# Prediction of Distant Residue Contacts With the Use of Evolutionary Information

**Spyridon Vicatos,[1,2] Boojala V.B. Reddy,[2] and Yiannis Kaznessis[1,2]***
[1]*Department of Chemical Engineering and Materials Science, Minneapolis, Minnesota*
[2]*Digital Technology Center, University of Minnesota, Minneapolis, Minnesota*

***ABSTRACT*** **In this work we present a novel correlated mutations analysis (CMA) method that is significantly more accurate than previously reported CMA methods. Calculation of correlation coefficients is based on physicochemical properties of residues (predictors) and not on substitution matrices. This results in reliable prediction of pairs of residues that are distant in protein sequence but proximal in its three dimensional tertiary structure. Multiple sequence alignments (MSA) containing a sequence of known structure for 127 families from PFAM database have been selected so that all major protein architectures described in CATH classification database are represented. Protein sequences in the selected families were filtered so that only those evolutionarily close to the target protein remain in the MSA. The average accuracy obtained for the alpha beta class of proteins was 26.8% of predicted proximal pairs with average improvement over random accuracy (IOR) of 6.41. Average accuracy is 20.6% for the mainly beta class and 14.4% for the mainly alpha class. The optimum correlation coefficient cutoff (cc cutoff) was found to be around 0.65. The first predictor, which correlates to hydrophobicity, provides the most reliable results. The other two predictors give good predictions which can be used in conjunction to those of the first one. When stricter cc cutoff is chosen, the average accuracy increases significantly (38.76% for alpha beta class), but the trade off is a smaller number of predictions. The use of solvent accessible area estimations for filtering false positives out of the predictions is promising. Proteins 2005;58:935–949.** © 2005 Wiley-Liss, Inc.

**Key words: correlated mutations; proximal residues; fold prediction; contact map; multiple sequence alignment**

## INTRODUCTION

The three-dimensional (3D) structure of a protein circumscribes its biological function, and is of crucial importance for studying biochemical processes in living organisms. Experimental methods such as X-ray crystallography and NMR spectroscopy can solve certain protein structures, but the processes involved are cumbersome and time consuming. This fact, combined with the necessity of analyzing millions of protein sequences determined by recent genomic projects, creates the need for efficient computational prediction methods.

Many of the existing prediction methods incorporate only sequence or local secondary structure information. Recent methods are focusing on predicting distant residue contacts in a protein, i.e., residue pairs that are distant in protein primary structure, but proximal in its native folded structure. Nonlocal, noncovalent interactions are necessary for secondary structure elements to be packed in a cohesive native structure, a structure that is favored energetically over alternative conformations.[1,2] Thus, prediction of such proximal residues is useful for protein fold prediction.[3] Among the most promising of those methods is the one that predicts residue contacts through occurrence of correlated mutations.

The basic hypothesis founding the concept of correlated mutations is that proximal residues tend to mutate in a covariant fashion. When random mutations occur in the genome of a living organism, the expressed proteins should maintain or improve their function, as well as their structural integrity. Otherwise, the living organism undergoing the mutation has fewer chances to survive in the biosphere due to undermined protein properties. Thus, when a residue playing a crucial role in the function or structure of a protein randomly mutates, the proximal residues are forced to compensate for the change by undergoing mutations covariant to the first one.[4]

Correlated mutations analysis (CMA) has been exploited in various ways. It has been widely used for residue contacts prediction of protein sequences. The predicted residue contacts were used for the construction of contact maps. Fariselli and coworkers[3,5,6] used CMA-derived residue pairs in neural networks for contact map prediction. Olmea and coworkers[7] combined information of CMA with a variety of applications such as protein docking and threading. CMA has also been used to identify functionally important residues[8] and residues involved in protein–protein interactions.[9] The most promising use of CMA results is in protein folding predictions. Ortiz and coworkers [10–12] used CMA-predicted proximal residue pairs

Grant sponsor: University of Minnesota Biotechnology Institute; Grant sponsor: American Chemical Society Petroleum Research Grant; Grant number: G7-38758.

*Correspondence to: Yiannis Kaznessis, Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Ave. SE, Minneapolis, MN 55455. E-mail: yiannis@cems.umn.edu

Received 9 June 2004; Accepted 2 September 2004

Published online 11 January 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20370

which after enrichment were used as distant constraints in Monte Carlo simulation. Overall, the method provides useful information for protein structure prediction.

Previous CMA methods have largely used amino acid substitution matrices for comparison of the protein residues and calculation of correlation scores. Those matrices were created by utilization of statistical analysis on sequence data, currently known at that time. This however could be a considerable setback for the predictive ability of the method, because statistically derived information about protein sequences is outdated with only but a minimum passing of time. The recent estimate of sequence output by genomic projects has grown to approximately 5.4 million entries per year,[13] which indicates the vast amount of change in sequence information.

Efforts have also been made to incorporate residue physicochemical properties for amino acid comparison.[14,15] Still, only certain properties such as hydrophobicity, side chain volume, and charge and polarity have been used, and the contribution of those properties to the method's predicting ability is inconclusive. So far, the average accuracy of residue contact prediction is low, less than 20%.[5,16] Singer and coworkers[16] have used structural information from Protein Data Bank (PDB)[17] to create a contact likelihood matrix which can be used for contact map prediction, but the average accuracy obtained is still low, around 15%. Such low accuracy for predicted residue contacts by CMA prohibits them for direct use in protein fold prediction.[7] Also, there has been no particular study regarding CMA's performance in different folds. Finally, questions such as under what conditions CMA should be implemented, and what the expected accuracy is for the predictions have never been addressed in a systematic manner.

In this paper, we investigate a new approach of correlated mutation analysis (CMA) and its application toward the prediction of distant residue contacts. Amino acid property vectors based upon experimental physicochemical properties have been constructed and tested for residue comparison. The predictive ability of the method is evaluated for 127 different protein families taken from PFAM database,[18] representing all major classes and architectures of CATH database.[19] We also discuss the conditions where CMA is expected to demonstrate highest accuracy for proximal residue prediction.

## MATERIALS AND METHODS

### Creation of Amino Acid Physicochemical Descriptors

The starting point of our method is the assignment of physical and chemical properties to amino acid residues. To do that, suitable amino acid physicochemical property vectors were created. These vectors, which will be called descriptors for the remainder of this work, are used for residue similarity comparison. We expect the comparison to be based on physicochemical properties, such as hydrophobicity, size, and residue charge. Additionally, application of the descriptors for residue comparison could provide us with an insight into the conditions for residue mutations and their influence on protein structure and function. Moreover, the simple forms of the descriptors make the entire method robust and easy to apply.

**TABLE I. The First Three Principal Component Descriptors**

| Amino acids | PRIN1 | PRIN2 | PRIN3 |
|---|---|---|---|
| A | −1.0169 | −7.9863 | 0.7662 |
| R | −8.4267 | 10.1801 | 0.7834 |
| N | −8.0017 | 0.3275 | −2.4554 |
| D | −10.2708 | −0.3838 | −2.7082 |
| C | 5.7345 | −3.3045 | −8.1438 |
| Q | −7.3957 | 2.5179 | −0.9235 |
| E | −8.8179 | 1.4271 | −1.3474 |
| G | −4.6415 | −10.7191 | 1.7519 |
| H | −0.9279 | 3.3268 | −2.2375 |
| I | 11.1181 | −1.9529 | 1.8220 |
| L | 9.3995 | −1.5643 | 1.8902 |
| K | −8.5038 | 7.2728 | 3.8666 |
| M | 7.8783 | 0.3062 | −2.5638 |
| F | 11.5142 | 2.6570 | −0.4756 |
| P | −1.3637 | −3.4392 | 8.9967 |
| S | −6.5985 | −5.2529 | −1.1304 |
| T | −4.0894 | −3.3886 | 0.1740 |
| W | 11.9190 | 7.9189 | 1.2488 |
| Y | 4.7959 | 6.3255 | 0.0512 |
| V | 7.6950 | −4.2680 | 0.6346 |

The AAindex database[20] contains a large number of experimental assignments, describing a large variety of physical and chemical properties of the amino acids. From these data, 142 experimental amino acid properties were manually extracted, excluding all empirically derived propensities of amino acids. In order to distinguish the important data, separate them and finally construct the descriptor vectors, we applied Principal Component Analysis[21] (PCA) on the selected set of properties. PCA reduces the dimensionality of a given set of data, and produces a new set of principal components, which account for the largest variation of the original data. It takes linear combinations of the data such that the first principal component accounts for the maximum variation, the second principal component has the next highest variation subject to being orthogonal to the first, and so on. JMP[22] statistical software package has been utilized on the set of experimental properties, previously normalized so that their mean value and standard deviation be zero and one respectively. Ninteen principal components were created that account for 99.99% of the variance in the dataset. The first three account for 72.3% of the experimental data variation. With the use of only three principal component vectors, shown in Table I, the entire original data set of properties is described with an approximate 28% loss of variation, thus the dimensionality of the original data is significantly reduced. The first principal component, PRIN1, which solely accounts for 44% of the data variation, has a strong correlation to hydrophobicity properties. The second component, PRIN2 is correlated to residue size. Finally the third component, PRIN3 is correlated to $pk_N$ values of the amino acids.

