

An Analysis of the Helix-to-Strand Transition Between Peptides With Identical Sequence

Xianghong Zhou,^{1,2} Frank Alber,³ Gerd Folkers,² Gaston H. Gonnet,¹ and Gareth Chelvanayagam^{4*}

¹Department of Computer Science, Eidgenössische Technische Hochschule, Zürich, Switzerland

²Department of Applied Bioscience, Eidgenössische Technische Hochschule, Zürich, Switzerland

³International School for Advanced Studies (SISSA) and Istituto Nazionale di Fisica della Materia (INFN), Trieste, Italy

⁴Department of Computer Science, University of Western Australia, Perth, Australia

ABSTRACT An analysis of peptide segments with identical sequence but that differ significantly in structure was performed over non-redundant databases of protein structures. We focus on those peptides, which fold into an α -helix in one protein but a β -strand in another. While the study shows that many such structurally ambivalent peptides contain amino acids with a strong helical preference collocated with amino acids with a strong strand preference, the results overwhelmingly indicate that the peptide's environment ultimately dictates its structure. Furthermore, the first naturally occurring structurally ambivalent nonapeptide from evolutionary unrelated proteins is described, highlighting the intrinsic plasticity of peptide sequences. We even find seven proteins that show structural ambivalence under different conditions. Finally, a computer algorithm has been implemented to identify regions in a given sequence where secondary structure prediction programs are likely to make serious mispredictions. *Proteins* 2000;41:248–256.

© 2000 Wiley-Liss, Inc.

Key words: structural ambivalence; protein secondary structure; structure prediction; sequence properties; sequence neighbours; long-range interaction; global environment

INTRODUCTION

α -helices and β -strands are the two most distinct elements of the protein secondary structure. An α -helix is formed mainly by local interactions while a β -strand is usually formed by long-range interactions (i.e., residue i to residue $(i+x)$, $|x|>4$). Experimental results, as well as statistical analysis, show that different amino acids and their combinations have different propensities for α -helical or β -strand formation.^{1–10} These propensity scales provide important tools for secondary structure prediction, and in particular, methods that use local sequence information.^{1,11,12} However, prediction methods based only on residue propensities are not foolproof¹³ and various experimental studies have pointed out that the secondary structure formation is strongly dependent on the environment.^{14–17} For example, naturally occurring peptides were found to adopt an α -helix conformation in organic solvent, but β -strand in nonmicellar SDS.^{15,18} Likewise, several

theoretical studies showed that sequentially identical peptides in the Protein Data Bank (PDB¹⁹) can adopt different secondary structures in different proteins.^{20–25} Even naturally occurring peptides as long as eight amino acids can be helical in one protein and a strand in another.²⁵ We term such peptides as structurally ambivalent. It is not known, however, just how far secondary structure formation is influenced by forces other than the sequence's own intrinsic propensity. Nor is it known if there is a minimum length for an autonomous folding unit based on the local interactions. Understanding the degree to which and the means by which the environment influences the structural ambivalence of peptides has implications for both protein design and the development of structure prediction methods. This information is also important for elucidating the mechanisms by which proteins fold.

Given the explosive growth of the PDB, the goal of this work is to assess the level of structural ambivalence among peptides with identical sequence in known structures and to examine the origin of their structural diversity.

COMPUTATIONAL METHODS

Database Survey

The Protein Data Bank (PDB) of June 1999 was used in this study. Secondary structure assignments were made automatically using the program package STRIDE.²⁶ One of our goals is the statistical analysis of peptide sequences with structural ambivalence. Thus, to avoid statistical bias caused by the large number of homolog proteins in the PDB, two protein sub-databases were used: one in which all protein-chain pairs have less than 25% sequence identity (DB1) and one in which all protein-chain pairs have less than 95% sequence identity (DB2). DB1 (1106 chains) and DB2 (3295 chains) were taken from the May 1999 version of the "PDB_Select database."²⁷

The selection of identical pairs of peptide sequences was performed as follows. First, we surveyed the complete PDB database selecting all possible sequence pairs with four identical residues (4-mer). Where possible, these were

Grant sponsor: Australian Research Council.

*Correspondence to: Dr. Gareth Chelvanayagam, Department of Computer Science, The University of Western Australia, Stirling Highway, Nedlands, Perth, W.A., 6009, Australia. E-mail: gareth@cs.uwa.edu.au

Received 13 March 2000; Accepted 8 June 2000

then extended to identify longer identical sequence pairs (5–9-mer pairs). Contiguous, overlapping 4-mers that form higher order n-mers were not retained in the 4-mer dataset and were assigned as the appropriate n-mer. Only the longest possible n-mer was considered. Thus, we only considered all 4-mer pairs that could not be extended to 5-mer pairs, or higher order n-mer pairs; 5-mer pairs that could not be extended to 6-mer pairs or higher order n-mers; and so on. These peptide pairs were then assigned to DB1 and DB2 if both constitute proteins were members of the corresponding sub-database.

We focused on peptide pairs with identical sequences that adopt an α -helix structure in one protein and a β -strand structure in another protein. We classify these secondary structure transitions, termed *helix-to-strand transitions*, into two categories: a *partial helix-to-strand transition* is defined when they contain at least a dipeptide unit adopting an α -helix in one protein and a β -strand, for the same residues, in another (e.g., CCHHH in one protein and TCCEE in another protein, where H represents helix, E strand, C coil, and T turn); and a *complete helix-to-strand transition* is defined when one peptide of a pair contains only an α -helix structure and the other only a β -strand structure (e.g., HHHHH in one protein and EEEEE in another protein).

Analysis of the Helix-to-Strand Transition

In order to avoid biases in the PDB, the following analysis was performed on the dataset of the structurally ambivalent peptides in the DB1, where no protein pair has sequence identity more than 25%.

