Protein Science (1993), *2*, 2134–2145. Cambridge University Press. Printed in the USA.

Origins of structural diversity within sequentially identical hexapeptides

BRUCE I. COHEN,[1] SCOTT R. PRESNELL,[1] AND FRED E. COHEN[1,2,3]

Departments of [1] Pharmaceutical Chemistry, [2] Medicine, and [3] Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94143-0446

(RECEIVED May 28, 1993; ACCEPTED August 23, 1993)

Abstract

Efforts to predict protein secondary structure have been hampered by the apparent structural plasticity of local amino acid sequences. Kabsch and Sander (1984, *Proc. Natl. Acad. Sci. USA 81*, 1075–1078) articulated this problem by demonstrating that identical pentapeptide sequences can adopt distinct structures in different proteins. With the increased size of the protein structure database and the availability of new methods to characterize structural environments, we revisit this observation of structural plasticity. Within a set of proteins with less than 50% sequence identity, 59 pairs of identical hexapeptide sequences were identified. These local structures were compared and their surrounding structural environments examined. Within a protein structural class (α/α, β/β, α/β, $\alpha+\beta$), the structural similarity of sequentially identical hexapeptides usually is preserved. This study finds eight pairs of identical hexapeptide sequences that adopt β-strand structure in one protein and α-helical structure in the other. In none of the eight cases do the members of these sequence pairs come from proteins within the same folding class. These results have implications for class dependent secondary structure prediction algorithms.

Keywords: protein folding; secondary structure prediction; sequence–structure correlates; tertiary structural class

The amino acid sequence of a protein codes for a unique three-dimensional (3D) structure. Unraveling this code has proven extremely difficult. Protein folding is determined by a complex interplay between the local and global preferences of individual amino acid residues within the sequence. In an effort to simplify the computational complexity of the protein folding problem, local structure has been sought as the nidus for the subsequent tertiary condensation of the chain. Secondary structure features may be building blocks for tertiary structures. This is supported by spectroscopic studies that detect stable helical structure (Marqusee & Baldwin, 1990; Wright et al., 1990; Lyu et al., 1991) and β-turns (Wright et al., 1990) in short, isolated polypeptides. Studies that recognize the preferential protection of amide protons that participate in secondary structure early in the folding process (Roder et al., 1988; Hughson et al., 1990) also lend credence to this approach. Thus, secondary structure is a logical intermediate in predicting protein tertiary structure. Moreover, some sequences contain sufficient information to independently code for regular local structure. In spite of the regularity of this local structure, the general success of secondary structure prediction for an individual sequence without additional experimental data has failed to exceed 65%.

In 1984, Kabsch and Sander noted that the Brookhaven Protein Data Bank (PDB; Bernstein et al., 1977) contained examples of sequentially identical pentapeptides that adopted substantially different structures (Kabsch & Sander, 1984). Another research group later found pairs of identical hexapeptides that adopted different structures (Wilson et al., 1985). The immediate impact of these findings on secondary structure prediction efforts was clear: exclusively local sequence composition and ordering were insufficient to accurately predict secondary structure. In an attempt to circumvent the pentapeptide dilemma, larger groups of amino acids (e.g., 9–17 residues) were considered as a possible folding unit. Wodak and colleagues demonstrated that the current database of proteins of known structure was, unfortunately, too small to supply statistically relevant information about longer sequence patterns (Rooman & Wodak, 1988).

Recently, the pace of protein structure determination by X-ray crystallography and NMR spectroscopy has ac-

Reprint requests to: Fred E. Cohen, Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94143-0446.

celerated. In light of this, we have revisited the question of the structural integrity of local sequences and extend the observations of Kabsch and Sander (1984) and Wilson et al. (1985) to include a large set of hexapeptide sequences. This study finds eight pairs of identical hexapeptide sequences that adopt β-strand structure in one protein and α-helical structure in the other. In none of the eight cases do these sequence pairs come from proteins within the same folding class. Thus, it is possible that class dependent secondary structure prediction algorithms may offer a practical route to circumvent the structural plasticity dilemma.

Results

Background

An *n-mer* is a subsequence of a protein containing n consecutive residues. For example, DEHKTL is a hexamer ($n = 6$) taken from a rhinovirus coat protein. An n-mer has many attributes that emerge from the association with a protein from which the subsequence is drawn. This is particularly true when the protein's tertiary structure is known. Attributes can include a set of phi-psi angles, a secondary structure assignment, and a solvent-accessible surface area for the fragment in the context of the folded protein. The subsequence DEHKTL is also found in hemerythrin. The set of DEHKTL, rhinovirus coat protein, and hemerythrin comprise an *n-mer pair*.[1] This study is concerned with n-mer pairs and the conformations adopted by each n-mer in their respective proteins. Some examples are shown in the kinemages.

The data set for this study is composed of 59 n-mer pairs (where n is at least 6) found in an examination of 316 proteins from the Brookhaven Protein Data Bank (version of July 15, 1990). Each pair contains an identical residue sequence found in two proteins that have no more than 50% sequence identity after the two protein sequences are aligned. Each n-mer pair is categorized by the degree of similarity between the local structures adopted by an identical sequence in each protein. Four categories—*same structures*, *distinct regular structures*, *distinct loops*, and *different structures*—are detailed in the Methods section. Briefly, n-mer regions with essentially the same local structures are grouped in the first category. Distinct regular structures have the same assigned secondary structure within the range of the local sequences, but the local structures are sufficiently distinct as judged by a root mean square deviation between matched α-carbon positions. Similarly, the distinct loops category has n-mer pairs with both n-mer segments in loop conformations where the loops adopt structurally distinct aperiodic conformations. Finally, there is a category for different structures. Within the category of different structures, there is a subcategory (*disparate structures*) of eight n-mer pairs that show alpha structure in one protein and beta structure in the other.

Tertiary structural class and local structure differences

Information on the tertiary structural class of a given protein can contribute to superior secondary structural class predictions (Kneller et al., 1990). One of the motivations for revisiting the identical n-mer problem was to see if tertiary structural class could be used to characterize the examples of identical peptide sequences that adopt significantly different structures in different proteins. The concept of "tertiary structural class" was introduced by Levitt and Chothia (1976), who defined five classes: α/α, β/β, α/β, $\alpha + \beta$, and irregular. Each protein can be assigned to one of these classes based on the composition, ordering, and the spatial placement of its secondary structure features. For example, α/α- and β/β-class proteins are primarily composed of α-helices and β-strands, respectively. α/β-Class proteins are generally composed of alternating α-helices and β-strands, while $\alpha + \beta$-class proteins tend to have clustered α and β areas. With each n-mer pair, the tertiary structural classes of both proteins are noted.

The relationships between the tertiary structural class attributes of each n-mer pair and the local structure characterizations are shown in Figures 1 and 2. Figure 1 shows a breakdown of n-mer pairs with respect to the four categories of local structure differences. Each group is further divided to show whether the n-mer pair represents proteins from one or two tertiary structural classes. The bar graph shows that n-mer pairs with different local structure generally represent a pair of proteins from different tertiary structural classes. At the other end of the local structure spectrum, more often than not, same structure pairs represent proteins from the same tertiary structural class.

