ELSEVIER

Biophysical Chemistry 119 (2006) 240 – 246

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

A simple tool to explore the distance distribution of correlated mutations in proteins

Raul Perez-Jimenez, Raquel Godoy-Ruiz, Antonio Parody-Morreale, Beatriz Ibarra-Molero, Jose M. Sanchez-Ruiz *

Departamento de Quimica Fisica, Facultad de Ciencias, Universidad de Granada, Fuentenueva s/n, 18071-Granada, Spain

Received 13 June 2005; received in revised form 15 September 2005; accepted 15 September 2005
Available online 18 October 2005

Abstract

The analysis of correlated mutations in protein sequence alignments is of considerable interest, since it may provide useful energetic and even structural information (ideally, residue contacts). However, a number of recent experimental studies support the existence of long-distance communication in proteins, a fact that may lead to correlation between distant residues. We introduce in this work a simple statistical procedure to describe the relation structure–alignments on the basis of the residue–residue distance dependence of the number of residue couples over given thresholds of a correlation measure (such as a covariance value). This procedure may lead to clear pictures of the distance distribution of correlated mutations and may provide a simple but efficient tool to explore the different structural features that are reflected in the sequence alignments.

Keywords: Sequence alignments; Protein structure; Long-distance energetic coupling; Correlated mutations

1. Introduction

The analysis of correlated mutations in protein sequence alignments is of considerable interest. For instance, a significant number of such correlations could correspond to close residue contacts in the native structure and, consequently, the possibility of using information derived from sequence alignments in structure prediction has been suggested and explored [1–9].

It appears likely, however, that one of the main factors that determine the relation between sequence alignments and protein structure is the existence of long-distance interactions (long-distance energetic coupling or "long-distance communication" between distant residues). For instance, cooperative coupling between distal mutations has been shown to be involved in the acquisition of multidrug resistance to HIV-1 protease inhibition [10]. Also, hydrogen/deuterium exchange studies show that the effect of a core mutation in *E. coli* ribonuclease HI is communicated to regions outside the core [11]. Long-distance intramolecular signalling in a tRNA synthetase complex has been reported to span distances of about 40 Å [12]. Many other examples of long-distance communication in native proteins are referenced in Ref. [13]. In addition, recent work shows that cooperative processes and long-distance interactions may also occur in protein denatured states [14–16] and, in fact, a role for residual structure within the denatured state in fine-tuning protein stability has been proposed [16]. It is, therefore, at least plausible that some correlated mutations in sequence alignments actually reflect denatured-state effects and provide information about denatured-state residual structure, as we have recently suggested [17].

Regardless of their origin (native or denatured states), the molecular mechanisms involved (see Ref. [13] and references quoted therein), and their functional role (alosteric regulation, for instance), long-distance communication may lead to correlated mutations between distant residues. In fact, long-distance coupling has been detected in analyses of sequence alignments addressed at determining pathways of energetic connectivity in proteins [18,19].

It appears then that correlated mutations in proteins may span a wide range of inter-residue distances; i.e., from short distances corresponding to close contacts in the spatial

* Corresponding author. Tel.: +34 958243189.
E-mail address: sanchezr@ugr.es (J.M. Sanchez-Ruiz).

0301-4622/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2005.09.010

structure of the protein to long distances that perhaps reflect alosteric regulation mechanisms. In view of this, it appears advisable to develop simple analytical tools to specifically explore the distance distribution of correlated mutations in sequence alignments. Such tools may help to ascertain the impact of long-distance communication over close contacts in sequence alignments, provide criteria for the selection of correlation measures and, potentially, lead to structure–alignment descriptors useful in structure prediction.

In this work, we introduce a statistical procedure to describe the relation structure–alignments based on the residue–residue distance dependence of the number of residue couples over given thresholds of a correlation measure. For the sake of simplicity, we use a simple covariance value as the measure of correlation, although more complex measures could be easily implemented. We explain our approach using a set of sequence alignments derived from a BLAST search using as query the sequence of *E. coli* thioredoxin. Subsequently, we illustrate its application with a set of 24 proteins extracted from the architecture representatives in the CATH database [20].

2. Methods

Calculations reported in this work have been performed for *E. coli* thioredoxin and for a set of 24 proteins extracted from the architecture representatives in the CATH database [20] by eliminating (for calculation performance reasons) those with more than 400 residues with reported atomic coordinates in the corresponding pdb files. It is important to note that, although the CATH database provides sequences for protein domains, we used in our calculations (i.e., as queries in database searches) the whole sequences for which atomic coordinates are reported in the pdb files.

