BIOINFORMATICS

Identification of hierarchy of dynamic domains in proteins: comparison of HDWA and HCCP techniques

S. O. Yesylevskyy

Institute of Physics of National Academy of Sciences of Ukraine
46, Prospect Nauky, Kiev, Ukraine, 03680

yesint3@yahoo.com

Aim. *There are several techniques for the identification of hierarchy of dynamic domains in proteins. The goal of this work is to compare systematically two recently developed techniques, HCCP and HDWA,on a set of proteins from diverse structural classes.* ***Methods***. *HDWA and HCCP techniques are used. The HDWA technique is designed to identify hierarchically organized dynamic domains in proteins using the Molecular Dynamics (MD) trajectories, while HCCP utilizes the normal modes of simplified elastic network models.* ***Results***. *It is shown that the dynamic domains found by HDWA are consistent with the domains identified by HCCP and other techniques. At the same time HDWA identifies flexible mobile loops of proteins correctly, which is hard to achieve with other model-based domain identification techniques.* ***Conclusion***. *HDWA is shown to be a powerful method of analysis of MD trajectories, which can be used in various areas of protein science.*

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

Introduction. The method of Hierarchical Clustering of Correlation Patterns (HCCP) was developed for identifying dynamic domains in proteins [1]. HCCP is the only existing technique, which identifies the hierarchy of dynamic domains. Each dynamic domain can be divided into smaller relatively independent subdomains of next hierarchical level and so on. The HCCP technique was successful in revealing the statistics of dynamic domain in PDB [2], in finding the candidate proteins for biosensor design [3] and in simulating domain closure in the hinge-bending proteins [4]. Despite these successful application the HCCP technique possesses a serious limitation. It depends on the matrices of residue-residue correlations of motion, which should be computed by other techniques. It was shown that the Gaussian Network Model (GNM) [5–9] is an optimal choice for constructing such matrices in the case when a single crystal structure of a protein is available. However, the usage of GNM (or any other technique based on the normal modes calculations) restricts the sampled protein motions to small-amplitude harmonic displacements around some reference structure [10, 11]. As a result only tiny part of the protein conformational space could be described. The dynamic domains computed from the correlations of such restricted motions may not correspond to the pattern of large-amplitude inharmonic dynamics of real proteins. Certain techniques, such as DynDom [12] utilize the differences between two alternative structures of a protein or between several frames from the trajectories of Molecular Dynamics (MD) simulations, which allows to take into account large conformational displacements.

Table 1
The details of molecular dynamics simulations

PDB code	Number of residue	Number of water molecule	Length of trajectory, ns	Length of equilibrated part of trajectory, ns	Number of frames used in HDWA
1FS3	124	8788	18	8	26
1CLL	144	3716	20	5	16
2LAO	238	8592	20	10	27
1AO6	578	23504	180	50	12

However these techniques do not reveal the hierarchical arrangement of dynamic domains.

Recently the Hierarchical Domain-Wise Alignment technique (HDWA) has been developed. It is conceptually similar to the HCCP technique [1], 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 identifies a hierarchy of dynamic domains from MD trajectories or any other sets of atomic coordinates and allows estimating stability and interdependence of domains.

In the current work we compare systematically the HDWA and HCCP techniques using the set of four test proteins of different structural classes. A comparison with the widely used DynDom technique is also performed.

Theory and methods. *Test proteins*. Four proteins were selected as a test set – human calmodulin (PDB code 1CLL) [13], human serum albumin (PDB code 1AO6) [14], lysine-, arginine-, ornithine-binding protein (LAOBP) (PDB code 2LAO) [15] and bovine pancreatic ribonuclease A (PDB code 1FS3) [16]. These proteins belong to different structural classes and cover a wide range of sizes (from 124 residues in 1FS3 to 578 in 1A06).

