Protein Sectors:
Structure, Organization and Variation at Multiple Scales

Maeva Fincker, Civil & Environmental Engineering Department, Stanford University

Leah Sibener, Program in Immunology, Stanford University

Brad Huang, Biology Department, Stanford University

Tatum Banayat, Bioengineering Department, Stanford University

Abstract

Statistical coupling analysis has been used to define coevolving residues and their functional implications. In this paper, we explore how a novel algorithm, Mutual Information for Sequence Coevolution, can further extend this approach by connecting clustering with structural and mutational analyses. First, we study the structures and compositions of sectors in eukaryotic transmembrane proteins. We are able to cluster transmembrane sectors into five groups of shapes, with a gradient from globular sectors to stretched sectors. Within these clusters, we can identify several types of shapes corresponding to the known alpha helices and beta strands located in transmembrane domains and channels. In addition, our algorithm applied onto the specific group of pore forming toxins reveals sectors wrapped around the glycine-rich region of beta-barrel toxin. Although alpha hemolysin and gamma hemolysin both contain Gly-rich regions, the positions of the glycines are not aligned. From observations to the wrapped sectors, we propose an alternative hypothesis for this homology, stating that the sector amino acids interact and coevolve with each other to localize the glycines to specific regions in the barrels that form the beta loop portion. Having confirmed the functional implications of the sectors generated from our algorithm, we apply our algorithm to study correlated mutations in tumor suppressor p53. From the database of p53 somatic mutations in tumor patients, we observe enrichments and depletions of mutations in sectors. Combined with information on coevolving residues, we generate a list of compensatory mutation candidates. Most compensated mutations fall in DNA binding and oligomerization domains. However, these mutations are compensated by mutations outside of sectors, and no obvious trends in amino acid replacement can be observed. Overall, our algorithm can be applied to multiple scales to provides insights to structural evolution and candidates for compensatory mutation analyses.

Author Summary

Coevolution refers to the concerted changes occurring in groups of organisms or biomolecules over time. Similarly, protein coevolution, the coordinated changes of small protein structures, occur both between and within proteins. These changes usually come in the form of refining functionality of the molecules. Since the functionality and structure of proteins are intrinsically linked, coevolution of protein is important in understanding the interactions between proteins and between residues. In our study, we provide a novel algorithm for finding sectors in proteins containing amino acid residues that coevolve with each other. By combining these sectors with structural information, we find structures that proteins share, revealing insights to protein interactions. We also discover how residues in pore-forming toxins can interact with each other to form functional regions that previous computational methods cannot find. Last but not least, we use our method to investigate the tumor-inducing mutations of p53. For certain mutations that exist in a tumor-bearing patient, we are able to find mutations in other species that potentially compensate the adverse effects. In general, our method allows the study of protein structures and variations on multiple scales, and can aid biomedical researches in various ways.

Introduction

The central dogma requires that all biological information necessary to the folding of a protein and its function is contained within its sequence [1]. Layers of organization such as the primary, secondary, and tertiary structures have been extracted from the raw amino acid sequence and give clues to understand how proteins fold and which residues are functionally important. However, questions pertaining to allostery or adaption to evolutionary pressure remains mostly unanswered. Recently, a new level of organization, called protein sectors, was identified [2]. Sectors are defined as group of coevolving residues. They are defined using Statistical Coupling Analysis (SCA), an algorithm able to deconvolve the intricate evolutionary history hidden in the multiple sequence alignment of a protein family. When looking at their structure in space, sectors usually take the form of sparse networks of physically contiguous residues [3]. Sectors have been shown to play an important role for structure, function and allostery [3] [2] [4] as well as contain important residues for adaptation and evolution of the function [5].

More studies are demonstrating the biological importance of protein sectors but they do so on a small scale, at the level of a single protein family. To our knowledge, no large scale comparison of sectors from multiple protein families has ever been undertaken, nor has there been any effort to join evolutionary data from sectors with structural data from crystallized proteins. Based on these observation, we decided to explore how structural visualization of sectors can aid the understanding of what implications sectors have by combining the sectors obtained from our algorithm and the structural information of the Protein Database (PDB).

In this study we used a novel algorithm for SCA called 'Mutual Information for Sequence Coevolution' (MISC). This method varies from past SCA algorithms by defining coevolving residues through disentangling the evolutionary history of a protein family contained in the multiple sequence alignment of the protein family. More specifically, the MISC algorithm first generated a co-evolution matrix between every pair of residues wherein co-evolution scores were not dependent on the conservation of particular residues. Mean Shift clustering [6] of this co-evolution matrix yielded groups of residues, which were defined as sectors if they exhibited larger than average evolutionary coupling. This allows for linking evolutionary data and structural data on a larger scale than ever before.

