PROTEINS: Structure, Function, and Genetics 118-129 (1987)

Prediction of Secondary Structure by Evolutionary Comparison: Application to the α Subunit of Tryptophan Synthase

Irving P. Crawford,[1] Thomas Niermann,[2] and Kasper Kirschner[2]
[1] *Department of Microbiology, University of Iowa, Iowa City, Iowa 52242;* [2] *Department of Biophysical Chemistry, Biozentrum, University of Basel, CH 4056 Basel, Switzerland*

ABSTRACT **The amino acid sequences of the α subunits of tryptophan synthase from ten different microorganisms were aligned by standard procedures. The α helices, β strands and turns of each sequence were predicted separately by two standard prediction algorithms and averaged at homologous sequence positions. Additional evidence for conserved secondary structure was derived from profiles of average hydropathy and chain flexibility values, leading to a joint prediction. There is good agreement between (1) predicted β strands, maximal hydropathy and minimal flexibility, and (2) predicted loops, great chain flexibility, and protein segments that accept insertions of various lengths in individual sequences. The α subunit is predicted to have eight repeated β-loop-α-loop motifs with an extra N-terminal α helix and an intercalated segment of highly conserved residues. This pattern suggests that the teritiary structure of the α subunit is an eightfold α/β barrel. The distribution of conserved amino acid residues and published data on limited proteolysis, chemical modification, and mutagenesis are consistent with the α/β barrel structure. Both the active site of the α subunit and the combining site for the β_2 subunit are at the end of the barrel formed by the carboxyl-termini of the β strands.**

Key words: homologous proteins, hydropathy index, chain flexibility, α/β barrels, tryptophan synthase alpha subunit

INTRODUCTION

Comparison of homologous amino acid sequences from widely divergent organisms can give the following information: (1) the position and nature of conserved residues important in maintaining either structure or function, (2) the likely positions of surface loops, indicated by polypeptide segments accepting insertions of extra amino acids, and (3) improved accuracy in the prediction of secondary structure.[1] The underlying rationale is that structural features essential for stability, correct folding, and function are conserved during divergent evolution, while those of lesser significance will vary.[2]

We have studied phosphoribosyl anthranilate (PRA) isomerase:indoleglycerol phosphate (IGP) synthase, a bifunctional enzyme involved in the biosynthesis of tryptophan.[3,4] The three dimensional structure of the protein from *E. coli* has recently been determined by protein crystallography,[5] showing that both the IGP synthase and PRA isomerase domains have the eightfold α/β barrel-folding topology known as the "TIM barrel".[6,7] When the sequences of PRA isomerase and IGP synthase from nine different microorganisms were aligned, we found that secondary structure predictions averaged on the basis of the amino acid alignments[1] resulted in much higher scores of correctly predicted residues in α helices, β strands, or loops than obtained from predictions for each individual sequence (T. Niermann and K. Kirschner, unpublished results). Moreover, segments accepting insertions were found exclusively between secondary structural elements.[2] Finally, averaged hydropathy values correlated strongly with β strands and averaged values of chain flexibility with surface loops.

Because a large number of protein sequences are known for the α subunit of tryptophan synthase, we decided to use the same approach to obtain a clearer picture of the secondary structure of this protein.

Tryptophan synthase is the enzyme catalyzing the reaction immediately following IGP synthase in the tryptophan synthetic pathway. It has an $\alpha_2\beta_2$ composition in all prokaryotes examined and a fused $(\alpha\beta)_2$ composition in ascomycetes.[8] This enzyme has received intensive genetic and biochemical investigation in several systems.[9,10]

The tryptophan synthase molecule performs two enzymatic reactions in a concerted fashion.[8] Mutational analysis has revealed that the first reaction, the aldolysis of IGP to form indole and glyceraldehyde-3-phosphate, is the primary responsibility of the α subunit. The second reaction, the combining of indole and serine to form tryptophan is catalyzed primarily by the pyridoxal-5'-phosphate-containing β_2 subunit. The physiological reaction, in which indole-

Received December 23, 1986; accepted May 4, 1987.

Address reprint requests to Dr. I.P. Crawford, Department of Microbiology, University of Iowa, Iowa City, Iowa 52242

Abbreviations used: PRA, N-(5'-phosphoribosyl)anthranilate; IGP, indoleglycerol phosphate; TIM, triose phosphate isomerase; GOR, method of Garnier, Osguthorpe, and Robson; CHF, method of Chou and Fasman.

glycerol phosphate appears to react directly with serine to form tryptophan and glyceraldehyde-3-phosphate, presumably results from a coupling of these two reactions, because indole produced in the first reaction does not exchange freely with indole in solution.