Molecular dynamics simulations. All MD simulations were performed using Gromacs 4.0 suit of programs [17]. All four test proteins were simulated under NPT conditions at the temperature of 300 K and the pressure of 1 bar maintained by the Berendsen thermostat and the Berendsen barostat respectively [18]. GROMOS G43a2 force field for the proteins [19] and the SPC model for water [12] were used. The bond lengths in protein were constrained using the LINCS algorithm [20]. The water molecules were constrained using SETTLE [21]. The fourth-order PME algorithm [22] with the cut-off of 1 nm was used for computations of electrostatic interactions. The time step of 2 fs was used in all cases except the human serum albumin, which was simulated with the time step of 4 fs after increasing the masses of hygrogen atoms to 4 a. u. and decreasing the masses of the corresponding heavy atoms [23]. The number of water molecules, the length of the trajectories and the number of frames used in HDWA for all studied proteins are summarized in Table 1. The frames used in HDWA were extracted from the equilibrated parts of the trajectories at equal intervals. The quality of equilibration was controlled by monitoring backbone RMSD and the secondary structure content of the proteins.

Choice of the reference structure. If the molecular system subjected to MD simulation is well-equilibrated, it samples the ensemble of states, which are all equally suitable as a reference structure for domain-wise alignment. The choice of any single frame as a reference means that HDWA will attempt to transform all frames of the trajectory to this selected structure, which will inevitably introduce a bias. Indeed, in this case the motions of domains, which describe the transitions between other trajectory frames, are not taken into account. In order to avoid such bias the structure averaged over whole trajectory is used as a reference. The common argument against the usage of average structures is their «unphysical» nature. Indeed, the average structure may contain sterical clashes of atoms, unusually long bonds, etc. This may constitute a significant problem in the methods, which rely on correctness of the protein topology. However, HDWA does not suffer from this problem because it uses only the geometrical positions of atoms regardless of any «unphysical» contacts or bonds.

Technical details. The HDWA algorithm was implemented in custom C++ program using Pteros molecular modeling library (https://sourceforge.net/projects/pteros/). VMD [24] is used for visualization.

Results and discussion. *Top-level domains*. The boundaries of top-level domains identified by the HDWA, HCCP and DynDom techniques were compared. In the case of DynDom, which needs two struc-

Table 2
Comparison of the domain boundaries obtained in HDWA, HCCP and DynDom techniques

Protein	HDWA		HCCP		DynDom	
	Domain 1	Domain 2	Domain 1	Domain 2	Domain 1	Domain 2
Ribonuclease A	Flexible loops	Body of the globule	Identified as a single domain protein		No domains found	
Calmodulin	1–69	70–144	1–74	75–144	7–75	76–144
Lysine-, arginine-, ornithine-binding protein	1–89, 192–238	90–191	1–90, 192–238	91–191	3–90, 191–236	91–190
Human serum albumin	5–208, 230–294, 464–468	209–229, 295–464, 469–582	5–200, 227–293, 463–469	201–226, 294–462, 470–582	7–194, 283–284	195–282, 285–580

tures to identify the domains, two alternative crystal structures were used for each of the test proteins (1FS3 and 4RAT for ribonuclease A; 1CLL and 1CDL for calmodulin; 2LAO and 1LST for LAOBP; 1AO6 and 2BXP for serum albumin). The results of comparison are summarized in Table 2.

In the case of calmodulin the boundary between the domains is correctly identified by all techniques to be between the residues 69 and 75. The discrepancy is easily explained by the fact that the long helix, which connects two domains, is rather featureless in terms of structure and dynamics.

LAOBP is a classical hinge-bending protein, which exhibits large displacement of domains around well-defined hinge. The domain boundary in LAOBP is very well defined, thus it is not surprising that all techniques find it correctly with the difference of 1–2 residues.