BLAST 2 (Gish, W. (1996–2003) http://blast.wustl.edu) was used to search the UniProt/TrEMBL database (http://www.ebi.ac.uk/trembl/) with the sequences of each given protein in the set as query and the default options of the search. The sequences found were aligned to the query sequence using the Smith–Waterman algorithm and those with similarity with the query higher than 0.25 were retained (similarities were calculated as the number of matches between the given sequence and the query divided by the number of residues in the latter). The 0.25 cutoff was chosen because it is usually accepted that protein from various species and having sequence similarity of at least 0.25–0.3 have similar tridimensional structures (see Ref. [21] and references quoted therein). Therefore, we assumed that most of the retained sequences share the fold of the query protein. The resulting sets of sequences (see Table 1 for the number of sequences in each set) were used in all the calculations reported in this work.

Table 1
Protein set analyzed in this work

pdb code	Number of residues	Number of sequences	σ_{opt}	f^*_{opt}	$\eta(\sigma_{opt})$
1pdc	45	83	$2.56\cdot10^{-3}$	0.24	31.5
1cdf	46	28	$4.68\cdot10^{-3}$	0.14	50.7
1npo	81	214	$2.17\cdot10^{-2}$	0.77	29.9
1ahl	49	44	$6.17\cdot10^{-2}$	0.55	49.2
2hgf	97	10	$3.13\cdot10^{-2}$	0.43	57.3
2trx	108	491	$1.03\cdot10^{-2}$	0.44	137.5
1jpc	108	99	$3.36\cdot10^{-2}$	0.62	102.8
1ytf	115	5	$6.25\cdot10^{-2}$	0.82	44.17
2cy3	117	7	$4.00\cdot10^{-2}$	0.77	172.0
1hcd	118	68	$2.08\cdot10^{-2}$	0.63	157.3
1rie	127	69	$1.85\cdot10^{-2}$	0.45	401.0
1div	149	48	$2.22\cdot10^{-2}$	0.45	156.2
1cuk	203	5	$6.25\cdot10^{-2}$	0.76	228.3
1lrv	233	5	0.125	0.63	988.6
3aah	237	20	$1.13\cdot10^{-2}$	0.36	778.0
1bg5	252	38	$7.81\cdot10^{-4}$	0.40	3951.0
1plq	258	34	$3.67\cdot10^{-2}$	0.79	526.9
1lxa	262	36	$1.68\cdot10^{-2}$	0.46	604.6
3daa	277	89	$3.18\cdot10^{-2}$	0.75	727.0
1fl2	310	285	$3.14\cdot10^{-2}$	0.75	657.6
1ppr	312	57	$4.39\cdot10^{-4}$	0.44	1062.4
1aa8	340	29	$3.87\cdot10^{-2}$	0.54	1196.0
3bcl	350	33	0.105	0.83	3344.5
1cem	362	9	$2.50\cdot10^{-2}$	0.93	346.0
1f8d	388	629	$6.71\cdot10^{-4}$	0.36	2483.0

The number of residues refers to the residues for which atomic coordinates are given in the pdb file. The number of sequences refers to those in the alignment with similarity with the query higher than 0.25. The value of the optimum covariance threshold for ρ calculation with $\Delta r = 5$ Å (Eq. (2)), the corresponding f^* value (Eq. (6)) and the value of the integral η (Eq. (5)) for the optimum threshold are also given.

3. Results and discussion

3.1. Covariance analysis of thioredoxin sequence alignments

We analyzed a set of 490 sequences each individually aligned (Smith–Waterman algorithm) to the *E. coli* thioredoxin sequence (see Methods for further details). For each couple of positions (X and Y) in the *E. coli* thioredoxin sequence, we calculated a covariance value defined as:

$$\sigma_{XY} = \sum_{\text{sequences}} \frac{(\delta_X - \langle\delta_X\rangle)\cdot(\delta_Y - \langle\delta_Y\rangle)}{N_S} = \langle\delta_X\cdot\delta_Y\rangle - \langle\delta_X\rangle\cdot\langle\delta_Y\rangle \quad (1)$$

where δ_P (with P equal to X or Y) takes a value of 1 for a given sequence if the aminoacid at position P in the sequence is the same as in the thioredoxin sequence (and takes a value of 0 otherwise), $\langle\delta_P\rangle$ is the average value of δ_P (i.e., $\Sigma\delta_P/N_S$), N_S is the total number of sequences and the sum is the overall sequences. This simple covariance calculation is similar in spirit to the χ^2 association test we have previously employed [17]. We note, however, that covariance values can be positive and negative. We take sufficiently large absolute values of σ_{XY} (regardless of their sign) as indication of correlated mutations at positions X and Y. We also note that the covariance value calculated according to Eq. (1) equals zero if one of the positions is perfectly conserved.

