Quantitative Sequence/Conformation Relationship

The Quantitative Relationship between Amino Acid Sequence and its Local Conformation

Chen-hsiung Chan[a], P.C. Lyu[a], and Jenn-Kang Hwang[b*]

[a] Department of Life Sciences, National Tsing Hua University, Hsinchu 300, Taiwan

[b] Department of Biological Science & Technology, National Chiao Tung University, Hsinchu 300, Taiwan

[*] Corresponding author

Department of Biotechnology and Institute of Bioinformatics
National Chiao Tung University
R205, Chu-Ming Building
75, Po-Ai Street, Hsinchu 300
Taiwan

Fax: +886-3-5729288

Telephone: +886-3-5712121-56936

E-mail: jkhwang@cc.nctu.edu.tw

Summary

We have developed an approach based on information theory to compute the structural information content, or structural entropy, of amino acid sequences of arbitrary lengths. The structural entropy of amino acid sequences is computed from their local structural distributions, using different structural representations. Structure entropy values could be used to indicate structure conservation quantitatively. As a practical tool in de novo protein design, our approach offers a quantitative measure of the effects of a single amino acid mutation on the change of local conformations. There are also qualitative correspondences between structure entropy values and proton exchange rates. Structure entropy gives more information than the primary sequence of protein, and provides insights to the study of protein local sequence/structure relationships.

Keywords: structural entropy; information theory; entropy profile; mutation analysis.

Introduction

It is well known that the conformation of a given peptide fragment depends on the context of their environments (Cregut et al., 1999; Minor & Kim, 1996). Peptides with identical sequences or common patterns may have different conformations in different protein environments. There are also experimental studies (Blanco & Serrano, 1995; Gegg et al., 1997; Hamada et al., 1995; Reymond et al., 1997; Tsuji et al., 1999; Yanagawa et al., 1993) on intrinsic folding/structure propensity of peptide fragments, i.e., the ability of a peptide fragment to maintain a conserved conformation regardless of the change of its surrounding protein environment. It has been shown that the intrinsic stability of substructures within a protein may play important roles in folding pathways (Kuhlman et al., 2002). Peptide fragments of a given amino acid sequence may have different degrees of structure variations; some peptide fragments maintain almost identical conformations in different proteins, whereas some others adopt a myriad of various conformations.

There are many works in the study of structure propensities of proteins. Wright and coworkers (Reymond et al., 1997) have determined the helical propensities of myoglobin in various solvents by NMR; Gegg *et al.* (Gegg et al., 1997) have studied the folding propensities of fragments of dihydrofolate reductase by fluorescence spectroscopy and CD, to name a few. These results are valuable in giving useful information about the ability of a fragment to maintain its native conformation. However, in these experiments, the peptides used to determine structural propensities are based on different criteria - some are based on the secondary structures (Reymond et al., 1997; Tsuji et al., 1999), while others are using structurally closed modules or functional domains (Yanagawa et al., 1993). It is not easy to give a general interpretation to these results, since these fragments vary greatly in their lengths and characteristics; and interpretations are highly dependent on peptides' properties. In the case of secondary structural elements, while it is easy to measure secondary structural content of the elements compared to the native structure; it is difficult to describe transitions among different secondary structures quantitatively, as reported in some studies (Cregut et al., 1999; Hamada et al., 1995; Minor & Kim, 1996).

It has also long been accepted that the conformation and structure of a protein are determined by its sequence (Anfinsen, 1973). The numerous examples mentioned above regarding conformation-switching sequences seem contradictory. It is reasonable to infer that sequences may have different propensities toward an invariant conformation. Some sequences have high propensity for a single conformation, and others may not. Those with a low propensity may adopt different conformations as environments change. This phenomenon has been observed for pentapeptides (Argos, 1987; Kabsch & Sander, 1984), in which some pentapeptides have conserved backbone conformation, and others vary in different proteins. However, this observation is qualitative and it is difficult to apply these information for automatic protein sequence/structure analysis.

Based on a statistical approach, we have developed a method to calculate the structure preference of a peptide. We search a non-redundant PDB (Berman et al., 2000) subset (termed NRPDB) for the occurrences of specified peptides or patterns. The probabilities of finding each secondary structure types or other structural classes at individual residual positions in the occurrences have been calculated. Three structural classification schemes based on secondary structure (SS), main-chain virtual κ angle (κ), and main-chain torsional angles (ϕ–ψ) have been used. Using the probability matrix constructed above, a value has been calculated using information theory formulations. This value is the structural information content of the specified peptide or pattern, and can be termed as structure entropy, ΔS . High structure entropy value implies variations in conformations, while low entropy value suggests an invariant conformation. Details and formulations of the entropy calculation could be found in Materials and Methods.

We expect structure entropy values to be an indication of the structural conservation for peptides. A high entropy value implies that the conformations of the given peptides or patterns are not conserved and change in different protein environments. A low structure entropy value, on the other hand, indicates an invariant conformation across different proteins. To see whether the results meet out expectations, we have calculated structure entropy values for two data sets and compared their entropy profile, $g(\Delta S)$. One set contains pentapeptides (termed NRPDB-5) and the distribution of the entropy values should be broad, since pentapeptides have different preferences for structure conservation (Argos, 1987; Kabsch & Sander, 1984). The other set contains patterns from PROSITE database (Falquet et al., 2002), and the distribution is expected to localize in the low entropy region, since most PROSITE patterns are structurally conserved (Kasuya & Thornton, 1999).

Besides structural conservation, we have also applied out structure entropy to other interesting subjects. We have selected some PROSITE patterns with low and high structure entropy values, and compared their conformation variations. We have attempted to find the effect of a single mutation on protein local conformation with structure entropy. We also tried to correlate proton exchange rates with structure entropy values. Overall, structure entropy could provide additional information beyond the primary structure of proteins. This may prove valuable in automatic analysis of protein sequence/structure relationships in the future.

