PROTEINS: Structure, Function, and Bioinformatics 63:832–845 (2006)

Correlated Mutations: Advances and Limitations. A Study on Fusion Proteins and on the Cohesin-Dockerin Families

Inbal Halperin,[1] Haim Wolfson,[2] and Ruth Nussinov[1,3]*

[1]*Sackler Institute of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel*

[2]*School of Computer Science, Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*

[3]*Center for Cancer Research Nanobiology Program, Basic Research Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland 21702*

ABSTRACT **Correlated mutations have been repeatedly exploited for intramolecular contact map prediction. Over the last decade these efforts yielded several methods for measuring correlated mutations. Nevertheless, the application of correlated mutations for the prediction of intermolecular interactions has not yet been explored. This gap is due to several obstacles, such as 3D complexes availability, paralog discrimination, and the availability of sequence pairs that are required for inter- but not intramolecular analyses. Here we selected for analysis fusion protein families that bypass some of these obstacles. We find that several correlated mutation measurements yield reasonable accuracy for intramolecular contact map prediction on the fusion dataset. However, the accuracy level drops sharply in intermolecular contacts prediction. This drop in accuracy does not occur always. In the Cohesin-Dockerin family, reasonable accuracy is achieved in the prediction of both intra- and intermolecular contacts. The Cohesin-Dockerin family is well suited for correlated mutation analysis. Because, however, this family constitutes a special case (it has radical mutations, has domain repeats, within each species each Dockerin domain interacts with each Cohesin domain, see below), the successful prediction in this family does not point to a general potential in using correlated mutations for predicting intermolecular contacts. Overall, the results of our study indicate that current methodologies of correlated mutations analysis are not suitable for large-scale intermolecular contact prediction, and thus cannot assist in docking. With current measurements, sequence availability, sequence annotations, and underdeveloped sequence pairing methods, correlated mutations can yield reasonable accuracy only for a handful of families. Proteins 2006;63:832–845.**

Key words: binding site; correlated mutations; residue covariation; fusion proteins; protein–protein interactions

INTRODUCTION

The Rational Basis of the Correlated Mutations Concept

The concept of correlated mutations is one of the fundamental ideas behind the theory of coevolution. Functional constraints are expected to limit the amino acid substitution rates, resulting in a higher conservation of functional sites with respect to the rest of the protein surface. Once a residue is changed, given the functional constraints operating on it, this mutation can be compensated by an additional mutation of a complementary residue across the interface. This enables the coevolution of two proteins that can lead to high specificity and high affinity. Coevolution is evident at the molecular level, where interacting proteins exhibit similar evolution rates.[1] Coevolution was well characterized in a variety of proteins systems.[2–5] The concept of coevolution is not restricted to functional constraints in interacting proteins. It can be expanded to include also intraprotein residue pairs stabilizing the protein fold as well as protein–nucleic acid and nucleic acid–nucleic acid interactions.

Biological Mechanisms Generating Correlated Mutations

Two major sources for covariation are currently known: mutations and gene conversion. In the first mechanism, two independently occurring single mutations are preserved by positive selection. This process should be independent of the genomic distance between the mutated positions (if the effect of mutational hot spots is neglected). It is expected to play an equal role in intra- and intergenic correlated mutations. In contrast, gene conversion can only contribute to intragenic correlated mutations. Gene

Grant sponsor: Federal funds from the National Cancer Institute, National Institutes of Health; Grant number: NO1-CO-12400. Grant sponsor: the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

*Correspondence to: Ruth Nussinov, NCI-Frederick, Building 469, Room 151, Frederick, MD 21702. E-mail: ruthn@ncifcrf.gov

Received 15 May 2005; Revised 21 October 2005; Accepted 1 December 2005

Published online 28 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20933

conversion is defined as a nonreciprocal recombination event of a limited DNA region (several hundreds of nucleotides) that involves homologous pairing and requires gap filling DNA synthesis. The donor sequence can be the homologous chromosome or the sister chromatid during meiosis or a duplication of the gene during mitosis. Gene conversion introduces co-occurrence of mutations by two mechanisms: mismatch repair within the heteroduplex region following branch migration and gap filling.[6] Gap filling is carried out by error-prone repair DNA polymerases. These polymerases can copy damaged DNA with high efficiency, incorporate nucleotides with poor accuracy, tend to form base mispairing rather than correct Watson-Crick, and catalyze incorporation of aberrant DNA primer ends.[7] DNA that was resynthesized by the repair system is not subject to further repair.[6,8] The co-occurring mutations can be from the gap filled region, from the converted region, or one from the converted region and the other from the gap-filled region. Hence, gene conversion can generate correlated mutations that are not distant from each other by more than a few hundred nucleotides, the size of the conversion tract. Intermolecular co-occurring mutations can, therefore, be created only by the first mechanism (co-occurring single mutations combined with positive selection). The host–parasite coevolution indicates that this mechanism is sufficient to generate intergenic correlated mutations.[9]

Applications of Correlated Mutations

Correlated mutations were applied to RNA before they were utilized for proteins. They were employed for secondary structure prediction,[10] RNA folding,[11] intron prediction,[12] and RNA–protein interface prediction.[13] Applications of correlated mutations in proteins can be roughly divided into two groups according to the genetic location of the correlated residues. The first group includes residue positions that are in the same gene, while the second group includes residues from two separated genes that their corresponding protein products interact. Intragenic correlated residue pair prediction[14,15] is used for protein folding,[16] guided mutagenesis experiments,[17] fold recognition,[18] domain boundary definition (by a drop of the predicted contact density signal[19]) for revealing energetic-coupling pathways,[17] and for finding regulatory allosteric communication.[20] On the other hand, intergenic correlated residue pairs are used for protein–protein interaction prediction[21,22] and to guide docking.[23]

Correlated Mutations Prediction Methods

Despite the fundamental and straightforward nature of the ideas behind the concept of correlated mutations, quantifying it is not an easy task. Conservation, which is less complex than correlated mutations, yielded no less than 18 different diverse and sophisticated methodologies (reviewed in ref. 24). The number of measurements for correlated mutations that have been proposed so far is still smaller. Because no thorough review gathers all of these, they are briefly described here. For clarity, the following notation is used throughout this description: N= number of sequences; i,j = columns i and j positions in MSA; Multiple Sequence Alignment); k,l = sequences k and l ($1 \le k,l \le N$); x,y = residues x and y (any of the 20 amino acids or B, Z, gap); and CM = the correlated mutation score.

Pearson Correlation, Correlation Coefficient, Gobel, Spearman, NSC, McBASC[25]

This method seeks co-occurrence of mutations of comparable similarities. It is based on the mathematical definition of correlation coefficient (denoted as r). At the heart of the method is a similarity matrix, which provides a score for every mutation. This score indicates how radical or conservative this mutation is with respect to the change in the amino acid physicochemical properties. A radical mutation in one residue position is expected to be accompanied by a radical mutation in a complementary position. For every position (i), all possible pairs of sequences (k,l) are compared. This yields N^2 comparisons for every position. The correlated mutation score between two positions (i,j) is given by comparing the similarity score of the N^2 comparisons, as indicated in Equation (1).

$$CM_{ij} = r_{ij} = \frac{1}{N^2}\sum_{kl}\frac{W_{kl}(s_{ikl} - \langle s_i\rangle)(s_{jkl} - \langle s_j\rangle)}{\sigma_i\sigma_j} \qquad (1)$$

where $\langle S_i\rangle$ is the average similarity score of N^2 comparisons in position i; S_{ikl} is the similarity score between the residue in position i in sequence k, and the residue in position i in sequence l; σ_i is the standard deviation of N^2 similarity scores of position i; W_{kl} is the fraction of nonidentical positions from the MSA length in sequences k,l, normalized to sum to 1. The purpose is to downweigh information from very similar sequences.

