DIMACS Technical Report 2001-08
February 2001

On the inter-residue correlation patterns and their role in classification of protein families

by

Boris Galitsky and Sergey Shelepin[1]

[1] iAskWeb, Inc., 681 Main Str. #B Waltham MA 02451.
http://dimacs.rutgers.edu/~galitsky
(galitsky@dimacs.rutgers.edu)

The first author gratefully acknowledges the grant NSF STC 91-19999 for the support while visiting DIMACS.

ABSTRACT

We build a novel method to calculate and analyze the correlations in mutational behavior between different positions in a multiple sequence alignment. The inter-dependence between the residues for a protein family is represented as a matrix of correlation values obeying the invariance with respect to specific amino acids, the number of sequences representing a family, the length of sequences, residue variability and the uniformity of data set representation. Common and distinguishing properties of the few protein families, including immunoglobulins, are revealed, based on the geometry of correlation matrices. We analyze the specific texture of these matrices, inherent to the specific families, and suggest a way to distinguish proteins from non-protein set of sequences.

The role of correlation matrix technique in classification is discussed. We suggest that the classification criteria should be based on the residues at the positions with the highest overall correlation with the other positions. Revealing the positions with various correlation strength helps to reconstruct the phylogeny of protein families.

Introduction

In this paper, we present a method to analyze correlations between different positions in a multiple sequence alignment. This study is motivated by an attempt to build the classification of protein families based on the data that is different from sequence and 3D structure. To build a reasonable classification of protein families, the feature set should be chosen to provide the capability to compare each pair of protein families, including absolutely different in terms of both sequence and structure. Therefore, we target the construction of the set of formal features of the protein families, providing certain invariant features relatively to the motifs, sequence length, number of sequences, residue variability, dataset non-uniformity, etc. This study addresses the construction of such invariant feature set based on the inter-position correlations. The correlation matrices are constructed as a mean to recognize the protein family by its generalized mutation patterns, as well as to distinguish the known protein family from the set of sequences of arbitrary nature.

Correlated mutations are those changes occurred during evolution to compensate for structural drift and/or to maintain stability. They can be found in multiple sequence alignments as groups of positions with similar patterns of variation. Functional constraints typically involve surface residues; mutational drift most easily occurs in loop regions not involved directly in functional interactions, and structural constraints are usually the strongest in the internal part of protein, or “core” (Taylor & Hatrick, 1994). The historical process is thought to involve a series of random point mutations, differential replication of genetic information depending on genotype and phenotype, an elimination of the cells containing dysfunctional or non-competitive sequences. When functionally negative point mutations are compensated for by other mutations, the cell may survive and with it the protein sequence information.

Correlated mutations are viewed as the networks of amino acids in protein structures with the specific coordinated changes among the individual sequences of protein family. These changes are most likely occur to keep the proteins within the certain limits of stability after the occurrence of residue variation by the natural sequence drift or adaptation towards a new functional environment. It is generally assumed that compensation by a mutation could be easier, and therefore occurs more frequently between the spatially close residues. Hence, the primary goal of the analysis of correlated mutations is to predict the contacts between the residues by means of estimation the physical distance (Shindyalov et al 1994). However, the direct intuitive approach to the contact matrix prediction by matching solely with the correlation matrix (of the same dimension) gives the limited results. The properties of the correlated matrix, obtained in our study, clarify this observation.

We build the invariant of protein family, which is based on revealing the correlation between residues, reflecting the specific diversity and evolution of this family. Building the correlation matrix with the desired invariance properties is important to verify the traditional assumption that amino acid substitutions subsequent in time occur independently of the substitutions previous in time, or that the schemes of mutations are similar at low and high sequence divergence. This assumption was shown to be incorrect; (Benner, Cohen & Gonnet 1994) showed that genetic code influences accepted point mutations strongly at early stages of divergence, while the chemical properties of the side chains dominate at more advanced stages.

Our motivations for building invariants to compare two protein families are linked to the "visual speech mapping" methodology for speech recognition. Necessity to construct an invariant representation to compare the utterances is raised by the task of recognizing a word by comparing with the prepared template. Comparison of two speech signals should occur independently on the amplitude, duration and certain frequency and temporal patterns. The invariant of speech signal (*energy(time, frequency)*) is based on calculation of Fourier transform in a sliding window; a signal is matched against the template using the topological features of the visual speech. These features include the mutual positions, relative sizes, shape and other topological patterns, invariant with respect to intensity and duration of a particular utterance.

Analogously, the paradigm of "visual mutation" serves as a uniform measure of similarity of protein families of different nature.

Method

The main difference between our computation of correlation and other approaches is that our definition of correlation is intended to achieve the invariance properties. Therefore the definition of correlation between the pair of positions is not directly based on statistical intuition; correlation is thought as a formal measure, which leads to obtaining the matrix with certain properties, planned a priory. Exact algorithm of pair-wise correlation is derived as an inverse problem, given the desired set of correlation matrices for the chosen protein families. Generally speaking, our style of the analysis of correlated mutations is more common for pattern recognition and artificial intelligence communities than for the physicist one.

There is a series of common features of correlation computation with other approaches. In one of them, the positions in the alignments are compared directly as in (Rost et al 1994); the other is based on computation of relative variation of sequences at each position (Gobel et al 1994). In the latter approach, correlated mutations are detected by the presence of similar patterns of variation between pairs of positions in the multiple sequence alignments. Our approach is discrete by nature and different from mutual information calculated as conditional probabilities for each type of residue pair at two different positions (Clarke, 1995).

The main difference between our computation of correlation and other approaches is that our definition of correlation is intended to achieve the invariance properties. Therefore the definition of correlation between the pair of positions is not directly based on statistical intuition; correlation is thought as a formal measure, which leads to obtaining the matrix with certain properties.

Exact algorithm of pair-wise correlation is derived as an inverse problem, given the desired set of correlation matrices for the chosen protein families. Our approach to the analysis of correlated mutations is more common for the pattern recognition and artificial intelligence style of analysis than for the physical one.

Suggested analysis is applicable to the results of multiple sequence alignment procedure. Our approach can serve as a good alignment verification tool: it exaggerates the alignment inconsistencies. The gaps of alignment are treated as void observations. To estimate the correlation between positions of residues for a protein family, one needs the amino-acid similarity data for this protein family. For immunoglobulin family, we used the substitution measures, obtained in the detailed linguistic-like analysis (Galitsky et al, 1998). For the rest of families we use the amino acid similarity data similar to that of the other studies of correlations between positions. We try to construct the algorithm of correlation computation, which depends in a lesser degree on the choice of classes of “equivalent” amino acids.

Traditional computation of correlation between residues

i=2
G**K**PGRP
G**R**PGRP
G**A**PGQM
G**A**PGPM
G**P**QGPR
G**P**AGSP
G**D**KGES
G**D**KGEG
G**P**AGKE
G**S**PGKD

0	1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	1
1	1	0	0	1	1	1	1	1	1
1	1	0	0	1	1	1	1	1	1
1	1	1	1	0	0	1	1	0	1
1	1	1	1	0	0	1	1	0	1
1	1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	0	1	1
1	1	1	1	0	0	1	1	0	1
1	1	1	1	1	1	1	1	1	0

Fig.1 Sample calculation of the residue exchange matrix *e* for the second position. The e_{pq} value ($p=1,\dots l$, $q=1,\dots l$, l is the number of sequences in a family) at the residue exchange matrix shows how the amino acid from the sequence *p* is different from amino acid from the sequence *q* (for the fixed position *i*). The sample matrix of all exchange pairs at position *i* (all transformations from the sequence *p* to sequence *p*) is built to represent the mutational behavior at that position. For simplicity, 0 means the same amino acids, and 1 means different amino acids for the pair of sequences *p* and *q* at the position *i.* For the actual correlation estimates, amino acid similarities from various sources are taken for corresponding families.

