

Effective Use of Sequence Correlation and Conservation in Fold Recognition

Osvaldo Olmea¹, Burkhard Rost² and Alfonso Valencia^{1*}

¹Protein Design Group, CNB-CSIC, Cantoblanco, Madrid E-28049, Spain

²CUBIC Columbia University Department of Biochemistry & Molecular Biophysics, 630 West 168th Street, New York, NY 10032, USA

Protein families are a rich source of information; sequence conservation and sequence correlation are two of the main properties that can be derived from the analysis of multiple sequence alignments. Sequence conservation is related to the direct evolutionary pressure to retain the chemical characteristics of some positions in order to maintain a given function. Sequence correlation is attributed to the small sequence adjustments needed to maintain protein stability against constant mutational drift. Here, we showed that sequence conservation and correlation were each frequently informative enough to detect incorrectly folded proteins. Furthermore, combining conservation, correlation, and polarity, we achieved an almost perfect discrimination between native and incorrectly folded proteins. Thus, we made use of this information for threading by evaluating the models suggested by a threading method according to the degree of proximity of the corresponding correlated, conserved, and apolar residues. The results showed that the fold recognition capacity of a given threading approach could be improved almost fourfold by selecting the alignments that score best under the three different sequence-based approaches.

© 1999 Academic Press

*Corresponding author

Keywords: correlation; conservation; contact prediction; incorrectly folded models; threading

Introduction

At the molecular level, the natural process of evolution is reflected in the accumulation of variation between sequences of the same protein in distinct organisms. Zuckerkandl & Pauling (1965) initially proposed that protein families vary at different rates and that sequence divergence accumulates in protein regions related to protein function. Since then, it has generally been accepted that functional and structural constraints on proteins lead to the conservation of the chemical character of amino acid residues in polypeptide chains, as observed in multiple sequence alignments of protein families (Benner, 1989; Benner & Gerloff, 1991; Cooperman *et al.*, 1992; Howell, 1989; Hwang & Fletterick, 1986). However, it is almost impossible to distinguish the structural and functional impacts from one another (Ouzounis *et al.*, 1998).

The information derived from multiple sequence alignments has been used successfully in problems of protein family identification (Bairoch, 1992;

Smith & Smith, 1992), the search for remote homologous sequences (Bork, 1989; Gribskov *et al.*, 1987; Koonin *et al.*, 1996; Thompson *et al.*, 1994) and for the comparison of protein structures (Artymiuk *et al.*, 1993; Bork *et al.*, 1995; Godzik & Sander, 1989; Holm & Sander, 1994, 1995; Holm *et al.*, 1994; Mauzy & Hermodson, 1992; Murzin, 1996; Pastore & Lesk, 1990). Progress in secondary structure and accessibility predictions has been attributed largely to the use of simple sequence profile (Rost & Sander, 1994). Both *ab initio* folding experiments based on sequence constraints (Taylor, 1991) and popular threading approaches use sequence information to some extent (Defay & Cohen, 1996; Fisher & Eisenberg, 1996; Jones *et al.*, 1992; Ouzounis *et al.*, 1993; Rost, 1995; Rost *et al.*, 1997). Sequence information is also commonly used by biologists in a non-systematic way during the analysis of specific protein families. Intuitive rules are followed; for example, conserved histidine or aspartate residues are often replaced by site-directed mutagenesis in the search for active sites.

Other, more sophisticated definitions of sequence conservation can be derived from the analysis of group-specific sequence information,

E-mail address of the corresponding author: valencia@cnb.uam.es

i.e. residues conserved in particular groups of sequences but not in the entire protein family. These conserved residues are probably related to functional adaptations specific to certain sequence groups. Various strategies have been adopted for the analysis of these residues (Andrade *et al.*, 1997; Casari *et al.*, 1995; Lichtarge *et al.*, 1996; Livingstone & Barton, 1993; Pazos *et al.*, 1997c). It is striking that no proper tools have yet been developed to use this information for prediction of protein structure.

Sequence correlation

Information other than conservation can possibly be extracted from multiple sequence alignments, i.e. cases of concerted patterns of variation between different positions in multiple sequence alignments (Altschuh *et al.*, 1987, 1988). Correlated changes are more likely to correspond to compensatory substitutions that occur independently in proteins of the same family, to maintain them within the limits of protein stability. The underlying assumption is that compensation may be favored between residues in physical contact; correlated mutation calculations were therefore proposed to be of use for contact prediction (Göbel *et al.*, 1994; Olmea & Valencia, 1997; Pazos *et al.*, 1997a,b). A variety of methods detecting correlated positions has been proposed (Göbel *et al.*, 1994; Neher, 1994; Shindyalov *et al.*, 1994; Singer *et al.*, 1995; Taylor & Harrick, 1994). Conclusions about the amount of information that can be extracted from the alignments differed substantially. This difference results from different definitions of correlated mutations (see Chelvanayagam *et al.*, 1997; Pollock & Taylor, 1997; Pollock *et al.*, 1999); our definition (Göbel *et al.*, 1994; Olmea & Valencia, 1997; Pazos *et al.*, 1997a,b) favors those cases in which one or a few sequences co-vary over those in which large groups (subfamilies) change in a concerted manner. This seemingly subtle technical point is the key to our approach.

Sequence information as a constraint for structure prediction

The protein folding problem remains unsolved, despite the enormous amount of effort spent on it. One of the approaches pursued by a number of groups has been to fold proteins using simulation techniques under defined force-fields of different natures. In this particular incarnation of the protein folding problem, one of the major difficulties is to obtain sufficiently accurate long-range constraints to guide the simulation toward the correct fold. It seems natural to believe that general properties such as polarity, bulk, compactness, angular or secondary structure preferences may provide globular structures but not definitively folded structures.

Correlation between residue pairs appears to be a more appropriate method for providing this specific constraint.

The current precision of inter-residue contact predictions is still insufficient to be used successfully as long-range constraints in distance geometry-related protocols. An exception may be the case of small proteins, in which considerable success has been reported using correlated mutations as the main source of long-range sequence constraints (Ortiz *et al.*, 1998a,b). Furthermore, we have used conservation and correlation in some cases to provide additional information for predicting protein structure from sequence (Valencia *et al.*, 1995; O.O. *et al.*, unpublished results†).

Here, we review the current status of contact prediction using conservation and correlation information, and assess the feasibility of using conservation and correlation to discriminate incorrectly folded proteins.

A standard protein structure prediction test consists of comparing real proteins with deliberately misfolded proteins. The first example was provided by Novotny and co-workers (Novotny *et al.*, 1984), in which the sequence of a protein of known all-helical three-dimensional structure (hemerythrin) was placed into the known structure of a completely different type and anti-parallel β -barrel (immunoglobulin) and *vice versa*. Holm & Sander (1992) applied a similar test extended to a larger number of protein pairs with the same number of residues but dissimilar structures, which were misfolded by swapping the sequence of each pair. Variations of this test set have been used by other authors (Hendlich *et al.*, 1990; Huang *et al.*, 1995, 1996; Maiorov & Crippen, 1992), among others. In particular in threading experiments, a related idea ("ungapped threading") is commonly used as a first test (e.g. see Ouzounis *et al.*, 1993; Sippl & Weitckus, 1992). We used an even larger set of intentionally misfolded proteins. Unlike previous studies, it is not necessary to refine the models in our case, since we are interested only in the distances rather than the detailed environment of the residues.

We present the results obtained extending the filtering procedure to the analysis of threading models. A specific threading algorithm (Rost, 1995; Rost *et al.*, 1997) was used to generate models. The corresponding structure-based threading scores were compared to the scores obtained with the corresponding information about correlated, conserved and apolar residues.

Results

Relevance of sequence conservation for protein structure

Predicting inter-residue contacts

Contact predictions can be as accurate as 35% for small proteins when invariant residues were

† <http://www.gredos.cnb.uam.es/CASP>

considered, and as low as 3% for large proteins (Figure 1). The most representative results were those corresponding to the protein size category of 103 to 166 amino acid residues. For these, contacts were predicted at 13% accuracy with invariant residues. This was 2.6-fold better than the prediction obtained with variable residues. When the data were analyzed in terms of X_d values, this trend became clearer. Higher values of conservation correlated well with inter-residue contacts. For the 103-166 size category, the X_d value was 5.12, whereas the prediction obtained with variable residues was -0.8 . Furthermore, conserved residues were found to be shifted toward smaller distances such as the 4 to 12 Å bins (Figure 2).

Sequence conservation distinguishes correctly from incorrectly folded models

A total of 94% of the pairs were discriminated with X_d values greater in the real protein than in the incorrect model (Figure 3(a): positive differen-

tial X_d values). For a specific correct-incorrect example model, the distance distributions were compared (Figure 2(a)). Conserved residues were clearly closer than others (Figure 2(a)i: X_d value of 8.27), whereas they are randomly distributed in the model protein (Figure 2(a)ii: X_d value of -2.83) and were essentially indistinguishable from all other residues. The difference of X_d values in this case was 11.1.

Relevance of sequence correlation for protein structure

Predicting inter-residue contacts

Contact prediction *accuracy* for correlated pairs is represented for the same 71 protein families in Figure 1(c). Strongly correlated pairs ($L/2$ set) were found to be better predictors of contact than weakly correlated pairs. The *accuracy* of the predictions clearly depended on protein size, as it is more difficult to predict contacts in large proteins