Perfectly conserved columns having a zero standard deviation are removed from the analysis. A couple of modifications over the basic Pearson correlation methods were proposed: (1) without sequence weighting;[26] (2) without identity comparison. The "no self comparisons" measurement, NSC, simply corrects the original Gobel method, which mistakenly compares a sequence with itself ($k = l$);[27] (3) with rank order. Numerical similarity values are replaced by their rank.[23] This method is commonly known as the Spearman correlation coefficient. A few of the modifications were designed to detect negative correlation; (4) with negative correlation detection. Negative correlation was incorporated by the addition of a clustering analysis[28] or by using an absolute value of Pearson correlation.[29]

Observed Minus Expected Squared, OMES[15,20]

This method compares the observed co-occurrence of each two residues in each two columns with their expected co-occurrence [Eq. 2(B)]. This method is rooted in the chi-square nonparametric test of statistical significance. The expected co-occurrence is based on the occurrence of each residue in the column, assuming no dependency between the two residue distribution in the two columns [Eq. 2(A)].

$$\text{(A)}\quad N_{ex} = \frac{N_{\text{ex}} N_{yj}}{N_{\text{valid}}} \quad \text{(B)}\quad r_{ij} = \sum_{l}^{L} \frac{(N_{\text{obs}} - N_{\text{ex}})^2}{N_{\text{valid}}} \qquad (2)$$

where L is the size of a list of distinct residue pairs in columns i and j; N_{valid} is the sequences without gaps in columns i,j; N_{obs} is the number of times each distinct pair was observed (x in column i and y in column j); N_{ex} is the number of times we would expect residues x and y to co-occur in columns i and j, respectively, given their single occurrences in columns i and j; N_{xi} is the number of times residue x occurs in column i; and N_{yj} is the number of times residue y occurs in column j.

Mutual Information, MI[30–32]

Mutual information measures how much information one random variable provides about another. It is a "distance" between the joint distribution, $P(x,y)$, and the product distribution, $P(x) \times P(y)$ [Eq. (3)]. In this context, the mutual information measures how much column i in a MSA increases our knowledge about column j. Because the probability of finding a residue in a column is the only attribute taken into account, the method is indifferent to amino acid identity. In contrast to method 1, similarity between amino acids is disregarded by the MI method.

$$MI_{ij} = \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}, \qquad (3)$$

where $P(x)$ is the probability to find residue x in column i; $P(y)$ is the probability to find residue y in column j; and $P(x,y)$ is the probability to find residues x,y in columns i,j, respectively. The probabilities are calculated from the amino acid distribution in the corresponding MSA columns.

Quartets[33]

This method adopts the basic concept of the Pearson correlation, but gives up the usage of a physicochemical-based similarity matrix. This suggestion is appealing in light of the differences between the similarities matrices. For example, the correlation coefficient between the Miyata matrix[34] and McLachlan[35] is only 0.32. Another reason for neglecting the usage of a physicochemical similarity matrix is the observation that in some families, like the immunoglobulin heavy chain, exchange of similar amino acids in certain positions requires a dramatic exchange of residues in other positions.[36] The method is based on quartets of residues. A quartet is made up of two pairs of amino acids from two columns (X_{ik}, Y_{il}, X'_{jk}, Y'_{jl}). For each pair of columns, a 20 $\times$ 20 matrix is built and the contribution of all quartets is summed up. A quartet contributes 1 if it fulfills the conditions of Equation (4), and 0 otherwise.

$$\left\{ \begin{array}{l} [(P_{ix}{*}P_{jy} \gg P_{iy}{*}P_{jx}) \text{ and } ((P_{ix} > D_{\min}) \text{ or } (P_{jy} > D_{min}))] \text{ or} \\ [(P_{ix}{*}P_{jy} \ll P_{iy}{*}P_{jx}) \text{ and } ((P_{iy} > D_{\min}) \text{ or } (P_{jx} > D_{min}))] \end{array} \right\}$$

$$\text{and} \left\{ \left(\frac{P_{ix}{*}P_{jy}}{P_{iy}{*}P_{jx}} > D^{Q}_{\min} \right) \text{ or } \left(\frac{P_{iy}{*}P_{jx}}{P_{ix}{*}P_{jy}} > D^{Q}_{\min} \right) \right\} \qquad (4)$$

where $D^{Q}_{\min}$ is the number of sequences providing evidence of correlation via a pair of amino acids, divided by the number of sequences voting against it, and $D_{\min}$ is the absolute number of sequences providing evidence of correlation.

At a first glance this definition may look like a "catch 22:" $D^{Q}_{\min}$ and $D_{\min}$ appear to be defined based on the result of Equation (4), while this same equation requires the use of $D^{Q}_{\min}$ and $D_{\min}$. However, this loop is shunned if we use the following interpretation for the "number of sequences providing evidence of correlation" term. This term equals the number of sequences having residue x in column i or residue y in column j if the first parts of the two logical terms that constitute Equation (4) are fulfilled. Otherwise, it equals the number of sequences having residue y in column i or residue x in column j.

Statistical Coupling Analysis, SCA[17,37–39]

The SCA method is based on perturbation of a MSA. The most prevalent residue in column j defines a subalignment. The subalignment is composed only of sequences in which this residue appears in column j. The correlated mutation score is expressed by an "energetic" term, $\Delta\Delta G_{i,j}$ [Eq. 5(B)]. It expresses the difference between the ΔG_i parameter of the full alignment and of the subalignment. The ΔG_i parameter is expressed in Equation (5A).

$$\text{(A)}\quad \Delta G_i^{\text{stat}} = kT{*}\sqrt{\sum_x \left(\ln \frac{P_i^x}{P_{\text{MSA}}^x} \right)^2} \qquad (5)$$

$$\text{(B)}\Delta\Delta G_{ij}^{\text{stat}} = kT{*}\sqrt{\sum_x \left(\ln \frac{P_{i|\delta j}^x}{P_{\text{MSA}|\delta \text{j}}^x} - \ln \frac{P_i^x}{P_{\text{MSA}}^x} \right)^2}$$

$$\text{(C)}\quad \Delta\Delta G_{ij} = \sqrt{\sum_x (\ln P_{i|\delta j}^x - P_i^x)^2}$$

where P_i^x is the probability of finding amino acid x in column i; $P^x{}_{\text{MSA}}$ is the probability of finding amino acid x in the entire MSA; $P^x_{i\,|\delta j}$ is the probability of finding amino acid x in column i in the subalignment pertubated with respect to column j.

Fodor and Aldrich[39] and Dekker et al.[38] introduced corrections to the original SCA method. First, the KT constant was avoided. Its only use was to make the equation appear as an energetic term, and it had no effect on the correlation calculation. Moreover, because the connection between the SCA score and energetic coupling was refuted,[39] it is no longer appropriate. Second, the normalization with the P_{MSA} term was removed [Eq. 5(C)].

Perturbation, Explicit Likelihood of Subset Covariation, ELSC[38]

The essence of the MSA perturbation method is similar to the SCA method. The main difference between these methods is the measurement of deviation of amino acid composition between the subset MSA and the total MSA. ELSC measures how many possible subsets of size n would have the composition found in column j [Eq. (6)].

TABLE I. Differences between Two Implementations of the Ancestral Sequence Correlation Coefficient Method

Category	Fukami-Kobayashi et al., 2002[41]	Fleishman et al., 2004[42]
Substitutions taken into consideration	reversed charge substitutions (K_E, K_D, R_E, R_D, E_K, E_R, D_K, D_R)	All
Bootstrap analysis	No	Yes
The tree building method	neighbor joining	maximum-likelihood method Tree-Puzzle
Ancestral sequences reconstruction method	Fitch parsimony	PAML

$$\prod_{x=1}^{20} \frac{\binom{N_{xj}}{n_{xj}}}{\binom{N_{xj}}{m_{xj}}} \quad (6)$$

where $N_{x,j}$ is the number of residues of type x at position j in the unperturbated MSA; $n_{x,j}$ is the number of residues of type x at position j in the subset MSA defined by column i perturbation.

Two-State Maximum Likelihood[40]

This method measures the degree to which coevolution between two sites explains the data better than a model of independent evolution. The 20 amino acid alphabet is translated into a two-state representation. Two kinds of two-state representations were proposed: based on size or charge. The LR statistics, described below, is used as a measure for the correlation between two sites. It does so by comparing the likelihood of two models: dependent and independent. The dependent model assumes constant coevolution relationship between sites.

$$LR = -2\ln\left(\frac{L_I}{L_D}\right) \quad (7)$$

where L_I is the maximum likelihood values of the independent model, and L_D is the maximum likelihood values of the dependent model.

