DIMACS Technical Report 2002-18
April 2002

The set of pair-wise correlated sequence positions contributes to determination of the immunoglobulin fold

by

Boris Galitsky[1], Andrey Dementiev[1] and Sergey Shelepin[1]

[1] KnowledgeTrail, Inc., 9 Charles Str. Natick MA 01760 USA.
http://dimacs.rutgers.edu/~galitsky

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

ABSTRACT

In this study, we are resolving the contradiction that rather limited number of residues or classes of amino acids (about 10) determines the fold (Immunoglobulin-like) for the sequence of about 100 residues long. Immunoglobulin fold comprises the protein super-families with rather distinguishing sequences with less than 10% identity; their sequence alignment can be accomplished only taking into account the 3D structure. Therefore, we believe that discovering the additional common features of the sequences is necessary to explain the existence of common fold for these (SCOP) superfamilies. The analysis of pair-wise interconnection between residues of the multiple sequence alignment helped us to reveal the set of mutually correlated positions, inherent to almost every super-family of protein fold. Hence, the set of constant positions plus the set of variable but mutually correlated ones can serve as a basis of having the common 3D structure for rather distinguishing protein sequences.

Introduction

It is well known that there exist protein families with rather low sequence similarity but almost identical fold. For example, the *globin* family with <15% sequence similarity have both very close structure and functions. In this study, we consider the protein families with Immunoglobulin-like (Ig) fold (beta-sandwich that consists from 7 strands in two sheets) with less than 10% sequence identities between the families. We are looking for novel sequence similarity features in addition to the identical or alike residues under multiple sequence alignment to resolve the contradiction that just a few (<10) residues determine the common fold for the sequences of about 100 residues long (Halaby et al 1999, Gelfand & Kister 1997, Gerstein & Altman 1995, Chothia and Lesk,A. 1986).

Ig family was the focus of our previous studies of correlations between residues. Extracted interdependences between positions served as a strong verification of our methodology of computation, analysis and visualization of correlation matrices (Galitsky & Shelepin 2000), 2001). The correlation between 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. It is becoming evident that not only conserved residues contribute to the stability of 3D fold, and correlation matrix is found to be an adequate tool to perform the comparative analysis of interdependent positions with high variability. Hence, the set of constant positions within the families plus the set of variable but mutually correlated ones can serve as a basis to explain the common 3D structure for rather distinguishing protein sequences (compare with Lindahl & Elofsson 2000, Mirny & Shakhnovich 1999).

When sequence homology of a new protein with a protein of known structure is high, standard alignment techniques allow the prediction of structure. The exact degree of homology necessary to make such inference depends on homology peculiarities and structural similarity for a particular family but in the range of 25-40% of sequence identity (Sanders and Schneider 1991). Below this limit is the “twilight zone”, in which sequence homology does not imply structural similarity. In this study, we resolve the problem of twilight zone, extracting the additional information about sequences (via the mutually-correlated positions). In our earlier studies, we have started the comparative analysis of Ig- families and Ig-fold (Shelepin and Galitsky 2001, Dementiev and Galitsky 2002).

Generally-speaking, we intend to find additional common features in the sequences sharing the same fold. At this moment, we are not aware of such potentially common features except the sequences themselves and correlation analysis. There is a series of alternative ways to represent the protein sequence data, but it does not seem to be applicable in our case.

The main hypothesis we pose and verify in this study can be formulated as follows:

The same fold <= the same core positions + **common set of mutually-correlated positions**

And the main result we are expecting is

The common set of mutually-correlated positions exists for Ig-fold

Method: searching the common mutually-correlated positions for Ig-fold

In this Section, we enumerate the steps, which need to be performed to reveal whether there are common correlated positions for each Ig fold family.

1. Enumeration of SCOP-superfamilies (families) of Ig fold and their PDB representatives.
2. Circumscribing the members of each family, and obtaining multiple alignment for each family, based on PFAM database. Verification that there are no such PDB representatives among these members that have the alien folds. Choosing the statistically-significant families (PDB representatives *1a02, 1lla, 1cyg, 1cvc, 1sox, 1cob, 1pdk, 1wsp, 1qg3, 1msp,1svb*).
3. Obtaining the fragments of sequences for comparative analysis. 10 strands A, A', B, C, C', D, E, F, G, G' are selected to represent the core residues for each family.
4. Building multiple alignment of the chosen strands of selected sequence representatives from each family, based on their secondary structure. This step is the most ambiguous because various sources provide different secondary structure assignment. Results are improved by iterations (Step 4 - Step 9).
5. Obtaining the overall alignment of strands of all families for the comparative analysis of correlation.
6. Computing the correlation matrices and highlighting the correlated pairs for each constructed family and revealing the set of *mutually correlated* positions for each family.
7. Finding the *intersection* of the sets of mutually correlated positions through the totality of families and estimating its statistical significance.
8. Improvement of alignment between families (Step 4) while matching the distributions of correlated residues.

Method: calculating pair-wise correlation between residues

In this section, we present the algorithm of calculation the pair-wise correlation. We take into account the following considerations:

- Suggested analysis is applicable to the results of multiple sequence alignment procedure. Correlation matrix serves as a good alignment verification tool: it exaggerates the inconsistencies of sequence alignment and protein family assembly.
- Residue substitution measure is special for immunoglobulin family, based on our Ig-family classification approach (Galitsky, Gelfand & Kister 1998), traditional for the other protein families. Algorithm of correlation computation should depend on the choice of classes of "equivalent" amino acids in a lesser degree.

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. In this matrix, *i*-residue corresponds to the rows and *j*-residue corresponds to the columns. 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.

	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
…		…	…	…	…

Fig.1: calculating the correlation between the pair of positions

Conditions of correlation for a pair of positions is expressed as follows:

$(D_{xy} D_{uv} >> D_{xv} D_{uy})$ *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* $(D_{xv} > D_{min}))$ *and*
$(D_{xy} D_{uv} / D_{xv} D_{uy} > D^{q}_{min}$ *or* $D_{xv} D_{uy} / D_{xy} D_{uv} > D^{q}_{min})$

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*. The control parameters D_{min} and D^{q}_{min} 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.

1. Relative parameter D^{q}_{min} : the number of sequences, providing the evidence of correlation via a pair of amino acids, divided by the number 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.

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, calculated as a cumulative sum of the values of correlation with the other positions.

Analysis of Immunoglobulins as representatives of Ig-fold

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.2). 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.

In this study we employ the common and distinguishing properties of correlation matrices obtained in our previous studies (Shelepin & Galitsky 2001). Analysis of visualization of correlation between residues yields a novel view of 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.

