ISSN 0233-7657. Biopolymers and Cell. 2010. Vol. 26. N 2

BIOINFORMATICS

New technique of identifying the hierarchy of dynamic domains in proteins using a method of molecular dynamics simulations

S. O. Yesylevskyy

Institute of Physics of National Academy of Science of Ukraine
46, Prosp. Nauky, Kyiv-28, Ukraine, 03680

yesint3@yahoo.com

Aim. *Despite a large number of existing domain identification techniques there is no universally accepted method, which identifies the hierarchy of dynamic domains using the data of molecular dynamics (MD) simulations. The goal of this work is to develop such technique.* ***Methods***. *The dynamic domains are identified by eliminating systematic motions from MD trajectories recursively in a model-free manner.* ***Results***. *The technique called the Hierarchical Domain-Wise Alignment (HDWA) to identify hierarchically organized dynamic domains in proteins using the MD trajectories has been developed.* ***Conclusion***. *A new method of domain identification in proteins is proposed.*

Keywords: dynamic domains, domain identification, Hierarchical Domain-Wise Alignment, molecular dynamics.

Introduction. The protein domain is a very important concept in the protein science [1, 2]. The motions and interactions of domains are crucial for the functioning of various proteins including enzymes [3–6]. There are numerous methods of domain identification. They can be based on observation of independent folding [2], similarity of the sequence motifs [7], the presence of a distinct hydrophobic core [8], functional activity [8, 9], contact classification [10], topology [11], structural homology [12], independent mobility [13–16] and other properties. In this work we focus on so-called dynamic domains. Dynamic domain is defined as relatively compact part of a protein that is characterized by its own pattern of internal collective dynamics, which can be distinguished from that of other domains [13–16]. This concept provides the most physically justified definition of the domain. The techniques, which identify the dynamic domains employ either the analysis of alternative crystal structures of the same protein [14] or the elastic network models [15–18] and the graph-theory approaches [19].

In contrast to conventional structural domains, dynamic domains could be subdivided into smaller relatively independent units using the same physical principles of the similarity of dynamic patterns. As a result, the hierarchy of dynamic domains, where each domain contains a number of smaller subdomains, could be identified [20]. The subdomains are characterized by their own distinctive patterns of motions within the parent domain. In our previous works we developed several techniques of identification and analysis of hierarchical dynamic domains [20–23]. It was shown that the concept of dynamic domains could be used with great success in the statistical analysis of

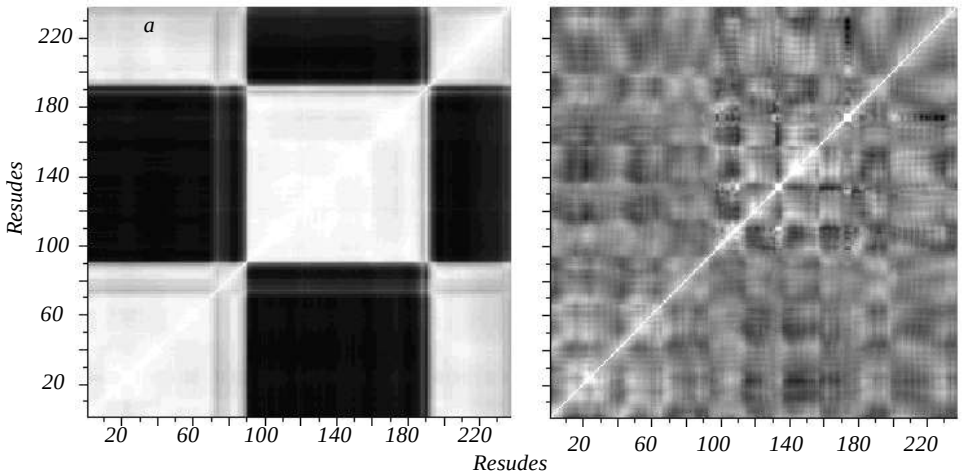

Fig. 1. The matrices of the residue-residue correlations of motions obtained by GNM (*a*) and MD (*b*) for lysine-, arginine-, ornithine-binding protein. White corresponds to 1, black to – –1. The block structure is clearly visible in (*a*), but absent in (*b*)

non-redundant protein structures [21], in finding the candidate proteins for biosensor design [22] and in simulating the conformational transitions in two-domain proteins [23]. These results are summarized in our recent methodological eview [24].

