both vertebrate and invertebrates sequences as well as a few bacterial and fungal enzymes. Psi-blast did not reveal any family S1B members. All sequences were chosen so that they include only one domain proteins (all membrane bound or multi-domain proteins were excluded from the alignment as their function could be different).

The alignment was made by using profiles of 80 structures of this family with Cn3D [127]. Then the alignment was extensively adjusted manually. Special consideration was given to aligning the loops which are not structurally conserved regions.

The diversity and quality of this alignment was significantly greater than the previous alignment constructed in the lab: it included many more invertebrate sequences and contained inactive proteins such as haptoglobins that are homologous to the active proteases. This increased diversity was useful as it enabled the detection of various sets of residues as will be shown later.

The first step to perform SCA is to calculate the coupling matrix as shown in Figure 87. The matrix reveals that this family has a sparse set of highly coupled positions as is typical of most protein families.

Figure 87: Coupling matrix of serine protease family



Next I will show the analysis via network graphs, hierarchical clustering, PCA and ICA.

The network graphs for all nodes with coupling values greater than one is shown in Figure 88 and for values greater than 0.75 is shown in Figure 89.

Figure 88: Network graph of serine protease S1A family. All nodes with coupling values 1 or greater are shown. The node numbers are based on the structure of 3TGI.

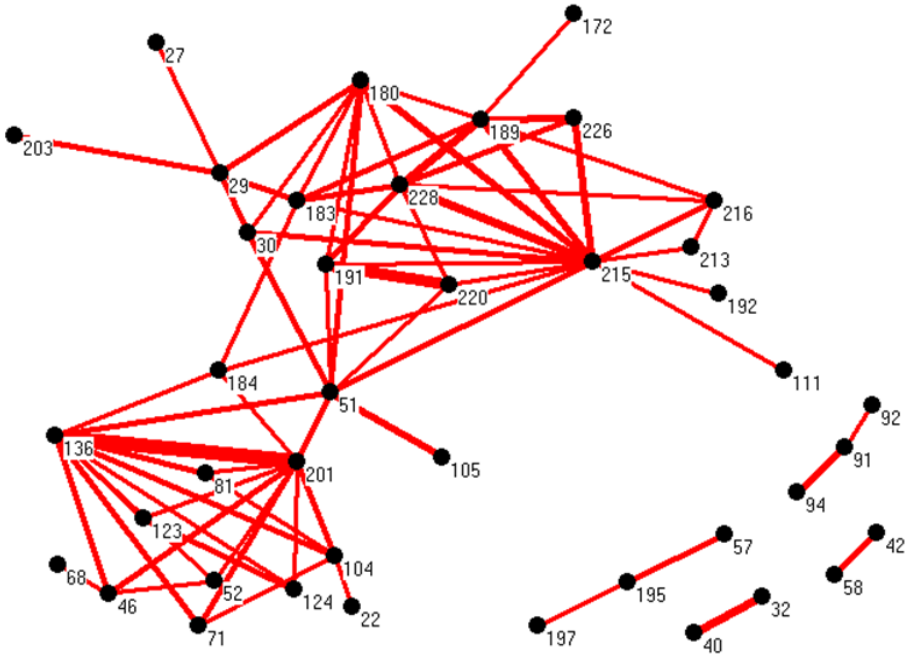
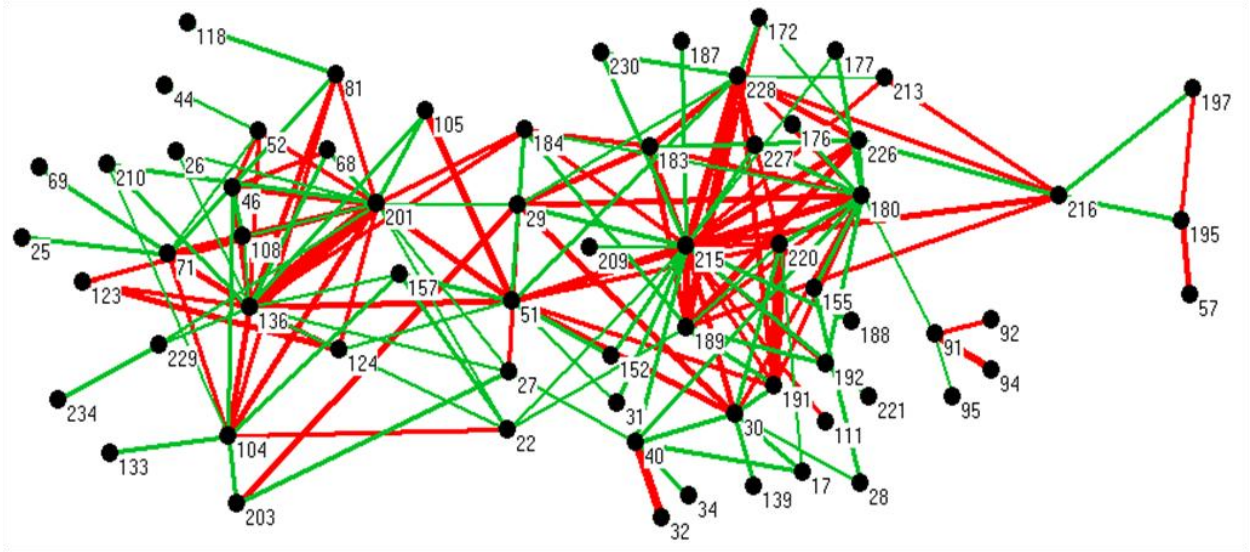


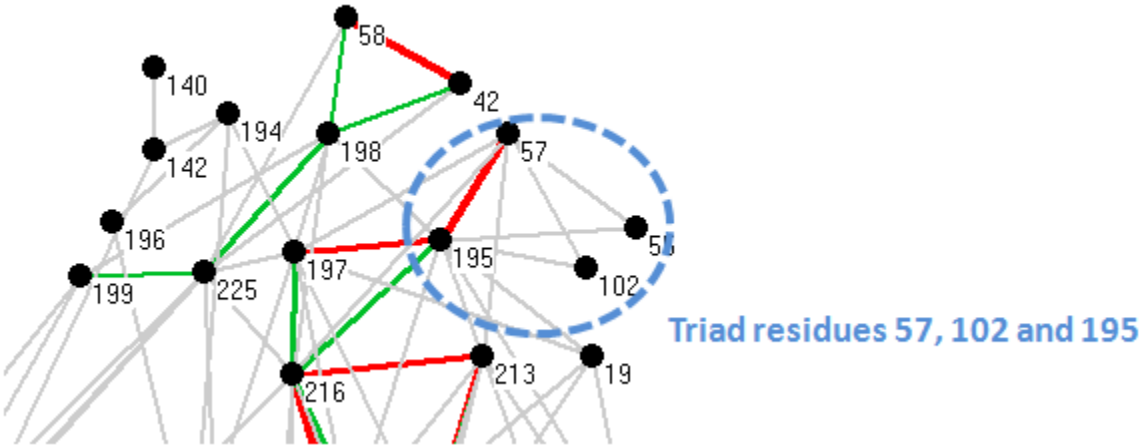
Figure 89: Network graph of serine protease S1A family. All nodes with coupling values 0.75 or greater are shown. Couplings with values greater than 1 are shown in red and couplings between 0.75 an 1 are shown in green. The node numbers are based on the structure of 3TGI.



The two network graphs reveal a striking network topology unlike those seen in other protein families. There is clearly sets of positions that form many more associations with each other than with other positions. There are at least three such positions. The sets are often bridged by particular residues. For example, one set of positions (residues 136,201,157,104,105,81,71,124) on the left of Figure 89 is connected to another set in the middle (residues 189,214,183,191,228,176,226) via a small set of bridging residues (184, 29, 51, 27 and 22).

In addition to the two sets making up the bulk of the connections, there is clearly a third set made of positions 195, 57 and 197 and bridged to the second set via position 216. What is striking about this set is that these positions are members of the catalytic machinery (S195 and H57) and absolutely conserved in the active proteases. In fact if I expand this network graph to couplings of 0.5 or higher the full catalytic triad comes out as linked, shown in Figure 90.

Figure 90: Network graph of serine protease family S1A. This graph shows all coupling greater than 0.5 but only a partial graph is shown focusing on the third set of positions which comprise the catalytic triad.



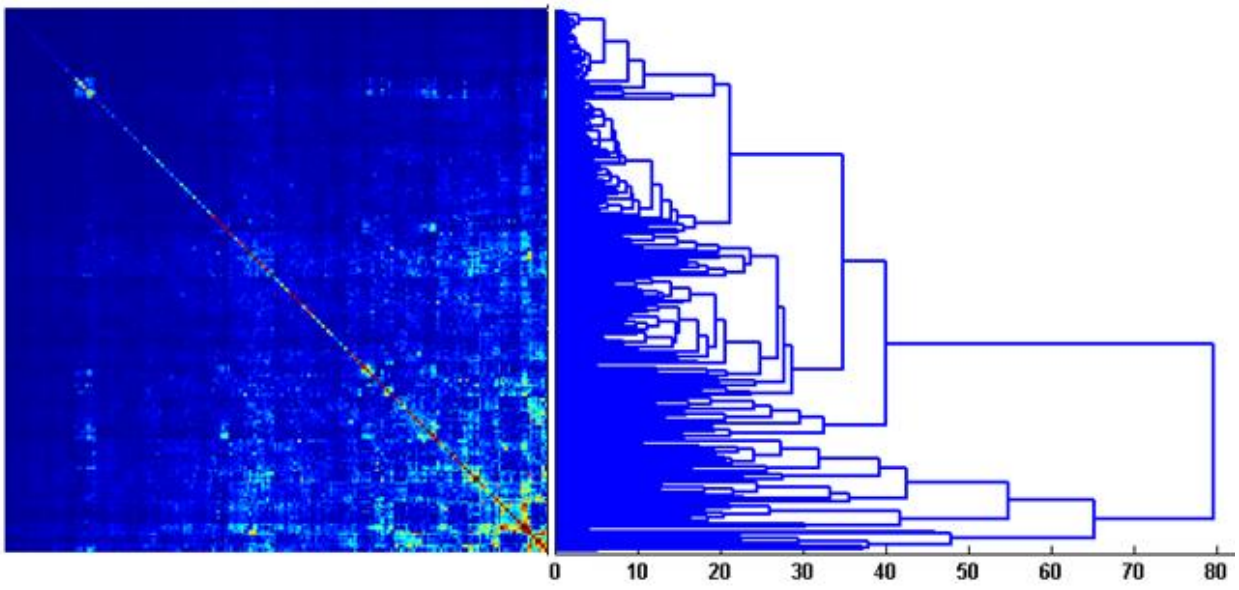
The biological significance of the catalytic machinery that appears out of the network graphs is evident. The significance of the other two sets will be discussed after further analysis of the coupling matrix.

The next stage of analysis is hierarchical clustering. Again many types of clustering are done and the results displayed in Table 13. As previously shown, the highest score correlation score is with Chebychev, average linkage whereas other types of distance and linkage methods give fairly good trees. The tree for cityblock distance and complete linkage is show in Figure 91.

Table 13: Hierarchical clustering of serine protease family. NaN means that the linkage or distance metric or a combination of linkage/distance does not apply to the data.

		linkage methods						
		'average'	'centroid'	'complete'	'median'	'single'	'ward'	'weighted'
distance methods	'euclidean'	0.96	0.95	0.86	0.91	0.92	0.48	0.91
	'seuclidean'	0.89	NaN	0.63	NaN	0.77	NaN	0.85
	'mahalanobis'	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	'cityblock'	0.91	NaN	0.74	NaN	0.89	NaN	0.84
	'cosine'	0.86	NaN	0.65	NaN	0.79	NaN	0.77
	'correlation'	0.79	NaN	0.44	NaN	0.60	NaN	0.67
	'spearman'	0.77	NaN	0.65	NaN	0.65	NaN	0.56
	'chebychev'	0.98	NaN	0.98	NaN	0.96	NaN	0.98

Figure 91: Cityblock distance with complete linkage tree for serine protease S1A family



The tree for cityblock distance/complete linkage shows at least two sets of coupled positions at the lower right which correspond to those observed in the network diagram.

I will now describe the principal components analysis.

The eigenvalue spectrum of the normal alignment and the random alignment (Figure 92) shows that there is one primary eigenvalue for each matrix that correlates with the magnitude of coupling (Figure 93).

Figure 92: Magnitude of the eigenvalues of the coupling matrix and the random coupling matrix.

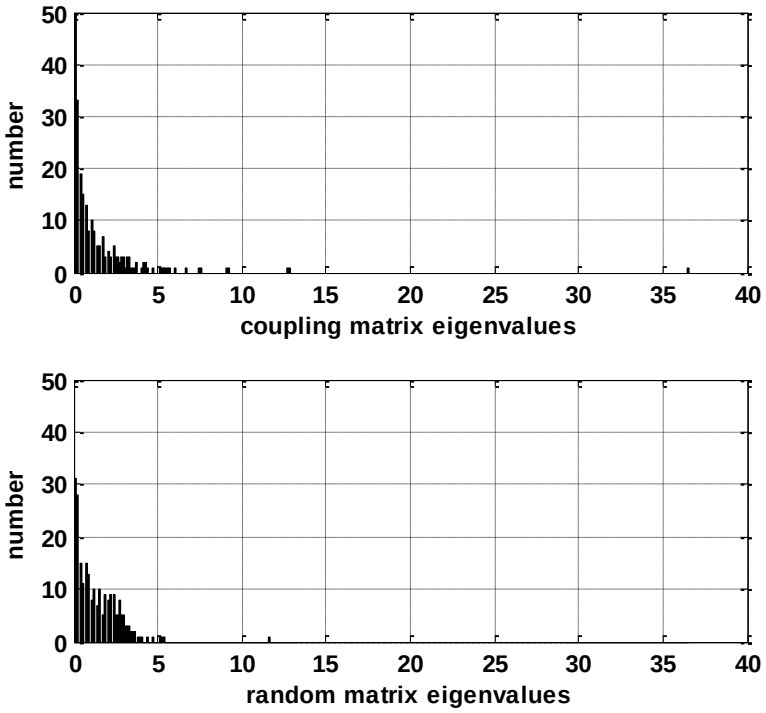
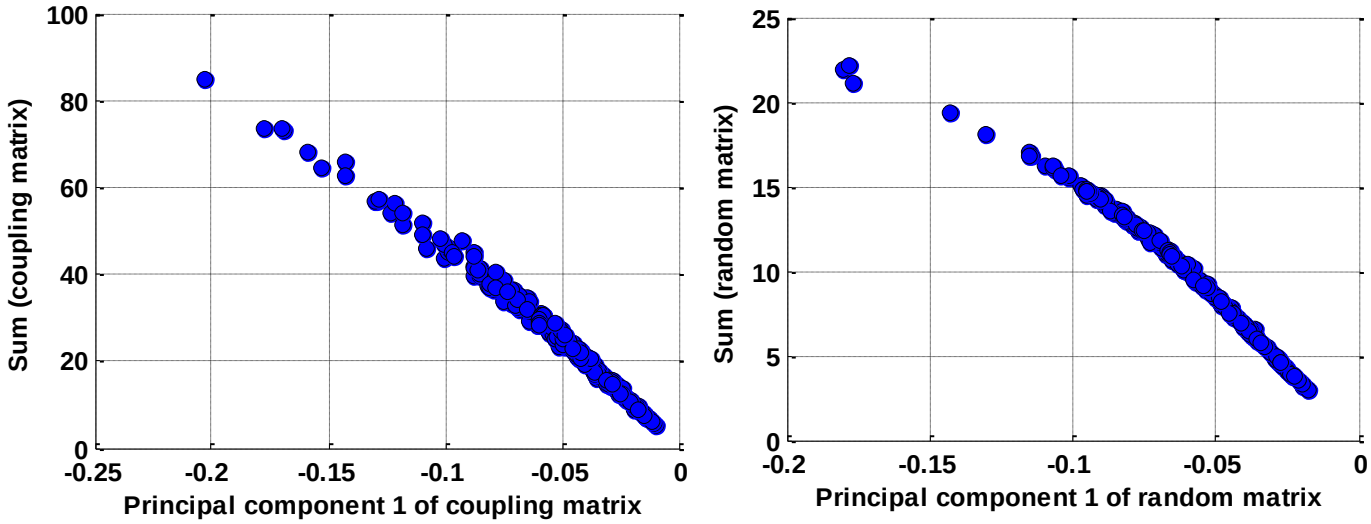


Figure 93: The first principal component correlates with the magnitude of coupling for both the normal coupling matrix and the random coupling matrix.



If the first principal component is removed, as it is not related to coupling, then, by comparison to the random matrix histogram it is clear that there are five eigenvalues with high values (Figure 94). I will take a close look at each eigenvalue. For clarity I will label the sets of positions that were identified in the network graph analysis with different colors. It will then be apparent if those sets occupy different regions of the variance space.

Figure 94: Eigenvalue histogram excluding the highest eigenvalue.

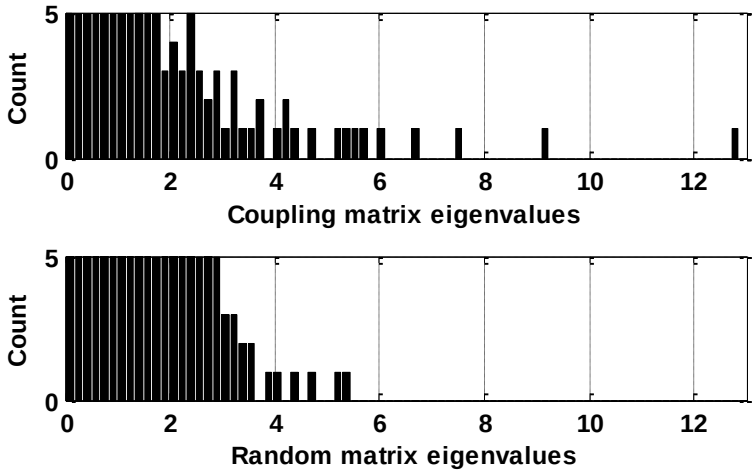
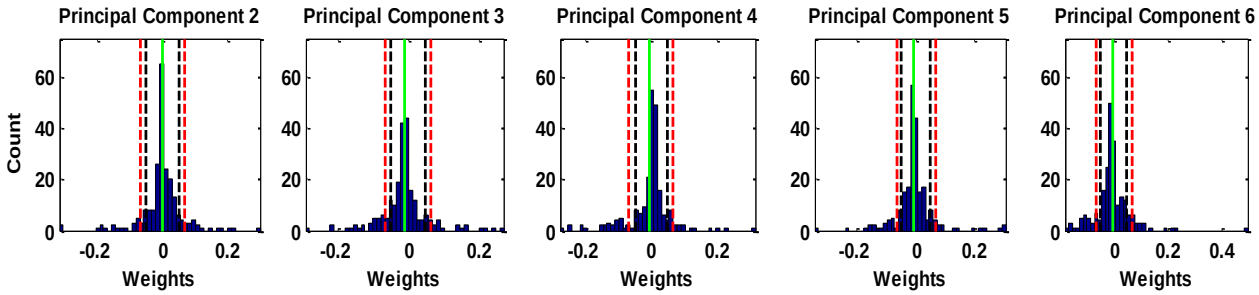


Figure 95: Histograms of the weights of the five significant eigenvectors. The green line is the mean; the red dashed line is one standard deviation and the black dashed line is at 0.05.



The next five figures show the weights of the six eigenvectors plotted against each other. These plots show what the network graphs and hierarchical clustering reveal (see for example PC2 vs. PC4 in Figure 96): that there are sets of residues that group together. What PCA does is it groups residues which have similar variance profiles as explained in chapter 3. It is clear that the set 1, set 2, set 3 and the bridge residues all occupy different regions of the variance maximizing space. For the bridge residues, while they are significantly different from random they are not all grouped together in all plots indicating that they do not form a distinct group of positions. Note also the position of residue 216 which sometimes appears as part of the red set and sometimes as part of the green set.

Figure 96: Principal component analysis plots: PC2 vs. PC3 and PC2 vs. PC4. Blue: set 2. Red: set 1. Green: set 3. Black: bridge residues. Yellow: All other residues.

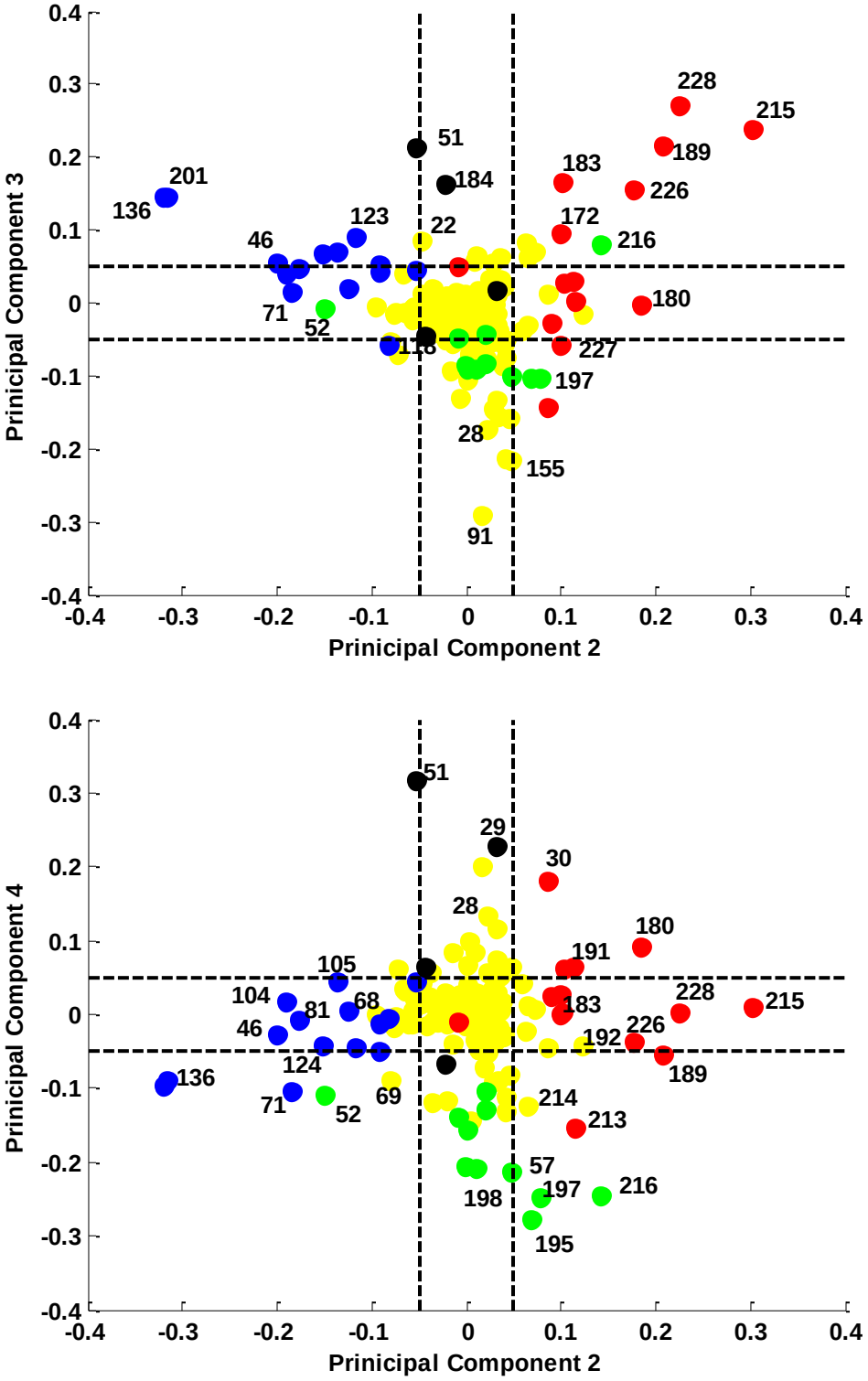


Figure 97: Principal component analysis plots: PC2 vs. PC5 and PC 2 v s PC6. Blue: set 2. Red: set 1. Green: set 3. Black: bridge residues. Yellow: All other residues.