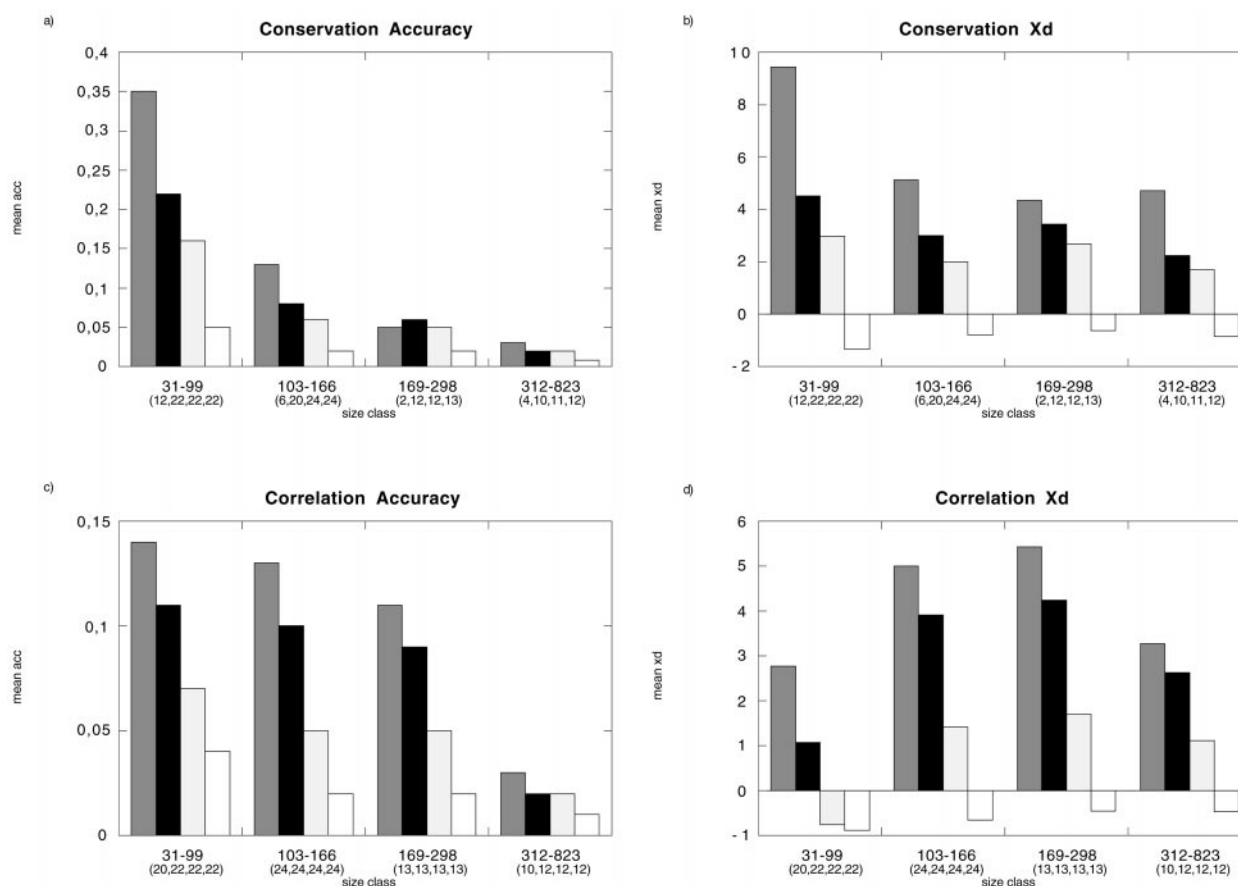


Figure 1. Performance of conservation and correlation as predictors of proximity in protein structures. Bar diagram of the average values of the predictions by (a) and (b) conserved or (c) and (d) correlated residues. The data are presented as grouping proteins into four protein length-categories (31-99, 103-166, 169-298 and 312-823 amino acid residues). The four bars correspond, from left to right, to variability levels of 0, 1-13, 14-18 and more than 21 in (a) and (b), number of correlated pairs equivalent to $L/2$, L , $7L$ and non-correlated residues in (c) and (d), where L is protein length. The number of proteins in each category of protein-size and variability or correlation are given on the y-axis. Predictions are measured by (a) and (c) Accuracy, and (b) and (d) X_d . Data correspond to the first set of 71 non-redundant proteins described in Methods as the first testing set.

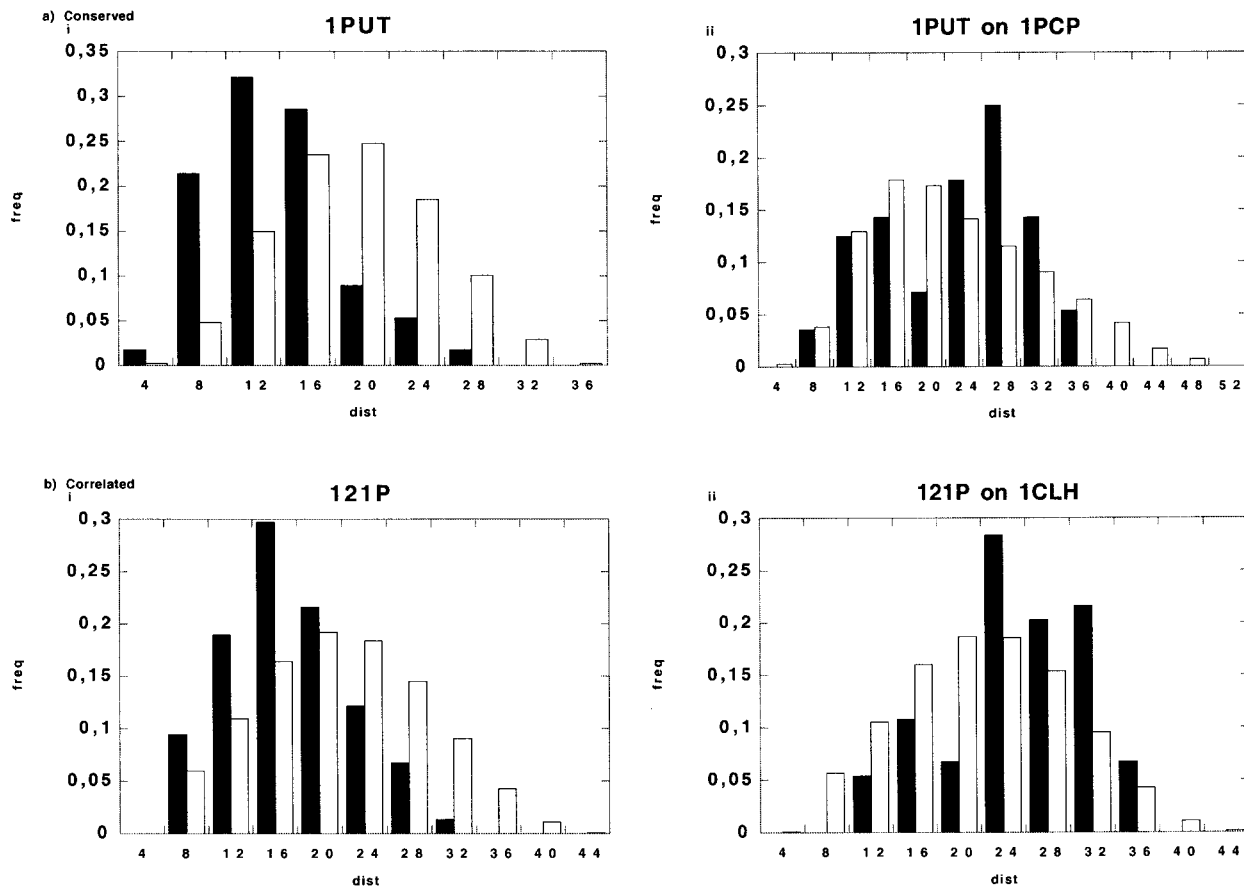


Figure 2. Representative examples of distance distributions in correctly and incorrectly folded proteins. The proportion of residue pairs at different distances are shown grouped in distance bins. Results for (a) conserved residues and (b) correlated residues. The proportion of conserved and correlated residues are represented by filled bars and all other residues with open bars. The same representation is used for two real proteins (redoxin 1PUT in (a)i and ras-p21 protein 121P in (b)i) and their corresponding incorrectly folded models ((a)ii 1PUTon1PCP -growth factor-; (b)ii 121Pon1CLH -cyclophilin). The number of observations was 56 distances between conserved residues in 1PUT *versus* 5050 for all other pairs, and 75 distances between correlated pairs in 121P *versus* 10,333 distances for all other residues.

than in small ones. For the best-represented set of proteins (103-166), *accuracy* was 13%, i.e. 6.5-fold better than contact prediction with all non-correlated pairs. The results evaluated with a measure of proximity (*Xd* value) revealed a more pronounced trend (Figure 1(d)), with clear separation of correlated pairs ($L/2$ and L) from the samples with almost no correlation ($L/7$ or no correlation). For the 103-166 protein size category, the $L/2$ most correlated pairs had an *Xd* value of 5. In comparison, non-correlated pairs, which represented the random prediction values, had an accuracy of 2% and an *Xd* value of -0.7 .

Distinguishing correct and incorrect models

In as many as 91% of the pairs, the incorrect model was effectively discriminated (Figure 3(b): *differential_Xd* values greater than 0). An example of the distance distribution obtained is shown in Figure 2(b). As in the case of conservation, we observed a clear shift of the distances between

pairs of correlated positions toward small distance (Figure 2(b)i). The *Xd* value of this displacement was 4.47, indicating a strong displacement of both populations. For the corresponding incorrectly folded model (Figure 2(b)ii), the distances between correlated residues were not smaller than between other residues (*Xd* value of -4.43), indicating a random distribution of correlated pairs when they were mapped onto the incorrect structure. The correlation information seemed only slightly less able to discriminate incorrect models than the conservation data. This trend became more obvious when a reduced set of proteins was analyzed. The reduced set contained only proteins with less than 25% pairwise sequence similarity (set 3, described in Methods). In this case, the percentages of pairs above the zero value of *Differential_Xd* was 85% and 87% for conservation and correlation, respectively (Figure 4). Although the smaller test set was "cleaner", it was difficult because the number of cases was drastically reduced.

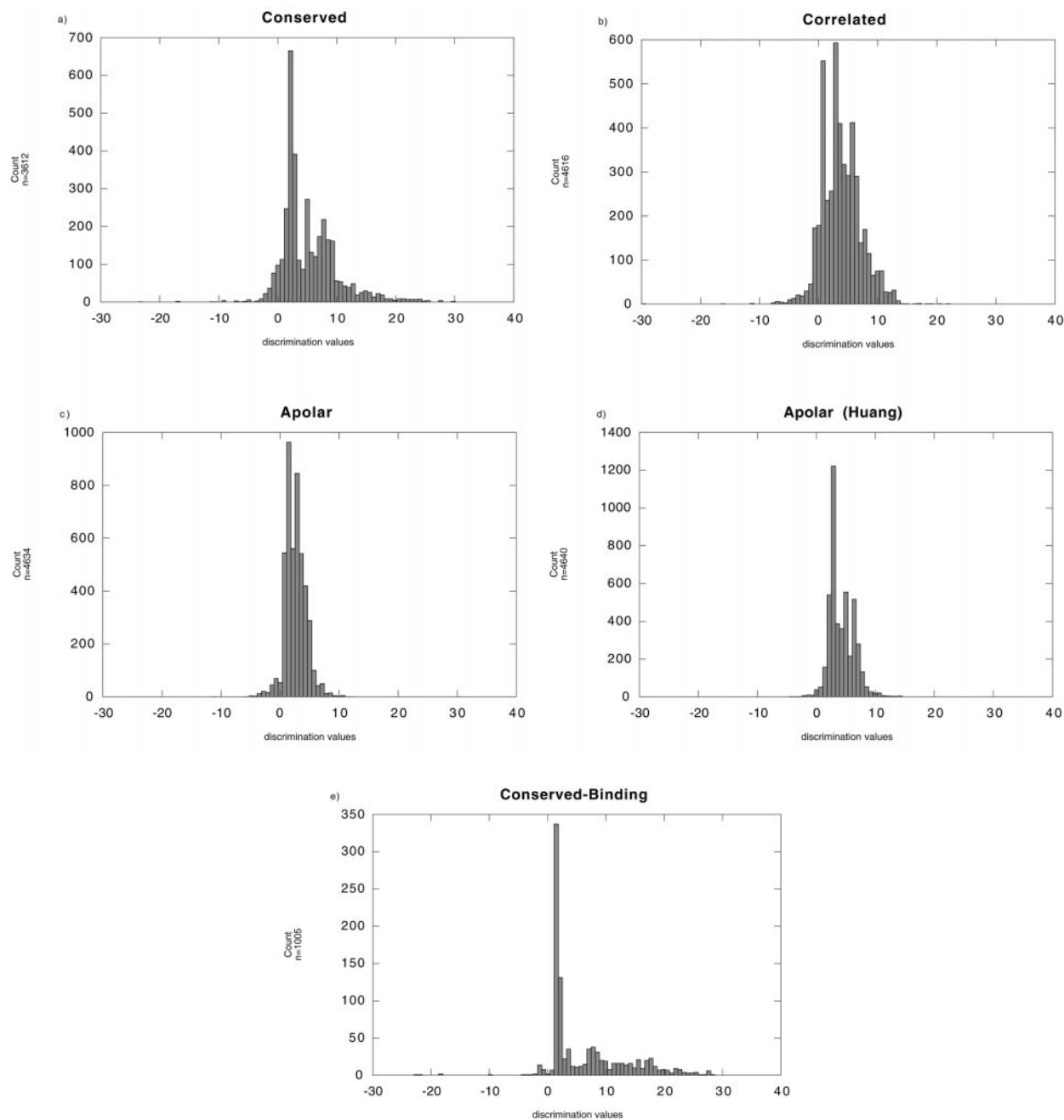


Figure 3. Sequence discrimination between real proteins and their corresponding incorrect models. The number of observations is shown on the y -axis and the $Differential_Xd$ value on the x -axis. (a) Conserved, (b) correlated, (c) apolar, (d) apolar-Huang, and (e) conserved-binding residues. Proteins were selected from the second set described in Methods.