Only few attempts were made so far to identify the hierarchy of dynamic domains in proteins using the data of Molecular Dynamics (MD) simulations [25]. MD simulations are the most precise way of tracking the motions of proteins because the trajectories of all atoms could be stored and analyzed [26]. The dynamic domains identified using MD trajectories would reflect the dynamics of the protein globule, which is quite close to reality.

In this work we present a new method of domain identification based on the analysis of MD trajectories. The method is called Hierarchal Domain-Wise Alignment (HDWA). It is conceptually similar to the recently developed Hierarchical Clustering of the Correlation Patterns (HCCP) technique [20], but uses different input data. HDWA exploits the hierarchical character of protein motions recorded in MD trajectories, while HCCP utilizes the patterns in the matrices of residue-residue correlation of motions, which are computed using GNM. HDWA is a model-free and parameter-free technique. It identifies a hierarchy of dynamic domains from MD trajectories or any other sets of atomic coordinates and allows estimating stability and interdependence of domains. Here we describe the theory of the HDWA method. The current work is a proof of the principle of our technique, thus the biological significance of our results for particular proteins is not discussed.

Theory and methods. *The rationale.* The correlations of motions provide important information about the protein dynamics. Computation of the correlations of motions in MD simulations looks very simple because the trajectories of all atoms are recorded, however this is not always the case. Let us leave aside the mathematical problems of computing non-linear correlations adequately [27] and concentrate on the interpretation of the correlation data. The correlations of motion obtained in the normal mode calculations (such as GNM) represent small harmonic displacements around the local energy minimum in vacuum. There are no spurious correlations caused by large-amplitude diffusive motions and thermal noise. As a result the matrix of the residue-residue correlations in the case of GNM (or other normal mode-based techniques) exhibits clearly visible «blocks», which correspond to dynamic domains [20] (Fig. 1, *a*). It is impossible to obtain such clear picture in the case of MD because of numerous factors, which produce spurious correlations or mask existing correlation patterns. Each atom in the protein participates in several motions in the course of MD, which are organized into the natural hierarchy. This includes the motion of the whole molecule, the motion of the domain, the motion of the relatively rigid subdomains inside the domain, the motion of the secondary structure element, the motion of individual residue side chain, etc.

The diffusion of the molecule as a whole is eliminated by aligning each frame of the MD trajectory with some reference structure, however all other motions are lumped together unpredictably in the resulting