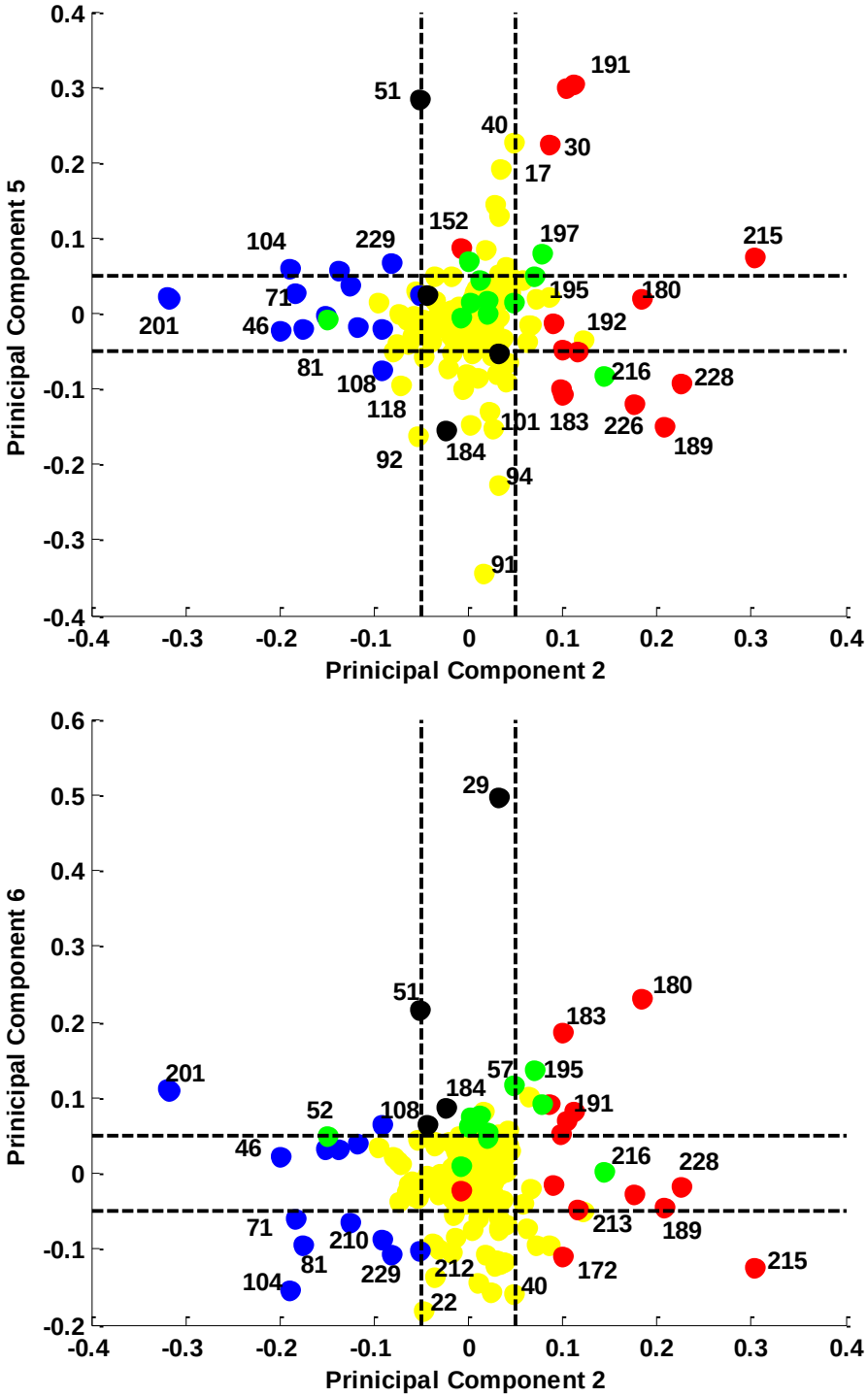


Figure 98: Principal component analysis plots: PC3 vs. PC4 and PC3 vs. PC5. Blue: set 2. Red: set 1. Green: set 3. Black: bridge residues. Yellow: All other residues.

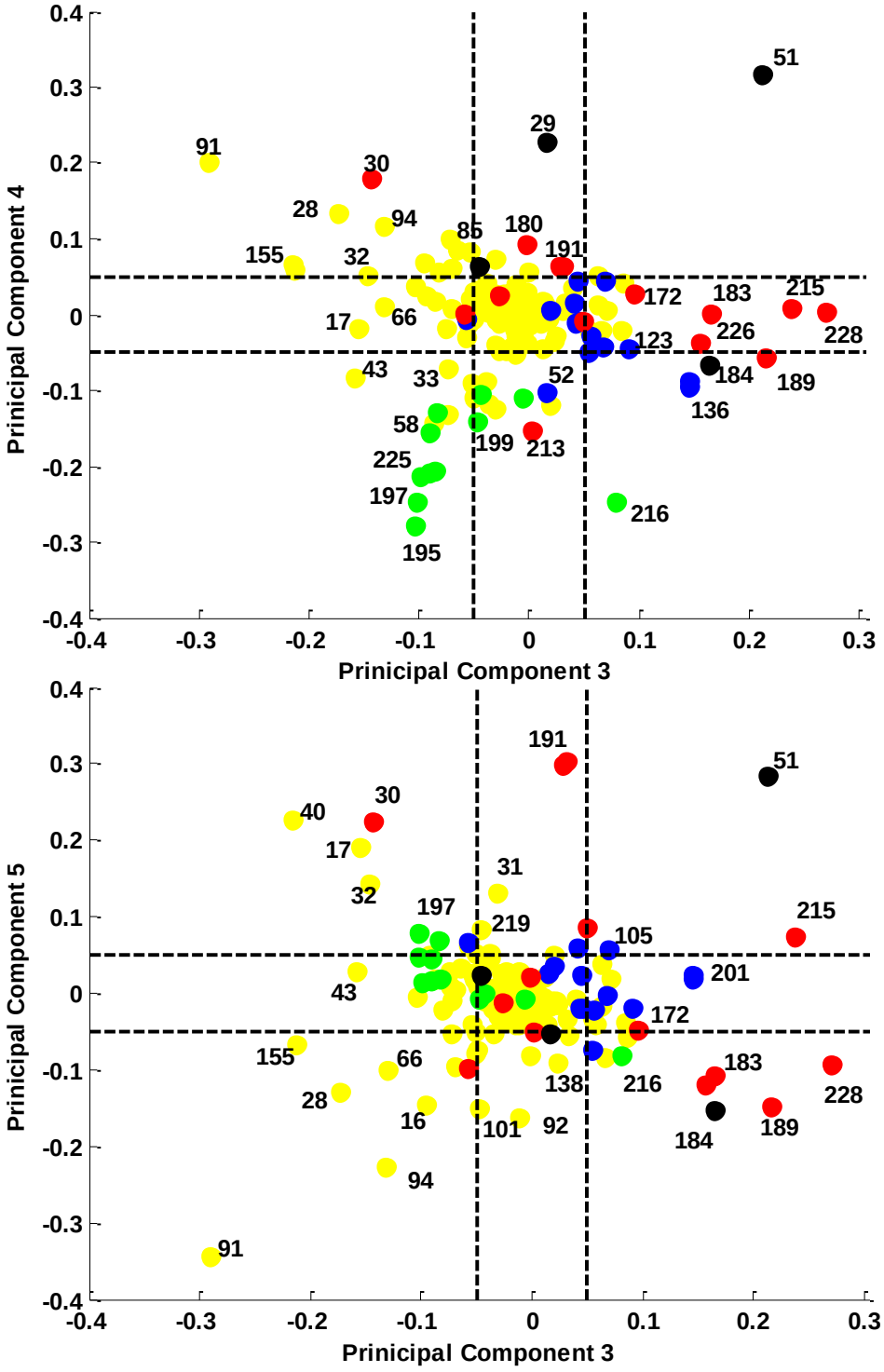


Figure 99: Principal component analysis plots: PC3 vs. PC6 and PC4 vs. PC5. Blue: set 2. Red: set 1. Green: set 3. Black: bridge residues. Yellow: All other residues.

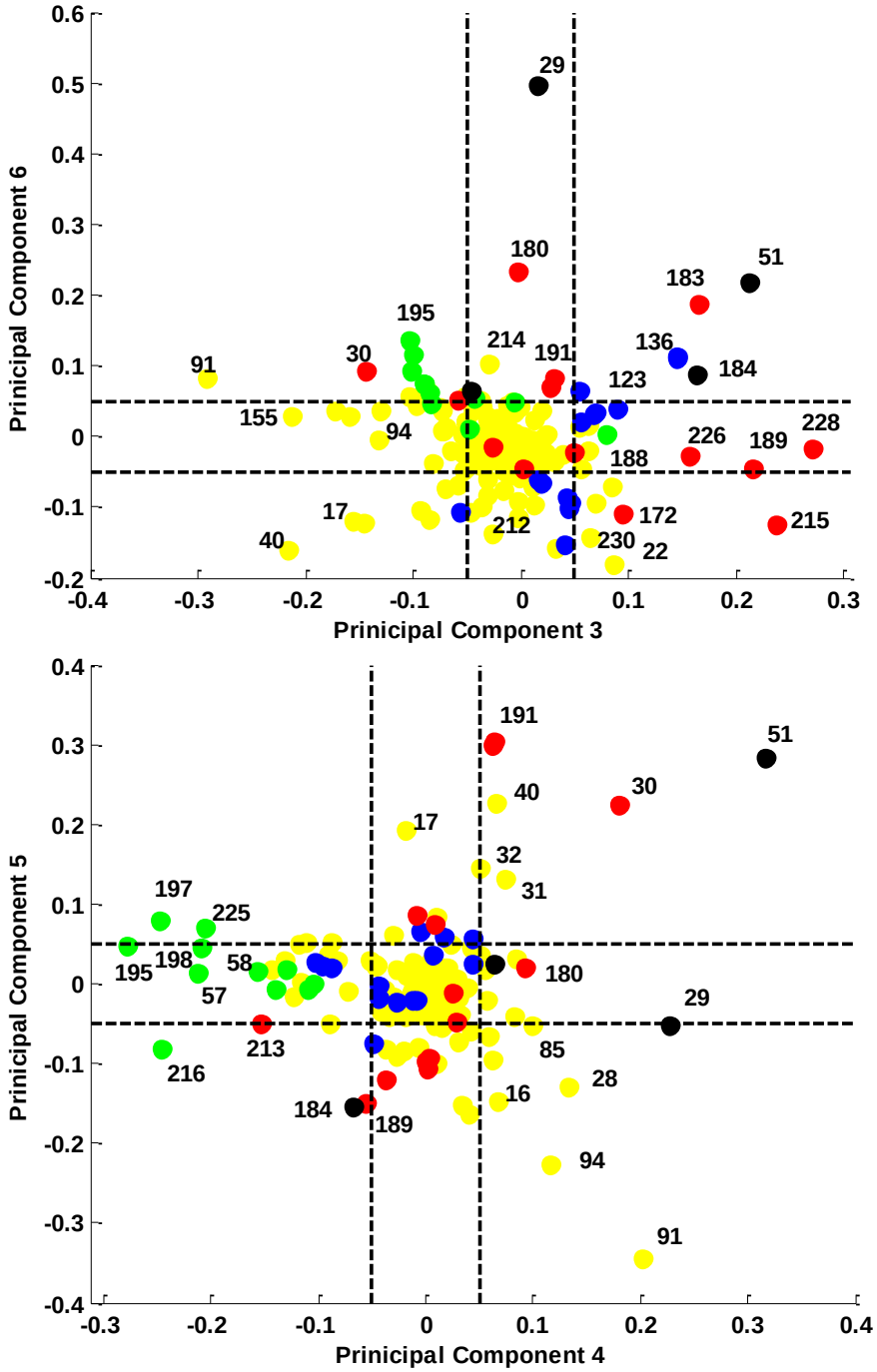
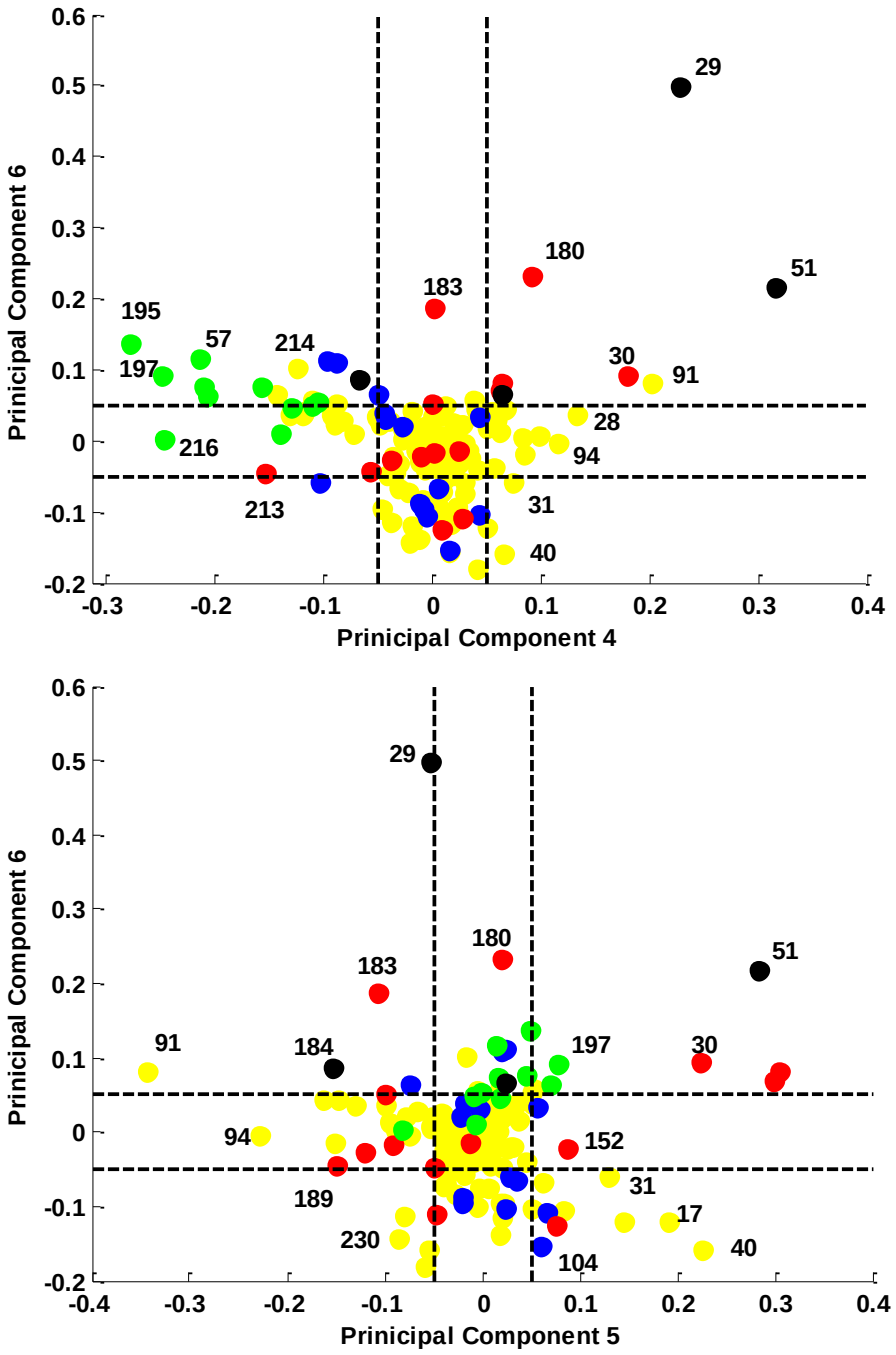


Figure 100: Principal component analysis plots: PC4 vs. PC6 and PC5 vs. PC6. Blue: set 2. Red: set 1. Green: set 3. Black: bridge residues. Yellow: All other residues.



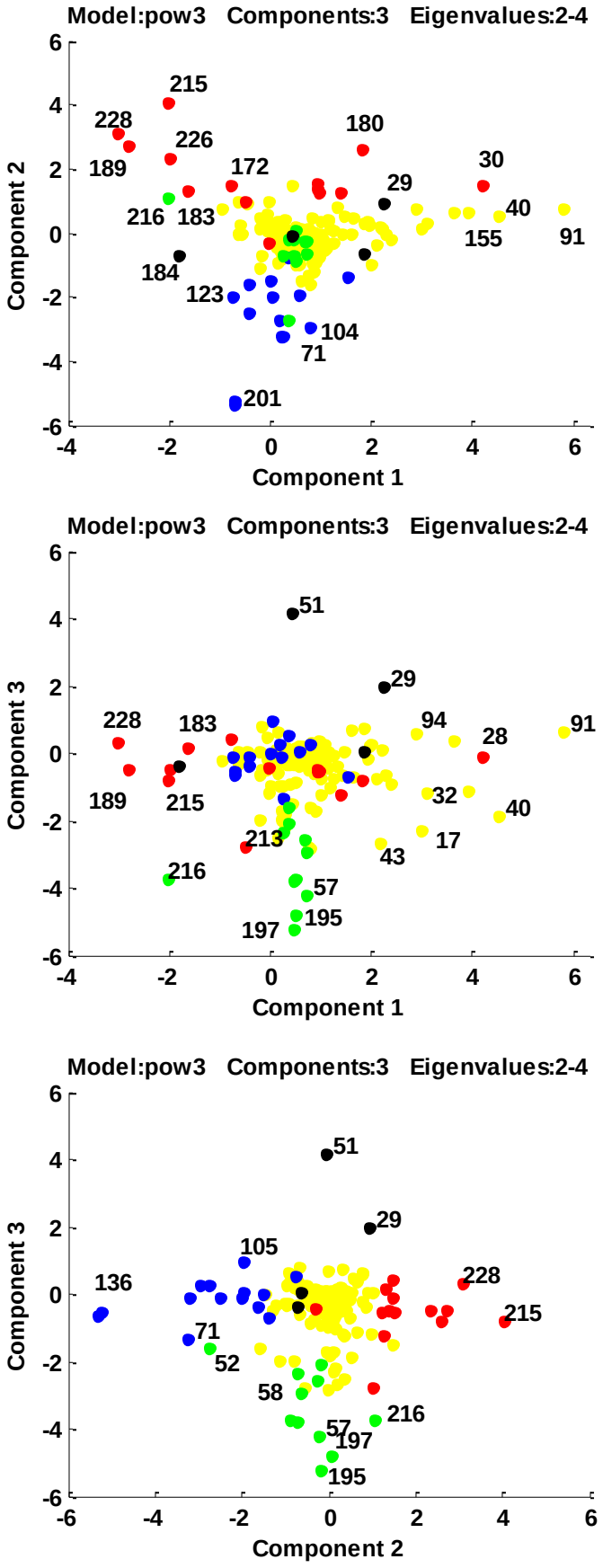
Although the PCA plots reveal that sets of positions cluster together, this is not evidence of independence. ICA can help determine if sets of positions are independent. ICA plots for one non-Gaussian model are shown in Figure 101. It is clear in the middle figure that set 3 (green) is perpendicular to set 1 (red) and set 2 (blue). On the other hand, the bottom plot shows that set 1 and set 2 are also almost perpendicular. This indicates that set 1 and set 2 and set 3 are independent or close to being independent. Other models for ICA also show the same trend (not shown). Position 216 seems to be part of both set 1 and set 3 and will be considered as a member of set 1 for mutation purposes. Also, position 29, 30 and 91 seem to be separate from other positions and will be considered as not members of any set. The positions in each set are in Table 14.

Now that there is evidence that set 1 and set 2 and set 3 are independent, these sets can be considered independently. Also, the sets will be termed sectors by analogy to economists who define economic sectors via the PCA approach that was taken here [166].

Table 14: List of positions in the different sets

Set 1	Set 2	Set 3
17	21	19
161	26	33
172	46	42
176	52	43
177	68	55
180	69	56
183	71	57
187	77	58
188	80	102
189	81	141
191	104	142
192	105	184
215	108	194
220	118	195
221	123	196
226	124	197
227	136	198
228	153	199
	157	213
	201	214
	210	216
	229	225
	237	
	242	
	245	

Figure 101: Independent component analysis of serine protease family S1A. Blue: set 1. Red: set 2. Green: set 3. Black: bridge residues. Yellow: All other residues.



The next two figures (Figure 102 and Figure 103) shows the sectors mapped onto the structure of bovine trypsin (pdb code: 2age). Each sector is made up of contiguous residues.

It is striking to see that members of different sectors could be in close proximity to each other. The biological function meaning of each sector will be discussed in detail in the next section.

Figure 102: The sectors mapped onto the structure of bovine trypsin (2AGE). Blue: sector 2. Red: sector 1. Green: sector 3. Yellow: substrate. Magenta sphere: calcium ion. Each structure is rotated 90 degrees relative to the axis indicated by the line.