Here we provide evidence that protein sectors can help in answering broad theoretical questions about protein organization as well as provide novel insights into biomedically relevant questions. We used MISC to generate a database of sectors of all membrane and transmembrane proteins for which there is a structure in the PDB. We were able to characterize sectors based on their shape and their amino acid composition. Additionally we found that the shape of sectors can be used as a novel way to characterize proteins. Furthermore, when we applied MISC onto the subgroup of beta barrel pore forming toxins we found that sectors wrapped around the glycine-rich region of beta-barrel toxin that is necessary for membrane penetration. Having confirmed the functional implications of the sectors generated from our algorithm, we applied MISC to study correlated mutations in the tumor suppressor p53. From the database of p53 somatic mutations in tumor patients, we observed specific enrichments and depletions of mutations in sectors. Combined with information on coevolving residues, we generated a list of compensatory mutation candidates.

Results

The membrane and transmembrane Sector DataBase (SDB)

Membrane proteins are notoriously difficult to purify and crystallize due to the high hydrophobicity of their outer surface, making them very insoluble. Structural information about membrane proteins is therefore very sparse. The development of a tool which can potentially lead to new knowledge on the function and structure of membrane proteins is worthy of pursuit. Based on this consideration, we decided to apply the MISC algorithm to membrane proteins and cross available structural data with evolutionary information contained in their sectors. Two hundred and fifty unique membrane proteins, whose 3D structure is available, were selected and used in the creation of the membrane sector database. They were sorted into three structural groups - alpha helical, beta barrel or monotopic membrane proteins - and sixty nine functional subgroups (Table 1). The alpha helical and beta barrel proteins contain a transmembrane region and are exposed to the inner and outer side of the membranes they cross. Conversely, the monotopic proteins are comprised of anchor domains attaching them to the membrane but interacting only with the cellular compartment on their side of the membrane. The distribution of proteins in functional subgroups was highly non-uniform; subgroups such as the oxygenase one had a single representative protein whereas the subgroup of G protein-coupled receptors comprised as many as 10 crystal structures. Comparisons at a functional level was therefore only possible for some of the categories. The MISC algorithm was applied to those proteins to define sectors of coevolving residues. The program identified one thousand two hundred and five sector, with an average of five sectors per protein. The length (number of residues) in each sector varied greatly (Figure S1), with a median value of 11 amino acid but a mean value of 29.

We were particularly interested in the properties of the sectors located in the transmembrane region of the proteins in our database. Sectors with more than seventy five percents of their residues in the transmembrane part of the proteins were extracted from the SDB and pooled together to form the transmembrane SDB. This database grouped eighty sectors coming from forty five different proteins and twenty one functional subgroups. Because the UniprotKB/trEMBl protein database does not list transmembrane region for monotopic proteins, the transmembrane SDB was only constituted with alpha helical and beta barrel transmembrane peptides.

Amino acid enrichment analysis shows differential preferences for amino acids in sectors

One of the first questions we sought to address was *how do amino acid distribution in sectors differ throughout subgroups?* It is unknown whether there is an enrichment or depletion of specific amino acids in co-evolving sectors, and therefore it was necessary to compare the amino acid distribution in sectors. Subgroups in the SDB are characterized by similar structure and function, and therefore we had decided to examine amino acid enrichment in each of the subgroups. We calculated the amino acid enrichment for each protein and took the median for all the proteins in a subgroup to yield subgroup enrichment. (Equation 1). We saw that there was no specific amino acid enrichment across all subgroups (Figure 1) Interestingly, four subgroups were highly enriched in cystine: 'Electron Transport Chain Complexes: Complex IV (Cytochrome C Oxidase)', Nitric Oxide Reductases, Oxidoreductases, and the Solute Sodium Symporter family. The heatmap indicates that in some subgroups there are different residues that are enriched or depleted. Additionally, many subgroups may have slightly more subtle changes in amino acid composition.

Looking at this question on the residue level, we were able to calculate the enrichment (Equation 1) of amino acids in all the sectors in a single protein. The amino acid frequency pattern for the transmembrane SDB shows a high enrichment for arginine and cysteine - 440% and 240% increase in frequency respectively. The presence of an increased frequency for cysteine can be easily understood by its ability to form a disulfide bond, hence lending rigidity to the transmembrane region {Bowie:vf}. The enrichment in arginine is more difficult to interpret but arginine and more generally positively charged amino acids have been found to ease the insertion of membrane proteins into the lipid layer [7].

Common sector structures are shared across protein families

Clustering of transmembrane sectors by their shape

Sectors in a protein are identified solely from the evolutionary history of the protein family. If we assume that evolutionary pressure might be comparable across different protein families - need for a different substrate specificity or an increased affinity for example - a possible hypothesis is that the impacts of selection might also be comparable at the structural level. To test this idea, principal component analysis was applied to the 3D coordinates of the centroids of the residues of each transmembrane sector. The resulting eigenvalues were used as an indicator of the shape of the the sector and the normalized eigenvalues - where for a given sector, each eigenvalue was divided by the sum of the eigenvalues of this sector - were used to compare relative deformation between sectors. Sectors stretched in a single direction will have a normalized primary eigenvalue close to 1 whereas globular sectors will have all 3 normalized eigenvalues with values close to 0.33. The distribution of eigenvalues (Figure S2) is continuous and does not exhibit any obvious clustering. The range of value for the primary eigenvalue is 10 times wider than the range for the second eigenvalues and 25 times wider than the one for the third eigenvalues. This depicts the broadness of the structural space and the variety of shapes sectors can adopt.