Results

The entropy profiles

The structural entropy profiles $g(\Delta S_{ss})$, $g(\Delta S_{\kappa})$, and $g(\Delta S_{\phi-\psi})$ of NRPDB-5 (solid line) and PROSITE (dotted line) are shown in Fig. 1. The characteristics of $g(\Delta S_{ss})$ and $g(\Delta S_{\kappa})$ entropy profiles in general agree with each other - the entropy profiles of PROSITE have much narrower distribution and show one prominent peak near the lower limit of entropy, while those of NRPDB-5 smoothly distributes over a wide range and do not have any particular preferences at specific entropy values. This behavior agrees well with those shown by previous studies (Argos, 1987; Kabsch & Sander, 1984). The $g\left(\Delta S_{\phi-\psi}\right)$ profile shows marked difference from the other two - it shows a broader, more structured profiles in both NRPDB-5 and PROSITE. In general, all three representations yield quite similar distributions for both NRPDB-5 and PROSITE patterns. To quantitatively compare these three representations, we can evaluate their information contents, or information gains (Solis & Rackovsky, 2000), given by $\sum_x \Delta S(x) p(x)$, where $p(x)$ is the probability of the occurrences of sequence x. However, information gain depends on the number and distributions of structural classes within a particular structure representation. To eliminate this dependency and be able to compare among the three representations directly, information gain has been divided by reference entropy $S^0(x)$ of each representation and yields the fraction of information gain. The fractions of the information gain for ΔS_{ss} and ΔS_{κ} are 0.59 and 0.54, respectively, whereas that of $\Delta S_{\phi-\psi}$ is 0.45. The results indicate that the secondary structure and the virtual angle representations of the protein backbones carry comparable amount of information, whereas the torsional angle representation exhibits extra information loss.

Evaluation of structure conservation

Structural entropy offers a useful quantitative measure of structural conservation of peptide sequence. Fig. 1 shows that most PROSITE patterns are structural conserved, as shown previously (Kasuya & Thornton, 1999). However, our results also show that there are a nontrivial number of patterns that have higher structural entropy values, which indicating structural non-conservation of these motifs. Table 1 lists examples of PROSITE patterns with low and high entropy values. 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. 2. We can see that the backbones of the low entropy patterns are structurally well overlapped (Fig. 2a), while those of the high entropy patterns contains quite varied conformations (Fig. 2b).

Application to conformational switch upon residue mutations

Though protein structure are usually stable enough to tolerate the effects of most small-scale amino acid mutations (Brown & Sauer, 1999; Kuroda & Kim, 2000), it is possible to induce completely different folds from existing folds without drastic mutagenesis. One well-known example is the Arc repressor, where a beta sheet structure fragment is switched to helices with two consequent mutations (Cordes et al., 1999). The wild type Arc contains a β-sheet peptide fragment FNLR (residue 10-13) in its DNA binding site (Baumann et al., 1994). An N11L mutation results in a new fragment, FLLR, was reported to adopt various conformations in different conditions (Cordes et al., 2000), and retains DNA binding capability. An additional mutation L12N leads to a double mutant fragment FLNR and results in a stable α-helix conformation (Cordes et al., 1999). The experimental results on the effect of the mutations on local conformations are summarized in Fig. 3.

Using a statistical approach we can also draw the same conclusion as the experimental results. The statistics on the secondary structure distribution of the local conformations of these peptide fragments are performed, and the results are summarized in Fig. 4. The structure queries of the peptide fragments on NRPDB have shown that wild type fragment FNLR occurred mostly in the form of β-sheet, N11L fragment FLLR is either in β-sheet or α-helix (and part of the fragment might in unstructured conformation), and double mutant N11L/L12N fragment FLNR has a dominant conformation of α-helix. Though we cannot infer that the N11L mutant keeps the DNA-binding capability, it is possible to correctly predict the conformation change caused by the point mutations.

The statistical approach is powerful, however, it is usually difficult to evaluate the effects of the mutations computationally with such an approach. We have computed the structure entropy values of the peptide fragments in question and the results are summarized in Fig. 5. The computed structure entropy values for the fragments from wild type, N11L mutant, and N11L/L12N double mutant are –0.6223, -0.2761, and –1.2596, respectively. The N11L mutant fragment FLLR has the highest structure entropy among the three, which implies a highly flexible, less rigid conformation. The difference on entropy between wild type and N11L mutant is small, compared to that between N11L mutant and N11L/L12N double mutant. Both the wild type and N11L mutant have DNA binding capabilities. Upon DNA binding, both undergo some conformational changes toward a similar DNA-bound conformation (Breg et al., 1990; Cordes et al., 2000; Raumann et al., 1994). The double mutant, as indicated by its low structure entropy, is confined to an invariant conformation, which, in this case, is helix. From Fig. 5 we can see that the transition from β-sheet to α-helix can be bridged by an intermediate.

Correlations with proton exchange study of proteins

The exchange rates of amide protons are faster in the β-sheet of Arc repressor, compare to those in the helices (Burgering et al., 1995). This has suggested that the β-sheet of Arc repressor may not be as rigid as the rest of the protein. The structure entropy of this region also suggested that the conformation is flexible. It should therefore be interesting to compare the result of proton exchange experiments and structure entropy. Here we present some results of the comparisons. The structure entropy distributions of several proteins along the sequences have been shown in Fig. 6, these proteins include Arc repressor, barnase, chymotrypsin inhibitor 2 (CI2), and cardiotoxin type III (CTX III). We will discuss the implications of the structure entropy distribution and their correspondence to proton exchange experiments in the following sections.

Arc Repressor

Arc repressor is a small, dimeric DNA binding protein formed by two identical monomers. The monomer of Arc repressor contains one β-strand (residue 8-14, termed β) and two α-helices (residue 16-29 and 35-46, termed α1 and α2, respectively); the β-strands from the two monomers formed an antiparallel β-sheet. The three secondary structure elements are labeled in Fig. 6a. As shown in Fig. 6a, the structure entropy values at α2 are lower than β and α1. The proton exchange study of Arc repressors has shown that protons in α1 and α2 are slow exchanging, whereas those in β are not (Burgering et al., 1995). The structure entropy values of β and α2 agree well with the proton exchange results. However, α1 have higher structure entropy values as opposed to our expectation. One possible explanation is that the two helices of Arc repressor are intertwined according to the NMR structure (Bonvin et al., 1994). Structure entropy is the preference of a peptide fragment toward an invariant conformation, the actual local conformation and its environments maybe affected by interactions within the substructures. Proton exchange measures the outcome of the interplays, and the protections of protons maybe contrary to our expectations from structure entropy computation.

