Helix Signals in Proteins

LEONARD G. PRESTA AND GEORGE D. ROSE*

The α helix, first proposed by Pauling and co-workers, is a hallmark of protein structure, and much effort has been directed toward understanding which sequences can form helices. The helix hypothesis, introduced here, provides a tentative answer to this question. The hypothesis states that a necessary condition for helix formation is the presence of residues flanking the helix termini whose side chains can form hydrogen bonds with the initial four-helix >N–H groups and final four-helix >C–O groups; these eight groups would otherwise lack intrahelical partners. This simple hypothesis implies the existence of a stereochemical code in which certain sequences have the hydrogen-bonding capacity to function as helix boundaries and thereby enable the helix to form autonomously. The three-dimensional structure of a protein is a consequence of the genetic code, but the rules relating sequence to structure are still unknown. The ensuing analysis supports the idea that a stereochemical code for the α helix resides in its boundary residues.

THE α HELIX WAS FIRST PROPOSED AS A MODEL STRUCTURE by Pauling *et al.* (*1*). Subsequent experimental support (*2*) has made the helix a familiar landmark in proteins. The key feature of the Pauling-Corey-Branson helical model is a pattern of iterated backbone hydrogen bonding between each >N−H donor and the >C=O acceptor located four residues previously. The resultant structure satisfies the hydrogen-bonding requirements of consecutive main-chain polar groups with a hydrogen-bond geometry that is nearly optimal.

Helices are classified as repetitive secondary structure because their backbone dihedral angles, ϕ and ψ, have repeating values near the canonical value of (−60°, −40°) (*3*). When the dihedral angles of a chain segment assume helical values, the backbone polar groups are automatically positioned to form hydrogen bonds with intrasegment partners. The situation is unlike that of β sheet, the other repetitive secondary structure, where backbone hydrogen bonds in each β strand are satisfied by extrasegment partners from an adjacent β strand that may be distant in sequence.

In globular proteins of known structure, approximately one-quarter of all residues are found in helices (*4*). The frequent occurrence of helices, and the fact that their hydrogen bonds can be localized to intrasegment partners, suggest that α helices may function as autonomous folding units in proteins. This suggestion is strengthened by recent experiments that demonstrate the stability of isolated protein helices in water (*5*, *6*).

We now show that the location of helices in water-soluble proteins is dependent on local sequence information alone. This finding is a result of the observation that the Pauling-Corey-Branson model accounts for only about half of the backbone hydrogen bonds in actual protein helices. In particular, the average protein helix, which is 12 residues in length (*7*), contains eight intrahelical >N–H ··· O=C< bonds, but >N–H donors in the first four residues and >C=O acceptors in the last four residues lack intrahelical partners (*8*) (Fig. 1). We hypothesize that a necessary condition for helix formation is the presence of residues flanking the helix termini whose side chains can supply hydrogen-bond partners for unpaired main-chain >N−H and >C=O groups. These boundary residues would then function as a stereochemical code for helix formation.

Unlike theories derived from statistical correlations, the helix hypothesis is based on simple physical chemistry and provides a mechanism for many well-known phenomena. For example, the tendency for helices to be situated at the molecular surface (*9*) and often to contain amphipathic sequences (*10*) is a consequence of

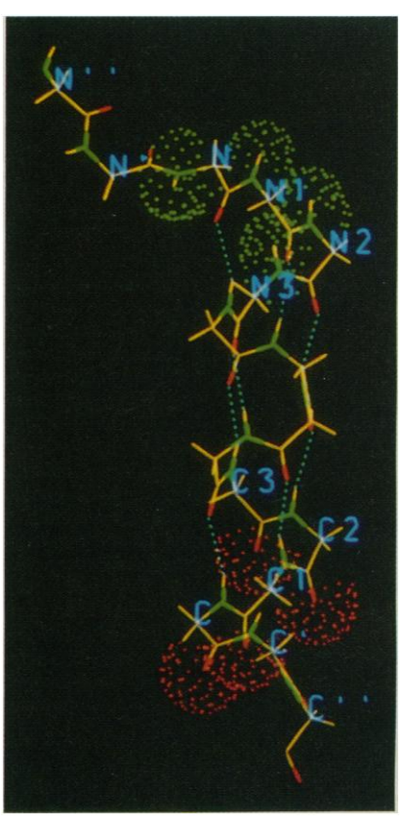


Fig. 1. A representative α helix, 12 residues in length, flanked by adjacent turns. Backbone nitrogen atoms are shown in green, backbone oxygen atoms in red. The eight intrahelical N–H ···O=C hydrogen bonds are indicated by broken lines. N1, N2, N3 are the initial three residues of the helix proper while C3, C2, C1 are the final three residues. Residues N and C have nonhelical dihedral angles but contribute one additional hydrogen bond to the helix. Residues N″, N′, N and C, C′, C″ are classified with the preceding and succeeding turns, respectively. Hydrogens in the initial four >N–H groups are indicated by stippled green surface; oxygens in the final four >C=O groups are indicated by stippled red surface. These eight groups cannot be satisfied by intrahelical main-chain partners.

Leonard Presta and George Rose are members of the Department of Biological Chemistry, Hershey Medical Center, Pennsylvania State University, Hershey, PA 17033.

*To whom correspondence should be addressed.

requiring polar residues at the termini and the concomitant desirability of having some apolar residues between the termini to promote hydrophobic association with rest of the protein. The statistical preference for acidic residues at the NH_2-termini of helices and basic residues at the COOH-termini (*11*) results from the Asp and Glu side chains being able to serve as hydrogen-bond acceptors while Lys and Arg can serve as donors; these hydrogen bonds would augment contributions arising from ionic interactions with the helix macrodipole (*6*). The puzzling examples of identical pentapeptides that are helical in one segment but not in another (*12*) can be reconciled if the segment termini are taken into account.

To test the helix hypothesis in proteins of known structure, we analyzed each sequence for potential helix boundaries and compared the results to the location of observed helices. Although the hypothesis is fundamentally simple, the analysis is complex. Briefly, a complete library of side-chain to main-chain hydrogen-bonding possibilities was compiled. Side chains from each of the 13 polar residues, together with two additional variants of His, were appended to each α carbon of a polyglycyl helix and a representative ensemble of adjacent turns. The conformations of these side chains were then uniformly sampled. Whenever a hydrogen bond could be formed, the residue, together with its conformation and position, was added to the library. The library was then used to evaluate actual protein sequences for sites at which residue side chains could satisfy the four terminal >N–H groups or >C=O groups of a helix. A window of only six consecutive residues turned out to be sufficient to identify such sites. Nevertheless, the enumeration of backbone to side-chain hydrogen bonds in actual proteins is computationally intensive because, for each six-residue sequence, every permutation of allowed conformations from the library must be assessed, as described below.

Analysis of x-ray–elucidated proteins. Our analysis required prior identification of helices and adjacent turns in proteins of known structure. The proteins (Table 1) included 26 high-resolution x-ray structures (resolution ≤2.0 Å; *R* factor ≤20 percent) from the Brookhaven protein database (*13*).

To identify helices, we determined all intramolecular main-chain to main-chain hydrogen bonds for each protein using criteria enumerated in Fig. 2 (*14, 15*). Backbone segments with (i, $i + 4$) or (i, $i + 3$) hydrogen bonds were then inspected for the presence of an α or 3_{10} helix. Helices were terminated at the final residue in which backbone >NH or >C=O groups participate in an (i, $i + 4$) or (i, $i + 3$) hydrogen bond while maintaining dihedral angles with helical values ($\phi \cong -60°$; $\psi \cong -40°$). This strict definition may differ slightly from assignments listed in the header records of the protein database (*13*) or those given by Kabsch and Sander (*4*) because the respective >C=O or >N–H groups in the residue immediately preceding or following a helix form one additional intrahelical hydrogen bond with dihedral angles having nonhelical values. Further ambiguity in the precise location of helix boundaries can be occasioned by adjoining type I and type III reverse turns that have dihedral angles near helical values.