Only two n-mer pairs, {KDLRRA, 5acn, 1gd1} and {VAGAAA, 1sbt, 2rub}, contain different local structures and come from the same tertiary structural class. All four proteins—aconitase, glyceraldehyde-3-phosphate dehydrogenase, subtilisin, and rubisco—are α/β-class proteins. As seen in Figure 2, 17% of the α/β–α/β pairs have different local structure. Because α/β-class proteins contain a mixture of helical and strand structures, α/β-class proteins might be expected to have more different structure pairs than any other tertiary structural class.

When the different structure group is further refined to include only pairs where one local structure is part of

[1] This n-mer pair may be referenced by a tuple notation, {DEHKTL, 1r08, 1hmq}, where DEHKTL is the n-mer sequence and 1r08 and 1hmq are the PDB entry identifiers for rhinovirus coat protein and hemerythrin, respectively. The ordering of the PDB entry names is arbitrary, so {DEHKTL, 1r08, 1hmq} and {DEHKTL, 1hmq, 1r08} refer to the same n-mer pair.

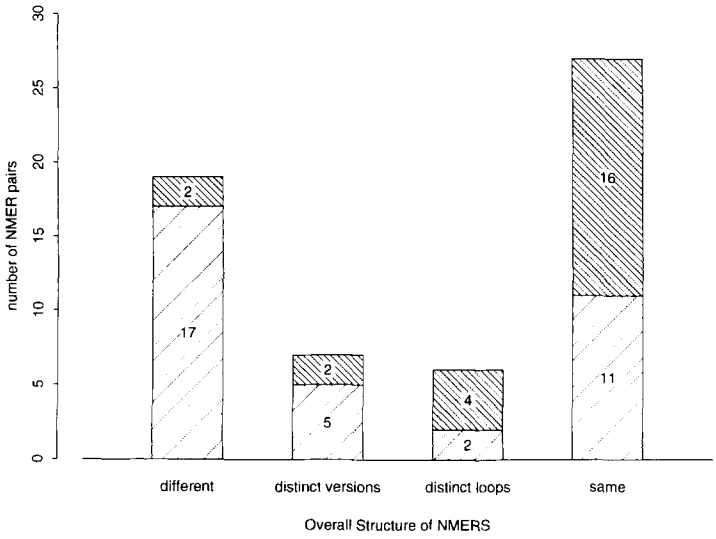


Fig. 1. Summarized tertiary structural class pairing and *n*-mer structure differences. This bar graph presents the tertiary structural class congruence for each of the four local structure groups of *n*-mer pairs described in the text. Each bar is labeled for one of the four groups and divided into two boxes. The upper boxes show the number of pairs where both constituent proteins are members of the same tertiary structural class. The lower boxes show the number of pairs where each protein is characterized by different tertiary structural classes.

an α-helix and the other a β-strand, all eight examples match proteins from different tertiary structural classes (see Table 1). Arguably, this result is not surprising because it is unlikely to find strands in α/α proteins or helices in β/β proteins. Closer examination of these eight examples shows only two cases where the tertiary structural class pairing involves both α/α and β/β. The other six examples pair one of the "pure" tertiary structural classes (α/α or β/β) with one of the mixed classes (α/β or $\alpha + \beta$). Tertiary structural class appears to be correlated with the influences played by "long-range" interactions.

Accessible surface area

It is likely that protein folding represents a balance between the hydrophobic effect driving the chain toward a compact organization and the conformational entropy lost by adopting a single native state (Kauzmann, 1959; Chan & Dill, 1990). Solvent-accessible surface area was introduced in an attempt to quantify the hydrophobic effect (Lee & Richards, 1971) and was subsequently shown to correlate with thermodynamic measures of a side chain's preference for an organic or aqueous phase (Chothia, 1974). For our purposes, the accessible surface area of a residue or set of residues (segment) can be used to describe the environment of a residue or segment. Most of the *n*-mer pairs have segments with similar accessible surface areas (see Fig. 3). This is sensible given the relationship between the hydropathy of a segment and the degree to which it is likely to be buried or exposed. One outlier, apparent in the lower right-hand corner of Figure 3, is the sequence IGHLAT found in both lactate dehydrogenase (6ldh) and tyrosyl-transfer RNA synthetase (2tsl). Although lactate dehydrogenase is a dimer, only

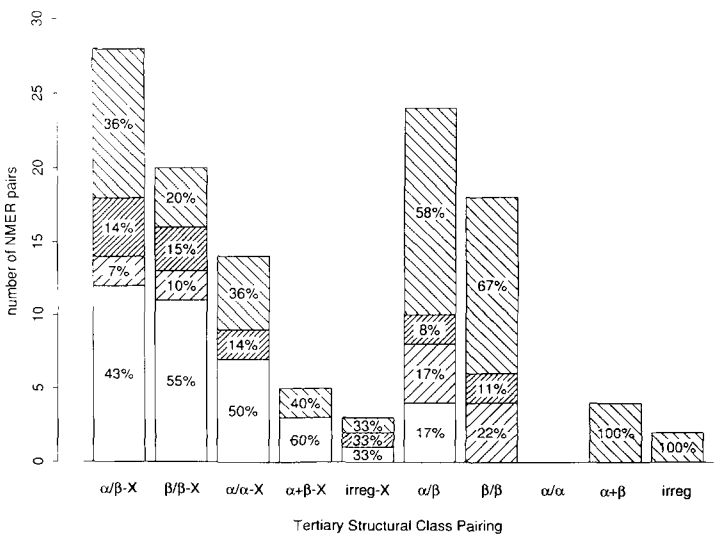


Fig. 2. Tertiary structural class pairing and *n*-mer structural differences. This bar graph contains 10 bars. Each bar is divided into four boxes. The left five bars represent the number of *n*-mer pairs where one constituent protein belongs to a given class (e.g., α/β in the first bar) and the other protein belongs to some other class. The right five bars represent the number of *n*-mer pairs where both proteins belong to the same tertiary structural class. Note that no *n*-mer pair contains two α/α-class proteins. Each *n*-mer pair is tallied twice—once for each protein in the *n*-mer pair. The box divisions (parts) of each bar show structure group counts and percentages. Using the first bar (α/β-X) as an example, the bottom part (▭) shows *different structure n*-mer pair counts, the lower middle part (▨) shows *distinct loops* counts, the upper middle part (▨) shows *different conformations of similar structure* counts, and the top part (▨) shows *same structure n*-mer pair counts. Some bars do not contain all four parts. The percentage breakdown between each of the four groups within each bar is printed in the corresponding box.