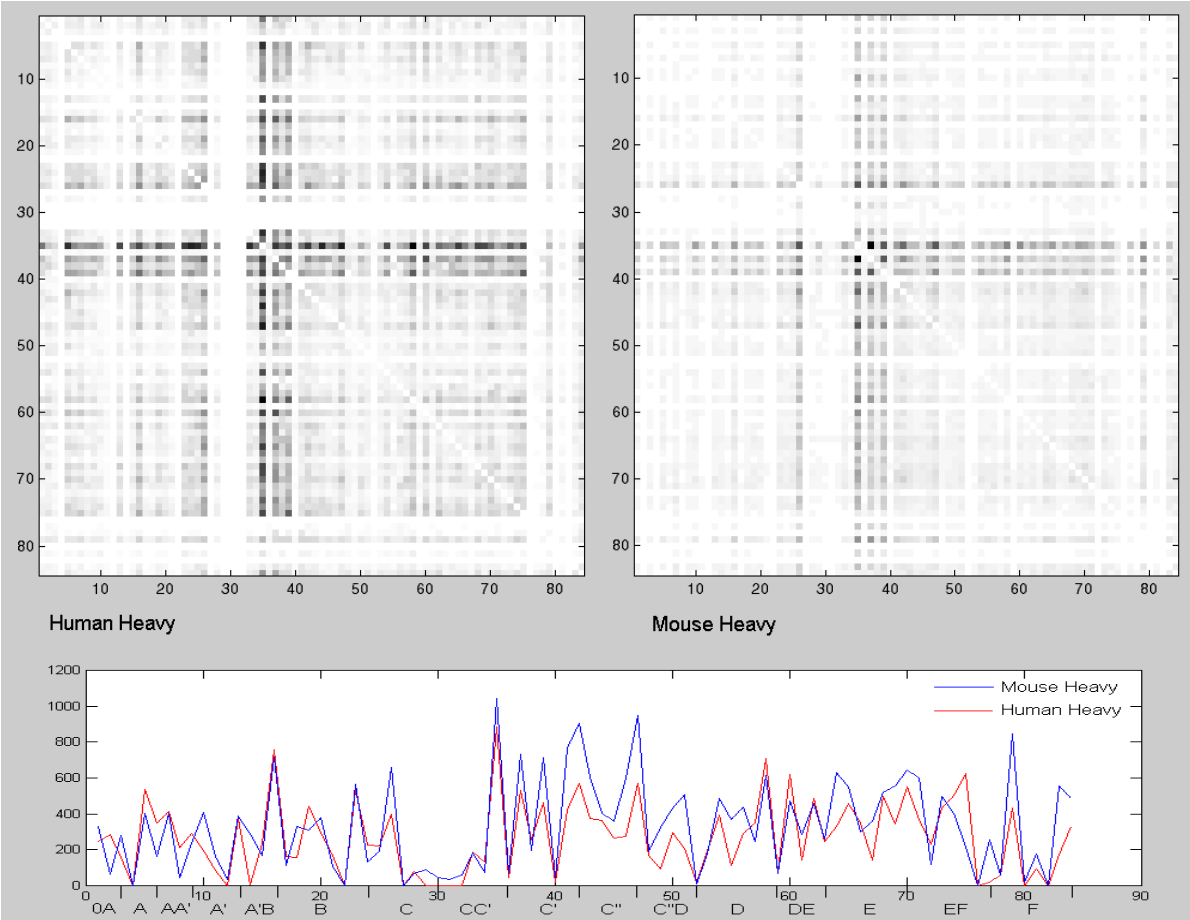


Fig.2: Correlation matrices for the Heavy chains of immunoglobulins (variable domain, Human and Mouse). On the bottom, overall correlation distribution is shown for this chain.

The common features of proteins are meant to help us with the second problem. The following 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 and 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 matrices of contacts between residues (Fig.3). Therefore, correlation information has a limited value for the direct fold prediction. Also, it is hard to find 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.

Correlation matrix presents a convenient tool to distinguish protein families. We have developed the image processing algorithm to extract the rectangular areas on the correlation matrices and analyze their structures. The procedure of automated match of areas with similar topology is currently under investigation. It helps to extract the common features of the protein families with distant properties.

Occurrence of crosses, grids, angle-shape and rectangular patterns 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.

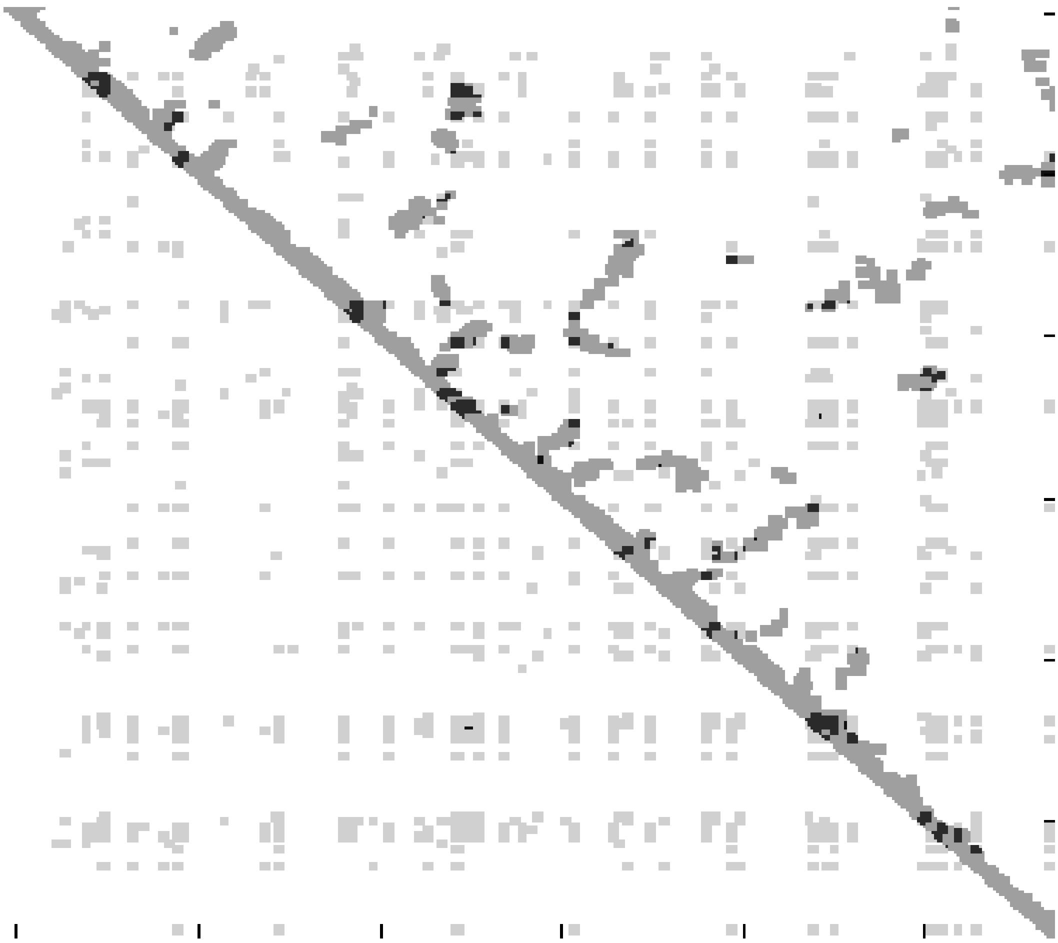


Fig.3: Contact matrix for the protein kinase family representative (shown above the diagonal, darker cells) and the correlation matrix (shown in the whole matrix, lighter cells), drawn in the same axis. Note that pairs in contact and correlated pairs not just hardly overlap, but the geometric structure of areas of contact and correlation are quite different.

Comparative analysis of the families of Ig-fold

We present the comparative analysis of selected strands (sequence and structure) for the superfamilies of Ig-fold. Secondary structure assignment from various sources is ambiguous (PDB data bank, PDBsum, and VAST). Our initial secondary structure assignment and cross-family alignment is presented at Table 1. Secondary structure assignment and the mutually-correlated positions, revealed in this study, are shown at Fig.4 and Fig.5. for two selected families.

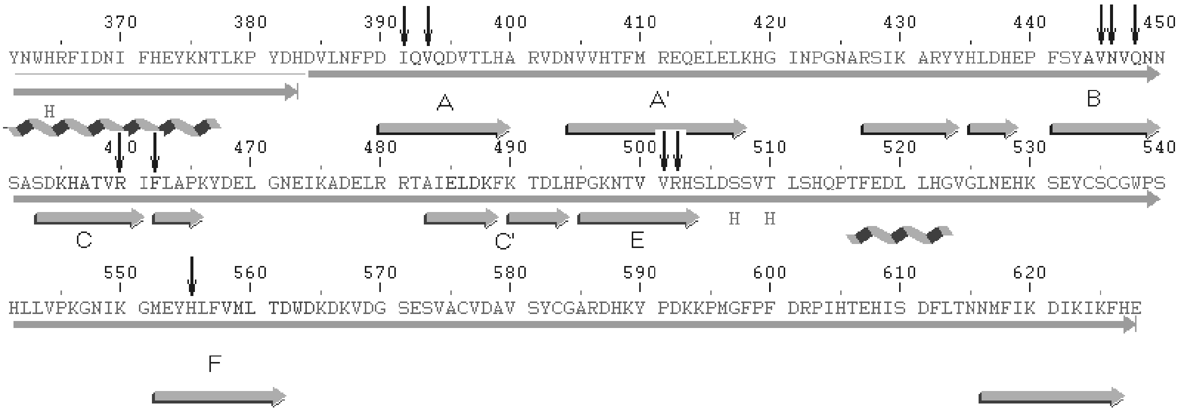


Fig.4: Sequence and secondary structure of *Hemocyanin* (*1lla*) protein, presented by cn3d software (NCBI). Arrows show the correlated (with most of others) positions in strands *A, A', B, C, C', E and F,* which are common for the whole Ig-fold. Strands D, G, G' are missing in Hemocyanin superfamily. Hemocyanin-specific correlated pairs are shown in the Table 1 for selected family representatives. Note that there is a deviation in the secondary structure assignment between that Table and VAST server results. It is compensated by the cross-family matching the resultant correlated positions, verifying the multiple alignment.