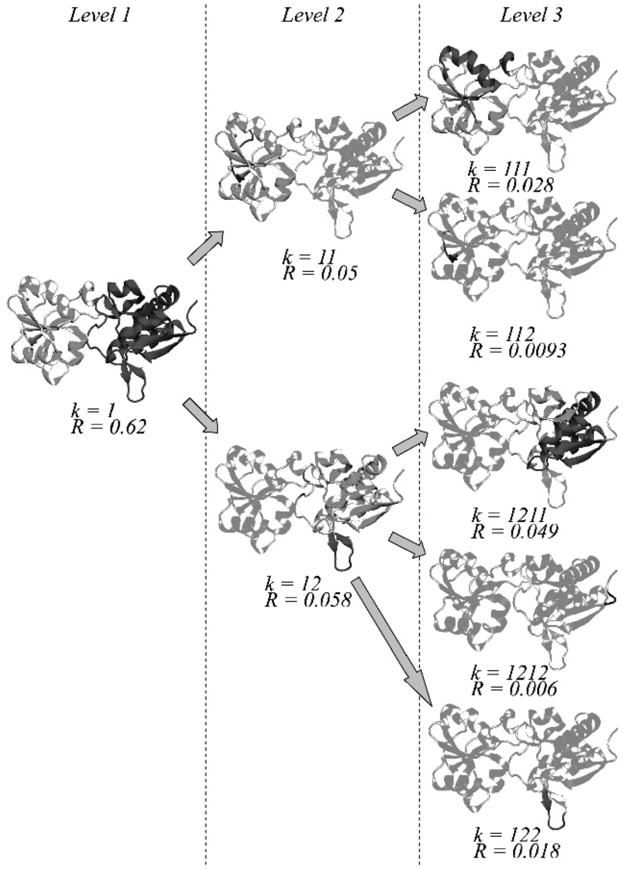
The human serum albumin is the most interesting among the studied proteins in terms of its domain organization. This protein is quite large and exhibits complex multicomponent dynamics. It also contains many flexible unstructured loops, which are important for its functioning. All three techniques find two top-level domains in serum albumin, however their boundaries are significantly different. HCCP and HDWA produce similar results with three continuous segments in each domain. The boundaries of these segments are shifted by up to 8 residues, but the overall arrangement is the same. DynDom identifies only two segments in each domain. It is necessary to note that the DynDom domain assignment for serum albumin is rather unreliable. It depends significantly on the choice of two alternative structures, which are used for domain identification (data not shown). This may be explained by high flexibility of serum albumin.

Subdomains. Both HDWA and HCCP technique are able to identify the subdomains of several hierarchical levels. However, it is impossible to compare these techniques on the level-by-level basis because of different algorithms of domain identification. The particular subdomain identified by HDWA at, say, level 3 may appear in HCCP at level 7 or does not appear at all. Thus the following procedure of comparison was used. HDWA was run with 6 hierarchical levels for all the proteins studied. Each subdomain found by HDWA on each level was matched with all the subdomains identified by HCCP for the same protein at the levels from 1 to 50. Matching was performed in terms of the Hamming distance between the binary vectors, which represent the domains. After this procedure, the mean mismatches for each HDWA hierarchical level were computed (Table 3).

The mismatches for different hierarchical levels differ substantially in different test proteins. In LAOBP and calmodulin the mismatch of the first-level domains is very small, while the domains of the levels 2–5 differ significantly in HCCP and HDWA. The mismatch decreases again for level 6. The same trend is observed for serum albumin. The mismatch of the first-level domains looks large (20 residues). However, this difference actually is not so dramatic because of large size of this protein and the fact that each of first-level domains consists of three pieces in terms of the sequence. The reason of this intriguing trend becomes evident after visual inspection of the subdomains identified by HCCP and HDWA. Typically small regions

Table 3
Mean mismatch (in residues) between HDWA domains of different level and corresponding HCCP domains

Protein	HDWA hierarchy level					
	1	2	3	4	5	6
Ribonuclease A	33.0	14.7	13.0	10.5	8.2	6.0
Calmodulin	3.5	9.2	13.6	9.4	7.9	5.7
LAOBP	1.0	11.0	13.9	14.2	15.6	8.2
Human serum albumin	20.0	40.5	43.6	25.7	16.6	10.3



HDWA domains in LAOBP for hierarchical levels *1–3*. The subdomains are colored black and white on each level. The parts of the protein, which do not belong to the current domain, are shown transparent. The domain indexes and the values of flexibility *R* are shown

around the hinge residues are cut off the largest domains on the second level of hierarchy in HCCP. The bodies of domains start to fragment into several subdomains on higher levels of hierarchy. These subdomains correlate rarely with the flexible loops and other highly mobile regions in the protein because of limitations of the underlying elastic network model. In contrast, HDWA subdivides the domains of the first level according to the mobility of their structural elements in the course of MD. Flexible fluctuating loops are assigned to one subdomain of the second level, while relatively rigid body of the domain is assigned to another subdomain (Figure). The same is true for subsequent levels of hierarchy until the subdomains become small enough to cover a single element of the secondary structure or individual loop. Such basic structural elements are identified by both HCCP and HDWA (although on different hierarchical levels). Thus the mismatch decreases for high levels of hierarchy.