Residue occurrence

To examine the residue occurrence in the structurally ambivalent peptides, the amino acid frequencies were calculated from all n-mers in DB1 that undergo *complete helix-to-strand transition*. Amino acids in common n-mer motifs, motifs that are found in more than one pair of sequences, are counted at each occurrence of the motif. These values were normalized against the amino acid frequency in the PDB.

Local sequence

To examine properties of the local sequence surrounding ambivalent peptides, the amino acid frequencies were counted over the four residues immediately flanking each of the n-mers in DB1 that undergo *complete helix-to-strand transition*. The amino acid frequencies that bound helices and strands were counted separately.

Local environment

Properties of the local environment of structurally ambivalent peptides in DB1 that undergo *complete helix-to-strand transition* were investigated by classifying the amino acids into 5 groups: aromatic residues (F,Y,W), non-polar residues (V,I,L,M,C,A,G,P), positive residues (H,K,R), negative residues (D,E), and non-charged polar residues (N,Q,T,S). The interactions that both helical and strand peptides have with their environments were then

examined by identifying all residues that form contacts with peptide residues that are at least four amino acids away in the sequence. Residue contacts were determined using the CSU software,²⁸ which considers residues to be in contact if a hypothetical solvent molecule can not be placed between them. The five amino acid classes gave rise to 25 possible interaction types. Local interactions were also examined by counting pairwise interactions between amino acids in the peptide and amino acids in the environment ($400 = 20 \times 20$ interaction types).

Tertiary structural class

Previously, it was proposed that tertiary structural class can be used to characterize examples of identical peptide sequences that adopt significantly different structures in different proteins.²³ This issue was revisited here using an enhanced dataset and also considering the peptide length. The five major classes provided in the SCOP database²⁹ (<http://scop.mrc-lmb.cam.ac.uk/scop/index.html>) were considered: All Alpha (α); All beta (β); Alpha and beta (α/β); Alpha plus beta ($\alpha+\beta$); and Multi-domain alpha and beta ($m(\alpha/\beta)$). All n-mers ($4 \leq n \leq 7$) in DB1 that have at least *partial helix-to-strand transition* and for which SCOP classifications exist for both proteins were used. The proportion of n-mer pairs that fall into the same and different tertiary structural classes was calculated and compared.

Alerting of structurally ambivalent sequence (ASAS)

We have developed a computer program, ASAS (Alerting of Structurally Ambivalent Sequence), accessible at <http://cbrg.inf.ethz.ch/ASAS.html>. It incorporates the statistical data on structurally ambivalent peptides described here. Given an input sequence, the program parses the sequence, matching its subsequences (4-mer to 9-mer) against our database of structurally ambivalent peptides, which is composed of the peptides showing *helix-to-strand transition* in DB1 and DB2. Due to the high structural flexibility of short peptides (vide infra) we only list strongly structurally ambivalent 4- and 5-mers. A peptide is defined to be strongly structural ambivalent if it has been found in n pairs of proteins in DB1 of which m pairs show *helix-to-strand transition*, where $m > 2$ and $m/n > 0.5$. Longer subsequences (6-mers to 9-mers) are returned if they have shown any structural ambivalence in DB1 or DB2.

All general computation is conducted within the Darwin programming environment.³⁰

RESULTS

Structural Ambivalence of Peptides With Identical Sequences

Table I shows that a high fraction (16–31%) of the n-mer pairs in DB1 undergo *helix-to-strand transition*. Of all 20^4 possible 4-mers, nearly a third are found in at least two structures in DB1. Of this third, about half undergo either *complete* or *partial helix-to-strand transition*. Even for longer peptides, a surprisingly high fraction shows *partial helix-to-strand transition*. In contrast, few longer peptides

TABLE I. Statistics of the Structural Ambivalent Peptides in the DB1 and DB2

	4mer	5mer	6mer	7mer	8mer
Survey of the DB1					
Total number of n-mer pairs	52,349	14,284	1,029	73	6
Number of n-mer pairs with partial helix-to-strand transition	16,710	2,661	181	16	1
	31.9%	18.6%	17.6%	21.9%	16.7%
Number of n-mer pairs with complete helix-to-strand transition	6,999	563	21	0	0
	13.4%	3.9%	2.04%		
Survey of the DB2					
Total number of n-mer pairs	91,533	62,727	7,809	1,932	1,162
Number of n-mer pairs with partial helix-to-strand transition	39,512	12,292	1,009	86	5
	43.2%	19.6%	12.9%	4.5%	0.4%
Number of n-mer pairs with complete helix-to-strand transition	18,572	2,580	117	2	0
	20.3%	4.1%	1.50%	0.10%	

in DB1 undergo *complete helix-to-strand transition*. One explanation for this comes from the fact that most strands in DB1 are five residues, or shorter, in length. Thus, it is more difficult to find long strands, let alone strands that undergo a transition to a helix. Because protein pairs with less than 25% sequence identity generally adopt different folds, the ambivalence of peptides found in DB1 explicitly reveals a high probability for peptides to fold into distinct secondary structures in different global environments.

In DB2, the percentage of n-mers that are structurally ambivalent decreases with increasing length. This is likely to be due to the fact that as n-mers increase in length, there is a greater chance that they derive from related sequences. Interestingly, comparing n-mers from DB1 to DB2 shows an increase in the fraction of short peptides that are structurally ambivalent. Both 4-mers and 5-mers are more likely to undergo *partial* or *complete helix-to-strand transition* in DB2, which allows 95% sequence identity between structure pairs. This indicates that short peptides are quite flexible, even in the context of very similar environments. This is far more difficult for longer peptides as a similar protein architecture imposes physical constraints limiting transitions. For example, the i to $i+4$ C α distance in a helix is about 6Å, but nearly 14Å for a strand.