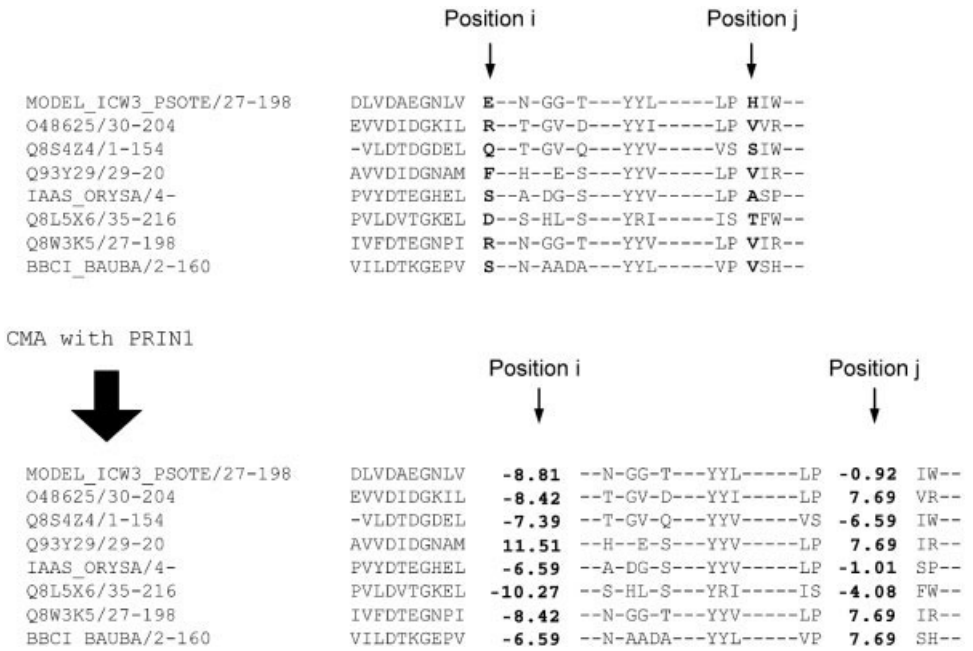


Fig. 1. Illustration of the calculation of the correlation coefficients. This example is taken from the Kunitz_legume multiple sequence alignment. The names of each protein sequence are stated at the far left of the illustration. The additional word "MODEL" has been added to the target protein, which is the first sequence of the MSA. Positions (i) and (j) of the MSA are taken, and each amino acid is replaced by its corresponding value of the descriptor. For this example, descriptor PRIN1 is used.

### Application of CMA: Evaluation of Correlation Coefficients

Following previous works[4,7,8,10,15,16,23] we calculated correlation coefficients $C_{ij}$ between pairs of positions (i) and (j) in a multiple sequence alignment (MSA). The absolute magnitude of the correlation coefficient characterizes the amount of covariant mutation of the amino acid residues located in the MSA column positions. An illustration of the application of CMA is shown in Figure 1. In our approach, each residue in an MSA position was substituted with its corresponding physicochemical property, taken from the property descriptor used for the residue comparison. For the gaps, the mean value of the descriptor for all twenty amino acids was calculated and assigned. Also, positions in the MSA with more than 10% gaps were omitted from the calculations.

The mean value and standard deviation for all column positions were calculated. Finally, correlation coefficients between two MSA positions are calculated with the use of a simple Pearson product moment correlation formula (shown in Equation 1).

$$C_{ij} = \frac{1}{N_{MSA}} \sum_{k=1}^{N_{MSA}} \frac{(q_i^k - m_i)(q_j^k - m_j)}{\sqrt{\mathrm{var}_i}\sqrt{\mathrm{var}_j}} \qquad (1)$$

Where :

- $c_{ij}$: The correlation coefficient between positions (i) and (j).
- $q_i^k$, $q_j^k$ : The values of the descriptor chosen for the aminoacids in the (i) and (j) position, for the protein sequence k.
- $m_i$, $m_j$: The mean values of the descriptor for the positions (i) and (j).
- $var_i$, $var_j$ : The variances of the descriptor for the positions (i) and (j).
- $N_{MSA}$: The number of sequences in the MSA.

The term $\frac{(q_i^k - m_i)}{\sqrt{\mathrm{var}_i}}$ is related to the relative deviation of the physicochemical property $q_i^k$ for each residue at position (i) from the mean value $m_i$, normalized by the standard deviation $\sqrt{\mathrm{var}_i}$. The total sum of the products of relative deviations between two MSA positions (i) and (j) indicates the mutation covariance between the two positions.

In this work, the magnitude rather than the sign of the correlation coefficient is considered to be indicative of the covariance between two positions. When a covariant mutation occurs, the properties of two interacting positions may simultaneously be reduced or increased, resulting in a positive correlation coefficient for this property. The positions however might undergo such mutation, that the property of one increases while the other decreases, so that the average value of this property in the two-residue cluster remains the same. In this case, a negative correlation coefficient occurs. The hypothesis behind CMA is that both cases are the result of interactions between proximal residues. We expect distant residues to have low magni-

tude of correlation coefficients, which allows us to separate them from the proximal ones.

### Multiple Sequence Alignments

All available multiple sequence alignments (MSA) were taken from the PFAM protein family database, release 12.0, January 2004. This choice was made because PFAM is considered to be state of the art database of protein families.

Each chosen PFAM MSA contains at least one protein sequence of known structure. CMA does not require a protein of known structure to be applied. However, this protein's atom coordinates would be the experimental reference with which the theoretical results from CMA are compared and evaluated. The corresponding atom coordinates of the target proteins for each family are obtained from the PDB databank, and checked for possible inconsistencies. In terms of inconsistencies, we refer to cases where several atom coordinates could not be experimentally obtained; therefore they were not reported and stored into the PDB file.

The multiple sequence alignments were chosen in such a way that all major protein architectures described in the CATH classification database were well represented. The actual choice of the families and therefore the alignments was arbitrary, because the entire range of protein folds in CATH could not be covered in its entirety. Major difficulties such as inconsistencies in the PDB files of certain families, the small number of protein sequences aligned, and the partial inclusion of the target protein's sequence in the MSA prohibited us from performing CMA on all folds in CATH. The goal was to cover all major architectures of CATH with a sufficient amount of families, so that the results obtained would provide us with a clear picture as to how the CMA method behaves in different folds. Eventually 127 families were chosen. From the mainly alpha class, 16 orthogonal bundles, 14 up-down bundles, two barrels and one solenoid architecture proteins have been chosen, a total of 33 mainly alpha protein families. From the mainly beta class, nine sandwiches, three distorted sandwiches, three beta ribbons, three beta rolls, six beta barrels, two trefoils, one beta aligned prism, one beta orthogonal prism, one three-layer sandwich, four propelors, one beta solenoid and two beta complex architecture proteins are chosen, a total of 39 mainly beta protein families. From the alpha-beta class, five rolls, one super roll, seven barrels, 12 two-layer sandwiches, 11 aba three-layer sandwiches, two bba three-layer sandwiches and eight alpha–beta complex architecture proteins are chosen, a total of 46 alpha–beta protein families. Finally, nine unclassified–irregular architecture proteins, a total of nine complex–unclassified protein families were chosen.

### Filtering of Multiple Sequence Alignments

When families are created and added in the PFAM database, they contain a large number of distant homolog proteins, which may not be structurally similar to the target protein. Also, plenty of the listed protein sequences may be highly homologous to each other, thus the evolutionary information corresponding to these sequence sets is redundant. To avoid using redundant or structurally incompatible information in CMA, filtering of the MSA is required to clear out the undesirable sequences.

Obtaining a filtered multiple sequence alignment with a sufficient number of sequences in it so that CMA can be performed is straightforward. All filtering steps should use the default threshold values, so that the final MSA should contain at least 25 sequences. Studied families where filtering gives a MSA containing less than 25 sequences should not be used for CMA since the expected accuracy is very low.

The filtering involves three steps:

1. In the first step, all protein sequences containing more that 50% gaps when aligned in the MSA, were excluded. PFAM alignments often contain small parts of protein sequences with considerably smaller numbers of residues than the target protein. These parts are introduced into the MSA with a large amount of gaps filling the remaining aligned space. We consider those small entries to provide false structural information to CMA, because we expect their nonaligned, larger parts of these sequences (which are not included in the MSA) to be of considerably different structure and function compared to the remaining sequences in the MSA. For a few MSAs a strict gap threshold value eliminates a fairly large number of sequences, leaving us with less than 25 sequences. In these cases, the less strict value of 60% has been used.
2. The second step is the exclusion of sequences with sequence identity to the target protein less than a certain threshold value. This threshold varies and depends upon the protein family that is applied. The threshold should be such that the number of the remaining sequences should be sufficiently large for CMA application, i.e., larger than 25. We find that sequence identity threshold of 20% is the most suitable for the filtering process. Again, only in a few cases a different sequence identity threshold has been used.
3. The final filtering step involves the identification and omission of very distant homologs of the target protein, and the sequences containing redundant evolutionary information. The starting point of this procedure is the calculation of the evolutionary distance between the target protein and the remaining proteins in the MSA. The Gonnet amino acid substitution matrix[24] had been used for the calculation of log odd similarity scores.

The evolutionary distance between target protein and any sequence included in the MSA is calculated as follows:

$$ed(i,tp) = \left(1 - \frac{\log odd\ score(i,tp)}{\log odd\ score(tp,tp)_i}\right) \cdot 100 \qquad (2a)$$

$$\log odd\ score(i,tp) = \sum_{k=1}^{alignment\ size} subst([aa_i]_k,[aa_{tp}]_k) \qquad (2b)$$

| MSA Position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Target Protein | M | E | N | --- | Q | N | --- | --- |
| Sequence 1 | M | E | --- | R | Q | N | A | E |
| Sequence 2 | M | E | K | G | V | P | S | T |

Fig. 2. An example of calculating evolutionary distances log odd score(tp, tp)$_i$. Target protein has gaps at position 4,7,8, while sequence 1 has only one gap at position 3 and sequence 2 has no gaps. For the calculation of log odd score(tp, tp)$_1$, which corresponds to sequence 1, positions 3, 4, 7, and 8 are omitted, while for odd score(tp)$_2$f, which corresponds to sequence 2, position 3 is included in the calculations, since there is no gap. Therefore: $\log odd\ score(tp,tp)_1 \neq \log odd\ score(tp,tp)_2$.

$$\log odd\ score(tp,tp)_i = \sum_{k=1}^{alignment\ size} subst([aa_{tp}]_k,[aa_{tp}]_k) \quad (2c)$$

Where:

- $ed(i,tp)$ is the evolutionary distance between protein sequence (i) and the target protein (tp)
- $\log odd\ score(i,tp)$ is the log odd score of sequence (i) if it mutates into the target protein
- $\log odd\ score(tp,tp)_i$ is the log odd score of the target protein if it mutates into itself
- $[aa_{tp}]_k$ is the amino acid of the target protein at the alignment position k
- $[aa_i]_k$ is the amino acid of sequence (i) at the alignment position k
- $subst(aa_1,aa_2)$ is the value of the Gonnet substitution matrix for the mutation of $aa_1$ amino acid, into the $aa_2$ amino acid

Positions containing gaps were ignored in the calculations. This means the parameter k for Equations 2b and 2c runs for all the positions in the MSA which contain no gaps for both the target protein as well as for the currently chosen sequence (i). It should be mentioned that $\log odd\ score(tp,tp)_i$ differs for different sequences (i) because of the different gap distribution between the target protein and each sequence entry (i) in the MSA. This is shown in the example illustrated in Figure 2.