(Gobel et al 1994) approaches the definition of correlation starting with the mutational behavior at one position. The observed mutations at a sequence position in the protein family are taken to describe the allowed variation of residue type at that position. The matrix $s(i,k,l)$ of all exchange pairs at position *i* (all transformations from the sequence *k* to sequence *l*) is taken to represent the mutational behavior at that position (Fig.1). For two positions *i* and *j*, we compute the correlation between the corresponding exchange matrices. The observed mutations at a sequence position in the protein family are taken to describe the allowed variation of residue type at that position.
The correlation measure between the pair of position *i* and *j* is chosen as

$$R_{ij} \sim \sum_{k,l} (\ s(i,k,l\ ) - \langle s_i \rangle)(s(j,k,l\ ) - \langle s_j \rangle) \ / \ \sigma_i \sigma_j\ ,$$

where σ_i, σ_j are the standard deviations of $s(i,k,l)$ and $s(j,k,l)$ about the mean values $\langle s_i \rangle$, $\langle s_j \rangle$ correspondingly. (Neher 1994) discusses a series of ways to raise the precision of the correlation estimate for statistically non-uniform and not fully-representative data. We outline the following disadvantages of the traditional (formal correlation) approach to explain the motivations for some features of our method.

- The traditional method cannot be adjusted to provide the invariant property. It is hard to parameterize this method to ignore the statistically insignificant data on one hand and to extract correlation from the residues with high variability on the other hand.
- There is a significant dependence on the choice of residue similarity. However, the clustering of some families revealed, in certain positions exchange of similar amino acids requires dramatic exchange of residues in the other positions (Galitsky et al 1999). Therefore, mutation analysis of the exchange of similar amino acids is as important as that one of the strongly distinguishing amino acids.

- Non-variable positions are formally correlated under traditional approach. Our approach extracts the pairs of positions with correlated mutations from both non-variable and high variable non-correlated positions. It helps to build the correlation representation with the desired invariant properties.

Residue correspondence diagram

In contrast to the traditional methodology, our approach to obtain the position correlation matrix for a protein family is based on the extraction of the patterns of dependent mutation rather than on the formal definition of correlation between abstract data. Extraction of position interdependent patterns is less sensitive to statistical non-uniformity of data and allows the adjustment of parameter to achieve the invariant properties.

Let us consider the following interconnection pattern (amino acids are encoded with variables). Each row is assumed to represent the statistically significant subset of sequences.

	Position *i*	Position *j*
Protein *l*	α	γ
Protein *k*	β	δ

If α mutates in β **then** γ mutates into δ. In other words, if in position *i* the residue is *α*, then in position *j* the residue is γ and analogously for β and δ. Correlation: **Yes**.

Protein *l*	α	γ
Protein *k*	α	δ
Protein *p*	β	δ

If position *i* the residue is *α*, then in position *j* the residue is γ **or** δ, and, respectively, if β then **only** δ. Correlation is present, but weaker than the previous case. Correlation: Yes.

Protein *l*	α	γ
Protein *k*	α	δ
Protein *p*	β	δ
Protein *q*	β	γ

If position *i* the residue is *α*, then in position *j* the residue is γ **or** δ, and, respectively, if β then γ **or** δ,. Correlation: No.

Exchange matrix for two positions

For each pair of positions *i* and *j* in the aligned sequences for a protein family, we construct the matrix 20x20 {A, V, L, I,…}x{A, V, L, I,…}of the relative number of sequences (Fig.2). In this matrix, *i*-residue corresponds to the rows and *j*-residue corresponds to the columns. For example, the relative number of sequences with *i*-residue *A* and *j*-residue *V* is 0.01. We highlight the rectangles in this matrix, where any two diagonal numbers are high, and at least one number on the other diagonal is low (see grayed cells). It means, that corresponding couples *{A,I} -> {A,L}* is correlated: if *i*-residue = *A*, then usually *j*-residue = *L* and not *A*.

	A	V	L	I	…
A	0.00	0.01	0.24	0.03	…
V	0.02	0.03	0.23	0.01	…
L	0.12	0.02	0.08	0.12	…
I	0.13	0.01	0.01	0.15	…
…	…	…	…	…	…

Fig.2: Sample exchange matrix for two positions

The correlation measure is the **number of rectangles** on the exchange matrix for two positions: the higher the number of "mutually caused" residue substitution, the higher the evidence of the correlation between the corresponding pair of positions. Each rectangle at the exchange matrix is subject to such kind of analysis; the higher the number of rectangles satisfying the correlation property below, the higher the evidence that the pair of positions is correlated. Therefore, our measure of correlation is indeed the multiplication of the implicit correlation value by the confidence level that this interdependence occurs.

	A	V	L	I	F
A	0.01	D_{xy}	0.01	D_{uy}	0.03
V	0.02	0.02	0.03	0.23	0.01
L	0.03	0.12	0.02	0.08	0.12
I	0.00	D_{xv}	0.01	D_{uv}	0.15
…	…	…	…	…	… α

Fig.3: Rectangle with the nodes D_{xy} , D_{uv} ,D_{xv} , D_{uy} at the exchange matrix for two positions.

The control parameters of the correlation extraction algorithm provide the invariant properties and allow comparing the correlation matrices for two different protein families, represented by the datasets with distinguishing statistical completeness (See Fig.4).

Two positions are considered correlated, if there is sufficient statistical evidence at the exchange matrix for these positions. The number of sequences in a protein family, delivering the fact of correlation, must obey the following conditions:

1. The relative number of sequences D^q_{min}, providing the evidence of correlation via a pair of amino acids δ and χ significantly exceeds the number of sequences, voting against the correlation by means of this pair of amino acids.
2. The absolute number of sequences D_{min}, providing the evidence of correlation exceeds the threshold (assuming these sequences are indeed different ones, not obtained from the same protein by the mutations in the certain positions).

We now proceed to the expression of correlation between the residues (Fig.3), based on the values D_{xy} , D_{uv} ,D_{xv} , D_{uy} (D_{xy} is the relative number of sequences with residue α (here, x='V') in the position *i* and residue β (y='A') in the position *j*.

$(D_{xy} D_{uv} >> D_{xv} D_{uy})$ and $((D_{xy} > D_{min})$ or $(D_{uv} > D_{min}))$

or

$(D_{xy} D_{uv} << D_{xv} D_{uy})$ and $((D_{xv} > D_{min})$ or $(D_{xv} > D_{min}))$.

Hence, we have a pair of family-dependent control parameters to compensate the data non-uniformity:

D_{min} and

D^q_{min}: $D_{xy} D_{uv} / D_{xv} D_{uy} > D^q_{min}$ or $D_{xv} D_{uy} / D_{xy} D_{uv} > D^q_{min}$.

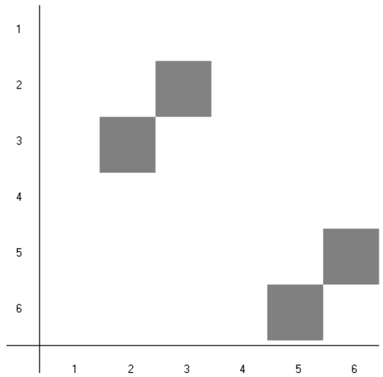
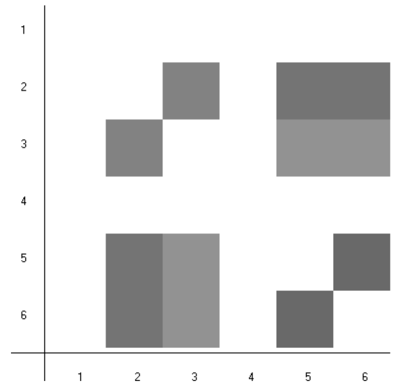
Sequences	New approach	Traditional approach
GKPGRP GRPGRP GAPGQM GAPGPM GPQGPR GPAGSP GDKGES GDKGEG GPAGKE GSPGKD		

Fig.4: Comparison of the sample correlation matrices for the dataset of Fig.1. On the left, the sample sequences are shown. In the middle, correlation matrix is built in accordance to suggested approach. On the right, correlation matrix is built in accordance to the traditional algorithm, based on the calculation of exchange matrices. Correlation matrix in the middle reflexes our intuition of correlation between positions in the higher degree, than that of the right. For positions 2 and 3, we have $D \rightarrow K$ and $P \rightarrow A$ pairs. Also, for positions 5 and 6, we have $R \rightarrow P$ and $P \rightarrow R,M$ pairs.