During the survey of PDB we identified a naturally occurring nonapeptide, “KGVVPQLVK,” that shows *partial helix-to-strand transition* in two proteins, namely Importin Alpha (PDB Code: 1IALA,³¹ residues 292–300) and Pyruvate Kinase (PDB code: 1PKYA,³² residues 413–421), that share only 11% sequence identity. The secondary structure assignments of these peptides are CCHHHH-HHH and TTEEEEECC, respectively. The two distinct structures of this peptide are illustrated in Figure 1. The protein Importin Alpha is an all alpha fold, and while the P296 and G293 in the peptide have the combined effect of disrupting the regular hydrogen bonding pattern of the helix, surprisingly, a number of ordered water molecules compensate for this. One (WAT80) mediates a hydrogen bond between the backbone O of G293 and the backbone N of Q297. All of the hydrophobic residues are at least

partially buried, while the hydrophilic residues are exposed to solvent. In pyruvate kinase, an alpha/beta fold, the peptide adopts a strand that lines the base of a cleft in the molecule. As an edge strand, P417 does not cause a loss of an inter-strand hydrogen bond as its nitrogen is pointing away from the adjacent strand. K413 is buried, but forms a salt link with D336. Both V416 and L419 are exposed, suggesting that the local structure is not optimal but has been “sacrificed” to best accommodate the rest of the fold. Although the existence of structurally ambivalent 9-mers has been previously speculated,²⁵ this is the first report identifying such a peptide. Additionally, two heptapeptides, “IKMFIKN” and “LITTAHA,” that show *complete helix-to-strand transition* (HHHHHHH/EEEEEEE) have been identified in protein sequence pairs that have less than 10% sequence identity: Phosphoribosyl Anthranilate Isomerase (PDB code:1NSJ³³) and Glycosylasparaginase (PDB code:1PGS³⁴) as well as Cyclodextrin Glycosyltransferase (PDB code:1CGT³⁵) and Beta-Galactosidase (PDB code: 1BGLA³⁶). Thus, peptides exhibit an intrinsic plasticity that allows them to adopt completely different structures that adapt to their environments.

Sequence Properties of the Structurally Ambivalent Peptides

Interesting features emerge from the analysis of the residue composition in n-mers with *complete helix-to-strand transition* (Fig. 2). Residues with relative frequencies greater than one are more likely to appear in structural ambivalent peptides than not. Interestingly, the residues with the highest frequencies are those with hydrophobic aliphatic side-chains (L,V,I,A). A possible explanation for this observation may be the rather unspecific nature of hydrophobic interactions, which allow a high degree of possible orientations in a hydrophobic environment, favoring structural transitions. Most polar side chains that are subject to directional hydrogen bonding have relative frequencies below one (see for instance N, D, Q, H, K, S, T in Fig. 2). The lowest frequencies (below 0.3) are found for residues C, P, and W. It is likely that the

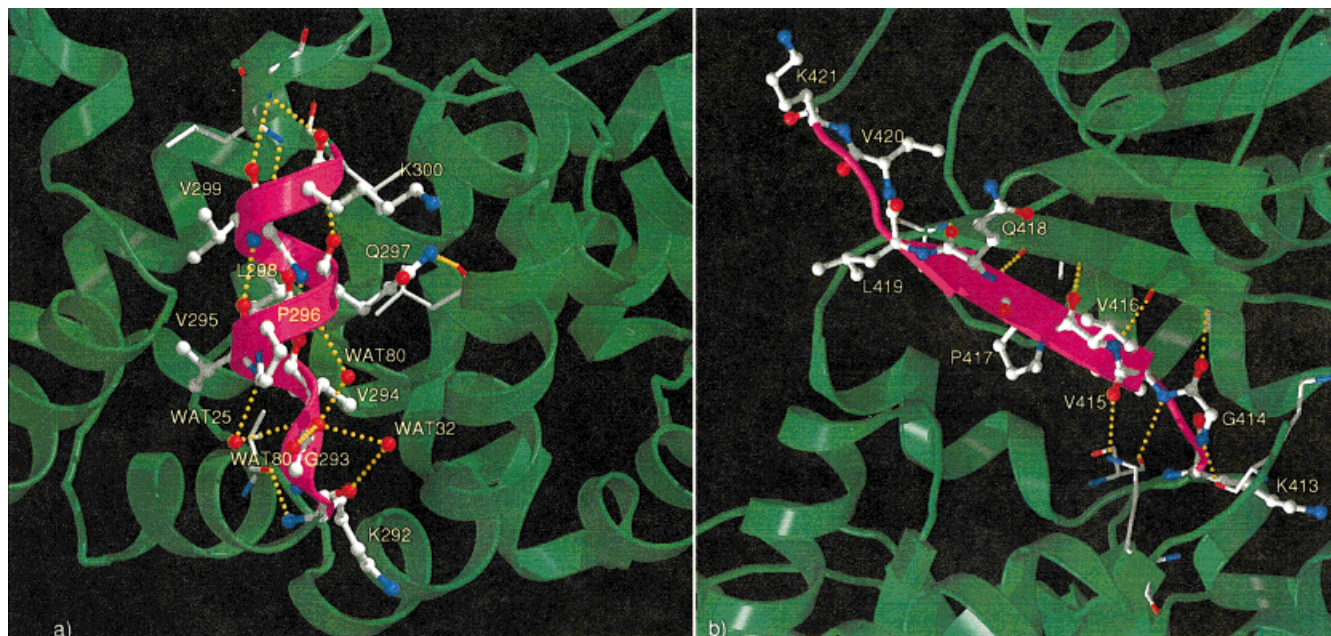


Fig. 1. Schematic illustration of the nonamer "KGVVPQLVK" in a 1IAL and in b 1PKY. Residues of the nonamer are in ball and stick mode with the ribbon of the nonamer highlighted with magenta. The rest of the protein is drawn in green. Hydrogen bonds of the nonamer are shown with

dotted lines. Residues and crystallographic water molecules, which hydrogen bond to the nonamer, are illustrated as sticks and balls, respectively. The figure was prepared with RASTER3D.⁶⁸

large volume of W and the potential of C to form S-S bonds reduce their flexibility in structure formation. The disruptive nature of P in the context of a helix or strand formation may be the reason for its low frequency in ambivalent peptides.

