The Organization of Protein Sectors

Tatiana Artemova*

Proteins perform the most important activities in a cell. Protein function is defined by its structure, which is determined by the sequence of amino acids. Different protein sequences and structures often maintain similar functions. It is believed that proteins possessing similar functions have diverged recently in their evolution. However, evolutionary constraints on this process remain unknown. In this paper, I report the analysis of the organization of the "protein sectors" of correlated mutations. A protein sector is a group of amino acids that are spatially close to each other in a folded protein. Amino acids display strong correlations within a sector and weak correlations between different pairs of sectors. Strikingly, in many cases only a few positions contribute most to the correlation pattern within a sector. However, these contributions are not large enough to disregard the correlations among the remaining positions. These results can be used to introduce coupling constants in the spin system model.

INTRODUCTION

It has been known for more than 40 years that distribution of amino acids in proteins is non random [1]. Indeed, an amino acid sequence defines the structure of a protein [2], which in turn determines its function [3]. It is critical for the survival of a cell that the proteins inside the cell function properly. Therefore, both the sequence and the structure of a protein are subject to strong selection, which restricts the initial variability of amino acid sequences to limited combinations. For instance, detailed sequence analysis has revealed functionally important single residues that can rarely be replaced without the loss of a function [4].

In folded proteins, not only single amino acids but also residue-residue interactions play a crucial role in maintaining structure and function [5, 6], making it sometimes costly to replace one amino acid without changing its neighbors. In other words, two mutants with single amino acid substitutions have a protein that functions poorly in contrast to that of the double mutant. The studies of the protein families have shown that the positions of these correlated mutations tend to be next to each other spatially in the folded protein[6]. Moreover, Halabi et al. have found the "protein sectors" of correlated mutations [7]. These sectors are parts of a protein that are spatially aggregated and different from the secondary or tertiary structure. The amino acids' residues are strongly correlated within a sector and weakly correlated between different sectors. Although these sectors are believed to play an important evolutionary role, their origin and their purpose remain unknown. Indeed, while protein domains are believed to be the evolutionary units of proteins [8], the existence of the independent sectors within domains raises a question about this belief.

The analogy of the states of folded proteins to the states of the spin system at low temperatures has been successfully used to understand physics of the protein folding [9]. The existence of the protein sectors suggests that nonrandom correlations within a group of positions in the alignment can correspond to the correlated domains of the spin systems at low temperatures. However, in order to construct a useful model, the distribution of coupling constants within a protein sector should be specified.

In this paper, I report the study of the organization of protein sectors. I show that the contribution of different positions significantly differs; i.e., the distribution of their contributions is not peaked around the mean contribution. Although some of these distributions have top contributors, which are separated by a gap from the other positions of a sector, the correlation pattern within a sector cannot be explained by the correlations of all positions with the top contributor only. The results are preliminary, because the process of sector identification is not robust to the data sets. For instance, identification of the sectors in this study slightly differs from that of Halabi et al[7].

METHODS

The multiple sequence alignments of the protein domains were obtained from the Pfam database[10]. The analysis of the domain families were conducted on the full data sets of the family PF00089 (corresponds to S1A in [7]). Although this alignment corresponds to the same family as in [7], they contain larger amount of proteins (15894 in this study versus 1470 in [7]).

Sector identification

The protein sectors were identified following the procedure described in [7]. First, binary statistical coupling analysis (SCA) matrices were constructed. (See the detailed description of the construction of the SCA matrices below.) The spectra of these SCA matrices were then compared to the spectra of matrices that were obtained by the randomization of the amino acids' distributions among the proteins at each position. Thus, the overall

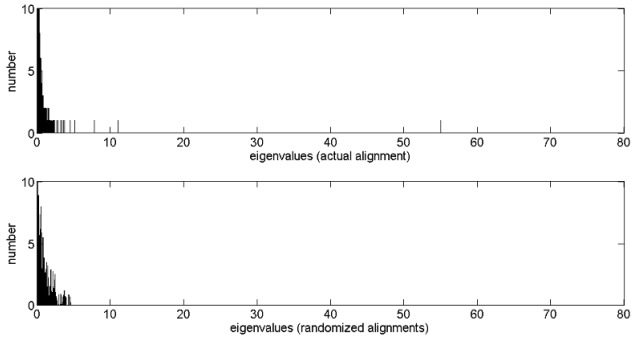

Figure 1: The spectrum of the SCA matrix includes five significant eigenvalues compared to the spectrum of the randomized matrix. The histograms of two matrices are shown: in the upper figure the histogram of the SCA matrix, in the lower figure the histogram of the randomized matrix. While the low-eigenvalue parts look similar for these matrices, the spectrum of the SCA matrix contains five significant eigenvalues, which correspond to statistically significant correlations between the pairs of the positions.