Barnase

Barnase is an α+β protein with three α-helices and five β-strands. The structure entropy distribution of barnase along the sequence is given in Fig. 6b. The structure and folding of barnase have been studied extensively (Fersht et al., 1992; Serrano et al., 1992). The proton exchange study of barnase has shown that all the helices and strands formed early in the protein refolding pathway (Matouschek et al., 1992). The structure entropy distribution in Fig. 6b can only correspond to the proton exchange experiment qualitatively. Several of the secondary structures indeed have low structure entropy values, e.g. α1, α2, β2, β3, and β4. However, the structure entropy values of α3, β1 and β5 are higher then the others. The folding of barnase is highly co-operative (Horovitz & Fersht, 1992). This co-operation of substructures may provide stabilizing environment for elements with higher structure entropy values.

Chymotrypsin Inhibitor 2

Chymotrypsin inhibitor 2 (CI2) is a small protein with two-state folding kinetics

(Jackson & Fersht, 1991). Fig. 6c illustrates the structure entropy distribution along the protein chain; secondary structure elements of CI2 are also labeled. The proton exchange studies of CI2 have been performed under different conditions, and the most slowly exchange protons are located on α, β3, and β4 of CI2 (Itzhaki et al., 1997; Neira et al., 1997), this corresponds well with the structure entropy computation. It is interesting to note that the reactive loop (residue 35-44) between β3 and β4 has low structure entropy values, even lower than those on the secondary structure elements. This loop has an extensive conformation; however, the low structure entropy values suggest that the sequence of this loop might have a moderate preference for this particular loop conformation.

Cardiotoxin analogue III

Cardiotoxin analogue 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 (Sivaraman et al., 1998). The structure entropy distribution of CTX III along the sequence is shown in Fig. 6d. It could be seen clearly that the structure entropy values on β3, β4, and β5 are generally lower than those on β1 and β2. The correspondences between proton exchange rates and structure entropy values are obvious in CTX III.

With the above examples we have illustrated that structure entropy values have correlated with proton exchange rates. This correlation is qualitative; however, the simplicity of structure entropy calculation makes it an interesting tool in assessing the local conformation preference of proteins. It may also provide complement information for protein local sequence/structure relationships.

Discussion

Structure entropy calculation is based on statistics and information theory. We have illustrated that structure entropy could act as a measure on the intrinsic structure preference of protein sequences. Our structure entropy calculation does not require a protein with known structure. Actually, even for protein with structural data available, the particular local structure is only one of the occurrences for the specified sequence.

Structure entropy provides a quantitative relationship between protein sequence and its local conformation. In a simplistic way, structure entropy is the degree of structural conservation of a sequence in different proteins. It is also the ability of a sequence to adapt different protein environments.

In this work we have used several structural classification schemes to assign residues to different classes and calculate structural entropy based on these schemes. Secondary structure assignments are actually summarized from hydrogen-bonding patterns of several residues (Kabsch & Sander, 1983) and contain lots of implicit structural information. The main-chain angles also composed of some additional information. κ-angles are virtual angles spanning across five residues, the angle class assigned to a residue does not represent the single residue but the local structure formed by five consecutive residues. ϕ- and ψ-angles are different; as main-chain dihedral angles, they only connect two residues. We have noted that ϕ- and ψ-angles together provide less information than secondary structure or virtual angle representations. A decomposition of the two angles has suggested that ϕ angles may be responsible for the loss of information (data not shown). It is interesting to note that though these classifications are different in nature, the structural entropy based on these classifications (except $\phi-\psi$ and ϕ alone) are consistent with each other. It is clear that these representations provide more structural information than the primary sequence of proteins.

Structural entropy can easily detect structural variations. In terms of PROSITE patterns, high structural entropy implies patterns with lower specificities. Cautions need to be made when attempt to use these patterns; combination with other patterns or use profiles may be a good alternative. To address the specificity issue, revisions of the patterns with high structural entropy may be necessary. As for peptides, low entropy implies consistent conformations in different protein environments; for example, AQLEK has helical conformation in all its occurrences. On the other hand, peptide ALGVE changes its conformation in different proteins and has high structural entropy. A dictionary of structural entropy for peptides could be constructed. Protein sequences could be scanned with such dictionary and regions on the sequence could be labeled with the entropy values. Interpretation of the entropy values may provide further insights to the stability and other fundamental aspects of proteins.

The application of structural entropy on Arc repressor provides new possibility to site directed mutagenesis. Structural entropy values provide hints to the effect of a mutation, whether the local structure will be stable or flexible and how the new peptide fits with the old environment. Our current approach has averaged out the environment factors during our entropy calculation. It is possible to further refine our method by considering the environmental factors separately (Reddy et al., 1998). For example, in an environment that lacks constraint on the conformation, peptides with low structural entropy will be more stable than peptides with high structural entropy. However, if an environment imposes large stress on a specific conformation like helix, then a peptide with high structural entropy may stabilize in such environment, but not a peptide with low entropy and high preference in sheet conformation. We have identified a number of peptide pairs, which only differ in one residue but have large entropy difference. For example, the entropy values for AAGAA and AAVAA are -0.5928 and -1.6414, respectively. AAGAA may adapt to both sheet and helix conformation, whereas AAVAA is dedicated to helix. Further investigations are required to reveal the implications and applications of these data.