Using the foregoing hydrogen-bond criteria and boundary conditions, we found all helical residues in a database of proteins. The average dihedral angles (mean ± SD) are $\phi = -63.8° \pm 6.6°$, $\psi = -41.0° \pm 7.2°$ for 1062 residues.

Helices and their flanking residues are labeled as follows:

N″–N′–N–N1–N2–N3– . . . –C3–C2–C1–C–C′–C″

where N1–N2–N3– . . . –C3–C2–C1 participate in the helix backbone hydrogen-bonding network and have helical backbone dihedral angles. Residues N and C participate in the hydrogen-bonding

Table 1. Proteins used. All are x-ray structures with resolution ≤2.0 Å and crystallographic R factor ≤20 percent.

Code*	Protein name	Helices†
351C	Cytochrome C551	3–9, 27–33, 40–50, 68–79
2ACT	Actinidin	25–42, 51–56, 70–79, 100–103, 121–128, 142–145
1AZA	Azurin (molecule B)	56–64
1BP2	Phospholipase A2	2–12, 18–21, 40–55, 59–63, 90–106
3C2C	Cytochrome C2	{4–10, [11–16]}, 50–58, 64–70, 74–82, 98–108
2CAB	Carbonic anhydrase B	131–134, {[155–157], 158–162, [163–166]}, 220–226
2CDV	Cytochrome C3	65–70, 79–87, 91–98
5CPA	Carboxypeptidase A	15–28, 74–89, 94–100, 113–121, 174–186, 216–230, 254–260, 286–305
1CRN	Crambin	7–16, 23–29
4DFR	Dihydrofolate reductase (molecule B)	25–35, 44–50, 78–83, 97–103
1ECD	Hemoglobin III, *Chironomos thummi thummi*	{3–13, [13–15]}, 20–30, 46–49, 53–71, 77–87, 94–111, 118–132
4FXN	Flavodoxin, semiquinone form	11–25, 66–72, 94–104, 125–135
1GP1	Glutathione peroxidase (molecule B)	48–62, 88–94, 120–128, 185–192
1HMQ	Hemerythrin (molecule D)	19–37, 41–64, 70–84, 91–102
1INS	Insulin, porcine (molecules C and D)	A2–A8, [A13–A18], B9–B19
2LHB	Hemoglobin V, lamprey	{13–24, 24–28, 30–44, [45–51]}, 61–65, 68–86, 91–107, 116–127, 132–145
1LZ1	Lysozyme, human	5–14, 25–35, 90–99, 110–114
1MBO	Myoglobin, sperm whale, oxidized	4–17, {21–35, [37–42]}, 52–56, 59–76, 83–95, 101–118, 125–148
2OVO	Turkey ovomucoid inhibitor	34–43
1PPT	Pancreatic hormone, avian	14–31
5PTI	Pancreatic trypsin inhibitor	48–55
5RSA	Ribonuclease A, bovine	4–12, 25–32, {51–55, [56–57]}
2SGA	*Streptomyces griseus* Protease A	{56–59, [62–63]}, 232–236
1SN3	Scorpion neurotoxin	{23–29, [30]}
3TLN	Thermolysin	68–87, 137–150, 160–179, {[234], 235–246}, 260–273, 281–296, 301–312
1TPP	Trypsin, bovine	165–171, 235–243

*Brookhaven protein database four-character name (*13*). †[] Denote segments of 3_{10} helix; {} denote segments considered as 1 helical unit.

network but have nonhelical dihedral angles. Residues N″, N′, C′, and C″ neither participate in helix backbone hydrogen bonding nor have helical dihedral angles, and they are classified with the preceding or succeeding turns, respectively (Fig. 1).

To construct a hydrogen-bond library, we required a representative ensemble of backbone conformations at the NH_2- and COOH-termini of helices. Six-residue segments consisting of residues N″–N′–N–N1–N2–N3 were tabulated for each helix NH_2-terminus in the 26 proteins. Six-residue COOH-terminal segments consisting of C3–C2–C1–C–C′–C″ were also tabulated. These data were then partitioned into classes. Thirteen classes were defined for turns at the NH_2-terminus (representing 60 percent of the observed distribution) (see Table 2), and six classes were defined at the COOH-terminus (representing 50 percent of the observed distribution) (see Table 3). Expansion to seven-residue segments that included the next residue before N″ or after C″ was attempted, but the variation was too large to be useful.

Table 2. Backbone conformations of residues N″, N′, and N in observed helices. The magnitude of the respective angle is given in degrees plus or minus the standard deviation (σ) of the respective angle in degrees. The number of residues in the sample is shown as *n*. No standard deviation was computed when $n = 1$.

C*	B†	N″ (deg ± σ)	*n*	N′ (deg ± σ)	*n*	N (deg ± σ)	*n*
A	φ	−60 ± 7	8	−91 ± 18	18	−100 ± 24	18
	ψ	−23 ± 12	8	−15 ± 18	18	139 ± 19	18
B	φ	−60 ± 6	5	−90 ± 17	5	−75 ± 9	5
	ψ	−27 ± 8	5	−6 ± 18	5	134 ± 17	5
C	φ	−108 ± 22	6	−91 ± 18	18	−100 ± 24	18
	ψ	141 ± 22	6	−15 ± 18	18	139 ± 19	18
D	φ	−65 ± 14	9	−71 ± 10	13	−94 ± 15	13
	ψ	−26 ± 13	9	−21 ± 16	13	−2 ± 16	13
E	φ	−90 ± 47	4	−71 ± 10	13	−94 ± 15	13
	ψ	152 ± 16	4	−21 ± 16	13	−2 ± 16	13
F	φ	−84 ± 22	8	−66 ± 11	16	−78 ± 17	16
	ψ	136 ± 25	8	142 ± 10	16	151 ± 20	16
G	φ	−93 ± 14	4	−66 ± 11	16	−78 ± 17	16
	ψ	5 ± 14	4	142 ± 10	16	151 ± 20	16
H	φ	−94 ± 21	2	−71 ± 11	3	−79 ± 17	3
	ψ	143 ± 25	2	119 ± 8	3	−20 ± 28	3
I	φ	−97	1	−71 ± 11	3	−79 ± 17	3
	ψ	12	1	119 ± 8	3	−20 ± 28	3
J	φ	−106 ± 25	5	−138 ± 15	8	−91 ± 12	8
	ψ	150 ± 16	5	164 ± 12	8	160 ± 20	8
K	φ	−68	1	−89 ± 6	4	−112 ± 8	4
	ψ	136	1	62 ± 10	4	17 ± 15	4
L	φ	−67 ± 7	4	−77 ± 8	5	−141 ± 24	5
	ψ	−17 ± 12	4	−17 ± 14	5	72 ± 12	5
M	φ	−92 ± 30	3	−73 ± 8	3	−159 ± 2	4
	ψ	141 ± 11	3	−24 ± 5	3	176 ± 9	4

*C lists the conformational class. †B is the backbone dihedral angle, φ or ψ.

Table 3. Backbone conformations of C, C′, and C″ in observed helices. In conformation N, residue C′ is usually a Gly. When ψ for residue C″ is a dash (−), several observed classes have been collapsed into one effective class, since the value of this angle does not affect the conformations of residues in the window [C3–C″]. Conformation R was used only when residue C′ is a Pro. In this case, the backbone conformation of residue C1 was $\phi = -75°$, $\psi = -28°$. In conformation S, residue C is usually a Gly.