Table 1. *Different structure n-mers*[a]

LLKANV	12.2%	2.78 Å		1
lllllllaaaaa	4mdh	99A-104A	α/β	
llblbbbbbbbb	1sgt	127-132	β/β	
FGVGSA	8%	2.84 Å		2
lbbbbbllbbbb	2er7	263E-268E	β/β	
aaaaalaaaaaa	1mba	97-102	α/α	
SGSSAT	12.3%	2.87 Å		3
bbbbllbbbbbb	3fab	61L-66L	β/β	
bbbllaaaaaaa	3bcl	110-115	α/β	
DEHKTL	6.7%	3.14 Å		4
lllllllbbbbb	1r08	2,161-2,211	β/β	
aaaaaaaaaaaa	1hmq	22A-27A	α/α	
NAAIRS	14.5%	3.54 Å		5
laaaaaaaaaaa	3pfk	16-21	α/β	
lllllbbbbbl	2tmn	96E-101E	β/β	
VDLLKN	14%	3.68 Å		6
llbbbbbbllbb	2hla	36B-41B	α + β	
aaaaaaaaaaaa	3wrp	15-20	α/α	
LKAAGA	14%	3.76 Å		7
bbbbbbllaaa	1ctf	5-10	α/β	
aaaaaaallxxx	1mba	140-145	α/α	
LGQLGI	16.9%	4.51 Å		8
aaaaaaaalll	7api	283A-288A	α/β	
lllllllbbbbb	2tbv	135A-140A	β/β	

[a] Each three-line entry includes the following information:

Row 1: The *n*-mer sequence, the sequence identity between the two chains, the RMS deviation between the two *n*-mers, an identification number used in Figure 4.

Rows 2 and 3: A secondary structure assignment, the PDB identifier, the residue numbers, the tertiary structural class of the relevant chain.

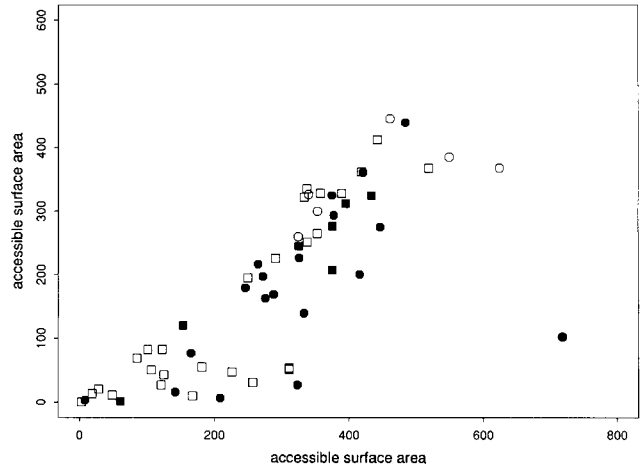


Fig. 3. Accessible surface areas of *n*-mer pairs. Each point of this plot shows a pair of accessible surface areas for an *n*-mer pair. A global accessible surface area is computed for each of the 106 proteins. An accessible surface area is calculated by adding the accessible surface area contributions for each of the *n* residues of one protein in an *n*-mer pair. Since there is no natural ordering of the proteins in an *n*-mer pair, the protein segment with the lower accessible surface area is assigned to the *y*-axis (vertical), placing all points at or below the line $y = x$. Open squares (□) represent same structure pairs. Filled squares (■) represent pairs with similar secondary structure but different tertiary structure. Open circles (○) represent pairs where neither region is in a helix or strand, but the two loops are not similar. Finally, filled circles (●) represent pairs with different local secondary (and generally tertiary) structures.

the monomer was considered in the original *n*-mer search. The *n*-mer segment (residues 7–12) in this enzyme forms a highly exposed loop in the monomer that is buried when forming the dimer interface. There is no apparent relationship between any local substructural group and the variation between paired accessible surface areas. Not surprisingly, the six points representing *n*-mer pairs of distinct loops have large accessible surface areas in both proteins in keeping with the general propensities of loops in proteins. Thus, it appears that these *n*-mer sequences have sufficient backbone conformational plasticity to adopt quite distinct structures as long as the side chain preferences for hydrophobic or hydrophilic environments can be met.

A more subtle measure of a local protein environment can be constructed from a combination of solvent accessibility and backbone conformational preferences. The 3D profile method (Bowie et al., 1991) has been successfully used to compare alternative model structures (Luthy et al., 1992). This method was used to compare the environments of segments described by *n*-mer pairs that adopt different structures. A 3D profile score is computed at each residue based on three factors: (1) the area of the residue buried in the protein and inaccessible to solvent; (2) the fraction of side chain area that is covered by polar atoms (oxygen and nitrogen); and (3) the local secondary structure. In examining alternative models, a 3D profile score is generated for each model and then compared. Moreover, by computing a local 3D score over a moving window along a modeled chain, a poorly modeled segment in an otherwise accurate model should stand out with a low score. Based on this example, the local 3D score of a native *n*-mer segment was compared with the score of an alternative segment—the same segment modified by replacing the secondary structure assignment with that of the paired segment. An example for the *n*-mer pair {VDLLKN, 2hla, 3wrp} is presented in Table 2. Four 3D profiles were generated for each instance of an *n*-mer of interest: (1) the *n*-mer itself; (2) the *n*-mer with the alternative secondary structure; (3) the *n*-mer plus three upstream and three downstream residues; and (4) the *n*-mer plus three upstream and three downstream residues with the alternative secondary structure.

The 3D profile method was applied to *n*-mers from the different structure subgroup. Unfortunately, the profile scores failed to provide adequate contrast over these short regions of the chain. In 6 of 11 cases the native

Table 2. *Example of 3D profile and secondary structure swap*[a]

A - Environment File for Normal Trp aporepressor for Residues 23-28
Environments of Residues in: 3wrp

ResN	Nam	Ab	Fp	SS	Env ..
23	VAL	83.0	0.76	H	P2
24	ASP	54.6	0.90	H	P2
25	LEU	65.6	0.67	H	P2
26	LEU	121.7	0.36	H	B1
27	LYS	69.6	0.82	H	P2
28	ASN	58.5	0.84	H	P2

B - Verify 3D Plot Results File for Normal Trp aporepressor for Residues 23-28
Quality: 1.330000

Sequence	position	accumulated Score	3D-1D Score ..
V	1	-0.48	-0.48
D	2	-0.20	0.28
L	3	-0.55	-0.35
L	4	0.72	1.27
K	5	1.33	0.61
N	6	1.28	-0.05

C - Environment File for Swapped Trp aporepressor for Residues 23-28
Environments of Residues in: 3wrp

ResN	Nam	Ab	Fp	SS	Env ..
23	VAL	83.0	0.76	C	P2
24	ASP	54.6	0.90	C	P2
25	LEU	65.6	0.67	S	P2
26	LEU	121.7	0.36	S	B1
27	LYS	69.6	0.82	S	P2
28	ASN	58.5	0.84	S	P2

D - Verify 3D Plot Results File for Swapped Trp aporepressor for Residues 23-28
Quality: 1.130000

Sequence	position	accumulated Score	3D-1D Score ..
V	1	-0.88	-0.88
D	2	-0.39	0.49
L	3	-1.69	-1.30
L	4	-0.56	1.13
K	5	0.03	0.59
N	6	-0.13	-0.16