Results

Analysis of correlation matrices for immunoglobulin family

There are two reasons we are interested in immunoglobulin family in connection with correlation matrix technique. Firstly, since we have explored the structure of correlation between residues by the different methods, immunoglobulin family may serve as a strong verification of the consistency of invariant properties. Secondly, the correlation matrix technique, evaluated by the multilevel clustering approach, in its turn increases the latter's accuracy of extraction of the highly correlated positions. There are many variable positions, where fragment-based clustering ignores relatively infrequent, but correlated residues; calculation of correlation distribution through positions then comes into play (Galitsky & Shelepin, 2001).

Before applying the correlation matrix analysis to exploration of the classification and phylogeny of immunoglobulins, we have used the techniques of clustering and structural pattern recognition (Galitsky et al 1998, 1999). Multi-level clustering targeted the extraction of correlation among the residues within the small fragments (first level) and between these clustered fragments themselves (second level). After the fragments, correlated with other fragments, are revealed, it is possible to "restore" the structure of mutual dependence between families. In this section, we briefly present the method of revealing correlation by the cluster analysis.

At the first level of analysis, the sequences are subject to multiple alignment and division into fragments. Then the parts of sequences (words) within these fragments are subject to clustering (finding most frequent motifs). The set of words for a fragment fall into distinguishing clusters if the positions within these fragments are correlated. (Otherwise, all exhaustive enumerations of combinations of residues are obtained as clustering results). At the end of first level of clustering, all words are represented by the cluster numbers, and the whole sequence is represented as the combination of cluster numbers. Centers of word clusters are chosen to minimize the number of words occurring in two or more clusters simultaneously.

At the second level of clustering, correlation between the fragments is revealed by means of enumeration of all frequent combinations of cluster numbers. Choice of the “best” combinations is motivated by the minimization of sequences, presented by more than a single combination of cluster numbers. After the second level clustering succeeds, obtained correlation between fragment clusters is transferred “back” to the level of individual residues.

Multilevel approach to correlation extraction is based on the hypothesis that there are “close” interdependence within fragments and “remote” correlations between the positions in the different fragments. Current study shows that it is not the case. We discovered that not only for immunoglobulins, but for all analyzed families, correlation distribution depends neither on the distances between the residues in a sequence nor on the distance in the protein fold.

Immunoglobulin family indeed served as a strong verification for the method consistency. Comparing the high/low correlation positions, revealed at the correlation matrix, with the results of multilevel clustering, we can reveal how situations with various correlation strength, interdependence configuration and variability are reflected by the visualization of correlation matrix.

To characterize each position with respect to the strength of correlation with the majority of other positions, we compare the overall correlation distribution through positions for each family. It is calculated for a given position as a cumulative sum of the values of correlation with the other positions. The shape of the overall correlation strength distribution can serve as the family identification and is used to reveal common and distinguishing properties of protein families.

Heavy chains of human and mouse are the example of strong coincidence of correlation picture, though there are noticeable differences, in particular, at the very beginning of the sequence (Fig.5). Also, common for the immunoglobulins drop of correlation at C-CC’ fragments occurs in a stronger degree than for the other chains. Heavy chain demonstrates stronger distribution alternating property than the other chains, though this property is inherent to the whole immunoglobulin superfamily.

For the most of fragments, if correlation configuration between Human and Mouse are similar, the same is true for the corresponding residues. There are situations, where the correlation distribution of a fragment is almost identical, but the actual residues are different. Typical C’-motifs are rather different: for *Human* they are EWVGXI, EWLGXT and for *Mouse* - ELVAVI, EWIGXI, EWIAAS, where X – an arbitrary residue (very variable position); at the same time, the curves are almost alike.

At the same time, inverse situations are frequent, when we have the distinguishing correlation configurations for the similar motifs. For example, distribution curves look rather different, but the corresponding 0A-motifs are almost the same: 0A={QVQ, EVQ} for Human and 0A={EVQ,QVQ,EVK} for Mouse. The second position is supposed to be non-correlated (conservative) but it is not the case for Human when the residue V mutates into I, depending on other positions in the sequence. We see from the correlation matrix that it occurs analogously to the first position.

Beta chains have higher variability of positions relatively to the Heavy, Kappa and Lambda chains (Fig.6). However, the correlation matrices for Human and Mouse obey the common properties of these chains: low correlated stripe around C strand, correlation alternating, significant portion of low-correlated variable positions and characteristic rectangular texture.

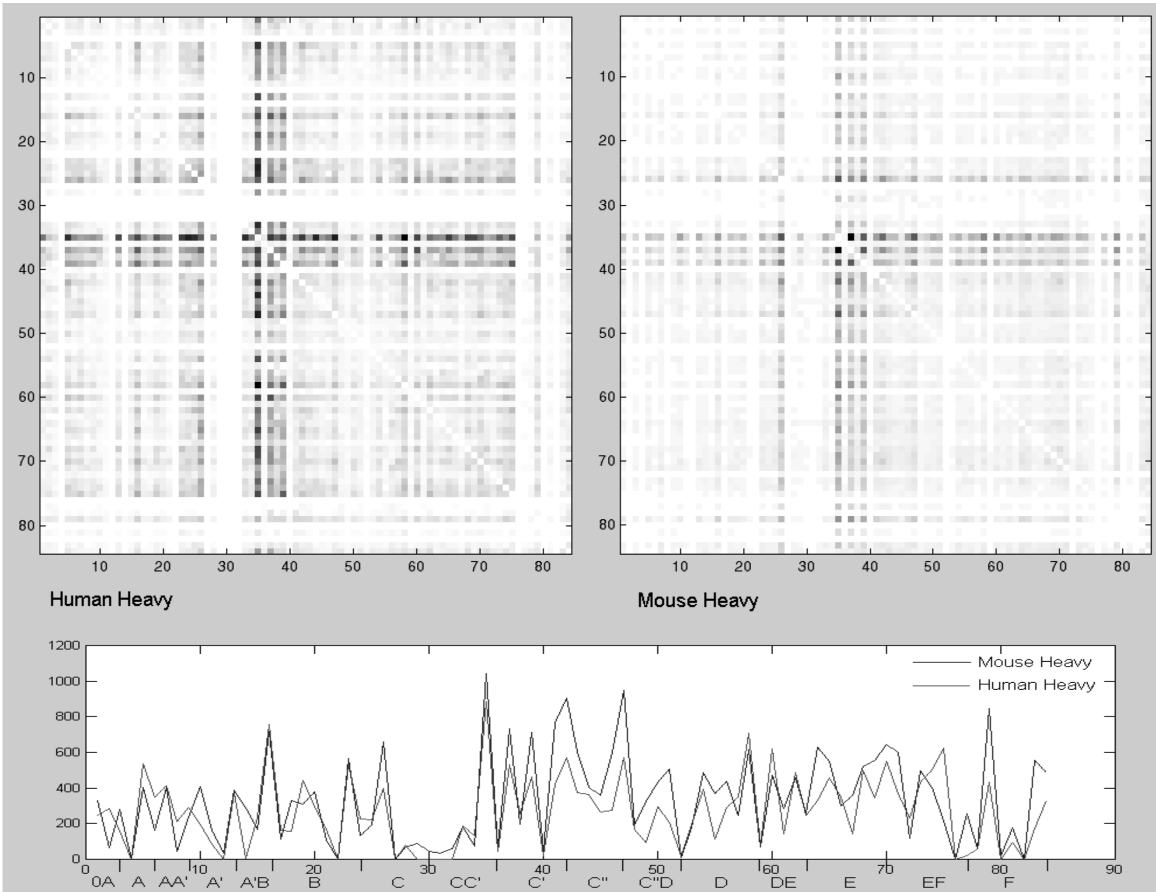


Fig.5: Correlation matrices for Heavy chains of immunoglobulins (Human and Mouse). On the bottom, overall correlation distribution is shown for this chain.

