ELSEVIER

Biophysical Chemistry 119 (2006) 84 – 93

Biophysical Chemistry

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

Hierarchical clustering of the correlation patterns: New method of domain identification in proteins

S.O. Yesylevskyy [a,b,*], V.N. Kharkyanen [a], A.P. Demchenko [b,c]

[a] Department of Biophysics, Institute of Physics, National Academy of Science of Ukraine, Prospect Nauki, 46, Kiev-03039, Ukraine
[b] A.V. Palladin Institute of Biochemistry, Leontovicha st. 9, Kiev-01030, Ukraine
[c] Research Institute for Genetic Engineering and Biotechnology, Tubitak, Gebze-Kocaeli, 41470, Turkey

Received 21 April 2005; received in revised form 4 July 2005; accepted 4 July 2005
Available online 25 August 2005

Abstract

New method of identification of dynamical domains in proteins — Hierarchical Clustering of the Correlation Patterns (HCCP) is proposed. HCCP allows to identify the domains using single three-dimensional structure of the studied proteins and does not require any adjustable parameters that can influence the results. The method is based on hierarchical clustering performed on the matrices of correlation patterns, which are obtained by the transformation of ordinary pairwise correlation matrices. This approach allows to extract additional information from the correlation matrices, which increases reliability of domain identification. It is shown that HCCP is insensitive to small variations of the pairwise correlation matrices. Particularly it produces identical results if the data obtained for the same protein crystallized with different spatial positions of domains are used for analysis. HCCP can utilize correlation matrices obtained by any method such as normal mode or essential dynamics analysis, Gaussian network or anisotropic network models, etc. These features make HCCP an attractive method for domain identification in proteins.

Keywords: Domain identification; Hierarchical clustering; Correlation patterns

1. Introduction

Motions of the protein domains play a crucial role in different biological processes ranging from enzyme catalysis and ligand binding to mechanical movement of the motor proteins. Thus identification of protein domains and their motions is an important task, which is addressed in a number of studies. There are several definitions of the protein domains, which are connected with different methods for their identification. Definition of domains can be based on their independent folding [1], sequence motifs [2], presence of distinct hydrophobic cores [1], functional activity [3,4], contact classification [5], topology [6] and many other characteristics. These definitions are based on the static properties of the protein and thus correspond to so-called *structural domains*. Alternatively, domains can be defined according to their dynamic properties. *Dynamic domains* are the parts of the protein molecule, which maintain structural integrity during the large-scale conformational changes and perform relatively independent collective dynamics (e.g. exhibit rigid-body-like motions) during such changes. The correspondence of structural and dynamic domains was shown experimentally for some proteins [7], however this feature cannot be considered universal. In this study we will focus on the dynamic domains.

There are several methods of dynamic domains identification [8–11,17], which differ by their theoretical robustness, computational complexity and the amount of information extracted. They can be roughly classified into two categories: the methods, which rely on the comparison of two different protein conformations [8], and those, which are based on the single conformation [8,9,18]. The methods

* Corresponding author. Department of Biophysics, Institute of Physics, National Academy of Science of Ukraine, Prospect Nauki, 46, Kiev-03039, Ukraine. Tel.: +38 44 5259851.
E-mail address: yesint3@yahoo.com (S.O. Yesylevskyy).

0301-4622/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2005.07.004

of the first group are usually quite accurate in identifying dynamic domains and quite cheap computationally, but they are applicable directly to the very limited number of proteins, which can be crystallized in two different conformations.

The methods of the second group estimate the dynamical properties of the globule from single conformation and thus seem to be applicable for the majority of proteins. These methods typically involve various modifications of the normal mode analysis [12–14,17] or essential dynamics analysis [19], followed by grouping of similarly moving residues into the dynamic domains. These techniques are usually quite intensive computationally. To reduce the computational demands various simplifications are introduced on the stage of the normal mode analysis. Usually the only C_α atoms are considered and the simplified force fields are used [15–17].

Recently the Gaussian Network Model (GNM) approach and its modification — Anisotropic Network Model (ANM) became very popular in determining the character of motions in the folded proteins [10,12]. GNM describes the protein as a network of harmonic springs, which connect the C_α atoms of the residues in the close spatial proximity regardless of their positions in the sequence. Normal modes of such elastic network can be computed easily. GNM can be viewed as an extremely simplified version of the normal mode analysis (NMA), where realistic potentials of the atom–atom interactions are substituted by the purely harmonic potentials [10]. It is shown, that GNM describes the motions of the folded proteins surprisingly well and produces the results, which are often indistinguishable from those of full-scale NMA. Not only the normal modes of the system, but also the cross-correlations between the motions of different residues can be easily calculated in GNM. The later can be used for domain identification by finding groups of residues, whose movement is strongly correlated.

In the present study we introduce the method of domain identification, which is based on the Hierarchical Clustering of Correlation Patterns (called HCCP hereafter). It is shown that our method identifies dynamical domains in the folded proteins successfully. The method allows obtaining identical results when different conformations of the same protein are analyzed (in the case of domains, which maintain their structural integrity during conformational changes). This means that it is quite insensitive to the small details of the protein structure and the domains are resolved reliably regardless of their positions and orientation. HCCP is applicable to any kind of correlation or covariation matrices produced, for example, by MD simulations, essential dynamics calculations, GNM or ANM models. HCCP is computationally cheap — the protein with up to 1000 amino acids can be analyzed in few minutes on the ordinary office PC. These features make HCCP quite attractive in identifying dynamical domains in the proteins with known structure.