Protein sequences having an evolutionary distance to the target protein greater or equal to 90 were omitted from calculations of CMA, because they were considered to be evolutionarily very distant.

The same procedure is followed for the omission of MSA sequences containing redundant information. The average evolutionary distance of all pair combinations of sequences is calculated, using Equations 3a and 3b.

$$ed(i,j) = \left(1 - \frac{\log odd\ score(i,j)}{\log odd\ score(j,j)}\right) \cdot 100 \quad (3a)$$

$$\langle ed(i,j)\rangle = \frac{ed(i,j) + ed(j,i)}{2} \quad (3b)$$

When a sequence pair (i, j) has an evolutionary distance score equal or lower to a threshold value of 5, then sequence (j) is omitted from the calculation of CMA. Again, the value of the threshold depends solely upon the tested family, and should be such that a sufficiently large number of protein sequences remains for CMA application. Concerning the last two steps, the filtering process is less sensitive to the value of the evolutionary distance threshold. The vast majority of the sequences in PFAM MSAs are very distant to each other, and a threshold value of 5 is sufficient to eliminate all evolutionary redundant sequences. A less strict value of 4 or 3 for the evolutionary threshold is applied only in a few extreme cases, when the final number of sequences after the first three filtering steps is very small.

### Residue Contact Distances

Residue distances for the target protein are required for the evaluation of the predictive ability of CMA. The distance between two residues can be defined in different ways. In this work, four types of distances are being investigated: Cα distances, Cβ distances, distances between residue centers of mass and finally the residues minimum distance between residues. All mentioned distances for the target protein were calculated from the coordinates reported in its corresponding PDB file. Hydrogen atoms have been omitted from the calculations.

## RESULTS AND DISCUSSION

### Correlation Coefficient Cutoff and Accuracy of CMA

As stated in previous works, residue pairs showing high covariant mutation tend to be proximal. Following the example of Goebel et al., Fariselli et al., Olmea et al., and Ortiz et al.,[3,4,7,10,11] an arbitrary positive cutoff value for the correlation coefficient has been defined. Residue pairs having an absolute value of correlation coefficient equal or larger to the cutoff value are predicted to be proximal.

Looking at the structures of proteins, pairs of residues separated from each other with eight or more consecutive residues, whose distance in the tertiary structure is equal or less than a threshold value of 6 Å, are considered to be proximal. Pairs of residues that are less than eight consecutive residues are omitted from the calculations. CMA was applied to 127 families shown in Table II–V. With the use of the first three descriptors shown in Table I, CMA is evaluated by comparing the predictions with the actual experimental data of the target protein for each family.

The accuracy (acc) of CMA for a certain cc cutoff is calculated as follows:

$$acc = \frac{TP}{TP + FP} \quad (4)$$

Where $TP$ is the number of truly proximal residue pairs predicted to be in contact and $FP$ is the number of falsely proximal residues predicted to be in contact by CMA.

Changing the cutoff value, the number of pairs for each category changes, thus accuracy is affected. In order to investigate the dependence of accuracy and the number of

**TABLE II. Performance of CMA for Mainly Alpha Families**

| Family name[a] | PDB[b] ID | Aligned length[c] | PRIN1 | | | | PRIN3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | cc cutoff[d] | Acc[e] | N[f] | IOR[g] | cc cutoff | Acc | N | IOR |
| IRF | 1IF1 | 105 | 0.65 | 0.3 | 10 | 2.25 | 0.55 | 0.1429 | 7 | 1.0714 |
| HTH_8 | 1FIP | 41 | 0.25 | 0.4286 | 7 | 2.7836 | 0.2 | 0.4444 | 18 | 2.8867 |
| Arg_repressor | 1AOY | 71 | 0.65 | 0.2857 | 7 | 4.06 | 0.65 | 0.18 | 11 | 2.58 |
| ACBP | 1ACA | 85 | 0.5 | 0.1875 | 16 | 3 | 0.5 | 0.2105 | 19 | 3.38 |
| FE_DEP_REPR_C | 2TDX | 71 | 0.5 | 0.2143 | 14 | 3.72 | 0.45 | 0.2143 | 14 | 3.72 |
| PEP-utilisers_N | 1ZYM | 125 | 0.5 | 0.1333 | 15 | 7.065 | 0.5 | 0.067 | 15 | 3.5327 |
| Endotoxin_N | 1JI6 | 226 | 0.7 | 0.1 | 10 | 4.75 | 0.6 | 0.22 | 9 | 10.55 |
| Bcl-2 | 1LXL | 99 | 0.5 | 0.1538 | 13 | 3.59 | 0.6 | 0.14 | 7 | 3.33 |
| Ribosomal_S7 | 1RSS | 135 | 0.8 | 0.1 | 20 | 2.87 | 0.55 | 0.055 | 36 | 1.6 |
| ribonuc_red_sm | 1R2F | 276 | 0.85 | 0.22 | 27 | 11.63 | 0.85 | 0.1951 | 41 | 10.21 |
| Phenol_Hydrox | 1MTY | 228 | 0.95 | 0.0346 | 231 | 1.53 | 0.95 | 0.0278 | 180 | 1.2293 |
| Cytochrom_C_2 | 1BBH | 126 | 0.65 | 0.1471 | 34 | 5.03 | 0.6 | 0.09 | 44 | 3.11 |
| Phospholip_A2_1 | 1BUN | 119 | 0.7 | 0.1429 | 42 | 3.02 | 0.75 | 0.2414 | 29 | 5.11 |
| TarH | 1VSL | 136 | 0.95 | 0.1351 | 37 | 4.72 | 0.85 | 0.0989 | 91 | 3.45 |
| ATP-cone | 1R1R | 85 | 0.75 | 0.1538 | 13 | 3.5632 | 0.6 | 0.909 | 11 | 2.1 |
| ATP-synt_ab_C | 1BMF | 105 | 0.7 | 0.058 | 17 | 1.84 | 0.55 | 0.0357 | 28 | 1.11 |
| Interferon | 1AU1 | 166 | 0.8 | 0 | 7 | 0 | 0.65 | 0.0385 | 14 | 1.3214 |
| Photo_RC | 1QOV | 258 | 0.9 | 0.0312 | 32 | 2.1 | 0.8 | 0.0429 | 70 | 2.88 |
| Hormone_recep | 1FCY | 181 | 0.5 | 0.0357 | 28 | 1.36 | 0.5 | 0 | 8 | 0 |
| LYASE_1 | 1FUP | 331 | 0.8 | 0.1538 | 13 | 10.69 | 0.8 | 0.1313 | 30 | 9.27 |
| Glyco_hydro_8 | 1CEM | 354 | 0.9 | 0.0556 | 18 | 2.95 | 0.9 | 0.083 | 12 | 4.43 |
| Glyco_hydro_9 | 1TFU | 436 | 0.95 | 0.0619 | 291 | 3.8079 | 0.95 | 0.0984 | 122 | 6.0552 |
| PCP | 1PPR | 146 | 0.9 | 0 | 16 | 0 | 0.9 | 0.0769 | 13 | 4.8707 |
| ATP-synt_DE | 1AQT | 45 | 0.5 | 0.125 | 8 | 1.6799 | 0.4 | 0.2 | 15 | 2.6878 |
| Chorismate_mut | 1ECM | 85 | 0.5 | 0 | 13 | 0 | 0.35 | 0 | 36 | 0 |
| Acyl_CoA_dh | 1IVH | 150 | 0.4 | 0.1429 | 28 | 5.2016 | 0.4 | 0.1333 | 15 | 4.8548 |
| COX3 | 2OCC | 256 | 0.6 | 0.2143 | 14 | 5.2878 | 0.6 | 0.1 | 10 | 2.4676 |
| ATP-synt_C | 1A91 | 70 | 0.5 | 0.1538 | 13 | 3.1476 | 0.4 | 0 | 11 | 0 |
| PARP_reg | 1A26 | 134 | 0.95 | 0.2222 | 9 | 8.4902 | 0.65 | 0.0536 | 56 | 2.0467 |
| HMG_box | 1QRV | 67 | 0.25 | 0.1111 | 54 | 4.0037 | 0.25 | 0.069 | 58 | 2.485 |
| CheR_N | 1AF7 | 57 | 0.45 | 0.2222 | 9 | 4.8611 | 0.45 | 0.2 | 10 | 4.375 |
| RGS | 1AGR | 114 | 0.55 | 0.3 | 10 | 4.503 | 0.45 | 0.4444 | 9 | 6.6711 |
| Gag_p17 | 1HIW | 115 | 0.7 | 0.1429 | 7 | 3.8798 | 0.55 | 0 | 11 | 0 |

[a]The PFAM name.
[b]The PDB ID name of the target protein, used for the predicting ability of CMA.
[c]The size of the target protein sequence, aligned in the MSA.
[d]The correlation coefficient cutoff where improvement over random is maximum and additionally the number of predictions is minimum seven pairs.
[e]Accuracy.
[f]Number of predictions.
[g]Improvement over random accuracy.

predicted pairs on the cc cutoff, accuracy plots were created. The true positive and false positive pairs are identified for a varying cc cutoff. Then the accuracy *acc* and the total number of predicted pairs $N_{pairs}$ are plotted against the cc cutoff.

### Accuracy Plots, Minimum Distance, Number of Predicted Contacts

Increasing the cc cutoff and selecting the residue pairs with higher or equal to the cc cutoff absolute value of correlation coefficient, we enrich our selection with proximal pairs. This happens for all four different residue distances, for all three descriptors, as it was expected since previous research has shown an increase of contact predictions by increasing cc cutoff. However, accuracy results show that from all four different residue distance definitions, the minimum distance is the one that gives the highest accuracy results. This is illustrated in Figure 3 for the protein target 1JUL of the IGPS family.

An increase of the cc cutoff positively affects the accuracy of the predicted pairs. The tradeoff however for the increase of the accuracy is the dramatic decrease of the number of the predicted pairs, as illustrated in Figure 4.

