行政院國家科學委員會專題研究計畫　成果報告

統計自由能偶合對應法的發展及應用(3/3)

計畫類別：個別型計畫
計畫編號：NSC91-2113-M-009-027-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學生物科技研究所

計畫主持人：黃鎮剛
共同主持人：林彩雲，劉銀樟，呂平江

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 8 月 17 日

91-2113-M-009-027

Calculation of Statistical Structural Entropy and its Applications to Protein Structures

Jenn-Kang Hwang

Department of Biological Science & Technology and Institute of Bioinformatics,

National Chiao Tung University, Hsinchu 300, Taiwan.

Abstract

We have developed a general method to compute the structure entropy of protein sequences. Structure entropy gives a quantitative measure of structure conservation. This relationship is similar to that between sequence entropy and sequence conservation. Experimental studies in protein folding have suggested that residues relevant to protein folding, or the so-called "hot spot" residues, are usually structurally conserved, though not necessarily conserved in sequences. Hence, the ability to compute structure entropy can help identify important residues related to protein folding. In this work, we have applied our approach to several model proteins frequently used in protein folding experiments. Our results suggest a close relationship between the structure entropies of residues and their rates of amide proton exchange, and we are able to identity regions of residues that are important in protein folding.

INTRODUCTION

The conformation and structure of a protein are determined by its sequence.[1] However, both designed and naturally occurring sequences are shown to adopt different conformations in different protein environments.[2-5] This observation suggests that the structures of some local protein subsequences are context dependant. Certain parts (the non-context-dependant parts) of the protein may be critical for protein folding and structure determination; these parts are usually termed as “folding nucleus” or “hot spot”.[6, 7] Whether these folding nucleus are conserved in sequence is controversy.[8-10] There are many cases where protein sequences with little sequence identities have very similar folds. One of the conspicuous examples is the triophosphate isomerase (TIM) fold,[11] which in part supports the non-conservation of folding nucleus. It is generally believed that the folding rates of proteins are closely related to their topology, though the choice among topological features varies.[12-14] Current analysis of protein topology requires experimentally determined or dynamically sampled structures of proteins.[14-16] A sequence based method that can take into account the non-conservativeness and is able to evaluate whether local protein fragments are structure determinants should prove valuable and useful.

For years the use of nuclear magnetic resonance (NMR) in hydrogen-deuterium exchange experiments have provided invaluable information on the structure of

proteins and their folding intermediates.[17, 18] Hydrogen-deuterium exchange experiments for amide protons can detect residues that are protected in different phases of protein folding processes. Therefore residues unfolded slowly can be identified. These residues are seen as important residues in protein folding processes. Their identification and consequent analysis are crucial to the study of protein folding mechanisms. NMR study is indispensable and has become an important tool in structural genomics.[19] However, NMR equipments and experiments are expensive. Also, pure protein samples and tedious operations are required for protein structure determination and analysis using NMR. A purely sequence based method for the analysis of protein conformation fluctuation and possibly folding mechanism is attractive. Though sequence based methods could never replace experiments, they may provide complementary information and hints for further experimental analysis. Starting from the sequence of a protein, with consequent querying of the subsequences' occurrences in structure databases, a new assessment for conformation fluctuation of proteins has resulted, i.e., the structure entropy of protein sequences. With structure entropy, it is possible to identify structurally conserved regions in protein sequences.

In this work, we have shown that the structure entropies of a protein are closely related to its proton exchange events. We investigated the conformation conservations

of local protein sequences; these conservations are measured with structure entropies.

~~A computational approach to calculate the structural conservation of local protein fragments is desirable. Such approach can provide insights to the stability and adaptability of protein local conformations. We have used an information theory based approach for the intrinsic structural entropy calculation. The method has been applied to peptides and patterns to observe the generalized characteristics of the conformation conservation. PROSITE patterns with non-conserved conformations have been identified. Some specific systems are also investigated, including mutations in Arc repressor and free energy of proton exchange in toxin proteins.~~Using an information theoretical approach,[20] we are able to calculate the structure propensity of a peptide. The structure propensity was calculated using categorized protein backbone conformations and was represented as a single value. This kind of practice is common in constructing sequence conservation (diversity) index.[21] It has also been applied to side-chain[22] and main-chain[23] conformation analysis of proteins. The resulting score could be termed as structure entropy, since it samples the local structure variations of protein sequences. This score is able to show the conformation conservation of PROSITE patterns. The score may also be used to construct the structure entropy profiles of proteins. The structure entropy profile of a protein corresponds well with its proton exchange events. These correspondences act as the first step to understand

protein folding with merely the presence of protein sequences.~~{Hamada, 1995 #25;Minor, 1996 #70;Cregut, 1999 #75}~~