2. Methods

2.1. The Gaussian network model

As it was described above, GNM treats the protein as a network of harmonic springs, which connect C_α atoms of the residues in the close spatial proximity. The detailed description of the GNM can be found elsewhere [10]. Here we present only the aspects, which are crucial for the further analysis. Since only C_α atoms are considered, the configurational partition function for a protein of N residues is defined as

$$Z_N = \int \exp\left(-\frac{E(\mathbf{R})}{k_B T}\right) d\mathbf{R}$$

where $\mathbf{R}$ is $3N$-dimensional vector of the positions of the C_α atoms, E is the potential energy of the protein. GNM considers only small harmonic fluctuations of the C_α atoms, which means that E can be approximated by a sum of harmonic potentials representing the fluctuations of individual residues. Thus partition function is expressed in GNM as

$$Z_N = \int \exp\left(-\frac{\gamma}{2k_\mathrm{B}T}\{\Delta\mathbf{R}^\mathbf{T}\}\Gamma\{\Delta\mathbf{R}\}\right) d\{\Delta\mathbf{R}\}$$

where ΔR is $3N$-dimentional vector of the fluctuations in the positions of individual residues, superscript T denotes the transpose (swapping of the rows and columns of the matrix), k_B is the Boltzmann constant, T is the absolute temperature, γ is the spring elasticity constant of harmonic residue–residue interactions, Γ is the $N \times N$ Kirchoff matrix. Individual elements of Γ are defined as

$$\Gamma_{ij} = \begin{cases} -1 & \text{if} \quad i \neq j \quad \text{and} \quad \mathbf{R}_{ij} \leq r_c \\ 0 & \text{if} \quad i \neq j \quad \text{and} \quad \mathbf{R}_{ij} > r_c \\ -\sum\limits_{i \neq j} \Gamma_{ij} & \text{if} \quad i \neq j \quad \text{and} \end{cases}$$

where r_c is the cut-off distance for residue–residue interactions. In the original formulation of GNM [10] r_c=7.0 Å. This value is obtained from statistical analysis of PDB (Protein Data Bank) entries as the averaged value of the radius of the first interaction shell of the individual "averaged" residue. Normal modes of the system are obtained by the eigenvalue decomposition of Γ matrix

$$\Gamma = \mathbf{U}\Gamma\mathbf{U}^\mathbf{T}$$

where $\mathbf{U}$ is an orthogonal matrix. Its columns are the eigenvectors u_i of Γ, which are usually sorted in ascending order according to the corresponding eigenvalues λ_i. Each eigenvector u_i reflects the shape of corresponding i-th normal mode, while the eigenvalue λ_i is the frequency of this mode.

The most important for us is the cross-correlation between the motions of individual residues. Covariation between the motions of two residues can be expressed as the sum of contributions from $N-1$ normal modes with non-zero eigenvalues

$$< \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j > = (3k_BT/\gamma)\sum_{k=1}^{N-1}\left[\lambda_k^{-1}\mathbf{u_k}\mathbf{u_k^T}\right]_{ij}$$

Correlation coefficient c_{ij} between the motions of the residues i and j is then expressed as

$$c_{ij} = \frac{< \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j >}{\sqrt{< \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_i > \cdot < \Delta\mathbf{R}_j \cdot \Delta\mathbf{R}_j >}}$$

If $c_{ij}=1$, then the motions of the residues i and j are perfectly correlated, and if $c_{ij}=-1$ they are perfectly anticorrelated; and if $c_{ij}=0$ they are completely uncorrelated.

In principle c_{ij} is sufficient for identifying the dynamic domains in the system. Indeed, the domain can be defined as the set of residues, which exhibits collective motions as a single unit i.e. in a highly correlated manner. Thus c_{ij} should be close to 1 for all the pairs inside the domain and close to 0 or negative for all the pairs, which involve one residue within the domain and another one outside the domain.

2.2. The hierarchical clustering of correlations (HCC)

The hierarchical clustering procedure can be used to identify the domains. Hierarchical clustering is the mathematical technique, which is routinely used to classify some objects according to their "similarity". The "similarity" could be any function that shows quantitatively the extent to which two objects are similar to one another. There are many modifications of the hierarchical clustering adapted for particular applications. In the present work we used the modified agglomerative clustering scheme with the average linkage. In this scheme the most similar clusters are merged (agglomerated) on each step to produce larger clusters. Pairwise similarity criterion is applied to all inter-cluster pairs and then averaged to calculate the similarity between the clusters. The details of the clustering algorithm are the following:

1. Each amino acid residue of the protein is assigned to be the simplest cluster of size 1.
2. Minimal $v_{\min}$ and maximal $v_{\max}$ elements of c_{ij} are found. The interval $(v_{\min}:v_{\max})$ is divided into M bins $v_{\max}>v_1>v_2>\ldots>v_{M-1}>v_{\min}$ ($M=1000$ in this study). The index of the current bin is set to $k=1$.
3. The pair of residues whose correlation $c_{ij}>v_k$ is found. If no such pairs exist, then index of the current bin k is increased by 1 and step 3 is repeated.
4. Residues from the matching pair are merged into single cluster.
5. Correlation matrix c_{ij} is recalculated by the following rule:

   $$c_{ij} = \frac{1}{m_im_j}\sum_{k\in\{M_i\}}\sum_{l\in\{M_j\}}c_{kl}$$

   where m_i and m_j are the numbers of elements in the clusters i and j; M_i and M_j are the vectors of sizes m_i and m_j, respectively, which contain the indexes of the residues in these clusters. In other words the average correlation of all inter-cluster pairs is calculated.
6. The step 3 is continued until all the residues are merged and the whole protein becomes the single cluster.

