CluD, a Program for the Determination of Hydrophobic Clusters in 3D Structures of Protein and Protein-Nucleic Acids Complexes

Andrei ALEXEEVSKI, Sergei SPIRIN, Daniil ALEXEEVSKI,

Oleg KLYCHNIKOV, Anna ERSHOVA[1], Mikhail TITOV[1], Anna KARYAGINA[1]

Belozersky Institute and Department of Computational Mathematics and Cybernetics, Moscow State University, Moscow, 119992, Russia; aba@belozersky.msu.ru

[1]*Institute of Agricultural Biotechnology, Timiryazevskaja street, 42, Moscow, 127550, Russia;* anna@iab.ac.ru

Annotation

In structures of macromolecules, hydrophobic clusters are spatial areas filled mainly by non-polar atom groups. Well-known hydrophobic cores of proteins are the most significant examples of hydrophobic clusters. Among other examples are hydrophobic clusters at inter-molecular interfaces.

We present an original approach and an algorithm for detecting hydrophobic clusters in a given 3D structure. In comparison with the approaches published earlier, an essential difference is the 'atomic' level of consideration instead of 'amino acid' level. The algorithm is realized as an on-line program 'Cluster Detector' (CluD). This program allows to divide a total set of non-polar groups (CH_3, CH_2, etc.) in a given structure into clusters having strong contacts inside each cluster, and weak or no contacts between the clusters.

Using the program, the conserved hydrophobic core of homeodomains and typical hydrophobic clusters at DNA – homeodomain interface were analyzed.

The work was partly supported by RFBR (grants 03-04-48476 and 03-07-90157) and Ludwig Institute for Cancer Research (CRDF GAP grant RB0-12771 MO-02).

Introduction

Non-polar molecules, such as carbohydrates, aggregate in water solution. This phenomenon is known as hydrophobic effect. The hydrophobic effect can be explained as an aspiration of a system including non-polar molecules and polar solvent (water) to minimize the contact area between non-polar molecules and water. There is a commonly used term 'hydrophobic interaction' between non-polar groups, but in the strict sense this 'interaction' can be realized only in water or in other polar solvent.

Non-polar atom groups in proteins and nucleic acids also aspire to minimize their contact area with water. Thus, non-polar side chains of amino acid residues group together to form a so-called hydrophobic core of a protein molecule [1]. Fig. 1 illustrates (a) a scheme of packing of small non-polar molecules in water and (b) a 6 Å slice of a protein structure across the hydrophobic core. In the latter case, the spatial distribution of non-polar atom groups distinguishes sufficiently from an ideal ball. More complicated picture of hydrophobic clusters organization in macromolecules as compared with the aggregation of small non-polar molecules in water can be explained by limited degree of freedom in amino acid side chains, which is a result of covalent bonds in a peptide chain and of stereochemical limitations provided by compact packing of atoms in the spatial structure of a macromolecule.

It is known that the hydrophobic effect is essential for protein folding, stabilization of protein structure, and plays an important role in interactions of proteins with others molecules and ligands. See [1, 2] for the discussion of the nature of hydrophobicity in the molecular biology context. In 3D structures of proteins and multimolecular complexes the hydrophobic interaction results in appearance of areas of space occupied mainly by non-polar residues. The detection of clusters of hydrophobic residues in 3D structures is important from different points of view: for the characterization of hydrophobic cores of proteins [3, 4] and the comparison of proteins with the similar topology [4], for the identification of structural domains [5], for the theoretical prediction of early folding intermediates [6, 7, 8, 9, 10]. Moreover, the analysis of the hydrophobic regions at interfaces is important for the understanding of mechanisms of protein-protein [11, 12], protein-nucleic acids [13, 14], and other types of intermolecular recognition.

At present, several automatic algorithms for the detection of clusters of hydrophobic amino acid (aa) residues or compact side-chain clusters, which often consist of hydrophobic residues, are described [6, 7, 4, 8, 9, 10]. The common feature of all these algorithms is the usage of aa residues as 'indivisible units' and detecting different groups of closely located residues.

The widely used 'aa residue' level of investigations of the hydrophobic effect in proteins, possesses some disadvantages. The main of them is a result of the complex nature of a number of aa residues, which include both polar and non-polar atomic groups: for example, in the side chain of lysine three non-polar groups formed by C_β, C_γ, C_δ atoms are followed by the positively charged polar nitrogen N_ζ group. Thus, hydrophobic clusters detected by the above mentioned algorithms either miss non-polar groups of aa residues like Lys (if clusters are restricted to only hydrophobic and aromatic aa residues) or include both non-polar and polar atom groups. Moreover, the ‘aa residue’ based algorithms can not be used for detection of hydrophobic clusters in non-protein molecules (e.g., in nucleic acids), as well as at protein – nucleic acid and protein – ligand interfaces.

To reflect the physical nature of hydrophobicity more adequately, we use 'an atomic' level of consideration. In this paper, an algorithm for detection of clusters of nearby located non-polar atomic groups in 3D structures of proteins, protein – protein complexes, and protein – nucleic acid complexes is described. This algorithm is realized in the CluD program implemented as an on-line service (http://math.belozersky.msu.ru/~mlt/HF_page.html).

As an example of the CluD program usage, we analyzed the hydrophobic clusters in structures of homeodomains and at homeodomain – DNA interfaces. Homeodomains are the common name for a family of DNA-binding eukaryotic protein domains, which are involved into the regulation of genes expression during the organism development. At present, the homeodomain family is one of the best structurally characterized group of DNA-recognition domains. The structure of homeodomains is extremely conserved, which implies a strong evolutionary pressure. It was interesting to imply an automatic algorithm to characterize the structural conservation of homeodomains at the level of hydrophobic clusters. The elaborated CluD program was used for the identification of the conserved hydrophobic core of homeodomains and revealing conserved hydrophobic contacts at the protein – DNA interfaces.

Materials and methods

Graph of the interactions between non-polar atomic groups

List of 'non-polar groups' in proteins and nucleic acids. In this work, the 'indivisible unit' of the hydrophobic interaction is an atomic group that includes one carbon or sulfur atom and covalently bound hydrogen atom(s). In proteins and nucleic acids these groups are -CH3, -CH2-, -CH=, -SH, and -S- groups.