C	B	C (deg ± σ)	*n*	C′ (deg ± σ)	*n*	C″ (deg ± σ)	*n*
N	φ	−92 ± 17	17	71 ± 14	17	−88 ± 26	13
	ψ	−2 ± 11	17	27 ± 14	17	—	
O	φ	−74 ± 16	32	−83 ± 19	32	−83 ± 14	18
	ψ	−19 ± 17	32	−13 ± 19	32	—	
P	φ	−74 ± 16	32	−83 ± 19	32	54 ± 6	3
	ψ	−19 ± 17	32	−13 ± 19	32	55 ± 10	3
Q	φ	−93 ± 13	10	−73 ± 13	10	−83 ± 22	6
	ψ	−6 ± 15	10	134 ± 15	10	−14 ± 25	6
R	φ	−137 ± 8	7	−60 ± 11	6	−74 ± 12	5
	ψ	69 ± 18	7	−23 ± 12	6	−19 ± 12	5
S	φ	83 ± 9	5	−119 ± 39	4	−91 ± 21	4
	ψ	14 ± 11	5	136 ± 22	4	137 ± 13	4

Construction of a side-chain to main-chain hydrogen-bonding library. A library of potential hydrogen bonds between side-chain and main-chain groups was compiled for polar residues within helices and flanking turns. The strategy for each residue type was to generate a set of 19 polyglycyl paradigms, including hydrogen atoms, consisting of an eight-residue idealized helix joined to each class of flanking turn (13 at the NH_2-terminus; 6 at the COOH-terminus). All polar side chains, each in turn, were appended separately to each α-carbon position of all paradigms, the side chain was allowed to rotate, and conformations forming hydrogen bonds to the main chain were added to the library.

Amino acid geometries were extracted from the Empirical Conformational Energy Program for Peptides (ECEPP) (*16*), but were modified to assign identical bond lengths and angles to all backbone structures except proline. Backbone dihedral angles for the idealized helix were set at ($\phi = -63.8°$; $\psi = -41.0°$), the mean value observed in helices of x-ray elucidated structures (Table 1). Backbone dihedral angles assigned to flanking turns are those listed in Tables 2 and 3.

All 13 polar amino acid residues were assessed: Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, Met, Ser, Thr, Trp, and Tyr. Three variants, of His were used: (i) neutral with Nε protonated, (ii) neutral with Nδ protonated, and (iii) (+1)-charged with both Nε and Nδ protonated. Although the hydrogen-bonding capabilities of Cys and Met are uncertain (*14, 17*), both were included for completeness.

Side-chain conformations were sampled uniformly. For practical reasons, the sampling interval increased as side-chain length increased. In particular, side-chain dihedral angles were sampled at rotation increments of 5 degrees for (Ser, Thr, Cys), 10 degrees for (Asp, Asn, His), 15 degrees for (Glu, Gln, Met), and 20 degrees for (Arg, Lys, Tyr, Trp).

Our hydrogen-bonding criteria (*14, 15*) are enumerated in Fig. 3. Conformations were rejected if at least one pair of atoms had an interatomic distance less than 80 percent of the Ramachandran "extreme" limit (*18*), regardless of hydrogen-bond presence, or if the number of contacts falling between the "extreme" limits and their 80 percent cutoff values exceeded the number of side-chain dihedral angles for the specific amino acid residues. Hydrogen-bond geometry and close-contact constraints were relaxed slightly from accepted values (*15, 18*) in order to compensate for the rigid backbone geometry and fixed rotation increments of side chains. The coefficient of 80 percent is derived from the following useful "rule of thumb." For two atoms with an interatomic distance less than their "extreme" limit, rotation about the subtending dihedral angles by 10 degrees can move each atom by approximately 0.25 Å (*19*), thereby increasing the overall interatomic distance by 0.50 Å (approximately 20 percent of the "extreme" limit). Rotation about distal dihedral angles can cause larger movement.

The resultant library contains the allowed hydrogen-bonding conformations for each polar residue at every side-chain position in all 19 paradigms. In practice, two amino acids, Trp and Tyr, cannot form hydrogen bonds at either end of the helix, and they were eliminated.

Proline was treated as a special case. Using the 19 polyglycyl paradigms, we modeled both *exo*- and *endo*-Pro at every position

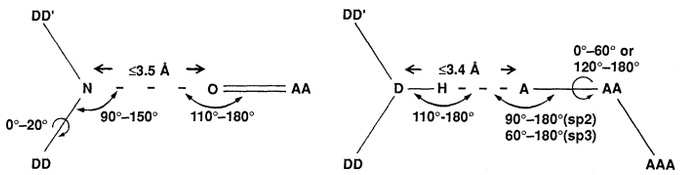


Fig. 2 (**left**). Hydrogen-bonding criteria for x-ray elucidated proteins. AA is the acceptor antecedent atom; DD and DD′ are the donor antecedent atoms. Torsion angle [N–DD–DD′–O] = 0° to 20°; this angle was used to measure the degree to which the oxygen acceptor is out of plane of the *sp*2 nitrogen donor. **Fig. 3** (**right**). Hydrogen-bonding criteria for hydrogen-bond library. D is the hydrogen-bond donor, DD and DD′ are the donor antecedent atoms; A is the acceptor, AA is the acceptor antecedent atom, AAA is the acceptor penultimate antecedent atom. Torsion angle [H . . . A–AA–AAA] = 0° to 60° or 120° to 180°.

having a compatible ϕ value (that is, $\phi \cong -75°$) (*16*). Structures with close contacts were considered sterically forbidden, but an exception was made in positions N″ or C″ where rotation about the Pro ψ dihedral angle can relieve a close-contact involving the Pro without affecting remaining positions.

Searching protein sequences for potential helix boundaries. We tested the helix hypothesis by searching the amino acid sequences of x-ray elucidated proteins for potential helix boundaries and comparing the results against the location of observed helices. The search procedure identifies two types of helix boundaries: NH_2-terminal bounds (NTB's) and COOH-terminal bounds (CTB's). Such boundaries are identified by moving a six-residue window along the protein sequence from its NH_2- to COOH-terminus. When being searched for NTB's, the window contains positions N″–N′–N–N1–N2–N3, and the side chains within the window must provide hydrogen-bond acceptors for at least three of the four main-chain >N–H donors of residues N–N1–N2–N3. When searching for CTB's, the window contains positions C3–C2–C1–C–C′–C″, and the side chains within the window must provide hydrogen-bond donors for at least three of the four main-chain >C=O acceptors of residues C3–C2–C1–C.

Side chains within each window position were appended to each of the six-residue backbone paradigms and coordinates were generated, including hydrogen atoms. All possible hydrogen-bonding arrangements were identified with the use of the hydrogen-bonding library. As the window is advanced, every residue in the sequence is positioned in turn at each locus in all paradigms. This search procedure is not only sequence-dependent but also structure-dependent because residues are assessed in an explicit backbone conformation that depends upon their position in the current window and the paradigm under consideration.

It is necessary to distinguish between hydrogen-bond combinations and conformations. A combination is defined as a distinct pattern of side-chain to main-chain hydrogen bonding for the six residues within the window. In effect, a combination is a hydrogen-bond "wiring diagram." A given six-residue segment could have many possible hydrogen-bond wiring diagrams or none at all. Each combination can assume multiple conformations, all of which preserve its wiring diagram. In effect, a conformation moves the "wires" but not their points of attachment. As the number of combinations, each comprised of an ensemble of conformations, becomes larger, the reduction in chain entropy needed to form the required hydrogen bonds becomes smaller.

The hydrogen-bonding library was used to retrieve all possible conformations with side-chain to main-chain hydrogen bonds for each combination. Since the library is derived from paradigms decorated with solitary residues, some permutations of these individually allowed hydrogen-bonding conformations may be mutually exclusive within the six-residue window. Consequently, all possible hydrogen-bonding conformations for each combination were tested to eliminate steric impossibilities. A conformation was rejected if at least one pair of atoms had an interatomic distance less than 80 percent of the Ramachandran "extreme" limit or if the number of contacts falling between the "extreme" limit and their 80 percent cutoff values exceeded the sum of the number of side-chain dihedral angles for the participating amino acids.