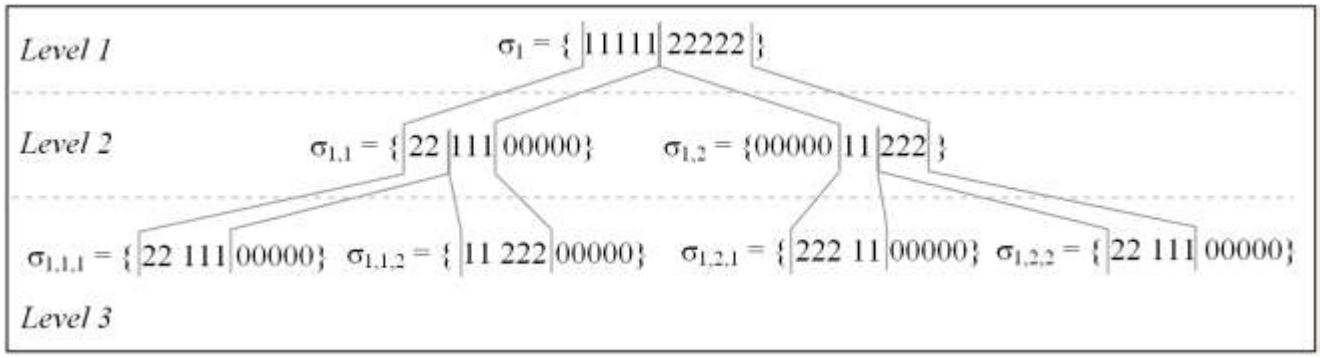

Fig. 2. The scheme of the hierarchical tree of domains with the corresponding domain encoding for each node. For clarity the regions to «1»s and «2»s are shown continuous. In reality this is not necessarily so

trajectory. This makes it very hard to extract the motion on particular level (i. e. the motions of domains). As a result the matrix of the residue-residue correlations in the case of MD usually has little in common with the matrix obtained in GNM (Fig. 1, *b*). Such matrix still contains useful information about large-scale protein dynamics, which could be extracted by means of the principal components analysis [28], however the absence of pronounced block pattern makes it not unsuitable for dynamic domains identification (at least in the same manner as in the case of GNM).

The terms «domain» and «subdomain» are used in the following meanings hereafter. Each domain consists of several subdomains of the next hierarchical level (or the child domains) and is a child of the domain of the previous level (the parent domain).

The difference in the correlation matrices between MD and GNM is caused by intrinsic difference of the character of motions in these techniques, thus attempts to obtain GNM-like correlation matrices in MD are unlikely to be successful.

Other technique, which is not based on the block pattern of the correlation matrix, is needed to identify the dynamic domains from the MD trajectories. The most logical solution is to exploit the natural hierarchy of motions in the protein globule.

The motions of the whole molecule in the course of MD could be eliminated by aligning the molecule to the reference structure in each trajectory frame. Following this idea the motion of individual domain could be eliminated by aligning this domain to the reference separately. The motions of subdomains could be eliminated in the same way by aligning them to the reference separately and so on, until random fluctuations of individual residues around their reference positions will remain. In other words such procedure eliminates all systematic motions from the MD trajectory in the step-wise manner. The most pronounced large-amplitude collective motions would be eliminated first. Remaining smaller scale motions would be eliminated on the next step, etc. However, the boundaries of domains and subdomains are not known in advance, thus it is not known which part of the structure should be aligned with the reference on each level.

Let us assume that the boundaries of domains are guessed correctly and each domain is aligned separately to the reference structure (domain-wise alignment, DWA) for each trajectory frame. The root mean square deviation (RMSD) between all trajectory frames and the reference structure would be quite small, because DWA eliminates large systematic domain motions. In contrast, if the domain boundaries are guessed incorrectly then the RMSD after alignment would be large because of the remaining systematic displacements from the reference structure. Thus the RMSD relative to the reference structure after DWA could serve as a criterion for domain identification.

The algorithm. These considerations lead to the following formal algorithm of domain identification. Let us assume that there are N residues in the protein. The motions of the residues are monitored by the motions of their C_α atoms. The dynamic domains are organized into M hierarchical levels. Each domain of the level i is subdivided into two subdomains of the level $i + 1$ in such a way that the whole hierarchy is described by the binary tree. Nodes of the tree are enumerated by the multi-component index k (the details of indexing are explained below). Each node is repre-

sented by the vector $\vec{\sigma}_\kappa$ of size N. The elements of take the values 0, 1 or 2. All non-zero elements correspond to the residues, which belong to the domain encoded by the index k. The elements equal to 1 belong to the first subdomain of this domain, while the elements equal to 2 belong to the second subdomain. The domain tree for the protein with $N = 10$, $M = 3$ is shown schematically in Fig. 2.

DWA for the node k could be performed as following for each of MD trajectory frames:

I. The set of residues, which correspond to «1»s in $\vec{\sigma}_\kappa$ is aligned to the same set from the reference structure. Other residues are ignored.