The CluD program uses two variants of the list of non-polar groups. The extended list includes all such groups in protein and nucleic acids, the restricted list includes the groups of extended one, main atom of which is covalently bound only with another carbon or sulfur atoms (and, thus, is not covalently bound with any polarized atom). These lists are available at http://math.belozersky.msu.ru/~mlt/help.html#list. In the CluD program, the selection of a list is a user choice.

The center of a non-polar group is the center of the carbon or sulfur atom. Hydrogen atoms are not considered in calculations.

Graph of the interactions of non-polar groups in a 3D structure. Each non-polar group corresponds to one vertex of the graph. Two vertices are bound by an edge (and are suppose to be in interaction) in the case if (a) the distance between the centers of the groups does not exceed the threshold value d; (b) the interaction of these groups is not

prevented by any other (polar or non-polar) group. We say that a group *C* prevents the interaction between non-polar groups *A* and *B*, if the ball with the center *C* and the radius R_C includes the intersection of the balls with the centers *A* and *B* and the radii R_A and R_B, respectively. In the CluD program all radii are equal to 2.7 Å, which corresponds to the minimal possible distance from a carbon-generated non-polar group to the oxygen atom of a water molecule. The value of *d* is a user parameter. Maximal value of *d* is 5.4 Å, which corresponds to the sum of R_A and R_B radii in the case of carbon non-polar groups (roughly speaking, the hydrophobic interaction occurs if a water molecule cannot be placed between two non-polar groups). By default the *d* value is 4.5 Å.

Edges of the graph reflect two sufficiently different types of relations between atomic groups: (i) 'fixed' groups (including covalently bound ones); (ii) groups that are nearby located due to the hydrophobic interactions. Two non-polar groups *A* and *B* are 'fixed', if the distance between them is fixed due to chemical bonds, or any other interaction with energy exceeding sufficiently the free energy of the hydrophobic interaction between the groups. The examples of fixed non-polar groups are two groups, the central atoms of which are covalently bound with some third atom, or any two groups of aromatic ring. In the described algorithm and program these two different types of edges are not distinguished.

The graph of the interaction of non-polar groups is denoted by Γ.

Definition of (*K*,*L*)-cut of the graph Γ

A (K, L)-cut of a graph is the basic notion used in the elaborated algorithm for a hydrophobic cluster detection.

The 1-*neighborhood of a subgraph* Δ *of the graph* Γ is the subgraph Δ' constructed from Δ by adding all edges that have at least one vertex belonging to Δ. The *L-neighborhood* is produced by *L* iterations of 1-*neighborhood*.

A connected subgraph Δ with ≤ *K* edges is called *a (K, L)-cut*, if the subgraph obtained by removing the edges of Δ from its *L*-neighborhood has two or more connected components (Fig. 2).

Algorithm of exhaustion of connected subgraphs of ≤ *K* edges

Assign indexes 1, 2, ... to all edges of Γ; the *n*-th edge will be denoted by E_n. Let Δ be a connected subgraph of $t<K$ edges, E_n be the edge of Δ with the maximal index *n*. The subgraph Δ' obtained by adding an edge E_x to Δ is called a 1-extension of Δ if (i) E_x belongs to the 1-neighbourhood of Δ; (ii) either $x>n$, or $x<n$ and the subgraph obtained from Δ' by deletion of the edge E_n is not connected.

The algorithm starts with the exhaustion of all subgraphs of one edge. Then for each subgraph Δ all 1-extensions of Δ are considered. The procedure iterates K–1 times.

It can be proved that each connected subgraph of $\leq K$ edges is considered by the algorithm exactly ones.

Parameters of hydrophobic clusters

The following parameters are calculated for each hydrophobic cluster. (i) *The number of non-polar groups in the cluster*. (ii) *The mean number of interactions of one non-polar group in the cluster*. This parameter indirectly characterizes the form of the cluster. More compact packing of hydrophobic groups and the form of the cluster closer to an ideal ball are characterized by a bigger value of mean number of interactions of a non-polar group. (iii) *Semiaxes of the ellipsoid of inertia of the cluster*. In calculating the ellipsoid, the groups of the cluster are considered as material points in space, with unitary masses. Ratios of semiaxes reflex the spatial form of the cluster. For example, if all three semiaxes are approximately equal, then the form of the cluster can be assumed to be close to a ball.

Algorithm for the detection of the conserved hydrophobic core of the homeodomain protein family

The 41 spatial structures of homeodomains and their DNA complexes deposited in PDB were used in this work. Those structures represent the total number of 21 different proteins of yeast, fruit fly, rat, mouse, and human. The proteins are grouped into 12 homeodomain subfamilies. Full list of used structures is presented in (http://www.dpidb.belozersky.msu.ru/Reviews/Homeo/homeo_eng.html). From each NMR entry with a number of models, only the first model was used.

The conserved hydrophobic core was found using the following procedure.

1. Some of the initial PDB files represent structures of complexes including DNA and two or more molecules of homeodomains. Such files were transformed to files, each containing a complex of one homeodomain molecule with a DNA duplex. Together with the PDB files that initially contain only one homeodomain, 16 files with single homeodomains and 43 files with homeodomain – DNA structures were obtained. In sequel, we refer to these 59 files as 'structures'.

2. The files obtained in the step 1 were used for spatial superposition of all 59 homeodomain structures. The amino acid sequences alignment agreed to the structural alignment was built. The amino acid residues in all structures were renumbered in a unified manner, according to the alignment [15]. For those structures that include DNA, the

DNA bases were also renumbered so that the reference pair A-T (see [15]) obtained the numbers 103-203 in every structure. The programs SwissPDBViewer [16] and Rasmol [17] were used to perform this step.

3. The hydrophobic core was found in each structure. The hydrophobic core was defined as the maximal hydrophobic cluster found by the CluD program (parameters were: K=L=1, the restricted list of non-polar groups, d=5.4).

4. In the amino acid sequences of the homeodomains, the residues participating in the formation of the hydrophobic core were determined.

5. Positions (columns) of the alignment containing the residues included into the hydrophobic core in all structures were marked (Fig. 3).