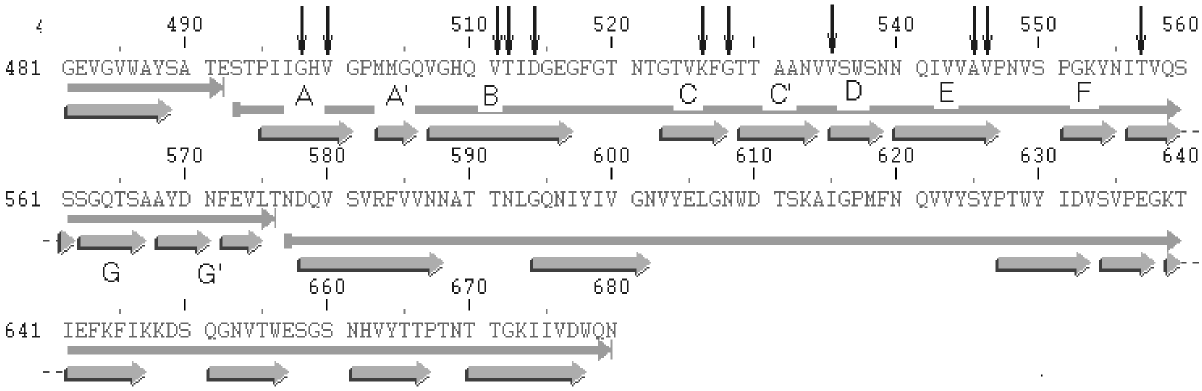


Fig.5: Sequence and secondary structure of *Cyclodextrin glycosyltransferase* (1cyg) protein.

Superfamilies/families		A	A'	B
		2 3 4 5 6 7 8 9 0 1 2 3 4 5	1 2 3 4 5 6 7 8 9 0	1 2 3 4 5 6 7 8 9 0 1 2 3 4
Fifroblast growth factor repressor, FGFR	1cvs	- V A P Y W T S - - - - - - -	- - - - K L H A V P	- - - V K F K C P S S G T -
Envelope glycoprotein, domain III	1svb	- - - - - F T W K R A P T D S	- - - - - - - - - -	- - T V V M E V T F - - - -
Cyclodextrin glycosyltransferase, dom E	1cyg	- - - - I G H V G - - - - - -	- - - M G G - - - -	- - - H Q V T I D G E - - -
Alpha-amylase II (TVAII), N-terminal dom	1bvz	- - - - - - Y A Y P I S - - -	- - - - - - - - - -	- - - T Q L R V R L A K - -
Hemocyanin, C-terminal domain	1lla	- - - - D I Q V Q D V T L H -	V V H T F M R E Q E	P F S Y A V N V Q N N - - -
Rho GDP-dissociation inhibitor 1	1rho	- - - - - V V V T G L T L V -	- - - F V L K - - -	E Y R I K I S F R V - - - -
Transcription factor NFATC	1a02	- - - M V E R Q D - - - - - -	- - - - - D S C L -	- - - - Q M I L T G Q - - -
p50 subunit of NF-kappa B transcr. factor	1svc	- - - - - - K I V R M D - - -	- - - - A G C V - -	- - E I Y L L C D - - - - -
Sulfite oxidase, C-terminal domain	1sox	- - Q S A V T Q - - - - - - -	- - - - G A A V - -	- - - G E L T V K G Y A W -
Integin beta-4 subunit	1qg3	- - - - - Q N P N A K A - - -	- - - - - - - - - -	- - - I H F N W L - - - - -
Growth hormone receptor	1axi	- - - - - - F T K C R - - - -	- - - - - - - - - -	- - - F S C H W - - - - - -
Erythropoietin (EPO) receptor	1ebp	- - - - - - E L L C F - - - -	- - - - - - - - - -	- - - L V C F W E E A - - -
Neocarzinostatin	1acx	- - - - - - A F S V S - - - -	- - - - - - - - - -	- - S V S V S V A - - - - -
Actinoxanthin	1noa	- - - - T A T V T - - - - - -	- - - - - - - - - -	- - - T V V K V A G A - - -
bovine	1cob	- - - - Q G T I H F E A K - -	- - - - - - - - - -	T V V V T G S I T - - - - -
Baker's yeast	1qup	- - A V R G L A R I V Q V - -	- - - - - - - - - -	K T L F D I T V N G V - - -
Alpa-adaptin AP2, N-domain	1qts	- - - - - L Q I G L K S E F R	- - - - - - - - - -	L G R M F I F Y G N - - - -
Pilus chaperone PapD, N-domain	1pdk	- - - - - - - - - - - - - - -	- - - - - A F V - -	- K S M T L D I S - - - - -
Major sperm protein, alpha isoform	1msp	- - - - - - - - - - - - - - -	- - - - - I V F N -	- - - Y H I K I T N A - - -
Purple acid phosphatase, N-domain	4kbp	- - - - - Q Q V H I T Q G - -	- - - - - - - - - -	- R A M I I S W V T - - - -
Invasin	1cwv	- - - L T L T A A V I - - - -	- - - - - G A P A -	- - - I T V E F T V A - - -

Superfamilies/families		C	C'	D
		2 3 4 5 6 7 8 9 0 1 2 3	1 2 3 4 5 6 7 8 9 0 1 2 3 4	1 2 3 4 5 6 7 8 9 0 1 2 3
Fifroblast growth factor repressor, FGFR	1cvs	- - T L R W L K N - - - -	- - - - K E F - - - - - - -	- - - - K V R Y - - - - -
Envelope glycoprotein, domain III	1svb	- - - I P V R A V A H - -	- - - - - - - - - - - - - -	- - M L I - - - - - - - -
Cyclodextrin glycosyltransferase, dom E	1cyg	- - - G T V K F G - - - -	- - - - - T T A - - - - - -	- - - - V V S W S - - - -
Alpha-amylase II (TVAII), N-terminal dom	1bvz	- - - V R C E V L Y A - -	- - - A H A L A G K A G S D	- - - - - - - - - - - - -
Hemocyanin, C-terminal domain	1lla	- K H A T V R I F L A P -	- - I E L D K F K T D L - -	- - - - - - - - - - - - -
Rho GDP-dissociation inhibitor 1	1rho	\ S G V K Y I E H T Y R -	- K I D K T D Y X V G S Y -	- - - - - - - - - - - - -
Transcription factor NFATC	1a02	- - - S K V V F T E K - -	- Q I W E M E A - - - - - -	- - - - - V D K D K S Q -
p50 subunit of NF-kappa B transcr. factor	1svc	- - - I Q I R F Y E E - -	- - - G V W E G F G - - - -	- - - - D V H R - - - - -
Sulfite oxidase, C-terminal domain	1sox	\ V R V D V S L D G - - -	- - - - - - W K V A - - - -	- - - - R L M - - - - - -
Integin beta-4 subunit	1qg3	- - G Y R V K Y W I - - -	- - - - - H L L D S - - - -	- - - - - - - - - - - - -
Growth hormone receptor	1axi	- I Q L F Y T R - - - - -	- - - - - K E C - - - - - -	- - - - D Y V - - - - - -
Erythropoietin (EPO) receptor	1ebp	- - N Y S F S Y Q L - - -	- - - - W K L C - - - - - -	- - - - - L H Q A P T - -
Neocarzinostatin	1acx	- E T Y Y I A Q C A - - -	- - - - - - A T S F T T - -	- - - - - - - - - - - - -
Actinoxanthin	1noa	- - A Y D V G Q C A W V D	- - - - - - F S S V T - - -	- - - - - - - - - - - - -
bovine	1cob	- G D H G F H V H - - - -	- G D L G N V T A - - - - -	- - - - - - - - - - - - -
Baker's yeast	1qup	- G N Y H A S I H E K - -	- G K V W H K F D E P I E C	- - - - - - - - N E S - -
Alpa-adaptin AP2, N-domain	1qts	- - F L N F T P T L I - -	- - - - - - - - - - - - - -	- - L N L Q T - - - - - -
Pilus chaperone PapD, N-domain	1pdk	- Y L A Q A W I E N - - -	- - - L R Q - - - - - - - -	- - I A T P - - - - - - -
Major sperm protein, alpha isoform	1msp	- I G W A I K T - - - - -	- - - L V G - - - - - - - -	- - - - - D - - - - - - -
Purple acid phosphatase, N-domain	4kbp	- S A V R Y W S - - - - -	- - - - - R I A K - - - - -	- - - - - - K M S T Y - -
Invasin	1cwv	- Q E V V I T T N - - - -	- - - - - N K I T E K T - -	- - - - - - - - - - - - -