The spread of deformation seemed very homogenous, making the direct analysis of the shape distribution impossible. Therefore an unsupervised hierarchical clustering method was employed on the z-scores of the three eigenvalues to elucidate the relation between the shape of each sector. As was hinted by the scatter plot in (Figure S2), the resulting heatmap (Figure 2A) does not show strong demarcation between groups of sectors. However, the algorithm was able to separate the sectors into 5 clades. Based on this number of clades, we grouped the sectors into 5 cluster using the K-mean algorithm. This effectively created groups of sectors sharing the same relative shape (Figure 2B). Cluster 1 consisted of the most globular sectors ($\lambda1 \approx \lambda2 \approx \lambda3$) whereas cluster 5 comprised the most stretched ones ($\lambda1 \approx 1$). Cluster 2, 3 and 4 are clusters whose shape is more and more extended in a single direction. The mapping of the sectors of each cluster onto their protein 3D structure revealed no cluster contains a single type of cluster shape, there is structural diversity in every one of them. However, a closer analysis reveals conserved sector structures in three of the five clusters. Cluster 5, wherein sectors residues are mostly distributed along one direction, consists mostly of sectors developing along an alpha helix or a beta strand where residues are contiguous in the structure as well as in the sequence. (Figure 3) Two main sector structures could be identified in cluster 4. In one type, the residues in the sector were wrapped around 2 main alpha helices over their whole length. The second type corresponded to sectors found in porins where the residues were numerous and spread around the beta barrel. Cluster 3 regrouped sectors with residues located on the inside of the protein channels, whether the channel is made of alpha helices or beta strands. No obvious shape could be recognized in clusters 1 and 2 where the sectors were much broader and often spanning the entire protein. Although the biological meaning of those common shapes could not be evaluated in this paper, the very existence of sectors sharing a similar structures across different protein families was intriguing and should be investigated in future works.

Functional subgroup 'Beta-Barrel Pore Forming Toxins' have similar sector structure

Another way to approach sector analysis is to specifically query a subgroup of proteins that have similar function. Two subgroups we were particularly interested in were beta-barrel pore forming toxins and alpha helical pore forming toxins. These subgroups are unique because across both subgroups they have the same function, but the two subgroups have distinct structural characteristics. Therefore we asked the questions: *How does the sector composition vary structurally and biochemically first within a structurally similar subgroup and second between functionally similar subgroups but structurally dissimilar?* We found that beta barrel toxins have similar sector structure and organization (Figure 4A). Specifically, the beta barrel toxins shown are Alpha Hemolysin, the HlyA Cytolysin, and Gamma-Hemolysin. We found that there are sectors in two distinct locations on protein . First there is a predominant "wrapped sector" in all of the toxins (shown in navy blue). These sectors span from the top of the beta-barrel down into the body of the protein (fig a). Second, many different protein sectors are contained in the body of the protein. It is clear that there is spatial organization of the co-evolving residues that is consistent throughout the beta barrel toxins in the membrane SDB. In order to further compare the differences of these sectors between the different proteins, we calculated the amino acid enrichment of the sectors in each of the proteins (Figure 4B, S3). It seems that the sectors are biochemically distinct with different enrichments associated with different proteins, where by there are no specific trends. Sectors in both alpha-hemolysin and HlyA were the most highly enriched for cysteine (1.33 and 1.57 respectively), where gamma-hemolysin was the most enriched for glutamine and phenylalanine (1.4559). in order to pursue how these sectors vary compared to one another we performed residue PCA on all of the proteins within the beta barrel subgroup (Figure 4C). These data show how the sector composition of residues co-varies with respect to one another. The top three modes (18, 19, 20) give 44.3% of the variance between the sector composition, whereby mode 18, 19, and 20 yield 11.3%, 14.5%, and 19% of the variance respectively. Within the mode 20 we see that cysteine and glycine positively covary with one another. As aforementioned, cysteine has been shown to be important in protein-protein within the transmembrane region. Glycine is well known as an amino acid that allows for conformational flexibility. Overall these results imply that there is a specific structural organization of sectors within beta barrel pore forming toxins.

Next, we were interested to compare sectors of two subgroups that were functionally similar to one another. As mentioned above we compared alpha-helix pore forming toxins to the beta-barrel pore forming toxins. Within the alpha-helix pore forming complex subgroup there were only two proteins however they were structurally similar. We first mapped the sectors onto Hemolysin E from *E coli* and compared it to the HlyA beta barrel toxin (Figure 5A). Interestingly of the four sectors in Hemolysin E, two of them in are restricted to a specific alpha helix (navy and green sector), whereas the other two are spatially interspersed (red and yellow sectors). Unsurprisingly, the sector shapes are very different in the alpha helix toxins compared to the beta barrel toxins. However, there is a specific sector (navy blue) at the top of the toxin similar to the wrapped sector in the beta barrel toxin. We compared the median amino acid enrichment of the two subgroups (Figure 5B). Interestingly although these two subgroups have the same function, the amino acid compositions are unique. Within the beta-barrel subgroup there are larger differences of amino acid enrichment. For example,the beta-barrel toxins are most highly enriched for Arg, Cys, and Leucine whereas the alpha-helix toxins are the most highly enriched in Asn. However, both were depleted the most in Met (Figure 5B). These data depict that there are both structural and amino acid differences within co-evolving sectors in subgroups that have the same biological function. It appears that the overall shape of sectors is guided by the shape of the protein.