6. The *conserved hydrophobic core* of a given structure was defined as all non-polar groups of the amino acid residues in the marked positions.

Results

(*K, L*)-cut algorithm for the hydrophobic clusters detection

The suggested algorithm solves the problem of the detection of hydrophobic clusters corresponding to spatial areas occupied exclusively or mainly by non-polar groups of atoms, in a given 3D structure of protein or macromolecular complex. It is rather difficult to give a strict definition of hydrophobic cluster, because of the complexity of configuration of non-polar groups in structures. Detection of hydrophobic clusters can be based on the fact that each non-polar group has many interactions inside a hydrophobic cluster, and at the same time between two hydrophobic clusters the interaction is weak. Detection of hydrophobic clusters in a planar variant is illustrated in Fig. 4.

The algorithm realizes the above mentioned intuitive concept concerning a hydrophobic cluster as follows: (i) weak interactions between the hydrophobic clusters (yet not found) are searched, in terms of graph of interactions these weak interactions correspond to (*K, L*)-cuts; (ii) the hydrophobic clusters are detected as components obtained after the removal of weak interactions, in terms of graph these components are connected subgraphs. The parameters of the algorithm are the integers *K, L* used in the definition of (*K, L*)-cut, the list of non-polar groups, the threshold for distance between interacting non-polar groups *d*, and the minimal number *m* of non-polar groups in a found hydrophobic cluster. The algorithm works as follows.

At the *first* step the graph Γ of interactions of the non-polar groups is constructed (see Materials and Methods and Fig. 4b).

At the *second* step all (*K*, *L*)-cuts in the graph Γ are found (Fig. 4c). If a subgraph with $t<K$ edges is a (*K*, *L*)-cut, then larger subgraphs including this (*K*, *L*)-cut are not considered: the process of connected subgraphs exhaustion (see Materials and Methods) stops extending a subgraph if it is proved to be a (*K*, *L*)-cut.

At the *third* step all edges of found (*K*, *L*)-cuts are removed from the graph (Fig. 4d).

At the *fourth* step the connected components of the obtained graph are found. The vertices of each connected component that contains not less than *m* non-polar groups form a hydrophobic cluster. For every hydrophobic cluster the additional parameters (see Materials and Methods) are calculated. The found hydrophobic clusters, ordered in accordance to the number of non-polar groups, form the *output* of the algorithm.

For graphs of interacting non-polar groups the algorithm is linear in the number *N* of atoms in a given structure, because the number of interactions of a non-polar group with other groups is obviously restricted. The constant in this linear function is rather big and depends on *K* exponentially. Fortunately, only small $K=1, 2, 3$ are expected to be sufficient for reasonable hydrophobic cluster detection. In the application described in sequel, even $K=1$ appeared to be acceptable.

Program CluD (Cluster Detector)

The (*K*, *L*)-cut algorithm of hydrophobic clusters detection is realized as a program named CluD that is equipped with a publicly available on-line interface (http://math.belozersky.msu.ru/~mlt/HF_page.html). In the current version the parameters $K=L=1$, $m=3$ are used.

The *input* of the program is PDB code of a structure and, if necessary, specified regions of protein or nucleic acid chains to be analyzed. The user's parameters are the maximal distance of hydrophobic interaction *d*, the list of non-polar groups (Extended or Restricted) and the minimal number of non-polar groups in a cluster *m*.

The *output* of the program includes:

- a table of found hydrophobic clusters, with the list of non-polar groups forming each cluster, and the parameters of the cluster. The table format provides the possibility to import the file into standard spreadsheet programs (like MS-Excel);

a script file for RasMol program [17] intended for the visualization of 3D structures of molecules; the script file allows the user to visualize the found clusters in a local PC;

an on-line visualization of the input structure and the found hydrophobic clusters, which can be administered by the user; the visualization can be used after installing the free Chime package (http://www.umass.edu/microbio/chime/) on the user's PC.

The command line version of the CluD program has the option for the detection of hydrophobic clusters at a protein – DNA interface (or at an interface of any molecules). If this option is on, then every resulting cluster includes only those groups of both molecules that interact with some group of the other molecule.

Conserved hydrophobic core of homeodomains

Homeodomains are components of a large number of eukaryotic transcription factors [15]. The majority of homeodomain-containing proteins participates in regulation of the key processes of development and homeostasis. The structure of homeodomains includes three α–helices with additional so-called N-terminal arm. The main function of homeodomains is recognizing and binding specific DNA sequences. The DNA-binding module of homeodomains is HTH (helix-turn-helix) structural motif. 43 structures of homeodomains from *Drosophila melanogaster*, human, mice, rat and yeasts are deposited in PDB.

The superimposition of all 3D structures of homeodomains results in high degree of spatial coincidence [15]. Therefore, it was expected that hydrophobic cores of homeodomains include conserved part, similar in all structures, and additionally specific for each structure non-polar groups. To identify the conserved hydrophobic core of homeodomains the procedure described in Materials and Methods was used. The conserved hydrophobic core of homeodomains is shown in Fig. 3. We suppose that in all homeodomains, even with unknown 3D structure, amino acid residues in the found positions are involved into the hydrophobic core formation. It should be noted that the conserved hydrophobic core includes not only side chains of aliphatic and aromatic residues, but also non-polar groups of polar residues, such as arginine and lysine.

A priori, the algorithm does not guarantee that conservative hydrophobic cores in all structures should occupy identical position, i.e., will coincide at all superimposed structures. However, testing this hypothesis had shown practically full spatial overlapping of the conserved hydrophobic cores in all structures of homeodomains. Actually, after superimposition of 59 homeodomain structures by C_α atoms, the spatial location of all conserved hydrophobic cores was checked. It was shown that hydrophobic cores of all except two homeodomains are located in the 3Å

neighborhood of the hydrophobic core of Msh homeodomain (PDB code 1IG7), structure of which was used as a basis for comparison. There were found only two exceptions. Both of them are NMR resolved structures (yeast Mata1 homeodomain, PDB code 1F43 and human Oct-2 homeodomain structure from the PDB entry 1HDP, which represents an NMR average structure).