Ancestral Sequences Correlation Coefficient[41,42]

To understand the logic behind this method, one first needs to recognize the inaccuracy introduced by the correlation coefficient method. In the correlation coefficient method all possible pairs of sequences are compared. Some of the sequences may be evolutionarily distant. Their comparison generates mutations that never took place. The ancestral sequences correlation coefficient method solves this inaccuracy by building an evolutionary tree and reconstructing the ancestral sequences. Instead of going over all possible pairs of sequences, only nodes (i.e., sequences and ancestral sequences) that are connected to each other (i.e., are evolutionary proximate) are compared. Tracing the reconstructed pathway reduces the errors generated from phylogenetically distant sequence comparison. Several differences between implementations of the ancestral sequences correlation coefficient method are detailed in Table I.

Sequence Sources for Correlated Mutations

For the success of the analysis, the selection of sequences is just as important as the selection of a correlated mutation measurement. For intramolecular position pairs one can collect orthologs and paralogs based on sequence similarity. When a protein consists of a repetitive domain and there is no difference between the roles of the repeats, it is advisable to treat each repeat as a separate sequence (see the Cohesin-Dockerin analysis). The issue of sequence selection is much more complex in intermolecular analysis. This type of analysis requires pairing of orthologs that belong to the same species. Several pairing methodologies can be adopted. Because the pairing step considerably limits the number of sequences, the selected pairing methodology is highly important. Nevertheless, the pairing methodology rarely gets the appropriate attention. In the pioneering work on hemoglobin, α and β sequence pairs were treated "as if they were a single protein with two domains by appending the sequences of the β-chain to their corresponding α-chains."[23] The question of paralogs is completely disregarded. A stringent approach to sequence pairing is offered by ADVICE, which uses the global pattern of two protein coevolution to validate protein–protein interaction.[43] Species where more than one orthologous sequence is found are excluded, because it is difficult to determine which one is the actual ortholog. This strict limitation effectively avoids making possible pairing mistakes, however, it leaves so few sequence pairs that correlated mutations analysis becomes impossible for all 15 fusion proteins explored in this work. Here, we propose three alternatives to overcome the reduced sequence availability due to sequence pairing. (1) Artificial sequences: artificial evolution experiments can provide a source for sequences that contain correlated mutations information.[44,45] This strategy is labor intensive, but provides a rich information source. (2) Fusion proteins: a combination of genes to produce a more complex protein is termed *gene fusion*. The reverse process, *gene fission*, generates multiple less complex proteins.[46] Fused proteins, from either a fusion or a fission event, can be considered as "naturally paired" sequences. They offer "pairs" free of pairing mistakes. (3) All paralogs: in some families, the interaction pattern of the relevant domains is many-to-many rather than one-to-one. That is, each paralog of one protein interacts with each paralog of the other protein. An example of such an interaction pattern is detailed below for the Cohesin-Dockerin families. Conveniently, the many-to-many interaction type allows increases in the number of

available sequences rather than decreases it. The three alternatives proposed above for dealing with the sequence pairing problem provides a solution only for a limited number of protein–protein interactions. A general automated method for sequence pairing, that counterbalances false pairings and sequence loss, is still required.

Accuracy Assessment of Intramolecular Correlated Mutations

Early reports on the accuracy of intramolecular contact prediction based on correlated mutations have been rather confusing. Gobel et al.,[25] in the work that set the basis for correlated mutations analysis in proteins, estimated the accuracy to range from 0.37 to 0.68. This estimate is high compared to later analysis. It may be a result of the small dataset that was used in this study (11 protein families).[25] In a later analysis on 173 protein families, the accuracy of correlated mutations was estimated to be as low as 0.09 using the same method (Pearson correlation coefficient).[21] Applying a neural network that includes, in addition to the correlated mutation data, conservation and secondary structure data, elevates the accuracy to 0.21. A recent thorough work that compares four methods, including a modification of the Pearson correlation, estimated the accuracy to be between 0.15 and 0.18 based on 224 protein families.[29] Similar results were obtained from 127 protein families with the physicochemical-based correlation coefficient method. The average accuracy was 0.26, 0.2, and 0.14 for αβ, β and α proteins respectively.[47] It appears that recent reports tend to agree with each other more than earlier reports. Larger datasets also make these reports more reliable. These set the accuracy level of intramolecular contacts prediction around 0.2. Although this accuracy level is higher than the estimated random level, 0.02, by 10-fold,[38] it is still too low to be used in large-scale blind prediction.

Previous Work on Intermolecular Correlated Mutations

There are far fewer studies on intermolecular contact prediction using correlated mutations compared to intramolecular studies. The first attempt was made on the hemglobin α1–β1 monomers.[23] Although it found correlated mutations a good discriminator between correct and incorrect docking solutions, no other attempts to predict intermolecular contact for two proteins, which are known to interact a priori followed. Moreover, the measurement used to distinguish between docking solutions based on correlated mutations, notated X_d, is somewhat questionable. The aim of this measurement is to detect a shift in the correlated mutation pair-population toward low distances. The X_d measurement uses 15 bins between 0 and 60 Å. In each bin the percentage of correlated pairs is compared with the percentage of all pairs as shown in Equation (8). If X_d is positive, the correlated pairs population is shifted towards low distances. However, this measurement can become positive as a result of long distance bins that are irrelevant for residue interaction. For example, in the papain protein [Fig. 1(b) in Pazos et al. 1997[23]] the major contributing bins are 12–24 Å. Residues cannot interact over such long distances, and therefore for most of the data (distances larger than 8 Å, that is, 13 out of 15 bins) the X_d representation is misleading. The weighting of bins with respect to their distance reduces the error to some extent. Correlated mutation predictions were also used indirectly to determine whether two proteins interact. The prediction relies on an "interaction index" that is obtained by comparing the distribution of intermolecular and intramolecular pair correlated values. Although the datasets used for testing the interaction index method were of limited size, encouraging results were obtained.[48]

$$X_d = \sum_{i=1}^{n} \frac{P_{ic} - P_{ia}}{d_i * n} \tag{8}$$

where P_{ic} is the percentage of correlated pairs in bin i; P_{ia} is the percentage of all pairs in bin i; d_i is the upper distance of bin i; and n is the number of bins.

Current Study Evaluates the Power of Correlated Mutations to Predict Intermolecular Contact Maps

To the best of our knowledge, this is the first attempt to directly assess the ability of correlated mutations to predict intermolecular contacts. This assessment became possible thanks to fusion protein families that bypass the obstacles that limit the dataset available for intermolecular analysis. These obstacles include complex structure availability, paralogs' discrimination, and sequence pair availability. In addition, we present a special case study of the Cohesin-Dockerin families. These families are well fitted for a correlated mutations analysis. We assess the performance of six previously suggested correlated mutation measurements. The contribution of a new method, that is based on complementarity, as wells as three modifications over existing methods is evaluated and discussed as well. Finally, we conclude that current methodologies of correlated mutations analysis are not applicable for large-scale intermolecular predictions, and therefore cannot assist docking.[49,50] The discussion emphasizes future directions for making correlated mutations applicable for this highly important task.

RESULTS

Dataset of Fusion Protein Families

Fusion proteins were selected for this work for two reasons. (1) Sequence pairing credibility and prevalence: one of the major obstacles in intermolecular correlated mutations analysis is obtaining a large number of homolog pairs with high pairing credibility. Fusion proteins provide a wealth of sequence pairs with the uppermost confidence level. In addition, special criteria can be employed for pairing sequences that belong to families of fusion proteins, but are split into two separate genes. Five criteria (separation index, fusion index, gene coverage, size ratio, and "baditude") are used in the fusion database FusionDB.[51] FusionDB assembles pairs of genes from different bacteria and archaea genomes that belong to the same cluster of orthogonal groups (COG).[52] The separation