Other types of sequence information: apolar residues as part of the protein core

We also investigated the information contained in two different sets of apolar residues. Figure 3(c) and (d) show the distribution of apolar residues (Asp, Leu, Ile, Val, Met, Trp, Phe, and Tyr) and apolar residues as defined by Huang *et al.* (1995) (Cys, Leu, Ile, Val, Met, Trp, and Phe). The $Differential_Xd$ values for the two sets were similar.

Apolar residues were closer in the real protein than in the incorrect models in 95% of the pairs. The subset apolar-Huang was slightly more discriminating in 98% of the cases. Interestingly, the distributions for $Differential_Xd > 1$ were similar between the sets of apolar (84%), conserved (89%), and correlated pairs (78%), but the set of apolar-Huang yielded a better discrimination, with 97% of the cases reaching values larger than one.

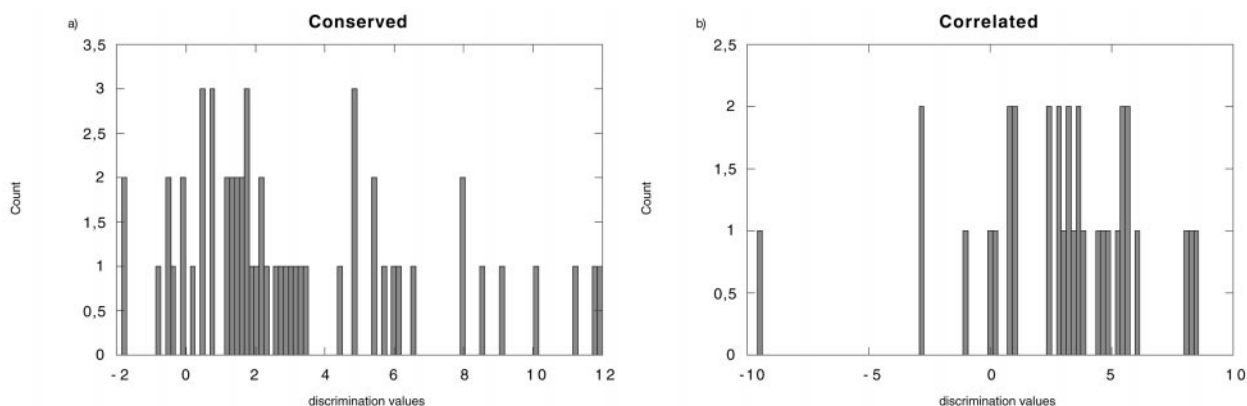


Figure 4. Sequence discrimination between correctly and incorrectly folded models in a non-redundant protein set. A representation has been used for a reduced set of models derived from a list of non-sequence redundant proteins (third set described in Methods) similar to that used for the first test set in Figure 1. The Figure represents Differential_Xd values for (a) completely conserved and (b) correlated residues.

Other types of sequence information: residues with a tendency to form part of protein-binding sites

We have previously studied the tendency of different residue types to form part of protein-binding sites and the relationship to conservation (Ouzounis *et al.*, 1998). By analyzing the frequency with which different types of invariant residues form part of binding sites, four sets of residues could be defined: strong tendency (Lys, Arg, Asp, Cys, and His); moderate tendency (Phe, Tyr, Trp, Gly, and Leu); slight tendency (Ser, Asn, Asp, Thr, Met, Pro, and Gln); and negative tendency (Val, Ile, and Ala). For the subset of invariant residues that included those with strong and medium tendencies to form part of binding sites, we observed a significant discrimination of misfolded models, represented as those with 96% of the pairs correctly separated (Figure 3(e)). This was comparable to the results obtained for other sets of conserved, correlated or apolar residues. The only improvement observed in analyzing these results was the displacement of Differential_Xd toward larger values.

Networks of conserved, correlated or apolar residues

Examples are shown of the spatial distribution of conserved (Figure 5(a)), correlated (Figure 5(b)) and apolar residues (Figure 5(c)) in real proteins and in the corresponding incorrectly folded models. Most of the conserved residues were part of a cluster, with only a single conserved residue located away from this cluster (Figure 5(a)i). In the growth factor example, cluster formation was less clear; actually, at least two clusters were present (Figure 5(a)ii). When this sequence information was translated to a different structure, it tended to render residues localized in the exterior of the protein and distributed randomly (Figure 5(a)iii and (a)iv).

For correlated pairs, all residues implicated in any of the correlated pairs have been represented. In this representation, adopted for simplicity, the relation between pairs of residues cannot be appreciated, although the concentration in the protein interior can be observed (first protein, Figure 5(b)i). In the second protein (Figure 5(b)ii), one of the external alpha-helices is the main element containing correlated pairs. In the incorrectly folded proteins (Figure 5(b)iii and (b)iv), the correlated residues are now clearly distributed mainly on the protein surface, occupying many of the external loops.

The results obtained with a selection of apolar residues (Figure 5(c)) showed the expected concentration in the protein core, accompanied by many exceptions (apolar residues on the surface: Figure 5(c)i and (c)ii). When these residues were represented in the corresponding incorrectly folded models (Figure 5(c)iii and (c)iv), they were randomly distributed and many were localized on the protein surface.

Distinguishing correct and incorrect models

The discrimination of incorrectly folded models through apolar residues (Figure 6(a)) appeared to be inferior to that through conserved residues (92% of the cases *versus* 95%). However, the two sources of sequence information were sufficiently distinct. Thus, both could be improved by selecting only pairs with good discrimination in at least one case. For example, most of the 5% of incorrectly folded models that were not identified with conserved residues could be recognized by the distance between apolar residues and *vice versa*, whereas nearly all the cases that were not discriminated using only apolar residues (8%) could be correctly assessed with a conservation cut-off. Overall, combining conservation and apolarity by using only values larger than zero (a logical “or”) discriminated 99% of the incorrectly folded models (Figure 6(a)).

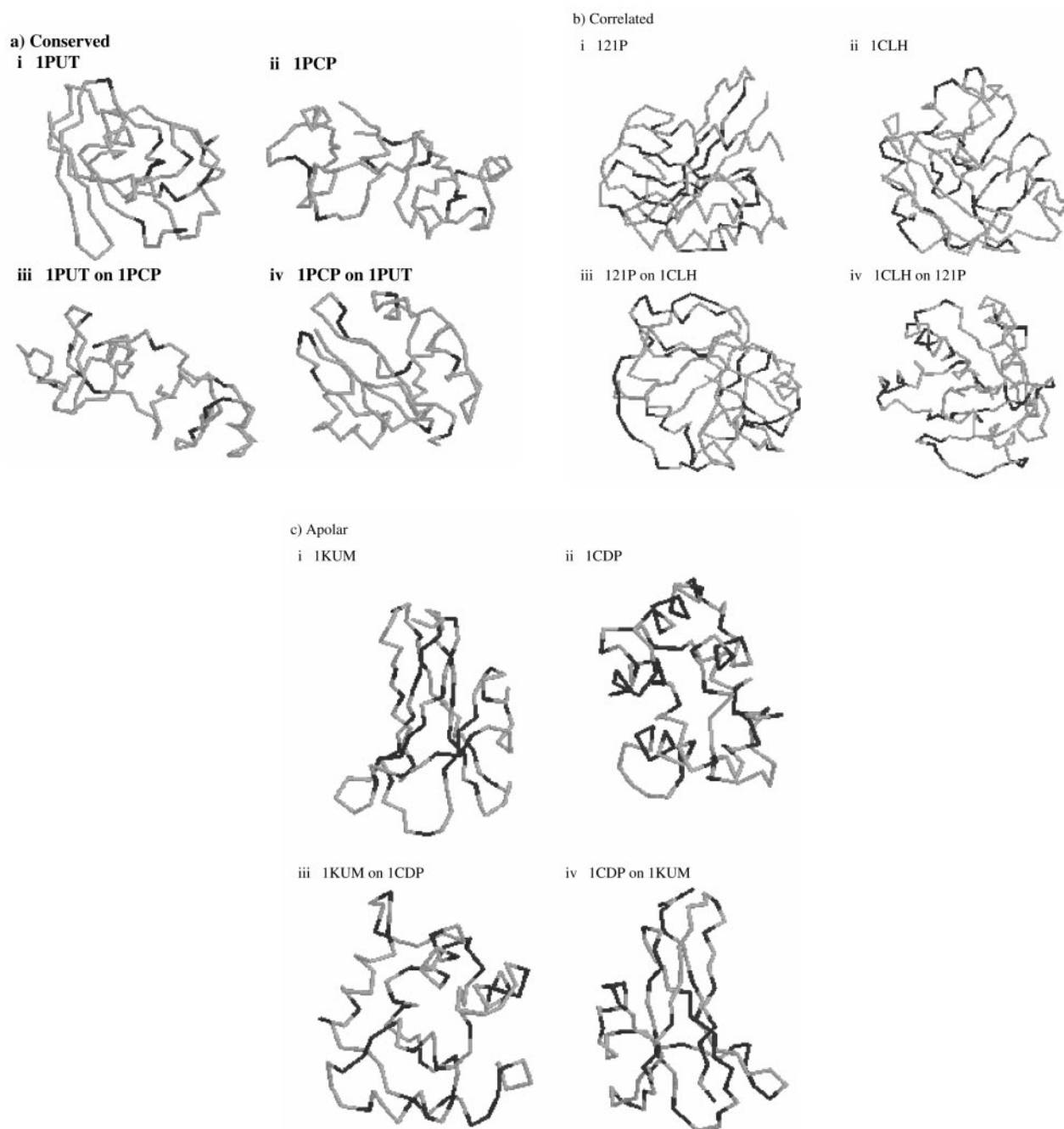


Figure 5. Examples of the distribution of residues in protein structures. The structures of the real proteins are shown in i and ii, and the corresponding swapped models in iii and iv. The different sets of residues are highlighted in dark shading. (a) Cluster of invariant residues in redoxin (1PUT) (i) and a growth factor (1PCP) (ii) and the corresponding models generated with the 1PUT sequence in the 1PCP structure (iii) and *vice versa* (iv). The corresponding X_d values are: 1PUT, 8.27; 1PCP, 1.88; 1PUTon1PCP, -2.83 ; and 1PCPon1PUT, -5.04 . (b) Best $L/2$ correlated pairs of residues in ras-p21 protein (121P) (i) and cyclophilin (1CLH) (ii) and the incorrect models of 121P in the 1CLH structure (iii) and *vice versa* (iv). For clarity, all correlated residues are shown without differentiating which particular pairs or groups of residues are correlated. The corresponding values of X_d are: 121P, 4.47; 1CLH, 2.34; 121Pon1CLH, -4.43 ; and 1CLHon121P, -1.28 . (c) Apolar residues in the structure of a hydrolase (1KUM) (i) and a calcium-binding protein (1CDP) (ii) and the models of 1KUM in the 1CDP structure (iii) and *vice versa* (iv). The corresponding values of X_d are: 1KUM, 4.39; 1CDP, 0.89; 1KUMon1CDP, -2.53 ; and 1CDPon1KUM, -2.20 .