METHODS

Representation of protein structure with secondary structure elements

In order to apply information theory to local conformations of proteins, one must categorize these conformations to a finite number. Protein structures in the Protein Data Bank (PDB)[24] are represented with the 3D coordinate of the atoms in each protein. These structure information are exact, but in some cases impose too much details; all-atom models of protein have too many degrees of freedom. It will be preferable to represent the conformation of each residue in a protein using a single symbol, much like the sequence of the protein. Therefore the routine sequence analysis techniques could be applied to the simplified structure representation as well. We have picked the secondary structure of a protein for such purpose. Other representations are also applicable.[25, 26] The secondary structure assignments of the proteins in PDB are available in DSSP database.[27] The local conformations of residues in a protein are categorized into 8 classes according to their hydrogen bonding patterns. The 8 secondary structure types are β-bridges (designated as B), extended β-sheet (E), 3_{10}-helix (G), α-helix (H), π-helix (I), bend (S), turn (T), and any others (U). With DSSP, the conformations of proteins could be represented in

one-dimensional sequences composed of their secondary structure assignments. These conformations in one-dimensional representations are used in the calculation of structure entropy.

Structure entropy

For a protein sequence x with arbitrary length l, we conduct query on the occurrences of x in structure database. The occurrences of x and their one-dimensional representations are recorded, as shown in Fig. 1. Fig. 1 illustrated two sequences, ELKEL and ELVGK. Both have multiple occurrences in different proteins. For each position in the protein sequence, there is a frequency distribution of the 8 secondary structure types. For example, the frequency to find helix (H) in the 3rd position of ELKEL is 1.0, and 0 for any other secondary structure types. The entropy at this position in sequence x could be calculated using:

$$S_{pos}^{x} = -\sum_{i} p_i \ln p_i , \quad (1)$$

where pos is the position in sequence x, i is the secondary structure types, and p_i is the frequency of i at position pos in sequence x. A conserved position will have a low entropy value, whereas a position with diverse conformations will have a high entropy value. The structure entropy of sequence x is estimated with:

$$S^{x} = \frac{1}{l} \sum_{pos=1}^{l} S_{pos}^{x} , \quad (2)$$

where l is the length of sequence x. We define the relative entropy of x as:

$$\Delta S^x = S^x - S^0 , \quad (3)$$

where S^0 is the reference entropy. The reference entropy was calculated using Equation (1), neglecting sequence and position. The frequency distribution used to calculate S^0 is based on the distribution of each secondary structure types in the entire structure database.

Construction of structure entropy profiles

For a given protein sequence, a sliding window of arbitrary length has been used to split the sequence into shorter fragments. These fragments are queried against the structure database, and the structure entropies are calculated accordingly. The structure entropies are assigned to the central residues in these fragments. It is not relevant whether the given protein presented in the structure database or not. These entropies form the sequence-based structure entropy profile of a protein sequence.

Identification of slow exchange and low entropy residues

The exchange rates of residues may be presented in various forms. For example, it could be represented as free energy,[28] as protection factor,[29] as rate (1/t, where t is time),[30] or as time.[31] Because of these variations, a consistent comparison between exchange rate and structure entropy is difficult. We have divided the residues in a protein into slow/non-slow exchange residues with certain criteria, which mainly follow Li and Woodward (1999).[32] The criteria used and residues identified as slow

proton exchange ones for various proteins are summarized in Table 1.

In order to assign a structure entropy cutoff value suitable for most proteins, correlations between slow exchange residues and residues identified by different cutoff values are calculated iteratively. For most proteins, the threshold value ΔS=-1.08 will yield maximal correlations between slow exchange and low entropy residues. However, in the case of Chymotrypsin inhibitor 2, this cutoff value needs to be relaxed to –0.88 for inclusion of more residues. The structure entropy cutoff values and residues identified by these values for various proteins are also listed in Table 1.

RESULTS

Structure entropy of PROSITE patterns

PROSITE is a database of functionally conserved sequence patterns.[33] Most patterns in PROSITE are structurally conserved.[34] However, we have found some PROSITE patterns with exceptionally high structure entropy, indicating non-conserved conformations. We have listed some typical examples of PROSITE pattern with high and low structure entropy in Table 2. Among the four low entropy patterns are malate dehydrogenase active site signature (PS00068), cutinase active sites signatures (PS00155), plant thionins signature (PS00271), and ferritin iron-binding regions signatures (PS00540). The high entropy patterns are EGF-like domains (PS00022), eukaryotic RNA recognition motif signature (PS00030), mitochondrial energy

transfer proteins signature (PS00215), and the Trp-Asp (WD-40) repeats signature (PS00678). The superimposed trace structures of these motifs are shown in Fig. 23. We can see that the backbones of the low entropy patterns are structurally well overlapped (Fig. 23A), while those of the high entropy patterns contains quite varied conformations (Fig. 23B). The computed structure entropies give a quantitative measure of conformational conservation of sequence patterns.

Structure entropy and proton exchange events

We have examined the structure entropy profiles for four proteins, and compare the results with their proton exchange events. These proteins are chymotrypsin inhibitor 2 (CI2), cytochrome *c* (cyt *c*), protein G B1 domain (GB1), and cardiotoxin analogous type III (CTX III). These proteins have all been extensively studied on their folding mechanisms.

