ISSN 0006-2979, Biochemistry (Moscow), 2011, Vol. 76, No. 9, pp. 1037-1042. © Pleiades Publishing, Ltd., 2011.
Published in Russian in Biokhimiya, 2011, Vol. 76, No. 9, pp. 1273-1279.

Identification of Latent Periodicity in Domains of Alkaline Proteases

Xiaofeng Ji[1], Jun Sheng[1], Fang Wang[1], Suzhen Zhang[2], Jianhua Hao[1], Haiying Wang[1], and Mi Sun[1]*

*[1]Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao 266071,
Shandong, China; fax: +86(532)8583-3525; E-mail: sunmi@ysfri.ac.cn*
[2]Department of Physics, Dezhou University, Dezhou 253023, Shandong, China

Received March 4, 2010
Revision received June 29, 2010

Abstract—Internal repeats in protein sequences have wide-ranging implications for the structure and function of proteins. A keen analysis of the repeats in protein sequences may help us to better understand the structural organization of proteins and their evolutionary relations. In this paper, a mathematical method for searching for latent periodicity in protein sequences is developed. Using this method, we identified simple sequence repeats in the alkaline proteases and found that the sequences could show the same periodicity as their tertiary structures. This result may help us to reduce difficulties in the study of the relationship between sequences and their structures.

DOI: 10.1134/S0006297911090082

Key words: latent periodicity, alkaline protease, correlation, primary sequence

The study of protein sequence periodicity is one of many approaches to protein sequence investigation. Study of the periodicity of proteins would be useful in the development of structural prediction methods and the understanding of mechanisms of protein evolution.

How and why do proteins exhibit obvious symmetry at the level of tertiary structures, and yet seldom periodicity in their primary sequences? A great effort has been made to explore this problem. Methods for the determination of distant repeats in protein sequences were introduced, which looked for internal periodicities by comparing the protein sequences to themselves with standard sequence–sequence alignment techniques [1-3]. Korotkov and coworkers developed information decomposition (ID), noise decomposition (ND), and cyclic alignment (CA) techniques to detect latent periodicity in protein families. It was particularly worth mentioning that a combination of ID and ND techniques revealed weak or latent periodicity [4-11]. Methods that used recurrence quantification analysis [12, 13] were also developed. Recurrence quantification analysis is a QSAR (Quantitative Structure–Activity Relationship)-related equivalent of a known sequence analysis tool that was originally called "distance chart analysis". Xiao's group used the method of modified recurrence plotting to detect periods in the sequences of β-trefoil [14], β-barrel [15], β-propeller [16], and Ig fold [17]. Over recent years some *de novo* repeat detection methods have been developed. Among all of these methods, REPRO [18], RADAR [19], TRUST [20], and HHrep [21] are especially efficient for detection of long repeats (more than ~10 residues long). However, they frequently fail to identify short repeats and do not distinguish between tandem and interspersed repeats. On the other side, XSTREAM [22] and MREPS [23] are well adapted for a large-scale search of protein repeats. However, these programs fail to identify some tandem repeats. Recently, Jorda and Kajava proposed a method called T-REKS [24] for protein tandem repeat identification, which was based on the analysis of distribution of short strings within the sequence by using a K-means algorithm.

To survey and examine the sequence characteristics in domains of the alkaline proteases, we report here a new method for detecting the latent periodicity in protein sequence. We demonstrate the presence of latent periodicity in sequences consistent with their tertiary structures. These results may support the hypothesis that large proteins are evolved by internal duplication and fusion.

* To whom correspondence should be addressed.

METHODS OF INVESTIGATION

Latent periodicity is defined here as latent similarity. For example, the sequence [RGNGIQINGK] [RGNGIQINGK] [RGNGIQINGK] is composed of three identical parts and we say that it has exact three-fold symmetry. Segments in the square brackets are the repeated segments. But real protein sequences do not have such exact symmetry. In fact, protein sequences appear nearly random [25, 26], although they exhibit periodicity at the level of tertiary structure.