Because the values of the pairwise correlations (elements of c matrix) are used in the clustering procedure we call this procedure Hierarchical Clustering of Correlations (HCC).

Domains can be identified as large clusters, which are merged on the last few steps of the clustering procedure. It is necessary to emphasize the difference between the terms "cluster" and "domain" used here. The clusters are "temporary" objects. They are agglomerates of the similarly moving residues, which appear on particular step (hierarchical level) of the clustering procedure. The number of clusters in the system decreases on each step of hierarchical clustering. Domains are the functional units of the real protein, which exhibit rigid body-like motions and which are the results of our search. Domains are the largest dynamic "blocks" in the protein that exhibit high correlation of motions. In terms of hierarchical clustering, domains correspond to the largest clusters of the very last hierarchical level.

2.3. The hierarchical clustering of the correlation patterns (HCCP)

As it is stated above, the matrix of the pairwise correlations can be used for domain identification. However, it has one serious drawback. Matrix c_{ij} contains only pairwise correlations. Thus only the motions of two residues are compared to each other, regardless to the motion of the rest of the protein. This makes the clustering procedure quite sensitive to small variations of the c_{ij} matrix. In the case of GNM, small changes of the protein structure can lead to the changes in the eigenvectors, which results in a different c_{ij} matrix. The overall structure of the matrix remains essentially the same, but individual values can change significantly. As a result the same residue can be assigned to different cluster according to this changed value of the pairwise correlation. In other words stability of the HCC algorithm is not very high. It is sensitive to small variations of the input data.

In order to improve the stability of the clustering procedure we considered the *correlation patterns* instead of pairwise correlations.

Let us consider single k-th column (or row) of c_{ij} matrix. It contains the correlations of the given residue k with all other residues in the system (including self-correlation, which is always 1). We will call such column-vector the *correlation pattern* of the residue k. The new matrix, the *correlation matrix of correlation patterns* p_{ij} can be defined as

$$p_{ij} = \frac{\frac{1}{N}\sum_{k=1}^{N}c_{ik}\cdot c_{jk} - \bar{c}_i\cdot\bar{c}_j}{\sigma_i\sigma_j}$$

where $\bar{c}_i$ is the mean of the i-th column of the matrix c, σ_i is the root mean square deviation of the i-th column of the matrix c. p_{ij} is the $N \times N$ matrix, whose elements show to what extent the correlation patterns i and j are similar in terms of linear correlation. The matrix p_{ij} provides much more robust way of comparing the motions of residues than does the conventional correlation matrix c_{ij}. Comparing the correlation patterns one compares the whole set of correlations of two given residues with the rest of the protein, but not only the pairwise correlation between them. Thus, if this analysis is performed for two different conformations of the same protein the results should be much more similar than the results obtained with the correlation matrix c_{ij}. Clustering procedure in this case remains the same as described above with the c_{ij} substituted by p_{ij} everywhere. We call this procedure Hierarchical Clustering of Correlation Patterns (HCCP).

HCC and HCCP analysis were performed using our own software written on Perl 5 and Fortran 90. Cut-off of 7.0 Å was used for Kirchoff matrix construction.

2.4. Test proteins

In order to test and compare the HCC and HCCP procedures we considered five test proteins: the lysine-, arginine-, ornithine-binding protein (LAO) from *Salmonella typhimurium* [20]; the glutamine binding protein (GLNBP) from *E. coli* [21]; phosphoglycerate kinase from *Trypanosoma brucei* [22]; human calmodulin [23] and 5-enolpyruvylshikimate-3-phosphate synthase from *Streptococcus pneumoniae* [24]. All test proteins consist of two well-defined domains, which undergo significant hinge-bending motion during the binding of ligand or substrate. For each protein the crystal structures of both "closed" (with the ligand) and "open" (without ligand) forms are available.

In the cases of GLNBP, phosphoglycerate kinase and calmodulin the number of resolved residues in the crystal structures of the open and closed forms is different. In order to simplify the comparison we truncated the longer structures, so that both forms contain the same number of residues. We extracted the Cartesian coordinates of the C_α atoms from both conformations of all proteins (the ligands were not included) and performed HCC and HCCP analysis on them. Domains were identified as the clusters on the last step of clusterization procedure.

3. Results

The correlation matrices c and correlation of the correlation patterns matrices p are shown in Fig. 1 for LAO protein and in Fig. 2 for GLNBP protein. Similar results are obtained for other studied proteins (data not shown).

Domains can be identified visually as large light "squares", which contain the residues, whose motion is well correlated. First domain is split into two parts because it is not continuous in sequence — it contains both N and C terminal ends of the protein, while the central part of the sequence forms the second domain. Dark areas contain the