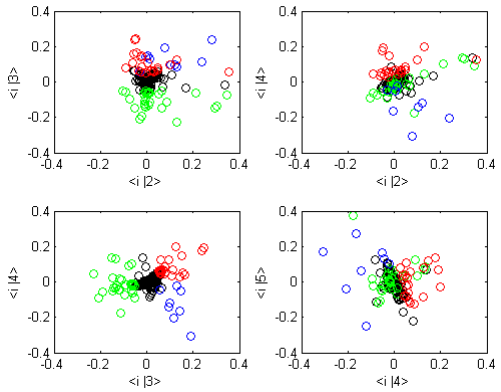

Figure 2: The contributions of different positions to the significant eigenvectors. Colors correspond to different sectors. Note that blue and red sectors correspond to the linear combinations of the third and the fourth eigenvectors.

positional conservation remained the same, while an inter positional correlation became random.

In order to find statistically significant correlations, the spectra of the SCA and the random matrix were compared. In the domain families studied, few eigenvalues significantly larger than the maximal eigenvalue of a random matrix were observed. (Fig 1.) The amino acids whose positions contribute the most to the eigenvectors corresponding to these high eigenvalues are correlated more than expected in a random matrix. However, the top eigenvalue comprises all the correlations and does not distinguish among sectors. Therefore, in order to identify statistically significant and independent sectors, the eigenvectors corresponding to significant eigenvalues (except for the top one) were considered. These eigenvectors were used to define the cleaned correlation matrix:

$$\bar{C}' = \sum_{k \in S} \lambda_k \mid k >< k \mid, \tag{1}$$

where $\mid k >$is the kth eigenvector of the SCA matrix, S is a multitude of the significant eigenvalues, except for the top one. Moreover, only positions that significantly contribute to the top eigenvectors were considered. The significance threshold for positional contribution was obtained from the comparison of actual significant eigenvectors with the corresponding vectors in the randomized matrix.

Although in [7] all the protein sectors corresponded to single eigenvectors, it is possible that in the different alignments the same protein sector corresponds either to a different single eigenvector or to the linear combinations of the significant eigenvectors. As shown in Fig 2, in the alignment used for this study, protein sectors of the PF00089 domain corresponded to linear combinations of the third and fourth eigenvectors. In general, each protein sector corresponds to a particular direction in the eigenspace of the significant eigenvectors. Thus, the process of sector identification is essentially the process of the identification of special directions in this space. The positions of the sectors are grouped along these directions, and there are no or few positions with significant projections in two independent directions simultaneously.

SCA matrix

I used the same form of SCA matrix as [7]. Following [7], I made a binary approximation and did not consider 20 amino acids. For each position a prevalent amino acid was identified, which became the first option for each position:

$$x_{i,s} = 1, \tag{2}$$

where i – position number and s – sequence number. All other amino acids at this position, including the absence of amino acids, were considered as a second option for this position:

$$x_{i,s} = 0. \tag{3}$$

The covariance matrix was then

$$C_{ij} =< x_{i,s}, x_{i,s} >_s - < x_{i,s} >_s < x_{j,s} >_s \equiv f_{ij} - f_i f_j. \tag{4}$$

An SCA matrix was obtained by weighting the covariance matrix with the gradients of the positional conservation vectors:

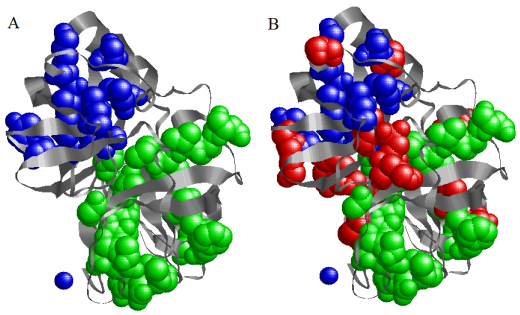

Figure 3: The space filling representation of the identified sectors proves physical proximity of the residues within a sector. In (A) green and blue sectors are shown. These two sectors are spatially separated. In (B) all three sectors are shown. Pictures were obtained using RasMol software[13].