The above calculation yields 5778 covariance values corresponding to the 5778 residue couples in thioredoxin. We show in Fig. 1 plots of residue–residue distance, as

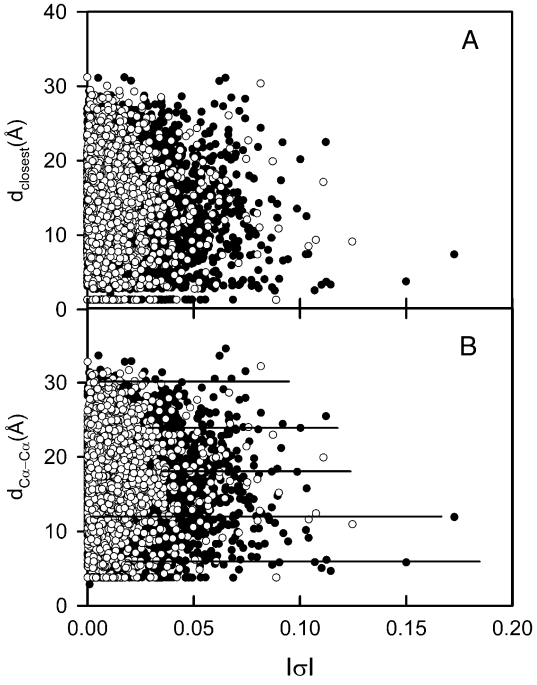

Fig. 1. Plots of residue–residue distance versus absolute value of covariance for *E. coli* thioredoxin. Covariance values are calculated from sequence alignments using Eq. (1) (see text for details). The negative covariance values are shown with open symbols. Residue–residue distances are measured by the distance between closest atoms (panel A) of between the α-carbons (panel B). The horizontal lines in panel B stand for distance values of: 6, 12, 18, 24 and 30 Å.

measured by the distance between closest atoms (Fig. 1A) and by the distance between α-carbons (Fig. 1B), versus absolute value of covariance ($|\sigma|$). Strikingly, these plots show a roughly triangular shape (see Ref. [1] for a similar result). That is, for a given value of $|\sigma|$ a certain range of residue–residue distances is observed, the average distance and the range itself showing a rough tendency to became smaller as $|\sigma|$ increases above. This result suggests that, in this case at least, the relation between structure and sequence alignments (at least, as characterized by our simple covariance calculation: Eq. (1)) is to a significant extent determined by long-distance coupling. Note that, if the relation structure–alignments were dominated by contact interactions, we could expect the plot of residue–residue distance versus $|\sigma|$ to be L-shaped, with most of the residue–residue distances about a certain $|\sigma|$ threshold being very short.

Another surprising feature is revealed by a close examination of the plots of residue–residue distance versus $|\sigma|$, in particular the plot based on the distance between α-carbons (Fig. 1B). The distances corresponding to the higher $|\sigma|$ values seem to be organized roughly in layers corresponding to about 6, 12, 18, 24 and even 30 Å. These inter-residue distances could be viewed as the integer multiples of a typical α-carbon distance for residues in close contact (about 5–6 Å), thus suggesting long-distance communication through roughly linear networks of connected residues in the native state (or, perhaps, in a denatured state with some native-like residual structure).

3.2. Statistical description of the relation between structure and sequence alignments: application to thioredoxin

In order to further explore the relation between protein structure and sequence alignments we introduce here a simple statistical analysis based upon the following function:

$$\rho(r,\sigma^*)=\frac{N(r\pm\Delta r)}{2\cdot\Delta r}\{P(|\sigma|\geq\sigma^*|r\pm\Delta r)-P(|\sigma|\geq\sigma^*)\} \quad (2)$$

where $N(r\pm\Delta r)$ is the number of residue–residue distances (as measured using the α-carbons) within the $\{r-\Delta r,\ r+\Delta r\}$ range, $P(|\sigma|\geq\sigma^*)$ is the probability of finding couples of residues with absolute values of covariance equal or higher than a given threshold (σ^*), and $P(|\sigma|\geq\sigma^*|r\pm\Delta r)$ is the same probability but conditional to the fact that the residue–residue distances are in the $\{r-\Delta r,\ r+\Delta r\}$ range. Strictly speaking, ρ should be defined as the $\Delta r\rightarrow 0$ limit of the right-hand-side of Eq. (2) and thus it would have the meaning of a number density (see further below for details). In practice, however, a finite Δr value must be used in the calculation. The probabilities in Eq. (2) are calculated as,

$$P(|\sigma|\geq\sigma^*)=\frac{N(|\sigma|\geq\sigma^*)}{N} \quad (3)$$

$$P(|\sigma|\geq\sigma^*|r\pm\Delta r)=\frac{N(|\sigma|\geq\sigma^*,\ r\pm\Delta r)}{N(r\pm\Delta r)} \quad (4)$$

where N is the total number of residue couples ($108\cdot 107/2=5778$ for thioredoxin), $N(|\sigma|\geq\sigma^*)$ is the number of couples with $|\sigma|$ equal or greater than the threshold (σ^*), and $N(|\sigma|\geq\sigma^*,\ r\pm\Delta r)$ is the number of residue couples for which simultaneously $|\sigma|\geq\sigma^*$ and the residue–residue distance is in the $\{r-\Delta r,\ r+\Delta r\}$ range.