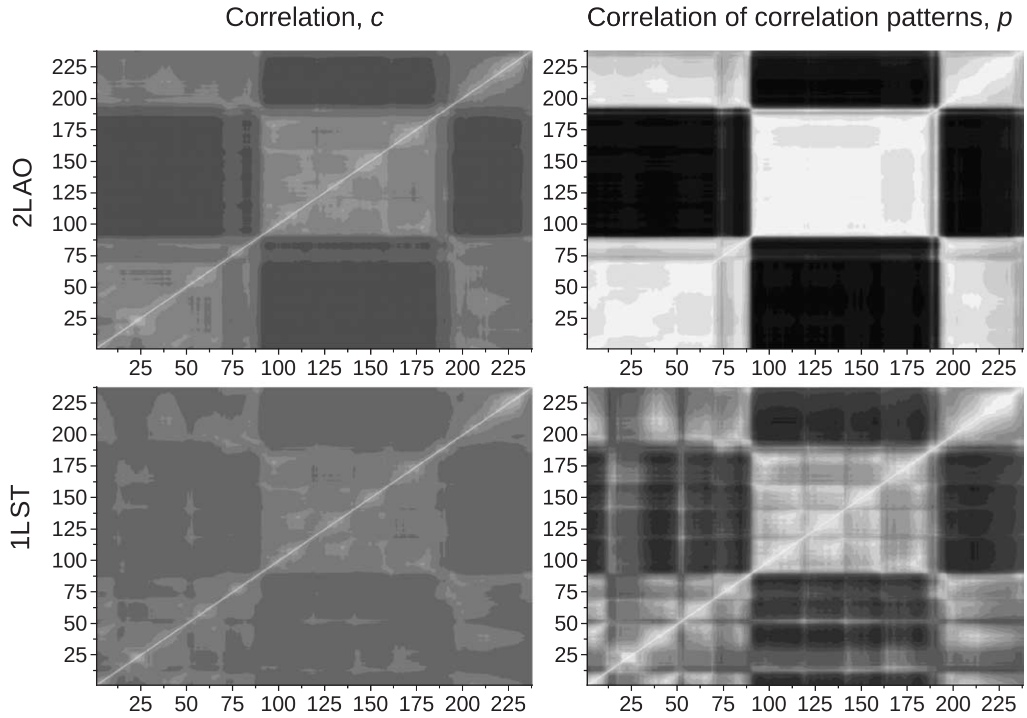

Fig. 1. c and p matrices for closed (2LAO) and open (1LST) forms of LAO protein. Color scale is from −1 (black) to 1 (white).

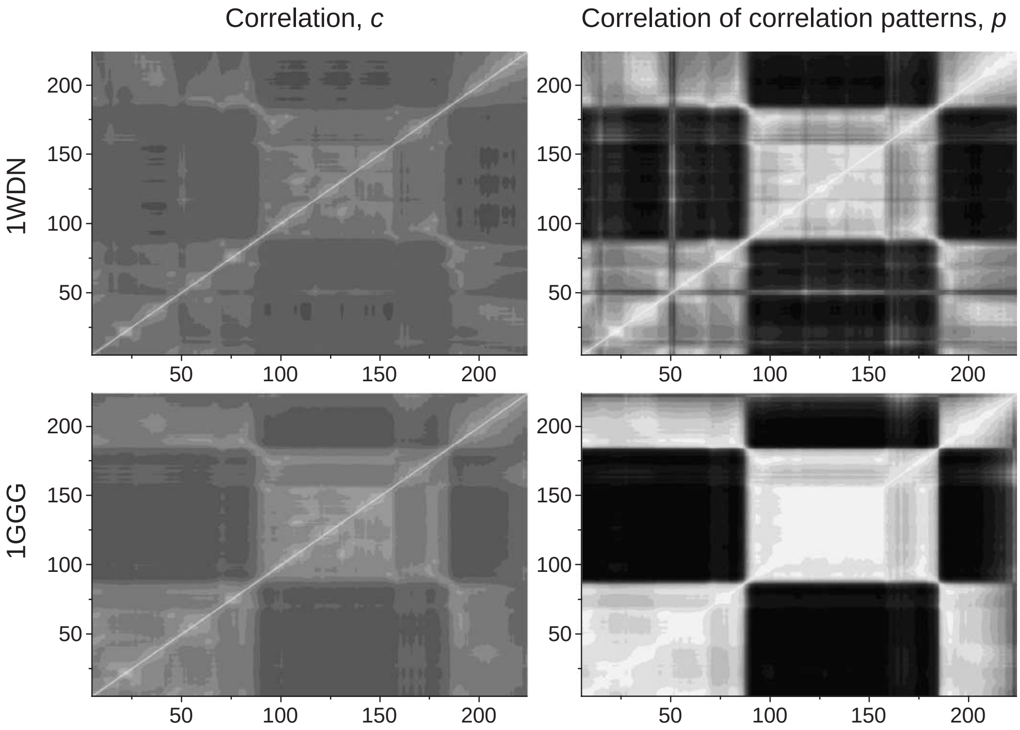

Fig. 2. c and p matrices for closed (1WDN) and open (1GGG) forms of GLNBP protein. Color scale is from −1 (black) to 1 (white).

pairs of residues from different domains, whose motion is in general anticorrelated. The diagonals of the matrices contain self-correlations, which are always 1. Both c and p matrices posses very similar features, however there is one significant difference. The values of c matrices are closer to zero (except the diagonal). Thus correlated and anticorrelated regions are not very well separated, which is seen visually as a "dim" picture. In contrast, the values of p matrices are close to -1 or 1. As a result, the correlated and anticorrelated regions are very well separated, which can be seen as essentially "black and white" picture with much higher visual contrast.