‘Wrapped sectors’ reveal coevolution of functional domains in beta-barrel toxins

Next we wanted to explore the biological significance of the ‘wrapped sectors’ in beta barrel toxins. When the first beta barrel toxin structure was solved it was hypothesized that the glycine rich region along the top of the beta barrel facilitated membrane entry into the membrane to cause efficient lysis (8). Additional studies have shown that site specific mutagenesis on the Gly-rich region led to decrease in or elimination of protein functionality, where by these proteins could no long cause hemolysis [8]. We noticed that these gly-rich regions colocalized with the ‘Wrapped sectors’ that are present in all of the beta barrel toxins within the membrane SDB (Figure 6A). The first analysis we did was to understand the biochemical composition of the wrapped sectors. Initially, we thought that because these sectors participated in the membrane insertion there would be a high percentage of acidic or basic residue. We saw that across the beta barrel toxins the wrapped sectors were relatively similar to “all proteins” (Figure 6B). All proteins (purple bars) represents the frequency of given amino acids classified as hydrophobic, acidic, or basic throughout all proteins in the NCBI database. Although the frequency of residues is similar the database calculated the alpha-hemolysin and HlyA wrapped sectors were classified as acidic, whereas gamma hemolysin sectors was classified as basic. These classifications are based on the composition of residues in all proteins . Next we investigated which of the residues in the gly-rich region were in the wrapped sectors. In all of the wrapped sectors across the beta barrel toxins there were the residues that are specifically within the gly-rich region (Figure 6C). It should be noted that only the gly-rich region that has been annotated in alpha-hemolysin. The HlyA and gamma-hemolysin “gly-rich” regions were identified by structural homology as well as close examination of the sequence. However although it was clear that there was a gly-rich region in each of these proteins located at the same place as alpha-hemolysin, there was no alignment or consensus sequence for the position of the glycines. Furthermore, not all of the glycines were in the co-evolving sectors. This may imply that the beta barrel sectors have co-evolved to position functional amino acids.

In order to explore this idea, we assessed the interactions of the sector amino acids. It appears that the polar interactions of the sector amino acids interact with each other to form the beta loop portion of the barrel (Figure 6D). Furthermore, two of the sector residues Asn 136 and Thr 118 form a hydrogen bond with one another. These allow for glycines (cyan) on two opposite beta strands to be localized to a specific portion on the middle and outside of the beta barrel. It is clear that the co-evolved wrapped sectors (yellow) may do not include the all of the functional glycine amino acids in the gly-rich region that allows for membrane penetration. However, these interactions may imply that these the wrapped sector allows for stabilization of the the beta sheet prior to oligomerization and upon oligomerization allows for glycine to be displayed and interact with the membrane.

After seeing that we were able to identify co-evolving sectors with similar shape within functional subgroups, and that the wrapped sectors encompass some of the functional domains within the beta barrel toxins, we were interested in pursuing using MISC to answer more biomedically significant questions.

Sector analysis of p53

Enrichment analysis of p53 sectors

One biomedical application we wanted to pursue with the MISC algorithm as a tool is screening millions of mutations for possible compensatory mutations. Tumor suppressor p53 is an example for us to test our strategy. First, we used the wild type Homo sapien p53 sequence as the reference sequence to obtain MSA of non-human p53 sequences. We used the MISC algorithm to compute the sectors. We then obtained database of somatic mutations observed in tumor patients from IARC TP53 [9]. We summarized the enrichment of mutations in non-human p53 sequences in Figure 7. Since the total number of mutations in each sectors is high, the odds ratio that a somatic mutation is in each sector is significantly different from the expected values based on sector length (Table 2). What's worth more attention is that sector 2 and 4 has enrichment as high as 4, while sector 5 and 6 has enrichment lower than 0.2 (Figure 7).

Statistical Analysis of mutations in p53 sectors

After confirming that the sectors do affect the distribution of residues with carcinogenic mutations, we continued to examine if these mutations do occur naturally in non-human species, and whether the existence of these mutations would imply compensatory mutation. Figure 8 displays the counts of each somatic mutations, grouped by sectors, present in non-human sequences.

Based on the results of Kruskal-Wallis test comparing the average counts of each sector, sector is a factor that shapes the distributions of somatic mutation abundance in the MSA (Table 3). However, multiple comparison tests do not show an apparent alternative hypothesis stating how the hierarchies of average abundance in sectors are (Table 4). The average abundance of mutations is highest in sector 6, and lowest in sector 2. Even though there is statistical significant difference between the abundance of these two sectors, there is no statistical significant difference between the abundance of sector 3 and 6, and between the abundance of sector 2 and 3. As a result, we cannot conclude a clear alternative hypothesis for the differences in abundance of mutations.