Chymotrypsin inhibitor 2

Chymotrypsin inhibitor 2 (CI2) is a small protein and has been studied extensively in terms of protein folding.[35] The proton exchange experiments of CI2[28, 36] have revealed several slow exchange residues (those with free energy of exchange ΔG_{ex}^{app} larger than 7.0 kcal/mol^{-1}). These residues located on hydrophobic region formed by the C-terminal of α-helix and central strand of the β-sheet (Fig. 3A, left). The low structure entropy residues are also located in these hydrophobic regions (Fig. 3A,

right), though not all slow exchange residues are found by structure entropy. It is notable that a number of residues on the reactive site loop region (the long loop in the right of the figure) are labeled as having low structure entropy values. These residues are I37, M40, E41, R43, and I44. However, the exchange rates for these residues are not available, and a comparison is not feasible for these residues.

Cytochrome *c*

Cytochrome *c* is an important component of the energy-harvesting complex on mitochondria, and its folding kinetics is also of great interest.[37] The proton exchange experiments showed that protected protons are mainly located on the terminal (N-terminal and C-terminal) helices (Fig. 3B, left).[29, 38] Slow exchange residues are those with protection factor P larger than 10^7, where $P=k_c/k_{ex}$, k_c is the intrinsic exchange rate, and k_{ex} is the measured hydrogen-deuterium exchange rate.[29] There is a contact region between the two terminal helices (lower part of Fig. 3B). The proton exchange experiment and structure entropy analysis both identified the two terminal helices. Another helix (60's helix; from it's sequence numbering) is also identified by both proton exchange experiment and structure entropy analysis. A number of residues are labeled as low structure entropy ones but not slow exchange residues. Half of these residues (D2, L35, and P44) do not have available proton exchange rates. Finally, proton exchange experiment and structure entropy analysis also correspond to

each other on residue L32.

Protein G B1 domain

Protein G is a multidomain cell wall protein with several immunoglobulin G (IgG) binding domains. The B1 domain of Protein G is one of these IgG binding domains. Protein G B1 domain (GB1) is a small protein with well-defined structures and has been studied extensively. The proton exchange experiments on GB1 have revealed several residues with slow exchange rates, these residues are categorized with rates smaller than 0.005 (h^{-1}).[30] These residues formed a compact hydrophobic core (Fig. 3C, left). Structure entropy analysis identified a number of slow exchange residues (F30, T44, and F52) in this hydrophobic core, but not all (Fig. 3C, right). Two residues outside the hydrophobic core, T2 and D22, were identified by structure entropy analysis. D22 acts as a helix cap and may be essential for helix formation,[39] whereas the exact role of T2 in the folding pathway is unclear.

Cardiotoxin analogous type III

Cardiotoxin analogous type III (CTX III) is a small, all β-sheet protein. CTX III contains two β-sheets, one is double stranded (formed by β1 and β2) and the other is triple stranded (formed by β3, β4, and β5). There are four disulfide bonds in CTX III. Proton exchange study of CTX III has revealed that the slow exchange protons are located on the triple stranded β-sheet (Fig. 3D, left).[31] The slow exchange residues are

those with time constant of refolding shorter than 15 ms. Structure entropy analysis identified most of these residues and more (most of these do not have proton exchange rates available, see Table 1) in the same region (Fig. 3D, right). Two residues were identified by structure entropy analysis exclusively, K2 and N55. It is interesting to note, that by loosen the time constant criterion to 25 ms, K2 and N55 will be included as slow exchange residues.

DISCUSSION

We have linked structure entropy analysis to proton exchange experiments. Previous study by Hisler and Freire (1996) has suggested that calculations based on protein structure may provide hints to protein folding pathway.[40] Our approach does not require the structure of the target protein. Though the correspondences we found between proton exchange experiment and structure entropy analysis are mostly qualitative, we believe our approach is more general and require much less resources than structure based approach.

The further improvement of structure analysis and its correspondences on proton exchange experiments relies on several issues. First, the available experimental data were collected under various conditions, and their interpretation requires careful calibrations. Second, the use of secondary structure to represent local conformation may not be optimal; alternative representations and perhaps combined ones may yield

better correspondences. Third, structure entropy measures local conformation conservation, thus it may or may not be able to capture tertiary interactions among subsequences of proteins. It is likely that structure entropy could never fully describe protein folding pathways in detail; but it may provide helpful structural information with merely the availability of protein sequences.

The major concern about information theory based scoring is that it cannot account for the distances among the symbols.[41] Both sequence and structure entropy suffer from this caveat. However, there are fundamental differences between sequence entropy and structure entropy. We have constructed both sequence and structure entropy profiles for CTX III (Fig. 4). In previous sections we have shown that structure entropy have close relationships with proton exchange events. Fig. 4A illustrated the sequence entropy profile of CTX III. It could be seen that the sequence of cardiotoxin is very conserved, but makes no distinction among the secondary structure elements. All these secondary structure elements are equally conserved in sequence entropy profile. On the other hand, structure entropy suggests that strands β3 and β5 are more stable than others (Fig. 4B), which agrees well with the proton exchange results (Fig. 3D and Table 1).[31] This result confirms that sequence conservations are not necessary corresponding to slow exchange residues or folding nucleus of proteins, which complements to the observation that folding nucleus are

not necessary more conserved in sequence identities.[8]

Structure entropy analysis is among many attempts to uncover the sequence-structure relationships. It is efficient, and may be proved valuable in genomics scale protein structure analysis. Current results have revealed some of the connections between protein sequences and structure conservation. Its applications to protein structure and folding analysis are promising and may be of great help for researchers in the named fields. A web page has been build to facility the usage of structure entropy. This web page was named “StEQ: Structure Entropy Query” and can be accessed at http://atp.life.nctu.edu.tw/~entropy/.