The difference between the two probabilities multiplied by $N(r\pm\Delta r)$ gives the number of couples with $|\sigma|\geq\sigma^*$ and within the $\{r-\Delta r,\ r+\Delta r\}$ range, in excess over the number expected if the probability of finding couples with $|\sigma|\geq\sigma^*$ did not depend on the distance range under consideration. For instance, the number of residue couples with absolute values of covariance larger than a threshold of $\sigma^*=0.01$ is 3331 and, consequently, $P(|\sigma|\geq 0.01)$ is $3331/5778=0.576$. The number of residue couples for which the residue–residue distance is within the 12 ± 0.5 Å interval is 290; if the probability of finding couples with $|\sigma|\geq 0.01$ did not depend on residue–residue distance, the probability value of 0.576 would apply to the 12 ± 0.5 Å interval and the number of couples with $|\sigma|\geq 0.01$ within that interval would be expected to be $0.576\cdot 290\cong 167$. However, the actual number is significantly higher (190), indicating an excess number of couples above the threshold of about 23. This is, of course, the same result that would be obtained by subtracting the two probabilities defined by Eqs. (3) and (4) and multiplying by the total number of couples within the distance interval [that is: $P(|\sigma|\geq 0.01\,|\,12\pm 0.5)=190/290=0.655$, $P(|\sigma|\geq 0.01)=3331/5778=0.576$, $N(12\pm 0.5\ \text{Å})=290$ and $290\cdot(0.655-0.576)\cong 23$]. Note that, although in this particular example the

excess number is positive, it may in general have either sign. The uncertainty associated to the calculated excess number of residue couples can be easily estimated by assuming binomial distributions for $N(|\sigma| \geq \sigma^*)$ and $N(|\sigma| \geq \sigma^*, \Delta r)$. For instance, the total number of residue couples is 5778 (number of trials for the binomial distribution), out of which 3331 have $|\sigma| \geq 0.01$ (number of successes); the success probability is then 0.576 and the standard deviation associated to $N(|\sigma| \geq 0.01) = 3331$ (see chapter 21 in Ref. [22]) is given by $[5778 \cdot 0.576 \cdot (1 - 0.576)]^{1/2} = 38$. A similar calculation yields $N(|\sigma| \geq 0.01 \,|\, 12 \pm 0.5) = 19 \pm 08$. Finally, using Eqs. (2)–(4) with standard error propagation procedures we obtain, for the example given here, $\rho = 23 \pm 8$.

In Eq. (2), the excess number is actually divided by the size of the distance range ($2 \cdot \Delta r$) in order to approach an excess number density (excess number of couples per Ångström). Rigorously, the number density would be given by the $\Delta r \to 0$ limit of Eq. (2). In practice, however, the estimated errors for ρ become very large for very low values of Δr. Most calculations reported here use Δr values of 5 and 0.5 Å, which we found to be an acceptable compromise between resolution and associated error (although, as expected, $\Delta r = 5$ Å produces smoother ρ versus r profiles).

Clearly, the outcome of the ρ calculation depends on the value chosen for the σ^* threshold. If $\sigma^* = 0$, the two probabilities (Eqs. (3) and (4)) become unity and the ρ value is zero for all distances. If σ^* is large, there will be few couples with $|\sigma| \geq \sigma^*$ and the ρ values will be necessarily low. We may expect the ρ values to be comparatively large (in absolute value) for some intermediate threshold. In order to determine the optimum value of $\sigma^* \cdot (\sigma_{opt})$, we calculate the integral of the absolute values of ρ:

$$\eta((\sigma^*)) = \int |\rho(r, \sigma^*)| \cdot dr \tag{5}$$

and take as σ_{opt} the σ^* value that maximizes $\eta(\sigma^*)$.