This difference becomes especially prominent if one calculates the frequency counts of values found in c and p matrices. To do so the range of values found in the matrix is divided into large number (100–200) of equally spaced

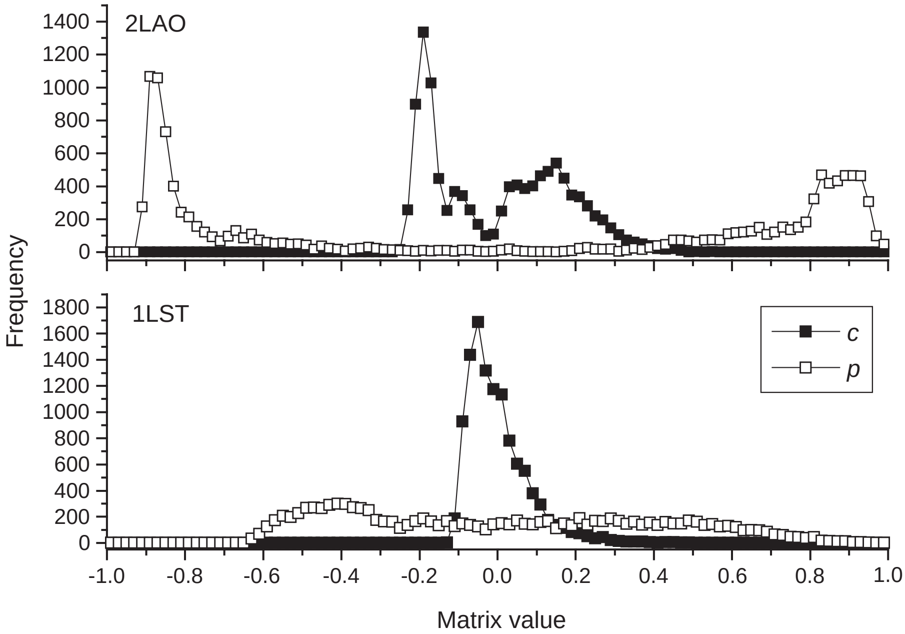

Fig. 3. Frequency counts for c and p matrices of LAO protein.

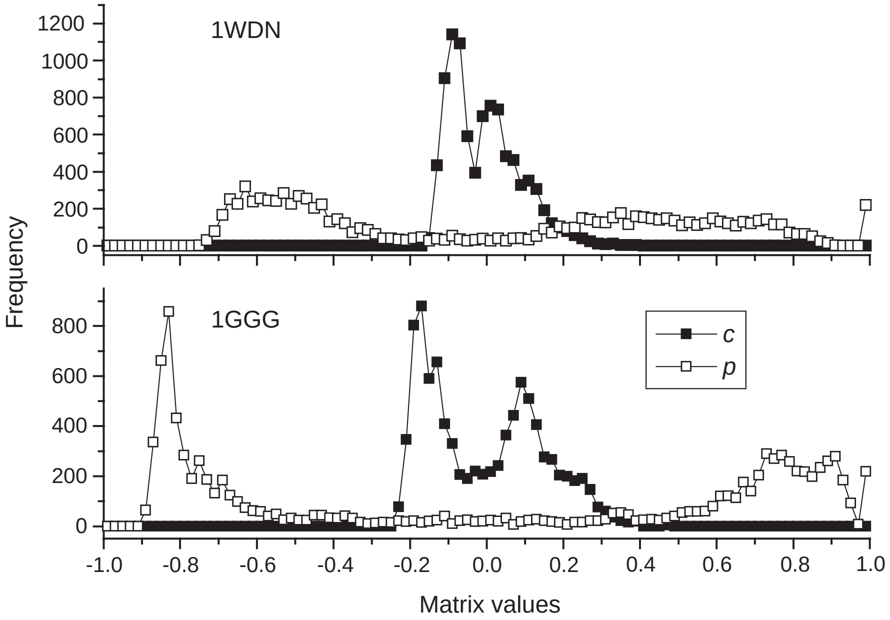

Fig. 4. Frequency counts for *c* and *p* matrices of GLNBP protein.

bins. Frequency counts are the histograms, which show the number of matrix elements, whose values fall into each bin. In other words, the histogram approximates the distribution function of the values of matrix elements. The frequency counts are shown in Fig. 3 for LAO protein and in Fig. 4 for GLNBP protein.

Peaks on the frequency count plots correspond to either highly correlated domain regions (positive values) or anticorrelated regions (negative values). We will call the distance between the peaks the *correlation gap* (CG). In the case of closed form of LAO (2LAO) the separation between these two peaks is rather small — CG ~0.4 for *c* matrix and ~1.8 for *p* matrix (Fig. 3). In the case of the open form of LAO (1LST), the peaks are not well resolved. For *c* matrix there is a single sharp peak near 0, while for *p* matrix there is a broad distribution of values with very poorly resolved maxima at ~ − 0.5 and ~0.6. The picture is very similar in the case of GLNBP. In the case of 1WDN

Table 1
Results of HCC and HCCP analysis

Protein	Clustering Method	Pdb structure	1st domain	2nd domain	Difference
Lysine-, arginine-, ornithine-binding protein	HCC	1LST	1–90; 186–238	91–185	8
		2LAO	1–89; 193–238	90–192	
	HCCP	1LST	1–89; 192–238	90–191	1
		2LAO	1–90; 192–238	91–191	
Glutamine binding protein (GLNBP)	HCC	1WDN	5–92; 188–223	93–187	2
		1GGG	5–90; 188–223	91–187	
	HCPP	1WDN	5–90; 188–223	91–187	0
		1GGG	5–90; 188–223	91–187	
Phosphoglycerate kinase	HCC	13PK	5–206; 406–418	207–405	20
		16PK	5–116; 170–193; 410–418	167–169; 194–409	
	HCCP	13PK	5–207; 409–418	206–408	10
		16PK	5–195; 409–418	196–408	
Calmodulin	HCC	1CLL	5–85	85–145	10
		1CDL	5–74	75–145	
	HCCP	1CLL	5–79	80–145	4
		1CDL	5–75	76–145	
5-enolpyruvyl-shikimate-3-phosphate synthase	HCC	1RF5	1–18; 230–427	19–229	9
		1RF6	1–16; 233–410; 415–427	17–232; 411–414	
	HCCP	1RF5	1–17; 230–427	18–229	0
		1RF6	1–17; 230–427	18–229	

The differences in positions of domain boundaries calculated by different methods are shown.