The analysis of the spatial location of the conservative hydrophobic core in homeodomains demonstrates that it is really located in the geometrical center of the domain, which is in accordance with theoretical suggestions. Fig. 5 illustrates the location of the conserved hydrophobic core between three α-helices. Besides, the arm of the homeodomain interacts with the conserved hydrophobic core by aa residue in the position 8. Apparently, this interaction fixes the arm, which is important for the contact of the domain with the DNA minor groove (Fig. 3).

Hydrophobic clusters at homeodomain – DNA interface

The analysis of 3D structures and site-directed mutagenesis experiments have demonstrated an essential role of the hydrophobic interactions between side chains of the recognition helix residues and the major groove of DNA in specific recognition of DNA by homeodomains [18, 19].

To reveal the role of certain homeodomain residues, forming the hydrophobic contacts with DNA, the hydrophobic clusters on the homeodomain-DNA interface were analyzed for each of 43 homeodomain-DNA structures (see Materials and Methods) using CluD program (version for detection the clusters on the interface of two molecules). Four clusters presented in 10 or more structures were identified (Table 1). The names of the clusters include the numbers of the key interacting aa residue and of the DNA base, according to the accepted numbering (see Materials and Methods). In several structures additional bases are involved into clusters.

Several 'recognition rules' can be derived from the analysis of the clusters.

1) If Val or Ile occupy the position 47 of a homeodomain, then T is in the position 104 of the recognition site and the hydrophobic cluster 47-104 is formed in the homeodomain-DNA complex. In the analyzed structures of complexes there are 16 homeodomais with Val residue in the position 47. In all these structures T is in the position 104 of DNA. The cluster 47-104 was detected in 15 of 16 cases (2 D. melanogaster Eve homeodomain structures from 1JGG PDB entry, 3 yeast Mat**a**1 structures from 1AKH, 1YRN and 1LE8, 3 D. melanogaster Prd structures from 1FJL, 7 Pou family structures: 5 human Oct-1 from 1CQT and 1OCT and 2 rat Pit-1 from 1AU7), and was not detected in 1 case (one of two human Oct-1 structures from 1CQT).

In 16 structures of homeodomains the position 47 is occupied by Ile. In 14 of 16 cases T is in the position 104 of DNA and the cluster 47-104 was detected (5 Antennapedia family homeodomain structures: 3 *D. melanogaster* Antp structures from 9ANT and 1AHD entries, 1 *D. melanogaster* Ubx from 1B8I, 1 human HoxB1 from 1B72; 8 *D. melanogaster* En from 1DUO, 1HDD, 2HDD and 3HDD, 1 mouse Msx-1 from 1IG7). In 2 structures (both of *D. melanogaster* NK-2 homeodomain from 1NK2 and 1NK3 entries) the position 104 of DNA is occupied by G and in this case a hydrophobic cluster was not detected.

In 11 structures the position 47 of the homeodomain is occupied by an Asn residue, and in all these cases the hydrophobic cluster 47-104 was not detected (8 yeast Matα2 structures from 1AKH, 1APL, 1MNM, 1YRN, 1K61; 3 Tale family homeodomains: 1 human Pbx-1 from 1B72 and 1 mouse Pbx-1 from 1LFU; 1 *D. melanogaster* Exd from 1B8I).

2a) If Cys is in the position 50, then C is in the position 200 of the recognition DNA site and the hydrophobic cluster 50-200 is formed. Cys in the position 50 were observed in 5 structures (5 POU family homeodomain structures: 2 rat Pit-1 from 1AU7 entry, 3 human Oct-1 from 1CQT and 1OCT), and in all cases a cluster 50–200 was detected.

2b) If Gln is in the position 50, then the hydrophobic cluster 50-200 is formed only if the position 200 of DNA is occupied by T. Among 16 structures with Gln50, in 4 cases the cluster 50-200 was detected: 3 cases with T in the position 200 (3 D. melanogaster En from 3HDD and 1HDD), and 1 cases with G in the position 200 (1 D. melanogaster Ubx structure from 1B8I). Other residues (Ala, Gly, Ile, Lys, Ser) in the position 50 do not form a cluster with a base in the position 200 (24 structures).

2c) The cluster 50-200 is accompanied by a cluster 47-104 (in 9 cases from 9).

3) If Ile or Met occupy the position 54 of homeodomain, then the position 200 of recognition DNA site is occupied by A or C and the hydrophobic cluster 54-200 is formed. In 3 structures Ile occupies the position 54 and in all cases the cluster 54-200 was detected (3 Tale family structures: human Pbx-1 from 1B72, D. melanogaster Exd from 1B8I, mouse Pbx-1 from 1LFU). In 10 structures Met occupies position 54. In 8 cases the cluster 54-200 was detected (5 Antennapedia family homeodomain structures: 3 Antp structures from 9ANT and 1AHD, 1 HoxB1 structure from 1B72, 1 Ubx structure from 1B8I, 3 Mat**a**1 structure from 1AKH, 1YRN, 1LE8). In 2 cases the cluster 54-200 was not detected (2 Eve structures from 1JGG).

Thus, in 11 of 13 cases the cluster 54-200 was detected. It should be noted that in 8 of 11 cases the base in the position 201 is also included into this cluster.

In 30 structures with another residues (Ala, Gln, Arg, Tyr) in the position 54 the cluster 54-200 was not detected.

4) *If Arg is in the position 54, then T is in the position 201 of the recognition DNA site and they form a hydrophobic cluster*. In 8 structures Arg occupies the position 54 (7 Mat-α2 structures from 1AKH, 1APL, 1MNM, 1YRN, 1K61). In 7 cases T is in the position 201 and the cluster 54-201 was detected. Hydrophobic clusters were not detected in 1 structure from 1K61, in which the position 201 is occupied by A.

Discussion

Algorithm and program implementation

The developed CluD program is a publicly available tool for detecting hydrophobic clusters in 3D structures of proteins and multimolecular complexes, including DNA-protein ones. We do not know any analogous Internet services or free-distributed computer programs.