For practical calculations, we found convenient to define the value of σ^* in terms of the fraction of residue couples below the threshold:

$$f^* = \frac{N(|\sigma| < \sigma^*)}{N}. \tag{6}$$

Obviously, the value of f^* completely specifies that of σ^*. The inset in Fig. 2A shows the plot of η versus f^* for thioredoxin with $\Delta r = 5$ Å. A clearly defined maximum is observed for $f^* = 0.44$ (corresponding to $\sigma_{opt} = 1.03 \cdot 10^{-2}$) and $\eta = 137.5$, which indicates a total number of excess residue couples of the order of a hundred. This is, of course, a statistical result, thus, we cannot assign that excess number to specific residue couples. The plot of ρ versus r for this optimum value of the threshold (Fig. 2A) indicates that ρ values tend to be positive below 15 Å and negative above that residue–residue distance. That is, short residue–residue distances are favored over long ones for high values of absolute covariance (above the threshold). Note, nevertheless that the maximum of the profile is about 10 Å, a value clearly above the contact distance.

The profile of ρ versus r calculated for $\Delta r = 0.5$ Å (Fig. 2A) is similar to that obtained with $\Delta r = 5$ Å although, as was

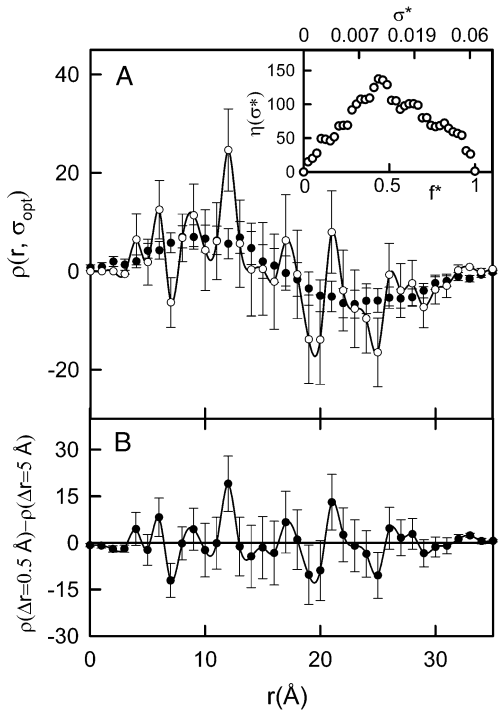

Fig. 2. Statistical analysis of the relation structure–alignments for *E. coli* thioredoxin on the basis of the ρ function (Eq. (2)). (Panel A) Plot of ρ value versus α-carbon inter-residue distance. The symbols refer to the width of the distance interval used in the calculation (closed symbols: $\Delta r = 5$ Å; open symbols: $\Delta r = 0.5$ Å). The error bars correspond to the uncertainties calculated assuming binomial distributions for $N(|\sigma| \geq \sigma^*)$ and $N(|\sigma| \geq \sigma^*, r \pm \Delta r)$. The value of σ^* used in the calculation is the optimum (see Table 1) given by the maximum of the plot of $\eta(\sigma^*)$ versus f^* shown in the Inset. (Panel B) Plot of the difference between the ρ values calculated with $\Delta r = 5$ Å and $\Delta r = 0.5$ Å versus inter-residue distance. The line connecting the points is a spline interpolation and is only meant to guide the eye.

to be expected, shows a higher apparent noise level. Interestingly, the difference between the two profiles (Fig. 2B) appears to be modulated, with maxima at about the preferred distances shown in Fig. 1B. An interesting possibility is then that some of the "noise" in the $\Delta r = 0.5$ Å profiles actually reflects the preferred distances for long-range interactions if these are transmitted through roughly linear networks of connected residues. Actually, we have undertaken an experimental mutational analysis into the residue–residue couplings suggested by Figs. 1B and 2B (work in progress).

3.3. Statistical analysis of the structure–alignments relation for a set of 24 proteins

In order to assess the general applicability of the structure–alignments analysis based on the ρ function, we have applied it, as described above, to a set of 24 proteins extracted from the 41 architecture representatives in the CATH database by eliminating (for calculation performance reasons) those with more than 400 residues (see Methods for further details). Names of pdb files, numbers of residues and other parameter of interest are collected for all the proteins analyzed here in Table 1. Plots of inter-residue distance