Superfamilies/families		E	F	G
		2 3 4 5 6 7 8 9 0 1 2 3 4	1 2 3 4 5 6 7 8 9 0 1 2 3 4	1 2 3 4 5 6 7 8 9 0 1 2
Fifroblast growth factor repressor, FGFR	1cvs	- - - - - W S I I M D - - -	- - - G N Y T C I V E - - -	- - - S I N H T - - - -
Envelope glycoprotein, domain III	1svb	- - - - G G F I E M Q - - -	- G D N I I Y V G - - - - -	- - - - - - L S - - - -
Cyclodextrin glycosyltransferase, dom E	1cyg	- - - N Q I V V A V - - - -	- G K Y N I T V Q S - - - -	- - - - Q T S A A - - -
Alpha-amylase II (TVAII), N-terminal dom	1bvz	F D Y F E A L L E C - - - -	- - K Y V F L L T G - - - -	- - - - E A V Y - - - -
Hemocyanin, C-terminal domain	1lla	- - - G K N T V V R H S - -	- - G M E Y H L F V M L T D	- - - - - - N M - - - -
Rho GDP-dissociation inhibitor 1	1rho	- - A E E Y E F L T P V E -	S Y S I K S R F T - - - - -	- - - - - H L S - - - -
Transcription factor NFATC	1a02	- - - - M L F V E I - - - -	- V K V N F Y V I N - - - -	- - - - K R S Q P - - -
p50 subunit of NF-kappa B transcr. factor	1svc	- - - - - F A I V F K - - -	- A S V F V Q L R R - - - -	- - - - L L S - - - - -
Sulfite oxidase, C-terminal domain	1sox	- A L W E L T V P V - - - -	- - E L E I V C K A V D S -	- - - Y N V Q H R - - -
Integin beta-4 subunit	1qg3	- - - - - P S V E L - - - -	- - D Y E M K V C A Y - - -	- G E G P Y S S L - - -
Growth hormone receptor	1axi	- - - - - N S C Y F - - - -	- - P Y C I K L T S - - - -	- - - G T V D E - - - -
Erythropoietin (EPO) receptor	1ebp	- - A V R F W C S L - - - -	- - - L E L R V T A - - - -	- - G A P R Y H - - - -
Neocarzinostatin	1acx	- - - - A A S F S F T - - -	- - - - N L G A G N - - - -	- - - L N L G H - - - -
Actinoxanthin	1noa	- - - A S T S L T V - - - -	- - - C Q V G L S - - - - -	- - - - G N G P E G - -
bovine	1cob	- - - - - A I V D I V D - -	- - - R T M V V H E - - - -	- - - S R L A C - - - -
Baker's yeast	1qup	- - L Y S G K T F L S A - -	- - G R S F V I S K S L - -	- - K D Y S F L - - - -
Alpa-adaptin AP2, N-domain	1qts	- - - A Q V Q Q V V N I E C	- P V L N I Q F R - - - - -	- - - - - F Q N - - - -
Pilus chaperone PapD, N-domain	1pdk	- - - - - S M V R L S T - -	- N E I W A Q A L Y - - - -	- - I A L Q T K - - - -
Major sperm protein, alpha isoform	1msp	- - K E K V L M A V S C - -	- - - - - I T I E W T N - -	- - M V R R K N - - - -
Purple acid phosphatase, N-domain	4kbp	- - - F I H H T T I R K - -	- - - Y Y Y E V G - - - - -	- - - - - W R R - - - -
Invasin	1cwv	- - - - V A R I A L - - - -	- G V T V V T A E V - - - -	- - - - - Q R Q S - - -

Table 1: Twenty-one representatives of protein families are enumerated. Only ten data sets were finally selected from the totality of Ig fold - families to provide the statistical significance (no alien folds, sufficient number of sequences, lower number of missing strands) for the correlation analysis. Preliminary alignment is obtained by means of comparison of the secondary and 3D structure. Grayed columns depict the positions that are mutually correlated for the majority of families being analyzed. Strand G' is not shown.

For *1lla* and *1cyg* sequences, additional (and missing common) correlation positions are shown in lighter gray. Evidently, alignment can to be corrected so that the strands are shifted to match the correlated positions. As a result, we obtain more precise set of common mutually correlated positions for Ig fold.

Revealing mutually-correlated positions

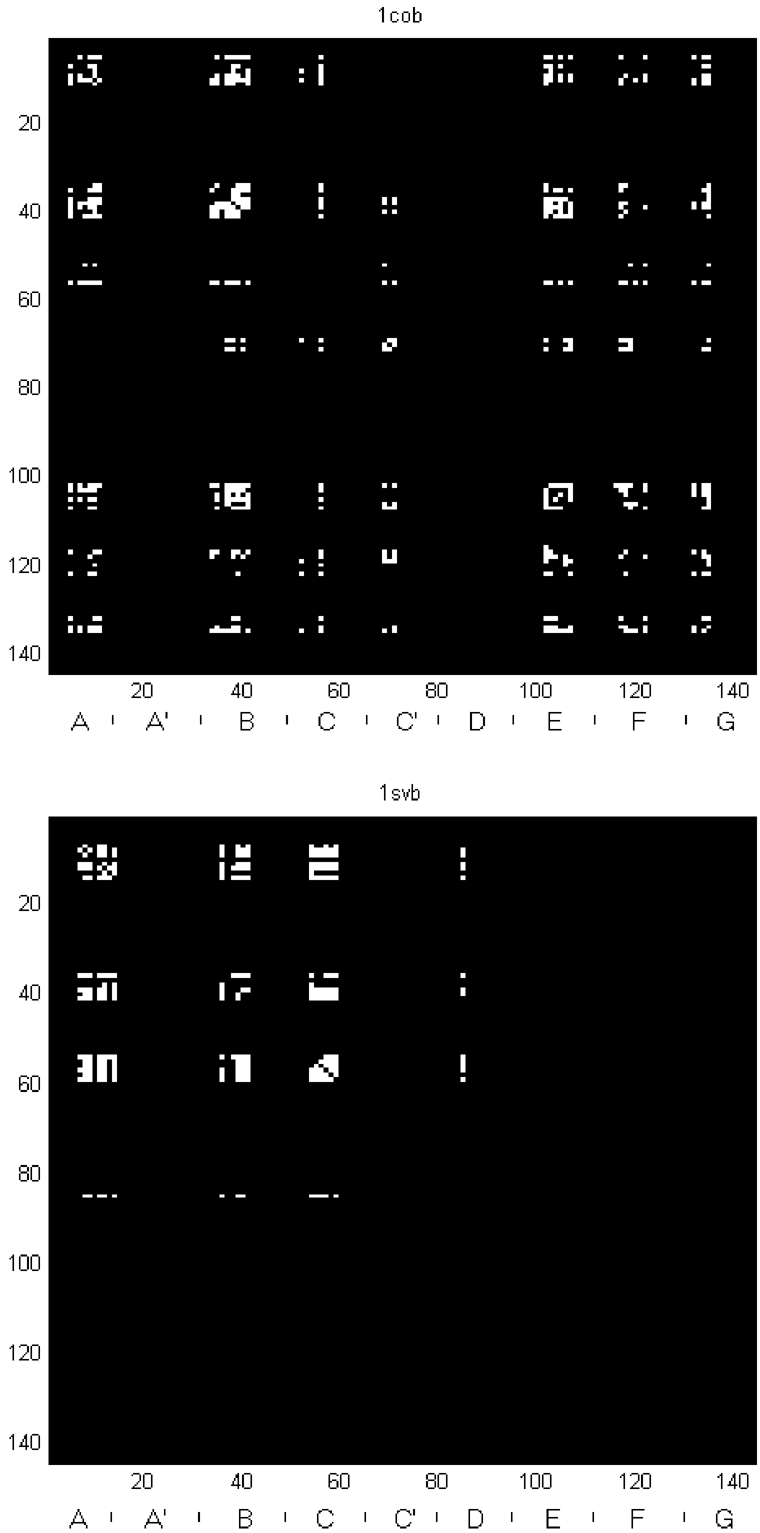


Fig.6 a: correlation matrices for 1cob and 1svb.