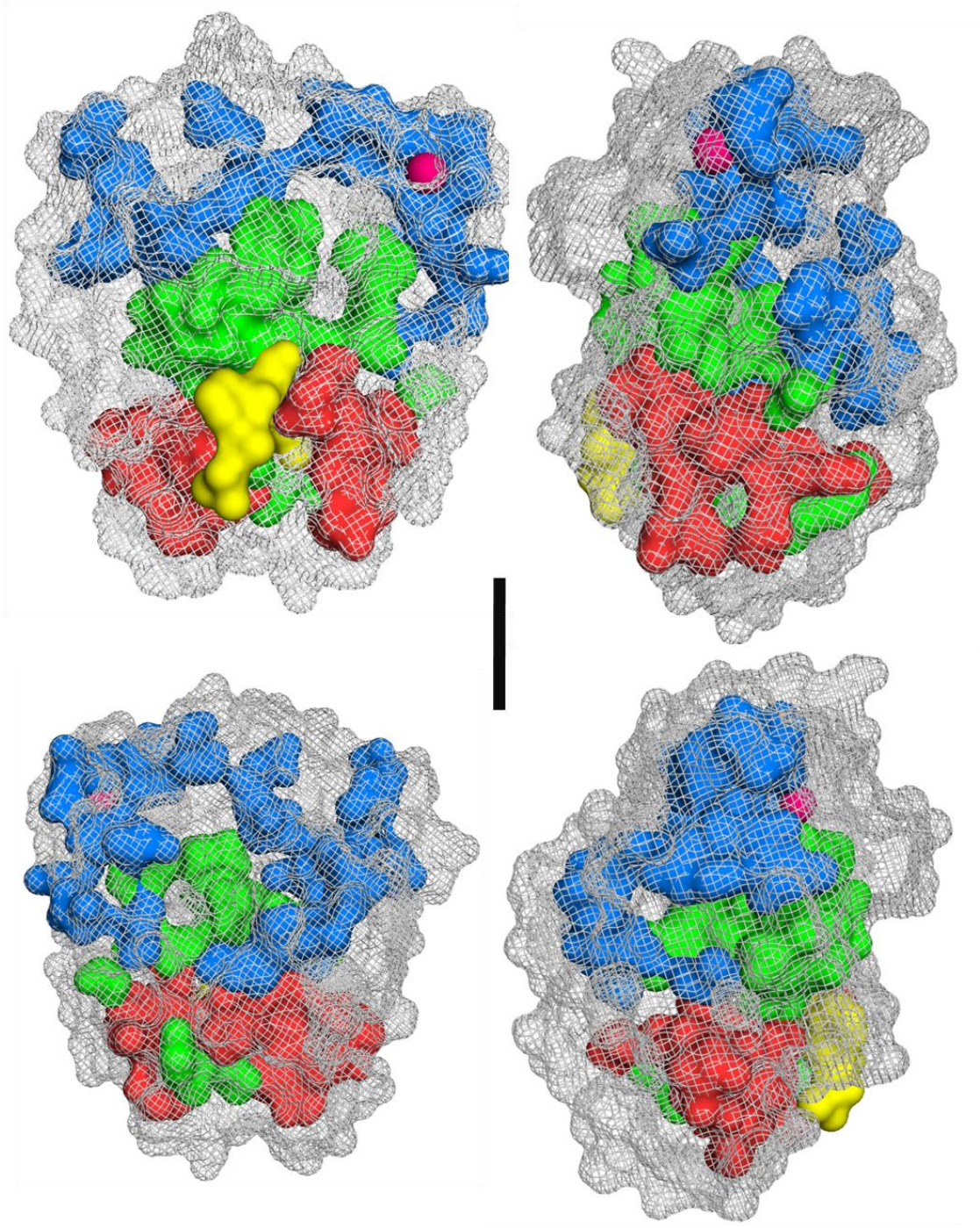
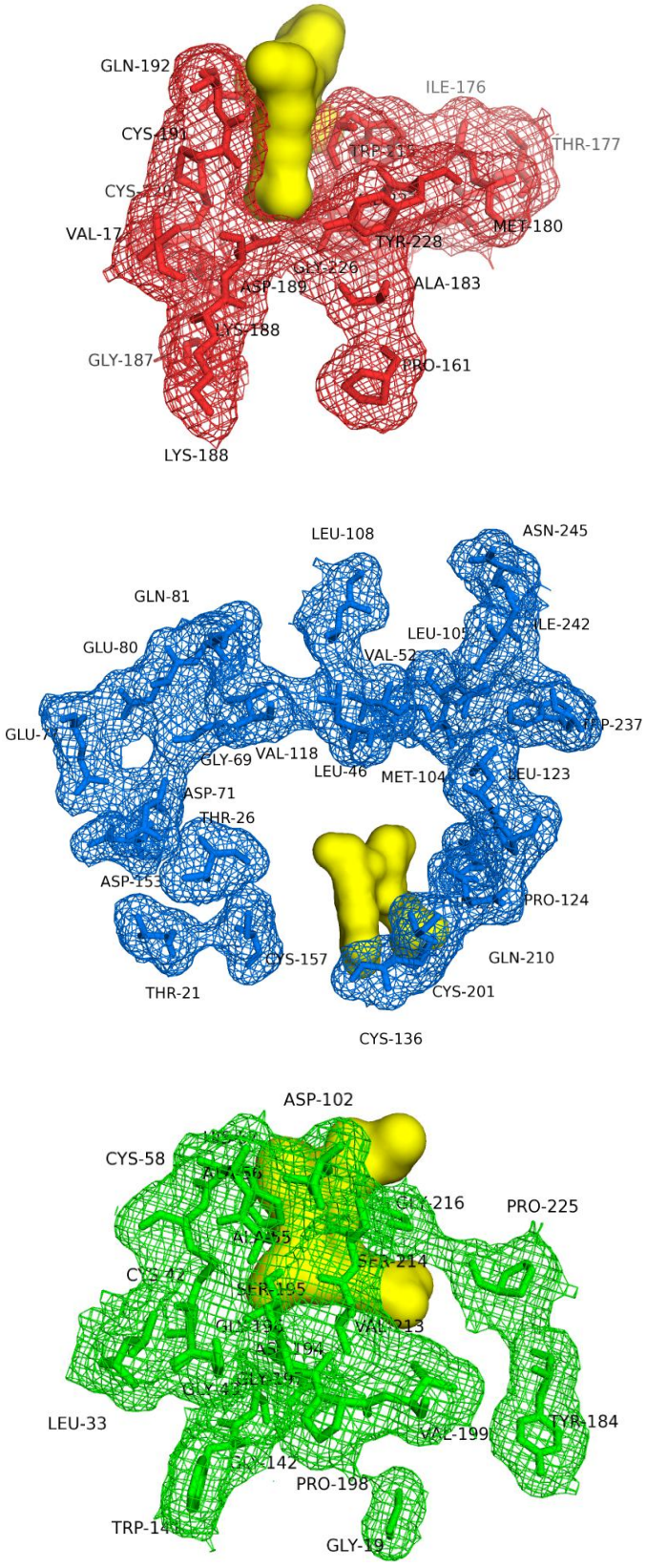


Figure 103: The three sectors individually mapped onto the structure of bovine trypsin (2AGE). The yellow surface is the surface representation of the covalently bound substrate. Red is sector 1. Blue is sector 2. Green is sector 3.



Section 3: Biological meaning of the sectors

The evidence for the evolutionary independence of sets of positions is clear. As the previous section shows, four different methods revealed that there are sets of positions where there are many more interactions within each set than with other positions in the protein. Furthermore, mapping onto the structure of a representative member of this family reveals that each sector forms a physically connected set. The evolutionary independence suggests strongly that each set may also be functionally independent. If that is true, then an evolutionary analysis would have decomposed the protein into evolutionarily and functionally independent units or sectors.

To examine the function of these sectors I will make use of the extensive literature available on this protein family and then report the results of experiments targeted at understanding the function concluding with a sector based distance method.

Literature study

First I will start with the third sector, shown in green in previous figures. The most striking positions in this sector are residues 195, 57 and 102. These are known members of the catalytic triad [132] where mutating any of them reduces the catalytic rate by a factor of 100,000 or more. In addition, residue 214 is part of the oxyanion hole because it plays a role in stabilizing the transition state of the reaction. This residue has been sometimes called as the fourth member of the catalytic triad. Some proteins have evolved mutations in these residues such as haptoglobin [146], hepatocyte growth factor [167], heparin binding protein [168] and clip-domain members [169] rendering them catalytically inactive.

Residues 194, 196, 197, 198 and 199 are also included in the green sector and these surround the catalytic serine 195 residue. Mutations in these residues (relative to the most frequent residue type) often confer resistance to natural inhibitors and are found naturally in brain proteases [170], snake venom proteases [171] and antifreeze proteins [172]. Residues 55 and 56 are also adjacent to the histidine 57 and may play a role in the orientation of that critical residue.

Position 225 is an interesting residue too as it forms a network extending from the surface to the interior of the protein via position 216. Extensive investigation by Di Cera's laboratory has shown that it is involved in switching the state of some serine proteases (thrombin is one example) from a slow enzyme to fast enzyme via a sodium ion [137, 138]. Specifically 225 is a proline in many protease classes except in the Vitamin-K dependent clotting pathway and the complement pathway. A member of the vitamin-K clotting pathway is thrombin. In thrombin, 225 is a tyrosine and mutation of this tyrosine to a

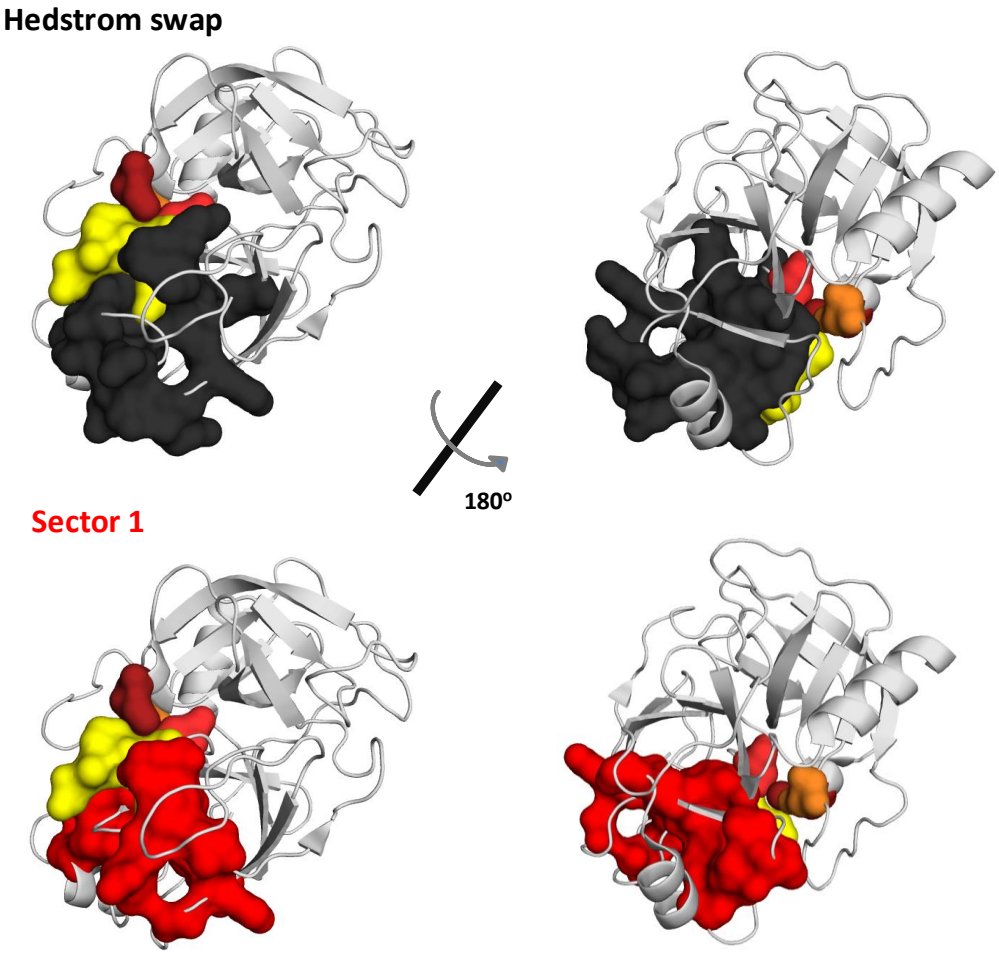
proline renders the enzyme incapable of being affected by sodium [138]. Changes in sodium during wound healing is thought to be important for the regulation of the clotting cascade [173].

The disulfide forming pair 42 and 58 is also interesting. It has been shown by Craik's group (a graduate from Rutter's group) that converting a serine protease to a threonine protease is aided by the C42A and C58A mutations (possibly by enlarging the active site to accommodate the larger side chain of threonine). Of note is that the threonine protease is still 100 to 10000 fold (depending on the substrate) lower in efficiency than the serine protease [174].

Information about the biological function of sector 1 (red) is also readily apparent from the literature and from looking at where the positions are on the structure.

Structurally, sector 1 positions are in the vicinity of the substrate binding pocket. Residues that make up sector 1 include residue 189 and 172 which are substrate specificity determinants. More generally the region specified by sector 1 includes many of the amino acids that needed to be swapped by Hedstrom to turn trypsin into chymotrypsin (see Figure 104).

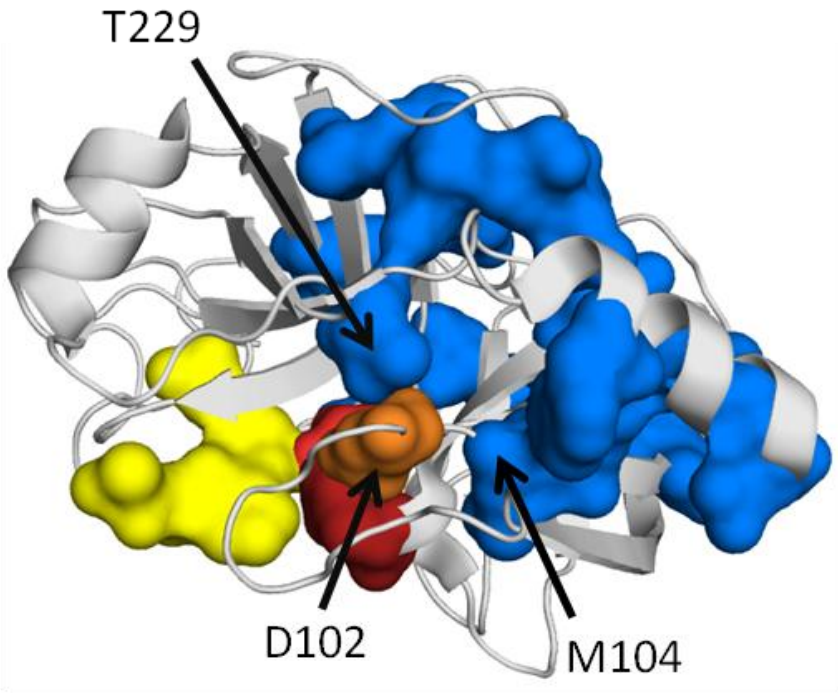
Figure 104: Hedstrom swap positions compared to Sector 1 positions. The Hedstrom swap positions include both positions mutated and positions that were identical between rat trypsin and cow chymotrypsin in the S1 pocket region.



Given that most sector 1 positions fall within the region required for specificity it is clear that set 1 positions are specificity positions. There are some positions (17, 161, 176, 177 and 180) that are not in the regions swapped by Hedstrom or considered as specificity determinants. Determination of their role in specificity awaits further investigation.

There is one more sector that requires functional characterization and that is the second sector (blue). The second sector is not close to the substrate and it has a more distributed architecture. It resembles somewhat a semi circle surrounding the protein. However, two residues (229 and 104) are in close proximity to the catalytic D102 (see Figure 105) and the first hypothesis I had was that these positions (and the sector) could also affect catalysis.

Figure 105: Sector 2 in relation to the catalytic triad. Blue: Sector 2 positions. Orange: D102. Yellow: Substrate covalently bound. Structure used is 2AGE.



However, a literature search turned up little data about the function of this set of residues. My project then was to obtain data about the function of this second set. Mutagenesis data has provided much insight into this protein family and so the strategy was to make mutants and assay for functions. Two functions were assayed: catalysis and fold stability. The next part of this section will describe the experimental assays and results and show further evolutionary analysis of the sectors. In the process of

assaying the second sector, I also assayed the function of sector 1 (red) which in some cases was a repeat of previously published data and confirmed that my methods were at least consistent with prior work in this field.

Experimental assays of protease function

The general experimental approach was to make alanine point mutants, express the recombinant protein in a yeast expression system, purify the active protein and then test the active enzyme for catalysis and fold stability. The methods used will first be described followed by the results and then a discussion of the results.

Mutagenesis and Cloning:

Mutagenesis was done using PCR site directed mutagenesis with overlapping oligoes encoding a mutation in one or two codons. Flanking primers amplified the full-length constructs. The full length PCR products were digested and ligated into pGEX vector and transformed into *E. coli*. Sequences of each construct were obtained. The colonies were then digested again for subcloning into the expression vector (yPT). The expression vector was then transformed into yeast.

Expression and Purification of enzymes:

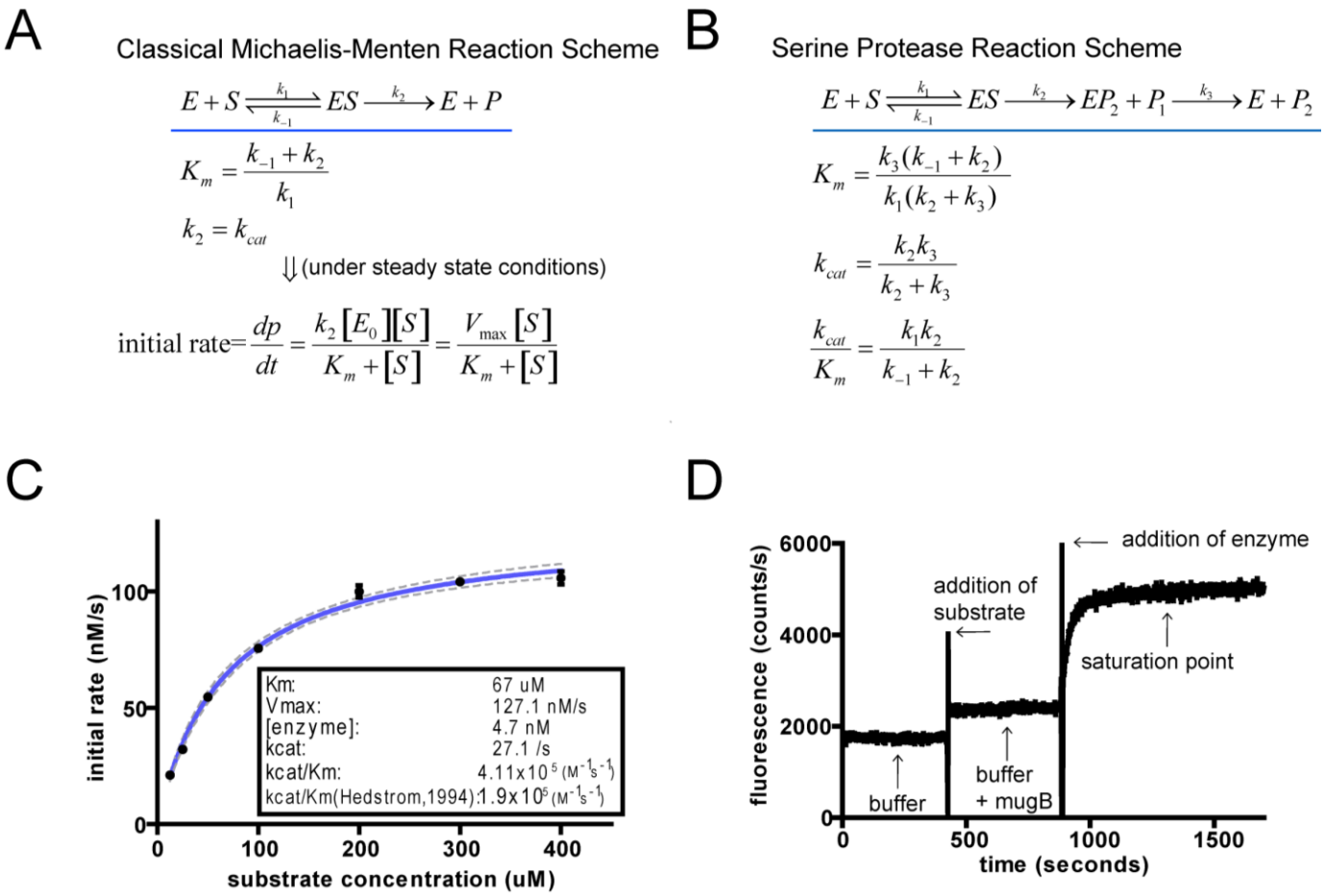
Purification of recombinant wild-type and mutant rat trypsins were adapted from a protocol kindly supplied by Lizbeth Hedstrom. Proteins were expressed in a *Saccharomyces cerevisiae* system (strain DLM101α) and induced via low glucose (1.5% w/v) where inactive zymogens are secreted into the culture medium. The culture is centrifuged at 7000g for 30 min to obtain cell-free supernatant. The supernatant is adjusted to pH 3.0 with 1 M HCL, gently stirred for 20 min at room temperature, and centrifuged for 120 min at 7000 g to pellet insoluble precipitates. Two to four ml of Toyopearl SP-650M cation-exchange resin is equilibrated in Buffer A (100 mM glacial acetic acid, 2 mM sodium acetate, pH 3.0) and added to the supernatant and stirred for a minimum of one hour. The resin is allowed to settle by gravity and most of the supernatant is decanted. The remaining resin + solution is loaded onto a Biorad polyprep chromatography column, washed with at least 100 bed volumes of Buffer A, and bound protein is eluted with 5 ml steps of Buffer A adjusted to pH 5.0, 6.0, 7.0 and 8.0 with 200 mM Tris pH 8.0. Eluted proteins are dialyzed for >8hrs into enterokinase buffer (50 mM Tris pH 6.5, 10 mM $CaCl_2$). For mutants that do not self-activate, enterokinase light chain (NEB) is added until 50% or more of the recombinant protein is activated. Activation is followed by SDS-PAGE. After activation, 1 to 3 mL of soybean trypsin inhibitor-agarose (Sigma-Alrdich) equilibrated in enterokinase buffer is added for at

least one hour with nutating. The activated enzymes bind specifically to this resin. After binding, the resin is loaded into a BioRad polyprep chromatography column and washed with 20-40 mL 50 mM Tris, pH 6.5 and 20-40 mL 50 mM Tris pH 6.5 + 0.5 M NaCl sequentially to wash away non specific binding. The protein is eluted with 2-4 ml of 0.1 M formic acid (pH 2.2). Proteins are stored in this buffer at 4 ^{o}C.

Catalytic assays:

Kinetic parameters of Vmax and K_m were measured using pseudo-first order kinetics as previously described [135]. The reaction scheme and assay of wild-type is shown Figure 106. The substrate used was Suc-Ala-Ala-Pro-Lys-PNA (Bachem) dissolved in dimethylformamide (DMF) to 50mM. Enzymes hydrolyze this substrate releasing p-nitroaniline, followed by absorption at 410 nm to detect hydrolysis (extinction coefficient of 10204 M^{-1} cm^{-1}). The enzyme reactions were done at 23^{o}C in 50 mM HEPES, 10 mM $CaCl_2$ and 100 mM NaCl, at a pH 8.0 (protease assay buffer, PAB). The total volume of reaction was 1 mL and the volume of substrate did not exceed 5%. A maximum of 20 ul of enzyme (in 0.1 M formic acid) was added to the reaction. In most cases, plots of initial velocity vs. substrate concentrations were fit to a hyperbola using non-linear regression to obtain K_m and V_{max}. R-square for all regressions was at least 0.9. To obtain k_{cat} (as V_{max}/active site concentration), active site concentration was measured by reacting the enzymes with 4-methylumbelliferyl p-guanidobenzoate (MUGB, Sigma-Aldrich), an enzyme inhibitor which releases a fluorescent compound, 4-methyl umbelliferone (4-MU) upon reacting with the enzyme. A standard curve of 4-MU was constructed to relate fluorescence counts to fluorophore concentration. Some enzymes did not react with MUGB and active site concentration was estimated by calculating the absorbance at 280 nm using the extinction coefficient of 33720 M^{-1} cm^{-1}. In some cases, the enzyme was not saturated with feasible concentrations of substrate so the approximation that K_m>>substrate concentration was used to calculate k_{cat}/K_m as the slope of the line of initial rate vs. substrate concentration.