### Accuracy and Improvement Over Random Accuracy for Different Folds

For each family of the set of the 127 families, the improvement over random accuracy (IOR) was calculated. As random accuracy, we refer to the fraction of the residue pairs of a protein sequence at least eight residues apart, which are proximal in its tertiary structure. The improve-

**TABLE III. Performance of CMA for Mainly Beta Families**

| Family name[a] | PDB[b] ID | Aligned length[c] | PRIN1 | | | | PRIN3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | cc cutoff[d] | Acc[e] | N[f] | IOR[g] | cc cutoff | Acc | N | IOR |
| Crystall | 1PRR | 82 | 0.6 | 0.1818 | 11 | 1.9331 | 0.65 | 0.2857 | 7 | 3.0377 |
| Dioxygenase_C | 3PCG | 167 | 0.4 | 0.1 | 10 | 2.1588 | 0.35 | 0.0769 | 13 | 1.6606 |
| F5_F8_type_C | 1GOF | 129 | 0.55 | 0.3 | 10 | 3.9704 | 0.35 | 0.1538 | 13 | 2.0361 |
| Polyoma_coat | 1VPS | 285 | 0.95 | 0.0388 | 103 | 1.4904 | 0.95 | 0.0227 | 88 | 0.8721 |
| RHD | 2RAM | 166 | 0.95 | 0.2857 | 7 | 5.4906 | 0.9 | 0.2727 | 11 | 5.241 |
| P53 | 1TSR | 196 | 0.95 | 0.0791 | 139 | 2.1068 | 0.95 | 0.0729 | 96 | 1.9412 |
| Gal-bind_lectin | 1IS6 | 133 | 0.5 | 0.375 | 16 | 5.5349 | 0.45 | 0.1 | 10 | 1.476 |
| Glyco_hydro_11 | 1C5H | 184 | 0.8 | 0.2222 | 9 | 4.7965 | 0.7 | 0.25 | 16 | 5.3961 |
| Sod_Cu | 1ESO | 143 | 0.55 | 0 | 7 | 0 | 0.55 | 0 | 14 | 0 |
| Glyco_hydro_7 | 1CEL | 431 | 0.95 | 0.061 | 82 | 3.0111 | 0.95 | 0.1077 | 65 | 5.3181 |
| Cu_amine_oxid | 1AV4 | 414 | 0.95 | 0.1818 | 22 | 9.2918 | 0.9 | 0.2 | 20 | 10.221 |
| PTS_EIIA_1 | 1F3Z | 105 | 0.6 | 0.1875 | 16 | 2.6231 | 0.5 | 0.3571 | 14 | 4.9963 |
| UPAR_LY6 | 1CDQ | 71 | 0.65 | 0.1429 | 7 | 1.3238 | 0.55 | 0.3636 | 11 | 3.3697 |
| Urease_beta | 1A5M | 100 | 0.75 | 0.125 | 8 | 2.0369 | 0.55 | 0.125 | 32 | 2.0369 |
| PDGF | 1PDG | 84 | 0.9 | 0.2222 | 9 | 2.5829 | 0.95 | 0.2222 | 9 | 2.5829 |
| PDZ | 1IU0 | 85 | 0.2 | 0.2222 | 9 | 2.9921 | 0.15 | 0.125 | 8 | 1.683 |
| PH | 2DYN | 106 | 0.5 | 0.5455 | 11 | 7.1059 | 0.4 | 0.0909 | 11 | 1.1843 |
| CAT_RBD | 1AUU | 55 | 0.55 | 0.1667 | 6 | 1.3737 | 0.45 | 0.3333 | 9 | 2.7475 |
| Serpin | 1ATU | 371 | 0.75 | 0.2727 | 11 | 8.6224 | 0.5 | 0.3333 | 9 | 10.5385 |
| Ribosomal_S8 | 1AN7 | 135 | 0.6 | 0.1667 | 12 | 1.4987 | 0.55 | 0.2222 | 9 | 1.9983 |
| SH3 | 1BBZ | 56 | 0.15 | 0.4688 | 32 | 3.7158 | 0.15 | 0.2 | 15 | 1.5854 |
| Glyco_hydro_45 | 2ENG | 199 | 0.9 | 0.1013 | 79 | 1.7 | 0.95 | 0.1 | 60 | 1.6788 |
| HCV_NS3 | 1A1R | 149 | 0.95 | 0.0645 | 186 | 1.3264 | 0.95 | 0.0818 | 159 | 1.6809 |
| Kringle | 1KDU | 80 | 0.65 | 0.1765 | 17 | 1.6906 | 0.65 | 0.2 | 10 | 1.916 |
| SNase | 2SNS | 134 | 0.45 | 0.2 | 15 | 3.0407 | 0.45 | 0.2857 | 7 | 4.3438 |
| Ribosomal_L14 | 1WHI | 122 | 0.55 | 0.0938 | 64 | 1.5108 | 0.5 | 0.0727 | 55 | 1.172 |
| Pro_isomerase | 1OCA | 158 | 0.65 | 0.2308 | 13 | 4.039 | 0.5 | 0.4 | 20 | 7.0009 |
| Kunitz_legume | 1FMZ | 172 | 0.7 | 0.3125 | 16 | 5.1234 | 0.55 | 0.625 | 8 | 10.2469 |
| IL1 | 6I1B | 143 | 0.7 | 0.3333 | 24 | 5.0198 | 0.8 | 0.2727 | 11 | 4.1071 |
| Jacalin | 1JAC | 133 | 0.55 | 0.1333 | 15 | 2.0806 | 0.5 | 0.1333 | 15 | 2.0806 |
| B_lectin | 1JPC | 107 | 0.8 | 0.3 | 10 | 4.413 | 0.65 | 0.1 | 20 | 1.471 |
| Rieske | 1RIE | 93 | 0.5 | 0.3333 | 9 | 3.7136 | 0.5 | 0.4286 | 7 | 4.7746 |
| PQQ | 1H4I | 38 | 0.35 | 0.125 | 8 | 0.7466 | 0.3 | 0.3636 | 11 | 2.172 |
| NEUR | 1MWE | 385 | 0.95 | 0.1219 | 361 | 4.6325 | 0.95 | 0.1573 | 267 | 5.9788 |
| Peptidase_S9 | 1QFS | 81 | 0.65 | 0.2727 | 11 | 6.7386 | 0.4 | 0.1333 | 15 | 3.2944 |
| Cytochrom_D1 | 1NIR | 395 | 0.8 | 0.2 | 5 | 4.1519 | 0.8 | 0.2222 | 9 | 4.6132 |
| Pec_lyase_C | 1AIR | 209 | 0.75 | 0.4 | 10 | 9.2211 | 0.7 | 0.1818 | 11 | 4.1914 |
| Topoisom_I_N | 1OIS | 219 | 0.95 | 0.2 | 70 | 7.7696 | 0.95 | 0.2941 | 34 | 11.4259 |
| Hom_end_hint | 1VDE | 177 | 0.85 | 0.1111 | 9 | 0.7646 | 0.8 | 0.1667 | 6 | 1.1469 |

[a]The PFAM name.
[b]The PDB ID name of the target protein, used for the predicting ability of CMA.
[c]The size of the target protein sequence, aligned in the MSA.
[d]The correlation coefficient cutoff where improvement over random is maximum and additionally the number of predictions is minimum seven pairs.
[e]Accuracy.
[f]Number of predictions.
[g]Improvement over random accuracy.

ment over random accuracy is therefore simply the ratio of the obtained accuracy for each cc cutoff, over random accuracy. For all but a few of the protein families tested, regardless of the size and the fold of their target protein, both accuracy and IOR increase when cc cutoff increases.

Not all folds demonstrate similar behavior in CMA. The cc cutoff used for the predictions was chosen so that the highest IOR is achieved, when at the same time CMA predicts at least seven pairs. These cc cutoff values and IOR are shown in Tables II–V. IOR and accuracy obtain very high values at high cc cutoff of around 0.8–0.9. However, as we stated above, the number of predicted contacts is very low at this cc threshold.

It is noted that there are families such as Phenol_Hydrox, Glyco_hydro_9, Xylose_isom, HCV_NS3 and NEUR, for which CMA predicts a very large number of residue contacts, and influence the average number of predictions and the standard deviation to higher numbers. A typical number of predictions for the rest of the families is, in actuality, around 20. No apparent connection between accuracy or IOR and the size of the fold was detected. The accuracy however is dependent upon