For the **Beta chain**, the difference between Human and Mouse lay in the beginning and in the end of the sequences (Fig.6). One can see the set of the continuous highly correlated positions in the second half of the sequences (fragments D-F) cut by the position E4, #54. For the Mouse Beta sequences, non-correlated positions of these fragments are almost absent. In contrast to other chains, beta chain has much higher number of classes of sequences within a family, staying apart of each other, so correlation is so strong.

Though the similarity between Human and Mouse for Beta chains is not as strong as that for the Heavy chains, the overall correlation strength correlations curves look rather alike. It is easy to reveal continuously high correlation area on the distribution curve, but each species have its own positions with the highest correlation strength (#56 for Mouse and #48 for Human. At the higher correlation area on the second half of the sequence, Mouse and Human are rather "desynchronized" relatively to the first half of the sequences. The first and second positions bring an exception to this observation. The value for correlation distribution for Human is higher than for Mouse, which is in agreement with the fact that 0A1 for Human ranges through {E,G,N,D,V} and for Mouse ranges through { G,N,D}. However, the second position is conservative for Human (0A2 = {A}) and rather variable and interdependent with 0A1 and 0A3 for Mouse (0A2 = {A,G,T,S,P}). Also note that Mouse distribution is subject to alternation in a higher degree than the Human one.

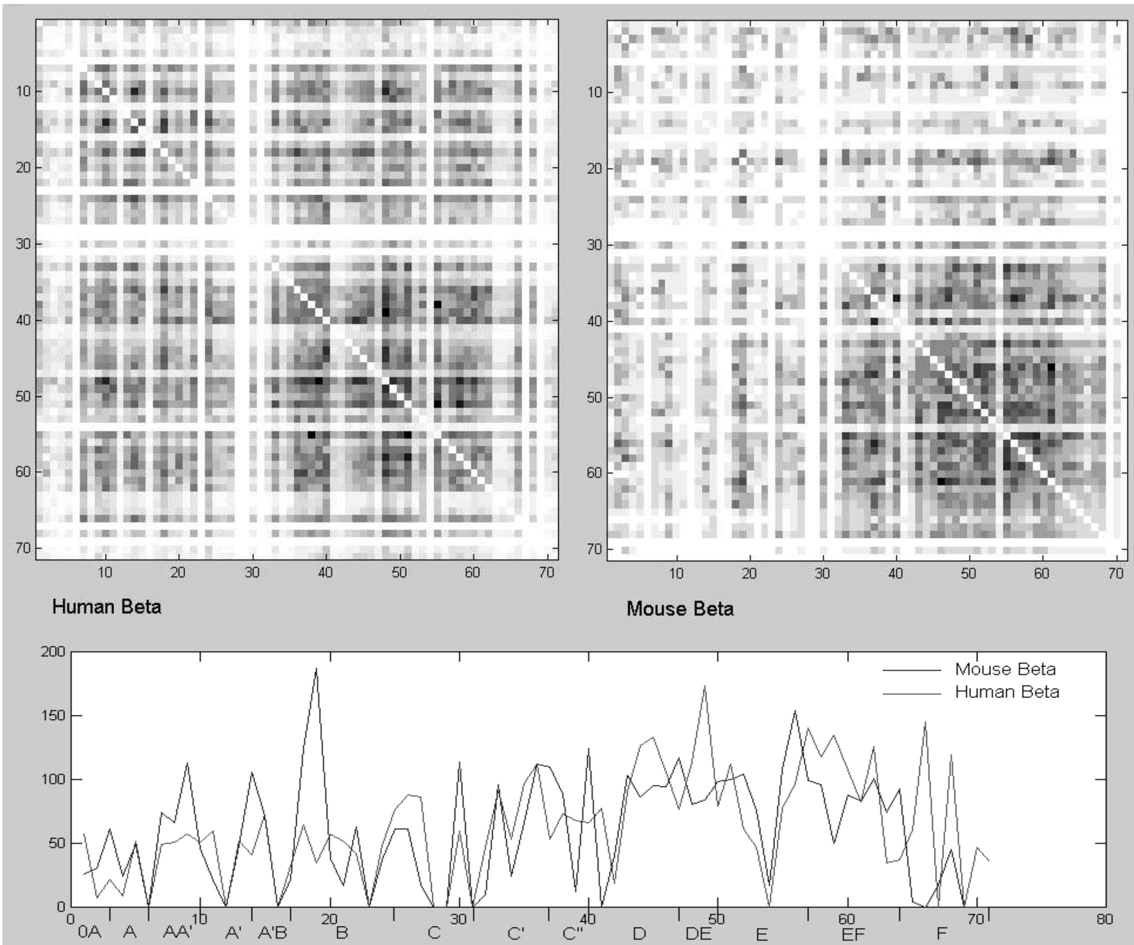


Fig.6: Correlation matrices and correlation distribution for Beta chains.

Generally speaking, correlation matrices of **Kappa and Lambda chains** obey the common properties of immunoglobulin chains (Fig.7). However, the area of lower correlation is longer than for the Heavy and Beta chains and is shifted from C strand of Heavy and Beta chains towards D strand (positions #50-60).

The geometric patterns of alternating correlated and not correlated lines are repetitive, but they do not follow the alternations of strands and loops. The strips in accordance to the scheme “low-high-low-high-low” are typical correlation patterns for immunoglobulins. There is the common set of highly correlated positions (for example, OA1, A’4, B7, C2). Spikes on the overall correlation distribution may shift by 1-2 positions because the length of sequence fragments varies correspondingly for Kappa and Lambda chains. We see that the sequences of the same chain but different species are close than that of the same species but different chains. Note, for example, similarity of areas around position 30-35 for the Lambda chains. The same observation is true for Heavy and Mouse chains. In general, the difference between Human and Mouse for Kappa and Lambda chains is stronger than that of the Heavy chains.

Note that correlation matrix allows performing the comparative analysis of families with varying statistical significance of data (up to 20 times more sequences for Mouse Kappa than for Mouse Lambda).

All considered above families show that the similarity between species for the same chain is higher than similarity between immunoglobulin chains.

Using the correlation matrix technique to reveal the positions with highest correlation strength gives a better accuracy, than the multilevel analysis of immunoglobulin sequences, based on their position-based and fragment-based clustering (Galitsky et al 1999). It is discovered, that there are positions with higher predictive power (in respect to constraints on the other residues) than those positions with lower variability used for classification of heavy chains of variable domain. In particular, the clustering study of heavy chains revealed the strand E as the location of positions, possessing the strongest correlation with the other ones. Therefore, strand E served as a basis for division into classes, and other fragments with the weaker overall correlation perform the division into subclasses. The reader can see classification of heavy, beta, kappa and lambda chains of immunoglobulins at http://dimacs.rutgers.edu/~galitsky/protein/Ig.

In terms of phylogeny it means that the following order is the good approximation of the family history. Assuming the family is originated by a single sequence, the position with the highest correlation is the point of the first mutation, such that the resultant sequences form two different classes, developing independently and splitting in the other positions.

Analysis of the correlation matrix for human heavy and mouse heavy suggested thet initial point of mutation is the position C2. Therefore, technique of correlation matrices helps to improve the classification accuracy, working complementary to the clustering approach.