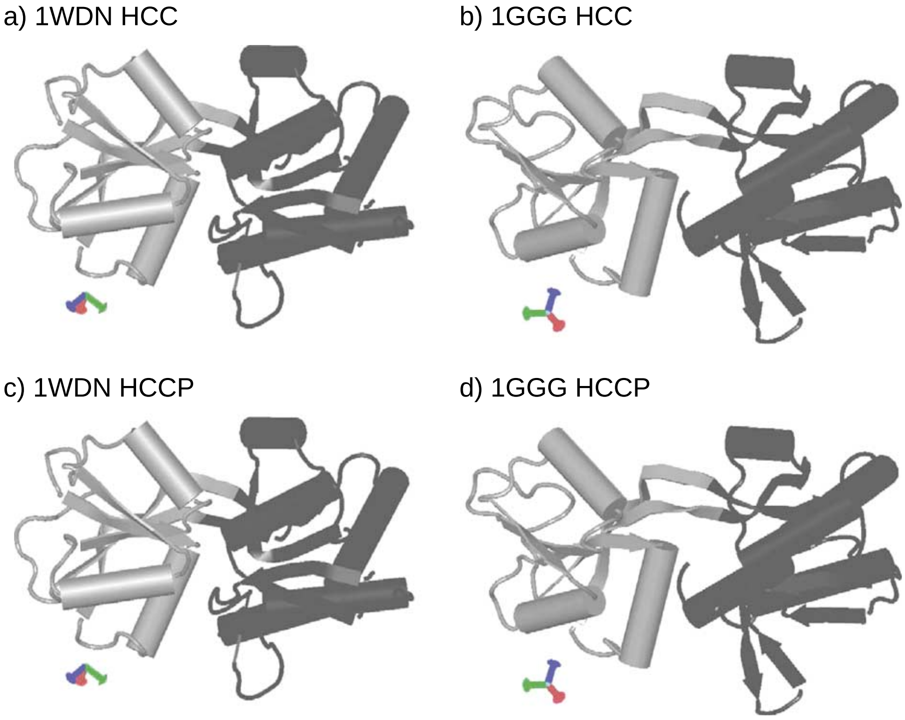

Fig. 5. Domains found by HCC and HCCP algorithms in closed (a, c) and open (b,d) forms of GLNBP protein. PDB code of the shown structure and method used for domain identification are indicated. First cluster is marked black, second cluster is gray.

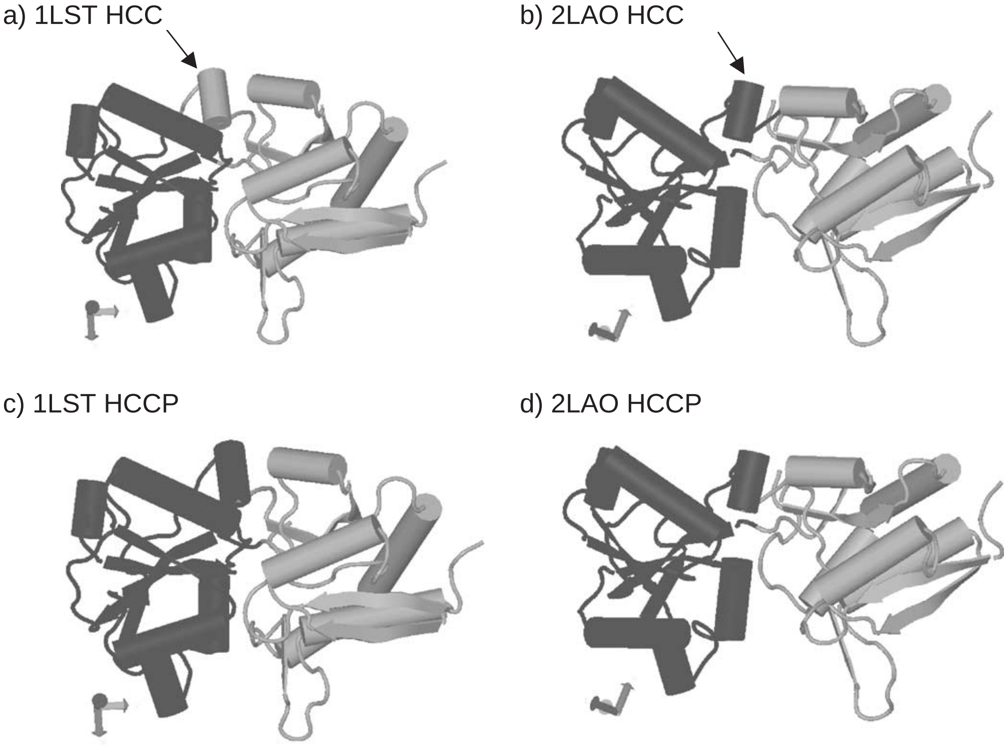

Fig. 6. Domains found by HCC and HCCP algorithms in closed (a,c) and open (b,d) forms of LAO protein. PDB code of the shown structure and method used for domain identification are indicated. First cluster is marked black, second cluster is gray. The arrow marks the helix, which is assigned to different clusters by HCC in open and closed forms.

two peaks are very close to 0 for *c* matrix (CG ~0.2), but there are two very broad well separated peaks for *p* matrix (CG ~1.3). In the case of 1GGG the peaks are more distinct and sharp (CG ~1.8).