$$\frac{\partial D_i^{(a)}}{\partial f_i^{(a)}} = \frac{\partial}{\partial f_i^{(a)}}(f_i^{(a)} ln \frac{f_i^{(a)}}{q^{(a)}} + (1 - f_i^{(a)}) ln \frac{1 - f_i^{(a)}}{1 - q^{(a)}}) \equiv ln \frac{f_i^{(a)}(1 - q^{(a)})}{q^{(a)}(1 - f_i^{(a)})} \quad (5)$$

where $q^{(a)}$ corresponded to the background sequences of amino acids.The SCA matrix was then constructed:

$$\bar{C_{ij}} = \frac{\partial D_i^{(a_i)}}{\partial f_i^{(a_i)}} \frac{\partial D_j^{(a_j)}}{\partial f_j^{(a_j)}} \mid C_{ij} \mid . \quad (6)$$

The absolute values of C_{ij} were used in order to disregard negative correlations due to the specific choice of amino acids. Absolute values still capture position-specific correlation, while do not consider amino acid-specific correlations.

RESULTS

In the domain family PF00089, three sectors were identified. Fig. 2 shows the weights of the positions along the top eigenvectors of the SCA matrix. As can be seen from several two-dimensional projections, the identified sectors formed separate clusters in the eigenspace of the SCA matrix.

Fig. 3 shows the localization of the identified sectors in the three-dimensional structure of pig trypsin (PDB 1YF4[11, 12]). While the blue and green sectors are spatially aggregated and separated, the red sector is mainly located between the green and blue sectors, while some of the red sector residues are spatially closer to the green and blue sector. This observation may also be done from Fig. 1, where only few of the red sector residues have a large projection on the second eigenvector.

Two of these sectors, the blue and green, corresponded to the green and red sectors found in [7]. The third sector (red) did not correspond to that of the previous study. This discrepancy may be explained by two factors. First, the choice of the leading directions, defining sectors, is probably not optimal in this study. Second, some of the sectors identified in a small set of data can disappear in a larger set of data. While in the original investigation, Halabi et al. studied 1470 members of S1A protease family, in this study I analyzed the alignment of 15894 protein domains.

In order to investigate the constraints that limit the size of a protein sector, I studied the distribution of the strengths of the significant correlations within a sector. (Table 1.) Strikingly, typical standard deviations of the correlations within a sector were usually almost the same as their mean values. Thus, the distributions of correlations were not peaked around some value. Moreover, a negative correlation between the size of a sector and its mean significant correlation and standard deviation was observed.

The observed negative correlation between the size of a sector and the mean significant correlation can be explained by the special organization of the position projections to the directions of the sectors. The red and blue sectors are organized in the following way: there are few leading positions that contribute most to the direction of a sector and many less significant positions, whose contributions are significantly smaller than that of the leading positions. (Fig.2.) The fewer less significant positions contribute to a sector, the higher the mean correlation within a sector.

The intuition about the role of the leading positions can be obtained from the analysis of the process of the identification of the sectors. For the sake of simplicity, I will consider two sectors with the assumption that if a particular position contributes to one sector, it does not contribute to the other. I will also assume that two domain directions correspond to the eigenvectors. The cleaned correlation matrix (equation 1) is then a block matrix that has two positively correlated diagonal blocks and negatively correlated or not correlated non-diagonal blocks. The entries of the diagonal blocks are organized as follows:

$$\lambda_m \mid m >< m \mid, \quad (7)$$

where $\mid m >$ is the eigenvector of the matrix, and λ_m is the eigenvalue of this vector. The entries of $\mid m >$ are the contributions of different positions to this direction. Thus, every position in a sector always has the strongest correlation with the leading position. With the addition of a gap between the leading position contribution and the remaining position contributions, the role of the leading position becomes even larger. Essentially, when this gap is large enough, a sector may be considered as

Table I: Summary of the properties of the proteins' sectors. Sector size is the number of positions composing a sector. Significant correlations include correlations from formula 1, while excluding self-correlation terms.