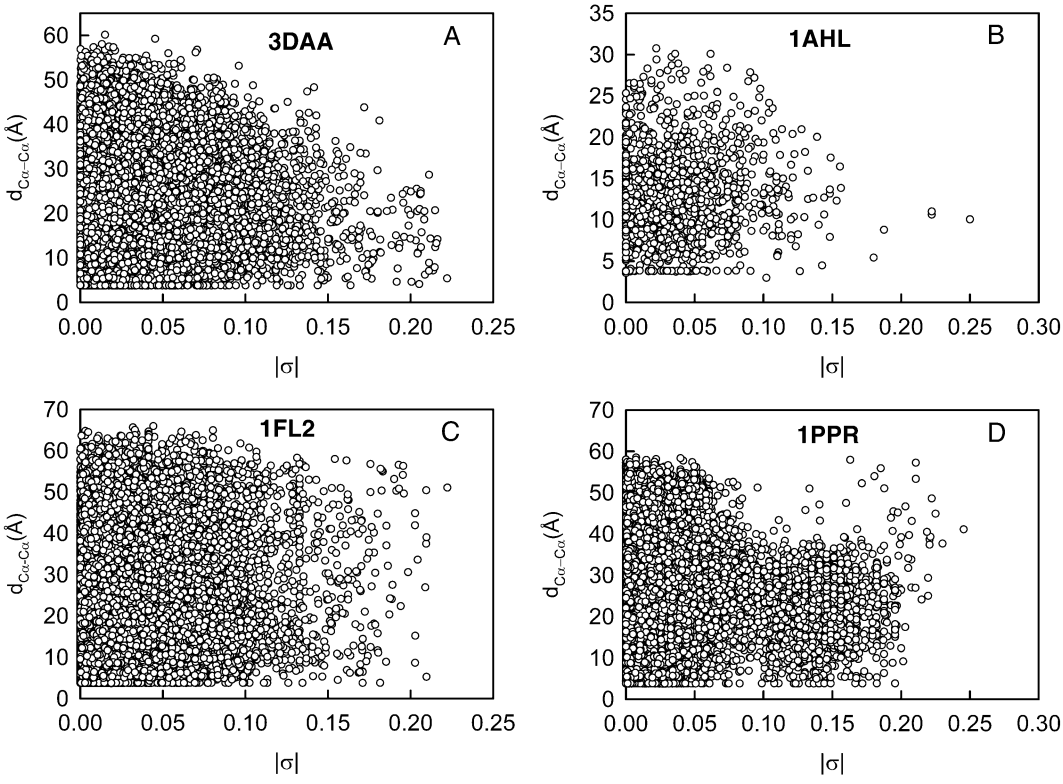

Fig. 3. Four illustrative examples of the plots of residue–residue distance versus absolute value of covariance obtained from the analysis of a set of 24 proteins extracted from the architecture representatives of the CATH database.

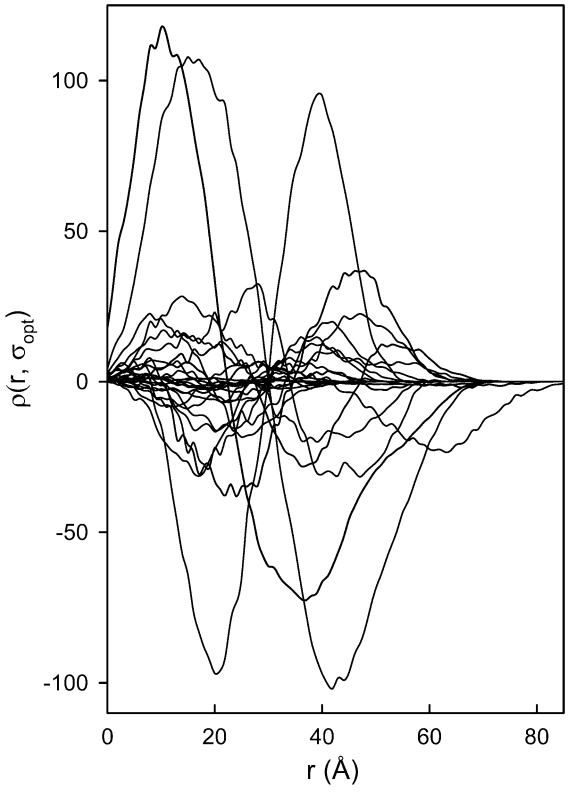

Fig. 4. Statistical analysis of the relation structure–alignments for a set of 24 proteins extracted from the architecture representatives of the CATH database. The profiles of ρ versus α-carbon inter-residue distance shown were calculated with $\Delta r = 5$ Å and the optimum value of σ^* according to the η criterion (Table 1). For the sake of clarity, associated uncertainties have been omitted.

versus absolute covariance for these proteins had a clearly triangular shape in some cases (but, not in all cases; see Fig. 3 for illustrative examples). The ρ/r profiles obtained with $\Delta r = 5$ Å and the optimum values of σ^* (given in Table 1) are collected in Fig. 4. It is interesting to note that there appears to be no strong correlation between the amount of information derived from the analysis [as measured by the $\eta(\sigma_{opt})$ value] and the size of the protein or the number of sequences in the alignment (Table 1). We observed the following four general types of behavior, which are illustrated by the specific examples given in Fig. 5 (where ρ/r profiles calculated with both $\Delta r = 5$ Å and $\Delta r = 0.5$ Å are shown): 1) For 6 proteins, the calculated ρ/r profile is flat within the uncertainty of the calculated ρ values (Fig. 5A). In this case, no information on the structure–alignments relation is derived from the analysis. 2) For 6 proteins, the profiles (Fig. 5B) are similar to that calculated for thioredoxin (Fig. 2); that is, a positive "peak" is observed for comparatively low distances and negative ρ values are obtained for the larger ones. 3) For 8 proteins the reverse behavior is observed (Fig. 5C): the positive ρ peak was observed for the larger distances and negative ρ values were obtained for the shorter ones. 4) For 4 proteins two separated positive "peaks" (at short and larger distances) are observed (Fig. 5D).