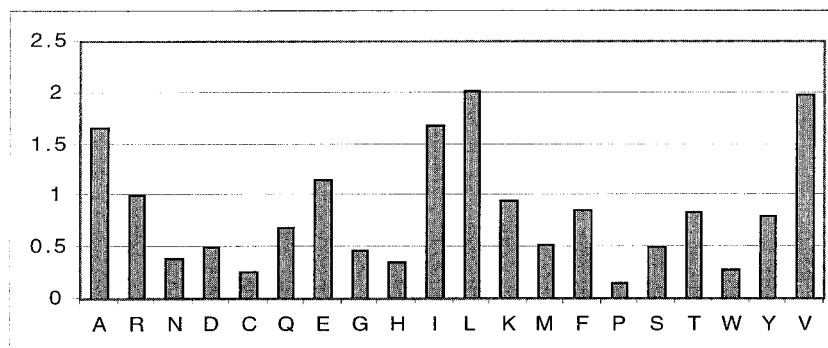
It is striking that all residues with frequencies larger than 1 are either strong α -helix formers (e.g., A and E) or strong β -strand formers (e.g., V and I) according to the Chou-Fasman parameters.¹ Leucine, the residue with the highest relative frequency, is known to have a strong propensity for both helix and strand formation. Moreover, our analysis reveals that the occurrence of strong α -helix and strong β -strand formers are highly correlated. In one third of all structural ambivalent 4-mers and nearly half of all 5-mers as well as about two thirds of 6-mer pairs, one of the two strong helix formers (A or E) and one of the two strong strand formers (V or I) appear simultaneously. We

have also calculated the most frequently appearing dipeptide combinations normalized by their natural frequency. Eight out of ten of the most frequently occurring dipeptides are again composed of a strong helix former and a strong strand former (namely LL, IA, AV, LV, LI, AI, EL, EI). However, the influence of the environment should not be underestimated as this can override any intrinsic helical or strand preferences for peptides. For example, the strongly helical peptides EAAAA forms a strand in the hydrolase inhibitor (PDB code: 1HLEA³⁷), while the peptide VVVIV, that contains strong strand forming residues, is helical in the lipoprotein (PDB code: 1SPF³⁸).

Local Sequence Properties of Structurally Ambivalent Peptides

The relative residue frequencies of the amino acids neighbouring structural ambivalent n-mers (Fig. 3) sug-

Fig. 2. The relative frequencies of the 20 amino acids in structurally ambivalent n-mers. The occurrence frequencies of the 20 residues are calculated based on the n-mers with complete helix-to-strand transition derived from DB1, and then normalized with their natural frequency. For example, A has a relative frequency 1.67, which means the occurrence frequency of A in the structural ambivalent n-mers is 1.67 times its natural frequency.



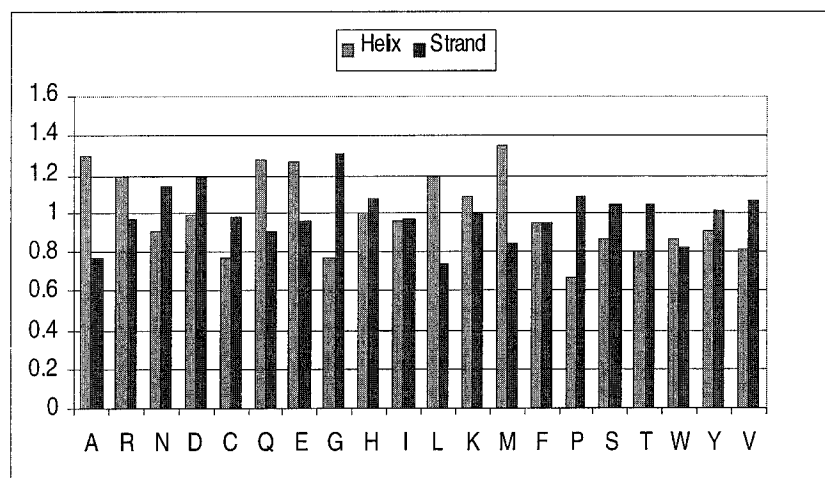


Fig. 3. The relative residue frequency of the sequence neighbours of the n-mers with *complete helix-to-strand transition* derived from DB1. The figure compares residue frequencies neighbouring helical and strand peptides.

TABLE II. Occurrence Frequency (%) of the Long-Range Interactions Between the Helical and Strand Peptides and Their Respective Environments

In the environment		Structurally ambivalent peptides				
		Aromatic	Non-polar	Non-charged polar	Positive	Negative
Aromatic	Helix	0.14	3.58	1.48	1.72	1.94
	Strand	0.16	3.48	1.54	1.52	1.85
Non-polar	Helix	1.06	27.09	9.31	10.01	10.27
	Strand	1.36	26.69	8.62	9.72	10.29
Non-charged polar	Helix	0.11	4.63	2.51	2.45	2.78
	Strand	0.13	4.84	2.74	2.52	3.04
Positive	Helix	0.09	3.68	2.18	2.02	4.48
	Strand	0.14	3.74	2.23	2.02	4.57
Negative	Helix	0.08	2.43	1.61	2.55	1.7
	Strand	0.12	2.72	1.68	2.74	1.56

gest that helices are normally flanked by strong α -helix forming residues (e.g., A, Q, E, M) while β -strands are less frequently bounded by strong strand formers. This is reasonable since helices are predominantly much longer than strands, as already discussed.