The goal of the work was to develop an algorithm for automatic detection of spatial hydrophobic regions in 3D structures. In the suggested algorithm, geometrical criteria and the graph-analytical approach were used. The graph of hydrophobic interactions models complex spatial relationships of non-polar atomic groups in macromolecules. The obtained results demonstrate that the elaborated algorithm and CluD program can be successfully used for the detection of hydrophobic spatial areas in proteins and at protein – DNA interfaces. The CluD program computes several parameters as well, which are useful for characterization of the geometrical form and density of hydrophobic groups packing.

The 'atomic' level used in the algorithm leads to an interesting observation. Polar residues like Arg, Lys, Gln, Glu, Thr are sometimes involved into formation of a conserved hydrophobic core. See Fig. 3 for the examples (Arg or Lys in position 52, Arg, Glu or Thr in position 44). Therefore, these residues can be considered as bifunctional ones, partially hydrophobic and partially hydrophilic. Thus, in a specific position even different in properties aa residues, for example, Val and Arg, can play the same role in formation of a hydrophobic core (Fig. 3, position 44).

The (*K, L*)-cut algorithm, used for detection of clusters, realizes an approach that can be named 'from whole to its units', in contrast to the approach 'growing units from germs'. We believe the former approach can have certain advantages against the latter in a number of specific problems, concerning both structural and sequence analysis.

Hydrophobic clusters in homeodomains

The detailed detection of the hydrophobic core in a protein structure is important to identify the function of certain amino acid residues and, hence, to predict the change of protein properties provided by directed mutagenesis, and, as well as, to study the peculiarities of evolution within protein families.

One of intriguing features of homeodomains is the precision of their backbone coincidence after fitting of 3D structures despite of wide evolutionary distribution (putatively, all or almost all branches of eukaryota) and not very high sequence similarity. Such coincidence, being natural for structural elements carrying catalytic centers of enzymes, is difficult to explain for transcription regulators [20, 21]. Hydrophobic core of proteins is known to be one of the main contributors into the folding and globule stability. The detection of the conserved hydrophobic core of homeodomains can give arguments for the explanation of the mechanism of the high similarity of homeodomain folds.

As it was shown in the work, the detected conserved hydrophobic core is placed in the center of the globule and tentatively it plays a significant role in positioning of three α-helices of a homeodomain. Hydrophobic core of all homeodomains occupies almost the same spatial region (see Results), which is in agreement with the above suggestion. Listing of all non-polar groups in cluster opens the way to measure the volume, the density of the hydrophobic core, the insertions of polar atoms in the same spatial region and their role, and may be essential for the predictions of spatial organization of a homeodomain on the base of sequence analysis.

In this work, rather strict criteria for detection of the aa residues contributing into conserved hydrophobic core were used (see Materials and Methods). We tend to detect hydrophobic core parts that are common for all homeodomains. Thus, even aa residues involved into hydrophobic core of the majority but not all structures were not selected as components of the conserved hydrophobic core (Fig. 3).

The problem of prediction of homeodomain specificity in DNA recognition is of special interest [22]. Suggested in the work partial 'recognition rules' for homeodomains are based exclusively on an analysis of 3D structures. The rules show a good precision for all analyzed structures (see Results) and are in the agreement with the observations made for homeodomains earlier [18, 19]. Nevertheless, the amount of available structural data is limited; the volume of sequence data is much greater. Thus, the rules can be used as new additional arguments in the attempts to predict the sequence of DNA site recognizing by a given homeodomain.

Conclusions

An algorithm for detection the clusters of nearby located non-polar atomic groups in 3D structures of proteins, protein-protein and protein-nucleic acid complexes is elaborated. This algorithm is realized in the CluD program implemented as on-line service (http://math.belozersky.msu.ru/~mlt/ HF_page.html). An example of successful usage of the CluD program for identification the conserved hydrophobic core of homeodomains and analysis of homeodomain-DNA hydrophobic contacts on the interface was described. The CluD program can be used by a wide range of researchers for the detection of spatial areas occupied mainly or exclusively by non-polar atomic groups, particularly, for precise detection of protein hydrophobic cores, hydrophobic clusters on protein-DNA interfaces, for investigations of protein-protein and ligand-protein hydrophobic interactions, and for other purposes.

References

1. *Finkelshtein A.V., Ptitsyn O.B.* // Protein Physics: a Course of Lectures, 2002, 350 pages, Academic Press.

2. *Chandler D.* // Nature. 2002. V. 417 (6888). P. 491.

3. *Umezawa Y., Umeyama H.* // Chem. Pharm. Bull. (Tokyo). 1988. V. 36 (12). P. 4652–4658.

4. *Swindells M.B.* // Protein Science. 1995. V. 4. P. 93–102.

5. *Swindells M.B.* // Protein Science. 1995. V. 4. P. 103–112.

6. *Plochocka D., Zielenkiewicz P., Rabczenko A.* // Protein Eng. 1988. V. 2 (2). P. 115–118.

7. *Heringa J., Argos P.* // J. Mol. Biol. 1991. V. 220 P. 151–171.

8. *Zehfus M.H.* // Protein Science. 1995. V. 4. P. 1188–1202.

9. *Tsai C.-J., Nussiniv R.* // Protein Science. 1997 V. 6. P. 24–42.

10. *Kannan N., Vishveshwara S.* // J. Mol. Biol. 1999. V. 292. P. 441–464.

11. *Tsai C.J., Lin S.L., Wolfson H.J., et al.* // Protein Sci. 1997. V. 6. P. 51–62.

12. *Tsai C.-J., Nussiniv R.* // Protein Science. 1997 V. 6. P. 1426–1437.

13. *Mandel-Gutfreund Y., Margalit H.* // Nucleic Acids Res. 1998. V. 26 (10). P. 2306–2312.

14. *Luscombe N.M., Laskowski R.A., Thornton J.M.* // Nucleic Acids Res. 2001. V. 29 (13). P. 2860–2874.

15. *Ledneva R.K., Alexeevskii A.V., Vasil'ev S.A., et al.* // Mol. Biol. (Mosk). 2001. V. 35 (5). P. 764–777 (in Russian).

16. *Guex N., Peitsch M.C.* // Electrophoresis. 1997. V. 18. P. 2714–2723.