**TABLE IV. Performance of CMA for Alpha–Beta Families**

| Family name[a] | PDB ID[b] | Aligned length[c] | PRIN1 | | | | PRIN3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | cc cutoff[d] | Acc[e] | N[f] | IOR[g] | cc cutoff | Acc | N | IOR |
| Fer2 | 1FRD | 77 | 0.35 | 0.4286 | 7 | 5.5934 | 0.25 | 0.0909 | 11 | 1.1865 |
| IF3_N | 1TIF | 75 | 0.7 | 0.125 | 8 | 1.8702 | 0.55 | 0.125 | 8 | 1.8702 |
| QRPTase_C | 1QAP | 96 | 0.65 | 0.0167 | 60 | 0.5359 | 0.7 | 0.125 | 16 | 4.0189 |
| Cyt-b5 | 1B5M | 75 | 0.4 | 0.2 | 10 | 2.9986 | 0.35 | 0.125 | 8 | 1.8741 |
| Carb_anhydrase | 1DMX | 236 | 0.8 | 0.4375 | 16 | 11.2559 | 0.7 | 0.6364 | 11 | 16.3722 |
| LBP_BPI_CETP | 1BP1 | 177 | 0.7 | 0.2727 | 11 | 6.6479 | 0.6 | 0.1818 | 11 | 4.432 |
| Aminotran_4 | 3DAA | 268 | 0.7 | 0.125 | 8 | 4.8777 | 0.6 | 0.125 | 8 | 4.8777 |
| Pterin_bind | 1AD4 | 186 | 0.8 | 0.5714 | 7 | 18.4796 | 0.65 | 0.2727 | 11 | 8.8198 |
| TIM | 1YDV | 246 | 0.7 | 0.1429 | 21 | 4.2987 | 0.45 | 0.2 | 55 | 6.0181 |
| Enolase_C | 1ONE | 292 | 0.8 | 0.5238 | 21 | 15.9043 | 0.75 | 0.5 | 12 | 15.1813 |
| Xylose_isom | 1CLK | 378 | 0.95 | 0.1056 | 161 | 7.3339 | 0.95 | 0.0882 | 102 | 6.1285 |
| Crystatin | 1CEW | 161 | 0.45 | 0.0526 | 19 | 0.6468 | 0.35 | 0.04 | 75 | 0.4916 |
| Potato_inhibit | 1CSE | 63 | 0.65 | 0.2857 | 7 | 2.7867 | 0.65 | 0.1429 | 7 | 1.3934 |
| 2-oxoacid_dh | 1EAF | 232 | 0.65 | 0.0568 | 88 | 2.2922 | 0.8 | 0.1429 | 7 | 5.7633 |
| Thymidylat_synt | 1TYS | 264 | 0.9 | 0.2143 | 14 | 8.3956 | 0.9 | 0.3333 | 9 | 13.0598 |
| zf-C4 | 1GLU | 74 | 0.75 | 0.5556 | 9 | 8.4132 | 0.75 | 0.5 | 8 | 7.5719 |
| Thionin | 1BHP | 45 | 0.5 | 0.2143 | 14 | 1.6378 | 0.65 | 0.2857 | 7 | 2.1837 |
| Ribosomal_S6 | 1RIS | 92 | 0.55 | 0.5 | 8 | 7.3692 | 0.55 | 0.1429 | 7 | 2.1055 |
| Acyl_transf_1 | 1MLA | 303 | 0.5 | 0.075 | 40 | 3.4406 | 0.45 | 0.1176 | 34 | 5.397 |
| PPV_E2_C | 1DBD | 81 | 0.75 | 0.2 | 10 | 3.5917 | 0.65 | 0.1111 | 9 | 1.9954 |
| GAF | 1F5M | 81 | 0.65 | 0.4375 | 16 | 5.3395 | 0.7 | 0.4615 | 13 | 5.6329 |
| Hexokinase_1 | 1DGK | 206 | 0.95 | 0.375 | 8 | 16.0294 | 0.85 | 0.625 | 8 | 26.7157 |
| Asparaginase_2 | 1APY | 161 | 0.7 | 0.2222 | 9 | 6.1927 | 0.65 | 0.0909 | 11 | 2.5334 |
| Inositol_P | 1DK4 | 252 | 0.6 | 0.4615 | 13 | 13.748 | 0.45 | 0.2143 | 28 | 6.383 |
| SCP | 1QNX | 144 | 0.75 | 0.4286 | 7 | 7.7109 | 0.7 | 0.4286 | 7 | 7.7109 |
| IGPS | 1JUL | 245 | 0.8 | 0.3824 | 34 | 10.1287 | 0.8 | 0.375 | 8 | 9.934 |
| PAN | 2HGF | 91 | 0.35 | 0.625 | 8 | 5.9128 | 0.35 | 0.375 | 8 | 3.5477 |
| GMC_oxred_N | 1GAL | 304 | 0.85 | 0.25 | 16 | 11.7081 | 0.85 | 0.3 | 10 | 14.0497 |
| Hist_deacetyl | 1C3P | 313 | 0.95 | 0.1875 | 32 | 5.0871 | 0.9 | 0.2121 | 33 | 5.7551 |
| Ribosomal_L9_N | 1DIV | 52 | 0.5 | 0.25 | 8 | 2.7074 | 0.45 | 0.375 | 8 | 4.0611 |
| PTSIIB_sorb | 1BLE | 151 | 0.65 | 0.0476 | 21 | 1.2763 | 0.7 | 0.1429 | 7 | 3.8289 |
| Dala_Dala_lig_N | 1IOW | 93 | 0.9 | 0.3333 | 9 | 5.4444 | 0.55 | 0.1875 | 16 | 3.0625 |
| NAD_binding_1 | 2PIA | 101 | 0.35 | 0.1111 | 18 | 2.0461 | 0.4 | 0.1429 | 7 | 2.6307 |
| CoA_binding | 2SCU | 112 | 0.7 | 0.3333 | 18 | 6.9061 | 0.55 | 0.069 | 29 | 1.4289 |
| VWA | 1LFA | 172 | 0.6 | 0.2308 | 13 | 3.8024 | 0.45 | 0.0625 | 48 | 1.0298 |
| CheR | 1AF7 | 192 | 0.6 | 0.2 | 10 | 5.8029 | 0.45 | 0.0645 | 31 | 1.8719 |
| OTCace | 1A1S | 158 | 0.85 | 0.2857 | 7 | 7.752 | 0.75 | 0.2857 | 7 | 7.752 |
| Xlink | 1TSG | 93 | 0.7 | 0.3 | 10 | 3.7157 | 0.6 | 0.1429 | 21 | 1.7694 |
| Pep_deformylase | 1DEF | 146 | 0.6 | 0.1 | 10 | 1.8331 | 0.65 | 0.4286 | 7 | 7.856 |
| Mur_ligase | 1UAG | 250 | 0.65 | 0.1667 | 18 | 6.8094 | 0.5 | 0.2222 | 9 | 9.0792 |
| Pyrophosphatase | 1YPP | 185 | 0.8 | 0.1579 | 19 | 3.7665 | 0.8 | 0.1667 | 12 | 3.9758 |
| Ldh_1_C | 1MLD | 165 | 0.8 | 0.1 | 10 | 2.8281 | 0.7 | 0.3 | 10 | 8.4842 |
| Y_phosphatase | 1YTN | 238 | 0.65 | 0.2609 | 23 | 7.517 | 0.5 | 0.1538 | 13 | 4.4331 |
| Fibrinogen_C | 1FIB | 241 | 0.85 | 0.3846 | 13 | 10.4548 | 0.75 | 0.2727 | 11 | 7.4134 |
| ECH | 1DUB | 167 | 0.35 | 0.3529 | 17 | 9.1087 | 0.3 | 0.2222 | 9 | 5.7351 |
| IU_nuc_hydro | 1MAS | 310 | 0.7 | 0.2857 | 7 | 13.0006 | 0.65 | 0.25 | 8 | 11.3755 |

[a]The PFAM name.
[b]The PDB ID name of the target protein, used for the predicting ability of CMA.
[c]The size of the target protein sequence, aligned in the MSA.
[d]The correlation coefficient cutoff where improvement over random is maximum and additionally the number of predictions is minimum seven pairs.
[e]Accuracy.
[f]Number of predictions.
[g]Improvement over random accuracy.

the random accuracy of the particular family. Random accuracy could vary from low values of only 1.5% to high values of 15%, and it is solely dependent upon the particular fold of the target protein. Larger folds tend to have lower random accuracies. Therefore, even though CMA increases IOR up to approximately four times for these families, the final accuracy obtained is low compared to other smaller folds. Of the 127 protein families, Chorismate_mut, Sod_cu and Crystatin families were the only ones which completely failed to give any good

**TABLE V. Performance of CMA for Unclassified–Irregular Families**

| Family Name[a] | PDB ID[b] | Aligned length[c] | PRIN1 cc cutoff[d] | PRIN1 Acc[e] | PRIN1 N[f] | PRIN1 IOR[g] | PRIN3 cc cutoff | PRIN3 Acc | PRIN3 N | PRIN3 IOR |
|---|---|---|---|---|---|---|---|---|---|---|
| ketoacyl-synt | 1DD8 | 246 | 0.55 | 0.0833 | 36 | 2.9783 | 0.4 | 0.0299 | 134 | 1.0668 |
| Gln-synt_C | 2GLS | 282 | 0.85 | 0.3333 | 15 | 10.0617 | 0.65 | 0.2667 | 15 | 8.0493 |
| Glu_synthase | 1EA0 | 369 | 0.85 | 0.1594 | 207 | 8.3268 | 0.8 | 0.092 | 87 | 4.8029 |
| Disintegrin | 2ECH | 45 | 0.4 | 0.7647 | 17 | 4.6238 | 0.4 | 0.5 | 10 | 3.0233 |
| COX6C | 2OCC | 73 | 0.95 | 0.0333 | 90 | 4.2157 | 0.7 | 0.005 | 201 | 0.6292 |
| Antifreeze | 1GZI | 58 | 0.55 | 0.1538 | 13 | 1.0631 | 0.65 | 0.125 | 8 | 0.8638 |
| LHC | 1KZU | 41 | 0.65 | 0.0357 | 28 | 1.9107 | 0.5 | 0.0556 | 18 | 2.9722 |
| CDI | 1JSU | 51 | 0.65 | 0.1 | 10 | 2.875 | 0.65 | 0.1667 | 6 | 4.7917 |
| Bac_DNA_binding | 1IHF | 90 | 0.4 | 0.1176 | 34 | 2.8729 | 0.55 | 0.25 | 4 | 6.1048 |

[a]The PFAM name.
[b]The PDB ID name of the target protein, used for the predicting ability of CMA.
[c]The size of the target protein sequence, aligned in the MSA.
[d]The correlation coefficient cutoff where improvement over random is maximum and additionally the number of predictions is minimum seven pairs.
[e]Accuracy.
[f]Number of predictions.
[g]Improvement over random accuracy.

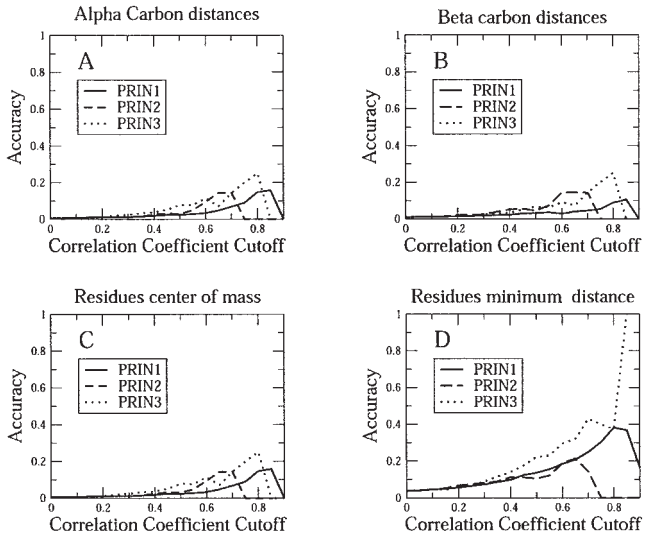


Fig. 3. Accuracy results for IGPS family. The Cα distances illustrated in A, Cβ distances in B and center of mass distances in C, give a weak prediction signal, even though accuracy increases with an increase of cc cutoff. The minimum distance of residue pairs however, shown in D, gives a strong prediction signal and its accuracy for a cc cutoff above 0.6 surpasses the other ones for all three descriptors.