[a] In this example, the *n*-mer pair {VDLLKN, 2hla, 3wrp} is considered. First, residues 15–20 of tryptophan aporepressor (3wrp) are run through the 3D profile programs. The 3D profile is done in two steps. First, the PDB entry is analyzed to produce the table shown in part A. This table is then fed to a second program that produces the score shown in part B. The "swapped" version of the *n*-mer is produced by replacing the secondary structure (SS) column of the part A table with the secondary structure of the *n*-mer in the paired protein. In this case, the secondary structure assignments from human class I histocompatibility antigen (2hla) residues 36B–41B are used as shown in part C. Part D shows the score for the swapped *n*-mer segment. In order to look at a larger piece of the polypeptide chain, this same method—3D profiles on a segment with the native and then swapped secondary structure—was used on segments that included the *n*-mer sequence and three residues upstream and downstream on the protein.

structure scored better than the swapped structure. With the larger window (three additional residues at each end of a segment), the native structures scored essentially the same. Here, in 6 of 10 cases (an *n*-mer sequence at the N-terminus of a protein eliminated one of the original 11 cases) no significant change was observed. Two of the 10 protein examples went either to or from higher native scores when the window was enlarged. A real change in secondary structure could also influence the other two factors. For example, an edge β-strand will expose a different set of residues than an α-helix with the same sequence located in the same general part of the 3D structure. The simple act of swapping secondary structure assignments does not account for other changes. Presumably, these sequences can be accommodated in both geometries, but are not ideally suited to either. Thus, the 3D profile method is not suited to recognize the correct match of local sequence and structure.

In an attempt to quantitate the conformational ambivalence of the *n*-mer pair sequences, we investigated the ac-

curacy of secondary structure prediction algorithms as applied to these peptides. Our analysis consisted of three parts. First, the Chou–Fasman secondary structure prediction algorithm (as implemented in the computer program CONFORM [Molecular Biology Information Resource, 1989]) was applied to the *n*-mer sequences. Next, a relationship between amino acid composition and secondary structure content suggested by several prediction methods was explored. Finally, the residue compositions of the segments of the structurally different *n*-mer pairs where tabulated and compared to the compositional biases.

Chou–Fasman prediction on n-mer pairs

Are the sequences of the most structurally disparate *n*-mer pairs composed of residues that are known to be structurally ambivalent? The Chou–Fasman prediction method is based on the observed propensities of amino acids to participate in particular types of secondary structure features (Chou & Fasman, 1978). This method uses these propensities as a measure in evaluating sequence windows for possible α-helix or β-strand structure. The CONFORM program only reports results for a given secondary structure type if the results are above a certain threshold. If the *n*-mer sequences in the proteins of the structurally different subcategory were conformationally ambiguous, then Chou–Fasman could report both α-helix and β-strand propensities above the threshold (exemplified by points in the upper right-hand quadrant of Fig. 4) or both propensities below the threshold. Incorrect predictions on these regions would still be interesting if the Chou–Fasman method encounters strong signals for both helical and extended structure. In one case, {NAAIRS, 3pfk, 2tmn-E} (points A5 and B5 on Fig. 4), Chou–Fasman correctly predicts an α-helix for the *n*-mer segment in 3pfk and a β-strand for the segment in 2tmn-E. This prediction requires the successful incorporation of conformational preferences from the flanking regions. In general, Chou–Fasman predicts an α-helix for six of the eight cases where an α-helix is found in the *n*-mer region. Similarly, Chou–Fasman predicts a β-strand in four of the eight cases where a strand is actually assigned in the region. These predictions are consistent with the 55–65% accuracy expected for secondary structure prediction. The Chou–Fasman method generally reports that both α-helix and β-strand are plausible in the 16 *n*-mer regions. In only one case does Chou–Fasman fail to report that both α-helix and β-strand were possible conformations for at least one member of the *n*-mer pair. In this one example, {FGVGSA, 2er7-E, 1mba} (points A2 and B2 on Fig. 4; structures shown in Kinemage 4), the Chou–Fasman predictions are totally reversed — α-helix for the segment containing a β-strand and β-strand for the segment containing an α-helix. These results highlight the need for structurally context dependent secondary structure prediction algorithms.

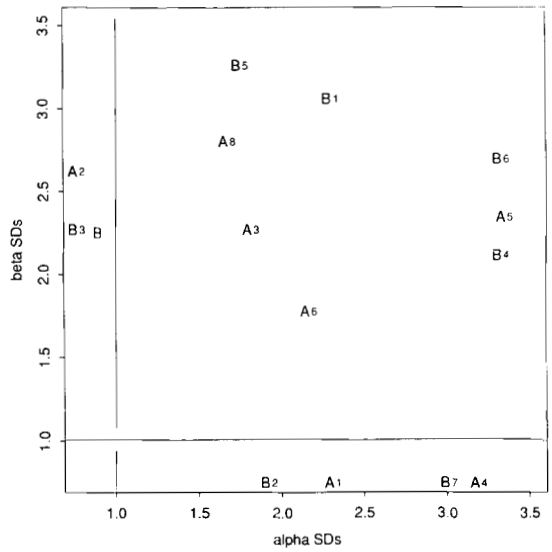


Fig. 4. Chou–Fasman results. This plot shows the results of running the program CONFORM on the eight *n*-mer pairs listed in Table 1. Rather than standard Chou–Fasman $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ results, CONFORM gives the number of standard deviations above a pure chance of an alpha or beta conformation. The letter A is used to plot a point for an *n*-mer found in a helical conformation while the letter B plots a point for an *n*-mer in a strand conformation. The number next to the point relates to the numbering scheme established in Table 1. Note that alpha and beta standard deviations (SDs) below a value of 1.0 have no meaning in the context of these results.

Amino acid composition and secondary structure content

Although secondary structure preferences of amino acids may not be strong enough to accurately predict locations of secondary structure features, amino acid composition has been used to successfully predict the secondary structure content of a protein, a surrogate for protein tertiary structural class. The major difference is that proteins with a mixture of α-helix and β-strand structures are not further divided into $\alpha + \beta$ or α/β classes. Within the last 10 years, statistical approaches to secondary structure content prediction based on amino acid composition have yielded relatively good (~80%) results (Sheridan et al., 1985; Klein & DeLisi, 1986). Even better results were reported by Muskal and Kim (1992), using a tandem computational neural network.

Why is amino acid composition information useful for the prediction of secondary structure content? To revisit this question, we examined the residues involved in the 80 structures from which the *n*-mer pair data set was derived. Residues that participate in regular secondary structure were grouped into one of three categories: alpha, parallel beta, or anti-parallel beta. These are collected in a triangle plot (Fig. 5) that shows the backbone conformational preferences of each amino acid. The triangle plot presentation leads to several observations. Not sur-

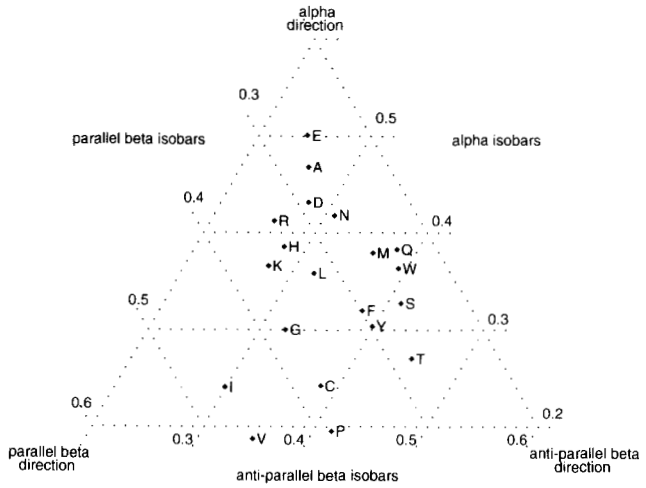


Fig. 5. Normalized amino acid attraction to secondary structure type. A triangle plot allows the easy plotting of three coordinates on a plane. In this figure, the coordinates are the normalized fractional population of each amino acid in α-helices, parallel β-strands, and anti-parallel β-strands. For example, isoleucine appearance in pieces of regular secondary structure is described by the triple [0.24, 0.46, 0.30] and is appropriately plotted in the triangle. The figure shows only the central region of a triangle plot. A full version of this triangle plot would have three types of secondary structure units (α-helices, parallel β-strands, and anti-parallel β-strands) represented at the vertices. This plot presents the normalized data points for each amino acid and isobars for each of the three implicit vertices.