Exhaustive search of conformations is highly computer-intensive, and several approximatioins were made. Non-hydrogen-bonding residues within the window were approximated by Ala. Specifically, an amino acid residue was represented by Ala unless (i) it was Gly or Pro, (ii) it was Asn, Asp, Gln, Glu, His, Ser, or Thr at an NTB, or (iii) it was Asn, Arg, Gln, His, Lys, Ser, or Thr at a CTB. For example, the sequence Ser-Tyr-Pro-Gly-Asn-Val would be represented by Ser-Ala-Pro-Gly-Asn-Ala. The use of an Ala proxy is based on the assumption that, for any given conformation, the non-hydrogen-bonding side chains can adopt conformations that do not perturb the hydrogen-bonding side chains. In this stage of the analysis, Cys and Met were not treated as hydrogen-bonding residues.

Practical considerations forced the three approximations discussed above: (i) use of an Ala proxy, (ii) relaxation of close contact limits, and (iii) rigid backbone paradigms. Even by resorting to these approximations, the complete search of a small protein such as ribonuclease requires ~30 weeks of VAX 11/780 processor time. In practice, typical proteins take 2 to 3 weeks of processor time with a dedicated CSPI 6430 array processor together with three microVAX II computers. With abundant processor capability an Ala proxy would not be necessary and all amino acids could be rotated and checked for steric conflicts. In addition, side-chain dihedral angles could be rotated in smaller increments, with the Ramachandran "extreme" limit as the sole criterion when screening steric conflicts. Ideally, the backbone would also be allowed to move, and the evaluation would explicitly include Pro at any position, thereby eliminating the need for backbone paradigms.

Application to proteins of known structure. A diverse set of 13 x-ray elucidated proteins was chosen for analysis by the above methods. The proteins and their parenthesized Brookhaven file names (*13*) are: carboxypeptidase A (5CPA), parvalbumin (3CPV), cytochrome c (4CYT), dihydrofolate reductase (4DFR), flavodoxin (4FXN), human lysozyme (1LZ1), myoglobin (1MBO), plastocyanin (1PCY), avian pancreatic peptide (1PPT), pancreatic trypsin inhibitor (5PTI), ribonuclease (5RSA), scorpion neurotoxin (1SN3), and triose phosphate isomerase (1TIM).

Included were representatives from each of the four classes (*3, 20*): (i) predominantly α-helical, (ii) predominantly β sheet, (iii) mixed helix and sheet, and (iv) segregated domains of helix and sheet.

The 13 proteins include 54 helices. For consistency, "kinked" helices (such as residues 21 to 42 in MBO) are counted as two distinct helical segments. These helical segments can be examined for NTB's and CTB's (Fig. 4). Direct comparison between the structures and the histograms is instructive, although it does not allow for the existence of possible folding intermediates that are either modified or eliminated entirely in the final crystal structures. Comparison is further complicated by ambiguity in the precise location of helix boundaries in the x-ray structures. Our definition, described above, differs slightly from assignments listed in the header records of the protein database (*13*) or those given by Kabsch and Sander (*4*).

Most helices are bracketed by a conspicuous cluster of NTB's or CTB's. Of the 54 helices, 44 have an NTB that overlaps the N1 residue (or, on occasion, an NTB that approaches N1 to within a

residue or two) or else a boundary within one helical turn of the protein NH_2-terminus, and 45 have either a corresponding CTB or terminate within one helical turn of a prolyl residue or the protein COOH-terminus. Alternatively, consecutive helices (for example, residues 4 to 40 in MBO), where a helical sequence is interrupted by four or fewer nonhelical residues, can be counted as single helical elements. In this case, 38 of 47 helix NH_2-termini are satisfied by an NTB or protein NH_2-terminus and 40 of 47 helix COOH-termini are satisfied by a CTB or protein COOH-terminus or else terminated by a Pro residue.

The protein termini have considerable flexibility and are not well represented by our rigid six-residue paradigms. For this reason, the protein NH_2- or COOH-terminus is considered to be able to provide hydrogen-bond acceptors or donors, respectively, for a helix boundary that is no farther than three residues away. Nevertheless, many helices with boundaries near the protein termini have NTB's and CTB's.

Of the ten helices lacking NTB's, six have between one and three glycyl residues within the initial six-residue window: CPA[254], DFR[97], FXN[94], LZ1[25], TIM[138], and TIM[215]. (Numbers in brackets indicate the N1 residue of the helix.) Such cases would not be adequately represented by the backbone conformations in Tables 2 and 3 because Gly residues, lacking side chains, have unusual conformational flexibility. Moreover, of these six, LZ1[25] and FXN[94] do have NTB's if Met is considered to form hydrogen bonds (*14, 17*), and TIM[138] could equally well be regarded as a "kink" in a longer helix beginning at residue 130. Three other helices lacking NTB's, DFR[25], MBO[101], and TIM[46], have Pro residues within the initial six-residue window, interposed between the ostensible helix boundary and an adjacent upstream helix or visible NTB. The tenth helix without an NTB, LZ1[110], has a "bridge" of 3_{10} turns to a nearby upstream NTB.

Similarly, of the nine helices lacking CTB's, five have either one or two Gly residues within the final six-residue window: CPV[87], FXN[104], TIM[85], TIM[100], and TIM[118]. (Numbers in brackets indicate the C1 residue of the helix.) A sixth helix, DFR[83], has both Pro and Gly within the final six-residue window. Moreover, TIM[100] can be regarded equally well as a "kink" in a longer helix that resumes after residue 104. A similar situation obtains for the seventh helix, MBO[17], that resumes after a disruption of three residues. The eighth helix, CPV[15], terminates in a series of near-helical turns. The ninth helix CPA[100], which lacks a CTB, does have hydrogen bonds stabilizing the C3, C2, and C1 residues, but all three hydrogens are contributed by donors distant in sequence. It is possible that helices lacking NTB's and CTB's can nevertheless be stabilized by tertiary interactions, although another explanation is also possible for CPA[100] (*21*).

It is known that Gly (*22*), like Pro (*23*), can function as a helix breaker under suitable circumstances. If so, all but 1 of the 108 helix boundaries have an NTB or CTB or can be convincingly rationalized.