TABLE II. The Dataset of 15 Fusion Protein Families

	Reference PDB	Seq. No.	Fusion No.	Seq. No. after 90% identity filter	Intramolecular surface couples	Intermolecular surface couples	Interface couples	Proteins' names
1	1IE7:AB	18	3	16	3828:5460	9240	20	Urease_&_subunits
2	1FFT:FH	30	5	27	74,691:13,203	63,081	72	Ubiquinol oxidase
3	1ACM:CD	19	1	13	24,976:8515	29,344	62	Aspartate transcarbamylase
4	1DTW:AB	43	14	37	44,253:25,878	67,944	155	Branched-chain α keto acid Dehydrogenase α &
5	1EFV:BA	22	2	21	21,736:29,403	50,787	285	Electron transferring Flavoprotein α &_heterodimer
6	1IIQ:AB	43	5	41	77,421:10,153	56,342	127	Anthranilate synthase components I & II
7	1FFU:DF	16	4	15	8256:24,310	28,509	108	Cuts & Cutm of carbon monoxide dehydrogenase
8	1BE3:CD	14	1	12	61,776:24,753	78,496	108	Cytochrome Bc1
9	1FM0:DE	26	4	23	2556:7875	9072	83	Molybdopterin convertin factor subunits 1 & 2
10	3EZA:BA	28	5	24	861:6903	4956	53	Phosphotransferase & HpR
11	1KQF:BA	15	2	15	33,930:219,453	173,043	218	Formate dehydrogenase iron–sulfur & major subunits
12	1O2F:BA	22	14	19	23,220:86,736	90,072	82	PTS system
13	1F6M:FH	56	5	50	35,778:3,828	23,584	76	Thioreduxin reductase & Thioreduxin
14	1K69:AB	33	16	26	52,650:2278	22,100	82	BCCP & acetyl CoA carboxylase
15	1POI:CD	33	15	32	27,495:21,115	48,410	186	Glutaconate coenzyme A-transferase

index takes advantage of the triple alignment of a fused gene with two split homologs. It measures the separation between the two split homologs in such an alignment, thereby indicating an unreasonable pairing of split homologs. Therefore, fusion proteins provide an excellent combination of cautious sequence-pairing with a large amount of sequence pairs. (2) High correlated mutation rate: currently known mechanisms that can introduce co-occurring mutations, operate at equal rates on intra- and intermolecular position pairs. It can, therefore, be argued that correlated mutations might be more ubiquitous between intra- compared to intermolecular residue pairs. If this argument is correct, then fusion proteins are expected to represent the upper limit of intermolecular correlation rate. The FusionDB provided 15 families (see Methods) suitable for correlated mutations analysis (listed in Table II).

Confirming the Suitability of the Fusion Protein Families Multiple Sequence Alignment (MSA) for Correlated Mutation Analysis

Many factors can influence the suitability of a MSA for a correlated mutation analysis. These factors can be an intrinsic trait of the protein family, like conservation background, or it can be of extrinsic origin, like a small number of sequences, uneven sampling of the sequence space or imperfect MSA. One way to confirm the suitability of a MSA for correlated mutation analysis is to examine sequential residue pairs. Because residues that are proximate in the primary sequence are always proximate in the 3D structure, we expect many mutations that change the size, or charge of an amino acid to co-occur with a compensatory mutation of a sequentially adjacent residue. We examined this assumption on 224 Pfam families (see Methods) that were previously shown to be suitable for predicting intramolecular contact maps.[38] In Figure 1 it can be clearly seen that a high correlated mutation score corresponds to an enrichment of sequential pairs. The same tendency can be clearly seen in Figure 2, which shows a similar plot for the 15 fusion protein families. Hence, the fusion dataset is suitable for further correlated mutation analysis.

Assessing the Performance of Correlated Mutation Methods on Fusion Protein Families

Given that the fusion protein families' MSAs are suitable for correlated mutation analysis, we can assess the power of such an analysis to predict intra- and intermolecular contact maps. We chose six leading correlated mutations methods: Pearson, OMES, MI, SCA, ELSC, and ancestral sequences correlation coefficient (detailed in the Introduction and Methods). The last method did not produce any results for most of the fusion families. The failure of the ancestral method to yield results in most

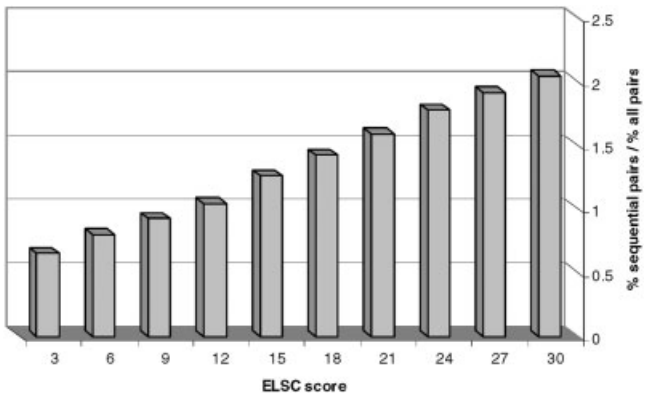

Fig. 1. Enrichment in sequential pairs as a function of correlated mutation score in 224 Pfam proteins. A correlated mutation score was calculated for all possible intramolecular pairs of 224 Pfam proteins using the ELSC method. The pairs were divided into 10 bins according to the correlated mutation score. The percentages of sequential and nonsequential pairs were calculated for each bin. The results are presented as the ratio of these two parameters, which indicates enrichment of the sequential pair. It can clearly be seen that a high correlated mutation score corresponds to an enrichment of sequential pairs.

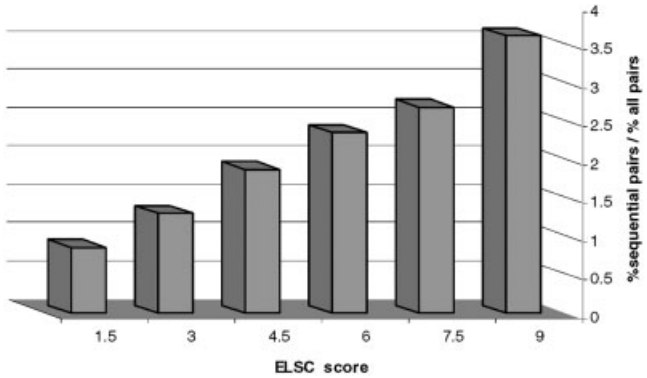

Fig. 2. Enrichment in sequential pairs as a function of correlated mutation score in 15 fusion protein families. A correlated mutation score was calculated for all possible intramolecular pairs of 15 fusion protein families using the ELSC method. The pairs were divided into five bins according to the correlated mutation score. The percentages of sequential and nonsequential pairs were calculated for each bin. The results are presented as the ratio of these two parameters, which indicates enrichment of sequential pair. The difference in the x-axis scale compared to Figure 1 is a consequence of the difference in the average number of sequences. Because ELSC is not a normalized score it tends to increase with the number of sequences. The tendency for a more profound enrichment of sequential pairs in higher scores is similar to the one observed in Figure 1.

cases is probably the outcome of employing the bootstrap analysis. This analysis provides credible results at the cost of requiring a large amount of data. The number of sequence pairs available in the fusion proteins' dataset (see Table II) falls behind the number of single sequences of potassium channel proteins on which this method was previously tested.[42] Consequently, accuracy levels are presented for the first five methods in Figure 3. Intramolecular contact maps are predicted with an average accuracy ranging from 0.15 to 0.18, with Pearson, OMES, and ELSC yielding the highest and almost indistinguishable performance. These accuracy levels are in excellent agreement with those found in the intramolecular analysis of 224 Pfam families.[38] The accuracy of intermolecular pairs is about 10 times lower than that of intramolecular pairs.

Could the Performance on Fusion Protein Families Be Improved?

The accuracy of predicting intermolecular contact maps is far from satisfying. A couple of different approaches were examined to improve the prediction accuracy. (1) Utilize other sequence-based methods for binding site prediction as a complement predictor (see section Conservation); (2) combine two correlated mutation methods to create a better united predictor (section Combining CM methods); (3) Improve the MSA on which all the abovementioned measurements rely (section Advanced MSA methods). (4) Improve the correlated mutation methods (sections: Pearson with the physicochemical matrix Miyata, Complementarity and mutual information with amino acid grouping). Because most of these strategies did not improve the accuracy on the fusion proteins' dataset, only some of the results are detailed here.

Conservation

Conservation and correlated mutations are somewhat complementary approaches. Correlated mutations are preferably detected in median levels of conservation by most correlated mutation methodologies.[39] A strictly conserved pair of residue positions as well as a highly variable pair of residue positions will get a low score by most correlated mutation algorithms. Low accuracy of predicting intermolecular contact maps can, therefore, be a result of a high conservation of interfaces. This possibility was explored by evaluating the conservation signal in the fusion protein family dataset. Figure 4 shows the scatter plot of Shannon's entropy,[30] a simple conservation measurement, for intermolecular residue pairs against their distance. In all 15 families, the polynomial function of first degree representing the scatter plot has a negative slope. This corresponds to a positive relationship between entropy and distance. However, this relationship is very weak, and does not seem to be obtained from pairs of short distance and low entropy. More advanced methods for measuring conservation rely on evolutionary trace.[53–55] One of these conservation measurements, the ConSurf method,[54] draws a similar picture (data not shown). In conclusion, it does not seem that the poor performance of the correlated mutations method is due to high conservation of interface couples in the fusion proteins dataset. Conservation does not appear to be an appropriate technique for complementing and enhancing the correlated mutation signal.