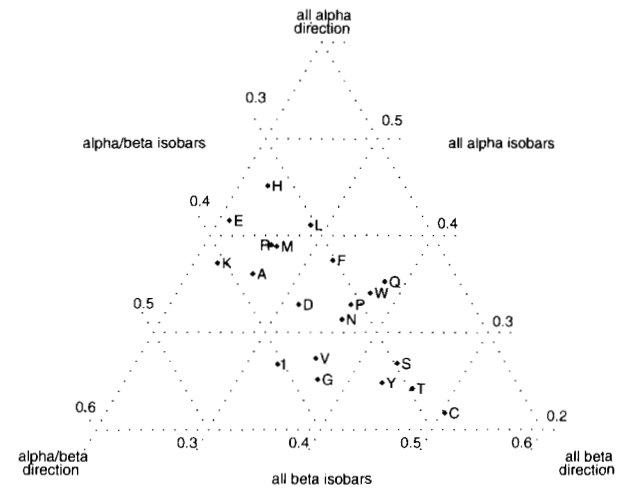


Fig. 6. Normalized amino acid attraction to tertiary structure content class. This is another triangle subplot as described in Figure 5. Three secondary structure content classes—all alpha, all beta, and alpha/beta—would be represented at the vertices of a full triangle. The data points for the 20 amino acids have been normalized so that each of the three secondary structure content classes occurs in equal numbers.

prisingly, proline lies most distant from alpha structure, whereas glutamate tends to gravitate toward alpha structure. A similar relationship can be seen for parallel beta structure: isoleucine and valine favor this structure, and glutamine and tryptophan avoid it. Threonine is preferentially seen in anti-parallel beta structure. Although isoleucine and valine are most commonly found in parallel beta structure, leucine is found in the center of the triangle, equally compatible with all three structure types. Care must be taken in interpreting these preferences for statistically rare amino acids. For example, tryptophan is the least commonly occurring amino acid in protein sequences. This sample contains only 288 tryptophan residues, well below the average of 1,094 seen for the other 19 amino acids. Similar care should be used in analyzing the structural biases of histidine, cysteine, and methionine.

The relative likelihood that each amino acid will appear in a protein within an α/α, β/β, or α/β protein is plotted in the triangle format in Figure 6. A comparison of Figures 5 and 6 suggests that enrichment in some amino acids may be the hallmark of a protein within a particular structural class. For example, lysine is uncommon in β/β proteins. On the other hand, threonine is common in anti-parallel β-strands and in β/β proteins. When the amino acid composition of the n-mer pair sequences is analyzed, leucine and alanine are seen to be the most common components (see Fig. 7). The relative paucity of cysteine, methionine, tryptophan, and tyrosine may relate to their infrequent occurrence in protein sequences. From an examination of Figure 6, it is clear that leucine and alanine are near the center of the triangle plot. This suggests that these amino acids are compatible with several folding motifs. Perhaps this explains their common occurrence within n-mer pairs. Presumably, the lack of proline residues in n-mer pairs reflects the power of the conformational restriction imposed by the pyrrolidine ring.

Discussion

The rapid expansion of the protein structural database has created an opportunity to reexamine a paradigm first posed by Kabsch and Sander (1984): Locally identical sequences are compatible with heterogeneous structures. Initially, this was clear for pentapeptides, but this work and the work of Wilson et al. (1985) extend the concept to hexapeptides. We expect that as the structural database grows, heptamers, octamers, and even longer identical sequences will be found with distinct structures. Zhong and Johnson (1992) synthesized peptides that would appear to favor α-helical structure based on a Chou–Fasman prediction but that actually are found in β-strand regions of proteins. Circular dichroism spectroscopy experiments show these peptides can be induced to fold into either α-helical or β-strand conformations by varying the solvent environment. Recent results from Shoham (pers. comm.) suggest that a 15-residue peptide from cholera-toxin adopts different structures in solution as determined by NMR, in the context of the protein by X-ray studies,

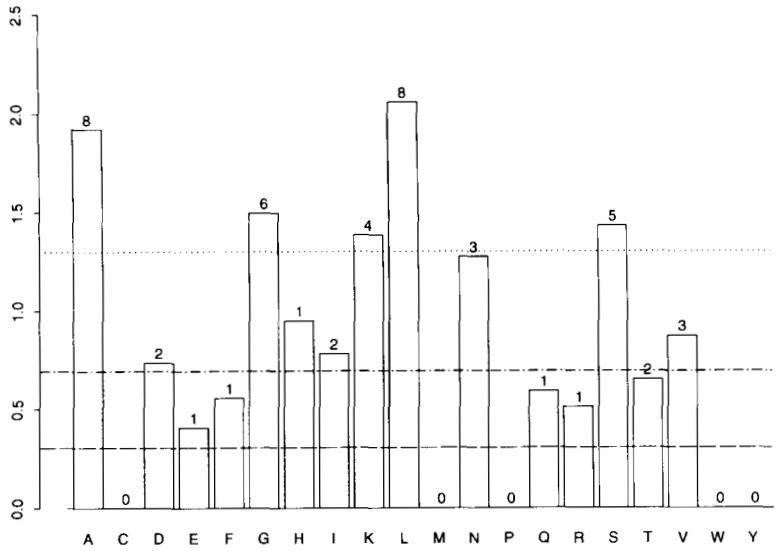


Fig. 7. Scaled frequencies of amino acids in different structure *n*-mer pairs. This bar graph shows normalized appearance frequency for each of the 20 amino acids in the identical sequences of the eight *n*-mer pairs in the subclass of disparately different for the set of different structure pairs. The normalized appearance frequency for each amino acid is calculated by counting the number of appearances of the given amino acid and dividing by the expected number of appearances (based on actual counts in the entire PDB) in a random sample equal to the size of the total number of residues in all different structure pairs. For example, histidine occurs nine times in disparately different structure pairs. Given a total of 48 residues in disparately different structure pairs (1/48 = 2.1%) and an expected appearance rate of 2.2%, histidine appears slightly less than would be expected. The number over each bar gives the actual appearance counts. The dash–dot line shows the mean normalized appearance frequency, while the dotted line above the mean line and the dashed line below give the 75 and 25 percentile normalized appearance frequencies, respectively.

and when bound to a monoclonal antibody. On the surface, this dissociation of sequence from structure would seem to create immense difficulties for fragment-based aspects of homology model building algorithms and for de novo structure prediction algorithms that exploit a buildup procedure (Vasquez & Scheraga, 1988) or secondary structure prediction methods. We suggest that the folding class of a protein provides a global constraint on the local conformation of these conformationally ambivalent peptides that favors a particular backbone geometry. Perhaps this explains why structurally distinct *n*-mers tend to be found in proteins from distinct folding classes. Alternatively, it may be possible to exploit a family of aligned sequences to avoid making a prediction for a conformationally ambiguous region of one particular sequence (Benner & Gerloff, 1991; Russell et al., 1992; Rost et al., 1993).