Key Words

Structure entropy; Proton exchange; Protein secondary structure; Protein folding; Protein sequence/structure relationships

REFERENCES

1. Anfinsen, C. B. *Science* **1973**, *181*, 223.

2. Minor, D. L. J.; Kim, P. S. *Nature* **1996**, *380*, 730.

3. Kabsch, W.; Sander, C. *Proc. Natl. Acad. Sci. U. S. A.* **1984**, *81*, 1075.

4. Mezei, M. *Protein Eng.* **1998**, *11*, 411.

5. Sudarsanam, S. *Proteins* **1998**, *30*, 228.

6. Fersht, A. R. *Curr. Opin. Struct. Biol.* **1997**, *7*, 3.

7. Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Rokhsar, D. S. *Curr. Opin. Struct. Biol.* **1998**, *8*, 68.

8. Plaxco, K. W.; Larson, S.; Ruczinski, I.; Riddle, D. S.; Thayer, E. C.; Buchwitz, B.; Davidson, A. R.; Baker, D. *J. Mol. Biol.* **2000**, *298*, 303.

9. Mirny, L.; Shakhnovich, E. *J. Mol. Biol.* **2001**, *308*, 123.

10. Larson, S. M.; Ruczinski, I.; Davidson, A. R.; Baker, D.; Plaxco, K. W. *J. Mol. Biol.* **2002**, *316*, 225.

11. Nagono, N.; Orengo, C. A.; Thornton, J. M. *J. Mol. Biol.* **2002**, *321*, 741.

12. Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985.

13. Fersht, A. R. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 1525.

14. Dokholyan, N. V.; Li, L.; Ding, F.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 8637.

15. Miller, E. J.; Fischer, K. F.; Marqusee, S. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 10359.

16. Bonneau, R.; Ruczinski, I.; Tsai, J.; Baker, D. *Protein Sci.* **2002**, *11*, 1937.

17. Bai, Y.; Sosnick, T. R.; Mayne, L.; Englander, S. W. *Science* **1995**, *269*, 192.

18. Englander, S. W. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 213.

19. Montelione, G. T.; Zheng, D.; Huang, Y. J.; Gunslus, K. C.; Szyperski, T. *Nat.*

Struct. Biol. **2000**, *7*, 982.

20. Shannon, C. E. *The Bell System Tech. J.* **1948**, *27*, 379.

21. Baczkowski, A. J.; Joanes, D. N.; Shamia, G. M. *J. theor. Biol.* **1997**, *188*, 207.

22. Creamer, T. P. *Proteins* **2000**, *40*, 443.

23. Solis, A. D.; Rackovsky, S. *Proteins* **2000**, *38*, 149.

24. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.

25. Hoffman, D. L.; Laiter, S.; Singh, R. K.; Vaisman, I. I.; Tropsha, A. *Comput. Appl. Biosci.* **1995**, *11*, 675.

26. Barlow, T. W.; Richards, W. G. *J. Mol. Graph.* **1996**, *14*, 232.

27. Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.

28. Itzhaki, L. S.; Neira, J. L.; Fersht, A. R. *J. Mol. Biol.* **1997**, *270*, 89.

29. Jeng, M.-F.; Englander, S. W.; Elöve, G. A.; Wand, A. J.; Roder, H. *Biochemistry* **1990**, *29*, 10433.

30. Orban, J.; Alexander, P.; Bryan, P.; Khare, D. *Biochemistry* **1995**, *34*, 15291.

31. Sivaraman, T.; Kumar, T. K. S.; Chang, D. K.; Lin, W. Y.; Yu, C. *J. Biol. Chem.* **1998**, *273*, 10181.

32. Li, R.; Woodward, C. *Protein Sci.* **1999**, *8*, 1571.

33. Falquet, L.; Pagni, M.; Bucher, P.; Hulo, N.; Sigrist, C. J. A.; Hofmann, K.;

Bairoch, A. *Nucleic Acids Res.* **2002**, *30*, 235.

34. Kasuya, A.; Thornton, J. M. *J. Mol. Biol.* **1999**, *286*, 1673.

35. Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R. *J. Mol. Biol.* **1995**, *254*, 260.

