Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 20, Issue Number 1, (2002)

Entropy of Protein Sequences: An Integral Approach

Alexei N. Nekrasov*

Shemyakin-Ovchinnikov Institute of
Bioorganic Chemistry
Russian Academy of Sciences
Miklukho-Maklaya St., 16/10, GSP-7
Moscow, 117997
Russia

http://www.jbsdonline.com

Abstract

Several classifications of protein spatial structures and their structural elements are known. This makes revealing of the relation between these structural elements and sequence fragments rather topical. The most important move in this direction would be the determination of positional sensitivity levels and ranges between the residues in protein sequences. In this work the Shannon-Weaver informational entropy was used as a disorder criterion for solving this problem. This entropy was computed as function of the distance between the amino acid residues in different sets of unhomological protein sequences. Similarity of this function for different sets of protein sequences was shown. Analysis of informational entropy allows detecting a long-range positional correlation (≥30) between the amino acid residues and oscillations with periods of 3.6 and 2.9. These oscillation periods correspond to periodicity of α- and $\mathbf{3_{10}}$-helices.

Introduction

At present the cumulated amount of protein structural data permits to make classifications of the spatial organization (1, 2) and to investigate the hierarchic organization of structural elements (3-5).

Due to this fact, the determination of relationship between spatial structural elements and the corresponding sequence fragments is very important. The first step in this direction is to find the levels and boundaries of inter-residues sensitivity (correlation) in protein sequences. This paper is focused on these issues.

The methodological basis of this study is Shannon-Weaver informational theory (6), already successfully applied to analysis of primary structures of proteins (7-11) and nucleic acids (12-15). In the present work, sets of the unhomological protein sequences were analyzed to find integral informational characteristics possibly reflecting the specificity of protein spatial structural organization levels in their sequences.

The search for integral characteristics of protein sequences must minimally depend on the functional and structural features inherent for individual protein families e.g., hemoglobins, lysozymes, cytochromes, etc., whose sequences are widely represented in the protein sequence databases. It is also desirable to confirm the reliability of the search for regularities using several UPSs significantly differing in size and composition. Hence, these UPSs must meet the following requirements:

(a) to contain only complete and reliable native protein sequences;
(b) to be large enough for the statistical reliability;
(c) not to include families of highly homologous protein sequences.

* Phone: 7 095 335 4366
Fax: 7 095 335 7103
E-mail: alne@ibch.ru, alexei_nekrasov@mail.ru

Abbreviation: UPS - unhomological protein set

The chapter *Annotated Entries* of the *Protein Identification Research (PIR)* database (16) satisfies these requirements.

Archive Data

Sets of sequences included into the chapter *Annotated Entries* of the *Protein Identification Research* database, releases 18, 27, and 49 were used. They include 5556, 12607, and 58089 protein sequences (1510026, 3417043 and 21699210 residues, respectively) and are designated in this work as BASE-I, BASE-II, and BASE-III.

Results and Discussion

In this study, the informational entropy was used as an integral criterion for determination the levels of the protein sequence organization. The informational entropy $\mathbf{S^k}$ (6) was computed from probability matrices $\mathbf{P^k_{i,j}}$ describing the occurrence of the **i-** and **j-**type residues separated by **k** positions in the sequences (see equation [1]). The computational scheme of probability matrices $\mathbf{P^k_{i,j}}$ for BASE-I, BASE-II, and BASE-III is presented in Figure 1.