There are two specific implications of this observation. First, secondary structure prediction algorithms are unlikely to improve unless class or motif dependent information is incorporated during their development. Until recently, there have not been enough examples in each class to constitute a statistically meaningful database. This has begun to change—algorithms directed at predicting the secondary structure locations in α/α proteins are more accurate than their unbiased counterparts (Kneller et al., 1990; Presnell et al., 1992). In a prediction of the structure of IL-4 (Curtis et al., 1991) in advance of the NMR structure determination (Smith et al., 1992), the helical regions of this α/α protein were predicted with 90% fidelity. Second, peptide fragment databases of the types constructed by Unger et al. (1989) or Rooman et al. (1992) could be improved. Presumably, fragment databases drawn from proteins within a given structural class will contain higher quality information for protein homology model building studies. This has been borne out in a preliminary fashion by the work of Chothia and colleagues (1989) on immunoglobulin structures.

It is somewhat puzzling that the *n*-mer pairs that adopted distinct structures occupied remarkably similar environments within the proteins from which they were extracted. Certainly, from the standpoint of solvent accessibility, these sequences achieved a comparable degree of burial in each protein. This is consistent with the importance of the hydrophobic effect in organizing protein tertiary structure. Even the conformationally sensitive 3D–1D profiles of Eisenberg and coworkers (Bowie et al., 1991) were unable to shed light on the environmental preferences of the conformationally ambivalent sequences. We conclude that proteins are sufficiently plastic that suitable environments can be created for specific peptide sequences even when they adopt distinct conformations. Perhaps the mechanism by which this accommodation is made will be clear from the study of additional examples of this phenomenon.

Methods

A data set was constructed from the Brookhaven Protein Data Bank (version of July 15, 1990) by selecting all unique protein chains. Where two chains have identical sequences (e.g., the A and C chains of hemoglobin or two versions of flavodoxin), only one structure was chosen. This choice was based on the resolution of the crystal structures and the availability of non-α-carbon coordinates. The initial data set consisted of 366 chains from 316 proteins.

To speed the search of the PDB sequences, a hash table algorithm was employed. Briefly, a hash system is composed of a table (a set of pigeonholes) and a hash function that assigns each object (in this case 5-mers) to a slot (see Sedgewick, 1983). A toy example of a hash system for 5-mers could include a table with 20 slots and a

hash function that assigned 5-mers to a slot based on the first residue of the 5-mer. Although different 5-mers might be assigned (hash) to the same slot, all identical 5-mers would necessarily hash to the same slot. Within each slot, the 5-mers were searched for 5-mer pairs. The 5-mer pairs were extended to longer n-mer pairs where possible, and initially all n-mer pairs where n was greater than 5 were considered.

It is not surprising that many long n-mer pairs can be found in pairs of homologous proteins (e.g., lactate dehydrogenases [2ldx, 5ldh] from different organisms). These n-mer pairs are not of interest for this study. After performing pairwise alignments (Smith & Smith, 1990), any n-mer pairs containing proteins that have at least 50% pairwise residue identity were excluded from further study.

The 50% pairwise residue identity cutoff removes intersequence similarities, but it fails to deal with interpair similarities. Redundant n-mer pairs occur when two or more sequentially similar proteins contain n-mers that pair with an n-mer in a different protein. For example, VDLLKN is found in human class I histocompatibility antigen (1hla), trp repressor (2wrp), and trp aporepressor (3wrp). The n-mer pair between 2wrp and 3wrp is eliminated because of the obvious sequence similarity, but two largely redundant n-mer pairs remain: {VDLLKN, 1hla, 2wrp} and {VDLLKN, 1hla, 3wrp}. One of these two n-mer pairs is assumed to contain redundant information and is removed from further consideration.

The 106 protein domains were each visually assigned to a tertiary structural class. These assignments are consistent with previous taxonomic studies (Sternberg & Thornton, 1978; Richardson, 1981). The pair of tertiary structural classes associated with each constituent protein serves as an attribute of each n-mer pair.

The n-mer pairs can be divided into four categories from a comparison of the 3D structures of the residues in the two proteins: *identical sequence, same structure*; *identical sequence, different structure*; *identical sequence, distinct structure version*; or *identical sequence, distinct loops*. These four categories are referred to as *local structure difference groupings*. The first consideration in the grouping process requires an assignment of regular secondary structure. Secondary structure assignments are made using the computer programs DSSP (Kabsch & Sander, 1983) for strand assignments and DEFINE (Richards & Kundrot, 1988) for helix assignments. The decision to use these two programs was based on our experience comparing algorithmic assignments with observations of structures on a graphics workstation. The use of α-carbon positions in DEFINE and hydrogen bonding patterns in DSSP accounts for different results from the two programs. For example, the lack of consideration of hydrogen bonds in DEFINE leads to some false positive β-strand assignments. Although we have generally found DEFINE to be superior for assigning helices (e.g., {NDSTVL, 1abp, 1hds-A}), there was one n-mer pair, {LKKSAD, 2ldx, 2lhb}, where the DSSP identified a break between two helices that DEFINE failed to find in a residue-based assignment (see Kinemage 2). N-mer pairs that have regular secondary structure in one member and a different secondary structure or aperiodic structure in the other member are clearly in the different structure group. On the other hand, similar secondary structure assignments for both members do not guarantee a same structure grouping. For example, two n-mers may contain β-structure, one in the N-terminal four residues and the other in the C-terminal four residues. In this case, the root mean square (RMS) deviations will be too large to consider the structures to be the same. When neither member of an n-mer pair contains regular secondary structure, the RMS deviation and a structural loop classification (Ring et al., 1992) provide tools for a quantitative comparison of the two structures.[2] Finally, assignments are checked by visual observations of the n-mer pair on a graphics terminal.[3] Figure 8 demonstrates that a 1.5-Å RMS deviation usefully separates structurally similar n-mers from their structurally distinct counterparts. The only exception to this dividing line is for the pair {CKSSQS, 1mcp, 1alc}. Here, a β-strand is found in an immunoglobulin fragment (1mcp), while the α-lactalbumin (1alc) subsegment forms a loop. The backbone structure of the region in 1alc is similar to a strand, but the local environment lacks a neighboring strand for hydrogen bonding (see Kinemage 1).

The Chou–Fasman predictions were performed on a computer with the CONFORM program. The entire sequence of a protein chain that contains the n-mer of interest was used as input to a CONFORM run. This means that residues upstream and downstream from the n-mer sequence could also contribute to the n-mer region's predicted secondary structure.