Figure 106: Description of catalytic assay. A) The classical Michaelis-Menton scheme assumes a one step reaction. B) The serine protease reaction is a two step reaction and the K_m and Vmax reaction have different meanings in terms of the microscopic rate constants. C) Assay to measure K_m and Vmax of wild-type rat trypsin. Dashed lines are the 95% confidence intervals. D) Assay to measure active site concentration using mugB, a fluorescent inhibitor. The text contains more details of the assays.



Fold stability assays:

The fold stability was measured using thermal denaturation and monitoring the intrinsic tryptophan fluorescence of enzymes (rat trypsin has four tryptophans as shown in Figure 107). Stability was assayed in 0.1 M formic acid (pH 2.2) to keep enzymes inactive and prevent proteolysis during the course of the assay [175, 176]. The fluorescence (excitation at 295 nm/emission at 340 nm) was measured in the range of 4^{o}C to 85^{o}C (at a rate of 4^{o}C/min; sampling interval 0.1^{o}C for most proteins) in a 3 ml quartz cuvette with stirring. The total volume was kept at 2.1 mL to ensure that the rate of temperature change was the same across different assays. Pre- and post-transition baselines were fit by linear regression, subtracted from the raw data, and the melting temperature, denoted by T_m, was calculated by the differential method [177, 178]. Briefly, baseline subtracted data were smoothed by the robust lowess method (MATLAB, Mathworks Inc.), differentiated, and the T_m measured as the extremum of the differential melt. C136A showed no observable transition in the temperature range of the experiment. All data were collected at least in triplicate. The steps for analyzing the fold stability curves are illustrated further in Figure 108.

One concern with using for a fold stability probe the tryptophans is that there are four tryptophans which means that it is possible to observe multiple transitions if different parts of the protein fold differently. However, in the enzyme data shown, the stability curves show that there is one main sharp transition. To confirm the suitability of this assay, I also performed differential scanning calorimetry (DSC) on selected mutants. Conditions were similar to the fold stability assays except that the protein concentration for DSC at 20 to 80 uM was higher than the tryptophan fluorescence. As shown in Figure 109, the melting temperature for the DSC data correlated very well to the melting temperature for the fluorescence melts (regression coefficient of 0.93). This correlation is striking especially when considering two mutants, Q81A and Y228A which by both DSC and fluorescence showed two noticeable transitions but which nevertheless correlate with each other. One thing to note about the fluorescence and DSC melting temperature is that the DSC melting temperatures are generally 2-3 degrees higher. One possible reason for this discrepancy is that the protein concentration is much higher in DSC than fluorescence (by a factor of at least 100) which could result in a stabilizing effect on the proteins. Regardless of the exact melting temperature, the DSC data confirms that tryptophan fluorescence is a suitable assay to monitor the folded state of the enzymes and to distinguish enzymes with different melting temperatures.

Figure 107: The four tryptophans in rat trypsin. Structure is 3TGI.

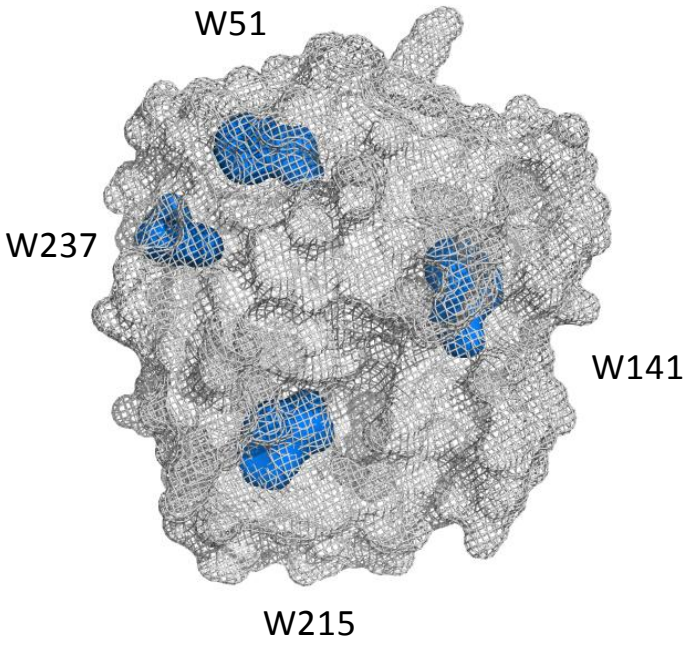


Figure 108: Analyzing tryptophan denaturation curves. This figure shows the four steps involved in the analysis. 1) Collect tryptophan denaturation data (raw data curve). Fit linear pre and post transition curves (extrapolated curves). 3) Subtract the raw data curve from the extrapolated curves (baseline subtracted curve). 4) Calculate the first derivative of the baseline subtracted curve. The baseline subtracted curve and the derivative of the baseline subtracted curves have been normalized between 0 and 1. The equation for subtracting the extrapolated linear lines (y1 and y2) from the raw data (x) is: x subtracted (i)=(y1(i)-x(i))(y2(i)-y1(i)). The data shown here are smoothed.

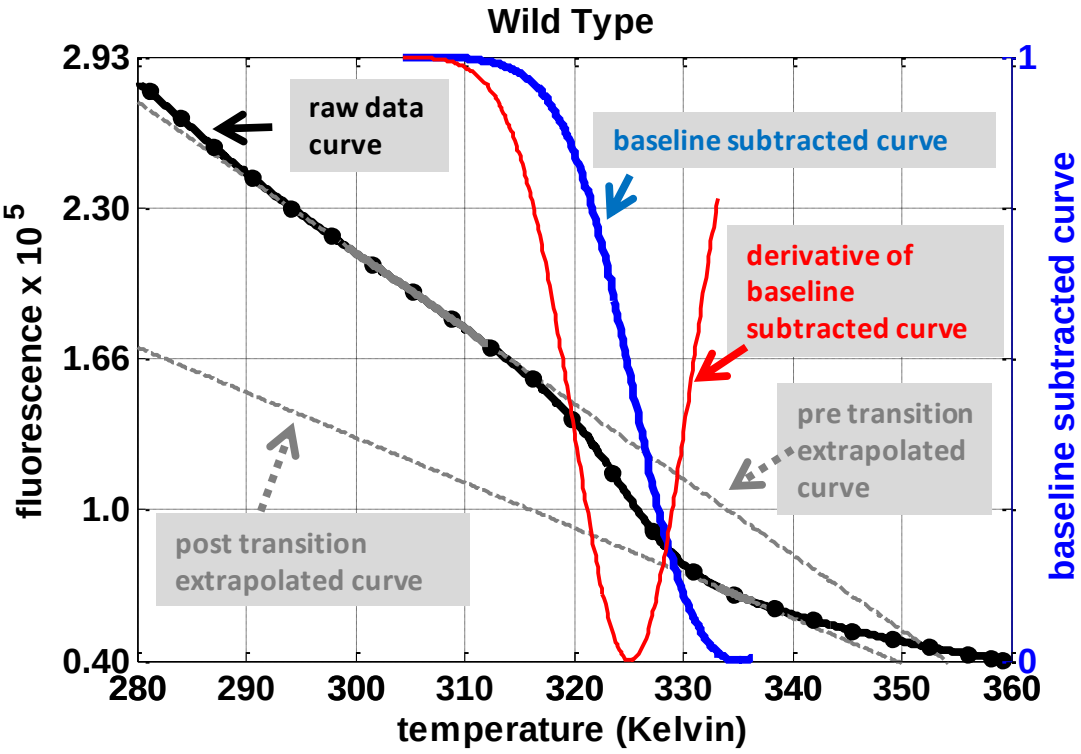
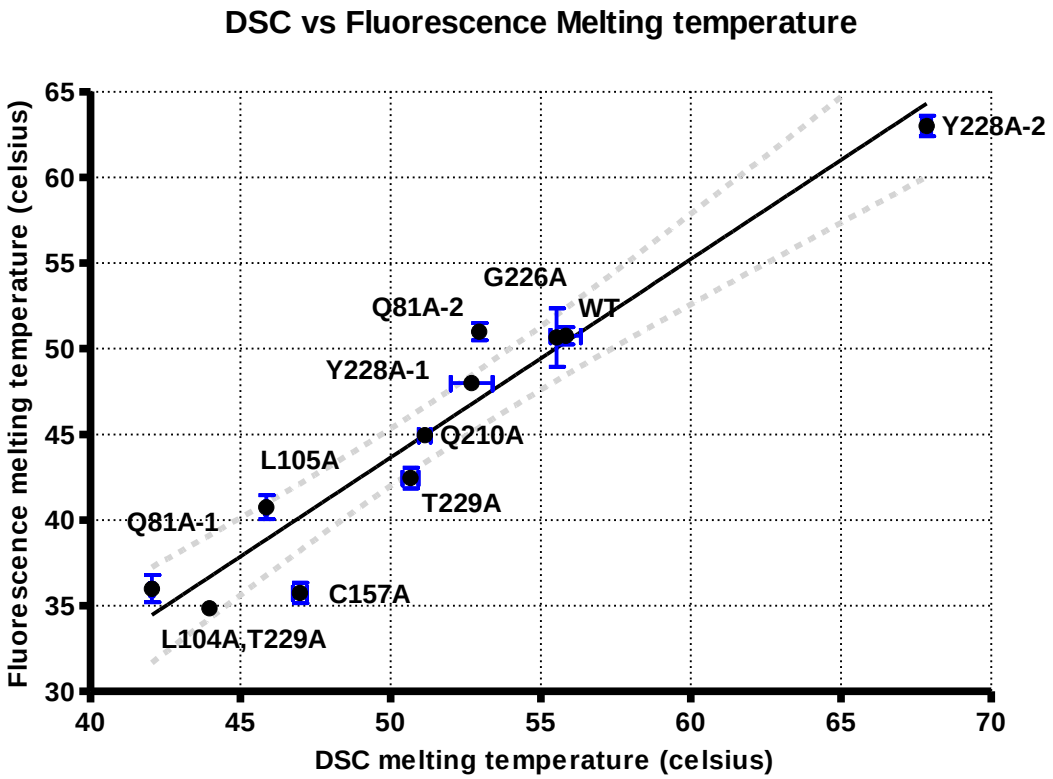


Figure 109: Correlation between melting temperature obtained with DSC and melting temperature obtained with tryptophan fluorescence. The black line is a linear fit (regression coefficient of 0.93) while the grey line shows the 95% confidence interval. The blue error bars show the standard errors.



Now that the methods have been explained, I will show in the next figures all the catalytic data (Figure 110 and Figure 111) and the fold stability data (Figure 112, Figure 113, Figure 114, Figure 115) for all the set of mutants tested. The mutants are from the first and second sectors (red and blue respectively) as well as a small set of control residues that are not in the three major sectors (Q30A, K230A, Y29A). There are also two mutants, G216A-Q210A and G216A-C157A in both the red and blue sectors to demonstrate independence between the sectors. The fold stability curve of C136A is not shown as there is no transition. The data for all mutants and all assays are summarized in Table 15 and Figure 116. A discussion of the results follows after the data is presented.

Figure 110: Catalytic reactions for single mutants in the blue, red and control sectors. The colors of the dots refer to the sector identity. Red is sector 1. Blue is sector 2. White is non sector positions. The errors are the standard errors. The solid line is a non-linear regression fit. The dashed line is the 95% confidence interval of the fit.

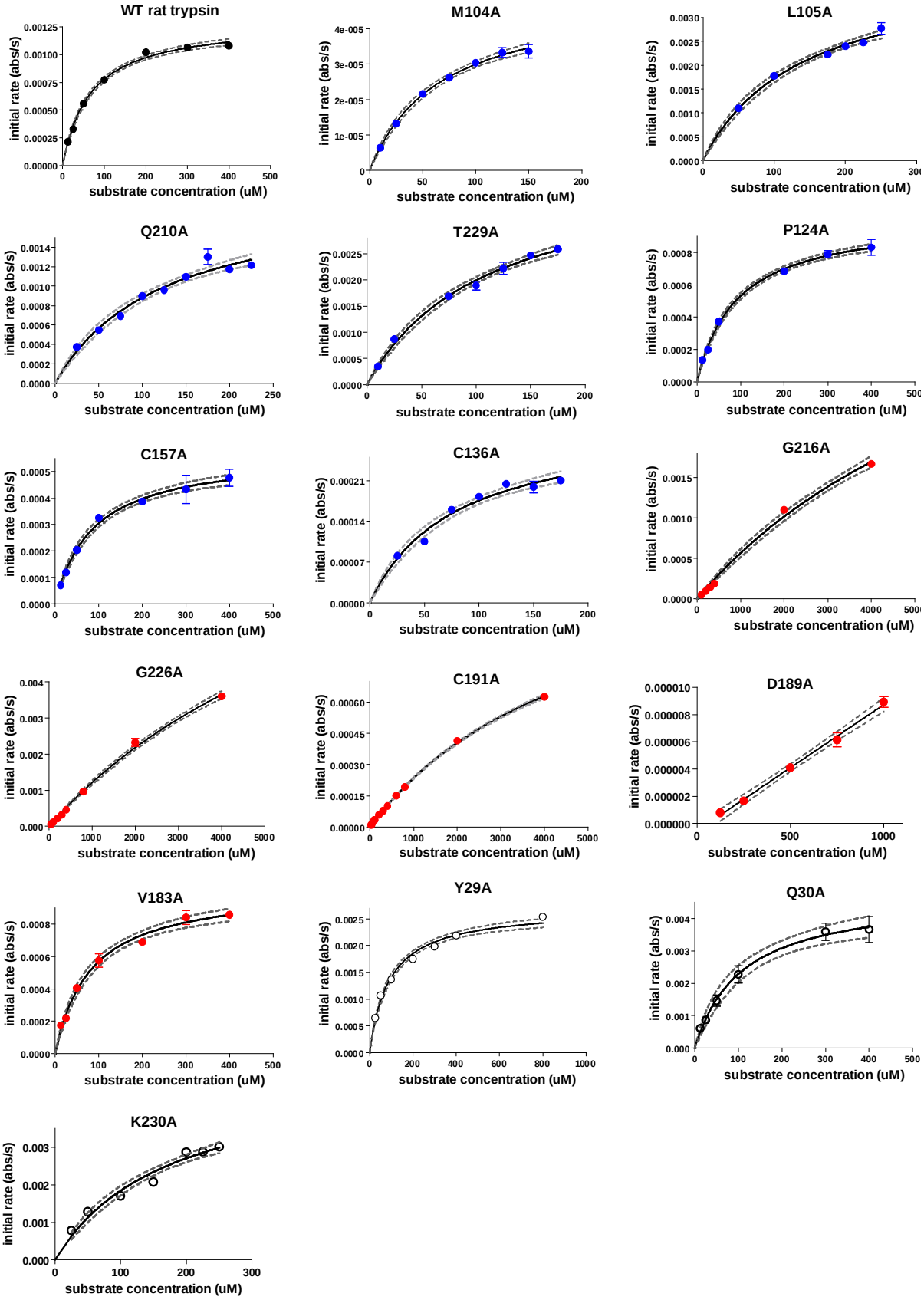


Figure 111: Catalytic reactions for double and multiple mutants. The colors of the dots refer to the sector identity. Red is sector 1. Blue is sector 2. White is non sector positions. The errors are the standard errors. The solid line is a non-linear regression fit. The dashed line is the 95% confidence interval of the fit.

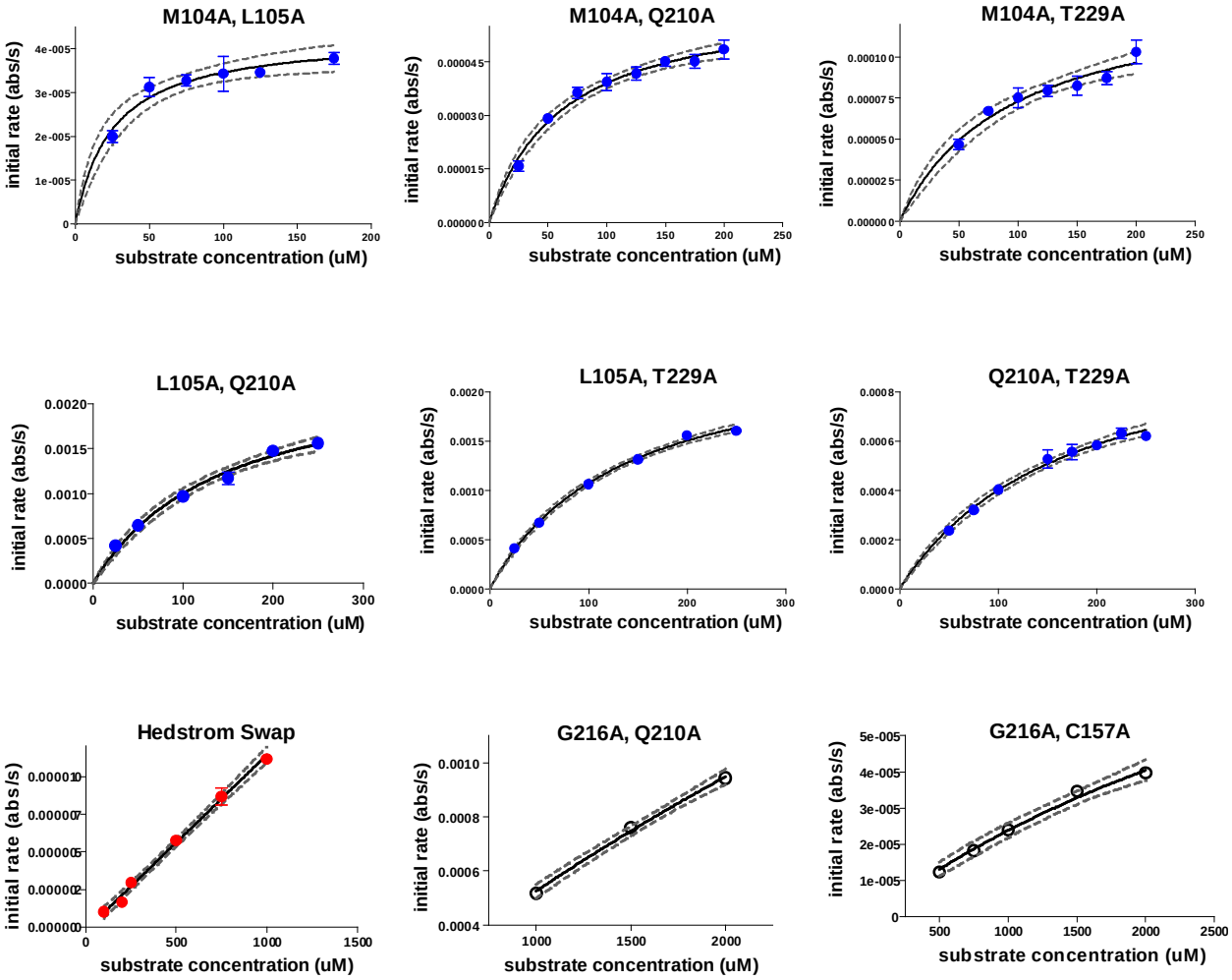


Figure 112: Fold stability curves: M104A, L105A, Q210A, T229A, P124A and C157A. Refer to Figure 108 for legend and explanation.

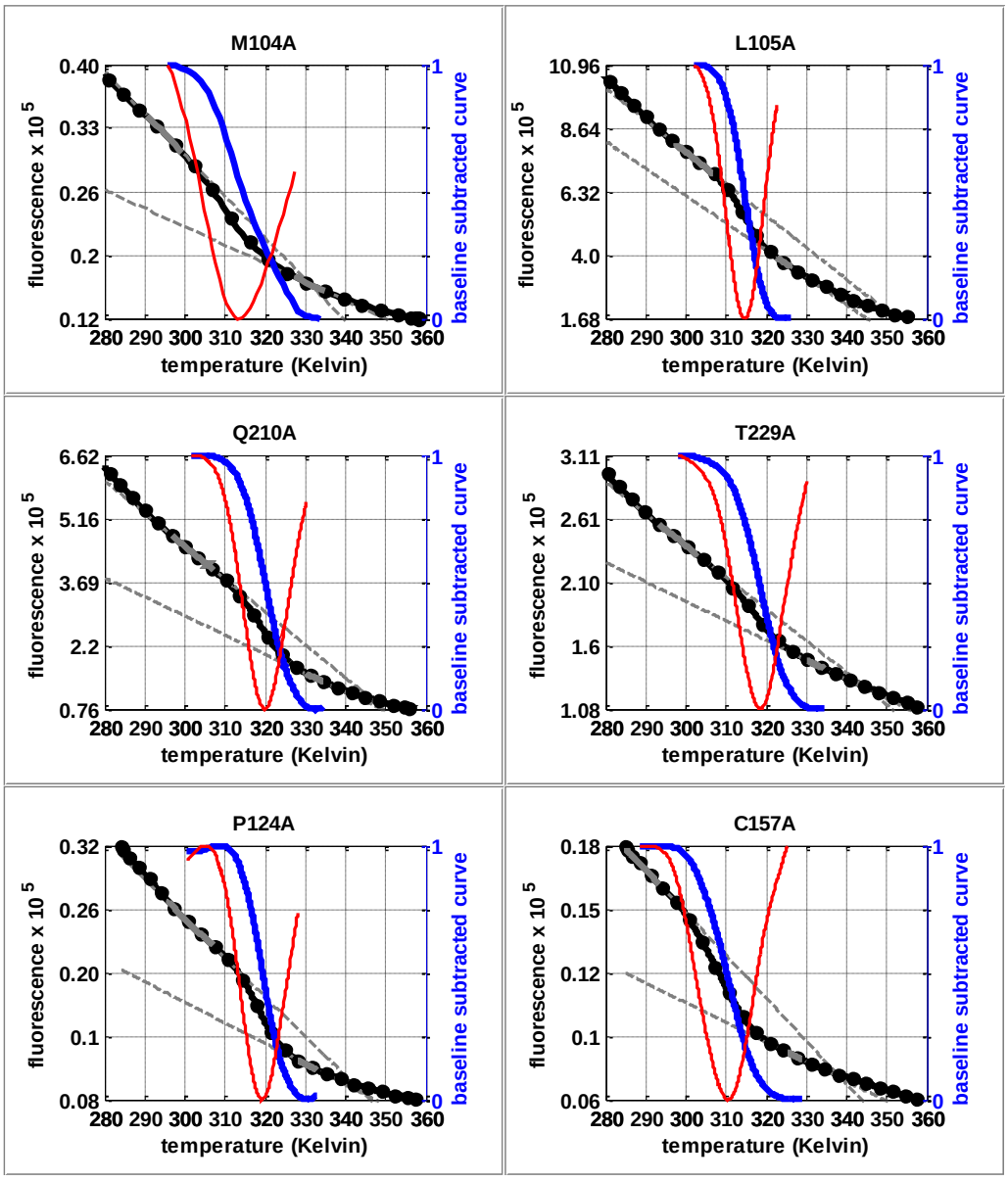


Figure 113: Fold stability curves. G216A, G226A, C191A, D189A, V183A, Y29A. Please refer to Figure 108 for legend and explanation.