Combining CM methods

The observation that correlated mutation predictors do not tend to agree with each other raised the possibility of improving the performance by combining them.[38] The combination of the two highest performing correlated methods on the fusion families' dataset, Pearson and OMES, was examined. The combination of the two methods is presented for the Molybdopterin Convertin factor

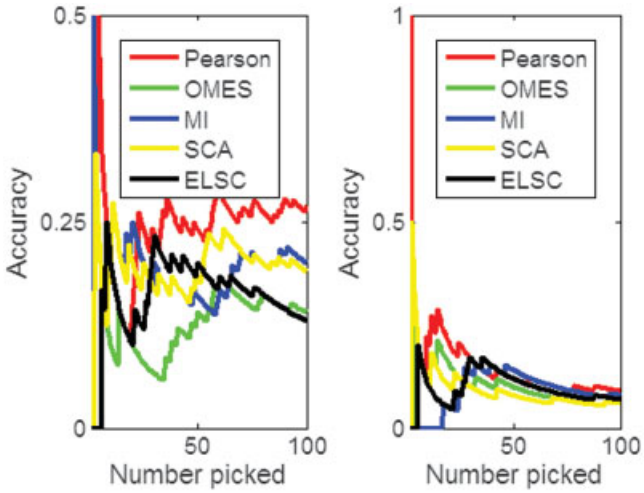

Fig. 3. Accuracy levels of five correlated mutation measurements in the fusion proteins' dataset. Average accuracy of the Pearson, OMES, MI, SCA, and ELSC on intra- (**A**) and inter- (**B**) molecular pairs.

family (PDB code 1FM0 chains DE), on which the prediction of intermolecular pairs by the correlated mutation methods was most accurate. Figure 5 plots the distance against the Pearson and OMES scores for intermolecular pairs of this family. It can be seen that none of the points corresponds to a high Pearson score, high OMES score, and a low distance. Thus, the combination of correlated mutation predictors does not seem to be a promising direction to improve the accuracy of inter-molecular prediction on the fusion proteins' dataset.

Advanced MSA methods

One of the major problems of correlated mutation analysis is its high sensitivity to multiple sequence alignment. Small errors in column assignment may have profound implications for correlated mutation scores. Therefore, the choice of the alignment method may be critical. MUSCLE is an advanced progressive alignment algorithm, which refines the MSA by dividing it to subsets, finding a profile for each subset and aligning the profiles.[56] We have used it

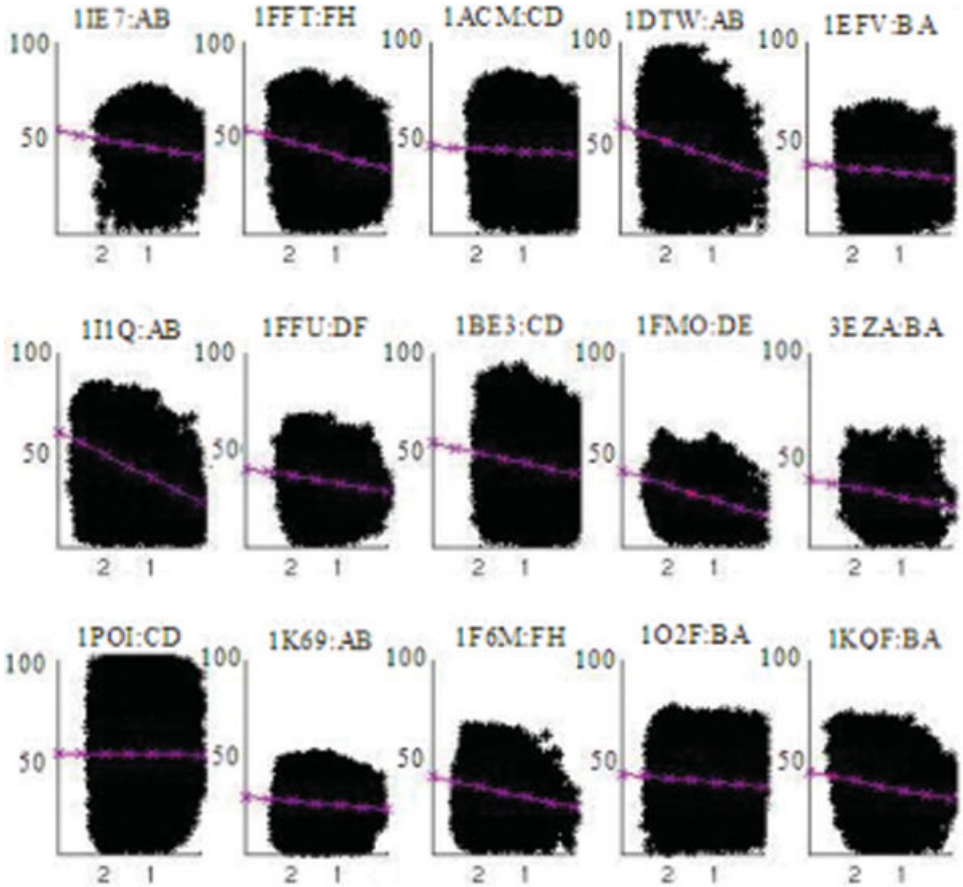

Fig. 4. Entropy and distance of intermolecular pairs for 15 fusion protein families. In each subplot a single fusion protein family is presented (PDB codes match those in Table I). Shannon's entropy, a simple conservation measurement, is plotted against the all-atoms distance for intermolecular pairs. The polynomial function of first degree representing the scatter plot is presented as a line. The polynomial functions are: ($-4.69X + 39.83$, $-6.93X + 33.80$, $-1.26X + 41.64$, $-8.64X + 30.52$, $-2.41X + 29.00$, $-12.04X + 23.18$, $-4.03X + 28.30$, $-5.43X + 37.42$, $-7.56X + 16.27$, $-4.41X + 21.05$, $-0.31X + 51.00$, $-2.32X + 22.81$, $-5.40X + 23.50$, $-2.14X + 34.93$, $-4.82X + 27.78$). All functions have a negative slope. This corresponds to a positive relationship between entropy and distance. However, this relationship is very weak and does not seem to be obtained from pairs with short distance and low entropy.

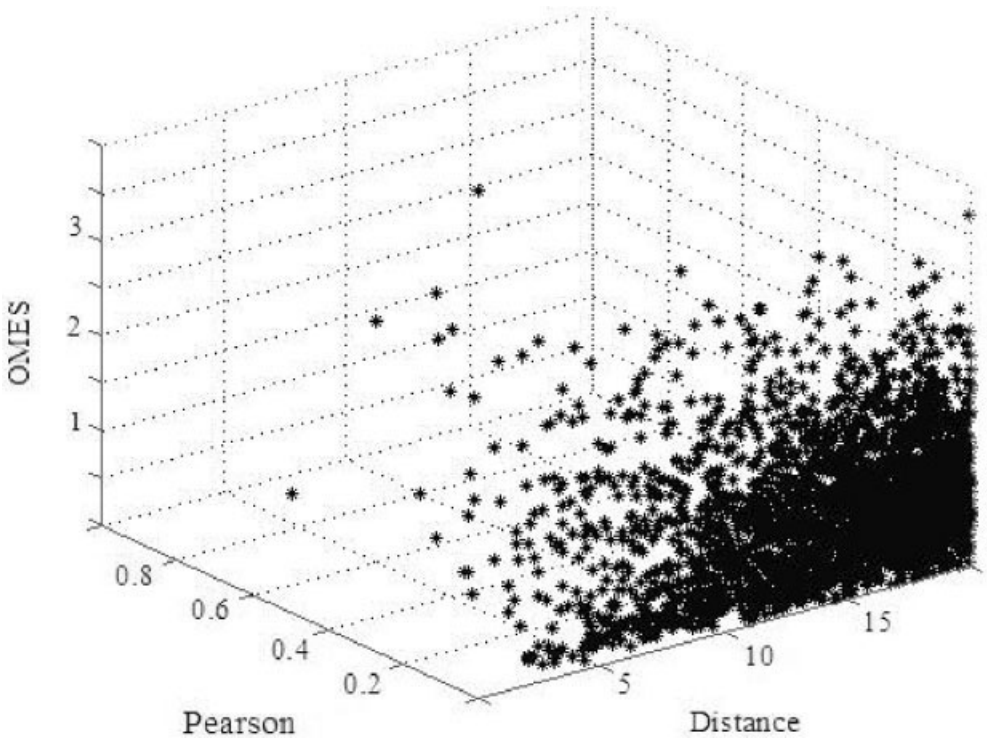

Fig. 5. Assessing the combination of Pearson and OMES scores on the Molybdopterin Convertin factor family. The scores of the two highest performing correlated methods, Pearson and OMES, are plotted against the distance for all possible intermolecular pairs of the Molybdopterin Convertin factor family (PDB code 1FM0 chains DE). The purpose is to spot pairs with a short distance and a high Pearson and OMES scores. It can be seen that no such pairs are found.