36. Neira, J. L.; Itzhaki, L. S.; Otzen, D. E.; Davis, B.; Fersht, A. R. *J. Mol. Biol.* **1997**, *270*, 99.

37. Elöve, G. A.; Bhuyan, A. K.; Roder, H. *Biochemistry* **1994**, *33*, 6925.

38. Roder, H.; Elöve, G. A.; Englander, S. W. *Nature* **1988**, *335*, 700.

39. Blanco, F. J.; Serrano, L. *Eur. J. Biochem.* **1995**, *230*, 634.

40. Hilser, V. J.; Freire, E. *J. Mol. Biol.* **1996**, *262*, 756.

41. Valdar, W. S. J. *Proteins* **2002**, *48*, 227.

TABLES

Table 1

List of residues with slow exchange rates and those with low structure entropy values in several proteins.

	Slow Exchange Criteria[1]	Residues with Slow Exchange Rates[2]	Structure Entropy Cutoff	Residues with Low Structure Entropy Values
Chymotrypsin inhibitor 2 (CI2)	ΔG_{ex}^{app} > 7.0 kcal/mol^{-1}	K11, V19, I20, L21, I30, L32, V47, R48, L49, F50, V51	aS < -0.88	K2[3], T3[3], E4, V19, I29[3], I37[3], M40[3], E41[3], R43[3], I44[3], R48, L49
Cytochrome *c* (cyt *c*)	$P > 10^7$	F10, L32, L68, L94, I95, A96, Y97, L98, K99	aS < -1.08	D2[3], Q12, H18, T19, L32, L35[3], P44[3], E66, Y67, K79, M80, L94, Y97, L98, K99
Protein G B1 domain (GB1)	rate < 0.005 h^{-1}	L5, I6, E27, F30, T44, T51, F52, T53, V54	aS < -1.08	T2, D22, F30, T44, F52
Cardiotoxin analogue type III (CTX III)	time constant < 15 ms	K23, I39, V49, Y51, V52, C53, D57, R58	aS < -1.08	K2, C21, Y22, K23, M24[3], F25[3], I39, D40[3], P43[3], Y51, V52, C53, C54[3], N55

[1] ΔG_{ex}^{app} is free energy of proton exchange; P is the protection factor, where $P=k_c/k_{ex}$,

k_c is the intrinsic exchange rate, and k_{ex} is the measured hydrogen-deuterium exchange rate.

[2] Experimental results are obtained from Itzhaki *et al.*,[28] Jeng *et al.*,[29] Orban *et al.*,[30] and Sivaraman *et al.*,[31] and reorganized for CI2, cyt *c*, GB1, and CTX III, respectively.

[3] The exchange rate of these residues are not determined or not probed in experiments.

Table 2

Summaries of some selected PROSITE patterns with low and high structure entropies.

Accession Number[1]	Entry Name[1]	ΔS	RMSD[2] (Å)
Sequence motifs of low Entropy			
PS00068	MDH	-1.68	0.35
PS00155	CUTINASE_1	-1.65	0.10
PS00271	THIONIN	-1.64	0.23
PS00540	FERRITIN_1	-1.66	0.19
Sequence motifs of high Entropy			
PS00022	EGF_1	-0.79	2.19
PS00030	RNP_1	-0.67	2.18
PS00215	MITOCH_CARRIER	-0.58	3.64
PS00678	WD_REPEATS	-0.84	3.59

[1] The accession number and the entry name are taken directly from PROSITE database.

[2] Averaged pair-wise Root-Mean Square Deviation.

FIGURE CAPTIONS

Figure 1 Protein sequences have different conformation preferences. Both sequences (ELKEL and ELVGK) occur multiple times in different proteins. The PDB ids for these proteins are provided. ELKEL is in helix (H) conformation in most cases, whereas ELVGK adapts different conformations in different protein environments.

Figure 2 Superimposed structures of PROSITE patterns with low and high structure entropies. The structures are shown in trace representation. **A:** the low entropy patterns, where the backbone of the occurrences of the patterns fit well. **B:** the high entropy patterns, for each high entropy pattern there are two or more distinctive conformations.

Figure 3 Correspondences between slow proton exchange residues and structure entropy in several proteins. Slow exchange regions in proteins are marked in red, so are the residues with low structure entropy. **A:** Chymotrypsin Inhibitor 2 (CI2). **B:** Cytochrome *c* (cyt *c*). **C:** Protein G B1 domain (GB1). **D:** Cardiotoxin analogous type III (CTX III). The PDB ids used to plot the structures are 2CI2 (CI2), 1HRC (cyt *c*), 1PGA (GB1), and 2CRT (CTX III), respectively.

Figure 4 Comparison of **A:** sequence entropy (ΔS_{seq}) and **B:** structure entropy (ΔS_{str}) profiles of CTX III. The secondary structures ($\beta 1 \sim \beta 5$) of CTX III are labeled on **B**. Note that the scales on the two figures are not identical as they refer to different

quantities. It is clear that sequence entropy cannot distinguish among these strands (sequences of all the strands are highly conserved), while structure entropies are markedly lower in β3 and β5.