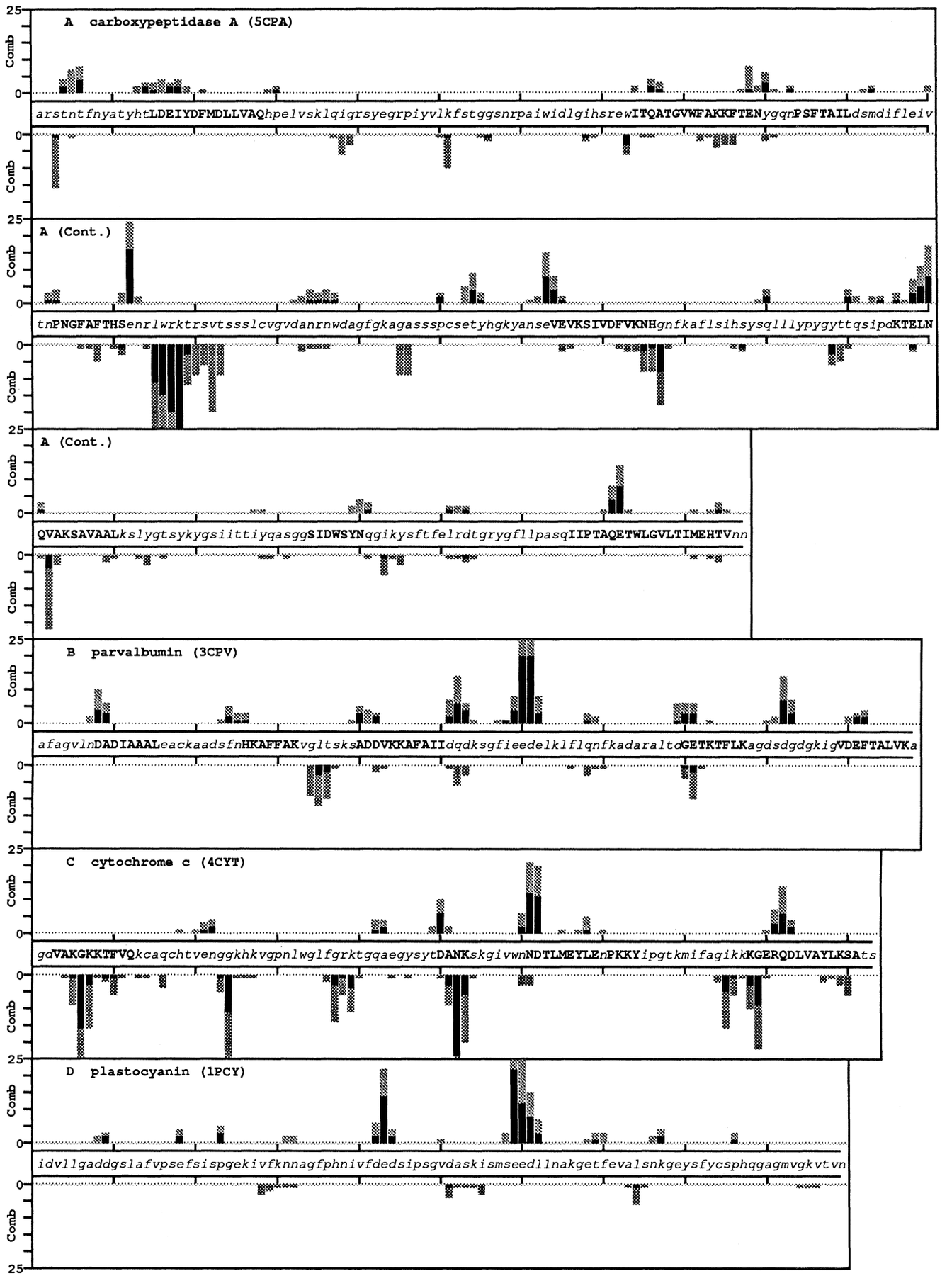
The 19 backbone paradigms used in our study represent only 60 percent of the observed distribution of turns at the NH_2-termini of helices and 50 percent at the COOH-termini. However, we confirmed that none of the preceding failures to find NTB's or CTB's would have been rescued by the inclusion of additional classes of turns. All failures were due instead to an insufficient number of polar residues.

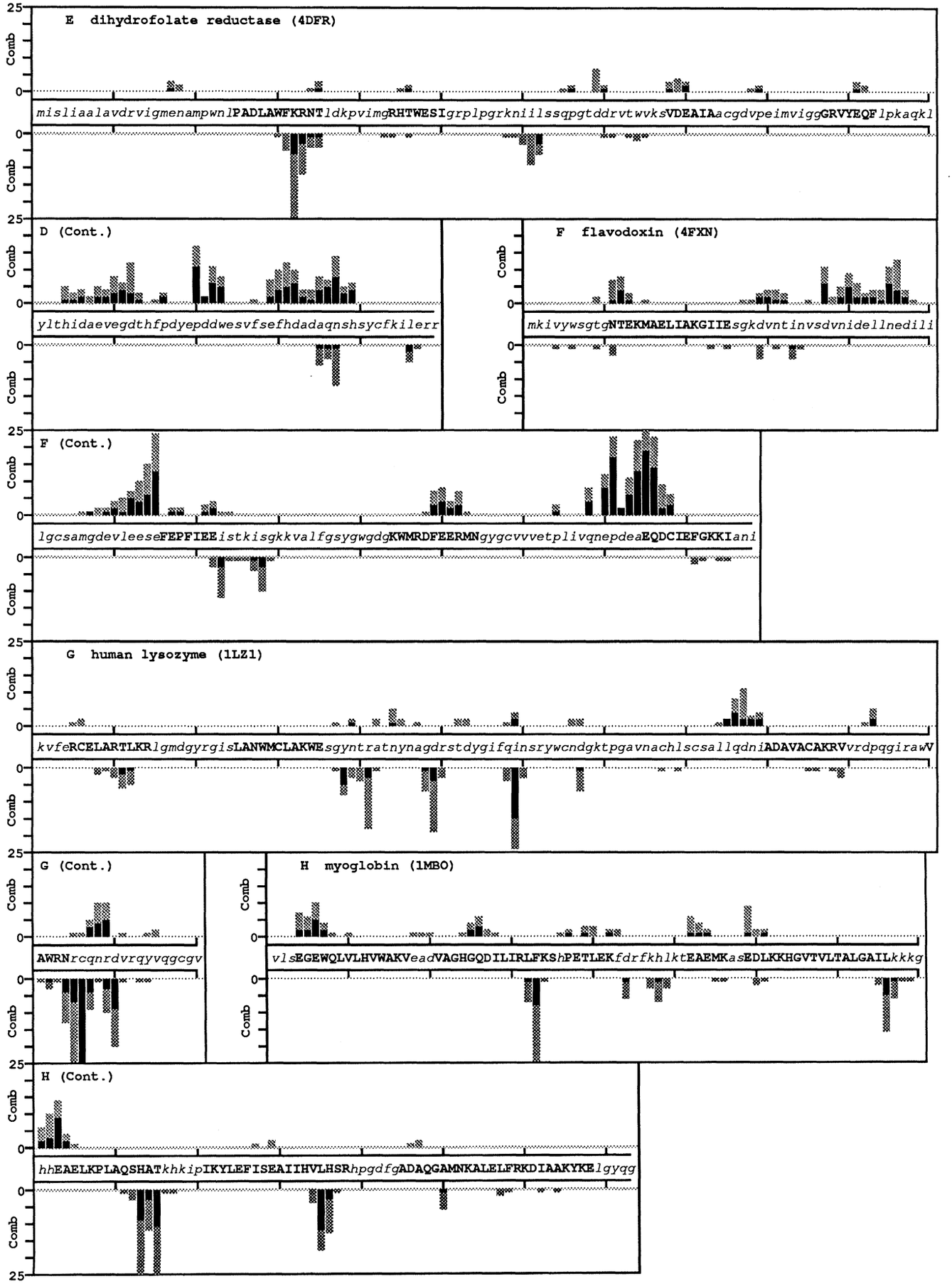
According to the helix hypothesis, the presence of NTB's and CTB's is necessary for helix formation. The degree to which their presence is also sufficient is uncertain. An analysis of sufficiency is complicated by five factors. (i) Conditions under which Gly or Pro (or both) will function as helix breakers must be made precise. (ii) The assessment of sufficiency is not simply a matter of comparing those sequences bracketed by NTB's and CTB's against the location of known helices. When an NTB and CTB overlap (Fig. 4), those residues that can serve either as donors or acceptors (for example, Ser, Thr, Asn, and Gln) could contribute to either the NTB or the CTB, but not simultaneously to both. Either the CTB would be abolished when these pivotal residues contribute to the NTB, or, conversely, the CTB would be established at the expense of the NTB. An example of an overlapping, mutually exclusive NTB and CTB occurs near residue 60 in RSA. In this example, the residues Ser^{59}–Gln^{60}–Asn^{62} presumably contribute to the CTB of the observed helix, precluding the existence of an ostensible NTB at residue 60. (iii) In our analysis, we have deliberately neglected any residues between the NTB and CTB, although these may contribute to helix stability as well. (iv) Side chains of opposite charge can compete with side-chain to backbone hydrogen bonds and diminish the strength of some NTB's and CTB's. These competing interactions cannot be assessed at present because our hydrogen-bond library does not include side-chain to side-chain hydrogen bonds. (v) NTB's and CTB's can often be augmented by main-chain to main-chain hydrogen bonds involving adjacent turns, but these interactions are also not included in the library at present.