There is a qualitative correlation between structure entropy values and proton exchange rates (Fig. 6). Proton exchange experiments are invaluable in the study of protein stability, folding, and dynamics (Raschke & Marqusee, 1998). Structure entropy is an indication of the preference of a local peptide fragment toward an invariant conformation. This preference does not account for environments provided by individual proteins and may not be able to reflect the physical property of the native conformation precisely. It is thus not surprising to find poor correlations for barnase and CI2, both folds in a highly co-operative manner (Horovitz & Fersht, 1992; Jackson & Fersht, 1991). However, the correlations between proton exchange rates and structure entropy values are obvious for Arc repressor and CTX III. This may due to the loosely held structures of Arc repressor and CTX III; CTX III is held by four disulfide bonds and the stacking among substructures are fewer than other proteins. Still, structure entropy may provide extra information and complement to proton exchange results.

In this work we have tried to show that structural entropy is generally applicable to vast number of structural biological problems. Structure entropy values provide additional information to the primary sequence of proteins. With structure entropy it is possible to study protein local sequence/structure relationship in a more fundamental manner. We hope structure entropy could become a routine facility for de novo protein design, mutation analysis, and even proteomics in the foreseeable future.

Materials and Methods

Non-redundant structural database

A non-redundant structure database, NRPDB, has been used through out this work. NRPDB is a subset of PDB containing non-redundant proteins with homologues filtered out by sequence comparison using BLAST (Altschul et al., 1990). NRPDB is compiled by National Center for Biotechnology and Information (NCBI) and can be found at http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html. The size of NRPDB varies with different cut-off values; the smallest one contains 2288 PDB entries, whereas the largest one contains 10212 entries. The statistics are obtained from the Oct. 3, 2001 release of NRPDB. There are no significant differences in using databases with different sizes.

Entropy calculation

For each residue x_i in a specific amino acid sequence x of given lengths l, there is an associated distribution of structural classes $\mathbf{p} = (p_1, p_2, \ldots, p_\sigma)$, where p_i is the probability of finding the i^{th} structure class in the set composed of these amino acid sequences, and σ is the total number of structure classes. The structural information content or entropy of x_i is calculated using the following equation (Shannon, 1948):

$$S(x_i) = -\sum_{j}^{\sigma} p_j \ln p_j$$

If there is only one structure available for the amino acid sequence x, the value $S(x_i)$ is zero. On the other hand, if the structures of the sequence are evenly distributed over all classes, we will obtain a maximum value of entropy, $\ln\sigma$. The average of the entropy of each residual position gives an estimation of $S(x)$. We define the relative entropy of a sequence by

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

where $S^0(x)$ is the reference entropy of x. The reference entropy is calculated using the probabilities of the structural classes in the database queried. For example, the reference entropy for secondary structure elements is calculated using the probabilities of finding each of the secondary structural classes in NRPDB. The value of $\Delta S(x)$ gives the relative measure of structural information content, or structure conservation, of sequence x. Using $\Delta S(x)$, we compute the entropy profile by

$$g(\Delta S) = \sum_x \delta(\Delta S - \Delta S(x))$$

where $\delta(y) = 1$, for $y = 0$ and $\delta(y) = 0$, for $y \neq 0$. The function $g(\Delta S)$ gives the distribution of structure entropies of the amino acid sequences in a data set.

Structure representations

Since we do not have a unique way to describe the local conformations of the sequences, we tried three structure representations in this work: the secondary structural elements, the backbone torsional angles, and the virtual angles, referred to as κ angles, defined by three alternative C_α atoms. In the secondary structure representation, we use the DSSP method (Kabsch & Sander, 1983) to assign secondary structural elements. DSSP defines eight secondary structural elements, seven of which are based on hydrogen bonding patterns between amino acid residues, and one is for the undefined structures excluded from the previous types. In the backbone torsional angle representation, we use the usual ϕ and ψ torsional angles used in the Ramachandran plot. As a third structure representation, we employ the virtual angles, referred to as κ angles, to represent the backbone conformation. The κ angle of residue i is a virtual angle defined by three alternate C_α atoms of residues $i-2$, i and $i+2$. The κ-angle has been used in structure search and comparison (D. Chang and J.-K. Hwang, unpublished results). The structure entropies of these representations are referred to as ΔS_{ss}, ΔS_κ and $\Delta S_{\phi-\psi}$, respectively.

Structure entropy along a protein chain

For a given protein chain, we may calculate the structure entropy of the fragments along the sequence. We have applied a fix width sliding window to the sequence, and compute the structure entropy of each fragment scanned by this sliding window. The entropy values computed are then assigned to the central residues of the fragments. For a protein chain, we will have a profile of the structure entropy along the chain. Since the structure entropy could be used to indicate the intrinsic structure variation of the local conformations formed by peptide fragments, this profile may provide indications to regions in the sequence where the structure formed early in the folding process or regions that have high tolerance to mutagenesis.

Acknowledgement

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* 181, 223-230.

Argos, P. (1987). Analysis of Sequence-similar Pentapeptides in Unrelated Protein Tertiary Structures Strategies for Protein Folding and a Guide for Site-directed Mutagenesis. *J. Mol. Biol.* 197, 331-348.

Baumann, B. E., Rould, M. A., Pabo, C. O. & Sauer, R. T. (1994). DNA recognition by beta-sheets in the Arc repressor-operator crystal structure. *Nature* 367, 754-757.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic*

Acids Res. 28, 235-242.

Blanco, F. J. & Serrano, L. (1995). Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *Eur. J. Biochem.* 230, 634-649.

Bonvin, A. M., Vis, H., Breg, J. N., Burgering, M. J., Boelens, R. & Kaptein, R. (1994). Nuclear magnetic resonance solution structure of the Arc repressor using relaxation matrix calculations. *J. Mol. Biol.* 236, 328-341.

Breg, J. N., van Opheusden, J. H., Burgering, M. J., Boelens, R. & Kaptein, R. (1990). Structure of Arc repressor in solution: evidence for a family of beta-sheet DNA-binding proteins. *Nature* 346, 586-589.

Brown, B. M. & Sauer, R. T. (1999). Tolerance of Arc repressor to multiple-alanine substitutions. *Proc. Natl. Acad. Sci. U.S.A.* 96, 1983-1988.