Since there were no apparent trends in the distribution of mutation abundance in sectors, we decided to take the mutation most abundant in non-human species and explore possible compensatory mutations of this sector 3 mutation (Figure 8). Figure 9 shows the visualization of sequences with the mutation on residue 133 from Met to Leu. Each of the 903 sequences visualized in this figure contains the mutation on residue 133. However, there is no apparent compensating mutation on residues of sector 3.

Compensatory mutation candidates

Based on the knowledge we have about distributions of mutations in sectors and non-human species, we attempt to generate a list of compensatory mutation candidates. Several factors are considered in the process: the coevolution between residues, the percentage of sequences able to compensate the mutations in coevolved residues, and the number of sequences with the same pair of compensatory mutation (see methods and materials for details). The resulting list of compensatory mutations contain 28 pairs (Table 5). The somatic mutations with compensating mutations under our screening only occur in sector 1 and 5, while all of the compensating mutations are outside of the clusters.

Discussion

Coevolution of structure may mediate function

Here we present MISC as a tool to analyse protein co-evoution and can be applied at multiple scales to provides insights to the evolution of protein structure-function relationship. We were able to characterize sectors in membrane proteins, cluster transmembrane proteins based on their shape, and specifically we were able to hypothesize how beta barrel pore forming toxins structurally evolved to insert into cell membranes.

By clustering sectors by their shape we eliminated a structural and functional bias when comparing sectors to one another. We were surprised to find that clustering based on coordinate PCA allowed for elucidation of similarly shaped transmembrane sectors, onto proteins that had similar function. For example, transmembrane cluster 5 which was the most similar based on the coordinate PCA revealed a distinct outer shape along the outside of the protein. However, the biological meaning of those common shapes could not be evaluated in this paper, but the very existence of sectors sharing a similar structures across different protein families should be investigated in future works.

When we looked at sectors of functional subgroups, we found that in all the beta barrel toxins within the membrane SDB residues that co-evolved together formed a specific structure we have called a wrapped sector. Interestingly this wrapped sector spans the length of the beta-barrel forming domain and continues into the body of the protein. These beta barrel toxins are secreted as monomers, insert into the membrane, and then they can oligomerize to form pore complexes [10]. A portion of the gly-rich region that is necessary for a monomer insertion into the membrane is within the wrapped sector. Additionally we found that there was no specific glycine sequence motif that was present in all of the proteins instead there was just a local concentration of glycines that were present either contained in the sector or in between sector residue components. This information taken together may imply that the specific position of the glycines of the beta-barrel may not matter, but the local concentration and proper positioning of glycines allows for proper membrane integration. Upon examining the interactions that were present within the wrapped sector portion of the beta barrel we notice that the residues within the sector form the intermolecular bonds that allows for the structural basis of the beta loop. We propose that the wrapped sector may act as a structural network that has co-evolved to orient the glycines properly for their interaction with cell membranes. The specificity of the interactions between the amino acids in the wrapped sector and the rest of the protein may be enthalpically driven in nature whereby the specific electrostatic interactions allow for proper folding of the beta loop [11]. Additionally the enthalpically driven interactions could consequently drive the positioning of the glycines for optimal interaction with the membrane. This hypothesis is speculative, and requires non-computational support via biological experiments such as mutagenesis studies coupled with structural studies as well as protein stability assays (e.g. GdnHCL unfolding assays) to understand the energetic contribution of the wrapped sector residues.

Sectors and Compensatory Mutations in p53

In the first part of our studies we approached questions pertaining to protein coevolution from the perspective of understanding the composition of sectors in membrane proteins, and specific structural and functional subgroups. We next pursued asking questions about protein co-evolution from the perspective of a biomedically relevant protein of interest: p53. We decided to study how mutations that affect the functions of p53 are associated with the sectors.

p53 sectors are continuous with protein functionality

The six sectors of p53 fall into one of three groups, each according to a particular domain of p53. Each sector can either be a DNA binding sector, a tetramerization sector, or a transactivation sectors. From our analysis, we observe that sectors one through four are DNA binding, sector five is a tetramerization sector, and sector six is a transactivation sector. After conducting the Kruskal-Wallis test, we concluded that we could not link any sectors to one another. However, this allows us to generate unconnected hypotheses for each sector. With this in mind, we delved into the functional consequences of the various sectors by visualizing the sectors and examining the mutational enrichment of the sectors.

Due to the limitations of the RCSB PDB database, we were only able to visualize sectors one through four, which all happen to span the DNA binding domain. These sectors varied in size both spatially and in the quantity of residues, as can be seen from the sector overlay onto the PDB file 3Q05. (Figure 10). Looking closely at the residues near to the DNA, it is clear that a significant concentration of sector two and four are adjacent to the DNA (Figure 11). Also, these sectors were already shown to exhibit a significant amount of enrichment compared to the other sectors (Figure 7). This suggests that sectors two and four are more significant for a functional DNA binding domain. In order to confirm this, functional studies on mutations occurring within sectors one through four could be used to compare factors such as stability and DNA binding affinity.