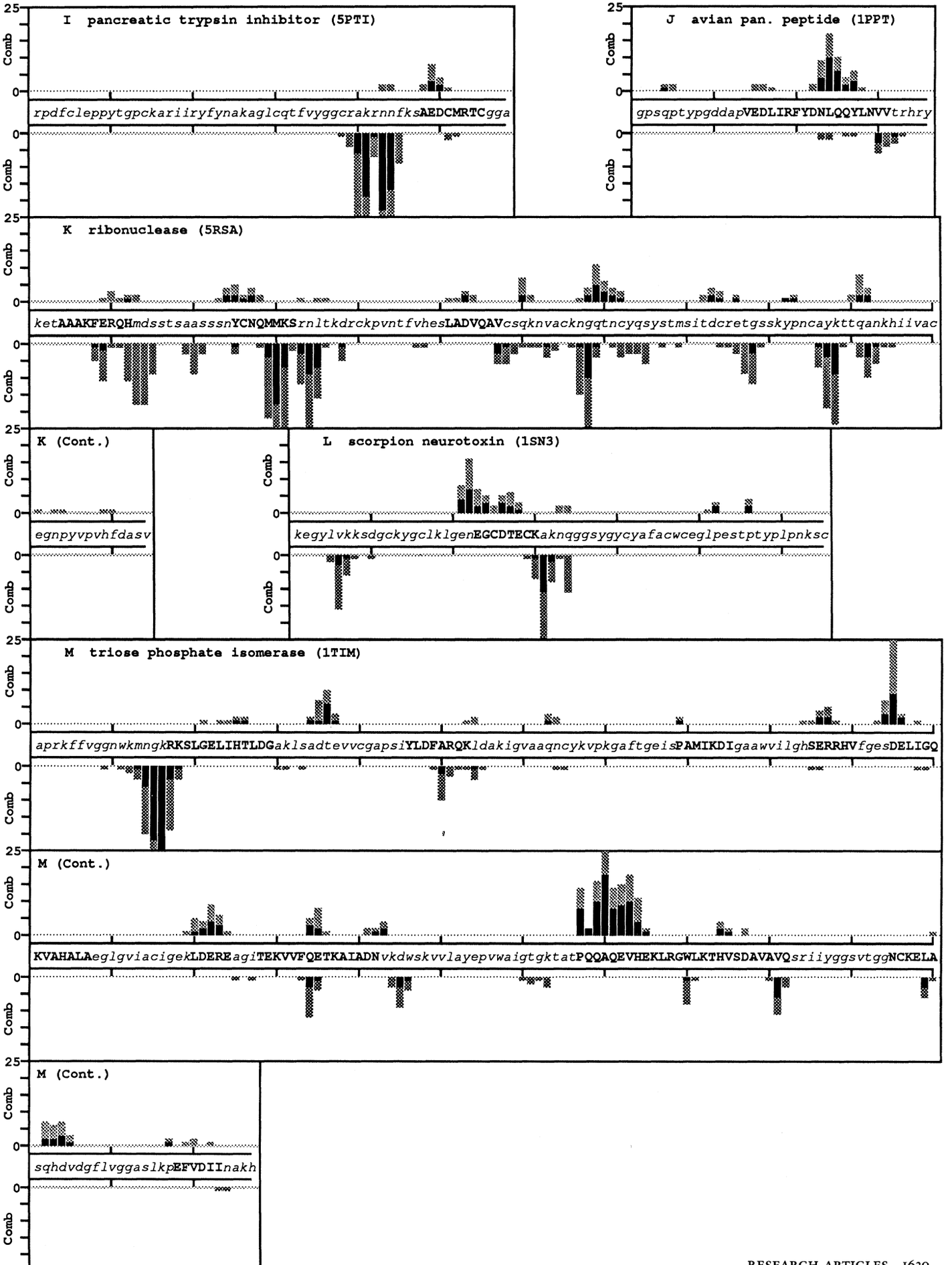
After these factors are taken into account, some nonhelical sequences bracketed by NTB's and CTB's remain. Most often, examination of such sequences reveals variant or perturbed helical structures, such as CPV(61–68), which adopts a series of 3_{10} turns or LZ1(103 to 109), which adopts kinked 3_{10} turns.

Protein folding and the helix hypothesis. A principal question raised by the helix hypothesis is whether segments bounded by NTB's and CTB's are helical in isolation. In studies of C peptide (residues 1 to 13 of ribonuclease A) and its analogs, Baldwin and coworkers (*5, 6*) demonstrated the presence of stable helices in aqueous solution. In their work, helix stability can be attributed in large part to flanking residues that interact with the helix macrodipole. For example (*6*), four peptides that differ only at the NH_2-terminal residue were synthesized: the alteration being Lys^1, Ala^1, acetyl-Ala^1, and succinyl-Ala^1. Consistent with the helix dipole model (*6, 24*), helix stability increases as the net charge at the NH_2-terminus becomes more negative in going from Lys(+2) → Ala(+1) → acetyl-Ala(0) → succinyl-Ala(−1). Stability undergoes a

Fig. 4. Histogram of NTB's and CTB's for 13 x-ray–eludicated proteins. The proteins and their parenthesized Brookhaven file names (*13*) are: (**A**) carboxypeptidase A (5CPA), (**B**) parvalbumin (3CPV), (**C**) cytochrome c (4CYT), (**D**) plastocyanin (1PCY), (**E**) dihydrofolate reductase (4DFR), (**F**) flavodoxin (4FXN), (**G**) human lysozyme (1LZ1), (**H**) myoglobin (1MBO), (**I**) pancreatic trypsin inhibitor (5PTI), (**J**) avian pancreatic peptide (1PPT), (**K**) ribonuclease (5RSA), (**L**) scorpion neurotoxin (1SN3), and (**M**) triose phosphate isomerase (1TIM). For each protein, the sequence is shown in one-letter code (*38*), with helical segments in boldface. Tic marks on the abscissa denote every tenth residue. The upper histogram indicates the results of searching the sequence for potential NH_2-terminal bounds (NTB's); each bar, positioned above window location N1, plots the number of backbone >N–H to side-chain hydrogen-bond combinations found within a given six-residue window. The lower histogram indicates corresponding information for backbone >C=O to side-chain hydrogen bonds at COOH-terminal bounds (CTB's), with bars positioned below window location C1. Dark bars in the histograms represent the number of combinations that satisfy all four consecutive >N–H groups (N, N1, N2, N3) or all four consecutive >C=O groups (C3, C2, C1, C), respectively. The superimposed light bars plot the number of combinations satisfying either the initial three (N1, N2, N3) or final three (C3, C2, C1) residues of the helix proper, that is, excluding residues N or C. Prolines, lacking an amide hydrogen, were considered to be satisfied automatically in an NTB. In an NTB, His was considered to be in the neutral form, with Nε protonated; in a CTB, His was considered to be in the neutral form with Nδ protonated. A small number of histogram bars that are isolated from neighboring bars by at least two residues on either side and that contain no combinations in which all four groups are satisfied have been omitted to enhance clarity.