Figure 1

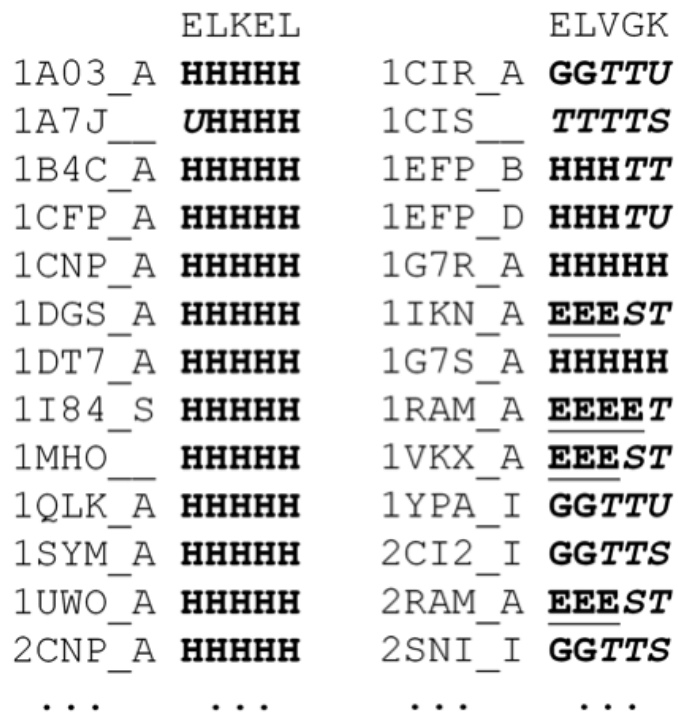

ELKEL
ELVGK
1A03_A HHHHH
1CIR_A GGTTU
1A7J__ UHHHH
1CIS__ TTTTS
1B4C_A HHHHH
1EFP_B HHHTT
1CFP_A HHHHH
1EFP_D HHHTU
1CNP_A HHHHH
1G7R_A HHHHH
1DGS_A HHHHH
1IKN_A EEEST
1DT7_A HHHHH
1G7S_A HHHHH
1I84_S HHHHH
1RAM_A EEEET
1MHO__ HHHHH
1VKX_A EEEST
1QLK_A HHHHH
1YPA_I GGTTU
1SYM_A HHHHH
2CI2_I GGTTS
1UWO_A HHHHH
2RAM_A EEEST
2CNP_A HHHHH
2SNI_I GGTTS
... ...