Common and distinguishing features of protein families, revealed using the correlation matrix

Analysis of visualization of correlation between residues yields a novel view of a protein families, using abstract geometric configuration instead of the explicit motifs or folding properties, inherent to each family. Suggested methodology can help in solving two following problems:

1) How to distinguish given protein family from the other protein families;
2) How to distinguish the set of aligned sequences from the sequences of non-protein nature.

The common features of proteins are meant to help us with the second problem. The following common protein properties have been revealed as the result of correlation analysis:

- For any position, if it is **correlated** with the **specific set of highly correlated positions**, than it is likely correlated with the majority of other positions. Vice versa, if it is not correlated with this set, it is unlikely correlated with the rest of positions.
- Correlation strength weakly depends on the residue **variability**.
- Correlation strength does not depend on the **distance** between positions in the sequence.
- Protein correlation matrix has the well-structured **rectangle block texture**, including grids and cross patterns.
- The geometrical patterns of correlation matrices usually **do not match** ones of the matrix of **contacts** between residues. Therefore, correlation information has a limited value for the direct fold prediction. Also, it is hard to find an interdependence between the overall correlation distribution and the secondary structure representation.
- All analyzed protein families demonstrated the **periodic** appearance of areas with low correlation. The period is 70±20 residues.

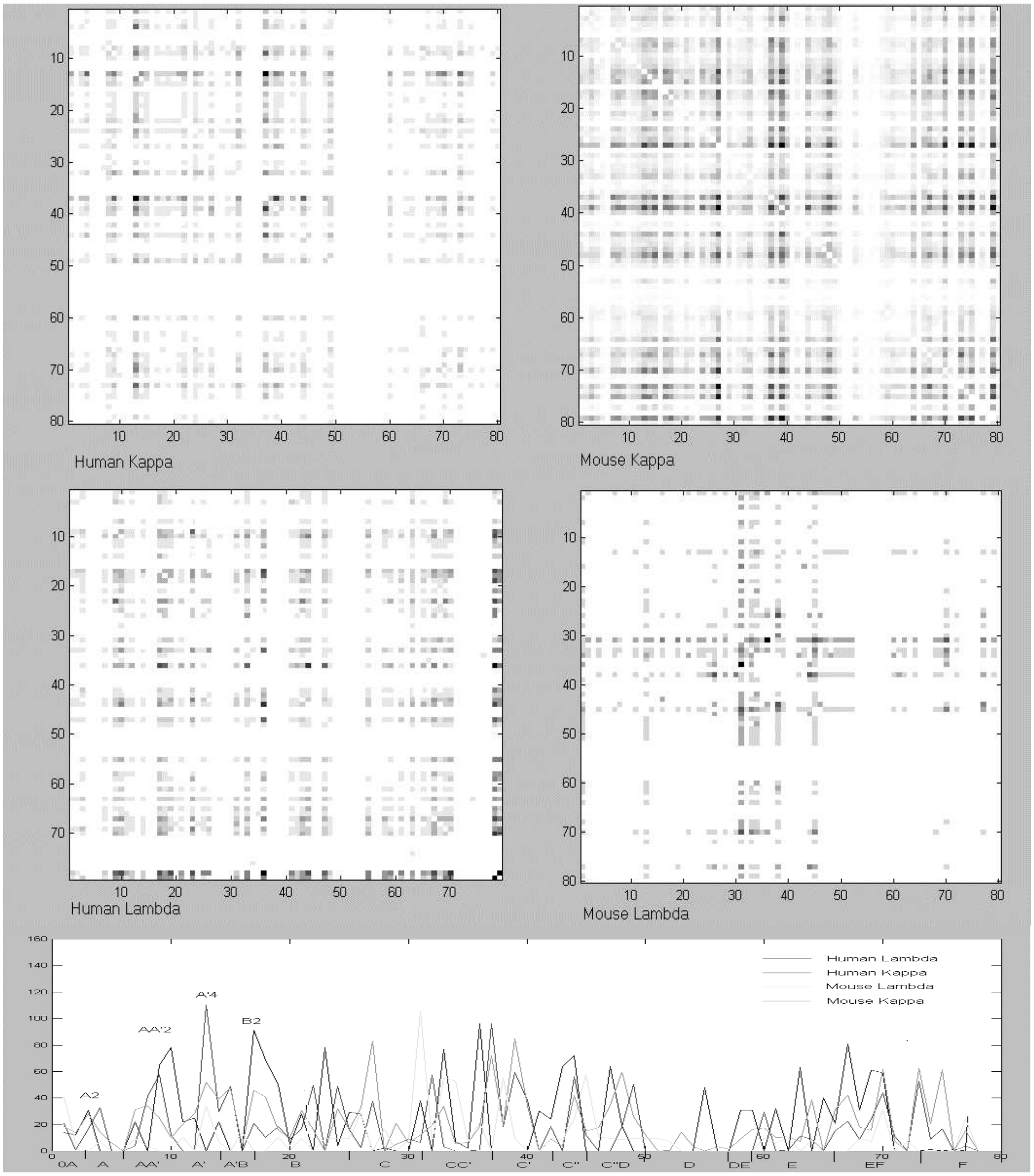


Fig.7: Correlation matrices and correlation distribution for Kappa and Lambda chains.

- Analysis of correlation distribution allows to determine the positions of the highest correlation strength. These positions are optimal to build the classification of protein families, serving as the criteria of division into classes. Besides, the positions with the highest correlation strength are likely to be the locus of the initial mutation point, assuming a family comes from a single sequence.

Correlation matrix presents a convenient tool to distinguish protein families. We developed the image processing algorithm to extract the rectangular areas on the correlation matrices and analyze its structure (Fig.8,9). Besides, the procedure of automated match of areas with similar topology is currently under development. It helps to extract the common features of the protein families with distant properties.

Occurrence of crosses, grids, angle-shape and rectangular patterns (Figs.10-13) and their combinations can serve as a strong criterion of the family "signature". In addition, uniform and non-uniform alternation of correlated and non-correlated areas demonstrates the individuality of protein families. Therefore, a set of sequences can be related to a protein family **without** knowledge of the specific residues. Besides, correlation matrices can be used as a framework for protein families classification, complementary to one based on the properties of sequence and structure.

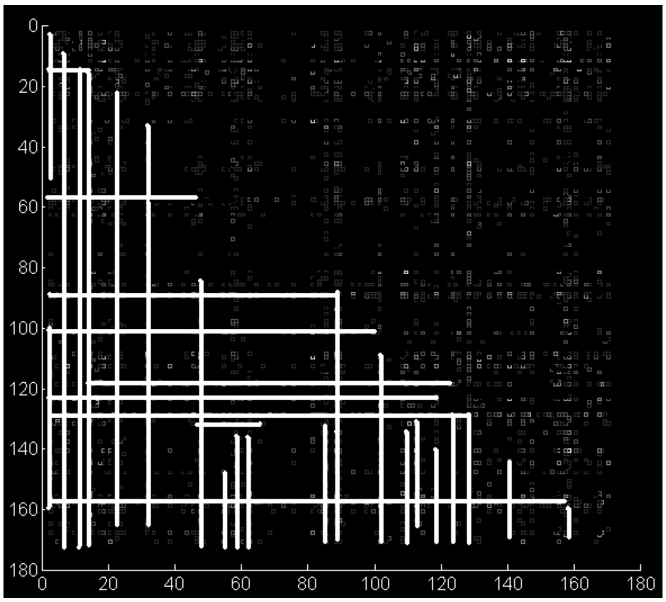


Fig.8: The correlation matrix with original top-right corner and extrapolated bottom-left corner for the ABC family (on the left). Note the matrix elements, deviated from the derived bars. In contrast to the other figures, here the lighter the cell the higher is the correlation value.