The proposed algorithm detects repeats without any prior knowledge and guided by the idea of recurrence quantification analysis. It relies on a scheme to access the correlation coefficient threshold and statistical significance test (*p* value). The process of the method can be found in Fig. 1, and we will introduce it step by step.

Consider an arbitrary sequence $S = x_1x_2x_3...x_N$, where N is the length of the sequence and x_i is one of the 20 amino acids. First of all, we use the Kyte–Doolittle hydrophobicity value [27] to denote the corresponding amino acid, and a vector representation of the protein sequence, as $A = a_1a_2a_3...a_N$, is achieved. One constructs a set of all $(N - d + 1)$ possible segments of d consecutive symbols:

$$
\begin{aligned}
&A_1(d) = a_1a_2 \ldots a_{\mathrm{d}},\\
&A_2(d) = a_2a_3 \ldots a_{d+1},\\
&\quad \ldots \qquad\qquad\qquad\qquad (1)\\
&A_{\mathrm{i}}(d) = a_ia_{i+1} \ldots a_{d+i-1},\\
&\quad \ldots\\
&A_{N-\ d+1}(d) = a_N - a_{N-d+1} \ldots a_N.
\end{aligned}
$$

Then we calculate the correlations between each segment $A_i(d)$ $(1 \le i \le N - d + 1)$ and $A_j(d)$ $(j = i + 1, i + 2, ..., N - d + 1)$. $A_j(d)$ denotes the remaining segments along the sequence. The Pearson correlation coefficient *r* was used to evaluate the correlation degree, and the details of the equation can be found in our previous paper [28]. We set 0.5 for the threshold of our program, i.e. if the correlation of the two segments is not less than 0.5, the two segments here are defined as correlation. Of course, the threshold alone is not enough to prove the results we get is statistically significant. Hence, we do a further statistical test. If the *p* value is lower than 0.01, we consider that two segments are statistically significant. Then the information

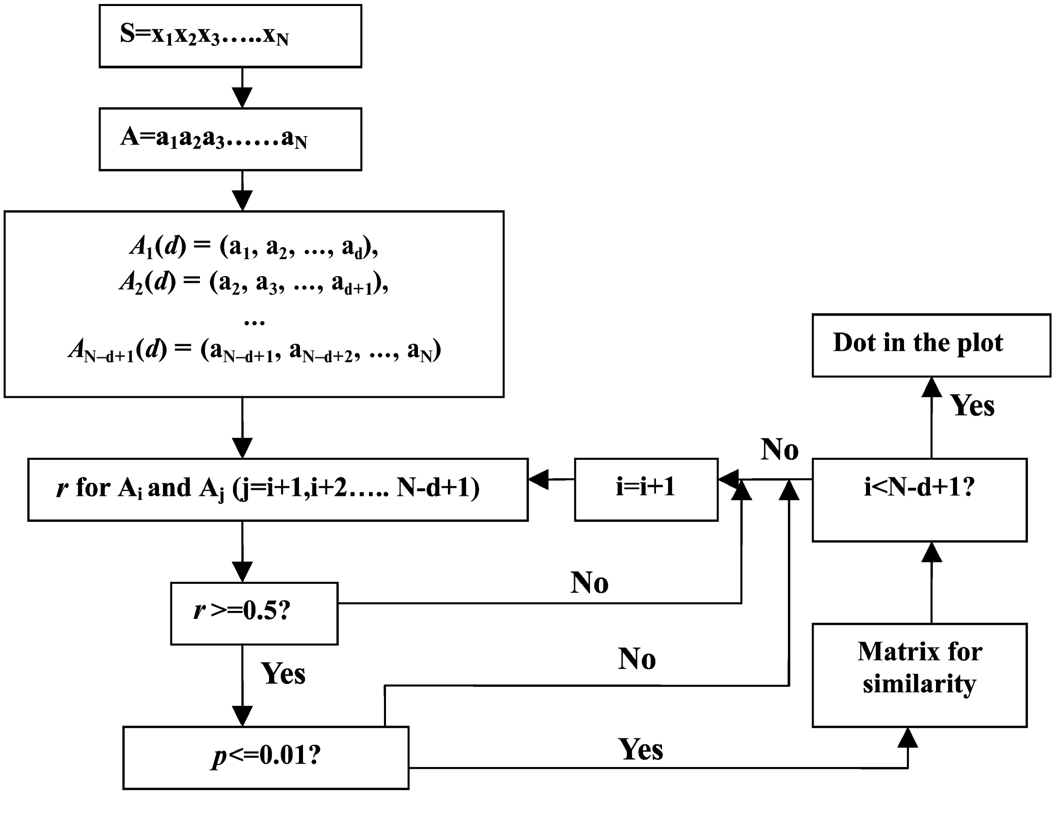
S=x1x2x3.....xN
A=a1a2a3......aN
A1(d) = (a1, a2, ..., ad),
A2(d) = (a2, a3, ..., ad+1),
...
AN–d+1(d) = (aN–d+1, aN–d+2, ..., aN)
r for Ai and Aj (j=i+1,i+2..... N-d+1)
i=i+1
No
i<N-d+1?
Yes
Dot in the plot
r>=0.5?
No
Yes
p<=0.01?
No
Yes
Matrix for similarity