Figure 2

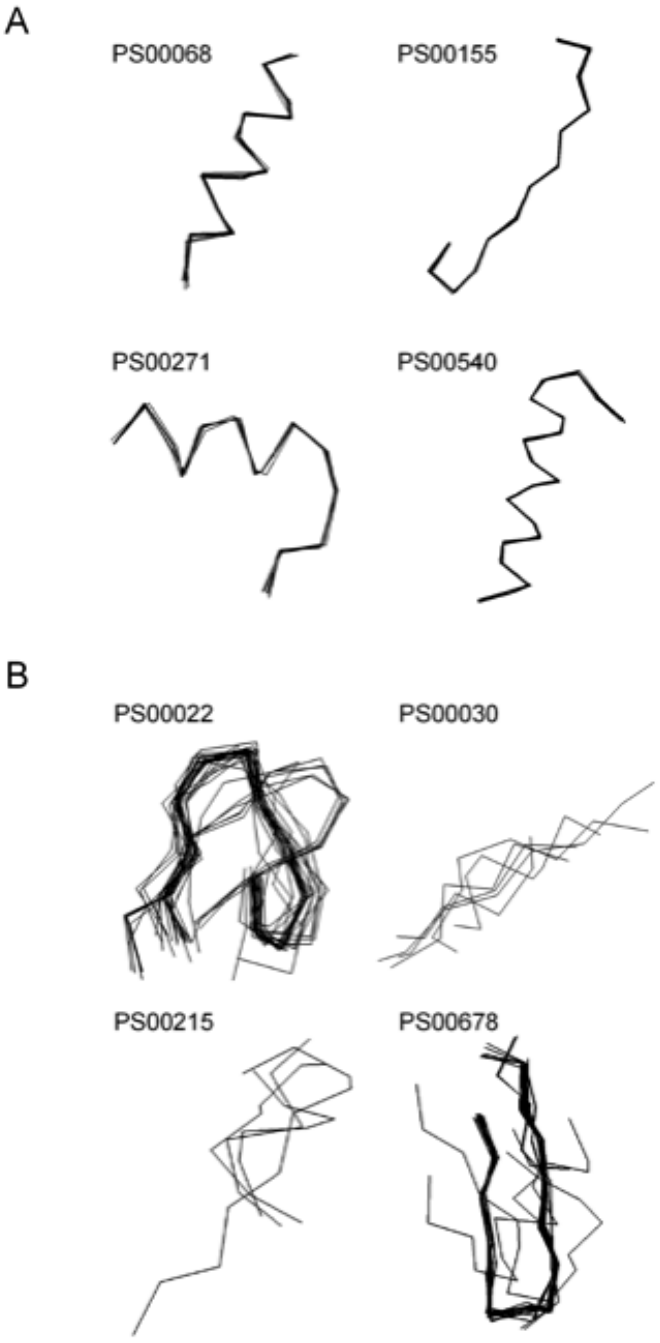
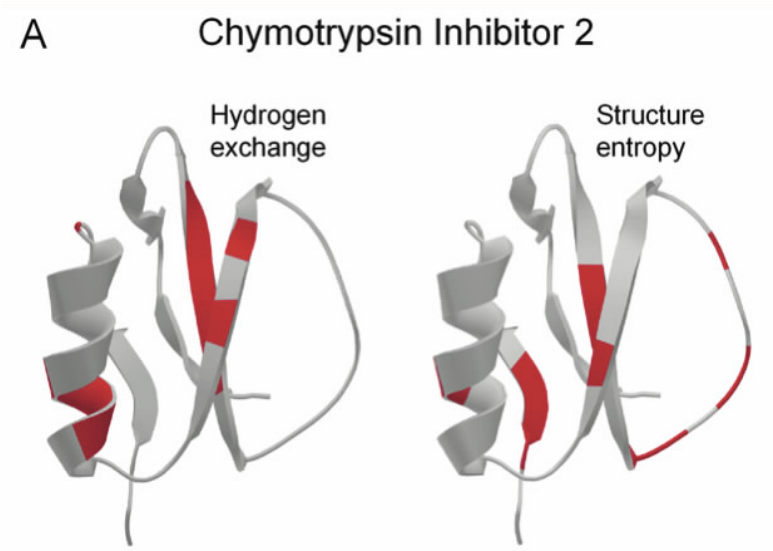
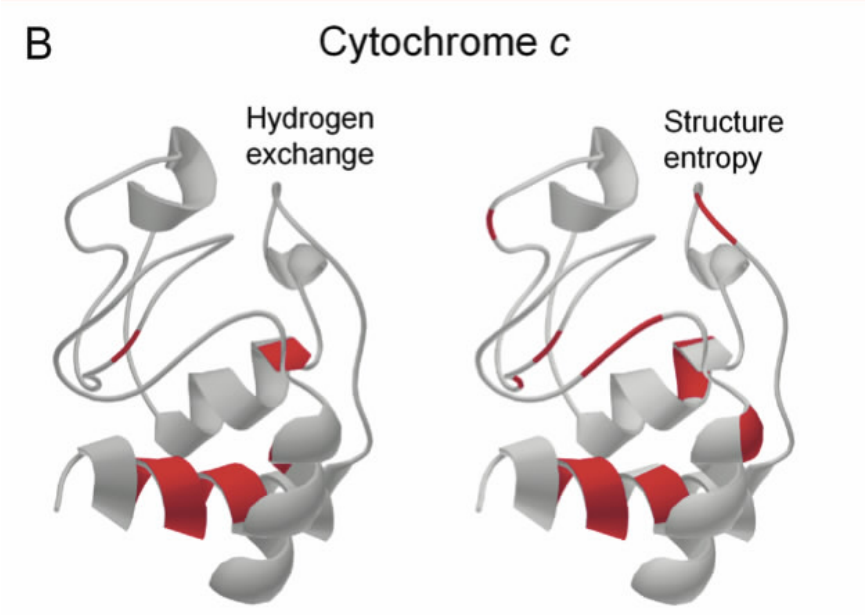
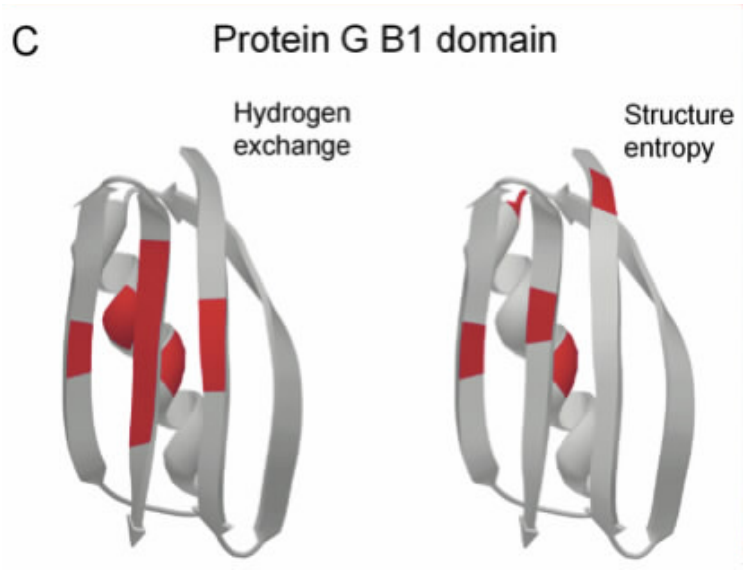