Protein domain sector color	Sector size	Mean significant correlation	Standard deviation of significant correlation
Blue	10	0.39	0.36
Green	30	0.10	0.16
Red	22	0.12	0.11

a collection of positions, highly correlated with the leading position and not correlated with each other. This conclusion remains the same if the direction of a sector is a linear combination of eigenvalues. However, if some positions contribute significantly enough to two sectors, the resulting distribution of the correlations in matrix (equation 1) significantly differs from the contributions evaluated in formula 7.

While a position that contributes most to any direction can always be found, an important question is whether the correlations of the less significant positions with the leading position is significantly larger than those between the pairs of the less significant positions. For instance, if all the correlations between pairs form a narrowly peaked distribution, a position contributing most to the direction still exists. However, I cannot say that other correlations, except for the correlations with this position, are insignificant.

In order to introduce a criterion of the significance of the leading positions, I will assume that there is a single leading position, whose contribution to the direction of the sector is y_0. The contributions to the direction of the sector made by these less significant positions is distributed with some mean (y) and standard deviation (Δy). The criterion of the significance of the top vector is then equivalent to the statement that the interaction of the leading position with the least significant position among the remaining is equal to, or larger than, the self-interaction of the most significant position of the remaining:

$$\frac{y_0(y-\Delta y)}{(y+\Delta y)^2} \geq 1. \qquad (8)$$

For the blue sector, y_0 equals 0.84; y equals 0.35; and Δy equals 0.17. (Two eigenvectors contribute to this sector. Therefore, their projections were multiplied by the square roots of the corresponding eigenvalues and summed as vectors.) Thus, the parameter in formula 8 equals 0.56 and the blue sector does not follow the criterion 8. For the red and green sectors the parameter are 0.62 and 0.63. Preliminary studies of the other protein domains show that the criterion is never valid. In some of the studied domains (PF07724), however, the parameter in formula 8 was very close to one (0.91).

DISCUSSION

The existence of the protein sectors can affect the understanding of physics or protein folding. One problem with the localization of protein sectors, however, is that the sectors can correspond to the directions in the eigenspace, different from eigenvectors. This problem can be resolved by probing different linear combinations of significant eigenvectors. The other problem is more fundamental. Although independent sectors have been observed in few protein domain families[7], the localization of these families is highly sensitive to the actual alignment of the protein domains. This observation raises a question about the independence and therefore existence of the protein sectors.

The results of this paper suggest that (i) in many protein sectors a leading position exists; (ii) however, in the most of the cases the interactions of the remaining positions with the leading position are not significant enough to disregard all the correlations between positions, except for those with the leading one. These results are preliminary. In order to study more protein domains, the algorithm for finding sectors should be improved and the discrepancies in the localizations of the sectors for different data sets should be explained.

* Electronic address: artemova@mit.edu

[1] A.V. Guzzo, Biophysical Journal, 5 (1965).
[2] C.B. Anfinsen, Science, 181 (1973).
[3] R.A Laskowski, J.D. Watson, J.M. Thornton, Nucleic Acid Research, 33 (2005).
[4] J.A. Capra, M. Singh, Bioinformatics, 23 (2007).
[5] I.N. Shindyalov, N.A. Kolchanov, C. Sander, Protein Engineering, 7 (1994).
[6] U. Gobel, C. Sander, R.Schneider, A. Valencia, Protein: Structure, Function, and Genetics, 18 (1994).
[7] N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Cell, 138 (2009).
[8] J. Thornton, C.A. Orengo, A.E. Todd, F.M. Pearl, Journal of Molecular Biology, 293 (1999).
[9] J.D. Bryngelson, P.G. Wolynes, Proceedings of the National Academy of Science in the United States, 84 (1987).
[10] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A.

Bateman, Nucleic Acid Research, 38 (2010)
[11] H.M.Berman, J.Westbrook, Z.Feng, G.Gilliland, T.N.Bhat, H.Weissig, I.N.Shindyalov, P.E.Bourne, Nucleic Acids Research, 28 (2000)
[12] B. S. Ibrahim, V. Pattabhi, Jouranl of Molecular Biology, 348 (2005).
[13] R. Sayle, E.J. Milner-White, Trends in Biochemical Sciences, 20 (1995).