II. The set of residues, which correspond to «2»s in $\vec{\sigma}_\kappa$ is aligned to the same set from the reference structure. Other residues are ignored.

III. The structures from steps 1 and 2 are combined. As a result the subdomains encoded with «1»s and «2»s are aligned to the reference structure separately. Domain identification is performed according to the following algorithm:

1. Start from the first hierarchical level (whole protein). There are no zero elements in at this level.

2. All non-zero elements of $\vec{\sigma}_\kappa$ are assigned to 1.

3. DWA is performed for each frame, mean RMSD over trajectory is computed.

4. For each non-zero element of $\vec{\sigma}_\kappa$:

a) elements are swapped (1 is substituted by 2 and vice versa);

b) DWA is performed for each frame; mean RMSD over trajectory is computed;

c) if RMSD decreases, keep new value of the element, otherwise revert to initial value.

5. Continue step 4 until RMSD decreases.

6. If current level of hierarchy is M then exit.

7. Go to next hierarchical level recursively:

a) make new vector $\vec{\sigma}_{\kappa,1}$, where all «2»s from $\vec{\sigma}_\kappa$ are set to zero and continue from step 2 with this new vector;

b) make new vector $\vec{\sigma}_{\kappa,2}$, where all «1»s from $\vec{\sigma}_\kappa$ are set to zero and continue from step 2 with this new vector.

Nodes of the tree are processed recursively starting from the first one, which represents the whole protein. For each node the parent domain (or the whole protein for the first node) is subdivided into two subdomains in such a way that the RMSD after DWA is minimized. Due to hierarchical application of the DWA the whole algorithm is called Hierarchical DWA (HDWA).

Flexibility of domains. The amount of internal flexibility in each of the domains can vary from zero (the domain is a rigid body) to almost free motions of several independent subdomains. This quantity for the particular domain k can be described by the flexibility coefficient R_k. The natural quantitative measure of flexibility in our technique is a difference between the whole-structure RMSDs before and after the DWA for given domain $R_k = RMSD_{before} - RMSD_{after}$. If $R_k \sim 0$ then the domain k is very rigid (there are no significant systematic internal motions, thus the division into subdomains does not eliminate them and does not change the RMSD). In contrast, if R_k is large then the domain is flexible and the motion of subdomains is significant. These motions are eliminated by the DWA which leads to large difference in RMSD before and after the alignment.

The flexibility usually decreases with an increase of the hierarchy level in our algorithm. This trend is explained by the fact, that the motions with the largest amplitude are eliminated on each level of hierarchy, thus the elimination of remaining motions on the next level usually leads to smaller change in RMSD.

The case of multiple subdomains. The number of subdomains of a particular domain is not known in advance. In principle it is possible to perform DWA for any number of subdomains, but the choice of an optimal number is problematic. Indeed, it is obvious that twenty subdomains would lead to smaller RMSD after DWA than two subdomains. The tests confirm that the RMSD decreases monotonously with an increase in the number of subdomains (data not shown). Thus, the division into two subdomains is the only justified option because it describes the dynamics of the parent domain by a minimal number of subdomains. In order to account domains with more than two subdomains the post-processing of the domain tree is performed.

As it was explained above the flexibility of domains usually decreases with an increase in hierarchical level. It is possible, however, that R of one of the subdomains is larger than R of the parent domain (for example $R_{11} < R_{112}$). In this case we assume that the parent domain

«11» is redundant and should be eliminated. Two subdomains of the redundant domain «111» and «112» are attached as children directly to the parent of eliminated domain (in this case to «1»). After such elimination domain «1» will have three subdomains, namely «12», «111» and «112». This procedure ensures that the flexibility decreases at higher hierarchical level and allows subdivision into more than two subdomains if necessary.