Fig. 1. Overview of the sequence of steps in the method for finding latent periodicity.

Table 1. Sensitivity and accuracy for different selected subfamilies of PROPEAT

Folds	Repeats				Residues			
	RADAR	TRUST	REPRO	our	RADAR	TRUST	REPRO	our
	Sensitivity							
β-Trefoil	28.79	28.79	65.15	98.48	30.74	27.81	99.26	79.96
Jelly-roll	22.50	13.85	96.92	98.46	–	–	–	–
Ig like	9.38	15.63	90.63	93.75	8.64	17.69	110.23	99.90
TIM-barrel	23.75	22.50	50.00	52.43	19.54	57.72	107.94	89.71
Ferredoxin-like	0	0	100	100	0	0	62.79	83.72
Total	20.18	19.01	73.01	79.53	16.65	24.00	93.12	86.56
	Accuracy							
β-Trefoil	63.16	68.42	42.50	74.63	76.16	76.88	63.20	81.19
Jelly-roll	–	–	–	–	–	–	–	–
Ig like	50.00	70.00	44.83	77.78	8.33	52.34	48.29	60.92
TIM-barrel	47.05	56.25	43.74	62.02	29.64	39.88	29.57	40.11
Ferredoxin-like	0	0	50.00	100.00	0	0	92.00	97.29
Total	50.00	59.57	47.37	80.28	48.15	54.58	45.06	71.28

of the two segments was recorded in the similarity matrix. All of the correlated segments are detected before the similarity matrix is found. Then we dot in the plot according to the similarity matrix. If the protein structure has recurring substructures, the corresponding segments should be related and this can be obviously seen from the result plot.

To assess the repeat finding performance of our method, we compared it with the REPRO, RADAR, and TRUST methods. These methods are *de novo* repeat detection methods, and they all have their own web servers. The benchmark dataset consists of 132 proteins that possess approximate structural symmetry. All the 132 proteins, with the most known folds (β-trefoil, Jelly-roll, Ig like, TIM-barrel, or ferredoxin-like), were selected from the PROPEAT database [29].

Compared with the RADAR, TRUST and REPRO programs, our method showed high accuracy across all of the selected proteins (Table 1) for repeats and residues. Our method also showed a higher sensitivity for repeat prediction, although the sensitivity was lower than REPRO if repeat residues were counted.