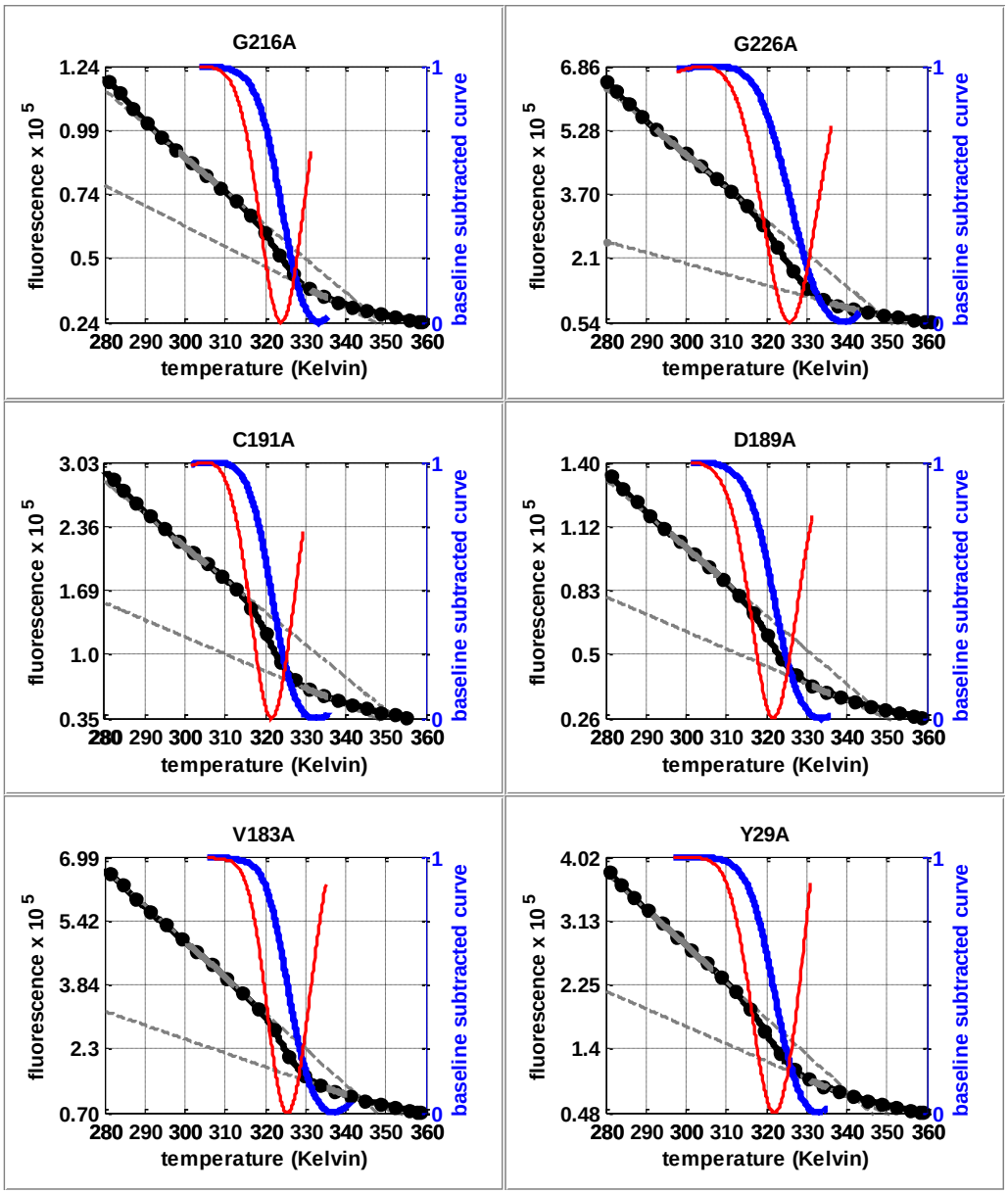


Figure 114: Fold stability curves: Q30A, K230A, M104A-L105A, M104A-Q210A, M104A-T229A, L105A-Q210A. Please refer to Figure 108 for legend and explanation.

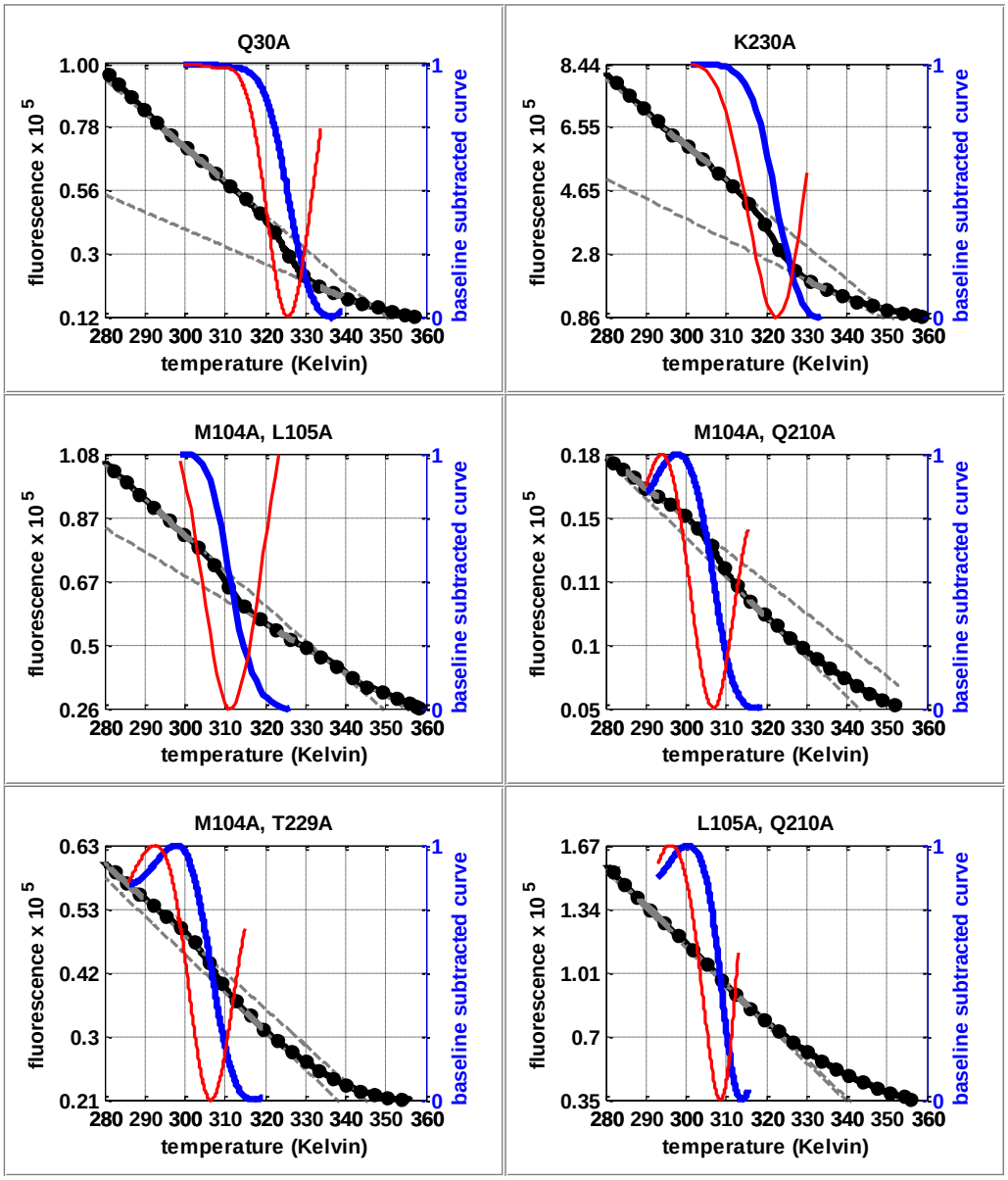


Figure 115: Fold stability curves: L105A-T229A, Q210A-T229A, Hswap, G216A-Q210A, G216A-C157A. Please refer to Figure 108 for legend and explanation.

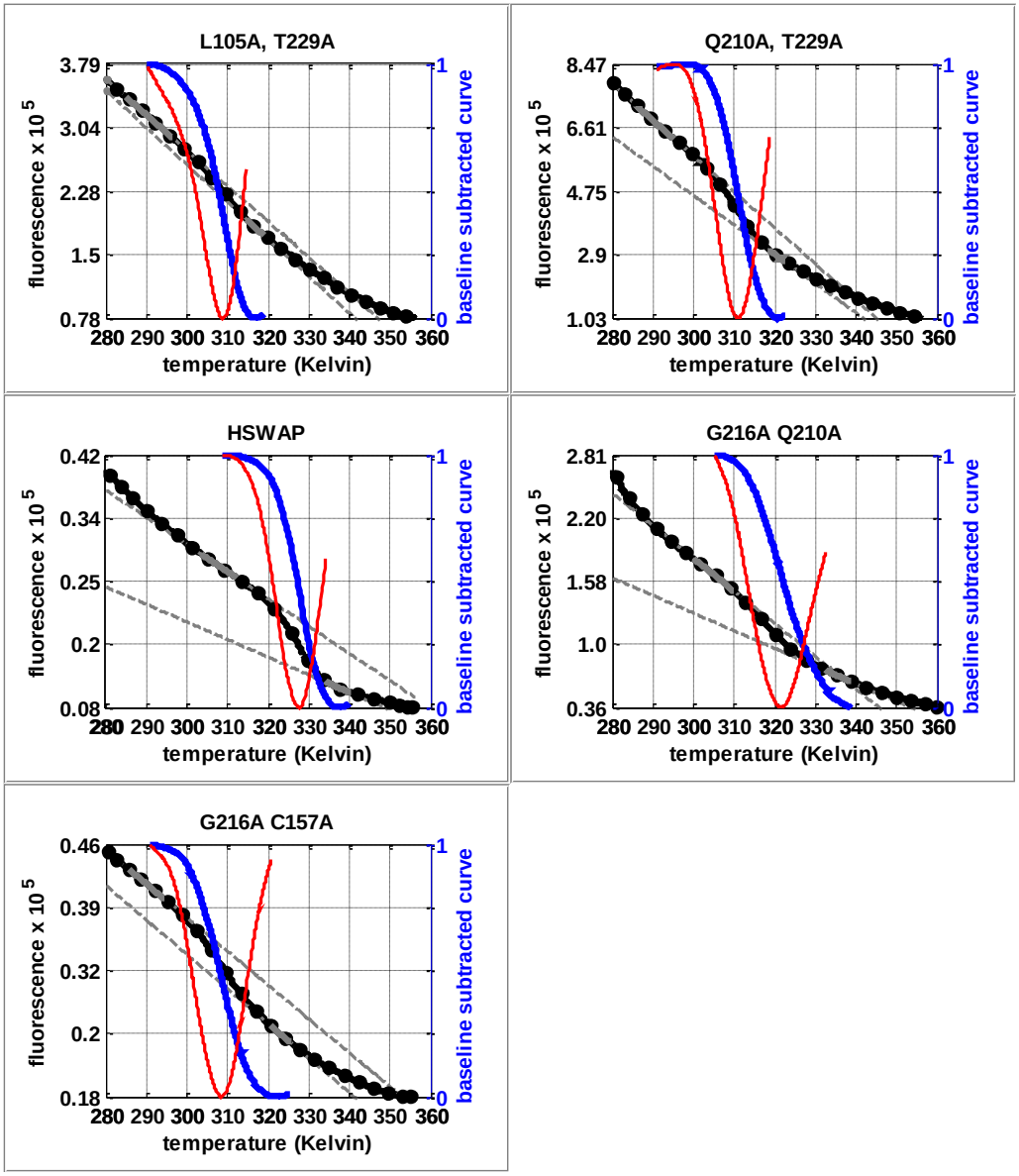
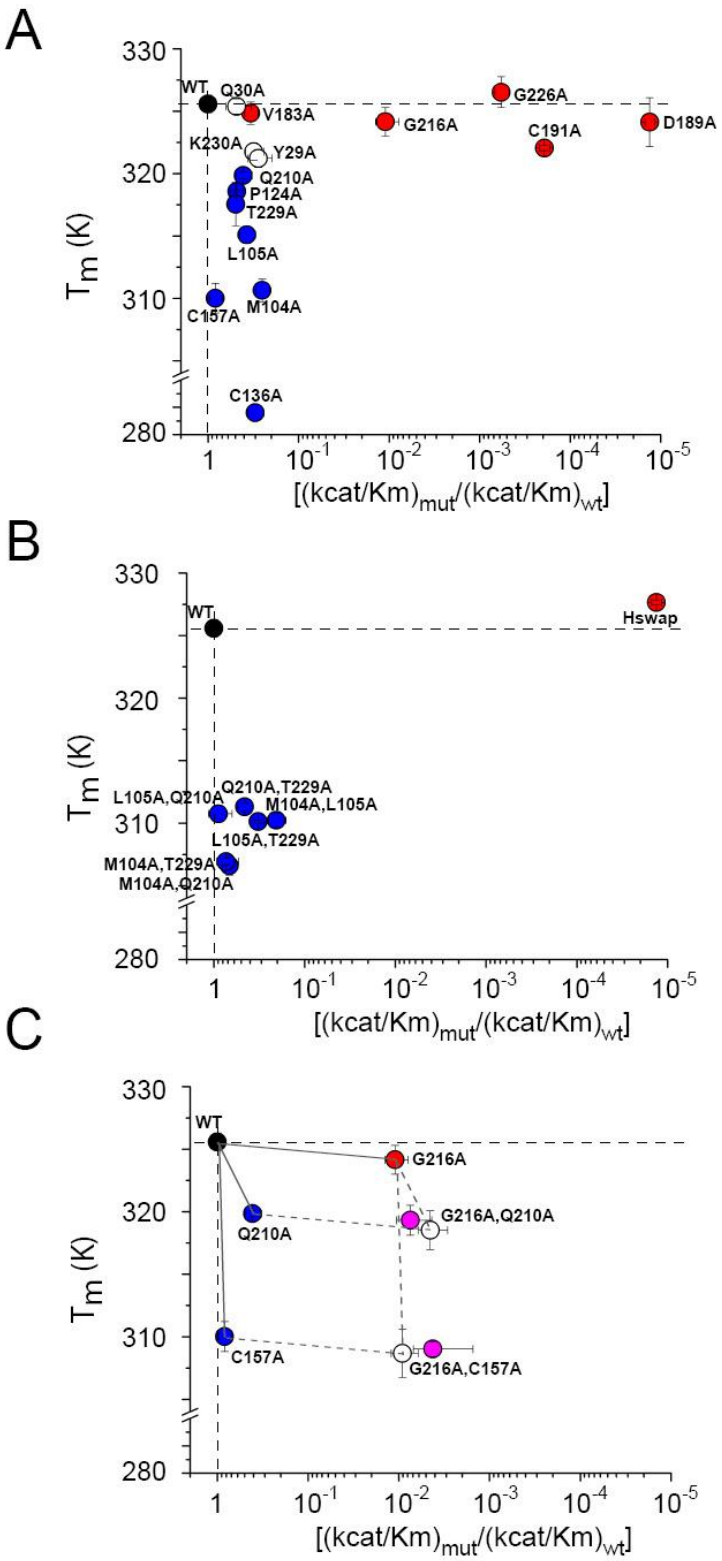


Table 15: Fold stability and catalytic parameters for sector and non sector positions. The errors in parenthesis are the standard errors.

		Thermal stability	Catalytic parameters		
		Tm (Kelvin)	Km (M^{-1})	kcat (s^{-1})	kcat/Km (M^{-1} s^{-1})
WT	Rat trypsin	325.58 (+/- 0.71)	6.6E-05 (+/- 4.3E-06)	27 (+/- 0.53)	4.1E+05 (+/- 2.8E+04)
Blue Sector	M104A	310.65 (+/- 0.92)	6.7E-05 (+/- 8.2E-06)	14 (+/- 0.76)	2.1E+05 (+/- 2.9E+04)
	L105A	315.10 (+/- 0.70)	1.4E-04 (+/- 1.9E-05)	38 (+/- 2.37)	2.7E+05 (+/- 4.1E+04)
	Q210A	319.85 (+/- 0.26)	1.2E-04 (+/- 1.9E-05)	43 (+/- 3.02)	3.4E+05 (+/- 5.8E+04)
	T229A	317.55 (+/- 1.74)	1.1E-04 (+/- 1.5E-05)	45 (+/- 2.93)	4.1E+05 (+/- 6.2E+04)
	P124A	318.58 (+/- 0.12)	1.0E-04 (+/- 1.0E-05)	40 (+/- 1.20)	4.0E+05 (+/- 4.2E+04)
	C157A	310.02 (+/- 1.19)	8.7E-05 (+/- 9.0E-06)	60 (+/- 2.09)	6.9E+05 (+/- 7.6E+04)
	C136A	284.0 (+/- 0)	7.7E-05 (+/- 1.2E-05)	19 (+/- 1.20)	2.5E+05 (+/- 4.0E+04)
	M104A,L105A	310.75 (+/- 0.69)	2.5E-05 (+/- 6.4E-06)	18 (+/- 1.19)	7.4E+05 (+/- 2.0E+05)
	M104A,Q210A	306.60 (+/- 0.76)	6.2E-05 (+/- 9.0E-06)	35 (+/- 1.78)	5.6E+05 (+/- 8.6E+04)
	M104A, T229A	306.92 (+/- 0.25)	8.2E-05 (+/- 1.9E-05)	50 (+/- 4.81)	6.1E+05 (+/- 1.5E+05)
	L105A, Q210A	310.23 (+/- 0.57)	1.4E-04 (+/- 2.4E-05)	24 (+/- 1.91)	1.7E+05 (+/- 3.1E+04)
	L105A, T229A	310.12 (+/- 0.12)	1.3E-04 (+/- 1.0E-05)	37 (+/- 1.31)	2.7E+05 (+/- 2.3E+04)
	Q210A, T229A	311.32 (+/- 0.32)	1.7E-04 (+/- 2.3E-05)	65 (+/- 4.52)	3.8E+05 (+/- 5.7E+04)
Red Sector	G216A	324.15 (+/- 1.15)	7.9E-03 (+/- 1.7E-03)	72 (+/- 11.14)	9.1E+03 (+/- 2.4E+03)
	G226A	326.52 (+/- 1.24)	8.1E-03 (+/- 1.1E-03)	4 (+/- 0.04)	4.8E+02 (+/- 6.4E+01)
	C191A	322.05 (+/- 0.20)	5.0E-03 (+/- 1.8E-04)	1 (+/- 0.02)	1.6E+02 (+/- 7.5E+00)
	D189A	324.12 (+/- 1.96)	N/A	N/A	1.1E+01 (+/- 5.7E-01)
	V183A	324.85 (+/- 0.89)	8.1E-05 (+/- 9.2E-06)	23 (+/- 0.89)	2.8E+05 (+/- 3.4E+04)
	Hswap	327.65 (+/- 0.20)	N/A	N/A	1.1E+01 (+/- 6.5E-01)
Blue-Red Sector	G216A, Q210A	319.32 (+/- 1.19)	8.3E-03 (+/- 2.5E-03)	51 (+/- 13.67)	6.2E+03 (+/- 2.5E+03)
	G216A, C157A	309.02 (+/- 0.60)	4.6E-03 (+/- 1.9E-03)	16 (+/- 7.47)	3.5E+03 (+/- 2.2E+03)
Non Sector	Y29A	321.22 (+/- 0.72)	9.1E-05 (+/- 7.1E-06)	47 (+/- 1.07)	5.2E+05 (+/- 4.2E+04)
	Q30A	325.38 (+/- 0.61)	1.1E-04 (+/- 2.3E-05)	39 (+/- 3.02)	3.7E+05 (+/- 8.2E+04)
	K230A	321.75 (+/- 0.71)	1.8E-04 (+/- 3.2E-05)	45 (+/- 4.14)	2.6E+05 (+/- 5.2E+04)

Figure 116: Plots of catalytic activity vs. melting temperature. A) Single mutants of sector 1 (red) and sector 2 (blue) and non-sector positions (white). B) Double mutants of sector 2 (blue) and the Hedstrom swap (red). C) Double mutants between the red and blue sectors. The white circles are that predicted if the two sites act independently whereas the pink circles is the actual red/blue double mutant. Errors depicted are standard errors of the mean of at least triplicate measurements.



Discussion of experimental data

Figure 116 illustrates several points that addressed the functional significance of the second sector (blue) and whether the second sectors acts independently of the first sector (red).

First, sector 2 positions clearly have little effect on catalytic function (A, Figure 116) including T229A and M104A which are in close proximity to the catalytic D102. This is in sharp contrast to the positions comprising sector 1. Even double mutants in the sector 2 (B, Figure 116) have little effect on catalytic function including the M104A, L105A.

The fact that sector 2 positions did not affect catalysis was surprising to me because structurally, it seems that some positions would have to affect catalysis. M104 and L105 are both near in sequence space as D102, an essential member of the catalytic triad. Even the double mutant, M104A, L105A did not affect catalysis to an extent as some sector 1 positions. Similarly, T229 is in structure space adjacent to the D102 and appears to be in van der Waals contact; yet the T229A mutant alone or in combination with M104A or L105A did not result in great loss of catalytic power. In fact, the data clearly indicate that no sector 2 positions affect catalysis to the extent that most sector 1 positions.

On the other hand what the data show is that sector 2 positions have large effects on fold stability. Specifically many mutations destabilize the enzyme with one mutation, C136A, having no detectable transition. This is especially noteworthy as the positions in sector 1 do not affect fold stability to any appreciable extent. It seems clear then, that the positions in sector 2 affect fold stability but not catalysis whereas positions in sector 1 affect catalysis but not stability. Double mutants in sector 1 also affect fold stability to a larger extent than the single mutants in some cases.

An important test of the independence of sector 1 and sector 2 positions is shown by the data on the double mutants G216A-Q210A and G216A-C157A (C, Figure 116). Both these sector 1/sector 2 double mutants show effects very similar to that expected if they were acting independently. Thus the data strongly suggests that sector 1 has a different and independent function from sector 2.

Also shown in the previous figures is the effect of three non sector positions. Q30A and Y29A have high, outlying coupling values yet they do not affect catalysis or fold stability to a large extent. K230A, adjacent to T229A, also does not affect catalysis or fold stability. This would argue that Q30A and Y29A and K230A do not belong in any of the sectors consistent with the PCA and ICA conclusions.

The meaning of this apparent functional and evolutionary independence will be discussed at the conclusion of this chapter. The next part will describe further analysis of the serine protease family with a new evolution-based method.