17. *Sayle R., Milner-White E. J.* // Trends in Biochemical Sciences (TIBS). 1995. V. 20 (9). P. 374.

18. Stepchenko A.G., Luchina N.N., Polanovsky O.L. // FEBS Lett. 1997. V. 412 (1) P. 5–8.

19. *Ades S.E., Sauer R.T.* // Biochemistry. 1994. V. 33 (31). P. 9187–9194.

20. *Suzuki M., Brenner S.E.* // FEBS Lett. 1995. V. 372 (2-3). P. 215–221.

21. *Wintjens R., Rooman M.* // J. Mol. Biol. 1996. V. 262 (2). P. 294–313.

22. *Damante G., Pellizzari L., Esposito G., et al.* // EMBO J. 1996. V. 15 (18). P. 4992–5000.

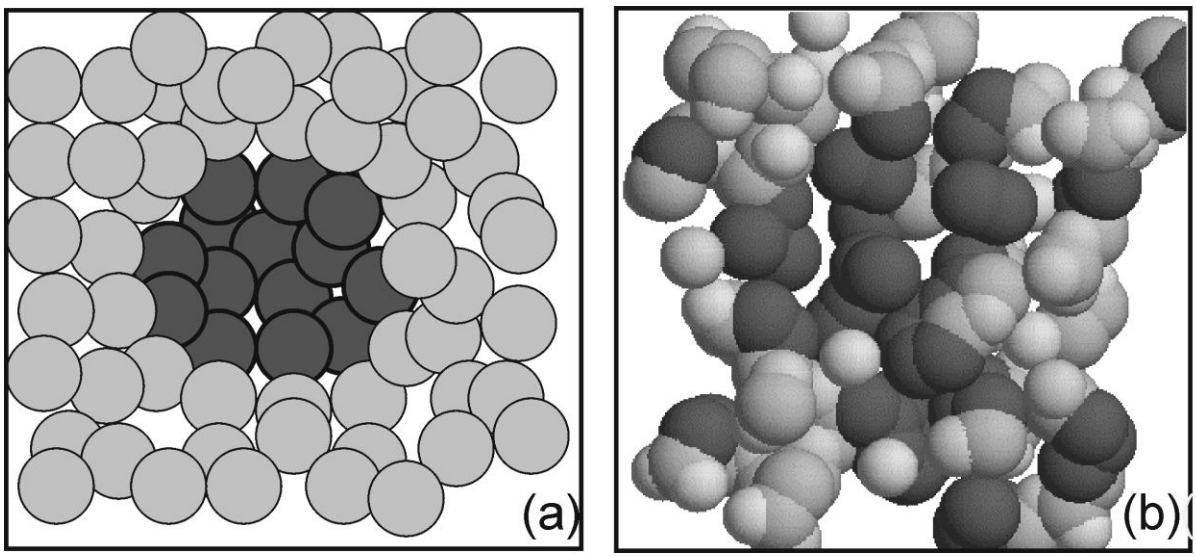

Fig. 1. (a) A scheme of packing small non-polar molecules in water. (b) A 6 Å slice of a protein structure (homeodomain 'Engrailed' from 1ENH PDB entry) across the hydrophobic core. Non-polar molecules and non-polar atomic groups are in dark gray, polar solvent and polar atoms in protein are in light gray.

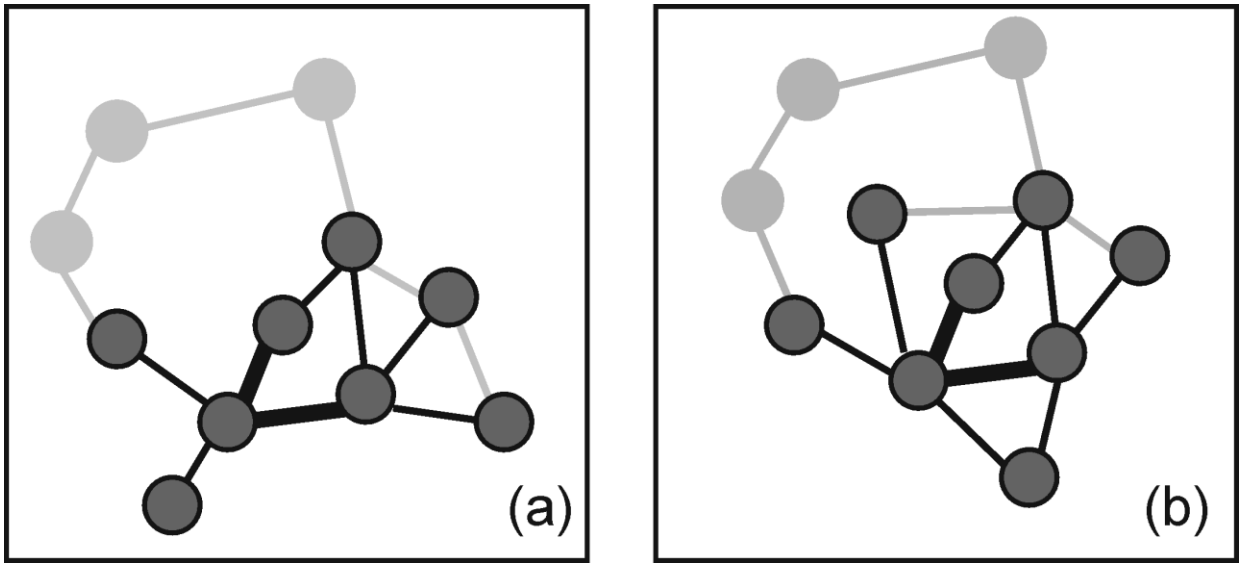

Fig.2. A Scheme of preventing the interaction between non-polar groups *A* and *B* by a group *C*. Atomic groups are denoted by black discs of radii corresponding to their van der Waals radii. Radius R of the unfilled circles corresponds to the minimal possible distance from a carbon-generated non-polar group to the oxygen atom of a water molecule.