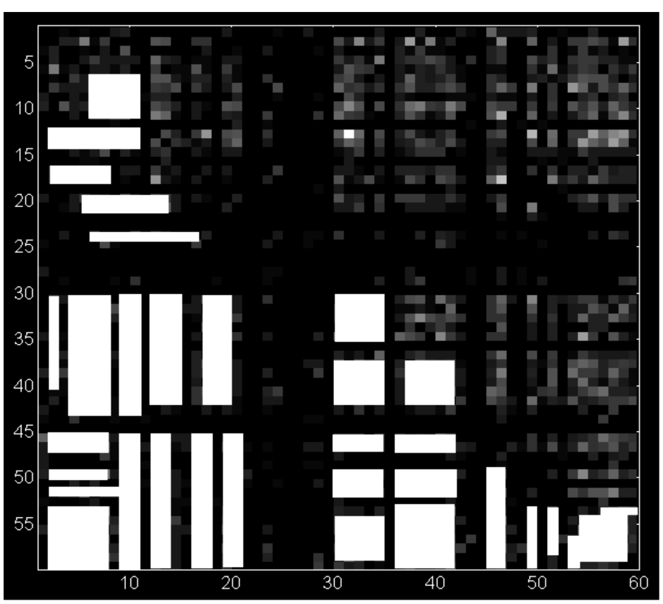


Fig.9: The correlation matrix with original top-right corner and extrapolated bottom-left corner for the C2 family (On the right). Combinations of rectangles and bars forms the C2-specific the correlation matrix for the protein family representation. Note the matrix elements, deviated from the derived bars.

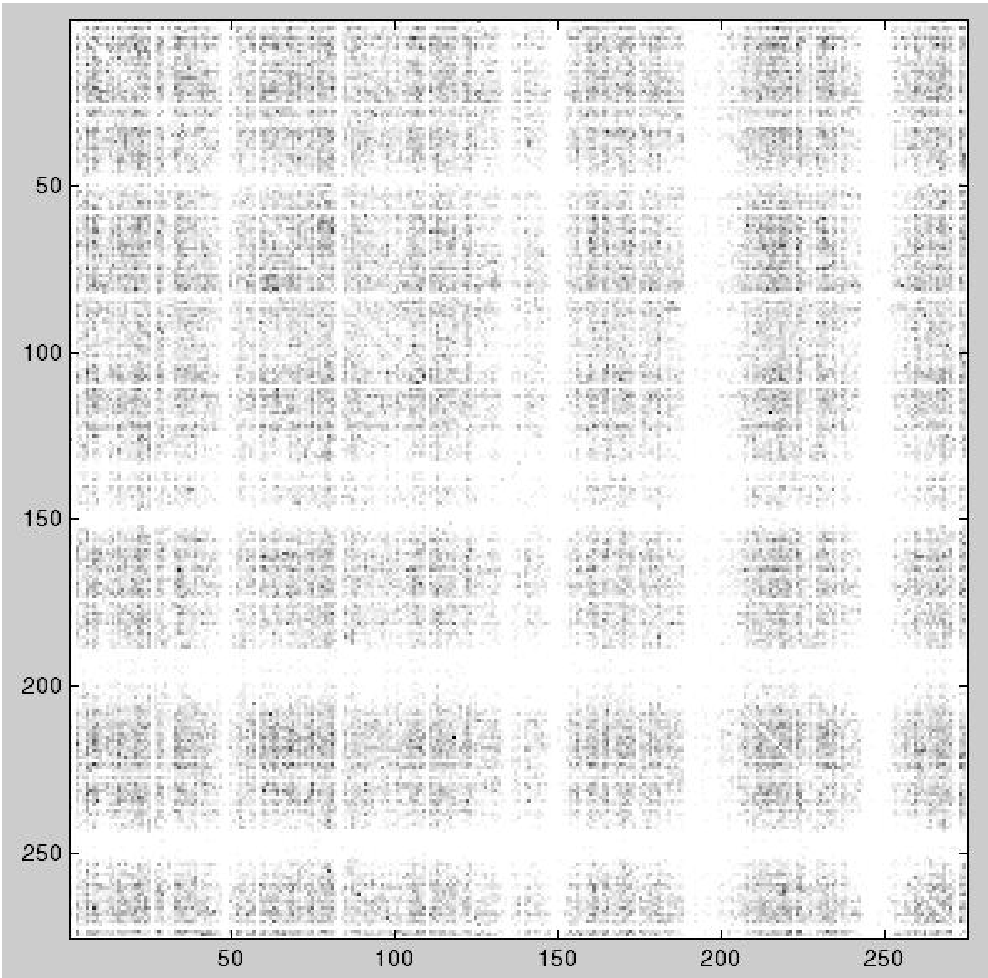


Fig.10: Correlation matrix for 7tm_1 (7 transmembrane receptor (rhodopsin family)). Distinguishing property of this family is the alternating non-correlated strips on the second half of sequence. At the same time, the correlation strength on the correlated strips on the rights is as high as on the left area with the narrow non-correlated strips, common for all protein families.

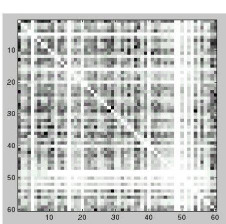


Fig.11: Correlation matrix for the homeodomain family (60 residues). It is shown in the same scale as the 7tm_1 (275 residues) family above to demonstrate the pattern similarity. Note the similarity in geometrical features (one period for homeodomain and 6 periods for 7tm_1) of alternation of correlated/non-correlated areas (approximately 50/10 residues) for rather different protein families (Compare with Clarke, 1995).

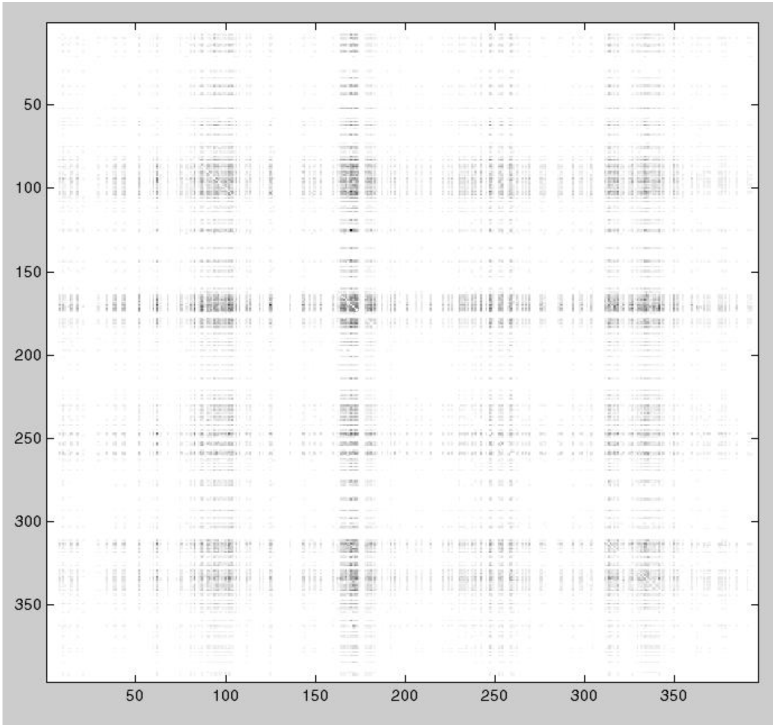


Fig.12: Correlation matrix of Chlamidia. Note the alternation periods of about 60 residues, similar to that of for 7tm_1 and homeodomain families. Chlamidia has the specific correlation pattern: the rectangular areas with “fences”.