Interestingly, structurally ambivalent peptides that adopt helical structure are seldom bounded by helix breakers such as P or G. Conversely, sequences surrounding strands show a preference for P and G. This suggests that many structurally ambivalent peptides fold into a β -strand only if α -helix folding is not possible due to the presence of a helix-breaker in the sequence neighbourhood. As helix formation is a local process, while strand formation needs global interaction, this result implies that at least in some regions, local folding happens prior to global folding, which supports the hierarchical model of protein folding.³⁹ For the structurally ambivalent peptides, the presence or absence of P and G in their neighborhood may give some clues as to which secondary structure might be preferred.

Local Environment Properties of Structurally Ambivalent Peptides

Classifying the residues into five types (aromatic, non-polar, positive, negative, and noncharged-polar), we inves-

tigated the long-range interactions between residues in the structurally ambivalent peptides and their spatial neighbors to search for trends in helix or strand formation. Our results (Table II) show that the occurrence frequency of each interaction type in helix formation and strand formation does not vary much in general. Counting all pairwise interactions between amino acids in the peptide and amino acids in the environment ($400 = 20 \times 20$ interaction types) also fails to give useful indicators for formation of specific secondary structure. Likewise, applying linear regression and more sophisticated statistical models such as “conditional logistic regression for matched pair analysis” did not uncover useful trends (data not shown). Although long-range interactions are considered important for the structural ambivalence of sequentially identical peptides, they seem to be too subtle to be summarized by the above statistics analysis.

Global Environments Implied by Tertiary Structure Class

We have compared the tertiary structural classes of protein pairs containing structurally ambivalent n-mers. Among the n-mer pairs with *complete helix-to-strand transition* in DB1, only 17.8% of 4-mer pairs, 14.4% of

TABLE III. Structural Ambivalence of the Same Protein Under Different Conditions[†]

Protein	Residue range	Local sequence	PDB ID	Local structure
Elongation factor Tu	57–58	APEER ARGIT	1EFT 1TUIA	CHHH HH HCCC CCCC EE TTEE
Cyclin-dependent kinase	150–151	FGLAR AF GVP	1B39A 1FINA	TTHHH HH HCCT TTTT EE TTTT
Inositol monophosphatase	42–44	SPVDL V TATDQ	1IMF 1IMDA	XXX C HHHHHHH ETTE EEEE HHHH
Antithrombin	114–115	TISE KT SDQI	1AZXI 1BR8I	TTCH HH HHHH GGT TE HHHH
Lactoferrin	324–325	LGS G YFTA I Q	1LFI 1LCT	HCHHH HH HHHH HHC CE CCXX
Tetranectin	49–52	ALQ T V CL KG T KV	1HTN 1TN3	HHHH HH HC EE CCCC EEEE TTTT
Hemagglutinin	61–62	E K T N E K F H Q I	1HGB 1HTMB	CCCC EE CCCC HHHH HH HHHH

[†]The column “residues” gives the residue numbers that show *helix-to-strand transition*. The column “local sequence” gives the peptide sequence around the residues having structural ambivalence (bold). The column “local structure” shows the local secondary structures around the structurally ambivalent residues (bold) in different PDB structures. The secondary structure assignments are: H, helix; E, strand; C, coils; T, turn; X, no secondary structure assigned since crystallographic data does not exist.

5-mer pairs, and 25.0% of 6-mer pairs derive from proteins with the same tertiary structural class. Similarly, for n-mer pairs with *partial helix-to-strand transition* only 20.2% of 4-mer pairs, 17.5% of 5-mer pairs, and 22.6% of 6-mer pairs are from proteins with the same tertiary structure class. Hence, in DB1 the probability of n-mer pairs adopting different structures in proteins from different classes is much higher than for proteins from the same tertiary structural class; 41.7% of 7-mer pairs with *partial helix-to-strand transition*, are from structures with the same tertiary class. Interestingly then, the probability for n-mer pairs to adopt different structures in proteins with the same tertiary class appears to increase with n-mer length. This is likely to stem from the peptides coming from related sequences. While this appears counter intuitive, it is also noted that the total number of n-mer pairs drops sharply with length. Thus, as suggested in earlier studies,²³ information of protein tertiary classes can contribute to predict the conformation of its subsequences, however not in all cases. Among the ambivalent n-mer pairs that derive from proteins from the same tertiary class, about two thirds are from the α/β class, suggesting that particular care should be taken when making secondary structure predictions for this class.

Structural Ambivalence Within the Same Protein

The PDB database includes many proteins that have more than one structure deposited. This redundancy allows an investigation as to whether secondary structure variation is possible within the same protein. During the survey of the whole PDB database, seven proteins were identified that exhibited *helix-to-strand transition* under different conditions (Table III). There appear to be two major reasons for this structural ambivalence: (1) the binding of different ligands or other proteins, such as in the structure pairs of Elongation Factor Tu (PDB codes: 1EFT⁴⁰/1TUIA⁴¹), Cyclin-Dependent Kinase (PDB codes: 1B39A⁴²/1FINA⁴³), Inositol Monophosphatase (PDB codes:

1IMF⁴⁴/1IMDA⁴⁴) and Antithrombin (PDB codes: 1AZXI⁴⁵ / 1BR8I⁴⁶); and (2) the cleavage of a protein sub-domain, such as in Lactoferrin (PDB codes: 1LCT⁴⁷/ 1LFI⁴⁸) and Tetranectin (PDB codes: 1HTN⁴⁹/1TN3⁵⁰). The structure pair 1HGB⁵¹/1HTMB⁵² of Hemagglutinin involved both different ligand-binding and sub-domain cleavage. For both 1LCT/1LFI and 1HTN/1TN3, the structurally ambivalent residues lie near the cleavage sites, which are within helices. After cleavage, what remains of the helices becomes a strand in the crystal structures of the cleaved molecules. Both the interaction with other molecules and subunit cleavage, result in changes to the environment of residues in the structures, again pointing to the dramatic influence that the environment has on the local structure. Such structurally ambivalent sequence region may promote conformational changes closely correlated with the biological function of the protein.