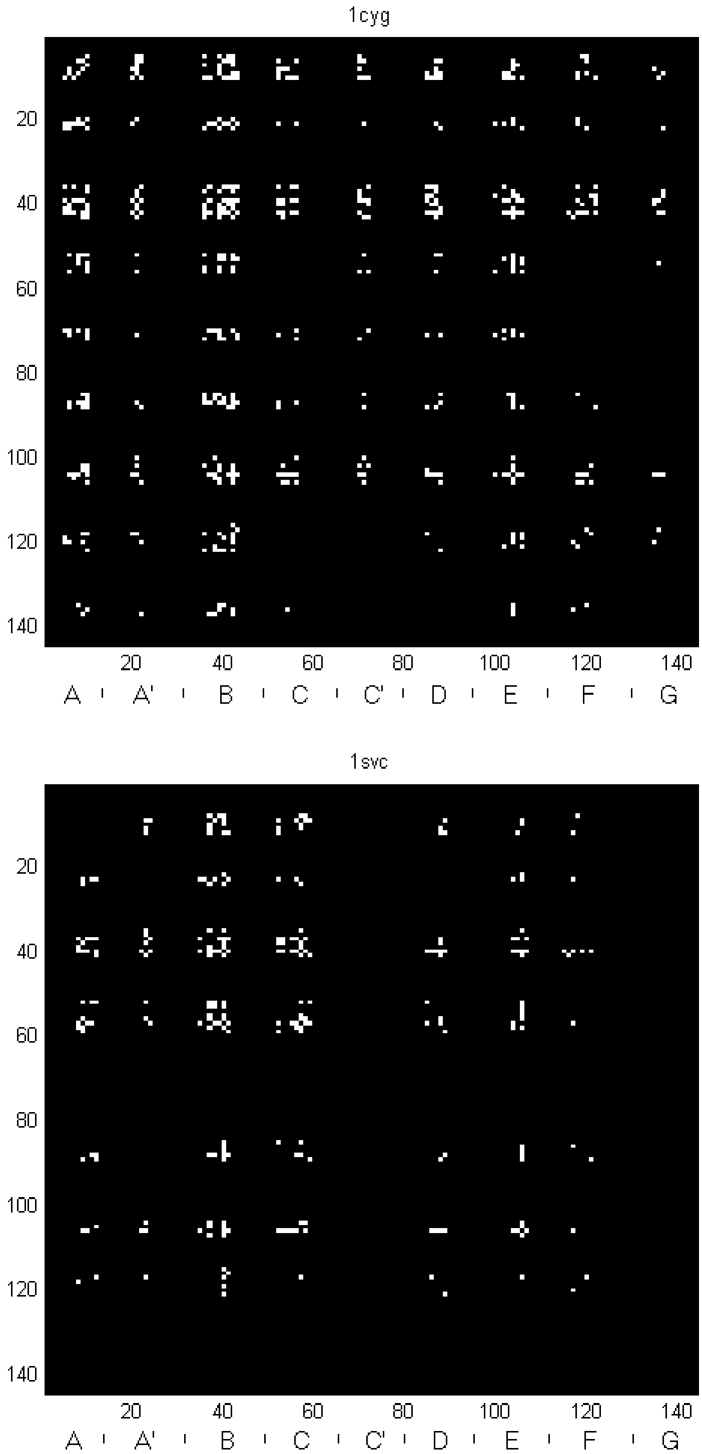


Fig.6 b: correlation matrices for 1cyg and 1svc.

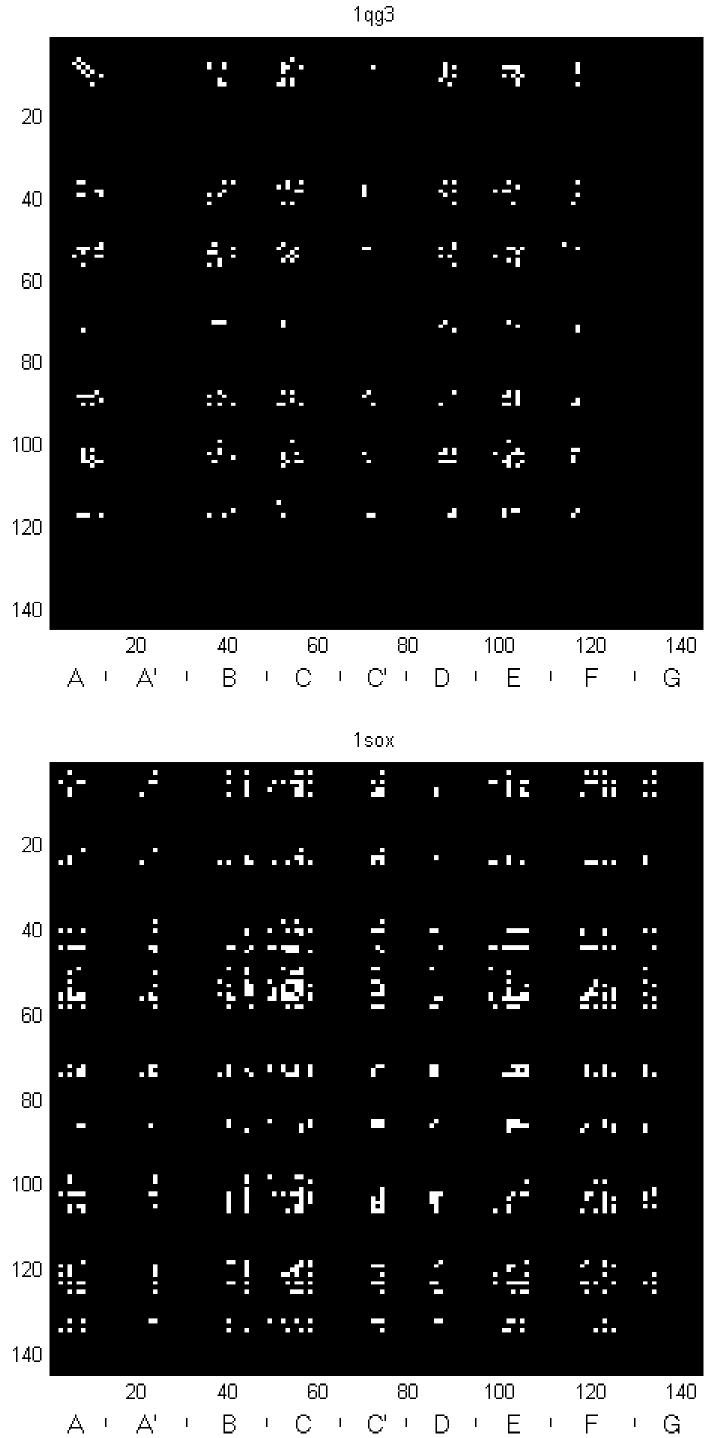
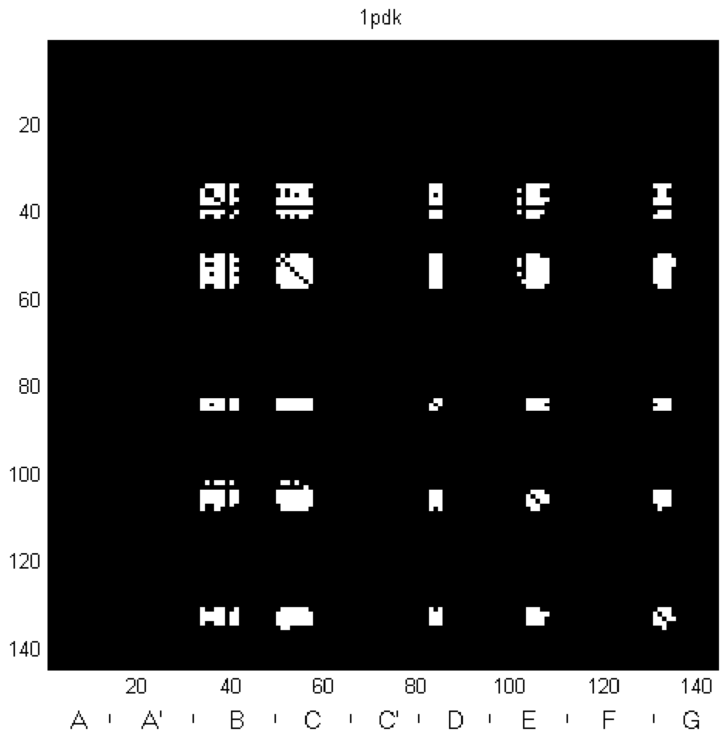


Fig.6 c: correlation matrices for 1qg3 and 1sox.