A
PS00068
PS00155
PS00271
PS00540
B
PS00022
PS00030
PS00215
PS00678

Figure 3


A
Chymotrypsin Inhibitor 2
Hydrogen exchange
Structure entropy


B
Cytochrome c
Hydrogen exchange
Structure entropy


C
Protein G B1 domain
Hydrogen exchange
Structure entropy

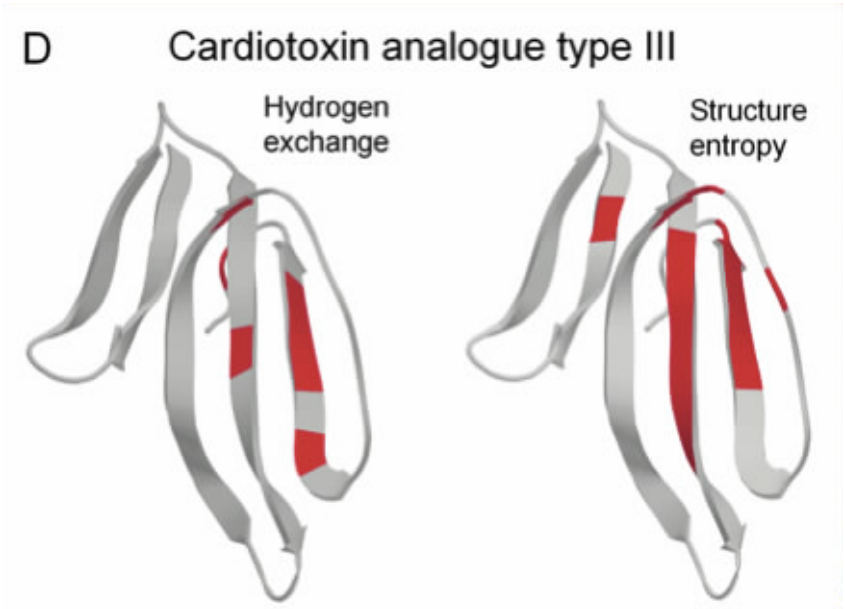
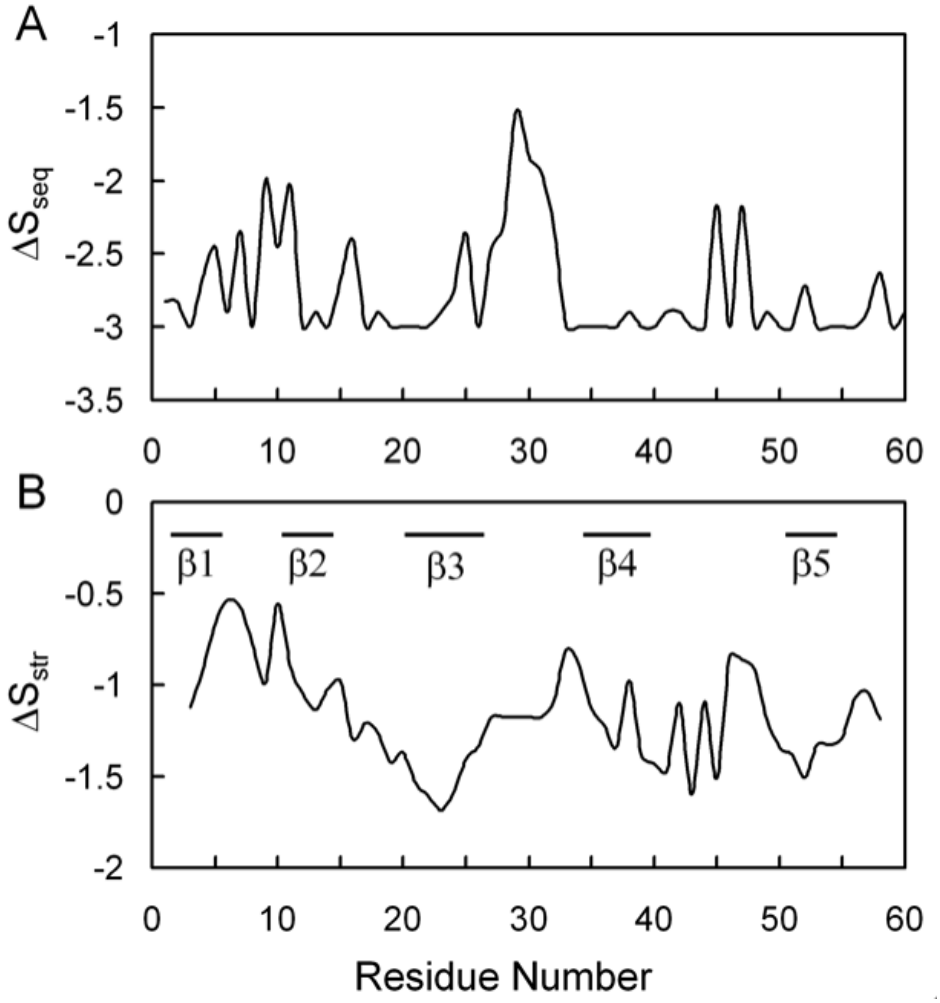

D
Cardiotoxin analogue type III
Hydrogen exchange
Structure entropy

Figure 4


A
-1
-1.5
-2
-2.5
-3
-3.5
ΔS_seq
0
10
20
30
40
50
60
B
0
-0.5
-1
-1.5
-2
ΔS_str
β1
β2
β3
β4
β5
0
10
20
30
40
50
60
Residue Number