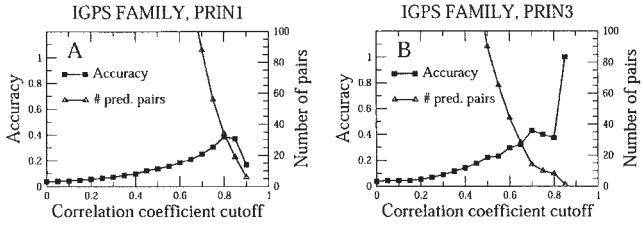


Fig. 4. Illustration of the decrease of predicted proximal pairs with the increase of cc cutoff for IGPS family. For predictors PRIN1, shown in A and PRIN3 shown in B, we observe an increase in accuracy, with a simultaneous decrease in the number of predicted contacts. For a cc cutoff of 0.8 where the accuracy is high for all demonstrated families, the use of PRIN1 gives only 34 contacts. PRIN3 predicts even less pairs than PRIN1, even though the obtained accuracy reaches similar values with those of PRIN1.

prediction with PRIN1 or PRIN3. The majority of the families give good results with both descriptors, and finally few families tend to give poor predictions with one descriptor but good predictions with the other one. Those cases need further study, because the mutational behavior could imply functional or structural importance.

The alpha-beta class of proteins gave the best predictions for both predictors with an average accuracy of 26.8% for PRIN1 and a slightly smaller accuracy of 23.6% for PRIN3. Therefore, there is an indication that alpha–beta proteins undergo more covariant mutations than the other classes. The predictions for the mainly alpha class are relatively poor, with an average accuracy of around 15% for both predictors, while mainly beta class gives better accuracy of around 20%. Also, the Irregular–unassigned proteins gave good predictions of around 20% for PRIN1 and 16.5% for PRIN3, even though the actual number of families used prohibits us from deriving valuable conclusions.

All average values and standard deviations are summarized in Table VI. From this table we also observe that all predictions had maximum IOR when a cc cutoff of around 0.65 is used for PRIN1 and around 0.61 for PRIN3. Therefore we conclude that pairs of residues with a correlation coefficient larger than 0.60, show a strong covariant mutation signal, and a cc cutoff not smaller than 0.60 should be used for CMA predictions. Larger cc cutoff would give higher accuracies than those presented in Table VI, but the number of predictions would also be much smaller.

Decreasing the necessary minimum number of predicted pairs from seven to three (thus increasing the cc cutoff), greatly increases accuracy and IOR. For the alpha–beta class families studied, this decrease resulted in a considerable increase of average accuracy to 38.76% and average IOR to 8.7, for an average cc cutoff

**TABLE VI. Average Performance of CMA for the Four Different CATH Classes of Proteins**

| Class | Accuracy | IOR | N-predictions | CC Cutoff |
|---|---|---|---|---|
| PRIN1 | | | | |
| Mainly Alpha | 0.144 (SD0.097) | 3.86 (SD2.63) | 32.8 (SD60.39) | 0.66 (SD0.20) |
| Mainly Beta | 0.206 (SD0.118) | 3.62 (SD2.42) | 37.4 (SD66.34) | 0.68 (SD0.21) |
| Alpha Beta | 0.268 (SD0.153) | 6.41 (SD4.23) | 19.84 (SD25.56) | 0.67 (SD0.16) |
| Irregular–Unclassified | 0.197 (SD0.230) | 4.32 (SD2.98) | 50 (SD63.70) | 0.65 (SD0.20) |
| PRIN3 | | | | |
| Mainly Alpha | 0.150 (SD0.174) | 3.43 (SD2.71) | 32.12 (SD37.38) | 0.60 (SD0.19) |
| Mainly Beta | 0.208 (SD0.128) | 3.67 (SD2.88) | 30.90 (SD49.69) | 0.62 (SD0.23) |
| Alpha Beta | 0.236 (SD0.147) | 6.06 (SD4.92) | 17.28 (SD18.82) | 0.61 (SD0.17) |
| Irregular–Unclassified | 0.165 (SD0.136) | 3.59 (SD2.56) | 53.66 (SD71.38) | 0.65 (SD0.20) |

of 0.72. The drawback is that CMA predicts lower number of predicted proximal pairs, with an average of 16.59.

### Validation Study

We randomly selected 100 protein families out of the investigated set. We calculated the average cc cutoff for maximum accuracy for all three descriptors of 0.85 for PRIN1 and PRIN3 and 0.8 for PRIN2. Maximum average accuracy for PRIN1 was 0.24, for PRIN2 0.24, and for PRIN3 0.32. Using the rest (27 families) as a separate testing set, we calculated an average accuracy of 0.22 for PRIN1, 0.34 for PRIN2, and 0.29 for PRIN3, at the same cc cutoffs found from the training set. The 27 testing set consisted of the following families: 2-oxoacid_dh, ATP-synt_C, cox3, cu_amine_oxid, cyt-b5, dioxygenase_C, ech, fibrinogen_c, gln-synt, glyko_hydro_8, hcv_ns3, hist-_deacetyl, hom_end_hint, hormone_rec_c, hth_8, irf, keto-acyl_synt, kunitz_legume, mur_ligase, pdfg, ppv_e2_c, pro-isomerase, pts_EIIA_1, pyrophosphatase, ribonuc_red_sm, ribosomal_L14, y_phosphatase. We observe that PRIN1 and PRIN3 had average accuracies very close to those of the training set, while for PRIN2 a much higher average accuracy has obtained.

We conclude that the predictions of PRIN1 and PRIN3 are more reliable than PRIN2, and further studies should be done to explain the predictive behavior of PRIN2.

### Performance of the PCA Descriptors

Figures 5 to 7 show the average accuracy obtained for different cc cutoffs for sets of proteins of different class, with the first three principal components being used as property descriptors. It is clearly shown that all three descriptors show similar behavior in terms of its predictive ability. The average accuracy increases with the increase of the cc cutoff, and all predictors reach high accuracy above the value of 0.20, for all protein classes. However, many key differences of their performance should be noted.

The important conclusion from these figures is that PRIN1, which as stated earlier is related to hydrophobicity, provides the most reliable results. This is due to the fact that PRIN1 gives in general the largest number of predictions compared to PRIN2 and PRIN3, with

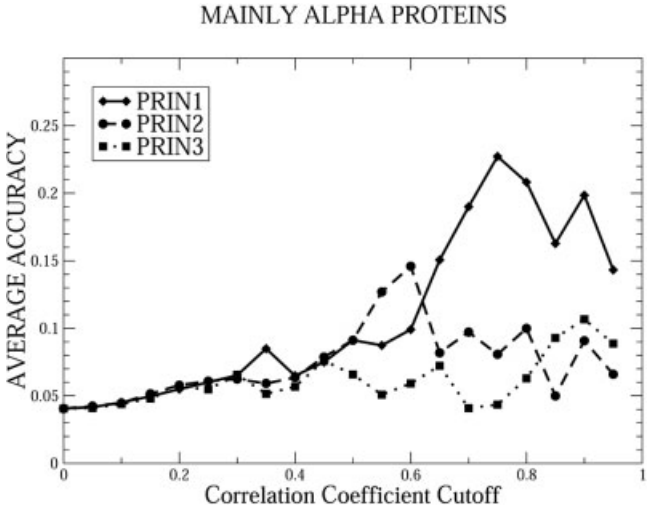


Fig. 5. Average accuracy for all sets of mainly alpha proteins. The first predictor Prin1 shows better predictions that the other two, for cc cut of higher than 0.7.

equally high accuracy compared to the other two descriptors. The average number of predictions for a cc_cutoff of 0.8 is 102.96 for PRIN1, 90.57 for PRIN2 and 74.54 for PRIN3. From Figure 5 we also observe that for the set of mainly alpha proteins PRIN2 and PRIN3 provide very poor predictions compared to PRIN1. This could be an indication that hydrophobicity is very important when correlated mutations occur between helices. Predictions from PRIN2 are also important for alpha–beta and mainly beta proteins. PRIN3 provides a much lower number of predictions although highly accurate for large cc cutoffs. It has also been observed (results not shown) that a substantial number of true-positive predictions from PRIN2 tend to be the same with PRIN1 for a large number of tested proteins. This is not the case for PRIN3, which gives generally a low number of predictions, but different from those of PRIN1 and PRIN2. The reason behind the similarity of the predictions between PRIN1 and PRIN2 is that even though these descriptors are correlated to hydrophobicity and size respectively, they may also share other physical and chemical properties, from the original 142 property set that was used to create them.

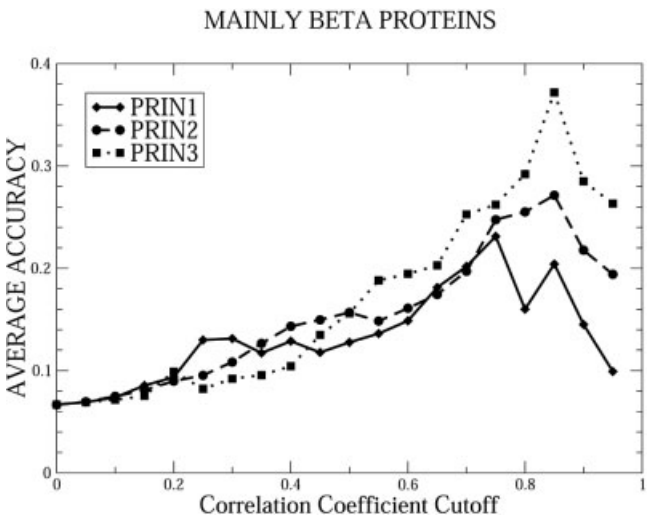


Fig. 6. Average accuracy for all sets of mainly beta proteins. The third predictor Prin3 is always more accurate than the other two predictors. We must not forget though that Prin3 is the one that gives the lowest number of predictions. Thus, results from Prin2 and Prin3 are always important.

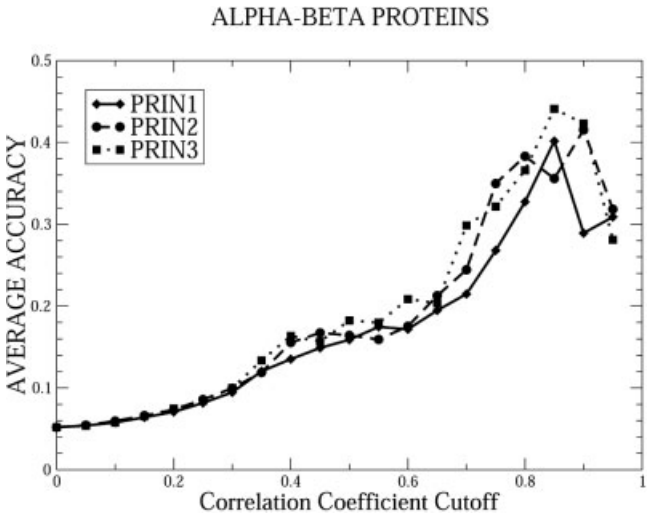


Fig. 7. Average accuracy for all sets of alpha–beta proteins. All three predictors show similar predictive ability. However, Prin1 gives always the largest number of predictions.