RESULTS AND DISCUSSION

We take the sugar binding protein (PDB id 1ouw) as an example to show the latent periodicity of sequence using the proposed method. A monomer of Banlec (banana lectin) forms 12 β-strands in a β-prism-I fold, which contains three 4-stranded antiparallel β-sheets shaped like a prism with pseudo 3-fold symmetry (Fig. 2a). But its primary sequence appears to be irregular (Fig. 2b). From Fig. 2c we can easily find that the whole zone is partitioned into three parts. It demonstrates 3-fold symmetries in the primary sequence of this protein. Meanwhile, Fig. 2d shows the correlation coefficients of the first segment and any other segments along the sequence. The graph clearly showed that there are two notable peaks in position of 50 and 106. Combining this result with Fig. 2c, we can easily conclude that 50 and 106 are the cut-off points, and the repeated segments are V1-T50, K51-T106, and N107-K152. The three repeats are marked red, green, and yellow respectively, and we also do a multiple sequence alignment (Fig. 2b). The residue G is identical in all of the three segments. This analysis showed that our method was usefully to search for latent periodicity in protein sequences. We also identified the latent periodicity in alkaline proteases. Alkaline proteases are the most widely used enzymes in the detergent industry. They remove protein stains such as grass, blood, egg, and human sweat [30]. The tertiary structure has two distinct domains. The N-terminal domain is the proteolytic domain; it has an overall tertiary fold and active site metal ligation. The C-terminal domain mainly consists of β-strands. Figure 3 gives the