Burgering, M. J., Hald, M., Boelens, R., Breg, J. N. & Kaptein, R. (1995). Hydrogen exchange studies of the Arc repressor: evidence for a monomeric folding intermediate. *Biochemistry* 35, 217-226.

Cordes, M. H. J., Burton, R. E., Walsh, N. P., McKnight, C. J. & Sauer, R. T. (2000). An evolutionary bridge to a new protein fold. *Nat. Struct. Biol.* 7, 1129-1132.

Cordes, M. H. J., Walsh, N. P., McKnight, C. J. & Sauer, R. T. (1999). Evolution of a Protein Fold in Vitro. *Science* 284, 325-327.

Cregut, D., Civera, C., Macias, M. J., Wallon, G. & Serrano, L. (1999). A tale of two secondary structure elements: when a beta-hairpin becomes an alpha-helix. *J. Mol. Biol.* 292, 389-401.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J. A., Hofmann, K. & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucleic Acids Res.* 30, 235-238.

Fersht, A. R., Matouschek, A., Sancho, J., Serrano, L. & Vuilleumier, S. (1992). Pathways of Protein Folding. *Faraday Discuss.* 93, 183-193.

Gegg, C. V., Bowers, K. E. & Matthews, C. R. (1997). Probing minimal independent folding units in dihydrofolate reductase by molecular dissection. *Protein Sci.* 6, 1885-1892.

Hamada, D., Kuroda, Y., Tanaka, T. & Goto, Y. (1995). High Helical Propensity of the Peptide Fragments Derived from β-Lactoglobulin, a Predominantly β-sheet Protein. *J. Mol. Biol.* 254, 737-746.

Horovitz, A. & Fersht, A. R. (1992). Co-operative Interactions during Protein Folding.

J. Mol. Biol. 224, 733-740.

Itzhaki, L. S., Neira, J. L. & Fersht, A. R. (1997). Hydrogen Exchange in Chymotrypsin Inhibitor 1 Probed by Denaturants and Temperature. *J. Mol. Biol.* 270, 89-90.

Jackson, S. E. & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* 30, 10428-10435.

Kabsch, W. & Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22, 2577-2673.

Kabsch, W. & Sander, C. (1984). On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. U. S. A.* 81, 1075-1078.

Kasuya, A. & Thornton, J. M. (1999). Three-dimensional Structure Analysis of PROSITE Patterns. *J. Mol. Biol.* 286, 1673-1691.

Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y. J. & Baker, D. (2002). Accurate Computer-based Design of a New Backbone Conformation in the Second Turn of Protein L. *J. Mol. Biol.* 315, 471-477.

Kuroda, Y. & Kim, P. S. (2000). Folding of Bovine Pancreatic Trypsin Inhibitor (BPTI) Variants in which Almost Half the Residues are Alanine. *J. Mol. Biol.* 298, 493-501.

Matouschek, A., Serrano, L., Meiering, E. M., Bycroft, M. & Fersht, A. R. (1992). The Folding of an Enzyme V. H/2H Exchange-Nuclear Magnetic Resonance Studies on the Folding Pathway of Barnase: Complementarity to and Agreement with Protein Engineering Studies. *J. Mol. Biol.* 224, 837-845.

Minor, D. L. J. & Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature* 380, 730-734.

Neira, J. L., Itzhaki, L. S., Otzen, D. E., Davis, B. & Fersht, A. R. (1997). Hydrogen Exchange in Chymotrypsin Inhibitor 2 Probed by Mutagensis. *J. Mol. Biol.* 270, 99-100.

Raschke, T. M. & Marqusee, S. (1998). Hydrogen exchange studies of protein structure. *Curr. Opin. Biotech.* 9, 80-86.

Raumann, B. E., Rould, M. A., Pabo, C. O. & Sauer, R. T. (1994). DNA recognition by beta-sheets in the Arc repressor-operator crystal structure. *Nature* 367, 754-757.

Reddy, B. V. B., Datta, S. & Tiwari, S. (1998). Use of propensities of amino acids to the local structural environments to understand effect of substitution mutations on protein stability. *Protein Eng.* 11, 1137-1145.

Reymond, M. T., Merutka, G., Dyson, H. J. & Wright, P. E. (1997). Folding propensities of peptide fragments of myoglobin. *Protein Sci.* 6, 706-716.

Serrano, L., Kellis Jr, J. T., Cann, P., Matouschek, A. & Fersht, A. R. (1992). The Folding of an Enzyme II. Substructure of Barnase and the Contribution of Different Interactions to Protein Stability. *J. Mol. Biol.* 224, 783-804.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Tech. J.* 27, 379-423, 623-656.

Sivaraman, T., Kumar, T. K. S., Chang, D. K., Lin, W. Y. & Yu, C. (1998). Events in the Kinetic Folding Pathway of a Small, All β-Sheet Protein. *J. Biol. Chem.* 273, 10181-10189.

Solis, A. D. & Rackovsky, S. (2000). Optimized Representation and Maximal Information in Protsins. *Proteins* 38, 149-164.

Tsuji, T., Toshida, K., Satoh, A., Kohno, T., Kobayashi, K. & Yanagawa, H. (1999). Foldability of Barnase Mutants Obtained by Permutation of Modules or

Secondary Structure Units. *J. Mol. Biol.* 268, 1581-1596.

Yanagawa, H., Yoshida, K., Torigoe, C., Park, J.-S., Sato, K., Shirai, T. & Go, M. (1993). Protein Anatomy: Functional Roles of Barnase Module. *J. Biol. Chem.* 268, 5861-5865.

Figure Captions

Figure 1

Distributions of structural entropy for NRPDB-5 (solid lines) and PROSITE patterns (dashed lines). The structural entropies are calculated using (a) secondary structure, (b) virtual κ-angles, and (c) torsional angles representations. Pentapeptides in NRPDB-5 have different structure entropy values, implying their different intrinsic structure variations. PROSITE patterns, on the other hand, located mostly on the lower entropy end; this has suggested that most PROSITE patterns are structurally conserved.