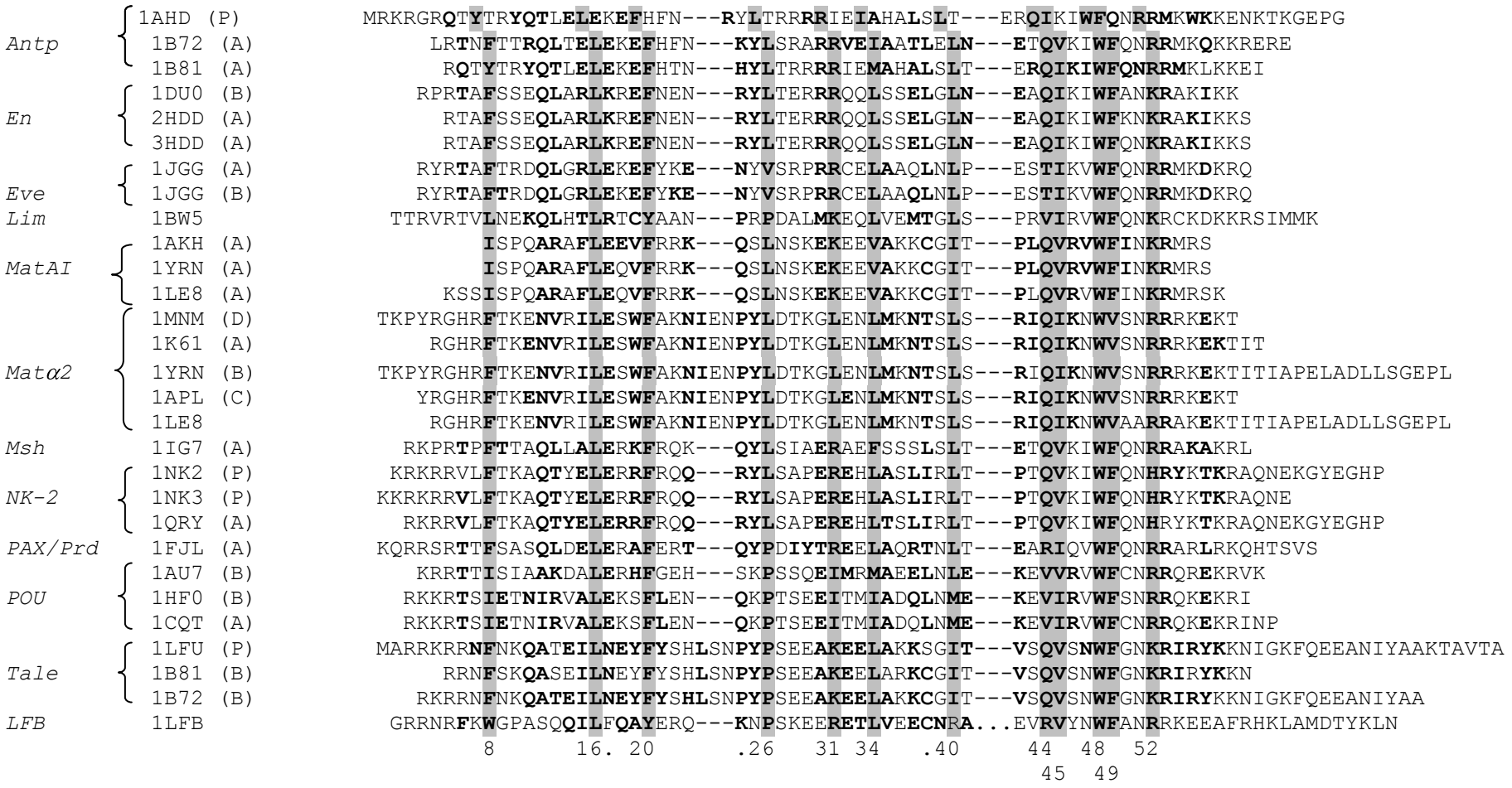

Fig. 3. Multiple alignment of selected homeodomains with marked hydrophobic core. A complete alignment with the sequences of all resolved 3D homeodomain structures is available at http://www.dpidb.belozersky.msu.ru/Reviews/Homeo/homeo_eng.html. Family names are given in the left column. The sequences are labeled by PDB codes and chain IDs. Residues involved into hydrophobic cores of proteins are in bold. The conserved hydrophobic core positions are highlighted, their numbers are given below. The hydrophobic core positions that are not conserved, but almost conserved in the complete alignment, are marked by dots. The insertion CIQRGVSPSQAQGLGSNLVTE in 1LFB after the position 42 is not shown.