Alerting the Structurally Ambivalent Sequences

The structural ambivalence of peptide sequence poses a dilemma for secondary structure prediction. For example, with the PHD method, confusion between helix and strand occur on average for about 8% of all residues.⁵³ We compared the prediction results made by the PHD method, for the 16 target proteins at CASP2, with the crystallographic secondary structure assignment. The 16 target sequences contain a total of 3,430 overlapping tetramer peptides, out of which we identify 167 as being strongly structurally ambivalent. Of the 167 peptides, PHD makes a serious misprediction, either partial or complete confusion between helix and strand, in 22 (13.2%) cases. For the rest of the tetramers, PHD only makes a serious error in 173 (5.3%) cases. Thus, serious mispredictions are nearly 2.5 times more likely to occur in strongly structurally ambivalent tetramers. Similar statistics are found for pentamer peptides; however, we note the statistics are quite sparse: one out of eight identified strongly structurally ambivalent pentamers contain a serious mispredic-

tion (12.5%) compared to 223 serious errors in the remaining 3,406 pentamers (6.5%). At CASP3 there are two sequence segments, GSSEL(CCCEE) in Cyanovirin-N (PDB code:2EZM⁵⁴) and LIAGG(EEEEET) in Flavin Reductase (PDB code:1QFJ⁵⁵), for which most prediction methods erroneously assigned the strand conformation as helix. These two sequences are also found in our database to be strongly structural ambivalent pentamers. Although secondary structure formation is partly determined by non-local residue interactions, this result confirms that there are also local sequence patterns related to structural ambivalence.

Our web service “Alerting of Structurally Ambivalent Sequence,” accessible at <http://cbrg.inf.ethz.ch/ASAS.html>, parses the submitted sequence and returns all of its subsequences that occur in our structurally ambivalent database. Information about the names of the constituent protein pairs, as well as the local secondary structures of the structurally ambivalent peptides, are made available for the user. This service provides a complementary means for secondary structure prediction.

DISCUSSION

We have surveyed the protein structure databases with different degrees of non-redundancy for peptide sequences of variable length ($> = 4$ residues), which adopt an α -helix structure in one protein and a β -strand structure in another protein. In comparison to previous work,^{20-23,25} we have found a dramatically increased number of structurally ambivalent pentamers, hexamers, and heptamers due to the rapid growth of the PDB database. This is despite the fact that in the previous studies, the database of protein structures that was considered allowed up to 50% sequence identities. Here, in DB1, only structures with mutually less than 25% sequence identity are allowed. Also, only peptides that undergo a transition between a helix and a strand are investigated. Such peptides have the potential to generate serious errors in secondary structure prediction.

Short peptide sequences, four or five amino acids long, appear most flexible. This may explain why the majority of insertions and deletions (indels) are of this length.⁵⁶ While, by definition, it is increasingly difficult to find longer identical peptides in unrelated sequences, the results presented here suggest that as many as one in six identical 8-mers form variable local structures. This points to the innate flexibility of peptide sequences to adapt to best match their environment. Moreover, we report the first naturally occurring nonapeptide showing *helix-to-strand transition* in two evolutionary unrelated proteins, suggesting that structural ambivalence may be even more prevalent than previously suspected. The seven proteins, which show *helix-to-strand transition* upon ligand binding and subunit cleavage, emphasize the importance of the global environment for the secondary structure formation. The engineered elevenmer “chameleon” sequence designed by Minor and Kim⁵⁷ strongly supports these findings.

The broad existence of structurally ambivalent peptides does not mean that sequence properties are completely indifferent to secondary structure formation. To the contrary, we find two sequence factors correlated highly with structural ambivalence. Firstly, hydrophobic residues with aliphatic side-chain (e.g., A,L,V,I) appear frequently in the structural ambivalent peptides. Secondly, we find strong helix formers and strong strand formers appearing together to be a good indication for structural ambivalence. Strikingly, the “chameleon” peptide (AWTVEKAFKTF) is comprised almost exclusively from amino acids that are shown to be prevalent in other structurally ambivalent peptides (Fig. 2) and includes both strong helix and strand formers. These results confirm that short-range interactions and the intrinsic properties of amino acids in peptide sequences are important in determining their structures. Many other elegant experiments have demonstrated structural changes in response to sequence mutations. For example, a strand has been converted to a helix by the introduction of single point mutations.^{58,59} Likewise, a helical 4-mer was caused to become part of a loop with the insertion of four residues in T4 lysozyme.⁶⁰ Thus, alternate structures can be induced by minor sequence change, and while intrinsic sequence properties can be important for structure formation, these preferences can be forgone to satisfy the structure of the molecule as a whole.

The structural ambivalence of peptides has ramifications for health issues. The amyloid protein associated with Alzheimer’s disease and the prion protein associated with scrapie diseases, are suggested to undergo a conformational change.^{61,62} Peptide sequences within these proteins that are known to have structural ambivalence would be key sites to examine in such molecules for potential therapeutic intervention. Recently a search for “molecular switches” has been undertaken for a number of proteins using the premise that if a region of local sequence lacks strong intrinsic secondary structure, then by changing its environment, its structure can also change.^{63,64}

The structural diversity of peptides presents difficulties for secondary structure prediction programs, particularly those that rely only on local sequence information.^{1,65} Likewise, tertiary structure prediction algorithms that work on the “spare parts” principle^{66,67} are also affected. The information gathered in this study has been used to build a filter that can identify regions of a sequence that are likely to be misassigned. Application of this filter is encouraging although the filter is also under the same limitations as all approaches that are based only on local sequence.