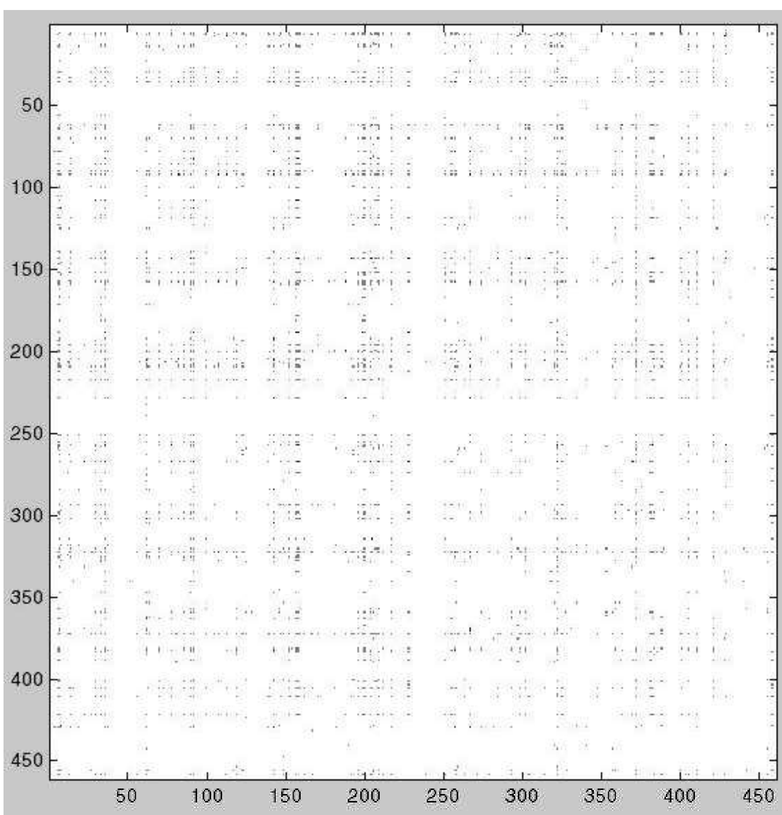


Fig.13: Cox1 protein family. Dot-type of patterns on the correlation matrix has the strong periodicity. Not only the alternation of areas with low and high correlation is periodic, but also the series of concentric crosses has the similar periodicity (55 residues).

Calculate the correlation matrix for your protein family

Please enter protein sequence to the form below in the plain text format (multiple alignment is required). Then enter two parameters (Quantity and Ratio, see method description) and click the Draw button.
You will obtain the visual representation of the correlation matrix (The darker the image cell, the higher is the correlation of the corresponding pair of positions).
You can get the sample set of sequences of various nature here. See here the computed correlation matrices for major protein families from PFAM. To calculate the correlation matrix using the standard method, go here.

Enter protein sequence here.

Quantity parameter 2 Ratio parameter 0.1

Automatic adjustment of parameters

Sample sequences of proteins and non-proteins for the method verification
You can open the file with the sequences and copy them into the form for obtaining the correlation matrix

Source	**Comments**	**Recommended parameters**
Pfam: Seed Collagen	Protein alignment from pfam.wustl.edu	Quantity =2, Ratio=0.1;
Pfam: Seed rvp	Protein alignment from pfam.wustl.edu	Quantity =2, Ratio=0.3;
Language sample	A set of related natural language sentences. All characters outside amino-acid alphabet are replaced	Quantity =1, Ratio=0.1;
Numerical sample	Digits in the double-format representation of π, 2π, 3π, …, converted into one-letter amino-acid code.	Quantity =3, Ratio=0.3;

Fig.14: Online computation of correlation matrix.

Discussion

At the same time, the visualization of correlation matrix can serve as a means to highlight alignment irregularities.
We outline the following advantages of the correlation matrix, delivering protein family invariant:

- Provides parameters, compensating residue variability.
- Provides parameters, compensating non-uniformity of the dataset (repetitions of the similar sequences, missing fragments).

Note the following objective limitation of the correlation matrix visualization for both traditional and suggested approaches. Correlation matrix as an image does not provide a means to link the neighboring cells. If the original set of sequences is “reshuffled” position-wise, the resultant correlation matrix would accordingly exchange the rows and columns. Therefore, correlation matrix does not encode the neighborhood relationships between the positions. Visualizations of the other nature (scenes, 2D distributions of parameters, visual speech signals) reflects the interconnection of the neighboring cells. Furthermore, statistical dependence between the close pixels is a basis for studies about real-world images.

To overcome this limitation, we explored some correlation algorithms, involving the computation of correlations between the triples of position, or correlation extraction within a sliding window. The resultant value would correspond to a first involved position in the first case and to the window center in the second case. However, these kinds of averaging led to distortion of the resultant invariant features of the correlation matrices.

We present the survey of some studies, addressing various issues of correlation between residies, residue conservation and variability and their connection with 3D structure and dynamics of folding.

Extraction of correlations

(Taylor & Hatrick, 1994) introduced a method to identify compensating changes between residues at positions in a multiple sequence alignment. A correlation-based method was used to measure the compensation found in the four residues at a pair of positions in any two sequences in a multiple alignment. All possible sequence pairings were measured at the pair of positions and the resulting matrix analyzed to give a measure of cooperative behavior among the pairs. Pairs of compensating positions were selected, and their mean separation (in a protein of known structure) was compared to both the mean pair-wise separation over all residues and the pair-wise separation over an equivalent sample of pairs of residues selected on the basis of their conservation alone.

Investigations of local sequence motifs that go beyond protein family boundaries showed the following (Han & Baker 1995). Naturally occurring protein sequences are significantly more clustered than the corresponding random sequences for window length ranging from one to 13 contiguous positions. The patterns of variation at the single positions are chemically similar amino acids, which tend to be grouped together. When the consideration window increases, the recognizable patterns of hydrophobic and hydrophilic residues are revealed, as well as highly conserved glycine residues. The abstraction methodology of this approach is similar to the way we compare different protein families by means of revealing common geometrical patterns on the correlation matrices.

(Taylor & Hatrick, 1994) introduced a method to identify compensating changes between residues at positions in a multiple sequence alignment. A correlation-based method was used to measure the compensation found in the four residues at a pair of positions in any two sequences in a multiple alignment. All possible sequence pairings were measured at the pair of positions and the resulting matrix analyzed to give a measure of cooperativity among the pairs. Pairs of compensating positions were selected, and their mean separation (in a protein of known structure) was compared to both the mean pair-wise separation over all residues and the pair-wise separation over an equivalent sample of pairs of residues selected on the basis of their conservation alone.

Sequence conservation and sequence correlation are two of the main properties that can be derived from the analysis of multiple sequence alignments. In accordance to (Olmea et al 1999) sequence conservation is related to the direct evolutionary pressure to retain the chemical characteristics of some positions in order to maintain a given function. Sequence correlation is attributed to the small sequence adjustments needed to maintain protein stability against constant mutational drift. It was discovered that sequence conservation and correlation were each frequently informative enough to detect incorrectly folded proteins. Furthermore, combining conservation, correlation, and polarity led to almost perfect discrimination between native and incorrectly folded proteins. This information was used for threading by evaluating the models suggested by a threading method according to the degree of proximity of the corresponding correlated, conserved, and apolar residues.

It was discovered (Wood & Pearson 1999) that while most protein families show high correlations between sequence similarity and structure similarity, the amount of structural change per sequence change (structural mutation sensitivity) varies almost fourfold; there is no direct connections between structural mutation sensitivity and either protein family mutation rates or protein size. These observations support our intentions in building the protein family invariant with respect to sequence length and mutation rates itself.

An adjustable fitness model for amino acid site substitutions is investigated by (Dimmic et al 2000). This model, a generalization of previously developed evolutionary models, has several distinguishing characteristics: it separately accounts for the processes of mutation and substitution, allows for heterogeneity among substitution rates and among evolutionary constraints, and does not make any prior assumptions about which sites or characteristics of proteins are important to molecular evolution.