to build alternative MSAs. The improvement of intermolecular contact map prediction accuracy was inconsistent: it assisted some families, while hindering others (data not shown). This result is not surprising. The aim of both of the compared MSA methods, ClustalW and MUSCLE, is to maximize conservation. Because this goal can sometimes contradict the attempt to increase the correlations between two columns, the outcome of differences in the conservation maximization methods is unpredicted.

Pearson with the physicochemical matrix Miyata

The Pearson method, which uses the McLachlan physicochemical matrix, was among the highest performing of all tested correlated mutation methods. The McLachlan matrix is in poor agreement with another frequently used physicochemical matrix, the Miyata matrix (see Introduction). Therefore, we compared the influence of the two matrices on the performance of the correlation coefficient method. Figure 6(A) shows a significant reduction in accuracy (from 0.2 to 0.02) with the Miyata matrix compared to the McLachlan matrix.

Complementarity

Until now, no correlated mutation method was designed specifically for detection of pairs of intermolecular positions. Because amino acids show preferences for coupling across interfaces, we can use knowledge-based matrices that reflect these preferences to elevate the score of "complementing" pairs. Two such matrices, Glaser[57] and Skolnick,[58] were examined as described in Methods. Figure 6(B) shows deterioration in performance with both of these matrices.

Mutual information with amino acid grouping

Reducing the alphabet size of amino acids in a meaningful way can potentially improve the prediction power of the mutual information method. However, even in an artificial example of a perfect correlation between positive (KR), negative (ED), and aliphatic (LI) residue pairs (Table III), the difference between the two alphabets scores is minimal. The mutual information score is 0.97 in a 20 symbols alphabet. Alphabet reduction that will group K with R, E, with D and L with I, will raise the score to 1 (indicating perfect correlation). Alphabet reduction according to Taylor's Venn diagram of amino acid properties[59] did not have a significant effect on the mutual information performance over the fusion proteins' dataset (data not shown).

A Case Study: Cohesin-I–Dockerin-I

The Cellulosome is a high molecular mass, multimolecular complex in anaerobic organisms. The Cellulosome degrades cellulose, found in nature mainly in the plant cell wall. One component of the Cellulosome is Scaffoldin, that has three types of domains: (1) one or two Dockerin-II domains that bind a Cohesin-II domain, thereby anchoring

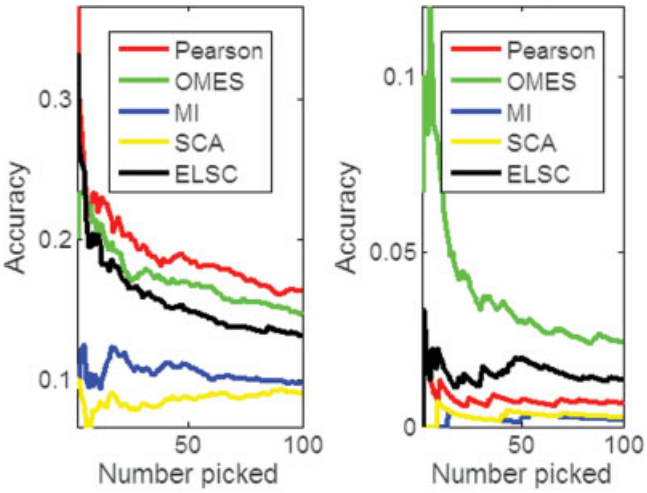

Fig. 6. Accuracy of modifications over the Pearson correlation coefficient method with similarity and complementarity matrices. (**A**) Accuracy of the Pearson correlation coefficient method with two similarity matrices: McLachlan[35] and Miyata.[34] (**B**) Accuracy of the Pearson correlation coefficient method alone and with two complementarity matrices: Glaser[57] and Skolnick.[58] Note: the lines representing the two complementarity matrices, Glaser and Skolnick, match up in a way that they cover one another.

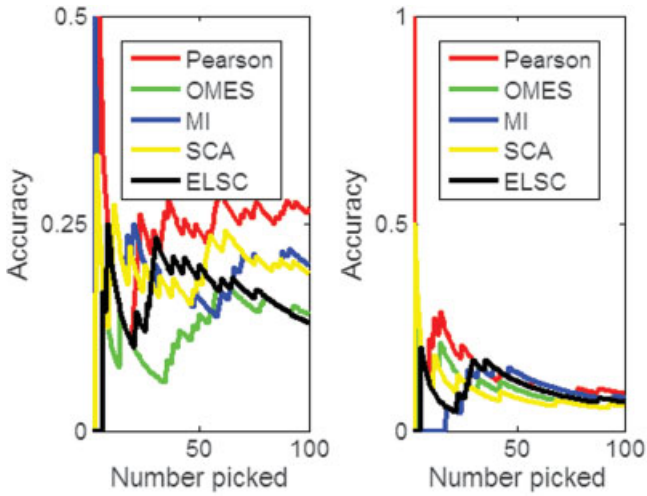

Fig. 7. Accuracy levels of five correlated mutation measurements in the Dockerin-Cohesin families. Average accuracy of the Pearson, OMES, MI, SCA, and ELSC on intra- (**A**) and inter- (**B**) molecular pairs.

TABLE III. Artificial Example of Two Columns Having Perfect Correlation Under a Reduced Alphabet

Column i	R	R	K	K	I	I	L	L	L	L
Column j	E	D	E	D	L	I	L	I	L	I

the cellulosome to the cell surface; (2) a cellulose binding domain, which binds the substrate; (3) several (1–11) Cohesin-I domains. Each of the Cohesin-I domains can bind a Dockerin-I domain, which is covalently bound to a cellulose catalyzing domain. This complex multi molecular organization of the Cellulosome contributes to high efficiency in cellulose degradation.[60]

Cohesin-I–Dockerin-I Has a High Potential for Correlated Mutations Analysis

A priori, the Cohesin-I–Dockerin-I families appear to fit well the correlated mutation analysis. Six reasons can be pointed out as the source for this suitability: (1) Because the Cohesin-I–Dockerin-I families exist in anaerobic bacteria, the low availability of genomes of higher organisms will not cause an unbalanced sampling of the sequence space. (2) The Cohesin-I and Dockerin-I domains interaction relies mainly on exposed surface residues, rather than a defined binding pocket.[61] (3) Many sequences are available (107 Cohesin and 215 Dockerin sequences are found in Pfam[62]). (4) Because each of these sequences contain several domain repetitions that differ from each other, the number of available diverse domain sequences is much larger. (5) As the Cohesin-I–Dockerin-I interaction appears to be rather nonselective,[61] we assume that every Cohesin-I interacts with every Dockerin-I domain within the same species. This assumption is supported by the "intraspecies conservation and interspecies dissimilarity" analysis.[63] This analysis divides the sequences in the MSA to groups according to their origin species. Columns which exhibit conservation inside each group and variability between these groups are expected to belong to the binding interface. The analysis proved successful in prediction of 10 species specific residue contacts in a blind docking experiment.[64] In our current analysis, available sequences Cohesin-I and Dockerin-I where divided to their domains. Each domain was treated as a separated sequence. Pairing every Cohesin-I domain with every Dockerin-I domain within the same species creates an extraordinary number, 2225, of protein–protein pairs for the correlated mutations analysis (Supplementary Material). (6) Furthermore, radical mutations (such as R → E) are prevalent in both Cohesin-I and Dockerin-I domains. All of these reasons make the Cohesin-I–Dockerin-I family an excellent candidate for correlated mutations analysis.