Evolutionary based distance analysis of the sectors

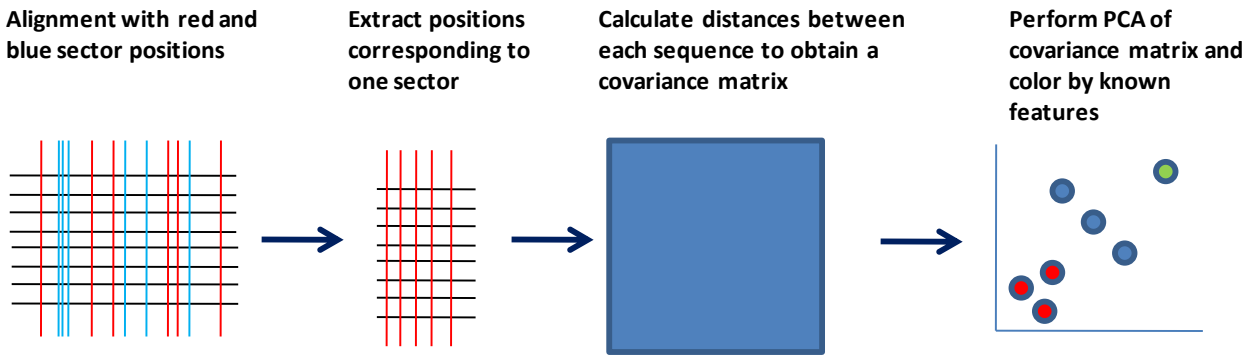
One shortcoming of the experimental data is that they only apply to one member of the serine protease family-rat trypsin. However the serine protease family, as previously discussed, includes many diverse organisms and many functions other than digestion. There are much data in the literature on the function and taxonomy of these proteases that could shed light on the meaning of the sectors if this information could be analyzed.

Following discussions with lab members at one lab meeting (in particular Chris Larson and Madhu Nataranjan), the idea was suggested to perform a PCA analysis on the distance matrix of the sequences but calculating distances based on the identified sectors. Calculating distances between sequences is nothing new and that is what is done whenever an evolutionary tree is made. The use of PCA analysis for sequence distance data is also not completely novel [179]. The usefulness of PCA for sequence data is to group sequences based on some unknown property encoded in their sequence. The novel part of this approach is to calculate a distance matrix considering only positions identified by the sectors. That is, three sectors are observed via the analysis presented before: one sector affects substrate binding, another affects stability and a third affects the catalytic machinery. Each sector can be used to calculate distances and then a PCA analysis can be done on that distance matrix. The idea is to see if separate groupings *within each sector* can be obtained which would make biological sense. This process is illustrated schematically in Figure 117.

In order for this to work there should be a lot of data available with different features as the more sequences with known features the stronger the conclusions are.

To obtain known features I made use of the NCBI database. I extracted for every sequence in the alignment the postulated substrate specificity as well as the taxonomy. I also extracted where in the

Figure 117: Schematic of sector based distance analysis



organism it was expressed (intestine, immune cell, blood, etc) if that information was known as well as what type of function (coagulation, complement, digestion, etc) and calculated the theoretical pI of every sequence.

Thus every sequence could have a feature set made up of substrate specificity, organism, general function, general expression site and isoelectric point. Then for each of the blue and red sectors, I calculated distances using sequence identity metric and then performed a PCA analysis on the distance matrix. I then colored each sequence based on the features above and visually looked for patterns. Two features - substrate specificity and taxonomy showed visible clustering patterns- as depicted in Figure 118 and Figure 119.

What those figures clearly show is that if one calculates distances based on sector 1 (the red/specificity sector) and then does a PCA based on those distances, (the first two components account for ~80% of the variance) one obtains a striking decomposition of sequences based on what specificity they have. Chymotrypsins with their distinct specificity cluster together; trypsins, tryptases and most kallikreins cluster together based on their specificity for charged residues such as lysine and arginines. Furthermore, granzymes which are a family of closely related immune proteases but which have different specificities [158, 180, 181] show different clusters within themselves. On the other hand, coloring the positions by vertebrate/invertebrate reveals no clustering.

The opposite is true for sector 2 (blue/stability sector). This sector shows no clustering based on specificity (with the exception of enzyme classes that exist only in vertebrates-but note that the granzymes have lost their specificity clustering) but a dramatic clustering based on phylogeny. The second sector has some memory of the evolutionary history of the protein but has no knowledge of the specificity.

Finally as Figure 120 shows, calculation of distances based on all sequences does not reveal any specificity or phylogeny clusters.

These results strongly add to the conclusions reached from the experimental work in rat trypsin and the SCA based analysis-that there are functionally and evolutionarily independent sets of positions within this protein family.

Moreover, these results have implications for using proteins to calculate evolutionary relationships. The data reveals that there are different evolutionary pressures on different parts of a proteins. Therefore it is necessary to determine which part of the proteins reflects evolutionary time. It is not clear from this work which sector or nonsector positions should be used but it is clear that different regions report differently on evolution.

Figure 118: PCA of distances calculated based on sector 1 (the red sector or the catalytic sector). A) Sequences are colored based on specificity. B) Sequences are colored based on whether they are vertebrate or invertebrate. C) Sequence motifs made using Weblogo from regions of the plot with squares. Open circles are proteases with unknown specificity.

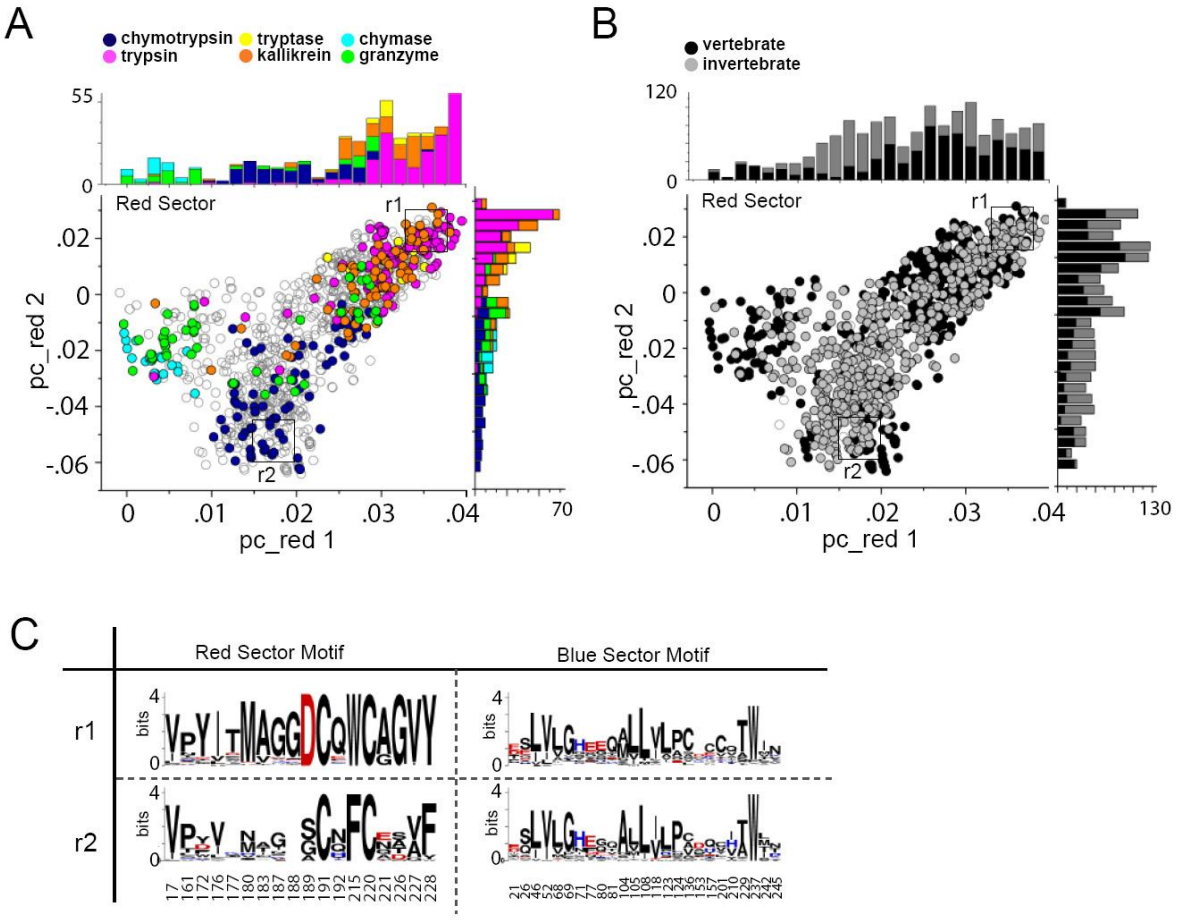


Figure 119: PCA of distances calculated based on sector 2 (the blue sector or the stability sector). A) Sequences are colored based on specificity. B) Sequences are colored based on whether they are vertebrate or invertebrate. C) Sequence motifs made using Weblogo from regions of the plot with squares. Open circles are proteases with unknown specificity.

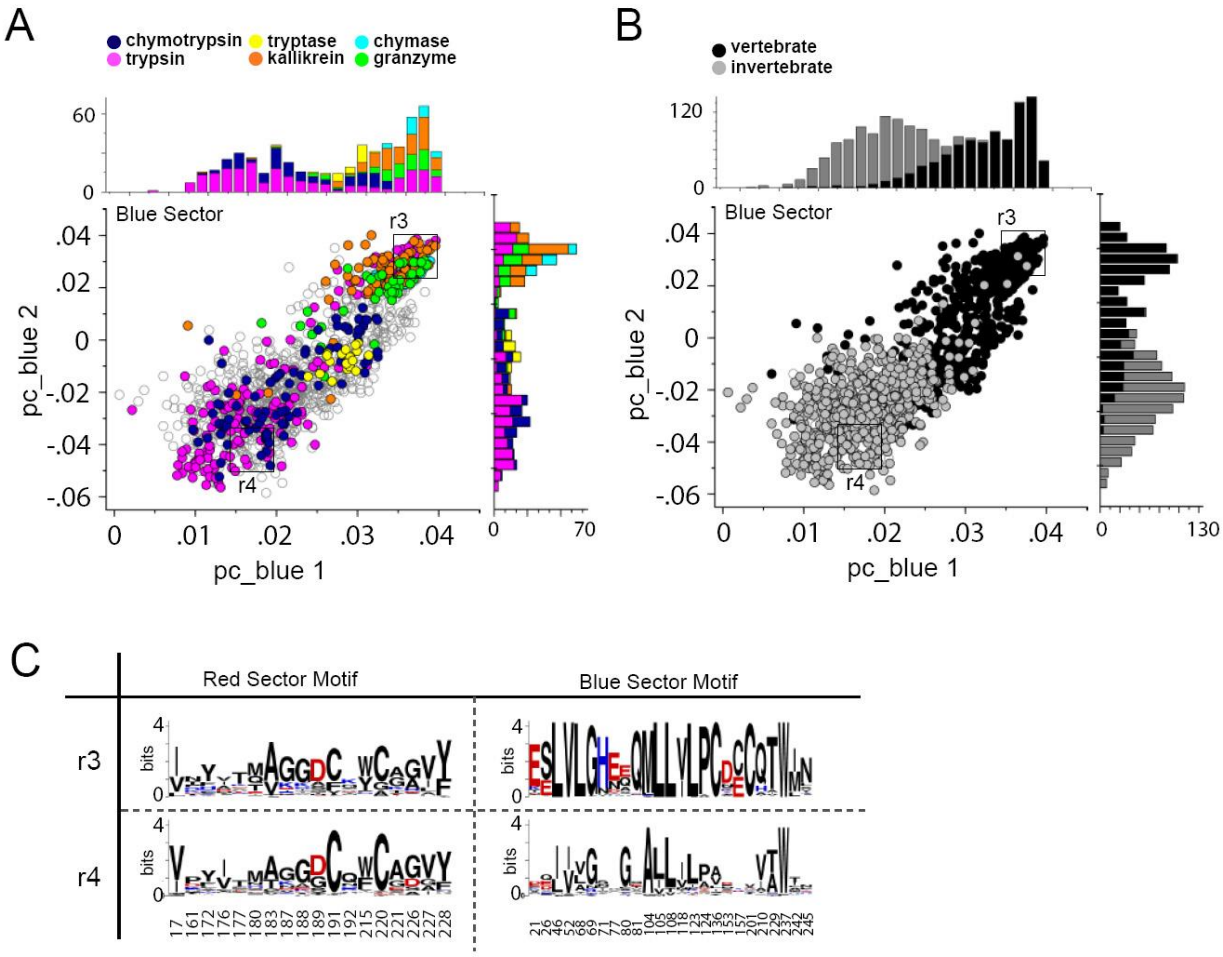
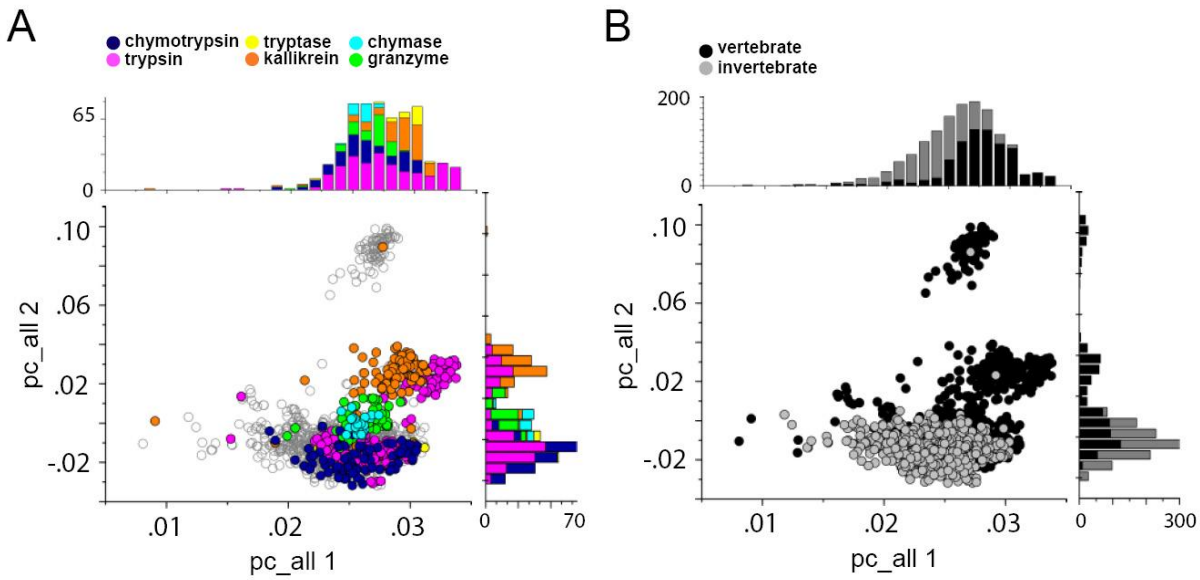


Figure 120: PCA of distances calculated using all positions. A) Sequences are colored according to specificity types. B) Sequences are colored according to whether they come from vertebrate or invertebrate organisms. Open circles are proteases with unknown specificity.



Conclusion:

The evidence presented in this chapter for the serine protease family S1A supports the idea that this family is made up of distinct and cooperative sets of positions that are evolutionarily and functionally independent. The data support the idea that one set encodes substrate specificity, one set encodes fold stability and a third set encodes the catalytic machinery. The evidence for this comes from analysis of the correlations between positions, experimental evidence for the functional independence of positions and analysis of sector based sequence distances.

The work presented here represents a novel decomposition of proteins into functional sectors. One question is why should different cooperative structures evolve in one protein family? As discussed in the introduction, cooperative systems are hard to build and harder to change as mutual constraints need to be fulfilled for function to be retained. A protein family that evolves independent units is likely to find itself in diverse environments as the organism can adapt one member of this family to a new function much faster than evolution of a new protein. In the serine proteases this is likely what occurred. One computational study that is of relevance to this work was done by Uri Alon's group. They showed that evolving computer programs display modular architectures when the target being sought is changing [182]. While proteins are not programs, evolutionary principles identified by this work apply.

Further questions that arose from this work that could form the basis for future studies are:

1. Analysis of family S1B (bacterial homolog of family S1A). It would be interesting to see what changes, if any, occur in the coupling pattern of a system that has evolved independently for 1.5 to 2 billion years.
2. Improving the theoretical approach to extracting correlations between positions. This work used several different ways to understand correlations within a system. However, there are more methods that can perhaps be used.
3. Calculating more accurate evolutionary trees in this protein family and others. A central problem in the extraction of correlations that reflect functional constraints is that the evolutionary history is convoluted with the functional constraints. Developing methods that could more accurately account for evolutionary rates would aid the deconvolution of function from evolutionary history.

4. Design of proteins with different specificities or stabilities. The work presented here suggests that it could be relatively easy to evolve this protein family to have different specificities and/or stabilities. Ongoing experiments by other lab members are aimed at this.
5. Studying physical mechanisms of independence between positions. One of the most challenging projects would be to understand the physical mechanism of coupling between positions. Perhaps one subject to explore is the dynamic nature of the amino acids of this protein family. Another project would be to understand the physical forces that act between residues such that different independent and cooperative units can exist. The next chapter will look at some methods that could be useful in the future to understand the detailed physical basis of protein function.

Chapter 6: Understanding the physical basis of cooperativity

In the first chapter I discussed how cooperativity underlies the function of many proteins and the different ways cooperativity can be observed. In the last four chapters I discussed the use of evolution based methods along with experimental verification to detect cooperative interactions within proteins. All methods, while capable of revealing cooperativity cannot address the mechanism by which atoms or amino acids form units of function. In this chapter, I discuss what possible methods could be used to further increase the understanding of cooperativity.

There are several experimental and theoretical methods that have the potential to allow detailed measurement or calculation of protein interactions. The experimental methods I will discuss are two force measurement techniques and two femtochemistry based approaches. Theoretical methods that could aid in the understanding of cooperativity are coarse-grained models of proteins that can model some cooperative features of proteins.

Experimental Force Methods

Atomic force microscopes were first developed only 22 years ago [183] from scanning tunneling microscopes (STM). In a scanning tunneling microscope a sharp conducting tip is brought very close to a conducting surface upon which electrons tunnel across the gap and induce a current in the detector attached to the conducting tip. The current flow through the tip is a function of the topology and electronic structure of the material being scanned. To circumvent the need for conducting surfaces and to measure extremely small forces, the first AFM used an STM not to measure the surface under investigation but to measure the deflection of a tiny cantilever that contacts the surface. The cantilever is affected by interaction forces with the surface. The original 1986 paper suggested that the maximal force resolution could be on the order of 10^{-18} N and at room temperature on the order of 10^{-15} N. Given that interatomic forces are on the order of 10^{-7} N for ionic bonds and 10^{-11} N for van der Waals forces, measuring interaction forces such as those that occur in proteins is well within the theoretical range of an AFM. In liquids, a measurement as low as 10^{-11} N was reported as water hydrogen bonds were broken [184]. Current methods use a laser focused on a mirror instead of an STM to measure the deflections of the cantilever.

The use of an AFM to measure the formation of a chemical bond was first reported in 2001 [185] when a covalent bond between two silicon atoms was measured. The authors to make this measurement had to operate at close to absolute zero to minimize the thermal noise that prevents precise measurements at higher temperatures. The authors also needed to measure the van der Waals

force to subtract it out from the bond formation force. AFM has also been used to identify closely related atoms based on their force profile [186], to measure the strength of a covalent bond [187], to measure magnetic properties of the system by making the tip magnetic [188], to measure the forces on atoms as they are being manipulated [189], to obtain time-resolved force spectra [190] and to obtain detailed force and energy spectra of fullerenes trapped in nanotubes [191].

Two recent studies from Julio Fernandez's lab from 2007 and 2008 warrant some discussion as the authors apply an AFM to probe catalysis and coevolving residues. In the first study [192], 8 titin domains with disulfide bonds between them were stretched with an AFM and the force of stretching was measured while thioredoxin, a disulfide bond reducing enzyme, was added to the media. Only upon the addition of thioredoxin were specific signals obtained corresponding to the reduction of the disulfide bond. The signals obtained were the lengthening of the polypeptide that could be measured with 0.1 nm accuracy. Furthermore, the authors tested the rate of disulfide bond breakage as the force was increased. They observed that there were two distinct responses to the force: increasing the force decreased the rate of catalysis up to a certain point; further increases in force increased the catalytic rate. They tested several force dependent catalytic mechanisms and found one model that best fit their data. They proposed a model for thioredoxin catalysis that consists of two paths; one path involves substrate shortening and another path substrate elongation. They further supported their model with mutational studies, molecular dynamics and available structural information. I think this work was interesting because it was the first time that force was used on a substrate to understand the mechanism of the enzyme.

The second study [193] follows up on the first but in this paper the authors mutated coevolving positions within the thioredoxin to see whether there is a relationship between the mutations of co-evolving residues and the mechanism. The authors had previously observed that a mutant has disrupted mechanism of catalysis. They carried out a coupling analysis and found the residue that was most correlated to the mutated position and the residue that was most anticorrelated. Both these residues were more than 10 angstroms away from the mutated one. They then made a double mutant enzyme with both the negatively correlated and positively correlated residues mutated. They then looked with the force experiments discussed previously at the mechanism of catalysis. They found that negatively correlated residues moved the enzyme to one with non-wild-type mechanism whereas positively correlated residues moved the enzyme to one with a wild-type mechanism. This paper is just as interesting as the previous one because it continues to try to understand in detail the mechanism of catalysis and how it is affected by mutations of co-evolving and therefore cooperative residues.

As the studies from the Fernandez lab show and the previously mentioned studies, AFM is a very versatile tool. The experiments that would be most interesting from the viewpoint in understanding the forces during protein function is if one can examine the force on a substrate as it undergoes catalysis or to measure the forces that act on a ligand as it approaches to and binds to its target receptor. One can also potentially measure the dynamical features of proteins if one can control the thermal noise on the tip. One experiment that would be interesting is to see if one could measure with an AFM the dynamic motions below and above the transition temperature of protein function (above and below 180 K). These experiments and others (see Carlos Bustamante's review [194] would provide much more mechanistic data about protein interactions that are available with current methods.

Optical trapping is another method, also developed in 1986, that can be used to apply and measure forces [195]. It works via a completely different principle than AFM however. In an optical trap a beam of light from a laser (typically infrared light for biological systems) is focused onto the smallest possible spot. A dielectric particle within that spot will encounter a force that steers it towards the center of the spot. The dielectric particle can then be moved by moving the light source. The dielectric particle can also be detected using the light scattered from the trapped particle and interferometry methods [196]. Optical traps have an advantage over AFM in liquid solution because lower forces, down to 0.1 pN can be applied by manipulating the laser. Optical traps have been most often used to understand the folding and function of DNA and RNA folding and function [196] as proteins are more easily studied with AFM methods. However, optical trapping could be useful if a small substrate or ligand can be bound to a particle upon which the forces acting on that substrate or ligand during biological processes could be measured.

Experimental femtochemistry methods

Other experimental techniques distinct from force measurements are those that use femtochemistry methods to probe reactions [197]. The utility of these is that they can be used to directly measure the potential of reactions by observing bond formation and bond breakage as the reaction occurs. This is possible via two major breakthroughs: the ability to synchronize multiple streams of atoms via quantum coherence and by being able to initiate the chemical reaction and then the detecting 'strobe' pulse with femtosecond light pulses [198].

One notable application to protein biology is the examination of ligand binding to human serum albumin [199]. The authors were able to follow the binding event by using a fluorescent ligand and a fluorescent tryptophan at the binding site. They could monitor the ligand and the tryptophan at different wavelengths. They observed via femtosecond resolved anisotropy that the ligand shows no

diffusive motion when bound up to 500 ps suggesting tight binding and a rigid binding pocket. This is in contrast to ligand motion in solvent which reorients itself with a half life of 45 ps. They were also able to observe the dynamics of the binding process and state that the protein can guide specific proton motions so that binding is efficient.