Relation to other techniques. The DynDom [14] domain identification method is probably the closest to HDWA. DynDom utilizes an alignment between two different structures of the same protein to extract information about domains, however the method of alignment differs from one used in HDWA. In DynDom short backbone segments of two structures are aligned to obtain the rotation matrices, which describe transformation from one structure to the other. The agglomerative clustering procedure is then used in the space of such rotations to identify the domains and the hinge regions. DynDom finds the largest dynamic domains, but can not reveal the hierarchy of subdomains inside them. In principle one may run DynDom on individual domains of the top level to search for subdomains, however no such attempts were made to our knowledge.

In contrast, HDWA is designed to describe the hierarchy of subdomains. The algorithm of alignment can also be classified as clustering, but the clustering is divisive rather than agglomerative. The largest domains are progressively subdivided into smaller subdomains until given depth of the hierarchy is reached. HDWA is better suited for proteins with complex conformational transitions, which are hard to describe by few domains on the single level of hierarchy. In contrast to DynDom, HDWA does not identify hinge axes and hinge residues automatically. Although an analysis of obtained domain hierarchy could easily provide the information about the hinge axes [23] we do not provide such analysis in this work.

HDWA is inspired by the HCCP technique, which also identifies the hierarchy of dynamic domains, but uses the matrices of residue-residue correlation of motions obtained from GNM normal modes calculations instead of MD trajectories. Both methods are conceptually similar, but operate on the data of different nature, which inevitably leads to different algorithms.

Another recently developed method, which is close to HDWA, is TIMME [25]. In this technique the hierarchy of the quasi-rigid clusters in the protein is identified using the fluctuations of the pair distances between the atoms. The HDWA compares the structures in terms of their RMSDs and uses divisive clustering, while the TIMME utilizes the distance fluctuations and agglomerative clustering. In principle, the approach used in HDWA may be more robust because the individual distance fluctuations could be more sensitive to the simulation setup and the force field than the average RMSDs. However, the systematic analysis of these issues is beyond the scope of this work.

Discussion. The problem of domain identification in proteins is more complex than it is usually thought. This complexity is clearly demonstrated by the large number of techniques, which are used to identify the domains and by the absence of one universally accepted technique. In this work we focused on the so-called dynamic domains, which represent the units of motion in the protein globules. In principle, the most precise method of identifying dynamic domains is the analysis of trajectories of MD simulations. However, currently there is no universally accepted method which is designed specifically for this task. The motions of proteins are extremely complex and occur on very different time and space scales. These motions constitute a natural hierarchy, which often correlates with the structural features of the protein globules. Accounting for this natural hierarchy is proven to be useful in domain identification techniques based on simplified elastic network models [20, 22]. In the current work we applied the idea of hierarchically organized dynamic domains to the analysis of MD trajectories. The idea of our HDWA technique is borrowed from the standard structural alignment algorithm, which is routinely used to eliminate rotations and translations of the whole molecules in MD trajectories. We eliminate certain systematic motions from the trajectory by aligning each dynamic domain of given hierarchical level separately to the reference structure. The remaining deviations from the reference structure represent the motions on the next hierarchical level, which could be eliminated by applying this procedure recursively. The division into subdomains is achieved by minimizing mean RMSD after domain-wise alignment. Such division is

optimal in the sense that it eliminates as many systematic motions inside the parent domain as possible.

The advantage of our technique is that it is completely model-free. The trajectories are analyzed without any assumptions. As a consequence the dynamic domains found by our technique are «intrinsic» to the studied protein under given parameters of MD simulations. In contrast to HCCP or other techniques, which are based on the single protein structure and utilize highly simplified models of the protein dynamics, the HDWA is based on the realistic motions, which occur during MD simulations. This allows studying the influence of various factors (such as ionic concentrations, pH, point mutations, etc.) on the dynamic domains.

Conclusion. New domain identification technique called the Hierarchical Domain-Wise Alignment has been developed. HDWA is designed for identifying the hierarchy of dynamic domains in proteins using the trajectories of MD simulations. HDWA is a model-free technique, which analyzes the motions in MD trajectories without introducing any simplified model of the protein dynamics.