Figure 2

Superimposed structures of patterns with low and high structural entropies. The patterns are illustrated as (a) the low entropy patterns, and (b) the high entropy patterns.

Figure 3

Local conformation changes among wild type Arc repressor, N11L, and N11L/L12N mutants. The structures are shown in ribbon representation. Secondary structures affected by the mutations are colored in red. The side chains of the fragment FNLR and FLNR are drawn. The PDB id for the wild type is 1ARR and the PDB id for N11L/L12N double mutant is 1QTG. There is no structure for N11L mutant available. The diagram illustrated that the local conformation of the fragment FLLR in N11L is a mixture of β-sheet (wild type) and α-helix (N11L/N12N double mutant).

Figure 4

Statistics of the secondary structure distributions of the three peptide fragment FNLR, FLLR, and FLNR from wild type, N11L mutant, and N11L/L12N double mutant, respectively. The secondary structure assignment follows DSSP (Kabsch & Sander, 1983); E denotes extended β-sheet, whereas H denotes α-helix. It can be seen that wild type fragment FNLR has dominant conformation of β-sheet, and the double mutant fragment FLNR has dominant conformation of α-helix. The fragment from N11L is either in β-sheet or α-helix. The statistics are done on each residual position, underlined bold text labels the fractions of the secondary structures higher than average.

Figure 5

Schematics for entropy values and local conformations of wild type and mutant Arc repressors. Each horizontal bar is leveled to the computed entropy value of the peptide fragment. The arrows indicate the corresponding mutations, and texts below the bars describe the conformation of each state. The transition from sheet conformation of wild type to helix of N11L/L12N double mutant could be bridged by the N11L mutant, which composed of various conformations at different solvent conditions.

Figure 6

Structure entropy distributions along the sequences for several proteins are illustrated. The secondary structure elements of each protein are also identified with horizontal lines. For consistency, all the secondary structure elements are numbered sequentially; for example, the first α-helix is termed α1, the second termed α2, and so on. The proteins used for structure entropy computations are (a) Arc repressor, (b) barnase, (c) chymotrypsin inhibitor 2 (CI2), and (d) cardiotoxin analogue III (CTX III), respectively. The entropy value at each residue position is calculated using the local peptide fragment around the specified residue.

Figure 1

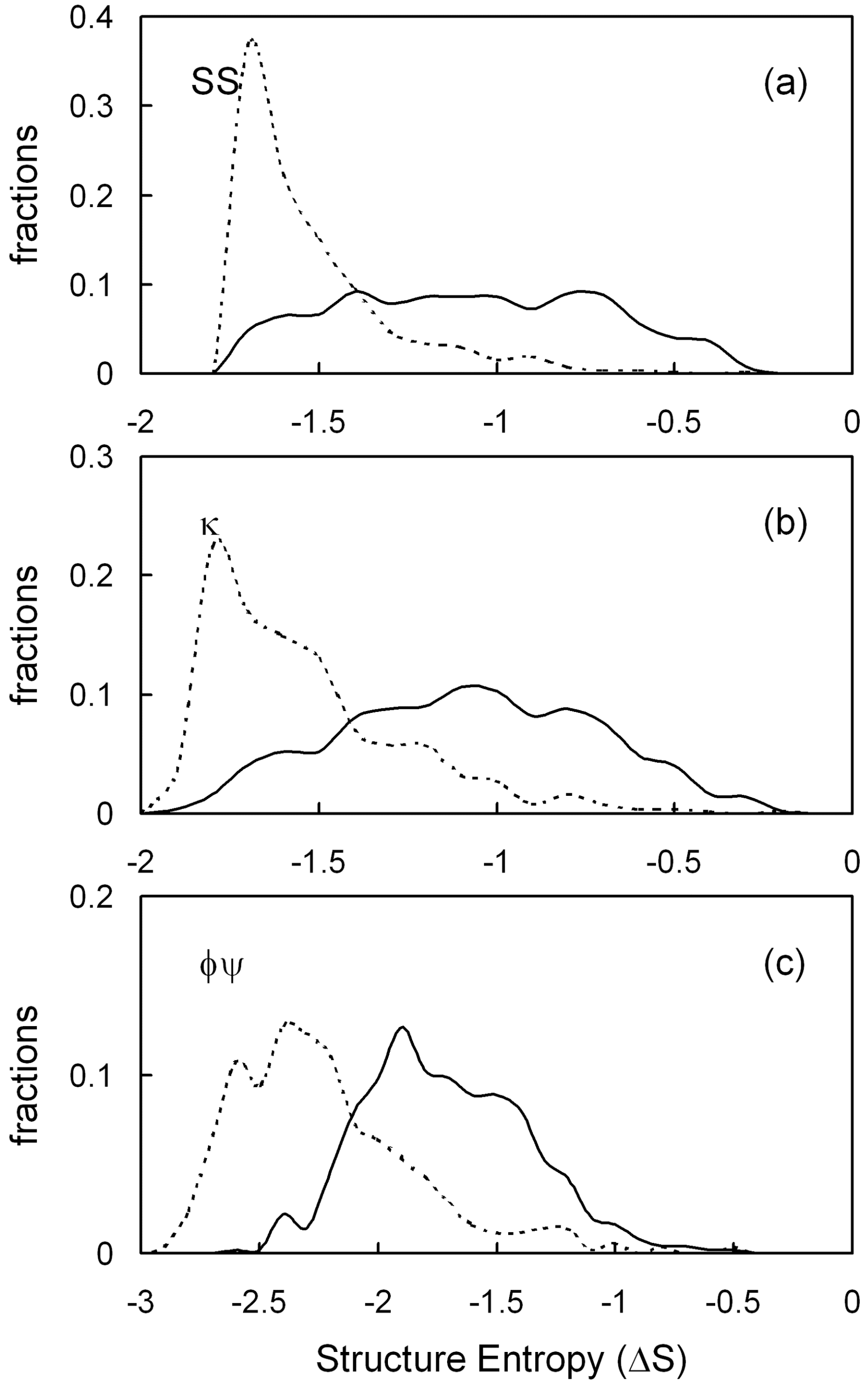

0.4
0.3
0.2
0.1
0
fractions
SS
(a)
-2
-1.5
-1
-0.5
0
0.3
0.2
0.1
0
fractions
κ
(b)
-2
-1.5
-1
-0.5
0
0.2
0.1
0
fractions
φψ
(c)
-3
-2.5
-2
-1.5
-1
-0.5
0
Structure Entropy (ΔS)