In general, the use of CMA with all three principal components leads to highly accurate predictions when a high cc cutoff is used. However, it must be noted that not all proteins tested show the same prediction behavior with any of those predictors. First, not all proteins tested give high values of correlation coefficients. For PRIN1, 40 protein families out of the total 127 tested, did not give cc values larger than 0.8, and 64 families, nearly half of the total set, did not have cc values larger than 0.9. The conclusion is that when CMA gives high cc values of above 0.7 for a protein family, the predictions will have a high probability to be proximal. However, we should not exclude the possibility that proteins having low cc results may also provide highly accurate predictions. That is the main reason that Tables II–V provide a better description of CMA results for PRIN1 and PRIN3, where families with low value cc's are included, rather than Figures 5 to 7.

### Performance of CMA for Different Multiple Sequence Alignments

When PFAM families undergo filtering, a considerable amount of protein sequences are omitted from the calculation. The question that arises is whether omitted sequences influence the original PFAM MSA in such a way that the accuracy of CMA is affected. To examine this case, six families from each of the first three CATH classes have been chosen, three of them showing very good performance with CMA. After filtering, the remaining sequences of those total 18 families were realigned with the use of HMMER.[25] The seeds for the original PFAM alignment were used to create the HMM model necessary for the creation of the new alignments. Then CMA was performed again for the new sets of MSAs.

The protein families used for this examination were the following: ACBP, Endotoxin_N, ribonuc_red_sm, Glyco_hydro_8, Lyase_1, Chorismate_mut, RHD, PH, Pro_isomerase, Dioxygenase_C, Sod_cu, Ribosomal_S8, QRPTase_c, Thionine, Crystatin, Enolase_c, IGPS, CoA_binding. The results for CMA showed that accuracy was virtually unchanged for all tested families (results not shown).

Additionally, CLUSTALW[26] has been used to realign the sequences remaining after filtering. CMA was again performed on the new alignments. The accuracy obtained from these sets of CLUSTALW derived MSAs was considerable lower than the ones obtained with PFAM alignments (results not shown).

We conclude that PFAM MSAs could be directly used for CMA, without realigning the filtered sequences with other methods.

### Evaluation of Contact Predictions—Delta Analysis

Following the example of Ortiz et al,[27] we employed delta analysis on the same set of 18 protein families we used to validate different multiple sequence alignments. The families were chosen from all three major CATH classifications. The set contains families like IGPS and PH, where CMA gives very accurate predictions, as well as families such as Sod_cu and Chorismate_mut, whose CMA predictions were very poor. With delta ($\delta$), we are referring to the vicinity of residues of a position i or j in the protein sequence we investigate, in order to find out whether an originally false pair prediction does include a true positive residue contact.

For example, consider residue positions, 10 and 40, which are predicted, to be proximal but in reality there are not (false positives). For $\delta = 1$, the residues $10 - 1$, 10 and $10 + 1$ are investigated with residues $40 - 1$, 40 and $40 + 1$ to find out whether there is a proximal contact among them. If there is one, then we say that the prediction is a success for $\delta = 1$.

We have used delta analysis for $\delta$ values varying from 1 to 5. When $\delta = 0$, the accuracy of the predictions are found, while when $\delta > 0$, the precision of the prediction is measured.[27] The results are shown in Table VII. The correlation

**TABLE VII. Delta Analysis for a Set of Protein Families, for Predictor PRIN1**

| Protein families | Cc cutoff | Accuracy Delta = 0 | Accuracy Delta = 1 | Accuracy Delta = 2 | Accuracy Delta = 3 | Accuracy Delta = 4 | Accuracy Delta = 5 |
|---|---|---|---|---|---|---|---|
| Fixed cc cutoff of 0.6 | | | | | | | |
| ACBP | | 0.33 | 0.33 | 0.33 | 0.33 | 0.66 | 1 |
| Endotoxin_N | | 0.05 | 0.17 | 0.21 | 0.33 | 0.5 | 0.52 |
| ribonuc_red_sm | | 0.04 | 0.12 | 0.19 | 0.28 | 0.34 | 0.39 |
| Glyco_hydro_8 | | 0.02 | 0.11 | 0.16 | 0.21 | 0.27 | 0.44 |
| Lyase_1 | | 0.03 | 0.09 | 0.17 | 0.22 | 0.25 | 0.30 |
| Chorismate_mut | | 0 | 0 | 0 | 1 | 1 | 1 |
| RHD | | 0.08 | 0.18 | 0.29 | 0.38 | 0.50 | 0.59 |
| PH | | 0.57 | 0.71 | 0.85 | 1 | 1 | 1 |
| Pro_isomerase | | 0.18 | 0.31 | 0.45 | 0.5 | 0.68 | 0.68 |
| Dioxygenase_C | | 0 | 0 | 0 | 0 | 0.5 | 1 |
| Sod_cu | | 0 | 0 | 0 | 0 | 0.33 | 0.33 |
| Ribosomal_S8 | | 0.16 | 0.33 | 0.5 | 0.58 | 0.66 | 0.75 |
| QRPTase_c | | 0.02 | 0.04 | 0.11 | 0.22 | 0.37 | 0.46 |
| Thionine | | 0.11 | 0.22 | 0.77 | 0.77 | 0.88 | 0.88 |
| Crystatin | | 0 | 1 | 1 | 1 | 1 | 1 |
| Enolase_c | | 0.24 | 0.48 | 0.6 | 0.73 | 0.81 | 0.87 |
| IGPS | | 0.18 | 0.36 | 0.51 | 0.63 | 0.68 | 0.71 |
| CoA_binding | | 0.19 | 0.35 | 0.51 | 0.60 | 0.75 | 0.75 |
| Cutoffs from Tables II to V | | | | | | | |
| ACBP | 0.5 | 0.21 | 0.57 | 0.71 | 0.85 | 0.92 | 1 |
| Endotoxin_N | 0.7 | 0.10 | 0.20 | 0.20 | 0.30 | 0.60 | 0.60 |
| ribonuc_red_sm | 0.85 | 0.22 | 0.40 | 0.51 | 0.66 | 0.66 | 0.66 |
| Glyco_hydro_8 | 0.9 | 0.05 | 0.22 | 0.33 | 0.33 | 0.38 | 0.44 |
| Lyase_1 | 0.8 | 0.15 | 0.38 | 0.53 | 0.61 | 0.61 | 0.61 |
| Chorismate_mut | 0.5 | 0 | 0 | 0.27 | 0.63 | 0.72 | 0.81 |
| RHD | 0.95 | 0.28 | 0.57 | 0.57 | 0.71 | 0.71 | 0.85 |
| PH | 0.5 | 0.54 | 0.63 | 0.81 | 0.90 | 0.90 | 0.90 |
| Pro_isomerase | 0.65 | 0.23 | 0.38 | 0.46 | 0.46 | 0.69 | 0.69 |
| Dioxygenase_C | 0.4 | 0.1 | 0.2 | 0.3 | 0.5 | 0.6 | 0.6 |
| Sod_cu | 0.55 | 0 | 0 | 0.14 | 0.14 | 0.28 | 0.42 |
| Ribosomal_S8 | 0.6 | 0.16 | 0.33 | 0.5 | 0.58 | 0.66 | 0.75 |
| QRPTase_c | 0.65 | 0.02 | 0.03 | 0.1 | 0.21 | 0.35 | 0.46 |
| Thionine | 0.5 | 0.21 | 0.28 | 0.71 | 0.78 | 0.85 | 0.92 |
| Crystatin | 0.45 | 0.05 | 0.25 | 0.3 | 0.45 | 0.7 | 0.95 |
| Enolase_c | 0.8 | 0.52 | 0.76 | 0.80 | 1 | 1 | 1 |
| IGPS | 0.8 | 0.38 | 0.55 | 0.58 | 0.73 | 0.76 | 0.79 |
| CoA_binding | 0.7 | 0.33 | 0.5 | 0.61 | 0.72 | 0.83 | 0.83 |

coefficient cutoffs used for those families is a fixed value of 0.6 for the first part of the table, and the values described in Tables II–V for the second part. For $\delta = 0$ the accuracy of CMA for the given cutoff is shown. From this table, we observe that with $\delta = 1$, there is an important increase of contact predictions. Predictions which are two or three times more accurate than those obtained at $\delta = 0$, such as in ACBP, RHD, Crystatin, were observed. Only Chorismate_mut, Dioxygenase, and Sod_cu fail to provide good results for $\delta = 1$, though they provide good results for $\delta \geq 3$. Qrptase_c also shows an increased ability for predictions with $\delta = 2$ and above. It is within our opinion that values of $\delta$ above 3 are not very important, simply because the vicinity of residues becomes large and that random probability of finding a proximal pair increases considerably. Hence, there is no meaning for searching residue contacts with $\delta$ larger than 2.

It should be mentioned that choosing a certain cc cutoff to describe the predictive ability of a certain family is not always adequate. As described before, not all protein families give high values of correlation coefficients. Therefore, it should be more appropriate to choose the predicted pairs with the highest values of cc's, rather than using a fixed cc cutoff. For the families shown in the second part of Table VII, if we use the cc cutoffs described in Tables II–V, we get superbly better results that those shown in the first part. Lyase_1 for example has an enrichment of 38% for $\delta = 1$ when a cc cutoff of 0.8 is used, rather than the relative poor value of 9% for cc cutoff of 0.6.