A carboxypeptidase A (5CPA)
Comb
arstntfnyatyhtLDEIYDFMDLLVAQhpelvsklqigrsyegrpiyvlkfstggsnrpaiwidlgihsrewITQATGVWFAKKFTENygqnPSFTAILdsmdifleiv
A (Cont.)
tnPNGFAFTHSenrlwrktrsvtssslcvgvdanrnwdagfgkagassspcsetyhgkyanseVEVKSIVDFVKNHgnfkaflsihsysqlllypygyttqsipdKTELN
A (Cont.)
QVAKSAVAALkslygtsykygsiittiyqasggSIDWSYNqgikysftfelrdtgrygfllpasqIIPTAQETWLGVLTIMEHTVnn
B parvalbumin (3CPV)
afagvlnDADIAAALeackaadsfnHKAFFAKvgltsksADDVKKAFAIIdqdksgfieedelklflqnfkadaraltdGETKTFLKagdsdgdgkigVDEFTALVKa
C cytochrome c (4CYT)
gdVAKGKKTFVQkcaqchtvenggkhkvgpnlwglfgrktgqaegysytDANKskgivwnNDTLMEYLEnPKKYipgtkmifagikkKGERQDLVAYLKSAts
D plastocyanin (1PCY)
idvllgaddgslafvpsefsispgekivfknnagfphnivfdedsipsgvdaskismseedllnakgetfevalsnkgeysfycsphqgagmvgkvtvn


E dihydrofolate reductase (4DFR)
misliaalavdrvigmenampwnlPADLAWFKRNTldkpvimgRHTWESIgrplpgrkniilssqpgtddrvtwvksVDEAIAacgdvpeimvigGRVYEQFlpkaqkl
D (Cont.)
ylthidaevegdthfpdyepddwesvfsefhdadaqnshsycfkilerr
F flavodoxin (4FXN)
mkivywsgtgNTEKMAELIAKGIIEsgkdvntinvsdvnidellnedili
F (Cont.)
lgcsamgdevleeseFEPFIEEistkisgkkvalfgsygwgdgKWMRDFEERMNgygcvvvetplivqnepdeaEQDCIEFGKKIani
G human lysozyme (1LZ1)
kvfeRCELARTLKRlgmdgyrgisLANWMCLAKWEsgyntratnynagdrstdygifqinsrywcndgktpgavnachlscsallqdniADAVACAKRVvrdpqgirawV
G (Cont.)
AWRNrcqnrdvrqyvqgcgv
H myoglobin (1MBO)
vlsEGEWQLVLHVWAKVeadVAGHGQDILIRLFKShPETLEKfdrfkhlktEAEMKasEDLKKHGVTVLTALGAILkkkg
H (Cont.)
hhEAELKPLAQSHATkhkipIKYLEFISEAIIHVLHSRhpgdfgADAQGAMNKALELFRKDIAAKYKElgyqg
Comb
25
0


I pancreatic trypsin inhibitor (5PTI)
rpdfcleppytgpckariiryfynakaglcqtfvyggcrakrnnfksAEDCMRTCgga
J avian pan. peptide (1PPT)
gpsqptypgddapVEDLIRFYDNLQQYLNVVtrhry
K ribonuclease (5RSA)
ketAAAKFERQHmdsstsaasssnYCNQMMKSrnltkdrckpvntfvhesLADVQAVcsqknvackngqtncyqsystmsitdcretgsskypncayktttqankhiivac
K (Cont.)
egnpyvpvhfdasv
L scorpion neurotoxin (1SN3)
kegylvkksdgckygclklgenEGCDTECKaknqggsygycyafacwceglpestptyplpnksc
M triose phosphate isomerase (1TIM)
aprkffvggnwkmngkRKSLGELIHTLDGaklsadtevvcgapsiYLDFARQKldakigvaaqncykvpkgaftgeisPAMIKDIgaawvilghSERRHVfgesDELIGQ
M (Cont.)
KVAHALAeglgviacigekLDEREagiTEKVVFQETKAIADNvkdwskvvlayepvwaigtgktatPQQAQEVHEKLRGWLKTHVSDAVAVQsriiyggsvtggNCKELA
M (Cont.)
sqhdvdgflvggaslkpEFVDIInakh
Comb
25
0

small increase when there is a shift from Lys to Ala, with larger increases occurring from Ala → acetyl-Ala → succinyl Ala (*6*). The acetyl group not only removes the charge at the NH_2-terminus, but also strengthens the NTB by providing a hydrogen-bond acceptor for the >N–H of residue 4. Moreover, the succinyl group can provide two or possibly three additional acceptors that satisfy >N–H donors in residues 4, 3, and possibly 2 (*25*). The Baldwin and Stewart analogs (*6*) will be valuable compounds for assessing the helix hypothesis in isolated peptides.

While most helices in Fig. 4 are bracketed by NTB's and CTB's, all of the postulated main-chain to side-chain hydrogen bonds may not persist in the crystal structure, although many do (*14, 26*). The crystal structures of the 13 proteins include 54 helices. Within these, 48 percent of the N–N1–N2–N3 residues and 35 percent of the C3–C2–C1–C residues are satisfied by side-chain or backbone hydrogen bond partners contributed by residues corresponding to the initial or final six-residue window of each helix. An additional 5 percent at the NH_2-termini and 9 percent at the COOH-termini are satisfied by sequentially distant side-chain or backbone intramolecular hydrogen-bond partners. Remaining groups are satisfied either by partners from neighboring protein molecules in the crystal lattice (*25*) or by solvent molecules.

We interpret these findings to mean that NTB's and CTB's, while required in the nascent helix, can often be liberated once the helix is "fixed" by the tertiary fold. During these postulated tertiary adjustments, the helix boundaries might be "peeled back" or extended by a few residues relative to the position of the NTB/CTB. Certainly the intermolecular hydrogen bonds that are apparent in the crystal structure (and possibly needed for successful crystallization) must, of necessity, involve a subsequent rearrangement of hydrogen bonds from the solution structure. A compilation of observed amino acid preferences at helix termini in 45 x-ray elucidated proteins is presented in the accompanying report by Richardson and Richardson (*26*), who arrive at similar conclusions about hydrogen bonding.

Protein sequence comparisons often reveal surprising relationships and unanticipated homologies (*27*). Strategies for comparison implicitly are based on the assumption that conservative substitutions are synonymous with chemical similarity (for example, Glu for Asp). However, conservation of structure is also an important factor, and maintenance of NTB's and CTB's among homologous structures is expected. Indeed, these sites do appear to be conserved in the hemoglobins (*28*).

Many statistical approaches to helix prediction have been proposed (*23*). For such procedures a database of known structures is used to derive an empirical probability that each residue type will be found in a helix. Most of the residues that participate in NTB's and CTB's are classified as helix breakers in statistical studies (*23*). Exceptions include Glu, Lys, and Gln; and, although classified as helix formers, they have an observed tendency to be localized near the helix NH_2- or COOH-termini (*23*). These empirical classifications are entirely consistent with our hypothesis that NTB's and CTB's comprise helix boundaries.

Initial results suggest that membrane-spanning helices do not require NTB's and CTB's, at least in the case of the photosynthetic reaction center (*29*). The extramembrane helices in this structure, which do have NTB's and CTB's, serve as a control. To pursue this observation further, we analyzed crambin, a hydrophobic molecule with two helices that crystallizes only in the presence of organic cosolvents (*30*). Such conditions are suggestive of a membrane-like environment, and crambin was chosen with this possibility in mind. Neither helix has NTB's, although one weak CTB is in evidence. In a membrane, a hydrophobic segment of a protein would favor a helix in order to satisfy backbone polar groups (*31*). Such a segment, lacking an NTB and a CTB, may undergo conformational transition from nonhelix in a polar environment to helix in an apolar environment, and, in fact, the LamB signal sequence appears to function in this way (*32*).