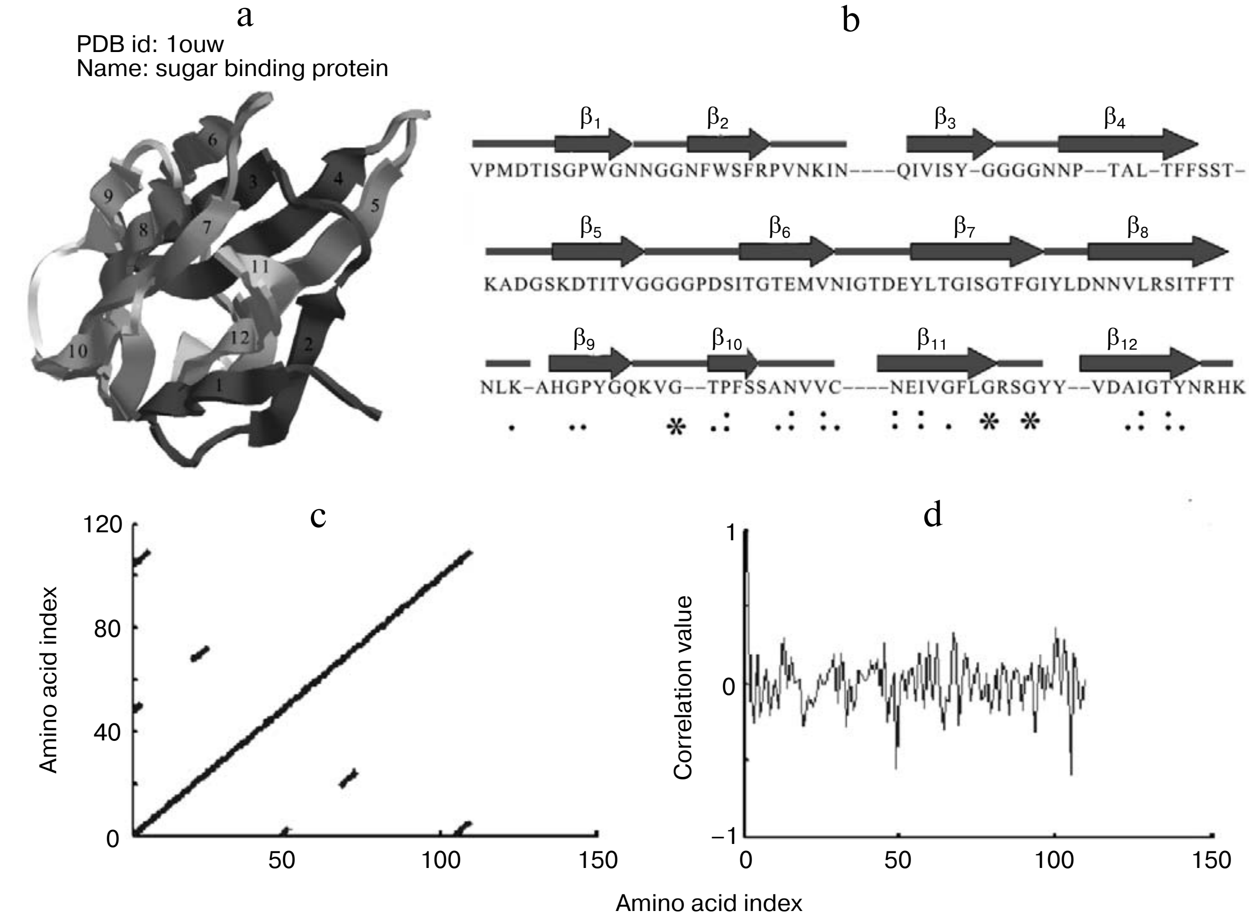

Fig. 2. Results plot of PDB id1ouw structure: a) tertiary structure; b) primary and secondary structures; c) recurrence plot; d) plot of Pearson's correlation coefficient between the first segment and all the other segments.

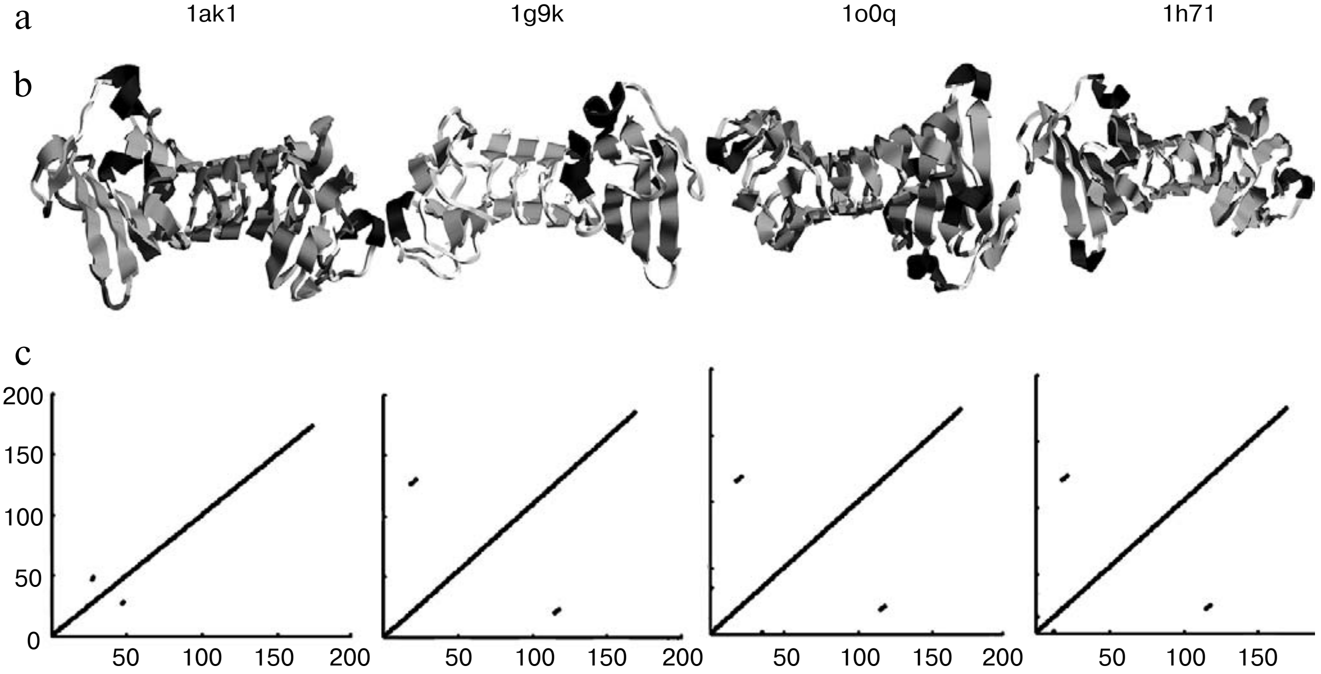

Fig. 3. Recurrence plot of four representative proteins: a) PDB id; b) tertiary structure; c) recurrence plot.