ACKNOWLEDGMENTS

The authors thank Walter Gander for his helpful advice in the mathematical modeling of long-range interactions. Steven A. Benner, Chantal Korostensky, Mike Hallett, and Ari Kahn are also gratefully acknowledged for their helpful suggestions. G.C. acknowledges support from the Australian Research Council.

REFERENCES

- Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* 1978;47:45-148.
- Munoz V, Serrano L. Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins* 1994;20:301-311.
- Lyu PC, Liff MI, Marky LA, Kallenbach NR. Side chain contributions to the stability of α -helical structure in peptides. *Science* 1990;250:669-673.
- O'Neil KT, DeGrado WF. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* 1990;250:646-651.
- Padmanabhan S, Marquese S, Ridgeway T, Laue TM, Baldwin RL. Relative helix-forming tendencies of nonpolar amino acids. *Nature* 1990;344:268-270.
- Kim CA, Berg JM. Thermodynamic β -sheet propensities measured using a zinc-finger host peptide. *Nature* 1993;362:267-270.
- Blaber M, Zhang XJ, Matthews BW. Structural basis of amino acid α helix propensity. *Science* 1993;260:1637-1640.
- Minor DL Jr, Kim PS. Measurement of the β -sheet-forming propensities of amino acids. *Nature* 1994;367:660-663.
- Smith CK, Withka JM, Regan L. A thermodynamic scale for the β -sheet forming tendencies of the amino acids. *Biochemistry* 1994;33:5510-5517.
- Creamer, TP, Rose GD. α -helix-forming propensities in peptides and proteins. *Proteins* 1994;19:85-97.
- Lim VI. Algorithms for prediction of α -helical and β -structural regions in globular proteins. *J Mol Biol* 1974;88:873-894.
- Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 1978;120:97-120.
- Levin JM. Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng* 1997;10:771-776.
- Zhong L, Johnson WC Jr. Environment affects amino acid preference for secondary structure. *Proc Natl Acad Sci USA* 1992;89:4462-4465.
- Waterhous DV, Johnson WC Jr. Importance of environment in determining secondary structure in proteins. *Biochemistry* 1994;33:2121-2128.
- Blondelle SE, Ostresh JM, Houghten RA, Perez-Paya E. Induced conformational states of amphipathic peptides in aqueous/lipid environments. *Biophys J* 1995;68:351-359.
- Minor DL Jr, Kim PS. Context-dependent secondary structure formation of a designed protein sequence. *Nature* 1996;380:730-734.
- Blondelle, SE, Forood B, Houghten RA, Perez-Paya E. Secondary structure induction in aqueous vs. membrane-like environments. *Biopolymers* 1997;42:489-498.
- Bernstein FC, Koetzle TF, Williams GJ, Meyer EE, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol* 1977;112:535-542.
- Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci USA* 1984;81: 1075-1078.
- Wilson IA, Haft DH, Getzoff ED, Tainer JA, Lerner RA, Brenner S. Identical short peptide sequences in unrelated proteins can have different conformations: A testing ground for theories of immune recognition. *Proc Natl Acad Sci USA* 1985;82:5255-5259.
- Argos P. Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis. *J Mol Biol* 1987;197:331-348.
- Cohen BI, Presnell SR, Cohen FE. Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci* 1993;2: 2134-2145.
- Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci USA* 1996;93:5814-5818.
- Sudarsanam, S. Structural Diversity of sequentially identical subsequences of proteins: identical octapeptides can have different conformations. *Proteins* 1998;30:228-231.
- Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins* 1995;23:566-579.
- Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. *Protein Sci* 1992;1:409-417.
- Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. *Bioinformatics* 1999;15:327-332.
- Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. *Nucleic Acids Res* 1999;27:254-256.
- Gonnet GH, Hallett MH, Korostensky C, Bernardin L. An interpretive programming language for biological application: Darwin. *Bioinformatics*, 2000;16:101-103.
- Kobe B. Autoinhibition by an internal nuclear localization signal revealed by the crystal structure of mammalian importin alpha. *Nat Struct Biol* 1999;6:388-397.
- Mattevi A, Valentini G, Rizzi M, Speranza ML, Bolognesi M, Coda A. Crystal structure of *Escherichia coli* pyruvate kinase type I: molecular basis of allosteric transition. *Structure* 1995;3:729-741.
- Hennig M, Sterner R, Kirschner K, Jansonius JN. Crystal structure at 2.0 Å resolution of phosphoribosyl anthranilate isomerase from the hyperthermophile *Thermotoga maritima*: possible determinants of protein stability. *Biochemistry* 1997; 36:6009-6016.
- Norris GE, Stillman TJ, Anderson BF, Baker EN. The three-dimensional structure of PNGase F, a glycosylasparaginase from *Flavobacterium meningosepticum*. *Structure* 1994;2:1049-1059.
- Klein C, Schulz GE. Structure of cyclodextrin glycosyltransferase refined at 2.0 Å resolution. *J Mol Biol* 1991;217:737-750.
- Jacobson RH, Zhang XJ, DuBose RF, Matthews BW. Three-dimensional structure of beta-galactosidase from *E.coli*. *Nature* 1994;369:761-766.
- Baumann U, Bode W, Huber R, Travis J, Potempa J. Crystal structure of cleaved equine leucocyte elastase inhibitor determined at 1.95 Å resolution. *J Mol Biol* 1992;226:1207-1218.
- Johansson J, Szyperki T, Curstedt T, Wuthrich K. The NMR structure of the pulmonary surfactant-associated polypeptide SP-C in an apolar solvent contains a valyl-rich alpha-helix. *Biochemistry* 1994;33:6015-6023.
- Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem Sci* 1999;24:26-33.
- Kjeldgaard M, Nissen P, Thirup S, Nyborg J. The crystal structure of elongation factor EF-Tu from *Thermus aquaticus* in the GTP conformation. *Structure* 1993;1:35-50.
- Polekhina G, Thirup S, Kjeldgaard M, Nissen, P, Lippman C, Nyborg J. Helix unwinding in the effector region of elongation factor EF-TU-Gdp. *Structure* 1996;4:1141-1151.
- Brown NR, Noble ME, Lawrie AM, Morris MC, Tunnah P, Divita G, Johnson LN, Endicott JA. Effects of phosphorylation of threonine 160 on cyclin-dependent kinase 2 structure and activity. *J Biol Chem* 1999;274:8746-8756.
- Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massague J, Pavletich NP. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature* 1995;376:313-320.
- Bone R, Frank L, Springer JP, Atack JR. Structural studies of metal binding by inositol monophosphatase: evidence for two-metal ion catalysis. *Biochemistry* 1994;33:9468-9476.
- Jin L, Abrahams JP, Skinner R, Petitou M, Pike RN, Carrell RW. The anticoagulant activation of antithrombin by heparin. *Proc Natl Acad Sci USA* 1997;94:14683-14688.
- Skinner R, Chang WS, Jin L, Pei X, Huntington JA, Abrahams JP, Carrell RW, Lomas DA. Implications for function and therapy of a 2.9 Å structure of binary-complexed antithrombin. *J Mol Biol* 1998;283:9-14.
- Day CL, Anderson BF, Tweedie JW, Baker EN. Structure of the recombinant N-terminal lobe of human lactoferrin at 2.0 Å resolution. *J Mol Biol* 1993;232:1084-1100.
- Smith CA, Baker HM, Shongwe MS, Anderson BF, Baker EN. Crystallographic studies on metal and anion substituted human lactoferrin. *Adv Exp Med Biol* 1994;357:265-269.
- Nielsen BB, Kastrup JS, Rasmussen H, Holtet TL, Graversen JH, Etzerodt M, Thogersen HC, Larsen IK. Crystal structure of tetranectin, a trimeric plasminogen-binding protein with an alpha-helical coiled coil. *FEBS Lett* 1997;412:388-396.
- Kastrup JS, Nielsen BB, Rasmussen H, Holtet TL, Graversen JH, Etzerodt M, Thogersen HC, Larsen IK. Structure of the C-type lectin carbohydrate recognition domain of human tetranectin. *Acta Crystallogr D Biol Crystallogr* 1998;54:757-766.