4. Concluding remarks

We have introduced in this work a simple statistical procedure to describe the relation structure–alignments based

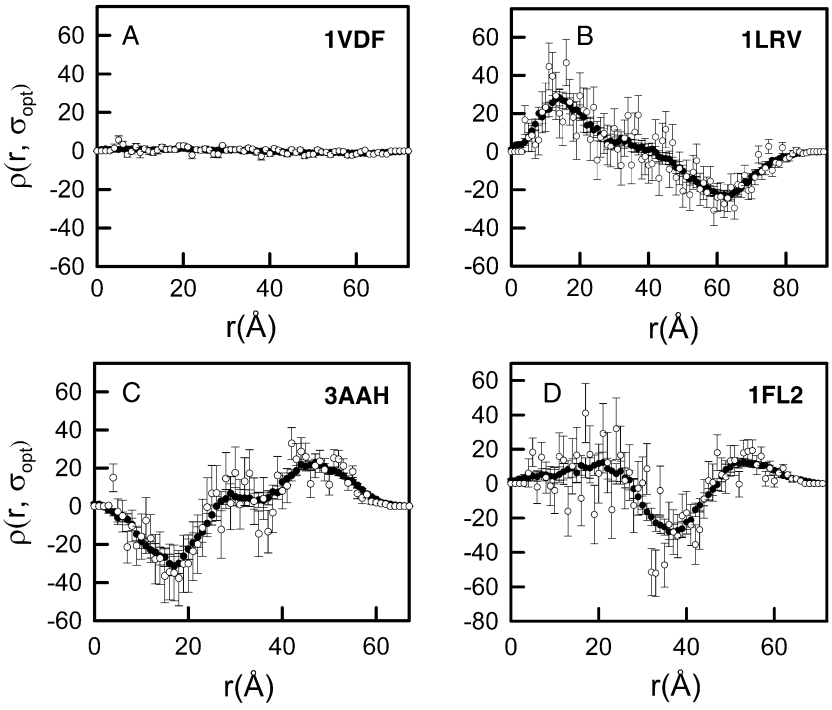

Fig. 5. Illustrative examples of the of the 4 different kinds of ρ versus inter-residue distance profiles found in the analysis of the set of 24 proteins extracted from the architecture representatives in the CATH database (Fig. 3). The symbols refer to the width of the distance interval used in the calculation (closed symbols: $\Delta r = 5$ Å; open symbols: $\Delta r = 0.5$ Å). The error bars correspond to the uncertainties calculated assuming binomial distributions for $N(|\sigma| \geq \sigma^*)$ and $N(|\sigma| \geq \sigma^*, r \pm \Delta r)$.

on the residue–residue distance dependence of the number of residue couples over given thresholds of a correlation measure. Thus, the ρ function (Eq. (2)) versus residue–residue distance profiles (Fig. 5) appear to provide clear pictures of the distance distribution of correlated changes in sequence alignments. We believe that this kind of statistical description of the structure–alignments relation may provide a simple but efficient tool for the exploration of the different structural features that are reflected in the sequence alignments. For instance, it could be argued that, in some cases, the statistical preference for correlations at the largest distances (see Figs. 4 and 5) could actually be due to the high level of conservation of a certain number of core residues (distances between these "central" residues and other residues in the protein will be comparatively shorter and covariance values for residue couples involving those residues will be comparatively smaller, due to the conservation). This structural feature could obviously be probed by eliminating highly conserved residues prior to covariance and ρ calculation. Along the same lines, one could multiply the covariance values by a function related with published values of inter-residue contact energies [23,24] in order to increase the weight of close interactions in the final statistical description. Also, with adequate definitions of the δ values in the covariance equation (Eq. (2)) it should be possible to investigate the distance distribution of correlations for specific couples of aminoacid types. In addition, other types of correlation measures (different than the simple covariance value defined by Eq. (1)) could be employed. Finally, different subsets of sequences may contain different structural information, a hypothesis which could be tested by, for instance, sequence clustering according to similarity prior to the statistical analysis.

Calculations of the kind described above may help to elucidate the structural features that can be extracted from sequence alignments and potentially lead to a set of sequence–alignment descriptors (such as η values) sufficiently sensitive to structure to be used in prediction.