The Zewail lab has also used femtochemistry to probe the folding of DNA structures [200], water dynamics [201] and light activation of the photoactive yellow protein chromophore [202]. Continued development of these methods will lead to observing protein processes at their most basic level.

Zewail's lab has also, given their unique expertise in femtochemistry methods, adapted femtochemistry methods for electron microscopy [203-205]. Electron microscopes such as transmission electron microscopes are familiar to most biologists as they are used to image fixed cellular structures to nanometer resolutions. Zewail's group have built electron microscopes that can take femtosecond time-resolved images [206]. The rate of progress of these methods has been very fast leading in recent years to several publications [207-212]. Electron microscopy is also integrated with electron diffraction methods which are able to achieve atomic level resolution as x-ray diffraction [213-215]. Although 4D electron microscopy has not yet been applied to biological problems, the potential of these methods once they are used in biological systems is that they will be able to resolve protein dynamics at the limit of bond formation. That would be as revolutionary as the invention of microscopes that allowed the visualization of cells.

Coarse-grained protein models

The previous methods that were presented were all experimental methods that promise to greatly increase our understanding of protein structure, function and energetics. In parallel to all these experimental methods has been much theoretical work to model protein function. Generally it is not possible to accurately calculate the interactions among all atoms starting from quantum mechanical principles. It becomes possible only with certain approximations the validity of which is questionable for protein systems. In addition, molecular dynamics simulations, due to computational constraints, cannot be carried out for the long periods of milliseconds where interesting biological functions such as catalysis or slow dynamics occur. However, there are coarse grained models that treat the protein not as atoms but as nodes with a certain connectivity. The remarkable aspect of these models is that they can predict some aspects of protein functions such as the B-factors that occur in proteins, aspects of fold stability and low frequency motions. If what is thought of as complex protein function can indeed be reduced in complexity, then it is possible that theoretical methods would allow one to understand and

therefore design arbitrary protein sequences soon. Of course, all atom molecular dynamics simulations would achieve the same goal but are limited by computational cost.

The first indications that it is not necessary to have all-atom models of proteins came from the work of Ariel Warshel and Andrew Levitt who modeled a protein as one side chain node and one main chain node [216], Ken Dill who developed a polymer based statistical mechanical theory of protein folding [217] and Bryngelson and Wolynes who treated proteins like spin glasses [218]. These authors showed that protein folding could be approximated by simplified representations. It is interesting that all three used different representations suggesting strongly that it is not difficult to make proteins that fold. Today, with experimentally verified predictions of coarse-grained folding models (reviewed in [27, 219-221]) there is agreement that proteins can be represented as coarse-grained structures that will fold. The essential characteristic seems to be the hydrophobic code of proteins in that the information in side chains can be reduced to hydrophobic or not hydrophobic.

Another type of coarse grained model that has been used is models for slow dynamics of proteins. These models are known as Gaussian network models or elastic network models. These models were developed to simplify the complicated and time consuming calculations of molecular dynamics for the purpose of calculating low frequency collective motions, known as normal modes. Normal mode analysis is basically a principal components analysis of a set of frequency dependent trajectories of atoms. The first normal mode describes the residues that move together at the lowest frequency. The second normal mode describes the residues that move together at a higher frequency and so on. The interesting feature of normal mode analysis is that the low frequency modes correlate with *biologically relevant* slow collective dynamics of proteins such as hinge motions or structure factors [222]. However, molecular dynamics are computationally difficult to so simplifications were attempted and proved successful.

The original study reported in 1996 by Monique Tirion [223]and further developed in 1997 by Bahar and colleagues [224, 225] showed that one can replace the complex potentials of molecular dynamics with a simple terms where interactions between atoms are treated as springs or Gaussian distributed fluctuations. A cutoff is also chosen so that only nearby atoms are connected and all interactions between all atoms are treated in the same way. The surprising results reported by Tirion and Bahar is that the low frequency normal mode analysis of these simplified potentials predicted the same low frequency motions as the normal mode analysis. That means that it is possible using a much simplified view of proteins to model cooperative behavior such as low frequency motions.

Both the Tirion and Bahar work calculate magnitude of motions without considering direction. Bahar's group reported in 2001 a model called the anisotropic network models that in addition to magnitude of fluctuations calculates a directionality that can be more biologically relevant [226]. The results from this work provided no additional insights into biology and showed that the calculation of low frequency collective models is insensitive to the exact model used.

The observation that normal mode analysis of a simplified model of interactions of proteins can model biologically relevant low frequency cooperative modes is worth pursuing further.

In conclusion, both slow scale dynamics and folding of proteins can be modeled with non-unique simplified expressions indicating that for important, collective motions, the details of interactions are not important. This makes the theoretical investigations useful as these models are not computationally expensive and their simplicity makes it easier to understand their features and to make predictions that one can test experimentally.

Conclusion

Force measurements once applied to the direct measurement of interaction forces in proteins will enable exact quantification of hydrogen bonds, ionic forces, van der Waals forces and other interactions revealing much that is now speculation. Femtochemistry based methods are potentially revolutionary because they enable the observation of chemical events on the time-scale of bond formation and can map chemical processes as they happen. Finally, it is possible that before force or femtochemical methods become widely used, the cooperative interactions within proteins could be at least partially understood with simpler models of proteins.

It is likely that the analysis of protein cooperativity will require the contribution of multiple methods before it is possible to understand proteins well enough such that one could design a protein sequence to carry out an arbitrary function.

References

1. Dill KA: **Additivity principles in biochemistry.** *J Biol Chem* 1997, **272:**701-704.
2. Stroppolo ME, Falconi M, Caccuri AM, Desideri A: **Superefficient enzymes.** *Cell Mol Life Sci* 2001, **58:**1451-1460.
3. Wolfenden R: **Thermodynamic and extrathermodynamic requirements of enzyme catalysis.** *Biophys Chem* 2003, **105:**559-572.
4. Hanson CV, Nishiyama Y, Paul S: **Catalytic antibodies and their applications.** *Curr Opin Biotechnol* 2005, **16:**631-636.
5. Kraut J: **How do enzymes work?** *Science* 1988, **242:**533-540.
6. Wolfenden R: **Transition state analogues for enzyme catalysis.** *Nature* 1969, **223:**704-705.
7. Green NM: **Avidin.** *Adv Protein Chem* 1975, **29:**85-133.
8. Weber PC, Ohlendorf DH, Wendoloski JJ, Salemme FR: **Structural origins of high-affinity biotin binding to streptavidin.** *Science* 1989, **243:**85-88.
9. Clackson T, Wells JA: **A hot spot of binding energy in a hormone-receptor interface.** *Science* 1995, **267:**383-386.
10. Moreira IS, Fernandes PA, Ramos MJ: **Hot spots--a review of the protein-protein interface determinant amino-acid residues.** *Proteins* 2007, **68:**803-812.
11. Creighton TE: *Proteins: Structures and Molecular Properties.* W. H. Freeman; 1992.
12. Monod J, Jacob F: **Teleonomic mechanisms in cellular metabolism, growth, and differentiation.** *Cold Spring Harb Symp Quant Biol* 1961, **26:**389-401.
13. Monod J, Wyman J, Changeux JP: **On the Nature of Allosteric Transitions: A Plausible Model.** *J Mol Biol* 1965, **12:**88-118.
14. Koshland DE, Jr., Nemethy G, Filmer D: **Comparison of experimental binding data and theoretical models in proteins containing subunits.** *Biochemistry* 1966, **5:**365-385.
15. Cui Q, Karplus M: **Allostery and cooperativity revisited.** *Protein Sci* 2008, **17:**1295-1307.
16. Anfinsen CB: **Principles that govern the folding of protein chains.** *Science (New York, NY)* 1973, **181:**223-230.
17. Jaswal SS, Sohl JL, Davis JH, Agard DA: **Energetic landscape of alpha-lytic protease optimizes longevity through kinetic stability.** *Nature* 2002, **415:**343-346.
18. Levinthal C, Levinthal C: **Are there pathways for protein folding?** *Journal de Chimie Physique et de Physico-Chimie Biologique* 1968, **65:**44?45-44?45.
19. Kubelka J, Hofrichter J, Eaton WA: **The protein folding 'speed limit'.** *Curr Opin Struct Biol* 2004, **14:**76-88.
20. Dill KA, Chan HS: **From Levinthal to pathways to funnels.** *Nat Struct Biol* 1997, **4:**10-19.
21. Oliveberg M, Wolynes PG: **The experimental survey of protein-folding energy landscapes.** *Q Rev Biophys* 2005, **38:**245-288.
22. Klapper MH: **On the nature of the protein interior.** *Biochim Biophys Acta* 1971, **229:**557-566.
23. Richards FM: **Areas, volumes, packing and protein structure.** *Annu Rev Biophys Bioeng* 1977, **6:**151-176.
24. Richards FM, Richmond T: **Solvents, interfaces and protein structure.** *Ciba Found Symp* 1977**:**23-45.
25. Liang J, Dill KA: **Are proteins well-packed?** *Biophysical Journal* 2001, **81:**751-766.
26. Dill KA: **Dominant forces in protein folding.** *Biochemistry* 1990, **29:**7133-7155.
27. Dill KA, Ozkan SB, Shell MS, Weikl TR: **The protein folding problem.** *Annual Review of Biophysics* 2008, **37:**289-316.

28. Chan HS, Bromberg S, Dill KA: **Models of cooperativity in protein folding.** *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences* 1995, **348:**61-70.
29. Dill KA, Fiebig KM, Chan HS: **Cooperativity in protein-folding kinetics.** *Proceedings of the National Academy of Sciences of the United States of America* 1993, **90:**1942-1946.
30. Privalov PL: **Stability of proteins: small globular proteins.** *Adv Protein Chem* 1979, **33:**167-241.
31. Privalov PL, Khechinashvili NN: **A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study.** *J Mol Biol* 1974, **86:**665-684.
32. Privalov PL, Tsalkova TN: **Micro- and macro-stabilities of globular proteins.** *Nature* 1979, **280:**693-696.
33. Feynman RP, Leighton RB, Sands ML: *Six Easy Pieces.* 1995.
34. Doster W, Cusack S, Petry W: **Dynamical transition of myoglobin revealed by inelastic neutron scattering.** *Nature* 1989, **337:**754-756.
35. Iben IE, Braunstein D, Doster W, Frauenfelder H, Hong MK, Johnson JB, Luck S, Ormos P, Schulte A, Steinbach PJ, et al: **Glassy behavior of a protein.** *Phys Rev Lett* 1989, **62:**1916-1919.
36. Ferrand M, Dianoux AJ, Petry W, Zaccai G: **Thermal motions and function of bacteriorhodopsin in purple membranes: effects of temperature and hydration studied by neutron scattering.** *Proc Natl Acad Sci U S A* 1993, **90:**9668-9672.
37. Rasmussen BF, Stock AM, Ringe D, Petsko GA: **Crystalline ribonuclease A loses function below the dynamical transition at 220 K.** *Nature* 1992, **357:**423-424.
38. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D: **Intrinsic dynamics of an enzyme underlies catalysis.** *Nature* 2005, **438:**117-121.
39. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D: **A hierarchy of timescales in protein dynamics is linked to enzyme catalysis.** *Nature* 2007, **450:**913-916.
40. Henzler-Wildman KA, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, Pozharski E, Wilson MA, Petsko GA, Karplus M, et al: **Intrinsic motions along an enzymatic reaction trajectory.** *Nature* 2007, **450:**838-844.
41. Labeikovsky W, Eisenmesser EZ, Bosco DA, Kern D: **Structure and dynamics of pin1 during catalysis by NMR.** *J Mol Biol* 2007, **367:**1370-1381.
42. Boehr DD, McElheny D, Dyson HJ, Wright PE: **The dynamic energy landscape of dihydrofolate reductase catalysis.** *Science* 2006, **313:**1638-1642.
43. Frederick KK, Marlow MS, Valentine KG, Wand AJ: **Conformational entropy in molecular recognition by proteins.** *Nature* 2007, **448:**325-329.
44. Balog E, Becker T, Oettl M, Lechner R, Daniel R, Finney J, Smith JC: **Direct determination of vibrational density of states change on ligand binding to a protein.** *Phys Rev Lett* 2004, **93:**028103.
45. Horovitz A, Fersht AR: **Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins.** *Journal of Molecular Biology* 1990, **214:**613-617.
46. Carter PJ, Winter G, Wilkinson AJ, Fersht AR: **The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus).** *Cell* 1984, **38:**835-840.
47. Horovitz A: **Double-mutant cycles: a powerful tool for analyzing protein structure and function.** *Folding & Design* 1996, **1:**R121-126-R121-126.
48. Horovitz A: **Non-additivity in protein-protein interactions.** *Journal of Molecular Biology* 1987, **196:**733-735.
49. Mark AE, van Gunsteren WF: **Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies.** *J Mol Biol* 1994, **240:**167-176.
50. Wells JA: **Additivity of mutational effects in proteins.** *Biochemistry* 1990, **29:**8509-8517.

51. LiCata VJ, Ackers GK: **Long-range, small magnitude nonadditivity of mutational effects in proteins.** *Biochemistry* 1995, **34:**3133-3139.
52. Zandany N, Ovadia M, Orr I, Yifrach O: **Direct analysis of cooperativity in multisubunit allosteric proteins.** *Proc Natl Acad Sci U S A* 2008, **105:**11697-11702.
53. Sadovsky E, Yifrach O: **Principles underlying energetic coupling along an allosteric communication trajectory of a voltage-activated K+ channel.** *Proc Natl Acad Sci U S A* 2007, **104:**19813-19818.
54. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36:**D281-288.
55. Ferguson AD, Amezcua CA, Halabi NM, Chelliah Y, Rosen MK, Ranganathan R, Deisenhofer J: **Signal transduction pathway of TonB-dependent transporters.** *Proc Natl Acad Sci U S A* 2007, **104:**513-518.
56. Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R: **Allosteric determinants in guanine nucleotide-binding proteins.** *Proc Natl Acad Sci U S A* 2003, **100:**14445-14450.
57. Suel GM, Lockless SW, Wall MA, Ranganathan R: **Evolutionarily conserved networks of residues mediate allosteric communication in proteins.** *Nat Struct Biol* 2003, **10:**59-69.
58. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286:**295-299.
59. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R: **Evolutionary information for specifying a protein fold.** *Nature* 2005, **437:**512-518.
60. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R: **Natural-like function in artificial WW domains.** *Nature* 2005, **437:**579-583.
61. Atkins PW: *Physical Chemistry.* 6 edn. Oxford: Oxford University Press; 1998.
62. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R: **Structural determinants of allosteric ligand activation in RXR heterodimers.** *Cell* 2004, **116:**417-429.
63. Crowson RA: **A systematist looks at cytochrome c.** *J Mol Evol* 1972, **2:**28-37.
64. Wong AK, Liu TS, Wang CC: **Statistical analysis of residue variability in cytochrome c.** *J Mol Biol* 1976, **102:**287-295.
65. Altschuh D, Lesk AM, Bloomer AC, Klug A: **Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus.** *J Mol Biol* 1987, **193:**693-707.
66. Altschuh D, Vernet T, Berti P, Moras D, Nagai K: **Coordinated amino acid changes in homologous protein families.** *Protein Eng* 1988, **2:**193-199.
67. Vernet T, Tessier DC, Khouri HE, Altschuh D: **Correlation of co-ordinated amino acid changes at the two-domain interface of cysteine proteases with protein stability.** *J Mol Biol* 1992, **224:**501-509.
68. Korber BT, Farber RM, Wolpert DH, Lapedes AS: **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* 1993, **90:**7176-7180.
69. Neher E: **How frequent are correlated changes in families of protein sequences?** *Proc Natl Acad Sci U S A* 1994, **91:**98-102.
70. Gobel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins.** *Proteins* 1994, **18:**309-317.
71. Taylor WR, Hatrick K: **Compensating changes in protein multiple sequence alignments.** *Protein Eng* 1994, **7:**341-348.
72. Hatrick K, Taylor WR: **Sequence conservation and correlation measures in protein structure prediction.** *Comput Chem* 1994, **18:**245-249.

73. Kovalenko O, Yifrach O, Horovitz A: **Residue lysine-34 in GroES modulates allosteric transitions in GroEL.** *Biochemistry* 1994, **33:**14974-14978.
74. Clarke ND: **Covariation of residues in the homeodomain sequence family.** *Protein Sci* 1995, **4:**2269-2278.
75. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257:**342-358.
76. Thomas DJ, Casari G, Sander C: **The prediction of protein contacts from multiple sequence alignments.** *Protein Eng* 1996, **9:**941-948.
77. Pollock DD, Taylor WR: **Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution.** *Protein Engineering* 1997, **10:**647-657.
78. Chelvanayagam G, Eggenschwiler A, Knecht L, Gonnet GH, Benner SA: **An analysis of simultaneous variation in protein structures.** *Protein Eng* 1997, **10:**307-316.
79. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A: **Correlated mutations contain information about protein-protein interaction.** *J Mol Biol* 1997, **271:**511-523.
80. Olmea O, Valencia A: **Improving contact predictions by the combination of correlated mutations and other sources of sequence information.** *Fold Des* 1997, **2:**S25-32.
81. Ortiz AR, Kolinski A, Skolnick J: **Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations.** *Proc Natl Acad Sci U S A* 1998, **95:**1020-1025.
82. Pollock DD, Taylor WR, Goldman N: **Coevolving protein residues: maximum likelihood identification and relationship to structure.** *Journal of Molecular Biology* 1999, **287:**187-198.
83. Olmea O, Rost B, Valencia A: **Effective use of sequence correlation and conservation in fold recognition.** *J Mol Biol* 1999, **293:**1221-1239.
84. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J: **Ab initio folding of proteins using restraints derived from evolutionary information.** *Proteins* 1999, **Suppl 3:**177-185.
85. Fariselli P, Casadio R: **A neural network based predictor of residue contacts in proteins.** *Protein Eng* 1999, **12:**15-21.
86. Tuff P, Darlu P: **Exploring a phylogenetic approach for the detection of correlated substitutions in proteins.** *Mol Biol Evol* 2000, **17:**1753-1759.
87. Wollenberg KR, Atchley WR: **Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap.** *Proc Natl Acad Sci U S A* 2000, **97:**3288-3291.
88. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.** *Mol Biol Evol* 2000, **17:**164-178.
89. Fariselli P, Olmea O, Valencia A, Casadio R: **Prediction of contact maps with neural networks and correlated mutations.** *Protein Eng* 2001, **14:**835-843.
90. Pritchard L, Bladon P, J MOM, M JD: **Evaluation of a novel method for the identification of coevolving protein residues.** *Protein Eng* 2001, **14:**549-555.
91. Filizola M, Olmea O, Weinstein H: **Prediction of heterodimerization interfaces of G-protein coupled receptors with a new subtractive correlated mutation method.** *Protein Eng* 2002, **15:**881-885.
92. Oliveira L, Paiva ACM, Vriend G: **Correlated mutation analyses on very large sequence families.** *Chembiochem: A European Journal of Chemical Biology* 2002, **3:**1010-1017.
93. Kass I, Horovitz A: **Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations.** *Proteins* 2002, **48:**611-617.
94. Tillier ER, Lui TW: **Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments.** *Bioinformatics* 2003, **19:**750-755.

95. Saraf MC, Moore GL, Maranas CD: **Using multiple sequence correlation analysis to characterize functionally important protein regions.** *Protein Engineering* 2003, **16:**397-406.
96. Fodor AA, Aldrich RW: **On evolutionary conservation of thermodynamic coupling in proteins.** *The Journal of Biological Chemistry* 2004, **279:**19046-19050.
97. Fodor AA, Aldrich RW: **Influence of conservation on calculations of amino acid covariance in multiple sequence alignments.** *Proteins* 2004, **56:**211-221.
98. Dekker JP, Fodor A, Aldrich RW, Yellen G: **A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments.** *Bioinformatics* 2004, **20:**1565-1572.
99. Fleishman SJ, Yifrach O, Ben-Tal N: **An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels.** *J Mol Biol* 2004, **340:**307-318.
100. Dutheil J, Pupko T, Jean-Marie A, Galtier N: **A model-based approach for detecting coevolving positions in a molecule.** *Mol Biol Evol* 2005, **22:**1919-1928.
101. Gloor GB, Martin LC, Wahl LM, Dunn SD: **Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions.** *Biochemistry* 2005, **44:**7156-7165.
102. Noivirt O, Eisenstein M, Horovitz A: **Detection and reduction of evolutionary noise in correlated mutation analysis.** *Protein Eng Des Sel* 2005, **18:**247-253.
103. Fares MA, Travers SA: **A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses.** *Genetics* 2006, **173:**9-23.
104. Halperin I, Wolfson H, Nussinov R: **Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families.** *Proteins* 2006, **63:**832-845.
105. Kundrotas PJ, Alexov EG: **Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives.** *BMC Bioinformatics* 2006, **7:**503.
106. Fuchs A, Martin-Galiano AJ, Kalman M, Fleishman S, Ben-Tal N, Frishman D: **Co-evolving residues in membrane proteins.** *Bioinformatics* 2007, **23:**3312-3319.
107. Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.** *Bioinformatics* 2008, **24:**333-340.
108. Sayar K, Ugur O, Liu T, Hilser VJ, Onaran O: **Exploring allosteric coupling in the alpha-subunit of heterotrimeric G proteins using evolutionary and ensemble-based approaches.** *BMC Struct Biol* 2008, **8:**23.
109. Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT: **Rewiring the specificity of two-component signal transduction systems.** *Cell* 2008, **133:**1043-1054.
110. Ortiz AR, Kolinski A, Skolnick J: **Tertiary structure prediction of the KIX domain of CBP using Monte Carlo simulations driven by restraints derived from multiple sequence alignments.** *Proteins* 1998, **30:**287-294.
111. Ortiz AR, Kolinski A, Skolnick J: **Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments.** *J Mol Biol* 1998, **277:**419-448.
112. Fariselli P, Olmea O, Valencia A, Casadio R: **Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations.** *Proteins* 2001, **Suppl 5:**157-162.
113. Benner SA, Gerloff D: **Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases.** *Adv Enzyme Regul* 1991, **31:**121-181.
114. Benner SA: **Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure.** *Adv Enzyme Regul* 1989, **28:**219-236.
115. Pearson K: **On lines and planes of closest fit to systems of points in space.** *Philosophical Magazine* 1901, **2:**559-572.
116. Strang G: *Introduction to Linear Algebra, Third Edition.* Wellesley Cambridge Pr; 2003.