Although sectors five and six could not be visualized, there is valuable information when comparing the enrichment of the sectors and their location within p53. Sector five resides almost entirely within the oligomerization domain, is only sixteen residues in length, and is depleted of mutations. (Figure 7) Biologically this makes sense since p53 has many more functions as a tetramer and dimer other than being a transcription factor. P53 alters metabolism and mediates many protein-protein interactions within the cell (Ref 9). Thus, mutations in the oligomerization site could lead to a host of other complications and diseases besides cancer, and perhaps increase morbidity. Hence it is no surprise that the tetramerization sector is depleted of cancerous mutations

Sector six is another interesting case, since it almost entirely within the transactivation domain. Like sector five, six is also depleted and is nine residues in length. In order to understand the sector’s role in protein structure, it is useful to analyze the members of sector six, particularly residues twenty two and twenty three. These residues are known to be essential for binding to transcriptional machinery and maintaining transactivation activity of p53 [12]. While these residues are important for cell function overall, studies have shown that decreased transactivation activity does not have a large effect on growth regulation and apoptosis [12]. With this knowledge, our method is validated since it recovered structurally important residues and did not contain highly oncogenic residues.

Selecting Compensatory Mutation

With the functional implications the sectors have, we are interested in finding whether mutations in these function are related to sectors, and if the sectors we obtain through our algorithm can facilitate finding correlated mutations with biomedical application potentials. The correlated mutations we are interested in are the pairs of mutations that compensate each other's effect on p53 functions, and our computational approach to finding compensatory mutations leads to some interesting findings.

The criteria we decided to select compensatory mutation from the somatic mutation include three consecutive conditions, first of which is that the two residues with compensatory mutations have to be coevolving to a certain degree. We select this criterion since the degree of coevolution, or the joint entropy, is an important indicator of interactions between residues. Selecting residue pairs with high joint entropy would imply mechanistically relationship between the residues. Although we cannot infer the biochemical interactions between the residues from our data, this possible link to protein structure and biochemistry is important for compensatory mutations.

The second condition is the proportion of sequences with the somatic mutation compensated. As indicated in Figure 8, the amount of sequences with somatic mutations naturally occurring varies with the sectors. The sequences of p53 protein in other species do not come from species of same expected age and life history. As a result, presence of a somatic mutation in some sequences does not necessarily imply the complete and correct functionality of p53 protein in those species, as it may be mitigated in early life stages. For a somatic mutation to be considered compensated in the other species, it is important that whenever a species bear the somatic mutation under consideration, the residues that coevolve with the mutation residue are also mutated.

The third condition is the number of sequences with the specific compensating mutation pair. Mutations on one residue may compensate the carcinogenic somatic mutation in different ways. Since our strategy does not include biochemical tests for the compensation mechanism, it is important to use the absolute number of sequences with the pair of compensating mutations as a way to ensure the potency of the compensation.

Compensatory Mutation Candidates

The compensatory mutations in the list generated from the above criteria are only compensating somatic mutations in sector one and five (Table 5). While sector one is in the DNA binding domain, sector two and four are the closest to the DNA and may be more functionally important. As shown in Table 4, somatic mutations in sector 2 and 4 are the least abundant in non-human sequences among those in the six sectors. As a result, mutations in sector two and four are more fatal and allow less possibility for compensatory mutations.

In addition, sector five contains the oligomerization domain of p53. Compensatory mutation in this domain should preserve the ability for p53 to oligomerize. There is only one residue with somatic mutations compensated in the sector, with a change from Arg to Lys. Arg and Lys are both basic amino acids, and this mutation may not be effectively changing how the p53 proteins oligomerize. However, there are many different mutations compensating this mutation. There is no apparent type of compensating mutation to this mutation, and most of the compensating mutations do not change a lot in the biochemical properties of the residues. Thus, the compensatory mutation pairs in sector 5 may be related to the change in physical space of protein oligomerization when mutation occurs, rather than the biochemical properties of the interface.

It is worth noting that none of the compensating mutations are within sectors. There are several possible explanations of this finding. First, the sectors may be functionally relevant, but may not be all of the functioning units of the protein. Residues that are completely conserved or highly variable are often not included in sectors, and there can be many different types of compensating mutations in the functional units not extracted as sectors. Second, we used three consecutive eliminating criteria to select the compensatory mutations. However, the criteria we used are only computationally convenient. It is possible to validate the criteria by comparing them with statistics on the compensatory mutations with biochemically understood mechanisms, but it is difficult to gather such data. At last, The compensatory mutations we found requires further biochemical tests to validate their compensations, and there may be unknown interactions between residues inside and outside of residues.