Acknowledgements

This work was supported by Spanish Ministry of Science and Education Grant BIO2003-02229 and Feder Funds. Raul Perez-Jimenez is a recipient of a predoctoral fellowship from the Spanish Ministry of Science and Education.

References

[1] U. Gobel, C. Sander, R. Schneider, A. Valencia, Correlated mutations and residue contacts in proteins, Proteins 18 (1994) 309–317.
[2] I.N. Shyndialov, N.A. Kolchanov, C. Sander, Can three dimensional contacts in protein structures be predicted by analysis of correlated mutations?, Protein Eng. 7 (1994) 349–358.
[3] A. Aszódi, M.J. Gradwell, W.R. Taylor, Global fold determination from small number of distance restraints, J. Mol. Biol. 251 (1995) 308–326.
[4] A.R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, J. Skolnick, Ab initio folding of proteins using restraints derived from evolutionary information, Proteins (Suppl. 3) (1999) 177–185.
[5] A.R. Ortiz, A. Kolinski, J. Skolnick, Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 1020–1025.
[6] S.M. Larson, A.A. Di Nardo, A.R. Davidson, Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions, J. Mol. Biol. 303 (2000) 433–446.

[7] P. Fariselli, O. Olmea, A. Valencia, R. Casadio, Progress in predicting inter-residue contacts with neural networks and correlated mutations, Proteins (Supp 1) (2001) 157–162.
[8] O. Graña, V.A. Eyrich, F. Pazos, B. Rost, A. Valencia, EVAcon: a protein contact prediction evaluation service, Nucleic Acids Res. 33 (2005) W347–W351.
[9] S. Vicatos, B.V.B. Reddy, Y. Kaznessis, Prediction of distant residue contacts with the use of evolutionary information, Proteins 58 (2005) 935–949.
[10] H. Ohkata, A. Schön, E. Freire, E. Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations, Biochemistry 42 (2003) 13659–13666.
[11] G. Spudich, S. Lorenz, S. Marqusee, Propagation of a single destabilizing mutation throughout the *Escherichia coli* ribonuclease HI native state, Protein Sci. 11 (2002) 522–528.
[12] N.T. Uter, J.J. Perona, Long-range intramolecular signaling in a tRNA synthetase complex revealed by pre-steady-state kinetics, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 14396–14401.
[13] M.W. Clarkson, A.L. Lee, Long-range dynamic effects of point mutations propagate trough side chains in the serine protease inhibitor eglin c, Biochemistry 43 (2004) 12448–12458.
[14] J. Klein-Seetharaman, M. Oikawa, S.B. Grimshaw, J. Wirmer, E. Duchardt, T. Ueda, T. Imoto, L.J. Smith, C.M. Dobson, H. Schwalbe, Long-range interactions within a non-native protein, Science 295 (2002) 1719–1722.
[15] M. Guzman-Casado, A. Parody-Morreale, S. Robic, S. Marqusee, J.M. Sanchez-Ruiz, Energetic evidence for the formation of a pH-dependent hydrophobic cluster in the denatured state of *Thermus thermophilus* ribonuclease H, J. Mol. Biol. 329 (2003) 731–743.
[16] S. Robic, M. Guzman-Casado, J.M. Sanchez-Ruiz, S. Marqusee, Role of residual structure in the unfolded state of a thermophilic protein, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 11345–11349.
[17] R. Godoy-Ruiz, R. Perez-Jimenez, B. Ibarra-Molero, J.M. Sanchez-Ruiz, Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations, J. Mol. Biol. 336 (2004) 313–318.
[18] S.W. Lockless, R. Ranganathan, Evolutionary conserved pathways of energetic connectivity in protein families, Science 286 (1999) 295–299.
[19] G.M. Süel, S.W. Lockless, M.A. Wall, R. Ranganathan, Evolutionary conserved networks of residues mediate allosteric communication in proteins, Nat. Struct. Biol. 10 (2003) 59–69.
[20] C. Orengo, A.D. Michie, S. Jones, T.D. Jones, M.B. Swindells, J. Thornton, CATH—a hierarchic classification of protein domain structures, Structure 5 (1997) 1093–1108.
[21] N.V. Dokholyan, E.I. Shakhnovich, Understanding hierarchical protein evolution from first principles, J. Mol. Biol. 312 (2001) 289–307.
[22] E. Steiner, The Chemistry Math Book, Oxford University Press, Oxford, 1996.
[23] D.A. Hinds, M. Levitt, Exploring conformational space with a simple lattice model for protein structure, J. Mol. Biol. 243 (1994) 668–682.
[24] S. Miyazawa, R.L. Jeringan, Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues, Proteins 34 (1999) 49–68.