117. Stone JV: **Independent Component Analysis.** In *Encyclopedia of Statistics in Behavioral Science* (Everitt BSaH, David C. ed. Chichester: John Wiley and Sons, Ltd.; 2005.
118. Stone JV: *Independent Component Analysis: A Tutorial Introduction.* The MIT Press; 2004.
119. Hyvarinen A, Oja E: **Independent component analysis: algorithms and applications.** *Neural Netw* 2000, **13:**411-430.
120. Ferguson AD, Deisenhofer J: **Metal import through microbial membranes.** *Cell* 2004, **116:**15-24.
121. Ferguson AD, Chakraborty R, Smith BS, Esser L, van der Helm D, Deisenhofer J: **Structural basis of gating by the outer membrane transporter FecA.** *Science* 2002, **295:**1715-1719.
122. Pawelek PD, Croteau N, Ng-Thow-Hing C, Khursigara CM, Moiseeva N, Allaire M, Coulton JW: **Structure of TonB in complex with FhuA, E. coli outer membrane receptor.** *Science* 2006, **312:**1399-1402.
123. Shultis DD, Purdy MD, Banchs CN, Wiener MC: **Outer membrane active transport: structure of the BtuB:TonB complex.** *Science* 2006, **312:**1396-1399.
124. Locher KP, Rees B, Koebnik R, Mitschler A, Moulinier L, Rosenbusch JP, Moras D: **Transmembrane signaling across the ligand-gated FhuA receptor: crystal structures of free and ferrichrome-bound states reveal allosteric changes.** *Cell* 1998, **95:**771-778.
125. Buchanan SK, Smith BS, Venkatramani L, Xia D, Esser L, Palnitkar M, Chakraborty R, van der Helm D, Deisenhofer J: **Crystal structure of the outer membrane active transporter FepA from Escherichia coli.** *Nat Struct Biol* 1999, **6:**56-63.
126. Chimento DP, Mohanty AK, Kadner RJ, Wiener MC: **Substrate-induced transmembrane signaling in the cobalamin transporter BtuB.** *Nat Struct Biol* 2003, **10:**394-401.
127. Wang Y, Geer LY, Chappey C, Kans JA, Bryant SH: **Cn3D: sequence and structure views for Entrez.** *Trends Biochem Sci* 2000, **25:**300-302.
128. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
129. Cobessi D, Celia H, Folschweiller N, Schalk IJ, Abdallah MA, Pattus F: **The crystal structure of the pyoverdine outer membrane receptor FpvA from Pseudomonas aeruginosa at 3.6 angstroms resolution.** *J Mol Biol* 2005, **347:**121-134.
130. Cobessi D, Celia H, Pattus F: **Crystal structure at high resolution of ferric-pyochelin and its membrane receptor FptA from Pseudomonas aeruginosa.** *J Mol Biol* 2005, **352:**893-904.
131. Rawlings ND, Morton FR, Barrett AJ: **MEROPS: the peptidase database.** *Nucleic Acids Res* 2006, **34:**D270-272.
132. Hedstrom L: **Serine protease mechanism and specificity.** *Chem Rev* 2002, **102:**4501-4524.
133. Hung SH, Hedstrom L: **Converting trypsin to elastase: substitution of the S1 site and adjacent loops reconstitutes esterase specificity but not amidase activity.** *Protein Eng* 1998, **11:**669-673.
134. Jelinek B, Antal J, Venekei I, Graf L: **Ala226 to Gly and Ser189 to Asp mutations convert rat chymotrypsin B to a trypsin-like protease.** *Protein Eng Des Sel* 2004, **17:**127-131.
135. Hedstrom L, Perona JJ, Rutter WJ: **Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant.** *Biochemistry* 1994, **33:**8757-8763.
136. Hedstrom L, Szilagyi L, Rutter WJ: **Converting trypsin to chymotrypsin: the role of surface loops.** *Science* 1992, **255:**1249-1253.
137. Di Cera E: **Thrombin: a paradigm for enzymes allosterically activated by monovalent cations.** *C R Biol* 2004, **327:**1065-1076.
138. Guinto ER, Caccia S, Rose T, Futterer K, Waksman G, Di Cera E: **Unexpected crucial role of residue 225 in serine proteases.** *Proc Natl Acad Sci U S A* 1999, **96:**1852-1857.
139. Gandhi PS, Chen Z, Mathews FS, Di Cera E: **Structural identification of the pathway of long-range communication in an allosteric enzyme.** *Proc Natl Acad Sci U S A* 2008, **105:**1832-1837.

140. Sherwood MW, Prior IA, Voronina SG, Barrow SL, Woodsmith JD, Gerasimenko OV, Petersen OH, Tepikin AV: **Activation of trypsinogen in large endocytic vacuoles of pancreatic acinar cells.** *Proc Natl Acad Sci U S A* 2007, **104:**5674-5679.
141. Waterford SD, Kolodecik TR, Thrower EC, Gorelick FS: **Vacuolar ATPase regulates zymogen activation in pancreatic acini.** *J Biol Chem* 2005, **280:**5430-5434.
142. Page MJ, Di Cera E: **Serine peptidases: classification, structure and function.** *Cell Mol Life Sci* 2008, **65:**1220-1236.
143. Rawlings ND, Morton FR, Kok CY, Kong J, Barrett AJ: **MEROPS: the peptidase database.** *Nucleic Acids Res* 2008, **36:**D320-325.
144. Dodson G, Wlodawer A: **Catalytic triads and their relatives.** *Trends Biochem Sci* 1998, **23:**347-352.
145. Polgar L: **The catalytic triad of serine peptidases.** *Cell Mol Life Sci* 2005, **62:**2161-2172.
146. Kurosky A, Barnett DR, Lee TH, Touchstone B, Hay RE, Arnott MS, Bowman BH, Fitch WM: **Covalent structure of human haptoglobin: a serine protease homolog.** *Proc Natl Acad Sci U S A* 1980, **77:**3388-3392.
147. Bowman BH, Kurosky A: **Haptoglobin: the evolutionary product of duplication, unequal crossing over, and point mutation.** *Adv Hum Genet* 1982, **12:**189-261, 453-184.
148. Radisky ES, Lee JM, Lu CJ, Koshland DE, Jr.: **Insights into the serine protease mechanism from atomic resolution structures of trypsin reaction intermediates.** *Proc Natl Acad Sci U S A* 2006, **103:**6835-6840.
149. Hedstrom L, Farr-Jones S, Kettner CA, Rutter WJ: **Converting trypsin to chymotrypsin: ground-state binding does not determine substrate specificity.** *Biochemistry* 1994, **33:**8764-8769.
150. Venekei I, Szilagyi L, Graf L, Rutter WJ: **Attempts to convert chymotrypsin to trypsin.** *FEBS Lett* 1996, **379:**143-147.
151. Graf L, Jancso A, Szilagyi L, Hegyi G, Pinter K, Naray-Szabo G, Hepp J, Medzihradszky K, Rutter WJ: **Electrostatic complementarity within the substrate-binding pocket of trypsin.** *Proc Natl Acad Sci U S A* 1988, **85:**4961-4965.
152. Craik CS, Largman C, Fletcher T, Roczniak S, Barr PJ, Fletterick R, Rutter WJ: **Redesigning trypsin: alteration of substrate specificity.** *Science* 1985, **228:**291-297.
153. Graf L, Craik CS, Patthy A, Roczniak S, Fletterick RJ, Rutter WJ: **Selective alteration of substrate specificity by replacement of aspartic acid-189 with lysine in the binding pocket of trypsin.** *Biochemistry* 1987, **26:**2616-2623.
154. Schechter I, Berger A: **On the size of the active site in proteases. I. Papain.** *Biochem Biophys Res Commun* 1967, **27:**157-162.
155. Perona JJ, Hedstrom L, Rutter WJ, Fletterick RJ: **Structural origins of substrate discrimination in trypsin and chymotrypsin.** *Biochemistry* 1995, **34:**1489-1499.
156. Vindigni A, Dang QD, Di Cera E: **Site-specific dissection of substrate recognition by thrombin.** *Nat Biotechnol* 1997, **15:**891-895.
157. Coombs GS, Bergstrom RC, Pellequer JL, Baker SI, Navre M, Smith MM, Tainer JA, Madison EL, Corey DR: **Substrate specificity of prostate-specific antigen (PSA).** *Chem Biol* 1998, **5:**475-488.
158. Ruggles SW, Fletterick RJ, Craik CS: **Characterization of structural determinants of granzyme B reveals potent mediators of extended substrate specificity.** *J Biol Chem* 2004, **279:**30751-30759.
159. Huntington JA: **How Na+ activates thrombin--a review of the functional and structural data.** *Biol Chem* 2008, **389:**1025-1035.
160. Huntington JA, Esmon CT: **The molecular basis of thrombin allostery revealed by a 1.8 A structure of the "slow" form.** *Structure* 2003, **11:**469-479.
161. Dang QD, Di Cera E: **Residue 225 determines the Na(+)-induced allosteric regulation of catalytic activity in serine proteases.** *Proc Natl Acad Sci U S A* 1996, **93:**10653-10656.

162. Prasad S, Cantwell AM, Bush LA, Shih P, Xu H, Di Cera E: **Residue Asp-189 controls both substrate binding and the monovalent cation specificity of thrombin.** *J Biol Chem* 2004, **279:**10103-10108.
163. Bobofchak KM, Pineda AO, Mathews FS, Di Cera E: **Energetic and structural consequences of perturbing Gly-193 in the oxyanion hole of serine proteases.** *J Biol Chem* 2005, **280:**25644-25650.
164. Page MJ, Carrell CJ, Di Cera E: **Engineering protein allostery: 1.05 A resolution structure and enzymatic properties of a Na+-activated trypsin.** *J Mol Biol* 2008, **378:**666-672.
165. Sahin-Toth M, Toth M: **High-affinity Ca(2+) binding inhibits autoactivation of rat trypsinogen.** *Biochem Biophys Res Commun* 2000, **275:**668-671.
166. Plerou V, Gopikrishnan P, Rosenow B, Amaral LA, Guhr T, Stanley HE: **Random matrix approach to cross correlations in financial data.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2002, **65:**066126.
167. Kirchhofer D, Yao X, Peek M, Eigenbrot C, Lipari MT, Billeci KL, Maun HR, Moran P, Santell L, Wiesmann C, Lazarus RA: **Structural and functional basis of the serine protease-like hepatocyte growth factor beta-chain in Met binding and signaling.** *J Biol Chem* 2004, **279:**39915-39924.
168. Iversen LF, Kastrup JS, Bjorn SE, Rasmussen PB, Wiberg FC, Flodgaard HJ, Larsen IK: **Structure of HBP, a multifunctional protein with a serine proteinase fold.** *Nat Struct Biol* 1997, **4:**265-268.
169. Piao S, Kim S, Kim JH, Park JW, Lee BL, Ha NC: **Crystal structure of the serine protease domain of prophenoloxidase activating factor-I.** *J Biol Chem* 2007, **282:**10783-10791.
170. Katona G, Berglund GI, Hajdu J, Graf L, Szilagyi L: **Crystal structure reveals basis for the inhibitor resistance of human brain trypsin.** *J Mol Biol* 2002, **315:**1209-1218.
171. Parry MA, Jacob U, Huber R, Wisner A, Bon C, Bode W: **The crystal structure of the novel snake venom plasminogen activator TSV-PA: a prototype structure for snake venom serine proteinases.** *Structure* 1998, **6:**1195-1206.
172. Chen L, DeVries AL, Cheng CH: **Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish.** *Proc Natl Acad Sci U S A* 1997, **94:**3811-3816.
173. Di Cera E: **Thrombin interactions.** *Chest* 2003, **124:**11S-17S.
174. Baird TT, Jr., Wright WD, Craik CS: **Conversion of trypsin to a functional threonine protease.** *Protein Sci* 2006, **15:**1229-1238.
175. Bittar ER, Caldeira FR, Santos AMC, A.R. Gn, Rogana E, M. SM: **Characterization of -trypsin at acid pH by differential scanning calorimetry.** *Brazilian Journal of Medical and Biological Research* 2003, **36:**1621-1627.
176. Brumano MH, Rogana E, Swaisgood HE: **Thermodynamics of unfolding of beta-trypsin at pH 2.8.** *Archives of Biochemistry and Biophysics* 2000, **382:**57-62.
177. John DM, Weeks KM: **van't Hoff enthalpies without baselines.** *Protein Sci* 2000, **9:**1416-1419.
178. Naganathan AN, Munoz V: **Determining denaturation midpoints in multiprobe equilibrium protein folding experiments.** *Biochemistry* 2008, **47:**6752-6761.
179. Olsson AY, Lilja H, Lundwall A: **Taxon-specific evolution of glandular kallikrein genes and identification of a progenitor of prostate-specific antigen.** *Genomics* 2004, **84:**147-156.
180. Kam CM, Hudig D, Powers JC: **Granzymes (lymphocyte serine proteases): characterization with natural and synthetic substrates and inhibitors.** *Biochim Biophys Acta* 2000, **1477:**307-323.
181. Bell JK, Goetz DH, Mahrus S, Harris JL, Fletterick RJ, Craik CS: **The oligomeric structure of human granzyme A is a determinant of its extended substrate specificity.** *Nat Struct Biol* 2003, **10:**527-534.
182. Kashtan N, Alon U: **Spontaneous evolution of modularity and network motifs.** *Proc Natl Acad Sci U S A* 2005, **102:**13773-13778.
183. Binnig G, Quate CF, Gerber C: **Atomic force microscope.** *Phys Rev Lett* 1986, **56:**930-933.

184. Hoh JH, Cleveland JP, Prater CB, Revel JP, Hansma PK: **Quantized adhesion detected with the atomic force microscope.** *Journal of the American Chemical Society* 1992, **114:**4917-4918.
185. Lantz MA, Hug HJ, Hoffmann R, van Schendel PJ, Kappenberger P, Martin S, Baratoff A, Guntherodt HJ: **Quantitative measurement of short-range chemical bonding forces.** *Science* 2001, **291:**2580-2583.
186. Sugimoto Y, Pou P, Abe M, Jelinek P, P,rez Rn, Morita S, Custance O: **Chemical identification of individual surface atoms by atomic force microscopy.** *Nature* 2007, **446:**64-67.
187. Grandbois, Beyer, Rief, Clausen S, Gaub: **How strong is a covalent bond?** *Science (New York, NY)* 1999, **283:**1727-1730.
188. Kaiser U, Schwarz A, Wiesendanger R: **Magnetic exchange force microscopy with atomic resolution.** *Nature* 2007, **446:**522-525.
189. Ternes M, Lutz CP, Hirjibehedin CF, Giessibl FJ, Heinrich AJ: **The force needed to move an atom on a surface.** *Science (New York, NY)* 2008, **319:**1066-1069.
190. Stark M, Stark RW, Heckl WM, Guckenberger R: **Inverting dynamic force microscopy: from signals to time-resolved interaction forces.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99:**8473-8478.
191. Ashino M, Obergfell D, Haluska M, Yang S, Khlobystov AN, Roth S, Wiesendanger R: **Atomically resolved mechanical response of individual metallofullerene molecules confined inside carbon nanotubes.** *Nat Nanotechnol* 2008, **3:**337-341.
192. Wiita AP, Perez-Jimenez R, Walther KA, Grater F, Berne BJ, Holmgren A, Sanchez-Ruiz JM, Fernandez JM: **Probing the chemistry of thioredoxin catalysis with force.** *Nature* 2007, **450:**124-127.
193. Perez-Jimenez R, Wiita AP, Rodriguez-Larrea D, Kosuri P, Gavira JA, Sanchez-Ruiz JM, Fernandez JM: **Force-clamp spectroscopy detects residue co-evolution in enzyme catalysis.** *J Biol Chem* 2008, **283:**27121-27129.
194. Bustamante C, Chemla YR, Forde NR, Izhaky D: **Mechanical processes in biochemistry.** *Annu Rev Biochem* 2004, **73:**705-748.
195. Chu S, Bjorkholm JE, Ashkin A, Cable A: **Experimental observation of optically trapped atoms.** *Phys Rev Lett* 1986, **57:**314-317.
196. Neuman KC, Nagy A: **Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy.** *Nat Methods* 2008, **5:**491-505.
197. Zewail AH: **Laser Femtochemistry.** *Science (New York, NY)* 1988, **242:**1645-1653.
198. Rosker MJ, Dantus M, Zewail AH: **Femtosecond Clocking of the Chemical Bond.** *Science (New York, NY)* 1988, **241:**1200-1202.
199. Zhong D, Douhal A, Zewail AH: **Femtosecond studies of protein-ligand hydrophobic binding and dynamics: human serum albumin.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97:**14056-14061.
200. Ma H, Wan C, Wu A, Zewail AH: **DNA folding and melting observed in real time redefine the energy landscape.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104:**712-716.
201. Pal SK, Zewail AH: **Dynamics of water in biological recognition.** *Chemical Reviews* 2004, **104:**2099-2123.
202. Espagne A, Paik DH, Changenet-Barret P, Plaza P, Martin MM, Zewail AH: **Ultrafast light-induced response of photoactive yellow protein chromophore analogues.** *Photochemical & Photobiological Sciences: Official Journal of the European Photochemistry Association and the European Society for Photobiology* 2007, **6:**780-787.
203. Shorokhov D, Zewail AH: **4D electron imaging: principles and perspectives.** *Physical Chemistry Chemical Physics: PCCP* 2008, **10:**2879-2893.

204. Zewail AH: **4D ultrafast electron diffraction, crystallography, and microscopy.** *Annual Review of Physical Chemistry* 2006, **57:**65-103.
205. Lobastov VA, Srinivasan R, Zewail AH: **Four-dimensional ultrafast electron microscopy.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:**7069-7073.
206. Gahlmann A, Tae Park S, Zewail AH: **Ultrashort electron pulses for diffraction, crystallography and microscopy: theoretical and experimental resolutions.** *Physical Chemistry Chemical Physics: PCCP* 2008, **10:**2894-2909.
207. Kwon O-H, Barwick B, Park HS, Baskin JS, Zewail AH: **4D visualization of embryonic, structural crystallization by single-pulse microscopy.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105:**8519-8524.
208. Kwon O-H, Barwick B, Park HS, Baskin JS, Zewail AH: **Nanoscale Mechanical Drumming Visualized by 4D Electron Microscopy.** *Nano Letters* 2008, **8:**3557-3562.
209. Barwick B, Park HS, Kwon O-H, Baskin JS, Zewail AH: **4D imaging of transient structures and morphologies in ultrafast electron microscopy.** *Science (New York, NY)* 2008, **322:**1227-1231.
210. Park HS, Baskin JS, Kwon O-H, Zewail AH: **Atomic-scale imaging in real and energy space developed in ultrafast electron microscopy.** *Nano Letters* 2007, **7:**2545-2551.
211. Lobastov VA, Weissenrieder J, Tang J, Zewail AH: **Ultrafast electron microscopy (UEM): four-dimensional imaging and diffraction of nanostructures during phase transitions.** *Nano Letters* 2007, **7:**2552-2558.
212. Grinolds MS, Lobastov VA, Weissenrieder J, Zewail AH: **Four-dimensional ultrafast electron microscopy of phase transitions.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103:**18427-18431.
213. Yang D-S, Lao C, Zewail AH: **4D Electron Diffraction Reveals Correlated Unidirectional Behavior in Zinc Oxide Nanowires.** *Science* 2008, **321:**1660-1664.
214. Ihee H, Lobastov VA, Gomez UM, Goodson BM, Srinivasan R, Ruan CY, Zewail AH: **Direct imaging of transient molecular structures with ultrafast diffraction.** *Science (New York, NY)* 2001, **291:**458-462.
215. Zuo JM, Vartanyants I, Gao M, Zhang R, Nagahara LA: **Atomic resolution imaging of a carbon nanotube from diffraction intensities.** *Science (New York, NY)* 2003, **300:**1419-1421.
216. Levitt M, Warshel A: **Computer simulation of protein folding.** *Nature* 1975, **253:**694-698.
217. Dill KA: **Theory for the folding and stability of globular proteins.** *Biochemistry* 1985, **24:**1501-1509.
218. Bryngelson JD, Wolynes PG: **Spin glasses and the statistical mechanics of protein folding.** *Proceedings of the National Academy of Sciences of the United States of America* 1987, **84:**7524-7528.
219. Dill KA: **Polymer principles and protein folding.** *Protein Science: A Publication of the Protein Society* 1999, **8:**1166-1180.
220. Baker D: **A surprising simplicity to protein folding.** *Nature* 2000, **405:**39-42.
221. Clementi C: **Coarse-grained models of protein folding: toy models or predictive tools?** *Current Opinion in Structural Biology* 2008, **18:**10-15.
222. Ma J: **Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes.** *Structure* 2005, **13:**373-380.
223. Tirion MM: **Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis.** *Phys Rev Lett* 1996, **77:**1905-1908.
224. Bahar I, Erman B, Haliloglu T, Jernigan RL: **Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations.** *Biochemistry* 1997, **36:**13512-13523.

225. Bahar I, Atilgan AR, Erman B: **Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.** *Fold Des* 1997, **2:**173-181.
226. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I: **Anisotropy of fluctuation dynamics of proteins with an elastic network model.** *Biophys J* 2001, **80:**505-515.

Appendix A: Distance metrics

Different methods can be used to calculate distances between two vectors, x and y with n components.

Euclidean: $d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_3)^2 + \cdots + (x_n - y_n)^2}$

Standardized Euclidean: Similar to the euclidean distance except x and y vectors are first divided by the variance. This is helpful when one vector is on average numerically much higher than the other which would result in the components of that vector dominating the distance calculation.

Mahalanobis: This distance measure takes into account the covariance between data points. This is easily expressed in matrix form: $m = (\mathbf{x} - \mathbf{y})\mathbf{V}^{-1}(\mathbf{x} - \mathbf{y})^T$, where $\mathbf{V}$ is the covariance matrix.

Cityblock: This distance is also known as manhattan distance or the taxicab metric. It represents the distance between points as if they were laid out on a grid with discrete distances. The formula is $cityblock = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$

Cosine: This is defined as 1 **–** [the dot product of two vectors of length n divided by the distance from the origin]. The formula is: $d = 1 - \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}\sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}}$

Correlation: This is similar to cosine distance except it is shifted by the means of the vectors so that: $\text{correlation distance}(x, y) = \text{cosine distance}(x - \bar{x}, y - \bar{y})$

Spearman: Spearman is a non-continuous distance metric. First the points in the vectors are ordered and then the square of the euclidean distance between each of the ordered points is calculated. One is subtracted from the square of the euclidean distance. This distance is capable of detecting linear and non-linear correlations.

Hamming: This measure treats the vectors as strings and finds the number of positions where the symbols constituting the string are different. If there are many symbols that are different then the distance would be large. The matlab implementation divides the distance by n.

Jaccard: This is similar to the Hamming distance is that it treats vectors as strings. It is more complex however. The formula is: $J = 1 - \frac{|A \cap B|}{|A \cup B|}$

Chebychev: This distance is the maximum of absolute differences between two vectors as in: $d = \max\left[|x_1 - y_1|, |x_2 - y_2|, |x_n - y_n|\right]$.

Appendix B: Linkage methods

This is a short description of linkage methods used in the text (based on the Matlab documentation):

1. Average linkage: This is calculated by taking the average values of distances calculated between all pairs of elements in each group.
2. Single linkage: The distance between the groups is the distance between the closest two elements.
3. Complete linkage: The distance between two groups is the distance between the farthest two elements.
4. Centroid linkage: This is calculated by taking the mean of each group and then computing the distance between the means.
5. Median linkage: This is similar to centroid linkage except the centroid means are weighted by the number of elements that make up each centroid.
6. Weighted linkage: This is calculated similar to average linkage but if there are different number of elements in each group then the groups with more elements contribute more; hence the weighted term in the name of the method.
7. Ward linkage: Ward linkage is the most computationally expensive of the methods here. The method calculates a distance by minimizing the increase in variance that occurs after grouping two groups together. It has similarities to analysis of variance (ANOVA).