Acknowledgement. Prof. Valery N. Kharkyanen is deeply acknowledged for discussion and the useful comments.

С. О. Єсилевський

Новий метод ідентифікації ієрархії доменів у білках на основі даних молекулярної динаміки

Резюме

Мета. *Незважаючи на велику кількість існуючих підходів до ідентифікації доменів у білках, на сьогодні немає єдиного загальновизнаного методу, який міг би визначати ієрархію динамічних доменів на основі даних молекулярної динаміки. Метою роботи є створення такого методу*. ***Методи***. *Домени визначають рекурсивним елімінуванням систематичних рухів у траєкторіях молекулярної динаміки*. ***Результати***. *Розроблено новий метод визначення ієрархії доменів на основі даних молекулярної динаміки, який отримав назву ієрархічного доменного вирівнювання (HDWA).*

Ключові слова: динамічні домени, ідентифікація доменів, ієрархічне доменне вирівнювання, молекулярна динаміка.

С. А. Есилевский

Новый метод идентификации иерархии доменов в белках на основе данных молекулярной динамики

Резюме

Цель. *Несмотря на большое количесто существующих подходов к определению доменов в белках, на сегодня нет единого общепризнанного метода, с помощью которого можно было бы идентифицировать иерархию динамических доменов на основе данных молекулярной динамики. Целью работы является разработка такого метода*. ***Методы***. *Домены определяют рекурсивным элиминированием систематических движений в траекториях молекулярной динамики*. ***Результаты***. *Создан новый метод определения иерархии доменов на основе данных молекулярной динамики, получивший название иерархического доменного выравнивания (HDWA).*

Ключевые слова: динамические домены, идентификация доменов, иерархическое доменное выравнивание, молекулярная динамика.

REFERENCES

1. *Janin J., Wodak S. J.* Protein modules and protein-protein interaction. Introduction // Adv Protein Chem.–2002.–**61**.–P. 1–8.
2. *Janin J., Wodak S. J.* Structural domains in proteins and their role in the dynamics of protein function // Prog. Biophys. Mol. Biol.–1983.–**42**, N 1.–P. 21–78.
3. *Ito K., Uyeda T. Q., Suzuki Y., Sutoh K., Yamamoto K.* Requirement of domain-domain interaction for conformational change and functional ATP hydrolysis in myosin // J. Biol. Chem.–2003.–**278**, N 33.–P. 31049–31057.
4. *Popp S., Packschies L., Radzwill N., Vogel K. P., Steinhoff H. J., Reinstein J.* Structural dynamics of the DnaK-peptide complex // J. Mol. Biol.–2005.–**347**, N 5.–P. 1039–1052.
5. *Zhang X. J., Wozniak J. A., Matthews B. W.* Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme // J. Mol. Biol.–1995.–**250**, N 4.–P. 527–552.
6. *Gerstein M., Anderson B. F., Norris G. E., Baker E. N., Lesk A. M., Chothia C.* Domain closure in lactoferrin. Two hinges produce a see-saw motion between alternative close-packed interfaces // J. Mol. Biol.–1993.–**234**, N 2.–P. 357–372.
7. *Falquet L., Pagni M., Bucher P., Hulo N., Sigrist C. J., Hofmann K., Bairoch A.* The PROSITE database, its status in 2002 // Nucl. Acids Res.–2002.–**30**, N 1.–P. 235–238.
8. *Nagar B., Bornmann W. G., Pellicena P., Schindler T., Veach D. R., Miller W. T., Clarkson B., Kuriyan J.* Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571) // Cancer Res.–2002.–**62**, N 15.–P. 4236–4243.
9. *Schmitt L., Tampe R.* Structure and mechanism of ABC transporters // Curr. Opin Struct. Biol.–2002.–**12**, N 6.–P. 754–760.
10. *Fischer K. F., Marqusee S.* A rapid test for identification of autonomous folding units in proteins // J. Mol. Biol.–2000.–**302**, N 3.–P. 701–712.
11. *Anselmi C., Bocchinfuso G., Scipioni A., De Santis P.* Identification of protein domains on topological basis // Biopolymers.–2001.–**58**, N 2.–P. 218–229.
12. *Nichols W. L., Rose G. D., Ten Eyck L. F., Zimm B. H.* Rigid domains in proteins: an algorithmic approach to their identification // Proteins.–1995.–**23**, N 1.–P. 38–48.
13. *Wriggers W., Schulten K.* Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates // Proteins.–1997.–**29**, N 1.–P. 1–14.
14. *Hayward S., Berendsen H. J.* Systematic analysis of domain motions in proteins from conformational change: new results