Table 2. Repeats and alignment results of repeated segments

PDB id	Source	Repeats	Alignment
1akl	*Pseudomonas aeruginosa* ifo 3080	F49-V98 A81-A132	-FS-QNQKINLNEKALSDVGGLKGNVSIAAGVTVENAIGGSGSDLLIGNDVANV AIGGSGSDLLIGNDVANVLKGGAGNDILYGGLGADQLWGGAGADTFVYGDIA-- :. ....: :.:.. . : * ** : .*: .:: **:*:* :: .*:*
1g9k	*Pseudomonas aeruginosa* tac ii 18	F14-T70 D109-G165	---FNSTADRDFYSATSSTDKLIFSVWDGGGNDTLDFS-GFSQN-QKINLTAGSFSDVGGMT DIIYGGGGADVLWGGTG-SDTFVFGAVSDSTPKAADIIKDFQSGFDKIDLTA--ITKLGG— :.. . ::..*. :*.::*.. ... .: *: .*... :**:*** ::.:**
1h71	*Pseudoalteromonas* tac ii 18	D16-M66 G111-G161	--DRDFYSATSSTDKLIFSVWDGGGNDTLDFS-GFSQN-QKINLTAGSFSDVGGM GGGADVLWGGTGSDTFVFGAVSDSTPKAADIIKDFQSGFDKIDLTA--ITKLG-- . *. . :.:*.::*.. ... .: *: .*... :**:*** ::.:*
1o0q	*Pseudoalteromonas* sp. tac ii 18	F11-N69 I108-V166	-FNSTADRDFYSATSSTDKLIFSVWDGGGNDTLDFS-GFSQN-QKINLTAGSFSDVGGMTGN IYGGGGADVLWGGTG-SDTFVFGAVSDSTPKAADIIKDFQSGFDKIDLTA--ITKLGGLNFV :.. . ::..*. :*.::*.. ... .: *: .*... :**:*** ::.:**:.
1TMQ	*Tenebrio* mealworm	G19-D47 V48-G76 V64-L92	VISGELSGGSCTGKSVTVG---D-NGSAD--ISLG---- TV-GD--NGS---ADISLGSAED-DGVLA--IHVNAKL- ---GS--QGF---VAFTNGG--DLNQNLNTGLPAGTYCD *. * .: * * : : .
1GNV	*Bacillus amyloliquefaciens*	V51-G101 A128-A178 V194-T244	VPSETNPFQDNNSHGTHVAGT-VLAVAPSASLYAVKVLGA--DGSG-QY-SWIING VSICST-LP-GNKYGA-KSGT-SMA-SPHVAGAAALILSKHPNWTNTQVRSSLENT AAVDKA-VASGVVVVA-AAGNEGTSGSSSTVGYPGKYPSV--IAVG-AVDSSNQRA .. . . . : :*. : :. . . . . * .

results plot of the representative domains and within these domains the successive β-strands compose mainly β_2-solenoids. It is surprising to find that most of these four proteins reveal the same two-fold symmetries as their tertiary structures. However, the slanting straight line, shown in result plot, is short when we detect long segment, and this trend becomes increasingly apparent as the selected fragment length increases. As we see in Fig. 3, the merged slanting straight line is very short; this is because the thing we are concerned about here is to get repeated segments as long as possible. The repeats and the alignment results of the repeated segments are listed in Table 2. As we know, repetitive units of the β-solenoid of the alkaline proteases have 18 residues, which is about the length of two β-strands, and our results show the length of about four strands. Structures 1TMQ and 1GNV possess the tertiary structure of β-sandwich and α-β-sandwich, respectively. The result plots also show certain sequence symmetries, and the results are shown in Table 2. From Table 2 we can easily find that the method we present here can find certain sequence symmetry signals, as their tertiary structures, in almost all of the selected sequences.

The results we showed here may suggest that protein sequences are not random and they exhibit periodic information in the properties of hydrophobicity. Moreover, the result may also give further evidence to the theory of that the symmetries at structure level are due to those at sequence level. This result is in agreement with the theory that modern proteins evolved by gene duplications and fusions. We hope that our method will be helpful for understanding the sequence–structure relationship of proteins.

This work is supported by the National Nature Science Foundation of China under Grant No. 30700613 and the Specific International Cooperation and Exchanges in Science and Technology (2011DFB30250).