The ribonuclease A is an exception among other studied proteins because it does not contain pronounced domains of the first level. Thus the mismatch is the largest for the first-level domains and decreases for higher hierarchical levels. In HDWA case the globule is subdivided into flexible loops and the rigid core at the first level of hierarchy. In the case of HCCP the mobility of loops is not detected and the domains of the first level do not correlate with the domains identified by HDWA.

The HDWA technique has some limitations. It is slow in comparison to other techniques due to expensive exhaustive search performed computationally for each domain subdivision. Typically, run time for the test proteins used in this work is between 5 and 30 min on fast office workstations for ~10–20 trajectory frames. This time increases rapidly with an increase in the number of frames.

However, the MD simulations themselves are typically 3–4 order of magnitude slower, thus the performance of HDWA is not critical. Another disadvantage is the character of domain subdivision. Each domain is subdivided into exactly two subdomains, which is not always the case in reality. However, as it was explained above, this is the only unbiased way of division (division into larger number of subdomains raises the problem of «overfitting»). The post-processing of the domain tree eliminates this problem partially by ensuring that the flexibility of domains increases with the increase of the hierarchical level. After the post-processing some domains may possess more than two subdomains.

HDWA can also be viewed as a powerful method of analysis of MD simulations, which extracts information about the hierarchy of the protein dynamics from the «mess of trajectories» for individual atoms. Our technique can be used in concert with the essential dynamics and other well established analysis techniques when the information about the hierarchy of domain motions is required. Our method is expected to be especially useful for large complex proteins. Such proteins possess the dynamics, which is unlikely to be described adequately at the single level of hierarchy. HDWA is the technique revealing the whole hierarchy of motions present in MD trajectories for such proteins.

Conclusion. The HDWA and HCCP methods of domain identification are tested on four proteins from different structural classes. It is shown that the number and the boundaries of large dynamic domains are consistent in both techniques and correspond well to the data of widely used DynDom technique. The hierarchy of dynamic domains in HDWA accounts for the presence of flexible loops and rigid regions, which is hard to achieve in other existing domain identification techniques. The domains found by HDWA may be considered as the most realistic units of the protein dynamics because they are identified using the data of atomistic MD simulations.

С. О. Єсилевський

Визначення ієрархії динамічних доменів у білках: порівняння методів HDWA та HCCP

Резюме

Мета. *Існує кілька методів для визначення ієрархії динамічних доменів у білках. Мета даної роботи полягала у проведенні систематичного аналізу двох нещодавно створених методів – HCCP та HDWA – на основі тестового набору білків з різних структурних класів.* ***Методи***. *Використано методи HDWA та HCCP. Перший розроблено для визначення ієрархії доменів з використанням траєкторій молекулярної динаміки, тоді як другий ґрунтується на нормальних коливаннях спрощеної еластичної моделі білка.* ***Результати***. *Встановлено, що динамічні домени, знайдені методом HDWA, добре узгоджуються з доменами, визначеними методом HCCP та із застосуванням інших підходів. У той же час HDWA правильно визначає рухливі петлі в білках, чого важко досягти іншим способом.* ***Висновки.*** *Показано, що HDWA є потужним методом аналізу траєкторій молекулярної динаміки для багатодоменних білків.*

Ключові слова:динамічні домени, ідентифікація доменів, HDWA, молекулярна динаміка.

С. А. Есилевский

Определение иерархии динамических доменов в белках: сравнение методов HDWA и HCCP

Резюме

Цель. *Существует несколько методов для определения иерархии динамических доменов в белках. Цель данной работы состояла в систематическом анализе двух недавно созданных методов – HCCP и HDWA – на основе тестового набора белков из разных структурных классов.* ***Методы***. *Использованы методы HDWA и HCCP. Первый разработан для определения иерархии динамических доменов с использованием траекторий молекулярной динамики, тогда как второй основан на нормальных колебаниях упрощенной эластичной модели белка.* ***Результаты***. *Установлено, что динамические домены, найденные методом HDWA, хорошо соответствуют доменам, определенным методом HCCP и с применением других подходов. В то же время HDWA правильно определяет подвижные петли в белках, чего трудно достичь другим способом.* ***Выводы***. *Показано, что HDWA является мощным методом анализа траекторий молекулярной динамики для многодоменных белков.*

Ключевые слова: динамические домены, идентификация доменов, HDWA, молекулярная динамика.