Assessing the Performance of Correlated Mutation Methods on Cohesin-I–Dockerin-I

The same correlated mutation methods that were examined on the fusion dataset were tested on the Cohesin-I–Dockerin-I family. Their accuracy is presented in Figure 7. Intr-molecular contact maps are predicted with an average accuracy ranging from 0.15 to 0.25, with Pearson yielding the best performance. This level is slightly higher than intramolecular prediction accuracy on the fusion families' dataset, which does not exceed 0.18. Interestingly, the accuracy of intermolecular contact maps prediction drops only by twofold. This is a noteworthy disparity from the 10-fold reduction found in the fusion proteins families. Thus, correlated mutations reach reasonable accuracy levels not only for intramolecular, but also for intermolecular contact maps prediction, in the Cohesin-I–Dockerin-I family.

DISCUSSION AND CONCLUSIONS

Over the years, the methodologies that were developed for improving the measurements of correlated mutations

(including those of the present study) were unfruitful. The basic method, the Pearson correlation coefficient, still remains the leading methodology. The accuracy of this method is reasonable but far from satisfying for large-scale blind automated prediction of intramolecular contacts. There is a drop in accuracy between intra- and intermolecular positions pairs, similar in magnitude to the decrease in signal that was observed between intra- and interdomain position pairs. The accuracy of intermolecular correlated mutations is inadequate for large-scale blind automated prediction of protein–protein interactions.

On the face of it, the results of this research call into question the habit of usage of intermolecular correlated mutations in research projects. Should (or shouldn't) correlated mutations be used? The answer should take into consideration the following: First, intra- and intermolecular correlated mutations were repeatedly demonstrated in RNA studies. Dismissing this area of research in proteins inevitably leads to the unreasonable assumption that the evolutionary rules applied to RNA and proteins are different. Second, the concept of compensatory mutations is a fundamental one in the theory of evolution. Pathogen–host interactions, for example, are partially explained based on this concept. Given the importance of the theory of evolution and our need to support it with as much evidence as possible is another substantial reason for exploring this area of research to its limit. Further, one may expect that systems that evolve at a slower rate are more likely to allow coevolution of binding partners compared to systems which evolve faster.

A certain reduction of the correlated mutations signal is anticipated due to the decrease in the ratio between true positives and the number of all possible pairs in inter- versus intramolecular pair analysis. Signal reduction may also be the result of the gene conversion mechanism, which generates correlated mutations only on the intramolecular level. This mechanism can generate a higher number of correlated mutations or it can produce a more apparent pattern of correlation, which can be detected more easily. The drop in accuracy may also be the result of the MSA. An improvement in the quality of the MSAs may allow detection of the weaker signal in intermolecular pairs. Further, the fusion protein dataset was constructed from bacteria and archaea, but not eukaryota genomes. Expanding the FusionDB by including eukaryota genomes could create a more detectable pattern of correlated mutation in two ways: first, raise the number of available sequence pairs; and second, include evolutionary distant sequences that introduce more diversity.

To date, not all strategies for measuring correlated mutations were fully explored. One example is the newly developed ELSC, a perturbation-based method. Current implementation of the ELSC method creates a single perturbation for each pair of positions.[38] The extension of this method is quite obvious: instead of a single perturbation per position pair, a few perturbations can be performed and their score weighed together. Nevertheless, this strategy is not expected to have a major effect on the ELSC accuracy in a way that will outperform the basic Pearson correlation coefficient method. Another example is the ancestral reconstruction-based methods. Current ancestral-based methods utilize only the protein sequence, while neglecting the corresponding DNA data. The DNA sequences of most sequenced proteins are available. The additional information that the DNA provides can be used to reconstruct a more accurate phylogenetic tree and ancestral sequences. Finally, two advances are proposed for the mutual information method: (1) it can be expanded to include more than statistics of two columns. Simultaneous analysis of multiple columns may be able to reduce the number of false positives. Nonetheless, performing such statistics requires raising the number of available sequences. (2) A more general definition of correlation, Tsallis entropy,[65] can be used. According to the generalized form of Shannon's information theory mutual information is just one of the possible definitions of correlation. Other forms of Tsallis entropy might outperform mutual information. However, minor modifications over existing methods are not expected to produce the significant progress needed to make correlated mutations applicable for large-scale prediction.

As we have noted above, improvement of the correlated mutations-based prediction may be possible not only by improving the correlated mutations measurement, but also by improving the MSA on which it operates. Two directions can be considered to improve MSA: (1) develop a more suitable alignment method: a multiple sequence alignment algorithm that not only maximizes the conservation of a position, but also maximizes the correlation of two positions, when evidence for such a correlation exists. This suggestion is extremely difficult, but because correlated mutations are presumed to be a fundamental mechanism of evolution, its contribution to our ability to study this process will indisputably be immense. (2) Develop a sequence pairing method. A pairing methodology that balances between pairing credibility and loss of sequence data is needed. For example, in the case of Cohesin–Dockerin, a simple automated pairing scheme, such as choosing only the most similar paralog, would have built only eight sequence pairs: one from each of the species having both a Cohesin-I domain and a Dockerin-I domain. The domain repeats would have been overlooked. This family would have been disqualified for further analysis due to insufficient sequence number, whereas a better acquaintance with this family led to an extremely large number of sequence pairs (2225) from which reliable predictions could be generated. Development of a reliable sequence pairing methodology is expected to increase the accuracy of predictions in families in which limited data availability prevents the detection of the low inter molecular signal.

Up to now the genetic mechanisms that introduce correlated mutations have been studied intensively at the molecular level, but largely overlooked in bioinformatics analyses. Knowledge of these mechanisms may contribute to a higher accuracy of a correlated mutation-based prediction. Meiotic recombination events are distributed unevenly throughout eukaryotic genomes.[66] Therefore, it might be worthwhile to examine if proteins that are

located in "meiotic hot spots" show a higher occurrence of correlated mutations. In addition, it might be valuable to examine if interacting proteins that their corresponding genes are proximate in the genome, and can undergo gene conversion, show a higher occurrence of correlated mutations.

METHODS

Fusion Proteins Dataset

The FusionDB[51] was queried for protein families with a known 3D complex structure. The sequence of the available complex was added as a reference sequence. In each family, all possible pairs of sequences were compared using ClustalW.[67] Only one sequence was retained from each sequence pair having more than 90% identity. Families that had more than 10 sequences after applying this filter set were retained for analysis. The remaining sequences were realigned using ClustalW (or MUSCLE[56] where it is specified in the Results). Overall, 15 families answered all these criteria. These are detailed in Table II.

Pfam Families' Dataset

The Pfam dataset,[62] containing 224 single protein families with at least one known structure, was previously constructed by Dekker et al.[38] Briefly, the Pfam 7.7 database was queried for families having at least one PDB structure. For each family, the structure having the longest match with at least 95% identity was selected. The sequence of the PDB structure was aligned with its closest homolog in the family using ClustalW.[67] Sequences with more than 90% identity to another sequence in the alignment were removed. Alignments that had fewer than 100 columns, which were composed of more than 50% gaps in the aligned sequences were removed. Alignments with at least 100 sequences were retained. This procedure yielded 224 protein families.

Surface

A description of the surface was created using the Molecular Surface program.[68] A residue was regarded as a surface residue if any one of its atoms had a molecular surface point.

Accuracy

Accuracy is defined as the number of correctly chosen residue pairs with an all-atom distance = 6 Å divided by the number of predictions made.

Conservation

Conservation was calculated by three alternative methods: (1) Shannon's entropy[69] according to Equation (9); (2) Consurf, using the MSA it provides;[54] (3) Consurf, using the MSA obtained for each fusion protein family as explained in the fusions proteins dataset.

$$H(x) = -\sum_{x \in X} p(x) \log p(x) \quad (9)$$

Alphabet Reduction

Alphabet reduction was performed according to Taylor's Venn diagram of amino acid properties.[59]

Complementarity Matrices

The Skolnick matrix[58] was used with opposite sign (high score = high complementarity). The Glaser matrix was used multiplied by a constant of 10^3 to create a clearer range of scores. The correlated mutation score weighted by complementarity was calculated according to Equation (10). Symbols that are not specified below are equal to those used in Equation (1).

$$CM_{i,j} = \frac{1}{N^2} \frac{\sum_{kl} (S_{ikl} - \langle s_i \rangle)(S_{jkl} - \langle s_j \rangle)}{\sigma_i \sigma_j} C_{ik,jk} C_{il,jl} \quad (10)$$

where $C_{ik,ij}$ is the complementarity of residue ik and residue jk, and $C_{il,jl}$ is the complementarity of residue il and residue jl.