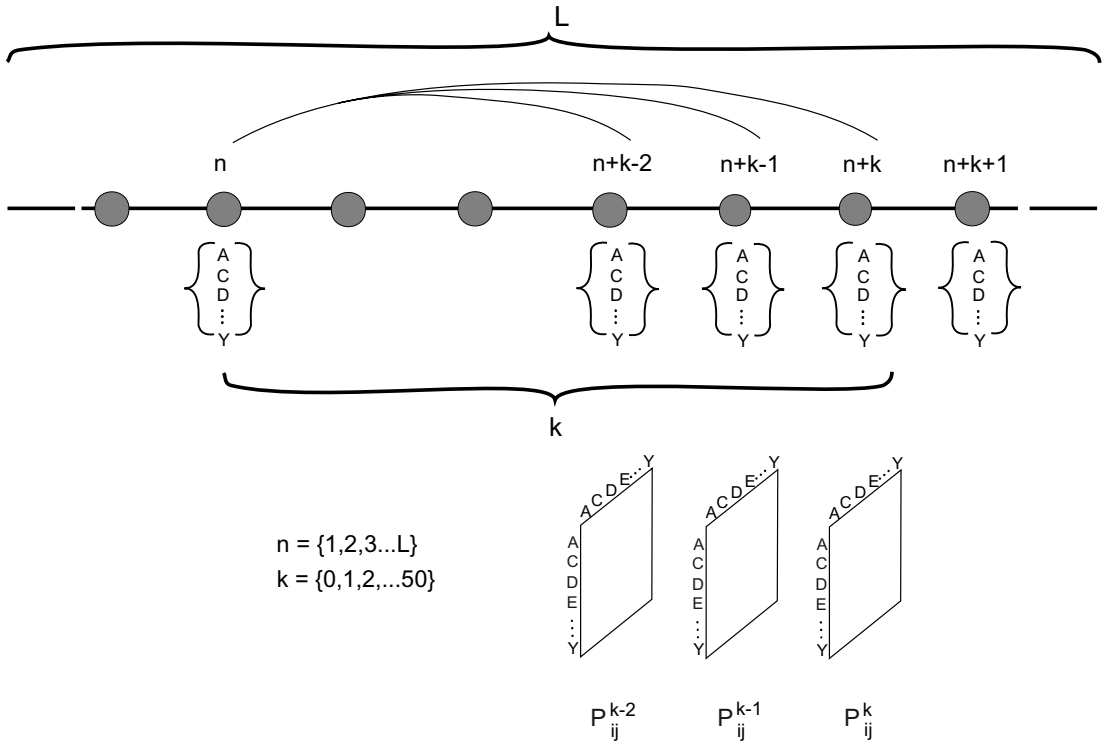


Figure 1: The computational scheme for determination the informational entropy $\mathbf{S^k}$ for BASE-I, BASE-II, and BASE-III is shown. **L** - the length of the protein sequence; $\mathbf{P^k_{i,j}}$ - the probability matrix of the amino acid residues of **i** and **j** types separated in the sequence by **k** positions.

The informational entropy $\mathbf{S^k}$ describes the informational significance of the probability matrix $\mathbf{P^k_{i,j}}$ of occurrence of the residues. Analysis of $\mathbf{S^k}$ alterations may allow determination of the characteristic values of **k** involved in the levels of organization realized in UPSs.

$$\mathbf{S^k} = -\sum_{i=1}^{20}\sum_{j=1}^{20} \mathbf{P^k_{i,j}} \log_2 \mathbf{P^k_{i,j}} \qquad [1]$$

The upper border of the range at **k=50** was selected basing on the fact that UPSs for small proteins (first of all, short inhibitors and toxins 50-60 residues long) would generate artifacts difficult for elimination at large **k** values. The informational entropy $\mathbf{S^k}$ versus distance **k** between amino acid residues for various UPSs was computed.

It was assumed *a priori* that the minimum value of the entropy $\mathbf{S^k}$ will be observed for the probability matrix $\mathbf{P^k_{i,j}}$ of the neighboring residues. For the convenience of analysis, dependencies of the informational entropy $\mathbf{S^k}$ of the distance **k** between residues were normalized at the value of informational entropy $\mathbf{S^1}$ of neighboring residues (Fig. 2).

The resulting absolute values of the informational entropy were ~8.4 and the absolute interval of their variation did not exceed 0.01. However, it is worth mentioning that the dependencies obtained were common for all tested UPSs and retained common characteristic features, e.g., the maximum and minimum positions. The coefficients of pair correlations for various dependencies $\mathbf{S^k/S^1}$ of the sequence sets ranged from 94% to 97%. This result seems rather important as these dependencies $\mathbf{S^k/S^1}$ were computed using the UPSs fundamentally differing in size and composition.