Similarly successful were the other pairwise combinations of conservation-correlation (Figure 6(b)) and correlation-apolar (Figure 6(c))

with a discrimination of 99.7% (conservation-correlation), and 98.8% (correlation-apolar) of the incorrectly folded models. When combining all three

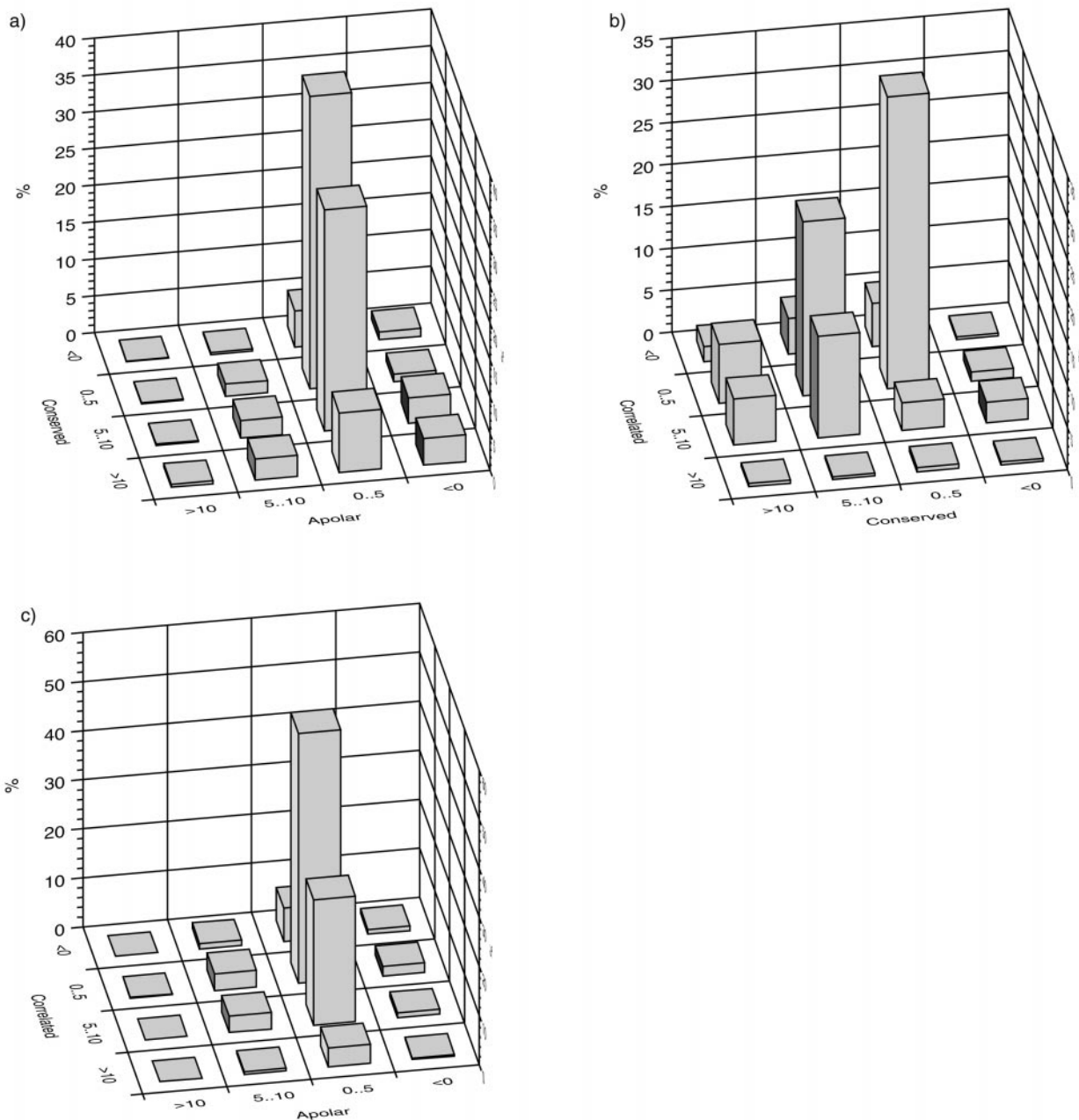


Figure 6 (legend opposite)

sources of sequence information, only one anomalous case was not correctly identified. This was the Tat protein (Figure 6(d)) which was solved under natural conditions, 1TVT (Willbold *et al.*, 1994) and in a trifluoroethanol solution 1TVS (Sticht *et al.*, 1994). In our test, the protein folded in a non-natural environment appeared like an incorrectly folded protein when compared with its structure obtained under normal conditions, and was clearly different from all other proteins in the test set. This observation concurs with our hypothesis that incorrectly folded models could be discriminated on the basis of the distribution of characteristic residues.

Combining threading results with sequence information, in practice

Evaluating accuracy for a large dataset

Since the sequence information proved successful in discriminating between incorrectly folded protein models, we applied it as a post-processing filter for threading models provided by the prediction-based threading method TOPITS (Rost, 1995; Rost *et al.*, 1997). As previously established (Rost, 1995), the accuracy of the threading method was correlated with the statistical Z-score describing the significance of a finding. At a level of TOPITS Z-score >1.5 , 59 correct solutions opposed 513

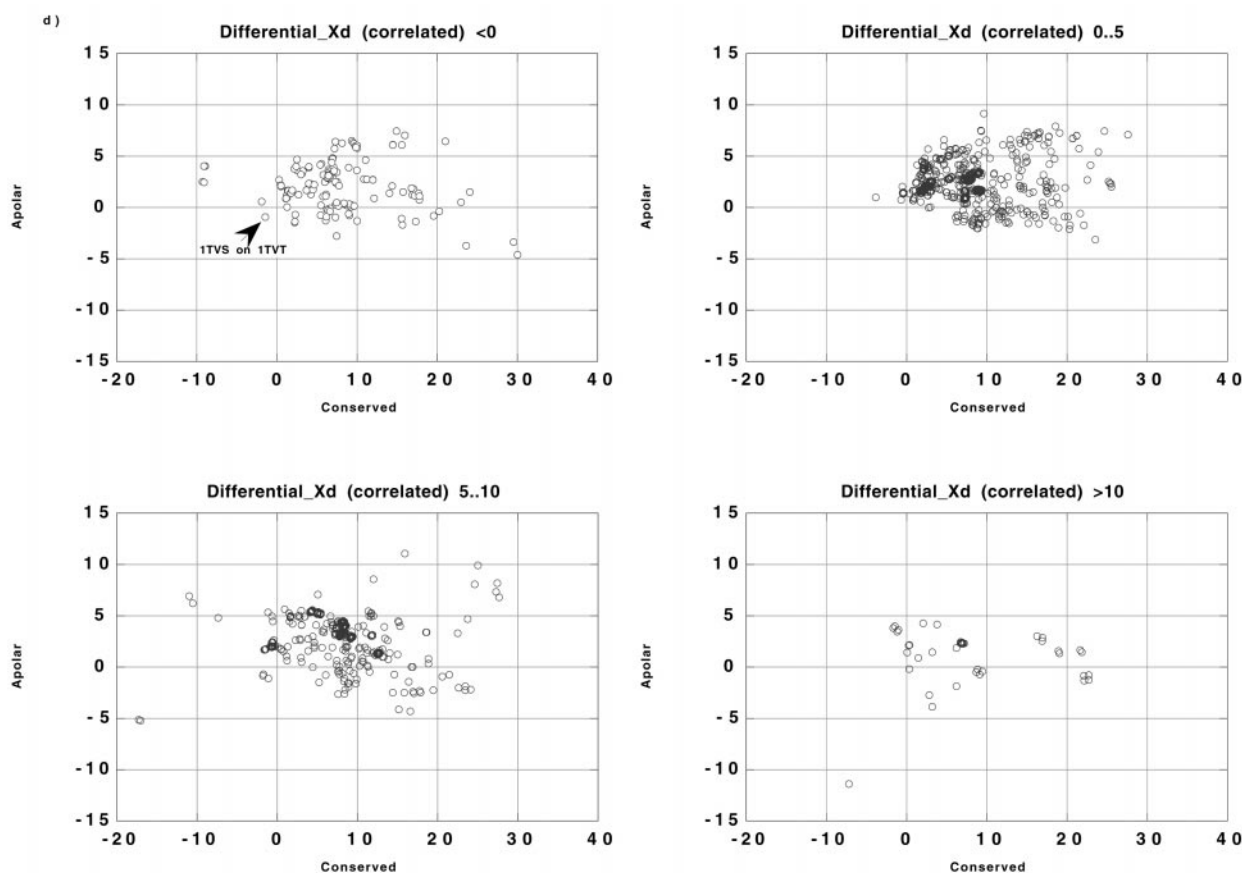


Figure 6. Combined results of conservation, correlation and apolarity. The percentage of pairs found at four *Differential_Xd* cut-offs are represented for the combinations of (a) conservation-apolarity, (b) conservation-correlation and (c) correlation-apolarity. The number of pairs with *Differential_Xd* value for both parameters is (a) 1884, (b) 1866 and (c) 4616. (d) The *Differential_Xd* values are represented for the combination of conservation and apolarity at three different values of correlation *Differential_Xd* (i, less than 0; ii, between 0 and 5; iii, between 5 and 10; and iv, greater than 10). In only one case (1TVSon1TVT) are *Differential_Xd* values less than 0 for the three parameters.

incorrect ones (*Precision* of 11.5, Figure 7); at a level of Z-scores >2.0 , the precision improved to 15% (Figure 7). Furthermore, structurally more similar proteins were found more reliably than borderline cases: structurally more similar pairs were found at a correct/incorrect solution ratio of 7 (Figure 7).