Materials and Methods

Statistical coupling analysis

Sectors were identified using a modified form of statistical coupling analysis. The novel SCA algorithm used in this study, called MISC, defined coevolving residues by disentangling the evolutionary history of a protein family contained in the multiple sequence alignment of the protein family. More specifically, the MISC algorithm first generated a co-evolution matrix between every pair of residues wherein co-evolution scores were not dependent on the conservation of particular residues. Mean Shift clustering [6] of this co-evolution matrix yielded groups of residues, which were defined as sectors if they exhibited larger than average evolutionary coupling.

Pipeline from pdb entry to clusters definition

For each protein of interest in the Protein Data Bank, a multiple sequence alignment was generated by blasting the amino acid sequence of the protein - the reference sequence - against the nr database of NCBI. The first 500 hits were kept and aligned using the clustal omega algorithm [13]. In order to constrain the phylogenetic space covered in the MSA to sequences of the same family as the protein of interest, the pairwise phylogenetic distance of each hit sequence to the reference sequence was calculated using the seqpdist distance metric from Matlab. The distribution of distances was then analysed and sequences with a distance higher than the mode + standard deviation of the distance distribution were discarded from further analysis. Furthermore, MSAs with less than a hundred sequences left after constraining the phylogenetic space were also removed from the dataset. The newly pruned MSA was then fed into the MISC algorithm to define the coevolving sectors in the protein of interest.

Relevant information about each sector, such as their protein of origin, their length or the coordinates of the centroid of each residue were extracted from the Protein Data Bank record and stored in a database for further access and analyses.

Membrane and transmembrane sector database construction

A recently updated list of membrane proteins whose crystal structure has been solved, could be found on the following website: http://blanco.biomol.uci.edu/mpstruc/ . We extracted the PDB identification number of the 477 unique crystallized membrane proteins as well as structural and functional information from this website. We then generated a MSA for each of them and ran the MISC algorithm on the MSAs to identify the sectors. More than half of the selected proteins had to be removed from the analysis because of inconsistencies in the PDB records and in the end, we had 250 unique membrane proteins in our database.

Information about whether a residue was part of the transmembrane region of a protein was extracted from the UniprotKB/trEMBl record of this protein, and more specifically from the region data in the sequence feature paragraph. The percentage of transmembrane residues in each sector was computed and sectors with more than 75% of their amino acids in the transmembrane region were labeled as transmembrane sectors pooled in a transmembrane sector database.

Analyses framework

Amino Acid Enrichment Analysis

Amino acid enrichment per protein was calculated by the frequency of a given amino acid in a sector divided by the enrichment of a the specific amino acid in the entire protein (equation 1.) The enrichment over the a subgroup was calculated as the median of all the protein enrichments in a given subgroup as annotated in the membrane sector database. In few cases there were zero specific amino acids in certain subgroups and therefore the subgroup enrichment was "infinite". In these cases, we denoted inf to equal no specific enrichment or a value of 1. Heatmaps were produced in Microsoft Excel.

Equation 1. SectorAAi SectorResidues/ TotalAAi ProteinResidues

Principal component analysis on the residue composition of the sectors

PCA on residues composition of sectors was done to find enrichment and depletion in co-evolving sectors of specific amino acids. The matlab script residuePCA.m was used to calculate the eigenvectors (or modes) that show variance within the sectors on the amino acid level. Modes 18-20 give the modes that contribute the largest variance to the sectors. The sign of the eigenvector values are arbitrary and in order to derive the meaning in "biological space" one must compare the PCA value of a given amino acid in SDB to the aforementioned enrichment.

Principal component analysis (PCA) on the 3D coordinates of the residues in the sectors

PCA on sector coordinates was used as a proxy for the shape of co-evolving sectors across multiple protein families in the transmembrane SDB. The matlab script coordinatePCA.m (found in the github repository) extracts the coordinates from the membrane SDB calculates the mean and difference average to produce a covariance matrix. From the covariance matrix the eigenvectors and eigenvalues were derived to yield the directions in space and stretch in that direction respectively. The eigenvalues were then used as a proxy for the shape of the sectors. In this analysis sectors containing less than five amino acids were removed in order to avoid inconsistent signal in the shape. Because we wanted to compare their deformation in space and relative shape, we normalized the eigenvalues of each sector by their sum.

Clustering of PCA on Sectors

To identify possible groups of sectors with a common shape, we employed an unsupervised hierarchical clustering method on the z-score of the normalized eigenvalues. We used the z-scores of the eigenvalues and their raw values because of the large difference in range of values and the relatively small variances of the normalized eigenvalues. Distances between each pair of sectors was calculated using the euclidean metric and a dendrogram showing the relation between sectors was computed using the average linkage function in Matlab. This unsupervised clustering method helped us define the number of clusters ; we then used the K-mean algorithm, with the default input parameters in Matlab, to define those clusters. The algorithm was run 10 000 times and the clusters with the lowest overall sum of distances were picked.

Biochemical Composition of Sectors

The biochemical composition of sectors was determined from the online resource: http://peptide2.com/N_peptide_hydrophobicity_hydrophilicity.php. Hydrophobic residues were considered: FILMVWAP. Acidic residues were: DE. Basic residues were: RKH and neutral residues were: GSTCNQP

p53 Mutation

Acquiring Database

The p53 Database was initially downloaded from the International Agency for Research on Cancer. We selected the human somatic cancer database, downloaded the text file, and pasted the text file into Microsoft Excel. Once the database was downloaded, we trimmed the dataset to include only the relevant information for our mutational analysis.