Some of the residues in NTB's and CTB's may also be involved in the binding of ligands or prosthetic groups. In such cases, changes in the structure of the apoprotein are expected. Examples include the binding of charged groups by peptide backbone (*33*) and the binding of heme by apomyoglobin (*34*).

The thermodynamics of helix formation remains to be assessed. Suitable models for cooperative hydrogen bonding in NTB's and CTB's could not be found. Presumably, such bonds are comparatively strong, especially those involving charged side chains. Our data also include the number of conformations for each combination shown in Fig. 4. Frequently these numbers are large, exceeding 10^6 in many instances. The corresponding decrease in conformational entropy needed to maintain side-chain to main-chain hydrogen bonds in NTB's and CTB's may turn out to be surprisingly small.

The helix hypothesis leads to a number of testable predictions. For example:

1) Transient side-chain to backbone hydrogen bonds at helix termini that do not persist in the x-ray elucidated structure may be detectable by nuclear magnetic resonance during protein folding.

2) Sequences bracketed by NTB's and CTB's that contain no prolyl or glycyl residues should be helical in aqueous solution (provided that the complete peptide dissolves and does not aggregate). Further, sequential elimination of the residues involved in either the NTB or CTB should lead to an incremental reduction in helicity.

3) Site-directed mutagenesis can be used in a variety of ways. For example, removal of Pro or Gly helix terminators that are situated upstream from a nearby CTB should extend the helix. Similarly, nonhelical sequences bracketed by NTB's and CTB's but containing an intervening Gly or Pro sequence should become helical when these helix-breaking residues are removed. Introduction of a companion CTB downstream from an isolated NTB, or a companion NTB upstream from an isolated CTB, should induce formation of an additional helix. (These possibilities may be disrupted by tertiary interactions.)

4) Charged side chains that form a hydrogen bond with backbone >N–H or >C=O groups should exhibit both a shift in pK and protection against hydrogen exchange.

5) Under suitable circumstances, neutral and protonated His should make pH-dependent contributions to an NTB and a CTB, respectively. Similarly, protonated Asp and Glu should participate in a pH-dependent CTB.

In addition to the specific tests mentioned here, the helix hypothesis should prove useful in protein engineering and design (*35*).

The question of whether secondary structure is formed before tertiary structure has yet to be resolved in protein folding studies (*36*). The helix hypothesis implies that helical secondary structure need not depend on tertiary interactions. In particular, helices with a strong NTB and CTB together with an appropriately stable intervening sequence may function as independent "seeds for folding" (*37*).

REFERENCES AND NOTES

1. L. Pauling, R. B. Corey, H. R. Branson, *Proc. Natl. Acad. Sci. U.S.A.* **37**, 205 (1951).
2. M. F. Perutz, *Nature* **167**, 1053 (1951).
3. J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981).
4. W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983).
5. A. Bierzynski, P. S. Kim, R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2470 (1982); P. S. Kim and R. L. Baldwin, *Nature* **307**, 329 (1984); M. A. Jimenez, J.

L. Nieto, J. Herranz, M. Rico, J. Santoro, *FEBS Lett.* **221**, 320 (1987).
6. K. R. Shoemaker, P. S. Kim, E. J. York, J. M. Stewart, R. L. Baldwin, *Nature* **326**, 563 (1987).
7. G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure* (Springer, New York, 1979).
8. J. A. Schellman, *Compt. Rend. Trav. Lab. Carlsberg (Ser. Chim.)* **29**, 230 (1955).
9. B. K. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971); C. Chothia, *ibid.* **105**, 1 (1976).
10. W. F. DeGrado, F. J. Kezdy, E. T. Kaiser, *J. Am. Chem. Soc.* **103**, 679 (1981); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982).
11. P. Y. Chou and G. D. Fasman, *Biochemistry* **13**, 211 (1974); D. E. Blagdon and M. Goodman, *Biopolymers* **14**, 241 (1975).
12. W. Kabsch and C. Sander, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1075 (1984).
13. F. C. Bernstein *et al.*, *J. Mol. Biol.* **112**, 535 (1977).
14. E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
15. R. Taylor and O. Kennard, *Acc. Chem. Res.* **17**, 320 (1984); A. Vedani and J. D. Dunitz, *J. Am. Chem. Soc.* **107**, 7653 (1985).
16. H. A. Scheraga, Quantum Chemistry Program Exchange, Program No. 286, Indiana University Chemistry Department, Bloomington (1975).
17. R. Srinivasan and K. K. Chacko, *Conformation of Biopolymers*, G. N. Ramachandran, Ed. (Academic Press, New York, 1967).
18. G. N. Ramachandran and V. Sasisekharan, *Adv. Prot. Chem.* **23**, 283 (1968).
19. J. Moult and M. N. G. James, *Proteins* **1**, 146 (1986).
20. M. Levitt and C. Chothia, *Nature* **261**, 552 (1976).
21. A. A. Kossiakoff, *Science* **240**, 191 (1988). Kossiakoff suggests that protein deamidation sites involve Asn-Ser sequences in which the Asn makes a hydrogen bond to the backbone at position $i + 2$ while the Ser hydroxyl makes a hydrogen bond to the Asn side chain. It is conceivable that the amino acid at CPA 101 was originally an Asn that became deamidated prior to sequence determination. If so, the one exceptional helix boundary lacking CTB, namely CPA[100], would then have a CTB.
22. C. Schellman, *Protein Folding*, R. Jaenicke, Ed. (Elsevier/North-Holland Biomedical Press, Amsterdam, 1980), p. 53.
23. P. Y. Chou and G. D. Fasman, *Annu. Rev. Biochem.* **47**, 251 (1978).
24. A. Wada, *Adv. Biophys.* **9**, 1 (1976); W. G. J. Hol, *Prog. Biophys. Mol. Biol.* **45**, 149 (1985).
25. L. G. Presta and G. D. Rose, unpublished results.
26. J. S. Richardson and D. C. Richardson, *Science* **240**, 1648 (1988).
27. R. F. Doolittle, *Of Urfs and Orfs* (University Science Books, Mill Valley, CA, 1987).
28. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
29. J. Deisenhofer, O. Epp, K. Miki, R. Huber, H. Michel, *Nature* **318**, 618 (1985).
30. W. A. Hendrickson and M. M. Teeter, *ibid.* **290**, 107 (1981).
31. D. M. Engleman, T. A. Steitz, A. Goldman, *Annu. Rev. Biophys. Biophys. Chem.* **15**, 321 (1986).
32. M. S. Briggs and L. M. Gierasch, *Adv. Prot. Chem.* **38**, 109 (1986).
33. F. A. Quiocho, J. S. Sack, N. K. Vyas, *Nature* **329**, 561 (1987).
34. S. C. Harrison and E. R. Blout, *J. Biol. Chem.* **240**, 299 (1965); J. T. J. Lecomte and G. N. LaMar, *J. Am. Chem. Soc.* **109**, 7219 (1987).
35. D. L. Oxender and C. F. Fox, Eds., *Protein Engineering* (Liss, New York, 1987).
36. P. S. Kim and R. L. Baldwin, *Annu. Rev. Biochem.* **51**, 459 (1982).
37. R. L. Baldwin, *Trends Biochem. Sci.* **11**, 6 (1986).
38. Single letter abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.
39. We thank R. Baldwin, P. Kim, S. Taylor, and C. R. Matthews for many useful discussions; B. Zimm and J. Lecomte for their critical reading of the manuscript; an anonymous referee for helpful suggestions, and E. Lattman for insightful comments at every stage of this work. Thirteen years ago, Kensal Van Holde urged that the key question is not which sequences will be found in helices, but rather why all sequences are not found in helices. Good questions, like good teachers, leave a lasting imprint. Supported by NIH grants GM 29458 and AG 06084 and by a Dean's grant from John Burnside.

25 February 1988; accepted 28 April 1988