on citrate synthase and T4 lysozyme // Proteins.–1998.–**30**, N 2.–P. 144–154.
15. *Hinsen K.* Analysis of domain motions by approximate normal mode calculations // Proteins.–1998.–**33**, N 3.–P. 417–429.
16. *Hinsen K., Thomas A., Field M. J.* Analysis of domain motions in large proteins // Proteins.–1999.–**34**, N 3.–P. 369–382.
17. *Tama F., Gadea F. X., Marques O., Sanejouand Y. H.* Building-block approach for determining low-frequency normal modes of macromolecules// Proteins.–2000.–**41**, N 1.–P. 1–7.
18. *Kundu S., Sorensen D. C., Phillips G. N., Jr.* Automatic domain decomposition of proteins by a Gaussian Network Model // Proteins.–2004.–**57**, N 4.–P. 725–733.
19. *Sista R., Brinda K. V., Vishveshwara S.* Identification of domains and domain interface residues in multidomain proteins from graph spectral method // Proteins: Structure, Function, and Bioinformatics.–2005.–**59**, N 3.–P. 616–626.
20. *Yesylevskyy S. O., Kharkyanen V. N., Demchenko A. P.* Hierarchical clustering of the correlation patterns: New method of domain identification in proteins // Biophys. Chem.–2006.–**119**, N 1.–P. 84–93.
21. *Yesylevskyy S. O., Kharkyanen V. N., Demchenko A. P.* Dynamic protein domains: identification, interdependence and stability // Biophys. J.–2006.–**91**, N 2.–P. 670–685.
22. *Yesylevskyy S. O., Kharkyanen V. N., Demchenko A. P.* The change of protein intradomain mobility on ligand binding, is it a commonly observed phenomenon? // Biophys. J.–2006.–**91**, N 8.–P. 3002–3013.
23. *Yesylevskyy S. O., Kharkyanen V. N., Demchenko A. P.* The blind search for the closed states of hinge-bending proteins // Proteins: Structure, Function, and Bioinformatics.–2007.–**71**, N 2.–P. 831–843.
24. *Yesylevskyy S. O., Kharkyanen V. N.* New approaches to slow dynamics of protein domains // Ukr. J. Phys.–2009.–**54**, N 1–2.–P. 109–116.
25. *Menor S. A., de Graff A. M. R., Thorpe M. F.* Hierarchical plasticity from pair distance fluctuations // Phys. Biol.–2009.–**6**, N 3.–P. 036017.
26. *Berendsen H. J. C.* Bio-molecular dynamics comes of age // Science.–1996.–**271**, N 5251.–P. 954–955.
27. *Lange O. F., Grubmuller H.* Generalized correlation for biomolecular dynamics // Proteins: Structure, Function, and Bioinformatics.–2006.–**62**, N 4.–P. 1053–1061.
28. *Amadei A., Linssen A. B. M., Berendsen H. J. C.* Essential dynamics of proteins. // Prot. Struct. Funct. Genet.–1993.–**17**, N 4.–P. 412–425.

UDK 577.322
Received 26.11.09