Figure 2

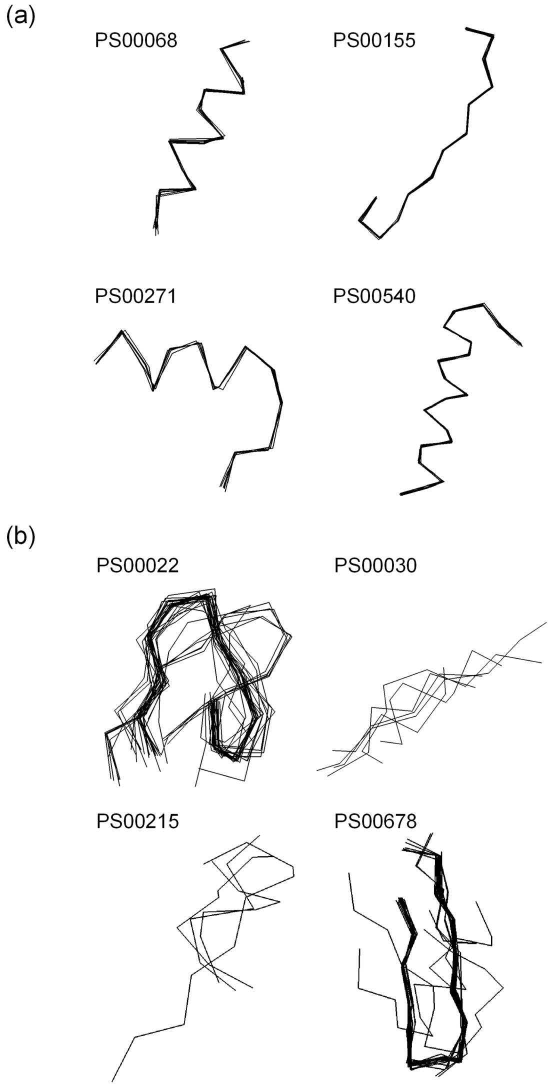

(a)
PS00068
PS00155
PS00271
PS00540
(b)
PS00022
PS00030
PS00215
PS00678

Figure 3

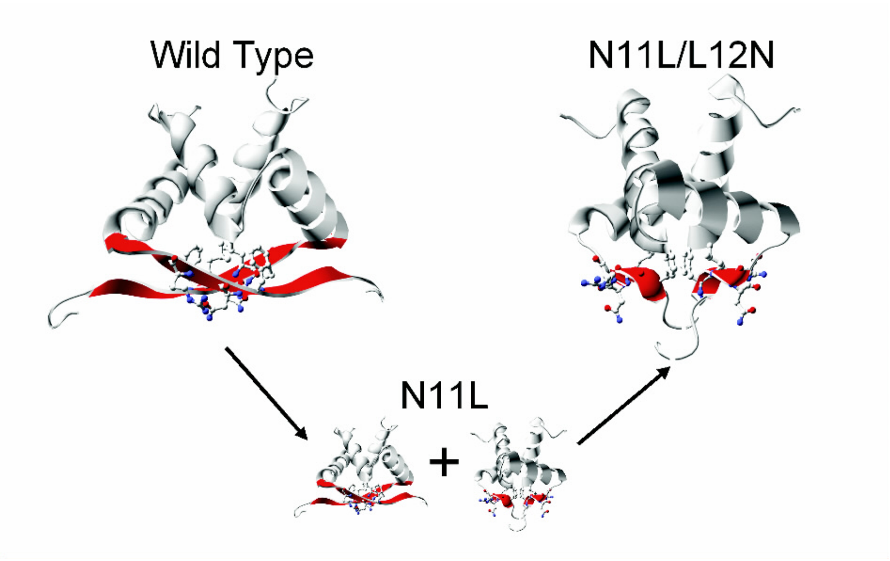


Figure 4

Secondary Structure	Wild Type				N11L				N11L/L12N			
	F	**N**	**L**	**R**	**F**	**L**	**L**	**R**	**F**	**L**	**N**	**R**
B	0.00	0.00	0.11	0.00	0.01	0.05	**0.41**	0.05	0.00	0.00	0.00	0.00
E	**0.70**	**0.70**	**0.59**	**0.59**	**0.23**	**0.20**	**0.16**	**0.16**	0.02	0.02	0.02	0.02
G	0.00	0.00	0.00	0.00	0.03	0.03	0.03	0.00	0.05	0.07	0.02	0.02
H	0.09	0.02	0.04	0.05	**0.55**	**0.23**	**0.22**	**0.20**	**0.91**	**0.89**	**0.89**	**0.93**
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.02	0.02	0.00	0.02	0.11	0.00	0.00	0.05	0.00	0.01	0.06	0.00
T	0.07	0.14	0.12	0.11	0.03	**0.35**	0.05	0.00	0.03	0.02	0.02	0.02
U	0.12	0.12	0.14	0.23	0.04	0.14	0.14	**0.53**	0.00	0.00	0.00	0.02