Enrichments of Mutations

Enrichment of mutations in each sector was measured by the following equation:

No. of Mutations in SectorNo. of Total Mutations/ No. of Residues in SectorNo. of Residues in the Protein

For each mutation, we checked whether the residue index was in one of the sectors, and the count of mutations in this index was added to the count of mutations in the respective sector. Mutations on residues outside of the sectors were neglected. Chi squared tests were performed on each sector to identify significantly enriched and depleted sectors.

Existence of Carcinogenic Mutations in non-human species

For each mutation observed in sectors, we examined if the same mutation occurred in non-human species. In the MSA of non-human sequences, we recorded the sequences that contain the specific mutated residue and the overall number of the sequences with the mutation. The counts of sequences with mutations were plotted for each sector to visualize the amount of clinically found mutations naturally occurring. Subsequently, we used the Kruskal-Wallis test and Tukey's HSD tests in MATLAB to compare the groups for significant differences in the counts.

Compensatory Mutation Candidates

To select compensatory mutations from the list of somatic mutations, we considered three computational features central to compensatory mutations: (a) the residue on which the mutation happens has to co-evolved with the residue with compensatory mutations (joint entropy > 0.4), (b) for a mutation to be compensated, there has to be a certain percentage (>90%) of non-human sequences with mutations on coevolving residues, and (c) among these mutations on coevolving residues, a certain number (bigger or equal to 15) of sequences should contain the same pair of compensatory mutations. As a result, we wanted to screen the MSA with these three criteria in sequence to obtain a list of compensatory mutations.

First, for each residue with somatic mutations recorded, we calculated the residues that had joint entropy over 0.4 with it. For each somatic mutation, we verified that there were coevolving residues with the residue that it was on. Then, we extracted the sequences with the somatic mutation, and observed if there were compensatory mutations on the residues coevolving with it, based on criteria (b) and (c). The remaining mutation pairs were stored in a cell array of MATLAB, with the first four columns the residue index of the somatic mutation, the wild type AA transformed to int using aa2int function in MATLAB, the mutant AA as an int, and the cluster index the residue was in. The fifth column of the array contained the sequences with the somatic mutation and its compensating mutation. The last four columns include the information of the compensating mutation, aligned the same way as the somatic mutation.

Sector Visualization

Sector coordinates were extracted from the Membrane Sector Database and run through a MATLAB function named sectorsOverlay.m (included in supplemental methods) to produce a PDB file that includes all co-evolving sectors in a single protein. The individual sectors were labeled by b-factor.

Github repository

All scripts, functions and datasets used in this study are available in the following Github repository: https://github.com/mfincker/protein_coevolution. Large MSA files were not uploaded on Github because of their size but are available upon request.

Acknowledgments

We would like to thank Alexandre Colavin for the instrumental development of MISC algorithm as well as providing guidance and discussion throughout the development of the project. Additionally, we would acknowledge KC Huang, Tiffany Vora, and Justin Sonnenburg for their advice throughout the development of this project.

References

1. Anfinsen CB (1973) Principles that Govern the Folding of Protein Chains. Science 181: 223–230.

2. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein Sectors: Evolutionary Units of Three-Dimensional Structure. Cell 138: 774–786. doi:10.1016/j.cell.2009.07.038.

3. Reynolds KA, McLaughlin RN, Ranganathan R (2011) Hot Spots for Allosteric Regulation on Protein Surfaces. Cell 147: 1564–1575. doi:10.1016/j.cell.2011.10.049.

4. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R (2004) Structural determinants of allosteric ligand activation in RXR heterodimers. Cell 116: 417–429.

5. McLaughlin RN Jr, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R (2012) The spatial architecture of protein function and adaptation. Nature 491: 138–142.

6. Sievers F, Wilm A, Dineen D, Gibson TJ (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems ….

7. Lerch-Bader M, Lundin C, Kim H, Nilsson I, Heijne von G (2008) Contribution of positively charged flanking residues to the insertion of transmembrane helices into the endoplasmic reticulum. Proceedings of the National Academy of Sciences 105: 4127–4132.

8. Valeva A, Palmer M, Hilgert K, Kehoe M (1995) Correct oligomerization is a prerequisite for insertion of the central molecular domain of staphylococcal α-toxin into the lipid bilayer. Biochimica et Biophysica ….

9. Petitjean A, Mathe E, Kato S, Ishioka C (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. … mutation.

10. Song L, Hobaugh MR, Shustak C, Cheley S, Bayley H (1996) Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. Science.

11. Aksel T, Majumdar A, Barrick D (2011) The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. … : Structure.

12. Boehden GS, Akyüz N, Roemer K, Wiesmüller L (2003) p53 mutated in the transactivation domain retains regulatory functions in homology-directed double-strand break repair. Oncogene.

13. Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24: 603–619. doi:10.1109/34.1000236.