Supplementary material on Diskette Appendix

File Cohen.kin (KINEMAGE directory) contains four kinemages. The first two give a sense of some of the details involved in making judgments on questions of structural similarity. Certainly, RMS deviation is one measure of structural similarity. Kinemage 1 presents an n-mer pair with a low RMS deviation between the backbones of the two relevant protein segments, but different secondary structure assignments (strand versus turn). Kinemage 2 shows a problem in relying on automated secondary struc-

[2] The loop alphabet is composed of four characters (U, J, Z, L) based on the virtual dihedral angles formed by the α-carbons of the tetrapeptide. An n-residue n-mer is characterized by $n - 3$ overlapping tetrapeptides. For example, the conformation of a hexamer (the usual case in this study) is described by a three-letter word. Distinct loop conformations form different words.

[3] The n-mer pairs were viewed on a Silicon Graphics Iris using UCSF MIDAS software (Ferrin et al., 1988). The n α-carbons of both members of each n-mer pair were superimposed using a least-squares fit. The RMS error is based on this fit.

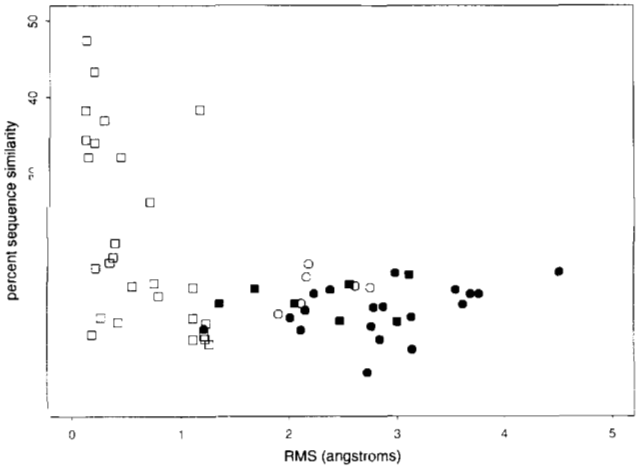


Fig. 8. Root mean square and sequence similarity. The RMS deviation between the proteins in each *n*-mer pair is plotted against pairwise aligned sequence similarity. The *n*-mer pairs are divided into four nonoverlapping groups based on the similarity between the local secondary and tertiary structure of the region composed of the *n*-mer residues in each protein. Unfilled squares represent same structure pairs. Open squares (□) represent same structure pairs. Filled squares (■) represent pairs with similar secondary structure but different tertiary structure. Open circles (○) represent pairs where neither region is in a helix or strand, but the two loops are not similar. Finally, filled circles (●) represent pairs with different local secondary (and generally tertiary) structures.

ture assignments. Here, one of the *n*-mer segments is given two different secondary structure assignments by two assignment programs. The final two kinemages offer representative results of this study. Kinemage 3 shows a pair of similar local structures embedded in proteins that belong to different tertiary structural classes. Kinemage 4 portrays an example of an *n*-mer in a helical conformation in one protein and a strand conformation in another.

An ASCII tab-separated listing of the 59 *n*-mer pairs appears in the SUPLEMNT directory (file Cohen.SUP). The first record has column headings.

Acknowledgments

We acknowledge with thanks many helpful discussions with Christine Ring and the support of the NIH (GM39900) and the National Library of Medicine (F37-LM00010). David Eisenberg's group at UCLA kindly supplied us with the software for the 3D profile method.

References

Benner, S.A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: A prediction of the structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul. 31*, 121–181.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Bowie, J., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*, 164–169.

Chan, H.S. & Dill, K.A. (1990). Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA 87*, 6388–6392.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature (Lond.) 248*, 338–339.

Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S., Air, G., Sheriff, S., Padlan, E., Davies, D., & Tulip, W. (1989). Conformations of immunoglobulin hypervariable regions. *Nature (Lond.) 342*, 877–883.

Chou, P.Y. & Fasman, G.D. (1978). Empirical predictions of protein conformation. *Annu. Rev. Biochem. 47*, 251–276.

Curtis, B.M., Presnell, S.R., Srinivasan, S., Sassenfeld, H., Klinke, R., Jeffery, E., Cosman, D., March, C.J., & Cohen, F.E. (1991). Prediction of the three-dimensional structure of human interleukin-4. *Proteins Struct. Funct. Genet. 11*, 111–119.

Ferrin, T., Huang, C., Jarvis, L., & Langridge, R. (1988). The MIDAS display system. *J. Mol. Graphics 6*, 13–37.

Hughson, F.M., Wright, P.E., & Baldwin, R.L. (1990). Structural characterization of a partly folded apomyoglobin intermediate. *Science 249*, 1544–1548.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition and geometrical features. *Biopolymers 22*, 2577–2637.

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. USA 81*, 1075–1078.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Protein Chem. 14*, 1–63.

Klein, P. & DeLisi, C. (1986). Prediction of protein structural class from the amino acid sequence. *Biopolymers 25*, 1659–1672.

Kneller, D., Cohen, F.E., & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol. 214*, 171–182.

Lee, B. & Richards, F.M. (1971). The interpretation of proteins structures: Estimation of static accessibility. *J. Mol. Biol. 55*, 379–400.

Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature (Lond.) 261*, 552–557.

Luthy, R., Bowie, J.U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature (Lond.) 356*, 83–85.

Lyu, P., Sherman, J., Chen, A., & Kallenbach, N. (1991). Alpha-helix stabilization by natural and unnatural amino acids with alkyl side chains. *Proc. Natl. Acad. Sci. USA 88*, 5317–5320.

Marqusee, S. & Baldwin, R.L. (1990). Alpha-helix formation by short peptides in water. In *Protein Folding* (Gierasch, L.M. & King, J., Eds.), pp. 85–94. AAAS, Washington, D.C.

Molecular Biology Information Resource. (1989). *Eugene User's Manual—Release 3.2.* Baylor College of Medicine, Houston, Texas.

Muskal, S.M. & Kim, S.H. (1992). Predicting protein secondary structure content—A tandem neural network approach. *J. Mol. Biol. 225*, 713–727.

Presnell, S.R., Cohen, B.I., & Cohen, F.E. (1992). A segment based approach to secondary structure prediction. *Biochemistry 31*, 983–993.

Richards, F.M. & Kundrot, C.E. (1988). Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins Struct. Funct. Genet. 3*, 71–84.

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem. 34*, 167–339.

Ring, C.S., Kneller, D.G., Langridge, R., & Cohen, F.E. (1992). Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol. 224*, 685–699.

Roder, H., Eloeve, G.A., & Englander, S.W. (1988). Structural characterization of folding intermediates in cytochrome *c* by H-exchange labelling and proton NMR. *Nature (Lond.) 335*, 700–704.

Rooman, M.J., Kocher J.P., & Wodak, S.J. (1992). Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry 31*, 10226–10238.

Rooman, M.J. & Wodak, S.J. (1988). Identification of predictive sequence motifs limited by protein structure data base size. *Nature (Lond.) 335*, 45–49.