Figure 5

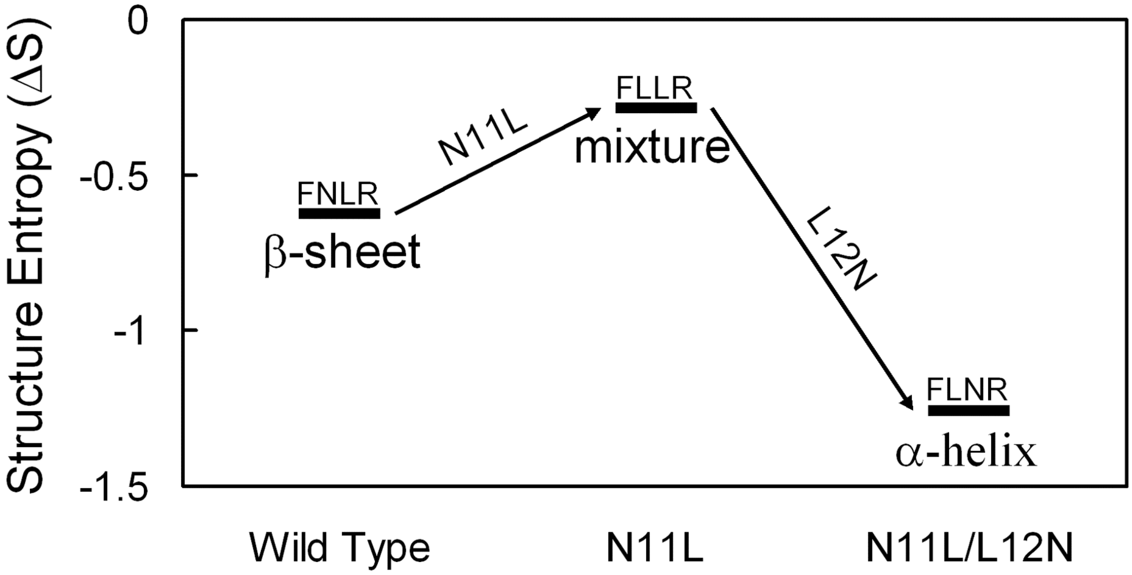
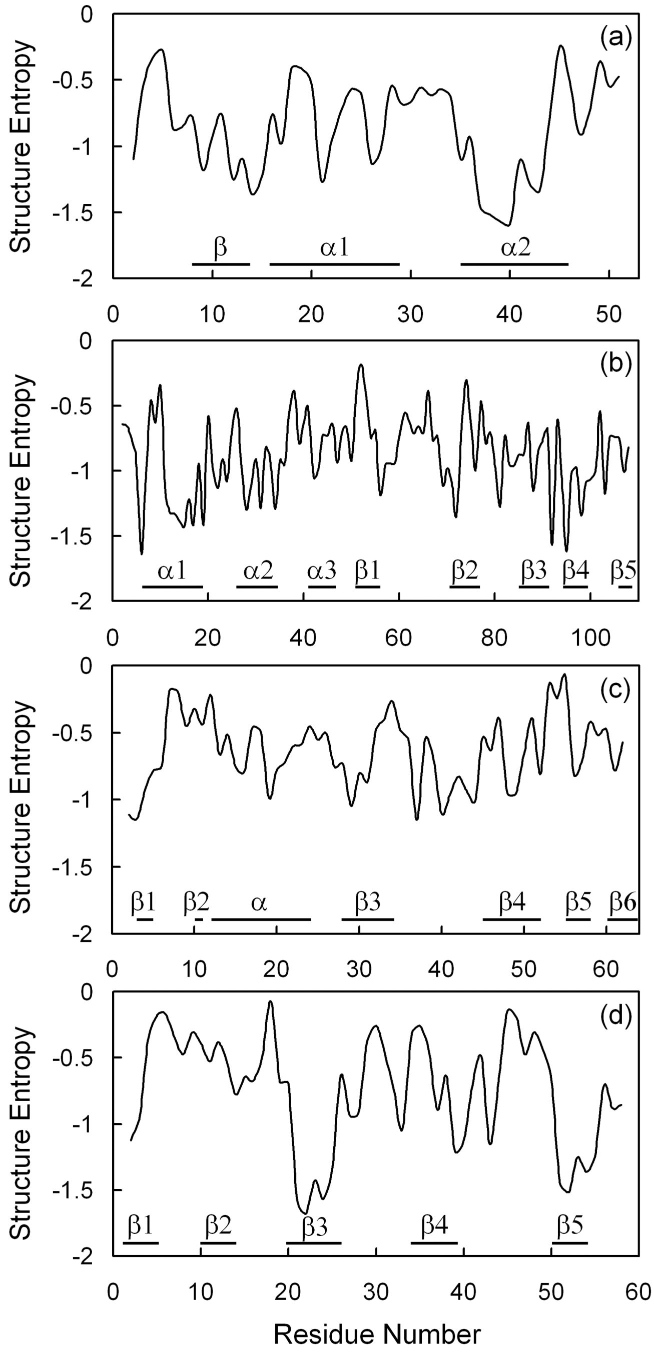

Structure Entropy (ΔS)
0
-0.5
-1
-1.5
FNLR
β-sheet
N11L
FLLR
mixture
L12N
FLNR
α-helix
Wild Type
N11L
N11L/L12N

Figure 6



(a)
(b)
(c)
(d)
Structure Entropy
0
-0.5
-1
-1.5
-2
β
α1
α2
0
10
20
30
40
50
α1
α2
α3
β1
β2
β3
β4
β5
0
20
40
60
80
100
β1
β2
α
β3
β4
β5
β6
0
10
20
30
40
50
60
β1
β2
β3
β4
β5
0
10
20
30
40
50
60
Residue Number

Tables

Table 1

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

Accession Number	ID	ΔS_{ss}	$\Delta S_{\phi-\psi}$	ΔS_{κ}	RMSD (Å)
Low Entropies					
PS00068	MDH	-1.6844	-2.3646	-1.7411	0.35
PS00155	CUTINASE_1	-1.6528	-2.5895	-1.6988	0.10
PS00271	THIONIN	-1.6416	-2.4570	-1.7068	0.23
PS00540	FERRITIN_1	-1.6645	-2.6062	-1.7730	0.19
High Entropies					
PS00022	EGF_1	-0.7905	-0.9334	-0.7849	2.19

PS00030	RNP_1	-0.6737	-0.9130	-0.5134	2.18
PS00215	MITOCH_CARRIER	-0.5753	-1.6548	-0.6839	3.64
PS00678	WD_REPEATS	-0.8367	-1.2915	-0.7468	3.59