In general, it is possible to conclude that the major effect of transition from correlations *c* to correlation patterns correlations *p* is a dramatic change of the distribution of values found in the matrices. As a result, the peaks of these distributions, which correspond to the groups correlated and anticorrelated residue pairs, become much more separated and CG increases. The influence of this fact on the clusterization procedure will be discussed below.

The results of application of hierarchical clustering analysis are summarized in the Table 1. It is clearly seen that in all the cases two domains in the protein structure are identified reliably. However, the exact position of the boundary between domains depends on the structure and the clustering technique. We compared the results obtained on different structures by the same clustering method. The difference in the position of the domain boundary (the number of residues, which belong to different domains if different starting structures are considered) is shown in Table 1. It is evident that in the case of HCC the position of the domain boundary depends strongly on the starting structure. However, in the case of HCCP this position is much less structure-sensitive. In the cases of GLNBP and 5-enolpyruvylshikimate-3-phosphate synthase on application of HCCP the absolutely identical domains are obtained for open and closed structures. The largest discrepancy for the position of domain boundary is observed in phosphoglycerate kinase, however, even for this protein it is two times smaller in the case of HCCP.

The domains for GLNBP, LAO and phosphoglycerate kinase are visualized in Figs. 5 and 6. It is clear that in the case of GLNBP both HCC and HCCP produce almost identical results (Fig. 5 a–d). Two mismatched residues 91 and 92 lie in the hinge region connecting two domains. This discrepancy is quite tolerable. However, the situation is different in the case of LAO (Fig. 6). If HCC clustering is performed on the open structure (Fig. 6b), the small helix near the hinge region, marked by an arrow, is assigned to the first domain (black). The same helix is assigned to the second domain (gray) if the closed structure is used (Fig. 6a). In the case of HCCP clustering, the helix is assigned to the first domain regardless of the used structure. Similar incorrect assignment in the case of HCC is observed in phosphoglycerate kinase for the residues 167–169 and in 5-enolpyruvylshikimate-3-phosphate synthase for the residues 411–414 (structures not visualized).

These examples show, that HCC clustering can lead to severe discrepancies in the domain pattern if different starting structures are used for analysis. On the other hand HCCP clustering produces much more consistent results, which are not strongly dependent on the structure used.

4. Discussion

As it was stated in the introduction, there are many methods of domain identification available. However, no systematic studies were performed to investigate the stability of domain identification if different conformations of the same protein are used for analysis (if the domains remain essentially unchanged in different conformations). If a particular method does not produce sufficiently similar results for the protein states, which differ only in positions of domains, it is dangerous to use it for the analysis of new proteins since the validity of domain identification becomes questionable.

In the current study we used five proteins of different size and spatial design, which are known to consist of two well separated domains undergoing substantial hinge-bending motion. These proteins are often referred as “classical” examples of the hinge-bending proteins. All these proteins are crystallized in open and closed forms, which allow comparing the results of domain identification based on different conformations of the same protein. This was the reason for their choice in our study.

The major prerequisite of precise and reliable domain identification by the clustering algorithm is a large difference in correlation coefficient between the residues of the same domain and the residues from different domains. This difference can be characterized by the correlation gap (CG) parameter introduced above. If CG is large, then small random variations in the correlation matrix are unlikely to be comparable with CG and thus they cannot influence the results of clustering procedure. In contrast if CG is small, then small random variations become comparable with it and can distort the result of clustering procedure. Two different conformations of the protein produce slightly different correlation maps. These differences can be viewed as small random variations described above. Therefore our first goal was to analyze the matrices *c* and determine CG for both open and closed conformations. In this context we will discuss the representative results for LAO and GLNBP proteins. Results for other studied proteins are quite similar. It is clear from Figs. 1 and 2 that the conformational changes do not lead to qualitatively important changes of *c* matrices, however the quantitative changes are significant. In the case of closed form of GLNBP correlated and anticorrelated parts of the matrix are not well separated (CG is close to 0.4, Fig. 4). The situation is even worse in the case of the open form of LAO, where these parts are not separated at all (CG = 0). Only one peak is observed in the frequency count histogram, which means that both positive and negative correlations are quite weak and there is no CG. This situation is quite unfavorable for clustering algorithm. Thus, it is quite expectable that HCC procedure produces different results for closed and open forms of LAO (Fig. 5). In the case of GLNBP the correlation gap exists for both open (CG ~0.3) and closed (CG ~0.1) forms. As a result the difference in clustering results is smaller than in the case of LAO. This shows that the value of CG is essential for reliable domain identification.

Construction of p matrices leads to dramatic increase of CG (Figs. 3 and 4). This is a direct consequence of the fact, that p matrix contains not only pairwise correlations but also an information about whole correlation patterns for all residue pairs. As a result, one should expect that HCCP, which is based on p matrices, produces almost identical results for both open and closed forms of the studied proteins, which is really so.

There is an additional advantage of HCCP algorithm, which makes it attractive for domain identification in unknown proteins. It contains no adjustable parameters (except the cut-off distance, which is a part of the GNM, but not of the domain identification scheme). Thus the results based on the application of HCCP cannot be biased by inadequate choice of empirical parameters.

It necessary to emphasize the difference of HCCP from recently developed GNM-based method of automatic domain decomposition of Kundu et al. [18]. The latter is based on the analysis of single eigenvector, which corresponds to the lowest non-zero eigenvalue of GNM. The shape of this eigenvector allows to detect the structural regions, which move in opposite directions along the slowest normal mode and assign them to different clusters. These clusters are post-processed ("filtered") to find the domains. Being very simple and intuitive, Kundu's method has several serious limitations:

1) it is limited to GNM;
2) only one normal mode is considered, which leads to considerable loss of information;
3) the analysis is qualitative — only direction but not the amplitude of motion is used for domain detection, thus the degree of internal correlation of motions in the cluster cannot be estimated;
4) no hierarchical features, like rigid sub-domains, can be found;
5) the filters applied to initial clusters contain many adjustable parameters.