New computational models of natural site mutations are developed that account for the different selective pressures acting on different locations in the protein. The number of adjustable parameters is greatly reduced by basing the models on the underlying physical-chemical properties of the amino acids. This allows (Koshi & Goldstein 1998) to use their method on small data sets built of specific protein type and to demonstrate that with this approach the evolutionary patterns in HIV envelope proteins can be represented far better than with more traditional methods.

Inter-residue correlations and fold prediction

Prediction of the protein folding class, based on global protein chain description, has been performed by (Dubchak et al 1995). Descriptors include overall composition, transition, distribution amino acid attributes, such as relative hydrophobicity, predicted secondary structure , and predicted solvent exposure.

As we mentioned in the introduction, the fold predictions generated with the use of residue correlation information were clearly better than random but not enough for their use in de novo protein folding experiments. (Olmea & Valencia 1999) assess the possibility of improving contact predictions combining information from the following variables: correlated mutations, sequence conservation, sequence separation along the chain, alignment stability, family size, residue-specific contact occupancy and formation of contact networks. The application of a protocol for combining these independent variables leads to contact predictions that are on average two times better than those obtained initially with correlated mutations. Correlated mutations can be effectively combined with other types of information derived from multiple sequence alignments. Among the different variables tried, sequence conservation and contact density are particularly relevant for the combination with correlated mutations.

The model of folding transition state formation is built in (Shakhnovich et al 1996), where the creation of a specific subset of the native structure (a folding nucleus) requires that some amino acids form most of their contacts in the transition state, and others only do so when reaching the native conformation. Therefore, mutations affecting the stability of the transition state nucleus should stronger affect the folding kinetics than other mutations, and the residues involved should be evolutionary conserved.

(Koshi and Goldstein 1997) investigated how the properties of individual amino acids result in proteins with particular structures and functions. The link between structure-dependent mutation rates and changes in various physical-chemical properties of the amino acids such as volume, charge, alpha-helical and beta-sheet propensity, and hydrophobicity, has been examined. Mutation matrices for the hyper-variable and framework regions of antibody light chain V regions were derived and analyzed, and surprisingly high conservation of hydrophobicity in the hyper-variable region was found, possibly indicating an important role for hydrophobicity in antigen recognition.

Relationship between stability changes caused by buried mutations and changes in 48 amino acid properties were analyzed in order to understand the mechanism of protein stability and to develop a simple method for predicting mutation-induced stability changes (Gromiha et al 1999). As expected from the importance of hydrophobicity, properties reflecting hydrophobicity are strongly correlated with the stability of proteins. It was found that subgroup classification based on secondary structure increased correlations significantly, and mutations within beta-strand segments correlated better than did those in alpha-helical segments, which may result from stronger hydrophobicity of the beta-strands. Consideration of sequence and structural information (neighboring residues in sequence and in space) did not significantly strengthen the correlations in buried mutations, suggesting that nonspecific interactions dominate in the interior of proteins.

The classification of the environment of residues within protein structures was studies by (Deerfield et al 1997). These environments grouped residues based upon their buried surface area, polarity of the surrounding environment, and secondary structure element in which the residue is found. The approach of environmental categories was expanded by both considering additional contacts in the calculation of the solvent-accessible molecular surface area and by subdividing the environmental plot into regions based upon its theoretical features. These new categories were able to discriminate between residues such as threonine, valine, and aspartic acid while reflecting the propensity of these residues to substitute for each other.

Conclusions

We built the visualization and comparison tool for protein families, based on the novel method of extraction correlation between the residues. Though correlation extraction was initially motivated by the attempts to predict the 3D structure, we did not find the significant link between the inter-residue correlations and contacts. However, using the correlation matrix, we constructed the convenient basis for describing the invariant properties of protein families. Some of these revealed properties were unknown earlier.

The reader can build the correlation matrix for an arbitrary protein family, using the Web-based software at http://dimacs.rutgers.edu/~galitsky/protein/corr (Galitsky, 1998), see Fig.14.

We outline the following problems for future study:

- To find a solution for the inverse problem: building a hypothetical family from the given correlation matrix. We would need to
- Automatic analysis of correlation matrices towards building the classification of the high number of protein families, based on the geometrical features.

- Using automatic text annotation as an additional source of estimating the differences between the protein families. We are planning to take advantage of our natural language processing technology to extract and analyze linguistic units, expressing with the protein functions. Machine learning technique is then supposed to come into play to find the formal correlations between the weight occurrence of extracted words with the protein features, reflected at a correlation matrix.

Acknowledgements

The authors are grateful to I. Muchnik, I Gelfand, A Kister, M Roytberg and B. Mirkin for fruitful discussions and to A. Dement'ev for the assistance in computations.

References

3) Taylor WR, Hatrick K 1994 Compensating changes in protein multiple sequence alignment *Protein Eng.* Mar; 7(3): 341-8.
4) Benner S.A., Cohen M.A., Gonnet, G.H. 1994 Amino acid substitution during functionally constrained divergent evolution of protein sequences *Protein Eng.* Nov; 7(11): 1323-32.
1. Gobel, U, Sander, C., Schneider, R, Valencia, A., 1994 Correlated mutations and residue contacts in proteins *Proteins* v18 309-317.
2. Rost, B., Sander, C. & Scneider, R. 1994 PHD – a mail server for protein secondary structure prediction. CABIOS. V10 53-60.
3. Clarke, N 1995 Covariation of residues in the homeodomain sequence family. *Protein science* v4 2269-2278.
4. Neher, E. 1994 How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci, USA v91, 98-102.
5. Dubchak, I., Muchnik, I., Holbrook, R., Kim, S.-H,1995 Prediction of protein folding class using global description of amino acid sequences *Proc. Natl. Acad. Sci*.V92 pp 8700-8704 September 1995.
6. Han KF, Baker D., 1995 Recurring local sequence motifs in proteins. *J Molec. Biol*. 4;251(1):176-87
7. Shakhnovich E., Abkevich V., Ptitsyn O. 1996. Conserved residues and the mechanismof protein folding. *Nature* 379(6560):96-8.
8. Wood T.C., Pearson W.R. 1999. Evolution of protein sequences and structures *J Molec. Biol.* 291(4):977-95.
9. Olmea O, Rost B, Valencia A. 1999 Effective use of sequence correlation and conservation in fold recognition. *Fold Des* 1997;2(3):S25-32.
10. Olmea O and Valencia A. 1997. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *J Mol Biol* 12;293(5):1221-39.
11. Koshi JM, Goldstein RA 1997. Mutation matrices and physical-chemical properties: correlations and implications. *Proteins* 27(3):336-44.
12. Koshi JM, Goldstein RA 1998 Models of natural mutations including site heterogeneity. *Proteins* 15;32(3):289-95.
13. Dimmic MW, Mindell DP, Goldstein RA. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac Symp Biocomput,* 18-29.

14. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai 1999. A Relationship between amino acid properties and protein stability: buried mutations *J Protein Chem* 1999 Jul;18(5):565-78.
15. Deerfield DW 2nd, Holland-Minkley AM, Geigel J, Nicholas HB Jr. 1997. Classification of the environment of protein residues. *J Protein Chem* 16(5):441-7.
16. Galitsky, B.A.,1998. For correlation matrices of various protein families and online computation, see http://dimacs.rutgers.edu/~galitsky/protein/corr.
17. Galitsky, B.A., Gelfand, I.M. and Kister A.E. Predicting amino acid sequences of the antibody human V_H chains from its first several residues *Proc. Natl. Acad. Sci*.V95 pp 5193-5198 April 1998.
18. Galitsky, BA, Gelfand I.M., Kister A.E.,1999, Class-defining characteristics in the mouse heavy chains of variable domains *Protein Eng* . Nov 12(11):919-25.
19. Galitsky, B.A., Shelepin, S.A. 2001. Using correlation matrices in the comparative analysis of immunoglobulin chains. In RECOMB-01, Montreal, Canada.