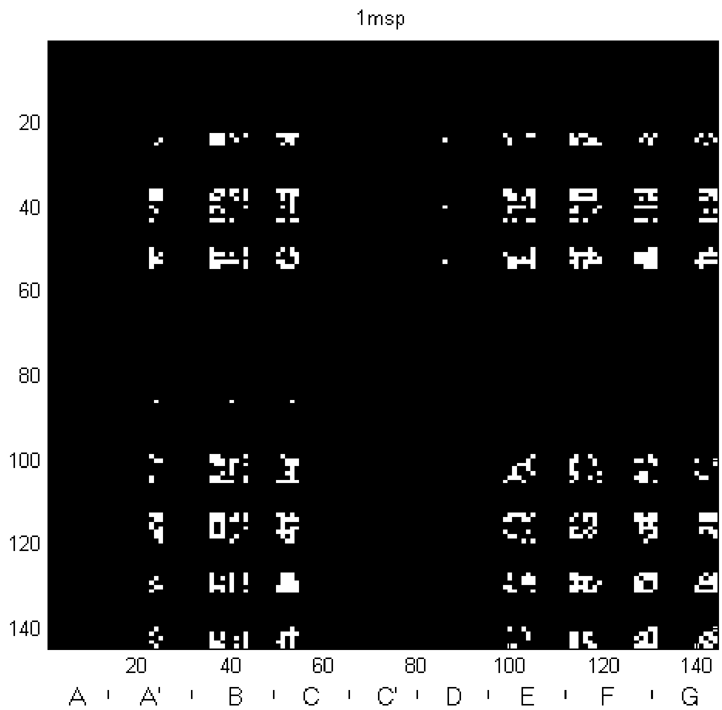
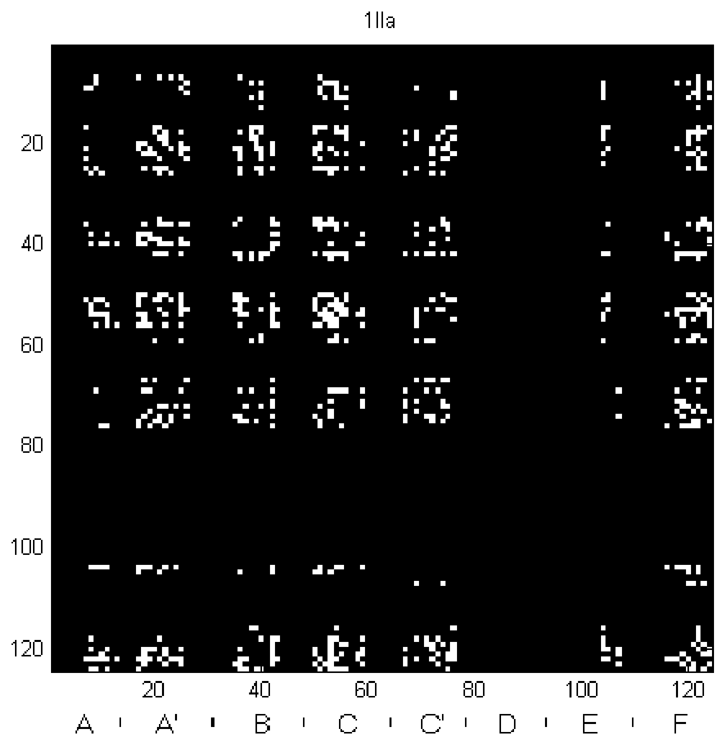


Fig.6 d: correlation matrices for 1pdk and 1msp



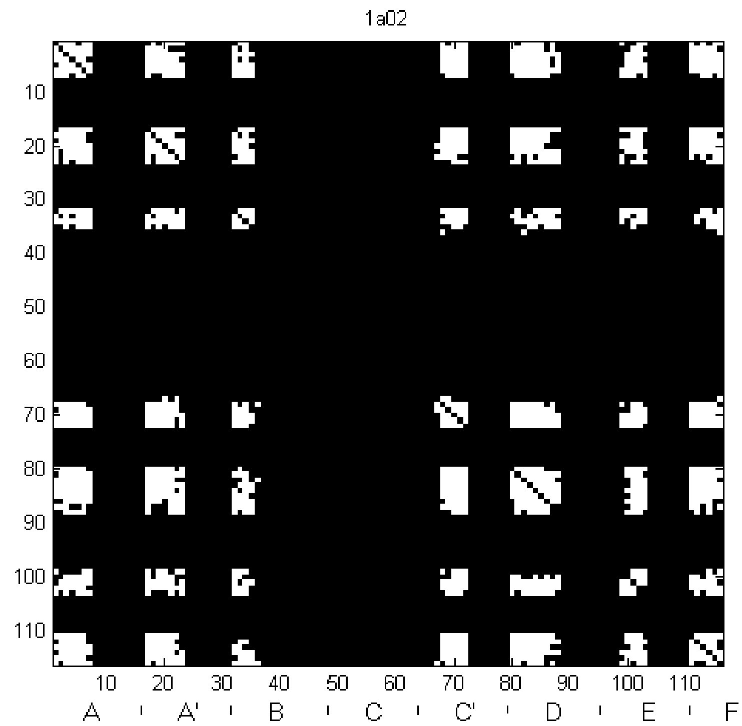


Fig.6 e: correlation matrices for 1lla and 1a02.

Correlation matrices for selected protein families are shown at Fig. 6a-6e.

Originally, to obtain the commonly correlated positions, we summed up the correlation matrices. To obtain the correlated positions, common to each Ig-family, with higher reliability, we allow a shift by one unit in each direction (including diagonal) while calculating the number of families the element of correlation matrix is positive.

As we discovered, mutually correlated positions served as additional means to verify the structural alignment of protein sequences from different families. The capability of the positions of correlation to serve as the criteria of matching the multiple sequence alignment was tested on the sets of sequences with sufficiently high similarity. If a set of protein sequences is randomly split into subsets and each subset is subject to multiple alignment, followed by the computation of correlated positions, then the respective positions can serve as a basis for overall sequence alignment *instead* of the residues themselves.

Therefore, we used the iterative approach to determine the mutually correlated positions for the totality of Ig-fold families. Initially, we use the preliminary structural alignment of the family representatives. The set of mutually correlated positions, obtained given that preliminary data, allowed us to verify and improve the alignment, which in turn led us to the higher number of revealed commonly correlated positions for the totality of Ig-families. Two such iterations led to the alignment such that the further iteration would not lead to alignment modifications. As a result, we did not have to take into account the shifts between correlated positions among the families of Ig-fold.

Enumeration of resultant positions is depicted at Table 2 (refer to Table 1 for position assignment).