ACKNOWLEDGMENTS

We thank Dr. Anthony Fodor and Prof. Richard Aldrich for providing their JAVA code for the SCA and ELSC methods. We thank Dr. Irad Ben-Gal for a fruitful discussion on information theory. This work was performed in partial fulfillment of the requirements for a Ph.D. degree of Inbal Halperin, Sackler Faculty of Medicine, Tel-Aviv University, Israel. The research of H. Wolfson and R. Nussinov in Israel has been supported in part by the "Center of Excellence in Geometric Computing and its Applications" funded by the Israel Science Foundation (administered by the Israel Academy of Sciences). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This research was supported (in part) by the Intramural Research Program of the NIH, and National Cancer Institute, Center for Cancer Research.

REFERENCES

1. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science 2002;296:750–752.
2. Moyle WR, Campbell RK, Myers RV, Bernard MP, Han Y, Wang X. Co-evolution of ligand–receptor pairs. Nature 1994;368:251–255.
3. van Kesteren RE, Tensen CP, Smit AB, van Minnen J, Kolakowski LF, Meyerhof W, Richter D, van Heerikhuizen H, Vreugdenhil E, Geraerts WP. Co-evolution of ligand–receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. J Biol Chem 1996; 271:3619–3626.
4. Hughes AL, Yeager M. Coevolution of the mammalian chemokines and their receptors. Immunogenetics 1999;49:115–124.
5. Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. Evolution of two-component signal transduction. Mol Biol Evol 2000;17:1956–1970.
6. Giver CR, Grosovsky AJ. Single and coincident intragenic mutations attributable to gene conversion in a human cell line. Genetics 1997;146:1429–1439.
7. Goodman MF. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. Annu Rev Biochem 2002;71:17–50.

8. Ninio J. Gene conversion as a focusing mechanism for correlated mutations: a hypothesis. Mol Gen Genet 1996;251:503–508.
9. Mitchell GF. Co-evolution of parasites and adaptive immune responses. Immunol Today 1991;12:A2–A5.
10. Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, Stephan W. RNA secondary structure and compensatory evolution. Genes Genet Syst 1999;74:271–286.
11. Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acid. Res 1992;20:5785–5795.
12. Williamson CL, Desai NM, Burkel JM. Compensatory mutations demonstrate that P8 and P6 are RNA secondary structure elements important for processing of a group I intron. Nucleic Acid Res 1989;17:675–689.
13. Osuna J, Soberno X, Morett E. A proposed architecture for the Central domain of the bacterial enhancer-binding proteins based on secondary structure prediction and fold recognition. Protein Sci 1997;6:6543–6555.
14. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15–21.
15. Larson SM, Di-Nardo AA, Davidson AR. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol 2000;303:433–446.
16. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignment. J Mol Biol 1998;277:419–448.
17. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–299.
18. Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 1999;293:1221–1239.
19. Rigden DJ. Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments. Protein Eng 2002;15:65–77.
20. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins 2002;48:611–617.
21. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14:835–843.
22. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins 2001;S5:157–162.
23. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein–protein interaction. J Mol Biol 1997;271:511–523.
24. Valdar WSJ. Scoring residue conservation. Proteins 2002;48:227–241.
25. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
26. Pollock DD, Taylor WR. Effectiveness of correlation analysis in identifying residues undergoing correlated evolution. Protein Eng 1997;10:647–657.
27. Neher E. How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 1994;91:98–102.
28. Taylor WR, Hatrick K. Compensating changes in protein multiple sequence alignments. Protein Eng 1994;7:341–348.
29. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 2004;56:211–221.
30. Cover TM, Thomas JA. Elements of information theory. New York: Wiley; 1992.
31. Clarke ND. Covariation of residues in the homeodomain sequence family. Protein Sci 1995;4:2269–2278.
32. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 2000;17:164–178.
33. Galitsky B. Revealing the set of mutually correlated positions for the protein families of immunoglobulin fold. In Silico Biol 2002;3:241–264.
34. Miyata T, Miyazawa S, Yasunaga T. Two types of amino acid substitutions in protein evolution. J Mol Evol 1979;12:219–236.
35. McLachlan AD. Tests for comparing related amino acid sequences. J Mol Biol 1971;61:409–424.
36. Galitsky B, Gelfand IM, Kister AE. Class-defining characteristics in the mouse heavy chains of variable domains. Protein Eng 1999;12:919–925.
37. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionary conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 2003;10:59–69.
38. Dekker JP, Fodor A, Aldrich RW, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary covariance in multiple sequence alignments. Bioinformatics 2004;20:1565–1572.
39. Fodor AA, Aldrich RW. On evolutionary conservation of thermodynamic coupling in proteins. J Biol Chem 2004;279:19046–19050.
40. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relashionship to structure. J Mol Biol 1999;287:187–198.
41. Fukami-Kobayashi K, Schreiber DR, Benner SA. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol 2002;319:729–743.
42. Fleishman SJ, Yifrach O, Ben-Tal N. An evolutionarily conserved network of amino acids mediates gating in voltage dependent potassium channels. J Mol Biol 2004;340:307–318.
43. Tan SH, Zhang Z, Ng SK. ADVICE: automated detection and validation of interaction by co-evolution. Nucleic Acid Res 2004;32:W69–W72.
44. Jespers L, Lijnen HR, Vanwetswinkel S, Van Hoef B, Brepoels K, Collen D, De Maeyer M. Guiding a docking mode by phage display: selection of correlated mutations at the staphylokinase–plasmin interface. J Mol Biol 1999;290:471–479.
45. Jucovic M, Hartley RW. Protein–protein interaction: a genetic selection for compensating mutations at the barnase–barstar interface. Proc Natl Acad Sci USA 1996;93:2343–2347.
46. Kummerfeld SK, Vogel C, Madera M, Pacoldl M, Teichmann SA. Evolution of multi-domain proteins by gene fusion and fission. Trends Genet 2005;21:25–30.
47. Vicatos S, Reddy BV, Kaznessis Y.Prediction of distant residue contacts with the use of evolutionary information. Proteins 2005;58:935–949.
48. Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 2002;47:219–227.
49. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;49:409–443.
50. Wodak SJ, Mendez R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 2004;14:242–249.
51. Suhre K, Claverie JM. FusionDB: a database for in-depth analysis of prokaryotic events. Nucleic Acid Res 2004;32:D273–D276.
52. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acid Res 2001;29:23–28.
53. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
54. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19:163–164.
55. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 2002;316:139–154.
56. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acid Res 2004;32:1792–1797.
57. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 2001;43:89–102.
58. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein–protein interactions. Biophys J 2003;84:1895–1901.
59. Taylor WR. The classification of amino acid conservation. J Theor Biol 1986;119:205–218.
60. Tavares GA, Beguin P, Alzari M. The crystal structure of a type

I Cohesin domain at 1.7 Å resolution. J Mol Biol 1997;273:701–713.
61. Shimon LJ, Bayer EA, Morag E, Lamed R, Yaron S, Shoham Y, Frolow F. A cohesin domain from *Clostridium thermocellum*: the crystal structure provides new insights into cellulosome assembly. Structure 1997;5:381–390.
62. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acid Res 2004;32:D138–D141.
63. Jindou S, Soda A, Karita S, Kajino T, Beguin P, Wu JH, Inagaki M, Kimura T, Sakka K, Ohmiya K. Cohesin–dockerin interactions within and between *Clostridium josui* and *Clostridium thermocellum*: binding selectivity between cognate dockerin and cohesin domains and species specificity. J Biol Chem 2004;279:9867–9874.
64. Inbar Y, Schneidman-Duhovny D, Halperin I, Oron A, Nussinov R, Wolfson HJ. Approaching the CAPRI challenge with an efficient geometry based docking. Proteins 2005.
65. Yamano T. A possible extension of Shannon's information theory. Entropy 2001;3:280–292.
66. Petes TD. Meiotic recombination hot spots and cold spots. Nat Rev Genet 2001;2:360–369.
67. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Res 1994;22:4673–4680.
68. Connolly M. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
69. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423.