51. Sauter NK, Hanson JE, Glick GD, Brown JH, Crowther RL, Park SJ, Skehel JJ, Wiley DC. Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography. *Biochemistry* 1992;31:9609–9621.
52. Bullough PA, Hughson FM, Skehel JJ, Wiley DC. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* 1994;371:37–43.
53. Rost B. Better 1D predictions by experts with machines. *Proteins* 1997; Suppl 1: 192–197.
54. Bewley CA, Gustafson KR, Boyd MR, Covell DG, Bax A, Clore GM, Gronenborn AM. Solution structure of cyanovirin-N, a potent HIV-inactivating protein. *Nat Struct Biol* 1998;5:571–578.
55. Ingelman M, Ramaswamy S, Niviere V, Fontecave M, Eklund H. Crystal structure of NAD(P)H: flavin oxidoreductase from *Escherichia coli*. *Biochemistry* 1999;38:7040–7049.
56. Pascarella S, Argos P. Analysis of insertions/deletions in protein structure. *J Mol Biol* 1992;224:461–471.
57. Minor DL Jr, Kim PS. Context-dependent secondary structure formation of a designed protein sequence. *Nature* 1996;380:370–374.
58. Yang WZ, Ko TP, Corselli L, Johnson RC, Yuan HS. Conversion of a beta-strand to an alpha-helix induced by a single-site mutation observed in the crystal structure of Fis mutant Pro26Ala. *Protein Sci* 1998;7:1875–1883.
59. Cordes MH, Walsh NP, McKnight CJ, Sauer RT. Evolution of a protein fold in vitro. *Science* 1999;284:325–328.
60. Heinz DW, Baase WA, Dahlquist FW, Matthews BW. How amino-acid insertions are allowed in an alpha-helix of T4 lysozyme. *Nature* 1993;361:561–564.
61. James TL, Liu H, Uljanov NB, Farr-Jones S, Zhang H, Donne DG, Kaneko K, Groth D, Mehlhorn I, Prusiner SB, Cohen FE. Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc Natl Acad Sci USA* 1997;94:10086–10091.
62. Jacchieri SG. Study of alpha-helix to beta-strand to beta-sheet transitions in amyloid: the role of segregated hydrophobic beta-strands. *Biophys Chem* 1998;74:23–34.
63. Kirshenbaum K, Young M, Highsmith S. Predicting allosteric switches in myosins. *Protein Sci* 1999;8:1806–1815.
64. Young M, Kirshenbaum K, Dill KA, Highsmith S. Predicting conformational switches in proteins. *Protein Sci* 1999;8:1752–1764.
65. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 1999;34:82–95.
66. Claessens M, Van Cutsem E, Lasters I, Wodak S. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng* 1989;2:335–345.
67. Levitt M. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol* 1992;226:507–533.
68. Merritt EA, Bacon DJ. Raster3D Photorealistic Molecular Graphics. *Methods Enzymol* 1997;277:505–524.