All the implicit threading models were simultaneously evaluated by the proximity of correlated, conserved and apolar residues using the *Xd* parameter. Any of the three sequence criteria improved finding remote homologs (Figure 7). In particular for TOPITS Z-scores between 1.5 and 2, many false positives could be excluded due to their low *Xd* values (Table 1A). For example, TOPITS alone for Z-score >1.5 recognized 11.5% of the correct folds (59 correct identifications over 513 pairs of alignments at this level of TOPITS score), four times more than at TOPITS Z-scores <1.5 (*Precision* 2.8%, 121 positive cases out of 4246 possible ones). The different sequence information parameters produce a clear improvement in the fold recognition results by themselves or combined with the TOPITS scores. For example, at TOPITS Z-scores better than 1.5 correlation leads to a 17.5%

precision, conservation to 16.6% and apolar residues to 15%. These results are significantly better than those obtained for the same sequence parameters at TOPITS Z-scores less than 1.5, and also better than the results obtained at *Xd* values smaller than zero.

When all three *Xd* values for conservation, correlation, and apolarity were required to be larger than zero, the performance to improved a 24% *Precision*, a value 3.4-fold better than the improvement obtained when any of the three *Xd* values was smaller than zero (*Precision* 7.1%). The significant improvement obtained by using three sources of sequence information was based on the large degree of independence among the three sequence-based parameters. The number of cases in which these three parameters were simultaneously larger than zero was 544 pairs, including 61 cases of positive fold recognition, *Precision* 11.2%. This *Precision* was similar to the result achieved by using only TOPITS at a Z-score of 1.5, 59 correct for 513 total number of aligned pairs of proteins, with a *Precision* of 11.5%. We concluded that sequence and structure-based methods were also mutually fairly

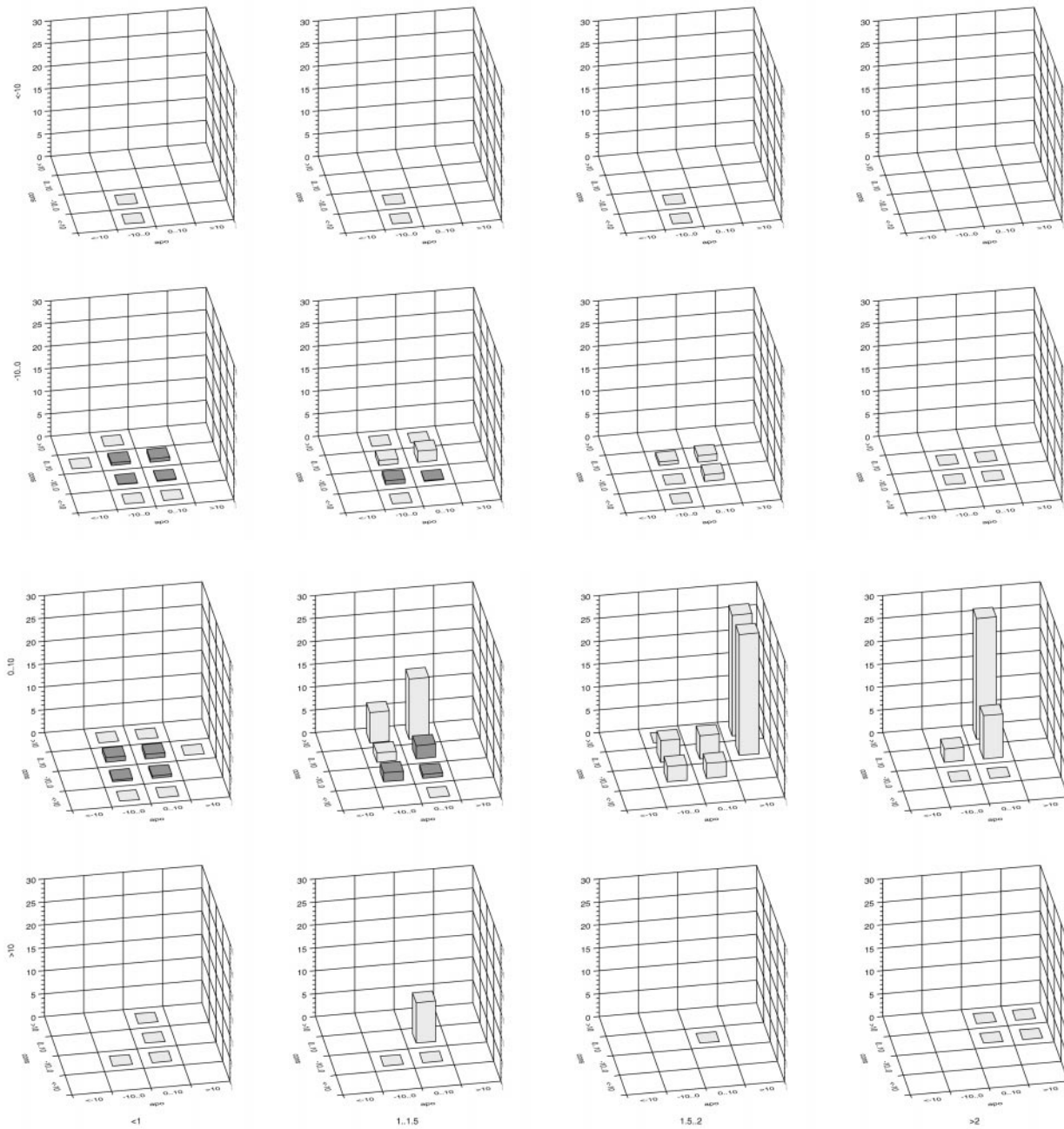


Figure 7. Evaluation of threading models with sequence information. The combined plot contains the information on the fraction of pairs whose fold was correctly recognized for different TOPITS Z-score values (large X-axis, ≤ 1 , $1 < \leq 1.5$, $1.5 < \leq 2$, > 2) and split into four different correlation X_d cut-offs (large Y-axis, < -10 , $-10 < \leq 0$, $0 < \leq 10$, > 10). At each one of the combinations, a plot of apolarity (small x-axis, < -10 , $-10 < \leq 0$, $0 < \leq 10$, > 10) against conservation (small y-axis, < -10 , $-10 < \leq 0$, $0 < \leq 10$, > 10) is given. The values correspond to the proportion of correct predictions over the total number of predictions compared with a random prediction, as described in Methods (*Discr* value in equation (4)). In the Figure, structural similarity is defined with a cut-off FSSP Z-score of 2. The actual number of observations in some of the most relevant combinations are given in Table 1.

independent, since they worked better in combination than any of them alone (*Precision* 24.1%, 32 positive cases for 133 possible ones).

Finally, we repeated the analysis for a more restrictive definition of correct fold identification (FSSP Z-scores > 3.5 , Table 1B). In this subset, the number of possible true positives was reduced from 180 to 74. The conclusion nonetheless remains

the same, since both threading and sequence-based criteria improved in this more conservative subset. In fact, the precision of TOPITS and combined sequence information was 6.3-fold more accurate than a random prediction in the set of FSSP Z-score > 2.0 , to be compared with the 14-fold improvement in the set of FSSP Z-scores > 3.5 .

Table 1. Fold recognition

A. Fold recognition in some of the most relevant combinations of threading, conservation, correlation and apolarity^a at a structural similarity cut-off of FSSP Z-score>2^b

		TOPITS < 1.5	Any TOPITS	TOPITS > 1.5
Correlation	Any <i>Xd</i>	121/4246 = 2.8 %	180/4759 = 3.7 %	59/513 = 11.5 %
	<i>Xd</i> > 0	73/1636 = 4.5 %	127/1945 = 6.5 %	54/309 = 17.5 %
Conservation	<i>Xd</i> < 0	48/2610 = 1.8 %	53/2819 = 1.8 %	5/204 = 2.5 %
	<i>Xd</i> > 0	79/1608 = 4.9 %	123/1873 = 6.5 %	44/265 = 16.6 %
Apolarity	<i>Xd</i> < 0	42/2638 = 1.5 %	57/2886 = 1.9 %	15/248 = 6.0 %
	<i>Xd</i> > 0	62/1777 = 3.5 %	105/2059 = 5 %	43/282 = 15.0 %
Combined: Corr & Cons & Apol.	<i>Xd</i> < 0	59/2469 = 2.4 %	59/2469 = 2.4 %	16/231 = 6.9 %
	All <i>Xd</i> > 0	29/411 = 7.1 %	61/544 = 11.2 %	32/133 = 24.1 %
	At least one <i>Xd</i> < 0	92/3835 = 2.4 %	119/4215 = 2.8 %	27/380 = 7.1 %

B. Improvement of fold detection results with the increase in structural similarity

Combined: Corr & Cons & Apol.	Any <i>Xd</i>	FSSP Z-score>2.0 ^b		FSSP Z-score>3.5 ^b	
		TOPITS >1.5	Any TOPITS	TOPITS > 1.5	Any TOPITS
		59/513 = 11.5 %	180/4759 = 3.8 %	36/513 = 7 %	74/4759 = 1.5 %
		3 folds ^d	background ^c	4.7 folds ^d	background ^c
	All three <i>Xd</i> > 0	32/133 = 24.1 %	61/544 = 11.2 % 2.9	28/133 = 21.1 %	39/544 = 7 %
		6.3 folds ^d	folds ^d	14 folds ^d	4.7 folds ^d

^a The results are given as positive cases of fold identification, total number of cases that could have been identified at each one of the cut-off levels and the corresponding percentage of fold identification (*Precision*, see equation (5)). An example could be taken from the right upper corner of the Table. There are 36 cases of correct fold identification by threading (TOPIT > 1.5 column) out of 513 possibilities in the set of clear structurally related pairs, the corresponding *Precision* is 7%. The additional analysis of those cases with correlated mutation information ("correlation" rows) identifies positively (*Xd* > 0) 54 cases out of the possible 309 cases (*Precision* 17.5%).

^b The fold recognition results were assessed at different levels of structural similarity. In A the cut-off level is of FSSP Z-scores of 2.0 and in B results at cut-offs of 2 and 3.5 are compared. All the other combinations of TOPITS scores, correlation, conservation and apolarity at different cut-off levels are represented in Figure 7 or given at <http://www.gredos.cnb.uam.es/olmea/models/combinations>.

^c Background level as given by the ratio between the number of related pairs of structures at a given FSSP Z-score and all pairs of alignments provided by TOPITS.

^d Increase in the fold identification precision as compared with the background level.