Table VII indicates that in many cases, even though CMA predictions are not very accurate, there is still a great chance that a proximal residue pair within $\pm$ 2 residues exists. These results could be very important for constrained protein folding simulations, because they provide us with an indication that a small sifting of the

**TABLE VIII. The Number of Total True Positive and False Positive Pair Predictions From PRIN1 and PRIN3 Before and After the Filtering[†]**

| | Original distribution[a] | | | Experimental s.a.a filtering[b] | | | Predicted s.a.a filtering[c] | | |
|---|---|---|---|---|---|---|---|---|---|
| Families | TP | FP | acc | TP | FP | acc | TP | FP | Acc |
| IL1 | 30 | 137 | 0.18 | 18 | 33 | 0.35 | 15 | 32 | 0.32 |
| PPV_E2_C | 15 | 143 | 0.09 | 11 | 32 | 0.25 | 5 | 71 | 0.06 |
| GAF | 32 | 171 | 0.16 | 16 | 56 | 0.22 | 7 | 50 | 0.12 |
| Gln-synt_C | 22 | 108 | 0.17 | 13 | 66 | 0.16 | 2 | 10 | 0.17 |
| Glu-syntase | 17 | 87 | 0.16 | 17 | 72 | 0.19 | 0 | 12 | 0 |
| Gmc-Oxyred_N | 25 | 114 | 0.18 | 22 | 86 | 0.20 | 10 | 20 | 0.33 |
| Hexokinase_1 | 17 | 133 | 0.11 | 10 | 71 | 0.12 | 8 | 60 | 0.12 |
| Hist-deacetyl | 19 | 106 | 0.15 | 19 | 71 | 0.21 | 4 | 13 | 0.23 |
| Inositol-P | 32 | 104 | 0.23 | 26 | 66 | 0.28 | 14 | 28 | 0.33 |
| Jakalin | 13 | 157 | 0.08 | 1 | 38 | 0.02 | 2 | 39 | 0.05 |
| Ketoacyl synt | 9 | 127 | 0.07 | 9 | 68 | 0.12 | 5 | 45 | 0.10 |
| Kunitz-legume | 27 | 96 | 0.22 | 20 | 41 | 0.33 | 18 | 37 | 0.33 |
| SCP | 28 | 125 | 0.18 | 23 | 85 | 0.21 | 16 | 61 | 0.20 |
| IGPS | 27 | 89 | 0.23 | 25 | 77 | 0.24 | 10 | 39 | 0.20 |
| Gal-bin-lectin | 21 | 103 | 0.17 | 8 | 29 | 0.22 | 6 | 32 | 0.16 |
| Glyco-hydro_11 | 18 | 140 | 0.11 | 11 | 44 | 0.20 | 3 | 30 | 0.09 |
| Pro-isomerase | 31 | 97 | 0.24 | 19 | 42 | 0.31 | 11 | 36 | 0.23 |

[†]TP are the true positive pairs, FP are the false positive pairs, and acc is the accuracy of the predictions.
[a]The original distribution of predictions for protein families, predicted by CMA, where the predictions have been taken by summing up the results from PRIN1 and PRIN3, excluding the very few common predictions.
[b]The distribution of predictions for the same protein families, after the use of the experimentally derived surface accessible area as a filter.
[c]The distribution of predictions for the same protein families, after the use of the surface accessible area predicted by Ahmad et al.[30]

predicted pairs could lead us to discover covered truly proximal residue pairs.

### Combining Results From CMA with Information of Surface Accessible Area of Target Protein

From the tests conducted of the set of PFAM families, and from the analysis of prediction data, it has been found that the majority of the true positive predictions of CMA lie in the buried region of the target protein. This observation has been quantified with the use of the surface accessible area (s.a.a.) of the target protein residues. The definition of Lee and Richards[28] has been used, and the Fortran code from Sali et al,[29] has been employed, in order to calculate the actual surface accessible area for the residues of the target proteins. Using the s.a.a data for the true positive and the false positive predicted residues, it has been found that a significant amount of false positive residues have s.a.a percentage value more than 20%, while the majority of the true positive pairs lie between 0–20%.

This observation can provide us with the opportunity to use CMA predictions with less than 0.65 values of correlation coefficient. From CMA, a large number of predictions are obtained for cc cutoff around 0.5 where pairs are considered to undergo lower degree covariant mutation, thus being less accurate. Filtering of residues whose s.a.a value exceeds 20% will significantly decrease the number of false positives, while at the same time the number of true positives will be virtually unchanged. The filtering of the target proteins of the families tested is illustrated in Table VIII. Using the predictions of both PRIN1 and PRIN3, a cc cutoff threshold has been chosen such that the first 100 pairs with the highest correlation coefficient from PRIN1 could be obtained. The same cc cutoff has been used to obtain all predictions above this value for PRIN3, which are less than those of PRIN1.

From Table VIII, it is clearly seen from the middle three columns that filtering the CMA predictions with calculated s.a.a information for lower cc cutoff thresholds, significantly decreases the number of false positives, with only but a small loss of true positives. Thus, the filtering is successful. Having the knowledge that filtering of the false positives can increase both the accuracy and the number of predicted pairs, we can combine CMA with surface-accessible area prediction methods. A method developed by Ahmad et al[30] has been used, and predictions of proximal pairs have been filtered using s.a.a prediction information. Unfortunately, this s.a.a prediction method under-predicts the true positives obtained from CMA, and does not provide good results. The filtering with the use of s.a.a prediction succeeds in clearing a fairly large portion of the false-positive pair predictions. At the same time however, it excludes a significant amount of the true positive predictions. We observe that the accuracy for most of the families after the filtering using predicted s.a.a data (last three columns of Table VIII) is smaller than the one obtained after the filtering using experimental s.a.a data. Still, in many cases, the accuracy obtained with predicted s.a.a. filtering, though it does not reach the values of the ideal s.a.a filtering is better than the initial one obtained from CMA (first three columns of Table VIII). The results obtained by coupling CMA predictions with s.a.a prediction look promising, since there is currently considerable development in s.a.a prediction methods.

## CONCLUSIONS

We have developed an algorithm to predict amino acid residues distant in the protein's primary structure, but proximal in its tertiary structure. Correlated mutation analysis (CMA) has been used to extract necessary evolutionary information from multiple sequence alignments taken from PFAM. Amino acid residue descriptors have been developed from experimental data obtained from AAindex database. Correlation coefficients have been calculated for all positions in the MSA. One hundred and twenty-seven (127) families representing all major classes and architectures from CATH have been arbitrarily chosen for CMA evaluation. Previously reported CMA accuracy[7,16] was around 15%, while in our method the vast majority of the tested families CMA perform well, with an average improvement over random prediction reaching the value of 4, and accuracy exceeding 20% for the mainly beta and alpha beta families. Using stricter thresholds than the current one of around 0.65, CMA can be up to 40% accurate, but with a very low number of predicted pairs. CMA can be performed for multiple sequence alignments, containing 25 or more sequences, with evolutionary distance from the target protein not more than 90. Still, CMA performance is highly dependent upon the quality of the multiple sequence alignment.

PRIN1, which has strong correlation to hydrophobicity, is the predictor that gives the most reliable results. PRIN3 also provides very accurate results, but the number of predicted contacts it provides is much smaller compared to PRIN1. Finally PRIN2, which correlates well with size, provides equally accurate results compared to the other two predictors and it should be used in conjunction to PRIN1, since many of its true positive predictions are already included in those of PRIN1.

Delta ($\delta$) analysis was performed in a set of protein families from all three major CATH classifications. It has shown that with a very close proximity of the residue positions that CMA predicts to be proximal, there is a very high probability that there exists a pair set of residues that is truly proximal. Therefore CMA predictions could be directly used for protein folding simulations.

The fact that true positive pair predictions lie in the buried region of the protein can be exploited with the coupling of CMA predictions and surface-accessible-area predictions. This is only but one out of many other methods that CMA can be coupled so that its results can be improved. The results from CMA-s.a.a coupling are not impressive, but they can show great potential in the future, since the accuracy of the s.a.a. method continues to increase.

The CMA algorithm described in this work gives good results for most of families taken from PFAM. The predicted proximal residue pairs can be used for simulations and protein-fold recognition methods. That gives us the initiative to create an automated version of the same algorithm, and apply it to other protein families, such as those stored in HSSP. CMA can be periodically evaluated and improved every time the used protein families database is updated. The direct relation between protein fold and CMA accuracy can be deciphered.

In future work, we expect that by creating an automated version of our algorithm, which is periodically applied into PFAM and other protein families databases, we can test more families which currently do not contain sufficient number of nonredundant sequences, and reevaluate those families already being tested with previous database versions. Comparative studies between families derived from different databases can also be performed and valuable conclusions could be drawn on the performance of CMA with different MSAs.

## ACKNOWLEDGMENTS

This work has been supported by the University of Minnesota Biotechnology Institute and the American Chemical Society Petroleum Research Grant (Award no. G7-38758). We also want to thank Dr. Eric Fauman for his help and insight.

## REFERENCES

1. Niggemann M, Steipe B. Exploring local and non-local interactions for protein stability by structural motif engineering. J Mol Biol 2000;296:181–195.
2. Abkevich VI, Gutin AM, Shakhnovich EI. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. J Mol Biol 1995;252:460–471.
3. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14:835–843.
4. Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
5. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15–21.
6. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins 2002;Suppl. 5:157–162.
7. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 1997;2:S25–32.
8. Oliveira L, Paiva ACM, Vriend G. Correlated mutation analyses on very large sequence families. Chem Bio Chem 2002;3:1010–1017.
9. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511–523.
10. Ortiz AR, Kolinski A, Skolnick J. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. Proc Natl Acad Sci USA 1998;95:1020–1025.
11. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. J Mol Biol 1998;277:419–448.
12. Ortiz AR, Kolinski A, Skolnick J. Tertiary structure prediction of the KIX domain of CBP using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. Proteins 1998;30:287–294.
13. Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, Nardone F, Stoehr P, Tuli MA, Tzouvara K, Vaughan R. The EMBL Nucleotide Sequence Database: major new developments. Nucleic Acids Res 2003;31:17–22.
14. Taylor WR, Hatrick K. Compensating changes in protein multiple sequence alignments. Protein Eng 1994;7:341–348.
15. Neher E. How frequent are correlated changes in families of protein sequences. Proc Natl Acad Sci USA 1994;91:98–102.
16. Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Eng 2002;15:721–725.
17. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein

Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
18. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
19. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
20. Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. Nucleic Acids Res 1999;27:368–369.
21. Bishop CM. Neural networks for pattern recognition. New York: Oxford University Press; 1995.
22. JMP. Version 4.0.4. Cary, NC: SAS Institute Inc.; 2001.
23. Pazos F, Olmea O, Valencia A. A graphical interface for correlated mutations and other protein structure prediction methods. Comput Appl Biosci 1997;13:319–321.
24. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science 1992;256:1443–1445.
25. HMMER. 2.3.2. St. Louis: Washington University, St. Louis; 2003.
26. CLUSTAL W. 1.83. Heidelberg: EMBL; 1996.
27. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 1999;Suppl 3:177–185.
28. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
29. Sali A, Blundell TL. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J Mol Biol 1990;212:403–428.
30. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629–635.