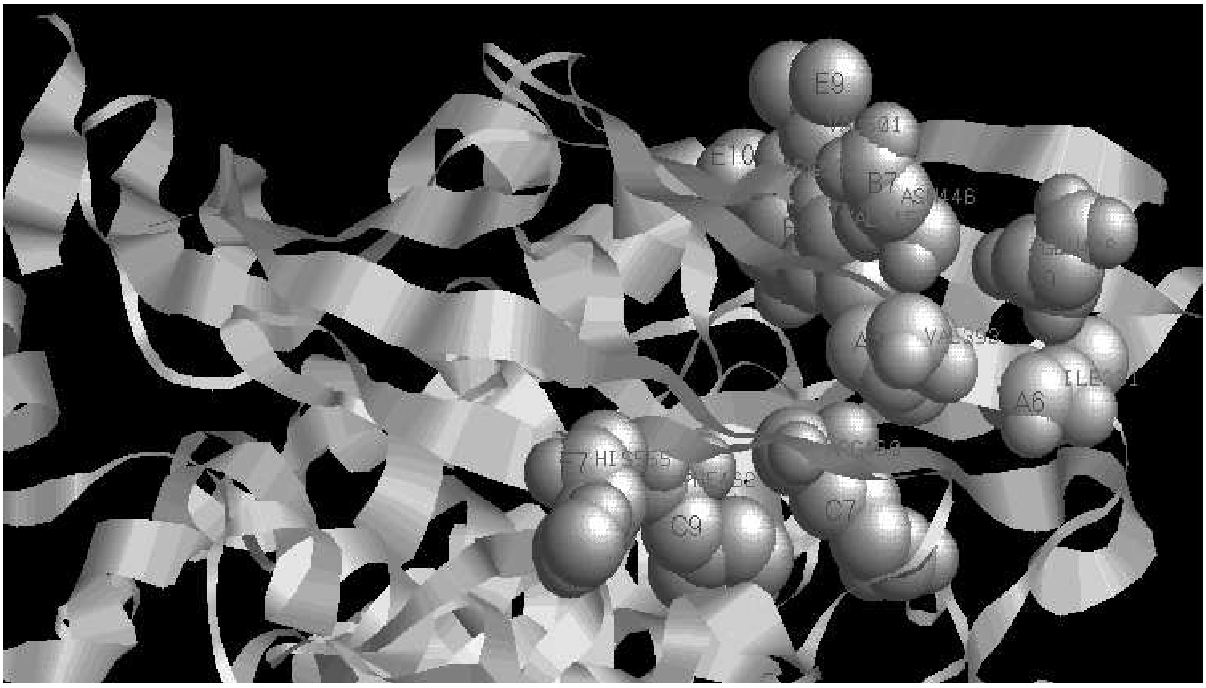


Fig.7: Mutually correlated residues for *1lla*. Note that not all mutually correlated residues are in contact with each other. Some of these residues are on the surface, some occupy inner space.

A6 A8
B6 B7 B9
C7 C9
D6
E9 E10
F7

Table 2: the set of mutually correlated positions for the Ig-fold.

We highlight the following observations concerning the set of mutually correlated positions:
1) The set of common mutually-correlated positions for selected families of Ig fold is found.
2) Correlated positions obey the following properties:

- Correlation *matrices* of Ig families obey the common properties of immunoglobulins, discovered and verified earlier, in particular, mutual instead of just pair-wise correlation.
- Roughly, about the half of mutually correlated position for each *Ig-family* forms the set of the mutually correlated position for the whole *Ig-fold.*
- The mutually-correlated Ig-residues are mostly *hydrophobic*.
- Mutually-correlated positions are almost *uniformly* distributed through the strands.

3) Revealed set of mutually-correlated positions allows to improve the structural alignment between the sequences of families of Ig fold.
4) The number of common mutually-correlated positions, obtained using the improved cross-families alignment, is higher than the original number.

In the conclusion, we discuss the roles of our findings in 3D structure assignment of a novel protein sequence. How can one take advantage of the proposed methodology of correlation analysis for Ig –fold? Having a set of sequences with unknown structure (particularly, from a genome), if it is possible to align them with the certain degree of confidence, then the set of mutually correlated positions can be matched against ones of the protein family with the known 3D structure. We call the matching templates *correlation patterns*. Application of the above method is successful when the following consecutive constraints are obeyed:

- There is no close homologue with known 3D structure available (otherwise, more simple homology-based assignment can be used).
- There are multiple sequences to be assigned a 3D structure;
- There is a reasonable multi-alignment accuracy achieved for these sequences;
- The set of mutually-correlated positions can be obtained;
- Obtained pattern of mutually correlated positions includes sufficient number of sequences for the adequate match;
- There is a successful match for the correlation patterns for the given set and a protein family/ super-family (with the known fold).

Software tools for the analysis of correlation

We present the software tools for the analysis of protein family: building correlation matrix and discovering the inter-dependence between the positions (Figs 8 and 9). Fig. 10 and 11 depicts the networks of interconnection for the selected highly-correlated and randomly taken sets of positions.

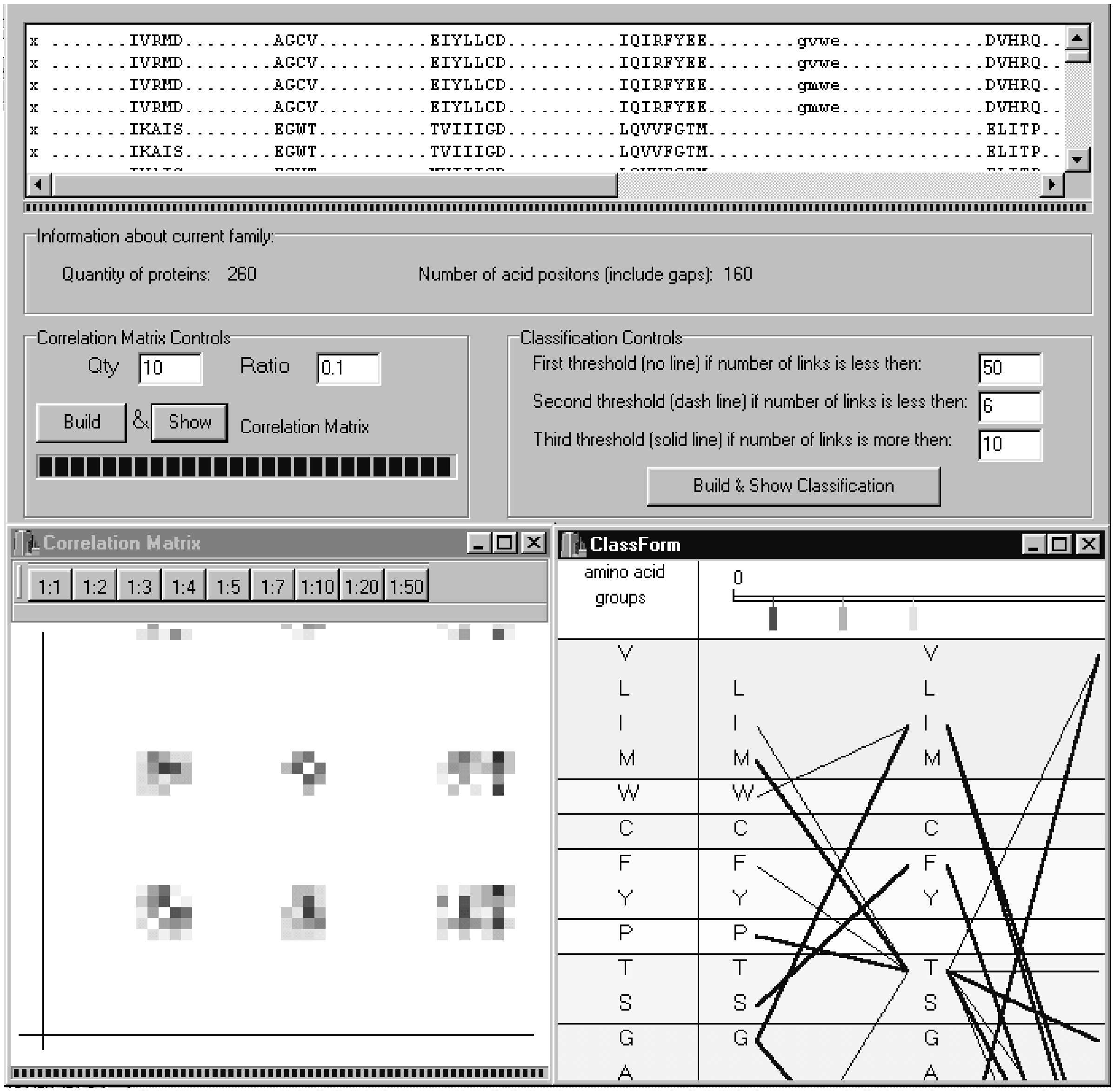


Fig.8: The screen–shot of the *ProtoTrail*, the software for analysis and visualizing the classification of protein family. Above is the main frame of the application, when the input file of protein family is specified. Also, the parameters of pair-wise calculation are specified on the left input area. On the right input area, the threshold parameters of resultant classification can be specified to control the scale of its details.

The fragment of the resultant correlation matrix is shown on the bottom-left. Grey-scaled cells depicts the pair of position with various degrees of correlation: the darker- the stronger. One can zoom in the visualization of correlation matrix. The fragment of inter-dependence visualization is shown on the bottom-right. Line thickness denotes the absolute number of sequences with the specified (by the edge's nodes) residues. Respective thresholds are set in the input area on the main application frame.