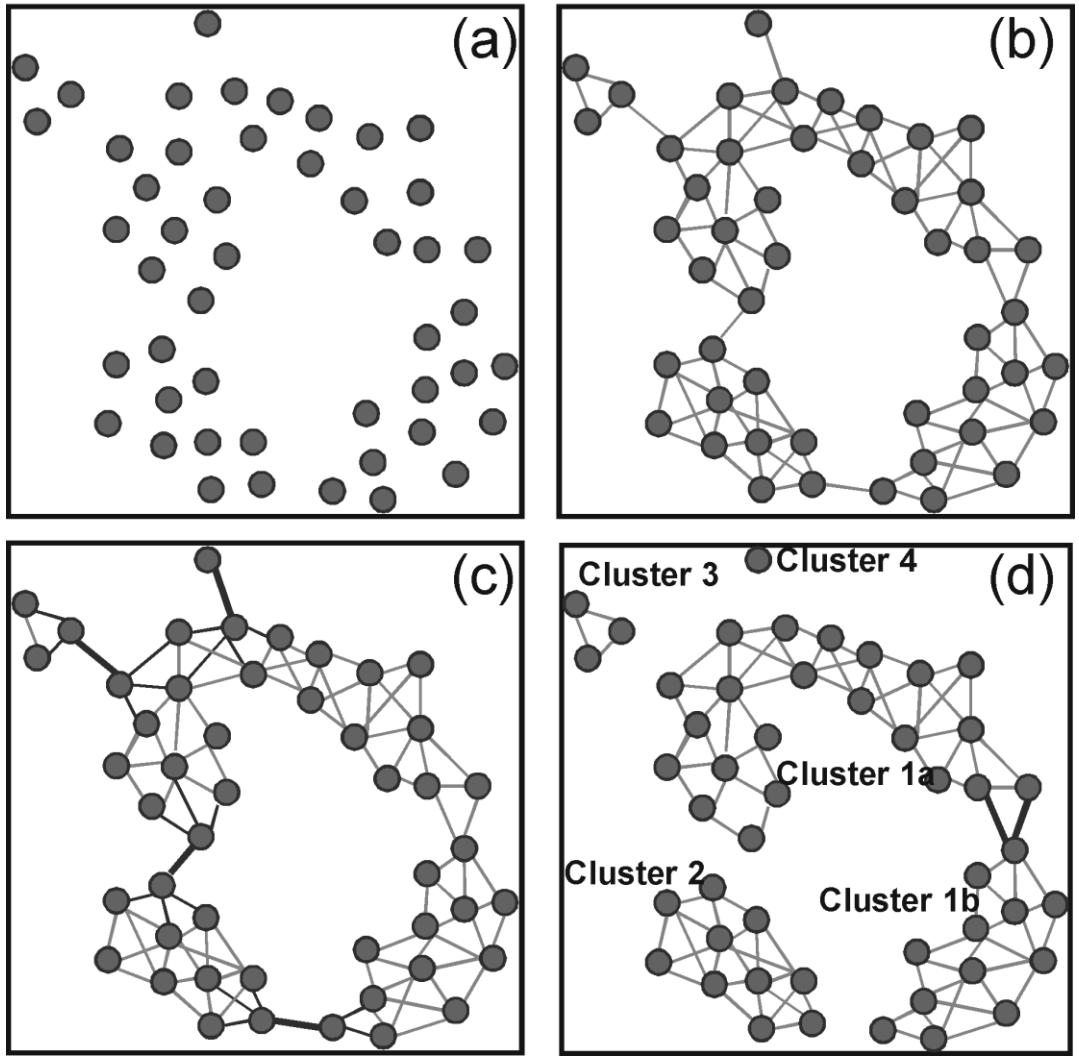

Fig. 4. Planar illustration of (*K*,*L*)-cut algorithm. (a) The initial distribution of non-polar groups. (b) The graph of interactions of non-polar groups. (c) All (1,1)-cuts in the graph. The edges of (1,1)-cuts are thick ones. (d) The graph after removal of all (1,1)-cuts. An example of one (2,1)-cut is presented (thick edges). Clusters, found after removal of (1,1)-cuts and one (2,1)-cut, are subscribed.

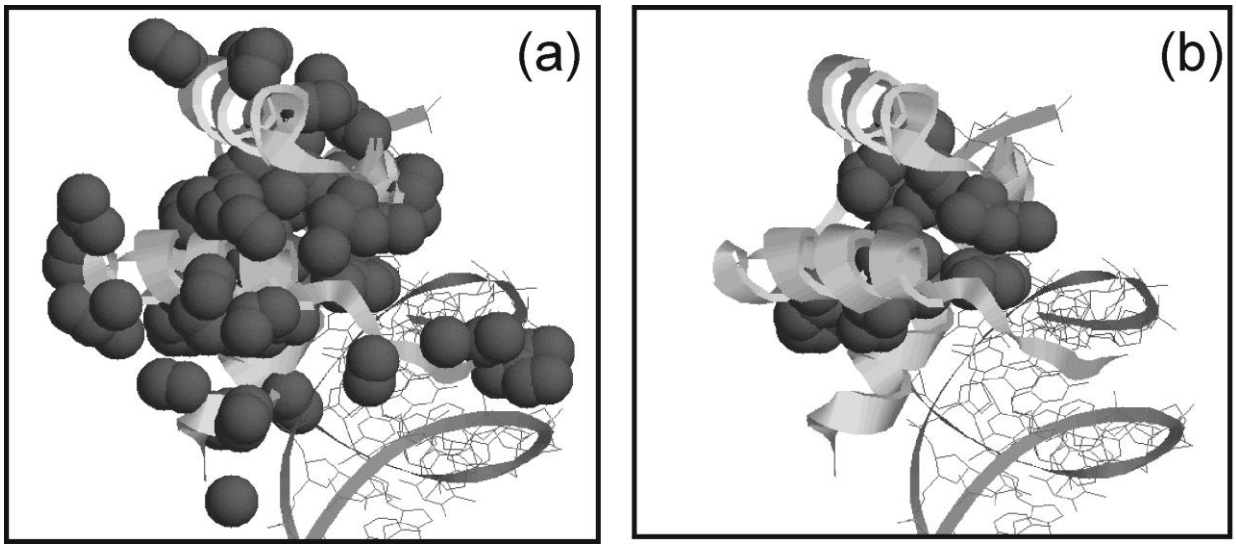

Fig. 5. (a) All non-polar groups in Matα2 homeidomain structure (PDB code 1APL) (b) Non-polar groups forming conserved hydrophobic core of the same structure. Non-polar groups are shown as dark gray balls.

Table 1

Cluster ID protein – DNA	Cluster compound		Number of structures in which the cluster		Total number of structures
	aa residues	DNA bases	was detected	was not detected	
47 – 104	Val	T	15	1	16
	Ile		14	0	14
	Val	not T	0	0	0
	Ile		0	2	2
	not (Val,Ile)	any	0	11	11
50 – 200	Cys	T	5	0	5
	Gln	T	3	2	5
	Cys	not C	0	0	0
	Gln	not T	1	11	12
	not (Cys,Gln)	any	0	21	21
54 – 200	Ile	A, C	3	0	3
	Met		8	2	10
	Ile	G, T	0	0	0
	Met		0	0	0
	not (Ile,Met)	any	0	30	30
54 –201	Arg	T	7	0	7
	Tyr	A	2	0	2
	Arg	not T	0	1	1
	Tyr	not A	0	0	0
	not (Arg,Tyr)	any	10	23	33

Conserved hydrophobic clusters at DNA – homeodomain interface. The clusters were detected and named by a key interacting pair aa residue–DNA base. Unified residue numbers and base numbers were used (see Materials and Methods). DNA bases of direct strand have numbers 100–106, of reverse strand 200–206. Bases 103 and 203 are complementary. The conserved adenine 103 is bound to the conserved Asn51 by hydrogen bonds in all structures.