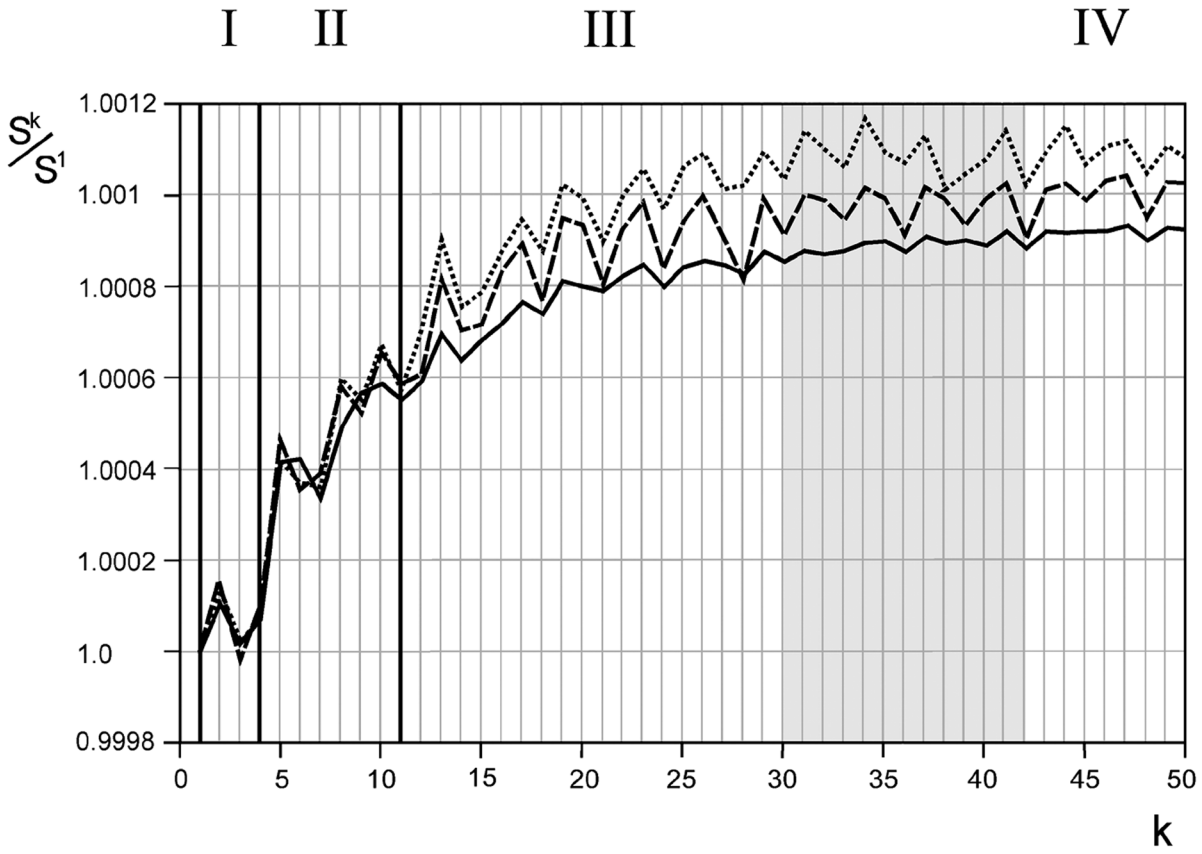


Figure 2: Dependence of informational entropy $\mathbf{S^k}$ normalized by entropy of neighboring residues $\mathbf{S^1}$ versus the distance **k** between amino acid residues. Solid, dash, and dotted lines show dependencies obtained using BASE-III, BASE-II, and BASE-I, respectively. Black vertical lines indicate the borders of REGION I, REGION II and REGION III. The $\mathbf{S^k/S^1}$ transition area (between REGION III and REGION IV) to the constant values and close to a maximum is given in gray.

As expected, minimum absolute values of the informational entropy 8.37451, 8.370002, and 8.360313 (for BASE-I, BASE-II, and BASE-III, respectively) were obtained for the neighboring residues in the sequence. A very close entropy value was observed for the residues separated by three positions. Their entropy in BASE-II was even lower than the entropy of the neighboring residues.

Let us consider the $\mathbf{S^k/S^1}$ dependencies shown in Figure 2. They can be regarded as a superposition of the oscillatory and **S**-like components. Fourier analysis of the $\mathbf{S^k}$ oscillating component allowed detecting two types of oscillation with periods of 3.6 and 2.9 residues. These values correspond to the periodicities of the α-helix and $\mathbf{3_{10}}$-helix. For the oscillating component, $\mathbf{S^k/S^1}$ maximum and minimum positions obtained correlate well for various UPSs. The only exception is the minimum at **k=9** observed for BASE-I and BASE-II and absent for BASE-III. This difference was caused by different occurrences of oscillations with the periods of 3.6 and 2.9 for these UPSs and was reflected in the ratio of the spectral densities for these oscillations. At **k≅9**, the oscillations with the periods 3.6 and 2.9 are in a phase

shift of 180°. The spectral densities of both oscillations for BASE-III are similar that leads to the absence of the minimum in the oscillating component of $\mathbf{S^k/S^1}$.

Within the **k** range tested, an amplitude of the oscillation component varies, decreasing with the rise of **k**. It is especially typical for **k≥11** and the BASE-III. It is remarkable that no periodicity was found inherent for the β-structure.