Studying the details of the combination for CheY

The search for a possible structure for 3CHY (CheY) with TOPITS returned a sorted list of models that were scored by the proximity of the correlated residues. In this case, the best model (2DRI, sugar transport protein, with an FSSP Z-score of 8.9) did not score as the first one in the TOPITS or correlated mutation list. Another interesting case was 5P21 (ras-p21, with an FSSP Z-score of 7.2), which would have been difficult to identify only by threading (TOPITS Z-score of 1.96) or by correlation (ranking position 6, with an *Xd*-correlation of 5.0). The combined score of the two parameters ranked among one of the best-scoring proteins for both parameters (best-scoring protein if restricting the hits to those with *Xd* correlation > 4). In detail, the model alignment revealed shifts typical for slightly incorrect models in different regions (Figure 8). Interestingly, the correlated pairs still were close in space, although the model contained errors. Obviously, the proximity of the correlated residues was better in the correct structural alignment (*Xd* value of 6.5 for the structural alignment *versus* 5.0 for the TOPITS alignment).

Studying the details of the combination for glutathione-reductase

In the threading search with the sequence of 3GRS (glutathione-reductase) among the 11 pro-

teins with similar structures the best candidates was 1FCD (flavocytochrome *c* sulfide dehydrogenase, Z-score of 23.3, in two alternative structural alignments). This protein ranked second by TOPITS and fourth by *Xd*-conservation. We selected another structurally similar protein (1PBE, p-hydroxy-benzoate hydroxylase, FSSP Z-score 11.4) as a typical example that would have been very difficult to identify by threading or conservation alone but became a much better candidate when the two parameters were considered (Figure 9). As for CheY, the proximity of the correlated residues was better in the correct structural alignment (*Xd* value of 11.9 for the structural alignment *versus* 3.1 for the TOPITS alignment).

Discussion

We showed that long-range inter-residue contacts could be predicted at low but significant levels of accuracy using information from sequence conservation and correlation. The tendency of conserved and correlated residues to cluster in space became more obvious when using real values for spatial distances rather than binary contacts: the distance histograms for conserved and correlated residues were shifted toward smaller values than for all other residue pairs.

Correlated mutations discriminate incorrectly folded proteins

We hypothesize that some of the correlations detected during the analysis of protein families corresponded to substitutions that were compensated during protein evolution by other replacements in the nearby structural regions. This compensation may relate to the limited window of stability of a protein structure. Thus, the signal detected as correlation can be expected to correspond to a general structural proximity. Although valuable, this signal is not strong enough for satisfactory prediction of inter-residue distances. However, correlated mutations are at least as successful in recognising incorrectly folded proteins as in sequence conservation.

Why are correlated mutations not more accurate?

In the following, we explain why we believe that it appears unlikely that it will be possible to achieve dramatic improvements in predicting residue contacts on the basis of the information contained in sequence conservation and correlation. In our understanding, the main reasons are:

Sampling errors in real protein families, such as alignment errors, unequal distribution of sequences in the sequence space (e.g. clusters of protein families). These factors may render strong correlation signals caused by the constant presence of one or a few sequences that are difficult to align with the rest of the protein family.

The natural signal of the compensation process is probably mixed with many other changes acquired by the structures (e.g. adaptation to new functions, new specificity or new folding requirements). We know too little of these processes to be able to account for them.

Our definition of correlation was restricted to residue pairs. However, it would be more natural to consider entire networks of co-varying residue positions.

Uncorrelated residue pairs may be part of the same contact network. For example, some of the invariant positions are typically involved in contact networks. However, uncorrelated pairs were not considered in our definition of sequence correlation. This may have distorted the observed distribution of distances between correlated positions. Similarly, positions influencing structure indirectly may have created signals of correlation not corresponding to direct physical proximity (Lapedes *et al.*, 1997).

Apolar residues and protein cores contain very limited information

Surprisingly, we found that contact predictions derived from subsets of apolar residues were of fairly limited accuracy. The best predictions were obtained with the subset of apolar-Huang residues. This definition of apolar residues has been used as

the basis for defining solvation potentials, effective in analyzing protein models (Huang *et al.*, 1995, 1996). In our analysis, they were only slightly superior to other types of sequence information, such as conservation or correlation.

Combining conservation, correlation, and apolar information successfully recognized incorrectly folded proteins

Despite the tendency of apolar and correlated positions to be conserved (Hubbard & Blundell, 1987; and our unpublished results), these variables contain sufficiently independent information to improve our results with their combination as independent variables. Indeed, the selection of those pairs identified by any of the three criteria led to the identification of all the cases, with only one exception (0.05% of the sample). Consequently, our results were similar to those obtained using more sophisticated approaches for finding incorrectly folded models; for example, semi-empirical energy force fields (Abagyan *et al.*, 1994; Gregoret & Cohen, 1990; Novotny *et al.*, 1988), mean field potentials derived from statistical analysis of structural databases (Bowie *et al.*, 1990; Bryant & Amzel, 1987; Lüthy *et al.*, 1992; Ouzounis *et al.*, 1993; Sippl, 1995; Sippl & Weitckus, 1992; see also Wodak & Rooman, 1993), regularities in protein structure according to tabulated sets of constraints (Clark *et al.*, 1991; Rooman *et al.*, 1992; Rooman & Wodak, 1992; Vriend, 1990) and solvation potentials (Holm & Sander, 1992; Huang *et al.*, 1995, 1996; Juffer *et al.*, 1995; Thanki *et al.*, 1991).

Improving threading through combined sequence information

The way we used sequence information was rather naive, in that we did not attempt driving threading algorithms based on the information we extracted. Rather, we simply post-filtered the output from one particular threading algorithm. We then answered the question: did the model that was predicted by threading appear correct using sequence information? This was a particularly difficult enterprise, since automatic threading methods have severe difficulties in producing correct alignments, even if fold identification is correct. Fortunately, the sequence information seemed to be strong enough to overcome this problem: the incorrect threading solutions could be rejected in many cases. This improved the ratio of correct *versus* incorrect folds significantly.

Each of the three sequence variables, correlation, conservation and polarity, improved the results for different proteins. This indicated again that the three evidences of sequence information were of distinct nature. The simple combination of the three features through a logical "or" operation resulted in a discrimination comparable to that produced by the score of the threading program.

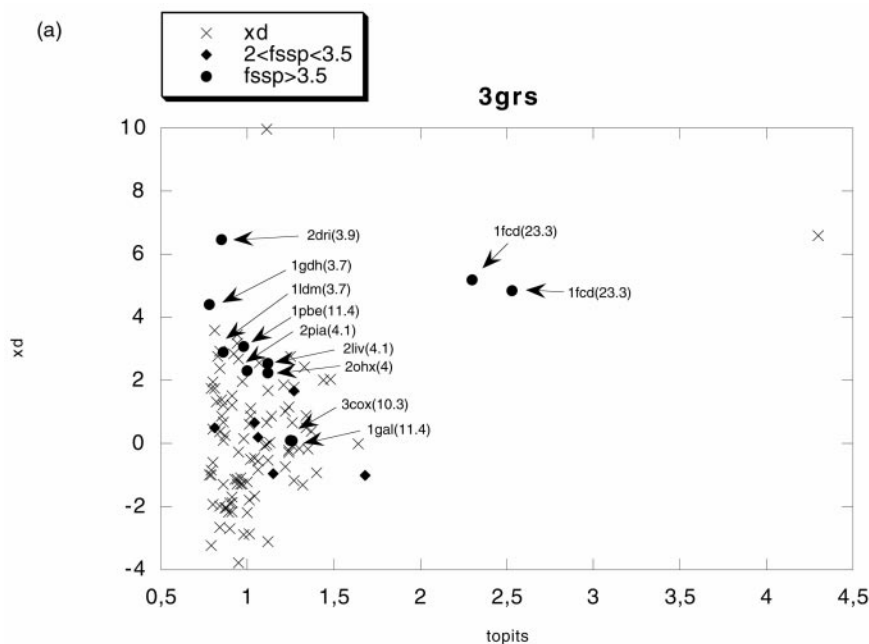


Figure 9 (legend opposite)

The best final results were obtained by combining threading with all three evidences of sequence information. Sequence information was particularly useful for retrieving correct models with low threading scores.

Future directions

Conservation and correlation obviously contain similar amounts of information. However, our success in post-processing combination of the two demonstrated clearly that they are not identical, as previously suggested (Taylor & Harrick, 1994). Thus, it may be beneficial to develop a strategy to combine these. Indeed, a simple linear combination of conservation and correlation indicates some improvement (Olmea & Valencia, 1997). Furthermore, it may be worth including correlation and conservation in other contact prediction approaches, such as statistical approaches (Galaktionov & Marshall, 1994; Galaktionov & Rodionov, 1981; Hubbard & Park, 1995; Thomas *et al.*, 1996) or neural network training (Fariselli & Casadio, 1999; Lund *et al.*, 1997).

In our analysis, the subset of invariant-binding residues produced a small improvement in the prediction of contacts. Future development of binding-site potentials will be needed to confirm this result.

Finally, we have presented a great variety of data illustrating how protein sequence information could improve fold recognition by filtering threading models. The results encourage a more sophisticated application, going beyond filtering schemes toward the direct incorporation of sequence corre-

lation and conservation information into threading algorithms.

Methods

Definition of conservation

Sequence conservation is defined in terms of the Variability scale (Sander & Schneider, 1993), which ranges from variability zero (invariant residues) to extremely variable residues with values greater than 50. Our results are presented as four conservation classes, 0, 1-13, 14-18, and more than 21 variability units.

Calculation of correlation

Correlated mutations were calculated as described (Göbel *et al.*, 1994). Each position in the alignment is coded by a distance matrix. This position-specific matrix contains the distances between all sequence pairs at that position. Distances are defined by the scoring matrix presented by McLachlan (1971). The association between each pair of positions is calculated as the average of the correlation for each corresponding bin of the position-specific matrices. Positions with more than 10% gaps or those that are completely conserved were not included in the calculation.