HCCP does not possesses these limitations:

1) it is based on the pair-correlation matrices of any origin and thus is not limited to GNM;
2) all normal modes are accounted if GNM is used to form the pair-correlation matrices;
3) the analysis is quantitative — not only the sign, but also the value of correlation is used for clustering.
4) hierarchical clustering allows to detect sub-structures of different levels and estimate their rigidity in terms of internal correlations;
5) no post-processing and adjustable parameters are needed.

It is shown that HCCP is quite insensitive to the starting conformation of the studied protein, providing that the domains really maintain their integrity in all the cases of change of protein conformation. Thus HCCP really finds structurally separated domains, despite the masking effect of numerous insignificant features of particular conformation. However, it is possible that domain boundary determined by HCCP changes significantly in different conformations of particular protein. In this case we have to suggest that the protein under study does not have well-defined domains, which can be treated as independently moving structural blocks. In this sense HCCP gives a unique opportunity to identify not only the domains themselves but also the extent of domain independence. All this makes HCCP an attractive and computationally cheap method of the analysis of protein structure.

References

[1] J. Janin, S.J. Wodak, Domains in proteins: definition, location and structural principles, Methods Enzymol. 115 (1985) 420–430.
[2] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hofmann, A. Barioch, The PROSITE database, its status in 2002, Nucleic Acids Res. 30 (2002) 235–238.
[3] B. Nagar, W.G. Bornmann, P. Pellicena, T. Schindler, D.R. Veach, W.T. Miller, B. Clarkson, B. Kuriyan, Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571), Cancer Res. 62 (2002) 4236–4243.
[4] L. Schmitt, R. Tampe, Structure and mechanics of ABC transporters, Curr. Opin. Struck. Biol. 12 (2002) 754–760.
[5] K.F. Fischer, S. Marqusee, A rapid test for identification of autonomous folding units in proteins, J. Mol. Biol. 302 (2000) 701–712.
[6] C. Anselmi, G. Bocchinfuso, A. Scipioni, P. De Santis, Identification of protein domains on topological basis, Biopolymers 58 (2001) 218–229.
[7] P.L. Privalov, L.V. Medved, Domains in the fibrinogen molecule, J. Mol. Biol. 159 (1982) 665–683.
[8] W. Wriggers, K. Schulten, Protein domain movement: detection of rigid domains and visualization of hinges in comparison of atomic coordinates, Proteins 29 (1997) 1–14.
[9] S. Hayward, H.J.C. Berendsen, Systematic analysis of domain motions in proteins from conformational change; new results on citrate synthase and T4 lysozyme, Proteins: Struct., Funct., Genet. 30 (1998) 144.
[10] I. Bahar, A.R. Atilgan, B. Erman, Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, Fold. Des. 2 (1997) 173–181.
[11] L. Holm, C. Sanders, Parser for protein folding units, Proteins 19 (1994) 256–268.
[12] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model, Biophys. J. 80 (2001) 505–515.
[13] M. Levitt, C. Sander, P.S. Stern, Protein normal-mode dynamics: trypsin inhibitor, crabmin, ribonuclease and lysocim, J. Mol. Biol. 181 (1985) 423–447.
[14] S. Hayward, N. Go, Collective variable description of native protein dynamics, Annu. Rev. Phys. Chem. 46 (1995) 223–250.
[15] K. Hinsen, The molecular modeling tollkit: a new approach to molecular simulations, J. Comput. Chem. (2000) 79–85.
[16] K. Hinsen, A. Thomas, M.J. Field, Analysis of domain motions in large proteins, Proteins 34 (1999) 369.
[17] K. Hinsen, Analysis of domain motions by approximate normal mode calculations, Proteins 33 (1998) 417–429.

[18] S. Kundu, D.C. Sorensen, G.N. Phillips Jr., Automatic domain decomposition of proteins by a Gaussian network model, Proteins 57 (2004) 725–733.

[19] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen, Essential dynamics of proteins, Proteins: Struct., Funct., Genet. 17 (1993) 412–425.

[20] B.H. Oh, J. Pandit, C.H. Kang, K. Nikaido, S. Gokcen, G.F.L. Ames, S.H. Kim, Three-dimensional structures of the periplasmic lysine-, arginine-, ornithine-binding protein with and without a ligand, J. Biol. Chem. 268 (1993) 11348–11355.

[21] C.D. Hsiao, Y.J. Sun, J. Rose, B.C. Wang, The crystal structure of glutamine-binding protein from *Escherichia coli*, J. Mol. Biol. 262 (1996) 225–242.

[22] B.E. Bernstein, P.A. Michels, W.G. Hol, Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation, Nature 385 (1997) 275–278.

[23] R. Chattopadhyaya, W.E. Meador, A.R. Means, F.A. Quiocho, Calmodulin structure refined at 1.7 angstroms resolution, J. Mol. Biol. 228 (1992) 1177–1192.

[24] H. Park, J.L. Hilsenbeck, H.J. Kim, W.A. Shuttleworth, Y.H. Park, J.N. Evans, C. Kang, Structural studies of *Streptococcus pneumoniae* EPSP synthase in unliganded state, tetrahedral intermediate-bound state and S3p-Glp-bound state, Mol. Microbiol. 51 (2004) 963–971.