REFERENCES

1. Rackovsky, S. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 8580-8584.
2. Heringa, J., and Argos, P. (1993) *Proteins*, **17**, 391-411.
3. Heger, A., and Holm, L. (2000) *Proteins*, **41**, 224-237.
4. Banerjee, N., Sarani, R., Ranjani, C. V., Sowmiya, G., Michae, D., Balakrishnan, N., and Sekar, K. (2008) *Bioinformation*, **3**, 28-32.
5. Korotkova, M. A., Korotkov, E. V., and Rudenko, V. M. (1999) *J. Mol. Model.*, **5**, 103-115.
6. Korotkov, E. V., Korotkova, M. A., and Kudryashov, N. A. (2003) *Physics Lett. A*, **312**, 198-210.
7. Turutina, V. P., Laskin, A. A., Kudryashov, N. A., Skryabin, K. G., and Korotkov, E. V. (2006) *Biochemistry* (Moscow), **71**, 18-31.
8. Laskin, A. A., Kudryashov, N. A., Skryabin, K. G., and Korotkov, E. V. (2005) *Comput. Biol. Chem.*, **29**, 229-243.
9. Turutina, V. P., Laskin, A. A., Kudryashov, N. A., Skryabin, K. G., and Korotkov, E. V. (2006) *J. Comput. Biol.*, **13**, 946-964.
10. Laskin, A. A., Skryabin, K. G., and Korotkov, E. V. (2007) *J. Proteome Res.*, **6**, 862-868.
11. Konopka, A. K. (2003) *Sequence Complexity and Composition* (Cooper, D. N., ed.) Nature Publishing Group Reference, London, pp. 217-224.
12. Giuliani, A., Benigni, R., Zbilut, J. P., Weber, C. L., Sirabella, P., and Colosimo, A. (2002) *Chem. Rev.*, **102**, 1471-1491.
13. Zbilut, J. P., Mitchell, J. C., Giuliani, A., Colosimo, A., Marwan, N., and Webber, C. L. (2004) *Physica A: Statist. Mech. Its Appl.*, **343**, 348-358.
14. Xu, R., and Xiao, Y. (2005) *Comput. Biol. Chem.*, **29**, 79-82.
15. Ji, X. F., Chen, H. L, and Xiao, Y. (2007) *Comput. Biol. Chem.*, **31**, 61-63.
16. Wang, X. C., Huang, Y. Z., and Xiao, Y. (2008) *J. Mol. Graph. Model.*, **26**, 829-833.
17. Huang, Y. Z., and Xiao, Y. (2007) *Proteins: Structure, Function, and Bioinformatics*, **68**, 267-272.
18. George, R. A., and Heringa, J. (2000) *Trends Biochem. Sci.*, **25**, 515-517.
19. Heger, A., and Holm, L. (2000) *Proteins*, **41**, 224-237.
20. Szklarczyk, R., and Heringa, J. (2004) *Bioinformatics*, **20**, 1311-1317.
21. Soding, J., Remmert, M., and Biegert, A. (2008) *Nucleic Acids Res.*, **34**, W137-W142.
22. Newman, A. M., and Cooper, J. B. (2007) *BMC Bioinformatics*, **8**, 382-400.
23. Roman, K., Ghizlane, B., and Gregory K. (2003) *Nucleic Acids Res.*, **31**, 3672-3678.
24. Jorda, J., and Kajava, A. V. (2009) *Bioinformatics*, **25**, 2632-2638.
25. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.*, **72**, 3907-3910.
26. Rahman, R. S., and Rackovsky, S. (1995) *J. Biophys.*, **68**, 1531-1539.
27. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.*, **157**, 105-132.
28. Ji, X. F., Wang, Y. J., Wang, H. Y., and Sun, M. (2008) *J. Theor. Biol.*, **255**, 316-319.
29. Shih, E. S. C., and Hwang, M. J. (2006) *Nucleic Acids Res.*, **34** (Web server issue), W95-W98.
30. Meagher, J. L., Winter, H. C., Ezell, P., Goldstein, I. J., and Stuckey, J. A. (2005) *Glycobiology*, **15**, 1033-1042.