The formula used for calculation of the correlation coefficient (r_{ij}) for each pair of positions i and j of a protein, with N proteins in the alignment, is:

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

For each position in the alignment, we have an $N \times N$ matrix in which each element (k and l from 1 to N) is the similarity (S_{ikl}) between the two residues (k and l) in this position (i) according to the given homology

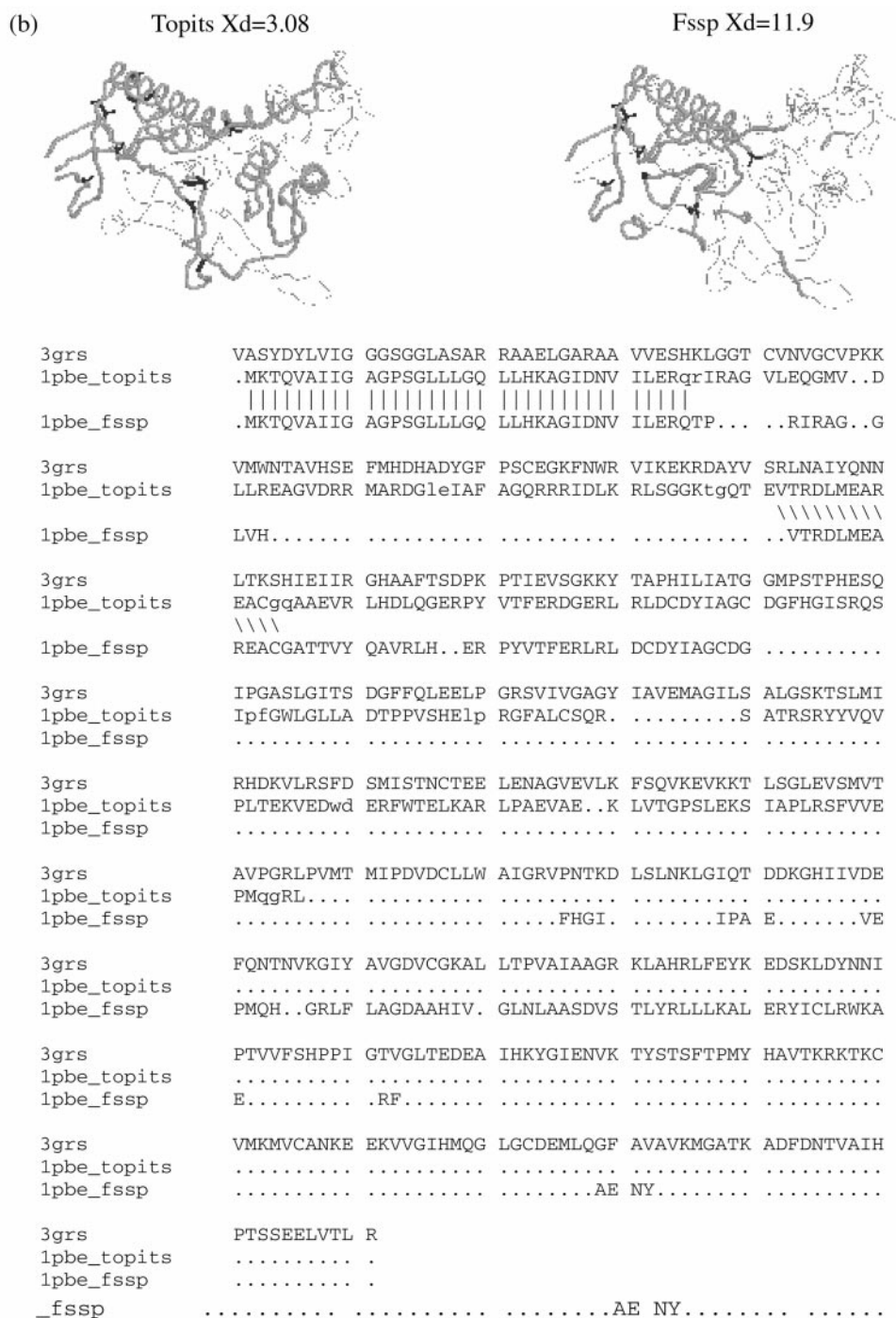


Figure 9. Example of the analysis of a TOPITS run combining threading and information about sequence conservation. The same representation as in Figure 8 is used to illustrate the combination of threading with sequence conservation. In this case, the example is taken from the threading run of 3GRS (glutathione-reductase) in which 1PBE (*p*-hydroxy-benzoate hydroxylase) is identified as a possible related protein. In this case, the structural similarity is clear, with an FSSP Z-score of 11.4. The structural and threading alignments differ by more than four residues in different regions. The position of the conserved residues in the 1PBE family is represented by the corresponding C^β atoms in the three-dimensional plots.

matrix. $\langle S_i \rangle$ is the mean of S_{ikl} ; σ_i is the standard deviation of S_{ikl} .

Given that the accuracy of contact prediction depends directly on the correlation values (Göbel *et al.*, 1994; Olmea & Valencia, 1997) the pairs of positions are sorted by their correlation value and the best M residues are

defined as predicted contacts, with M proportional to the protein size. For this study, we present results at different proportions of sequence length (L): $L/2$, L , $7L$ and greater than $7L$.

Common measurements of contact prediction such as *Accuracy* are highly dependent on protein size. Part of

the analysis is therefore presented by dividing proteins into four size categories (31-99, 103-166, 169-298, and 312-823 amino acid residues long), chosen to contain similar numbers of proteins.

Threading method

The threading experiments have been carried out with the TOPITS program (Rost, 1995), in the implementation of Olmea (unpublished).

Each of the threading runs was carried using equal weighting for the comparison of the single sequences and for the matching between predicted and observed secondary structures. Sequence similarity was evaluated with the Blousum matrix and the similarity between predicted and observed secondary structures and accessibility with the matrix derived by Rost (1995).

Test set

For the comparison of correctly and incorrectly folded proteins, three different protein test sets were used. The first set comprises 71 non-redundant protein families from the PDB-select list of August 1995 (Hobohm *et al.*, 1992), which was used to assess the accuracy of contact prediction with conserved, correlated or simply apolar residues.

The second set contains a larger set of protein pairs of identical size and different structure, and includes possible redundant sequence families. The additional restriction used is that each protein family must have at least 10% of positions in the category under study; that is, more than 10% conserved or apolar residues. The experiments were performed with 3612 protein-model pairs in the test with conserved residues, 4616 with correlated pairs, 4634 with apolar, 4640 with apolar according to Huang *et al.* (1995) (apolar-Huang) and 1005 with conserved-binding residues. For the double test with two variables, the number of pairs was 1884 for apolar-conserved, 4616 for apolar-correlated and 1866 for conserved-correlated and for the triple combination of conservation, correlation, and apolarity.

The third test set was additionally restricted to contain only sequences with less than 25% sequence similarity to any other sequence in the October 1997 set (Hobohm *et al.*, 1992). This set contains 56 protein pairs with enough conserved residues, and 33 with enough correlated pairs, to be analyzed.

Protein families were selected from the HSSP database (Sander & Schneider, 1993). To guarantee that protein pairs of identical size had different structure (sets 2 and 3), it was required that they not be recognized by a common algorithm of structural comparison (Holm & Sander, 1993), such that they are never contained in the same structural alignment as deposited in the FSSP database (Holm & Sander, 1996).

The threading experiments correspond to individual runs of each of the proteins in a non-redundant set (Hobohm *et al.*, 1992) with more than 15 alignments in the corresponding HSSP file (release 31, 1995, Sander & Schneider, 1993). In addition, only those alignments with more than 10% of positions in the corresponding category of conserved, apolar or correlated were used, giv-

ing a different number of cases than described for test set 2. The runs were carried out against a non-redundant database of 582 protein chains, of which those with more than 20 sequence similarity to each of the query proteins were excluded in the corresponding experiment†.

Assessment of the distance between pairs of positions

A strict definition of contact was first used, considering a prediction correct when two C^β atoms are closer than 8 Å, and defining *Accuracy* as the percentage of correctly predicted over the total number of predicted pairs.

Second, we used a definition in better agreement with the idea of residue networks in spatial proximity, but not necessarily in physical contact. In practice, the weighted difference was computed between the binned populations of distances between all pairs and correlated or conserved pairs (Pazos *et al.*, 1997a).

The distances between residue pairs are grouped in bins of 4 Å and the distribution represented as relative proportions of pairs of contacts. Two different distributions of binned data are obtained for the predicted pairs and for all other pairs of positions. The difference between the two distributions is calculated bin by bin and weighted by a factor inversely proportional to the normalized distance of the corresponding bin. The weight factor is introduced to increase the importance of closer distances. Distances between residues correspond to C^β - C^β distances (C^α for Gly):

$$Xd = \sum_{i=1}^{i=n} \frac{P_{ic} - P_{ia}}{d_{in}} \quad (2)$$

where n is the number of distance bins; there are 15 equally distributed bins from 4 to 60 Å. d_i is the upper limit for each bin, e.g. 8 for the 4-8 bin (normalized to 60). P_{ic} is the percentage of correlated pairs with distance between d_i and d_{i-1} , P_{ia} the same percentage for all pairs of positions. Defined in this way, $Xd = 0$ indicates no separation between the two distance populations, $Xd > 0$ indicates positive cases in which the population of predicted pairs is shifted to smaller distances with respect to the population of all pairs.

Evaluation of the discrimination between real and incorrectly modelled proteins

Pairs formed by proteins and incorrect models of the same sequence modelled on a different structure are compared with the simple measure of:

$$\begin{aligned} \text{Differential_Xd} &= Xd \text{ of the real protein} \\ &- Xd \text{ of the modelled protein} \end{aligned} \quad (3)$$

Differential_Xd values greater than zero indicates that the conserved or correlated residues form a network of proximity sufficient to discriminate between real and incorrectly folded proteins.

Given the type of low-resolution information being used and the fact that we are searching only for distances and not for specific molecular environments, building and refinement of the protein models was not necessary, thus avoiding any doubts about the modelling. This question is always present in other studies, such as those of solvation potentials, in which the potentials depend on the specific protein environment.

† The list of proteins and the corresponding set of threading results are provided at: <http://www.gredos.cnb.uam.es/olmea/models/threading>.

Evaluation of the discrimination of threading models

For each threading alignment, the corresponding structural similarity level in FSSP (Z-score) (Holm & Sander, 1993, 1996), threading score (TOPITS Z-score), and Xd value of conservation, correlation, and apolarity were computed. The results are given in terms of proportion of fold correctly identified, that is, pair with an FSSP Z-score better than a given threshold:

$$Discr(i, j, k, l) = \frac{N(i, j, k, l)_{(Z-FSSP > threshold)}}{N(i, j, k, l)} \quad (4)$$

where i, j, k and l are different boundaries of TOPITS Z-score, Xd -correlation, Xd -conservation and Xd -apolarity. $N(i, j, k, l)$ is the number of pairs in a given combination of scores (e.g. Z-score TOPITS between 1 and 1.5, Xd -correlation between 0 and 10), $N(i, j, k, l)_{(Z-FSSP > threshold)}$ is the number of pairs in the same boundaries that are structurally similar, corresponding to the number of positive identifications. The denominator represents the background probability of a correct fold identification in the experiment, where $N(total)$ is the number of pairs in the experiment (total number of alignments provided by TOPITS at a given Z-score cut-off), $N(total)_{(Z-FSSP > threshold)}$ is the total number of pairs considered to be structurally similar, that is, having FSSP Z-scores larger than the threshold.

In some cases, the actual number of correct $N(i, j, k, l)_{(Z-FSSP > threshold)}$ and total cases $N(i, j, k, l)$ cases and the corresponding percentage (*Precision*) is given:

$$Precision(i, j, k, l) = N(i, j, k, l)_{(Z-FSSP > threshold)} / N(i, j, k, l) \quad (5)$$

Acknowledgements

We are indebted to Chris Sander (Whitehead Institute Cambridge, MA) for the initial discussions on the use of conserved residues and to C. Sander and Georg Casari (Lion-AG, Heidelberg) for the suggestion of using the set of incorrectly folded proteins during a meeting in Madrid, 1994. This work was supported, in part, by a grant from the CICYT-Spain BIO94-1067.

References

Abagyan, R., Totrov, M. & Kuznetsov, D. (1994). ICM-a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488-506.

Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**, 693-707.