User interface allows easily changing the positions and tracking their correlation with the other two selected ones. Interactive features also allows to modify the sequence of positions that are in use for classification to approach a tree-like structure as close as possible. Fig.9 depicts the various ways of specifying the position for analysis.

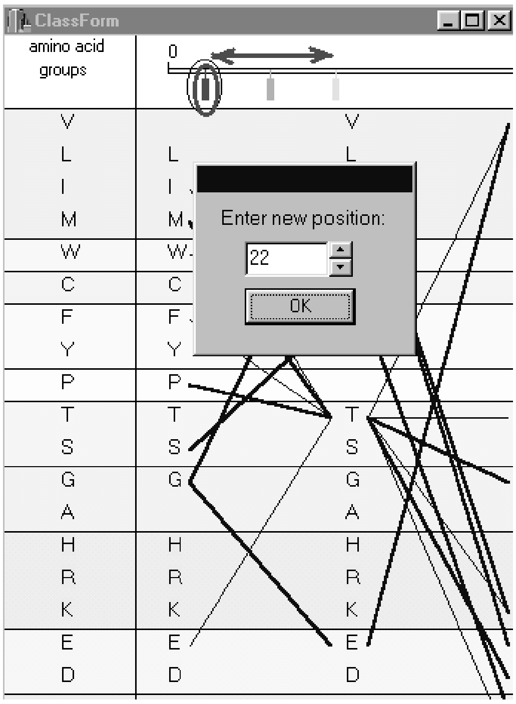


Fig.9: Selecting the positions for correlation analysis.

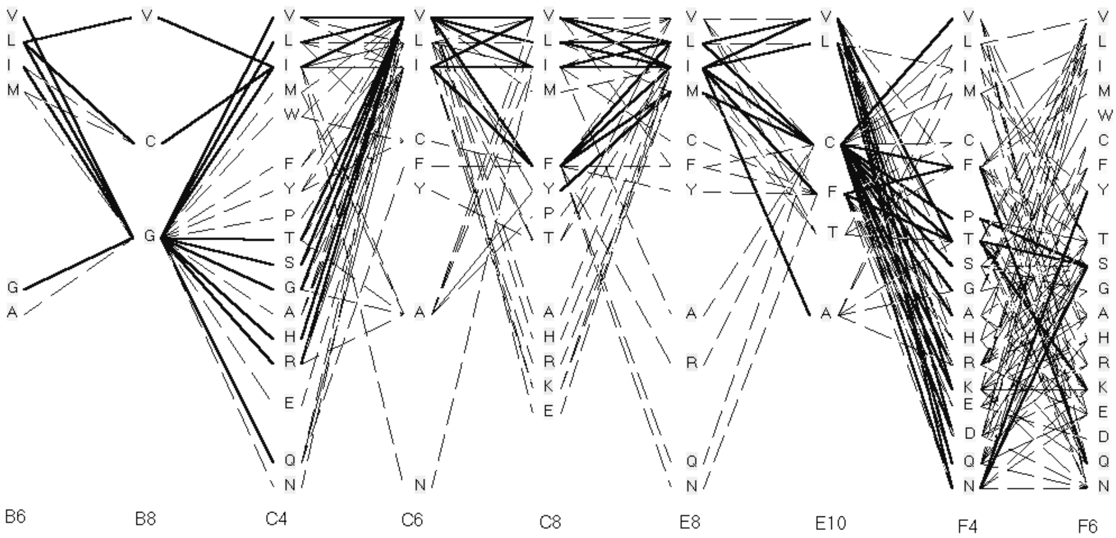


Fig.10: Graphic representation of the residues and correlation between positions for the selected mutually correlated ones of Ig-fold.

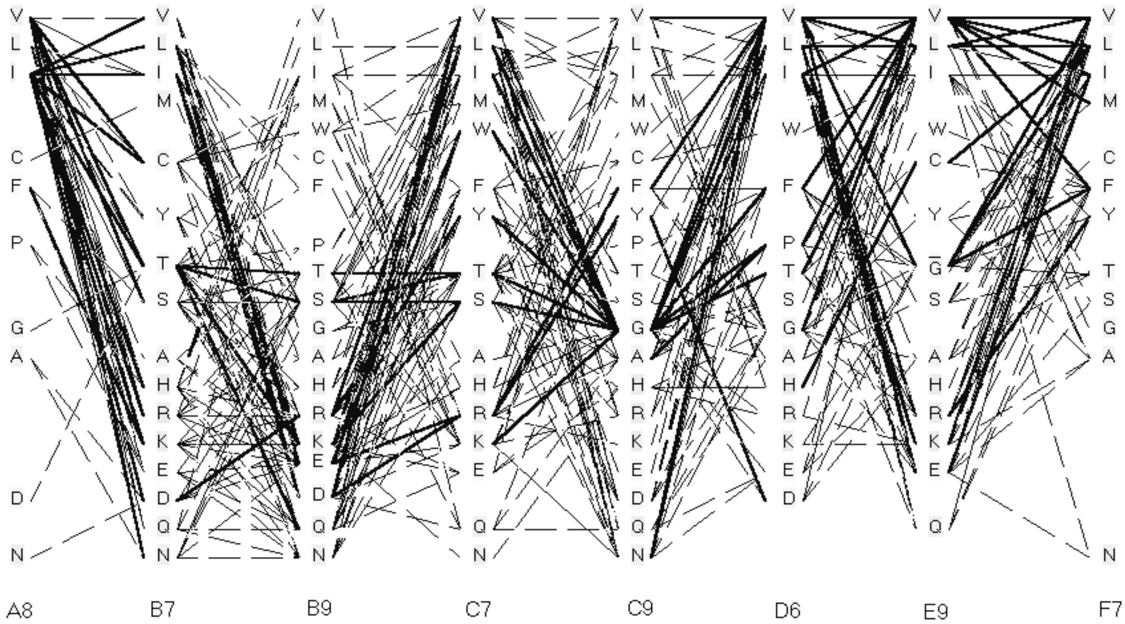


Fig.11: Graphic representation of the residues and correlation between randomly selected positions.

Acknowledgements

The authors are grateful to A.Kister and I.Gelfand for introduction to the field of Immunoglobulin proteins, initial comments on the problem, providing the data on structural alignment (Table 1) and valuable advises. Also, fruitful discussions with M. Roytberg, A. Finkelstein and others helped to increase the computational soundness of this study.

References

1. Neher, E. (1994) How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci., USA v91, 98-102.
2. Gerstein, M.& Altman, R.B. (1995) Average core structures and variability measures for protein families: application to the immunoglobulins *J. Mol. Biol.,* **251**, 161-175.
3. Mirny,L.K & Shakhnovich, E.I (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol.* 291(1):177-196.
4. Lindahl E, Elofsson A. (2000) Identification of related proteins on family, superfamily and fold level. *J Mol Biol.* 295(3):613-625.
5. Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
6. Halaby DM, Poupon A, Mornon J. (1999) The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng* **12**(7):563-571.
7. Chothia,C., Lesk,A. (1986) The relation between the divergence of sequence and structure in proteins *EMBO J.* **5**:823-826.

8. Galitsky, B.A., Gelfand, I.M. and Kister A.E. (1998) Predicting amino acid sequences of the antibody human V_H chains from its first several residues *Proc. Natl. Acad. Sci.*V95 pp 5193-5198.
9. Galitsky, B.A., 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.
10. Galitsky B.A., Shelepin S.L. (2001) On the inter–residue correlation patterns and their role in classification of protein families. DIMACS technical report 2001-09 February 2001.
11. Shelepin, S.L. and Galitsky B.A. (2001) Using correlation matrices in the comparative analysis of Immunoglobulin chains. In *Currents in computational molecular biology , RECOMB-01,* Montreal, Canada.
12. Dement'ev, A. and Galitsky, B.(2002) Discovering common mutually-correlated pairs of positions for the families of immunoglobulin fold (2002) In *Currents in computational molecular biology , RECOMB-02,* Washington DC.
13. Gelfand, I.M. & Kister, A.E., (1997). "A very limited number of keywords (main patterns) describes all sequences of the human variable heavy (V_H) and κ ($V_κ$) domains" *Proc. Natl. Acad. Sci. USA* 94, 12562-12567.