REFERENCES

1. *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.
2. *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.
3. *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.
4. *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.
5. *Atilgan A. R., Durell S. R., Jernigan R. L., Demirel M. C., Keskin O., Bahar I.* Anisotropy of fluctuation dynamics of proteins with an elastic network model // Biophys. J.–2001.–**80**, N 1.–P. 505–515.
6. *Bahar I., Atilgan A. R., Erman B.* Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential // Fold Des.–1997.–**2**, N 3.–P. 173–181.
7. *Yildirim Y., Doruker P.* Collective motions of RNA polymerases. Analysis of core enzyme, elongation complex and holoenzyme // J. Biomol. Struct. Dyn.–2004.–**22**, N 3.–P. 267–280.
8. *Keskin O.* Comparison of full-atomic and coarse-grained models to examine the molecular fluctuations of c-AMP dependent protein kinase // J. Biomol. Struct. Dyn.–2002.–**20**, N 3.–P. 333–345.
9. *Doruker P., Atilgan A. R., Bahar I.* Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to -amylase inhibitor // Proteins.–2000.–**40**, N 3.–P. 512–524.

10. *Levitt M., Sander C., Stern P. S.* Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme // J. Mol. Biol.–1985.–**181**, N 3.–P. 423–447.
11. *Hinsen K.* Analysis of domain motions by approximate normal mode calculations // Proteins.–1998.–**33**, N 3.–P. 417–429.
12. *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.
13. *Chattopadhyaya R., Meador W. E., Means A. R., Quiocho F. A.* Calmodulin structure refined at 1.7 A resolution // J. Mol. Biol.–1992.–**228**, N 4.–P. 1177–1192.
14. *Sugio S., Kashima A., Mochizuki S., Noda M., Kobayashi K.* Crystal structure of human serum albumin at 2.5 Å resolution // Protein Eng.–1999.–**12**, N 6.–P. 439–446.
15. *Oh B. H., Pandit J., Kang C. H., Nikaido K., Gokcen S., Ames G. F. L., Kim S. H.* Three-dimensional structures of the periplasmic lysine-, arginine-, ornithine-binding protein with and without a ligand // J. Biol.Chem.–1993.–**268**, N 15.–P. 11348–11355.
16. *Chatani E., Hayashi R., Moriyama H., Ueki T.* Conformational strictness required for maximum activity and stability of bovine pancreatic ribonuclease A as revealed by crystallographic study of three Phe120 mutants at 1.4 Å resolution // Protein Sci.–2002.–**11**, N 1.–P. 72–81.
17. *Hess B., Kutzner C., van der Spoel D., Lindahl E.* GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation // J. Chem. Theor. Comp.–2008.–**4**, N 3.–P. 435–447.
18. *Berendsen H. J. C., Postma J. P. M., van Gunsteren W. F., DiNola A., Haak J. R.* Molecular dynamics with coupling to an external bath // J. Chem. Phys.–1984.–**81**, N 8.–P. 3684–3690.
19. *van Gunsteren W. F., Kruger P., Billeter S. R., Mark A. E., Eising A. A., Scott W. R. P., Huneberger P. H., Tironi I. G.* Biomolecular Simulation: The GROMOS96 Manual and User Guide.–Groningen; Zurich: Biomos/Hochschul AG, 1996.–1044 p.
20. *Hess B., Bekker H., Berendsen H. J. C., Fraaije J. G. E. M.* LINCS: A linear constraint solver for molecular simulations // J. Computational Chem.–1997.–**18**, N 12.–P. 1463–1472.
21. *Miyamoto S., Kollman P. A.* Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models // J. Comp. Chem.–1992.–**13**, N 8.–P. 952–962.
22. *Tom D., Darrin Y., Lee P.* Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems // J. Chem. Phys.–1993.–**98**, N 12.–P. 10089–10092.
23. *Feenstra K. A., Hess B., Berendsen H. J. C.* Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems // J. Comp. Chem.–1999.–**20**, N 8.–P. 786–798.
24. *Humphrey W., Dalke A., Schulten K.* VMD – Visual Molecular Dynamics // J. Mol. Graph.–1996.–**14**, N 1.–P. 33–38.

UDC 577.322
Received 03.02.10