As for the **S**-like component, its variation interval can be divided into the following regions:

REGION I, the area of $\mathbf{S^k/S^1}$ values close to minimum and constant (for all UPSs) at **1≤k≤4**;

REGION II, the area of $\mathbf{S^k/S^1}$ values increasing and similar (for all UPSs) at **4<k≤11**;

REGION III, the area of increasing $\mathbf{S^k/S^1}$ values but different for various UPSs at **11<k≤30-42;**

REGION IV, the area of $\mathbf{S^k/S^1}$ values close to the maximum and constant (**k≥30** for BASE-I; **k≥36** for BASE-II; and **k≥42** for BASE-III).

It should be noted that the maximum $\mathbf{S^k/S^1}$ values significantly depended on sequence quantities included in UPS. This dependence manifested itself as a decrease in the maximum of $\mathbf{S^k/S^1}$ values when the sequence set size increased. For a larger set, the $\mathbf{S^k/S^1}$ values close to maximal values of informational entropy were observed at higher **k** values.

In conclusion, it would be important to demonstrate the identical nature of the informational entropy ($\mathbf{S^k}$) minima observed in different UPSs. For this purpose, differential maps were used which display deviations of the probability matrices $\mathbf{P^k_{i,j}}$ from the $\mathbf{P^0_{i,j}}$. The matrix $\mathbf{P^0_{i,j}}$ was calculated under the assumption that probability values correlate only with the amino acid composition of UPS. The regions where the probability values are higher or lower than the expected $\mathbf{P^0_{i,j}}$ are marked in black or white, respectively (Fig. 3). When the amino acid residues are similarly arranged along both axes, some characteristic 2D-patterns are formed on the differential maps. The similarity of 2D-patterns at identical **k** and different UPSs (Fig. 3) indicates that the positional variation of informational entropy is due to variations (of similar nature) in the probability matrices $\mathbf{P^k_{i,j}}$.

The differential probability maps for **k = 1, 3,** and **14** clearly show an identical nature of deviations in the paired probabilities of the occurrence of amino acid residues $\mathbf{P^k_{i,j}}$ from those of $\mathbf{P^0_{i,j}}$. The degree of coincidence of the probability matrix elements for any pair of the UPSs is within the interval 74.25% to 91.75% at various **k**. This confirms that the observed characteristic deviations of informational entropy $\mathbf{S^k}$ are inherent for any UPSs.

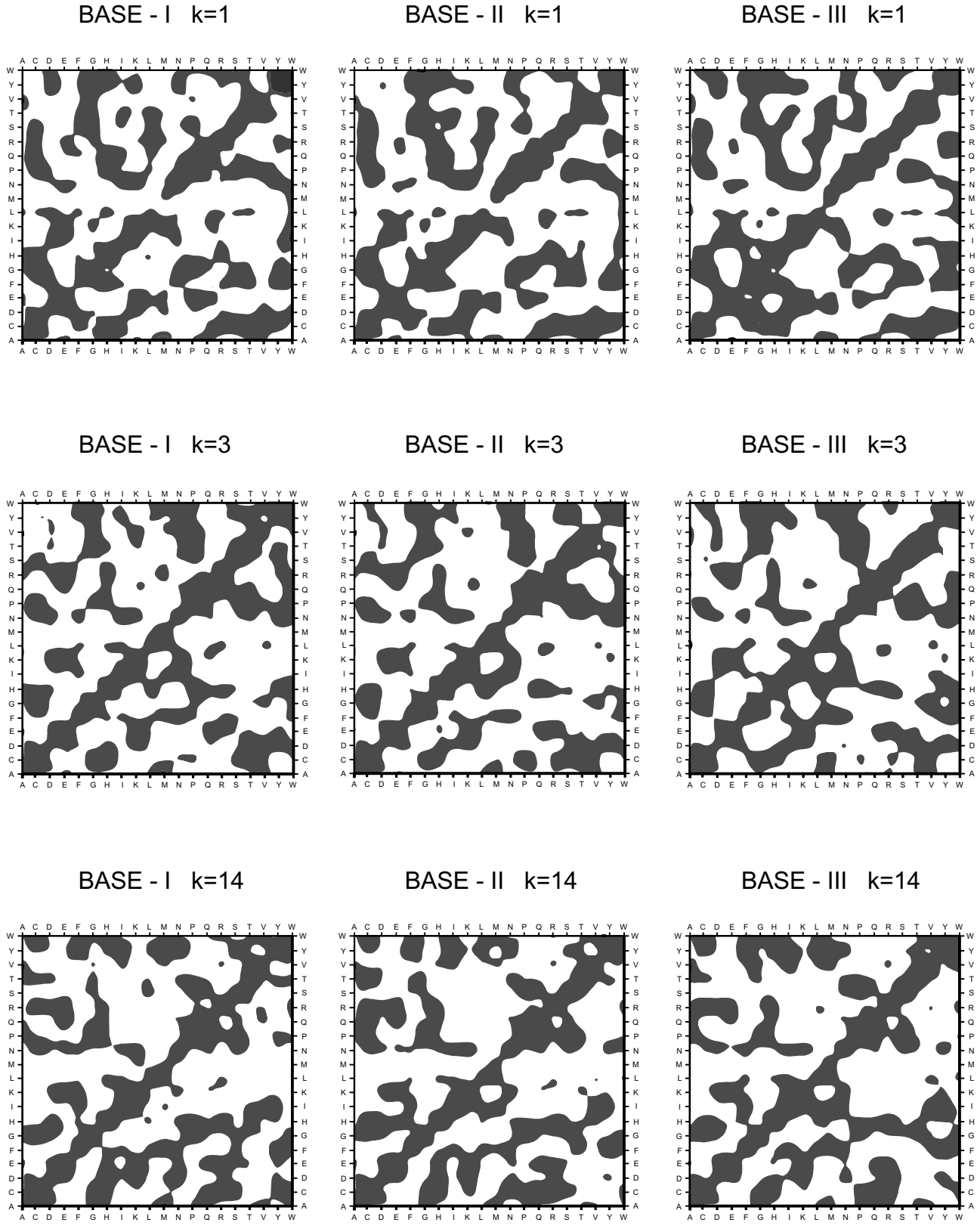
Analysis of informational entropy $\mathbf{S^k}$ in this paper leads to the following conclusions:

1. All UPSs have identical entropy characteristics;

2. A long-range (**≥30**) positional correlation (sensitivity) among amino acid residues occurs in protein sequences;

3. Only helix-like structural elements are encoded in protein sequences.

Figure 3: Differential maps describing deviation of the occurrences of pairs of residues probability matrix $\mathbf{P^k_{i,j}}$ from matrix $\mathbf{P^0_{i,j}}$, which were obtained at the assumption of their correlation only to the amino acid composition of the UPSs (BASE-I, BASE-II and BASE-III) at **k=1, 3** and **14**. The regions of $\mathbf{P^k_{i,j}}$ with higher or lower probability relative to the expected $\mathbf{P^0_{i,j}}$ were marked black or white.



Acknowledgements

The author is grateful to Prof. Yuri A. Berlin for his help in preparation of this manuscript.

References and Footnotes

1. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton, *Structure 5*, 1093-1108 (1997).
2. L. Lo Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin and C. Chothia, *Nucleic Acids Res. 28*, 257-259 (2000).
3. T. J. Yuschok and G. D. Rose, *Int. J. Peptide Protein Res. 21*, 479-484 (1983).
4. A. V. Efimov, *Prog. Biophys. Mol. Biol. 60,* 201-239 (1993).
5. R. Sowdhamini and T. L. Blundell, *Prot. Science 4*, 506-520 (1995).
6. C. F. Shannon and W. Weaver, *The Mathematical Theory of Communication,* University of Illinois Press, Urbana (1949).
7. H. P. Yockey, *J. Theor. Biol. 67*, 345-376 (1977).
8. J. F. Gibrat, J. Garnier and B. Robson, *J. Mol. Biol. 198*, 425-443 (1987).
9. P. S. Shenkin, B. Erman and L. D. Mastrandrea, *Proteins 11*, 297-313 (1991).
10. R. Farber, A. Lapedes and K. Sirotkin, *J. Mol. Biol. 226*, 471-479 (1992).
11. R. M. Williamson, *J.Theor. Biol. 174*, 179-188 (1995).
12. M. I. Granero-Porati, A. Porati and L. Zani, *J. Theor. Biol. 86*, 401-403 (1980).
13. H. Almagor, *J. Theor. Biol. 117*, 127-136 (1985).
14. L. F. Luo, L. Tsai and Y. M. Zhou, *J. Theor. Biol. 130*, 351-361 (1988).
15. T. D. Schneider, *Methods Enzymol. 274*, 445-455 (1996).
16. W. C. Barker, L. T. Hunt, D. G. George, L. S. Yeh, H. R. Chen, M. C. Blomquist, E. I. Seibel-Ross, A. Elzanowski, J. K. Bair and D. A. Ferrick, *Protein Seq. Data Anal. 1,* 43-98 (1987).

Date Received: March 11, 2002

Communicated by the Editor Valery Ivanov