Rost, B., Schneider, R., & Sander, C. (1993). Progress in protein structure prediction? *Trends Biochem. Sci. 18*, 120–123.

Russell, R.B., Breed, J., & Barton, G.J. (1992). Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Lett. 304*, 15–20.

Sedgewick, R. (1983). *Algorithms*. Addison-Wesley, Reading, Massachusetts.

Sheridan, R.P., Dixon, J.S., Venkataghavan, R., Kuntz, I.D., & Scott, K.P. (1985). Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. *Biopolymers 24*, 1995–2023.

Smith, L., Redfield, C., Boyd, J., Lawrence, G., Edwards, R., Smith, R., & Dobson, C. (1992). Human interleukin 4. The solution structure of a four-helix bundle protein. *J. Mol. Biol. 224*, 899–904.

Smith, R.F. & Smith, T.F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA 87*, 118–122.

Sternberg, M.J.E. & Thornton, J.M. (1978). Prediction of protein structure from amino acid sequence. *Nature (Lond.) 271*, 15–20.

Unger, R., Harel, D., Wherland, S., & Sussman, J.L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins Struct. Funct. Genet. 6*, 355–373.

Vasquez, M. & Scheraga, H. (1988). Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data. *J. Biomol. Struct. Dyn. 5*, 705–755.

Wilson, I.A., Haft, D.H., Getzoff, E.D., Tainer, J.A., Lerner, R.A., & Brenner, S. (1985). Identical short peptide sequences in unrelated protein can have different conformations: A testing ground for theories of immune recognition. *Proc. Natl. Acad. Sci. USA. 82*, 5255–5259.

Wright, P.E., Dyson, H.J., Waltho, J.P., & Lerner, R.A. (1990). Folding of peptide fragments of proteins in water solution. In *Protein Folding* (Gierasch, L.M. & King, J., Eds.), pp. 95–102. AAAS, Washington, D.C.

Zhong, L. & Johnson, W.C., Jr. (1992). Environment affects amino acid preference for secondary structure. *Proc. Natl. Acad. Sci. USA 89*, 4462–4465.

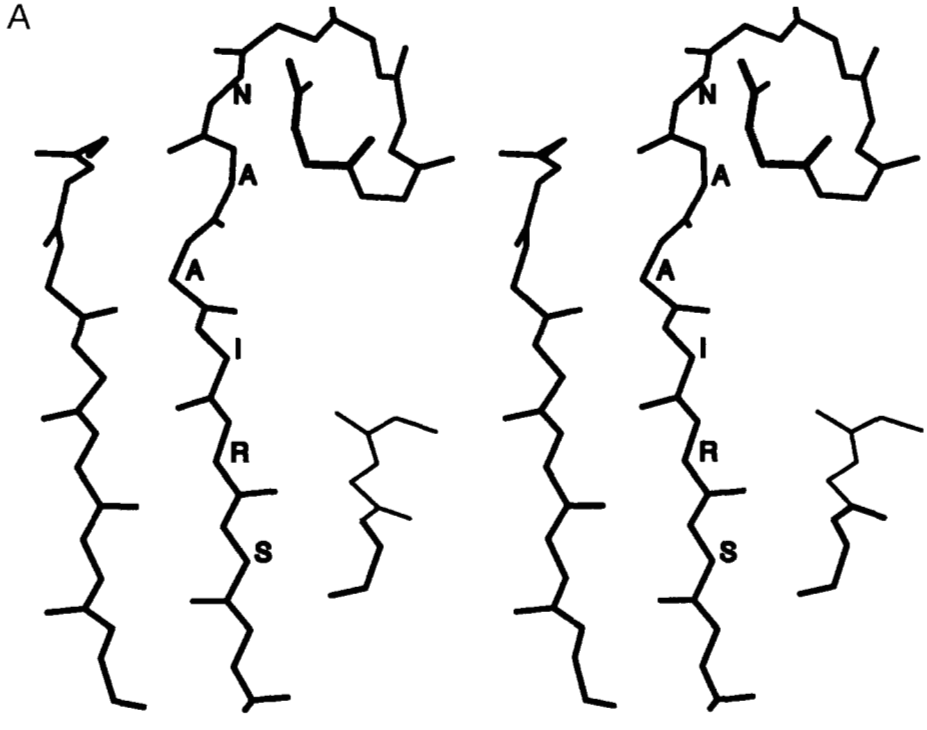
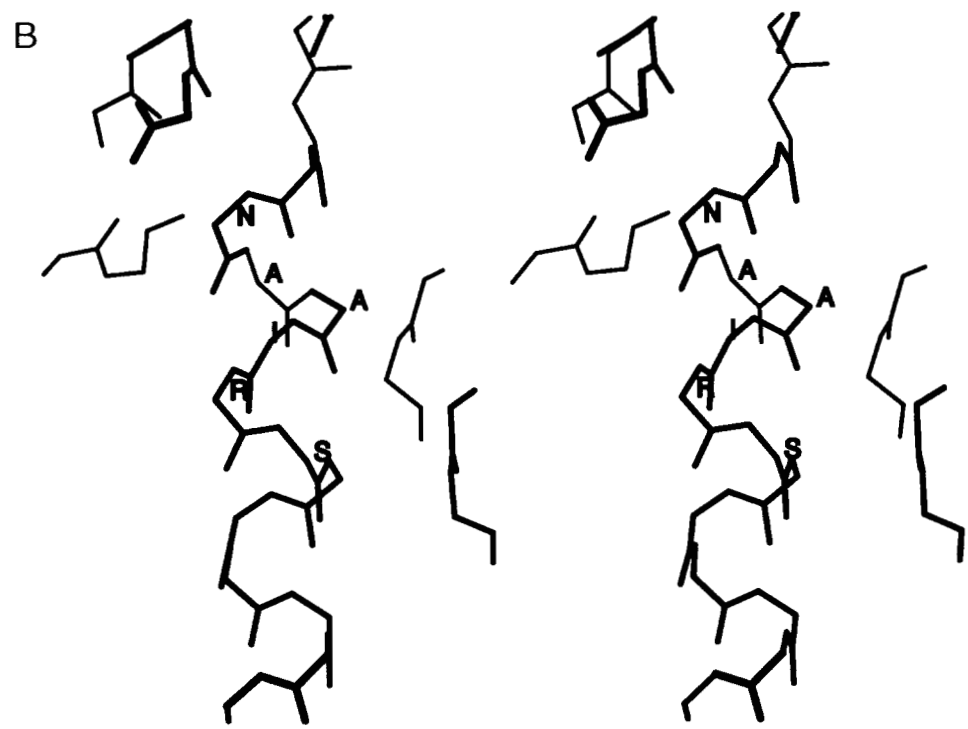


Figure added in proof (*see also facing page*). An example of a helix and strand with the same hexapeptide sequence. The sequence NAAIRS is found in both (A) thermolysin (2tmn) and (B) phosphofructokinase (3pfk). Stereo pairs of the relevant fragment of the protein together with the surrounding environment are shown. **A:** In the β/β domain of thermolysin, the NAAIRS subsequence begins a β-strand.



B: In the α/β-class protein, phosphofructokinase, NAAIRS can be seen as part of an α-helix.