Altschuh, D., Vernet, T., Moras, D. & Nagai, K. (1988). Coordinated amino acid changes in homologous protein families. *Protein Eng.* **2**, 193-199.

Andrade, M. A., Casari, G., Sander, C. & Valencia, A. (1997). Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol. Cybern.* **76**, 441-450.

Artymiuk, P. J., Grindley, H. M., Kumar, K., Rice, D. W. & Willett, D. W. (1993). Three-dimensional structural resemblance between the ribonuclease H and connection domain of HIV reverse transcriptase and the ATPase fold revealed using graph theoretical techniques. *FEBS Letters*, **324**, 15-21.

Bairoch, A. (1992). PROSITE: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **20**, 2013-2018.

Benner, S. A. (1989). Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Advan. Enzyme Regul.* **28**, 219-236.

Benner, S. A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advan. Enzyme Regul.* **31**, 121-181.

Bork, P. (1989). Recognition of functional regions in primary structures using a set of property patterns. *FEBS Letters*, **257**, 191-195.

Bork, P., Gellerich, J., Groth, H., Hooft, R. & Martin, F. (1995). Divergent evolution of a *b/a* barrel subclass: detection of numerous phosphate-binding sites by motif search. *Protein Sci.* **4**, 268-274.

Bowie, D., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins: Struct. Funct. Genet.* **7**, 257-264.

Bryant, S. H. & Amzel, L. M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. *J. Int. Pept. Protein Res.* **29**, 46-52.

Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171-178.

Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G. H. & Benner, S. A. (1997). An analysis of simultaneous variation in protein structures. *Protein Eng.* **10**, 307-316.

Clark, D. A., Shirazi, J. & Rawlings, C. J. (1991). Protein topology prediction through constraint-based search and the evaluation of topological folding rules. *Protein Eng.* **4**, 751-760.

Cooperman, B. S., Baykov, A. A. & Lahti, R. (1992). Evolutionary conservation of the active site of soluble inorganic pyrophosphatase. *Trends Biochem. Sci.* **17**, 262-266.

Defay, T. R. & Cohen, F. E. (1996). Multiple sequence information for threading algorithms. *J. Mol. Biol.* **262**, 314-323.

Fariselli, P. & Casadio, R. (1999). A neural network based predictor of residue contacts in proteins. *Protein Eng.* **12**, 15-21.

Fisher, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947-955.

Gabrielian, A. E., Ivanov, V. S. & Kozhich, A. T. (1990). On searching for the active sites in proteins and peptide hormones. *Comput. Appl. Biosci.* **6**, 1-2.

Galaktionov, S. G. & Marshall, G. R. (1994). Properties of intraglobular contacts in proteins: An approach to prediction of tertiary structure. In *27th Hawaii International Conference on System Sciences* (Waialea, H. L., ed.), pp. 326-335, IEEE Society Press, USA.

Galaktionov, S. G. & Rodionov, M. A. (1981). Calculation of the tertiary structure of proteins on the basis of analysis of the matrices of contacts between amino acid residues. *Biophysics*, **25**, 395-403.

- Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309-317.
- Godzik, A. & Sander, C. (1989). Conservation of residue interactions in a family of Ca-binding proteins. *Protein Eng.* **2**, 589-96.
- Gregoret, L. M. & Cohen, F. E. (1990). Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.* **211**, 959-974.
- Gribskov, M., McLachlan, M. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355-4358.
- Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167-180.
- Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409-417.
- Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93-105.
- Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.
- Holm, L. & Sander, C. (1994). Searching protein structure databases has come of age. *Proteins: Struct. Funct. Genet.* **19**, 165-173.
- Holm, L. & Sander, C. (1995). DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily. *Trends Biochem. Sci.* **20**, 45-347.
- Holm, L. & Sander, C. (1996). The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucl. Acids Res.* **24**, 206-210.
- Holm, L., Sander, C. & Murzin, A. (1994). Three sisters, different names. *Nature Struct. Biol.* **1**, 146-147.
- Howell, N. (1989). Evolutionary conservation of protein regions in the protonmotive cytochrome *b* and their possible roles in redox catalysis. *J. Mol. Evol.* **29**, 157-169.
- Huang, E. S., Subbiah, S. & Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* **252**, 709-720.
- Huang, E. S., Subbiah, S., Tsai, J. & Levitt, M. (1996). Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular simulations. *J. Mol. Biol.* **257**, 716-725.
- Hubbard, T. J. L. & Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159-171.
- Hubbard, T. J. & Park, J. (1995). Fold recognition and ab initio structure predictions using hidden Markov models and β -strand pair potentials. *Proteins: Struct. Funct. Genet.* **23**, 398-402.
- Hwang, P. K. & Fletterick, R. J. (1986). Convergent and divergent evolution of regulatory sites in eukaryotic phosphorylases. *Nature*, **324**, 80-83.
- Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86-89.
- Juffer, A. H., Eisenhaber, F., Hubbard, S. J., Walther, D. & Argos, P. (1995). Comparison of atomic solvation parametric sets: applicability and limitations in protein folding and binding. *Protein Sci.* **4**, 2499-2509.
- Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996). Protein sequence comparison at genome scale. *Methods Enzymol.* **266**, 295-322.
- Lapedes, A. S., Giraud, B. G., Liu, L. C. & Stormo, G. D. (1997). Correlated mutations in protein sequences: phylogenetic and structural effects. In *AMS/SIAM Conference on Statistics and Molecular Biology*, Seattle.
- Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342-358.
- Livingstone, C. D. & Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **6**, 645-756.
- Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J. & Brunak, S. (1997). Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* **10**, 1241-1248.
- Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83-85.
- Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876-888.
- Mauzy, C. A. & Hermodson, M. A. (1992). Structural homology between *rbs* repressor and ribose binding protein implies functional similarity. *Protein Sci.* **1**, 843-849.
- McLachlan, A. D. (1971). Test for comparing related amino acid sequences. *J. Mol. Biol.* **61**, 409-424.
- Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386-394.
- Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA*, **91**, 98-102.
- Novotny, J., Bruccoleri, R. E. & Karplus, M. (1984). An analysis of incorrectly folded models. Implications for structure prediction. *J. Mol. Biol.* **177**, 787-818.
- Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins: Struct. Funct. Genet.* **4**, 19-30.
- Olmea, O. & Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.* **2**, S25-S32.
- Ortiz, A. R., Kolinski, A. & Skolnick, J. (1998a). Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* **277**, 419-448.
- Ortiz, A. R., Kolinski, A. & Skolnick, J. (1998b). Native-like topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc. Natl Acad. Sci. USA*, **95**, 1020-1025.
- Ouzounis, C., Sander, C., Scharf, M. & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.* **232**, 805-825.
- Ouzounis, C., Perez-Irratxeta, C., Sander, C. & Valencia, A. (1998). Are binding residues conserved? In *Proceedings of the Fifth Annual Pacific Symposium on Biocomputing* (Hunter, L., ed.), pp. 399-410, World Scientist, Hawaii, USA.
- Pastore, A. & Lesk, A. M. (1990). Comparison of the structures of globins and phycocyanins: evidence

- for evolutionary relationship. *Proteins: Struct. Funct. Genet.* **8**, 133-155.
- Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997a). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **272**, 1-13.
- Pazos, F., Olmea, O. & Valencia, A. (1997b). A graphical interface for correlated mutations and other structure prediction methods. *Comput. Appl. Biosci.* **13**, 319-321.
- Pazos, F., Sanchez-Pulido, L., García-Ranea, J. A., Andrade, M. A., Atrian, S. & Valencia, A. (1997c). Comparative analysis of different methods for the detection of specificity regions in protein families. In *Biocomputing and Emergent Computation* (Lund, D., et al., ed.), pp. 132-145, World Scientific, Singapore, New Jersey, London, Hong Kong.
- Pollock, D. D. & Taylor, W. R. (1997). Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.* **10**, 647-657.
- Pollock, D. D., Taylor, W. R. & Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.* **287**, 187-198.
- Rooman, M. J. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: consensus regions with preferred conformation in homologous proteins. *Biochemistry*, **31**, 10239-10249.
- Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactants. *Biochemistry*, **31**, 10226-10238.
- Rost, B. (1995). TOPITS: threading one-dimensional predictions into three-dimensional structures. In *Third International Conference on Intelligent Systems for Molecular Biology* (Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. & Wodak, S., eds), pp. 314-321, AAAI Press/Menlo Park, CA, Cambridge, England.
- Rost, B. & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Genet.* **20**, 216-226.
- Rost, B., Schneider, R. & Sander, C. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471-480.
- Sander, C. & Schneider, R. (1993). The HSSP data base of protein structure-sequence alignments. *Nucl. Acids Res.* **21**, 3105-3109.
- Shindyalov, I. N., Kolchanov, N. A. & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations. *Protein Eng.* **7**, 349-358.
- Singer, M. S., Oliveira, L., Vriend, G. & Shepherd, G. M. (1995). Potential ligand-binding residues in rat olfactory receptors identified by correlated mutation analysis. *Recept. Chann.* **3**, 89-95.
- Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229-235.
- Sippl, M. J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins: Struct. Funct. Genet.* **13**, 258-271.
- Smith, R. F. & Smith, T. F. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng.* **5**, 35-41.
- Sticht, H., Willbold, D., Ejchart, A., Rosin-Arbesfeld, R., Yaniv, A., Gazit, A. & Rösch, P. (1994). Trifluoroethanol stabilizes a helix-turn-helix motif in equine infectious-anemina-virus *trans*-activator protein. *Eur. J. Biochem.* **225**, 855-861.
- Taylor, W. R. (1991). Towards protein tertiary fold prediction using distance and motif constraints. *Protein Eng.* **4**, 853-870.
- Taylor, W. R. & Harrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 342-348.
- Thanki, N., Umrana, Y., Thornton, J. M. & Goodfellow, J. M. (1991). Analysis of protein main-chain solvation as a function of secondary structure. *J. Mol. Biol.* **221**, 669-691.
- Thomas, D., Casari, G. & Sander, G. (1996). The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* **9**, 941-948.
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.* **10**, 19-29.
- Valencia, A., Hubbard, T. J., Muga, A., Bañuelos, S., Llorca, O., Carrascosa, J. L. & Valpuesta, J. M. (1995). Prediction of the structure of GroES and its interaction with GroEL. *Proteins: Struct. Funct. Genet.* **22**, 199-209.
- Vriend, G. (1990). WHAT IF: a molecular modelling and drug design program. *J. Mol. Graph.* **8**, 52-56.
- Willbold, D., Rosin-Arbesfeld, R., Sticht, H., Frank, R. & Rösch, P. (1994). Structure of the equine infectious anemia virus Tat protein. *Science*, **264**, 1584-1587.
- Wodak, S. J. & Rooman, M. J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.* **3**, 247-259.
- Zuckerandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving Genes And Proteins* (Bryson, V. & Vogel, H. J., eds), pp. 97-166, Academic Press, New York.
- Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J. Mol. Biol.* **195**, 957-961.

Edited by J. M. Thornton

(Received 8 June 1999; received in revised form 16 September 1999; accepted 20 September 1999)