Each subunit's activity is stimulated upon combination with the other subunit.[8] Most missense mutants of either subunit, devoid of catalytic activity in that subunit's reaction, are nevertheless able to stimulate the other subunit's activity. However, a different type of α subunit missense mutant can perform its own reaction quite well but cannot stimulate the β_2 subunit in the indole-to-tryptophan reaction, presumably because it is defective in some aspect of complex formation or subunit interaction. [11] These findings suggest that the sites responsible for catalytic activity of the α subunit and its binding to the β_2 subunit are not identical but interact somehow.

Crystals of *Salmonella typhimurium* tryptophan synthase suitable for X-ray diffraction were obtained recently,[12] but as yet no detailed model of the three-dimensional structure of the enzyme exists. Small-angle X-ray [13] and neutron[14] scattering studies have established that the α and β_2 subunits of tryptophan synthase are generally globlar to cylindrical in shape and that the two α subunits in the $\alpha_2\beta_2$ complex appear to be in contact with the β_2 subunit but not with each other.

The predicted secondary structure of the α subunit of tryptophan synthase clearly follows the pattern of an eightfold α/β barrel. Additional information gleaned from the *E. coli* enzyme by partial proteolysis, chemical modification, and mutation supports this tentative model of its tertiary structure.

MATERIALS AND METHODS

The following α subunit (or α domain) sequences were available from published sources or were provided prior to publication by the investigators named: *E. coli*,[15] *S. typhimurium*,[15] *Pseudomonas aeruginosa*,[16] *Bacillus subtilis*,[17] *Saccharomyces cerevisiae*,[18] *Klebsiella aerogenes*,[19] *Brevibacterium lactofermentum*,[20] *Lactobacillus casei* (Y. Natori, Y. Kano, and F. Imamoto, personal communication), and *Caulobacter crescentus* (partial sequence, C. Ross and M. Winkler, personal communication). The sequence of *Vibrio parahemolyticus* α subunit was recently obtained by one of us (M. Silverman and I. Crawford, unpublished results). The sequences were first aligned in pairs,[21] using the amino acid similarity matrix of Dayhoff et al.[22] The multiple sequence alignment was done by eye, based on the pairwise alignments and minimizing the number of gaps (insertions).

The consensus secondary structure prediction according to the GOR method (1) was obtained as follows. Predictions were made for each individual protein separately, first using unbiased decision constants. After it was clear that the secondary structure of the α subunit consists mainly of alternating α helices and β strands, the definitive prediction was made with the decision constants optimized for α/β proteins ($DC_\alpha = DC_c = DC_t = 0$, $DC_\beta = 20$[23]). The secondary structures predicted for individual sequences are presented in Figure 4. In the next cycle the probabilities of the four states (α, β, turn, and coil) provided by each of the homologous residues were added according to the alignment of Figure 1 and the sums were divided by the number of homologs. The state with the highest probability was taken as the consensus conformational state at that sequence position. The averaged probabilities are presented in Figure 5.

The consensus secondary structure prediction according to the CHF method[24,25] was obtained in a different manner. Again the predictions were first made for each protein sequence separately by using the computer program provided by Intelligenetics via BIONET. Ambiguities arising from overlapping assignments were resolved as follows: partially overlapping regions were attributed to the state (α, β, or turn) with the larger average propensity; completely overlapping regions were scored as ambiguous by arbitrarily attributing 50% of one and 50% of the other of the two major states to that region. The secondary structures predicted for individual sequences are presented in Figure 6. Second, at every sequence position in the aligned set of Figure 1, the state with the highest score in the column was taken as the consensus conformational state at that position. Where individual sequences contained unassigned residues, these were ignored in determining the consensus state.

The average hydropathy index was determined as follows. First the hydropathic index profile of each individual protein sequence was calculated with the algorithm developed by Kyte and Doolittle,[26] using a span setting of 5 residues. Second, the values for each sequence position according to the alignment of Figure 1 were added and the sums divided by the number of homologs. The average relative chain flexibility at each sequence position was calculated in an analogous manner according to the algorithm of Karplus and Schulz,[27] using the neighbor-correlated B_{norm} values given in that paper.

RESULTS

The alignment of nine complete (and one partial) amino acid sequences of tryptophan synthase α subunits from various microorganisms is shown in Figure 1. Twenty-nine positions (10.8%) have identical residues, but the number increases to 50 (18.7%) if a single difference is allowed. This latitude allows infrequent substitution of a similar residue,[22] or a possible isolated sequencing error, and emphasizes the clustering of conserved residues. There are 7 gaps at different positions (beginning at residues 96, 147, 161, 188, 207, 232, and 249), four of which are imposed by

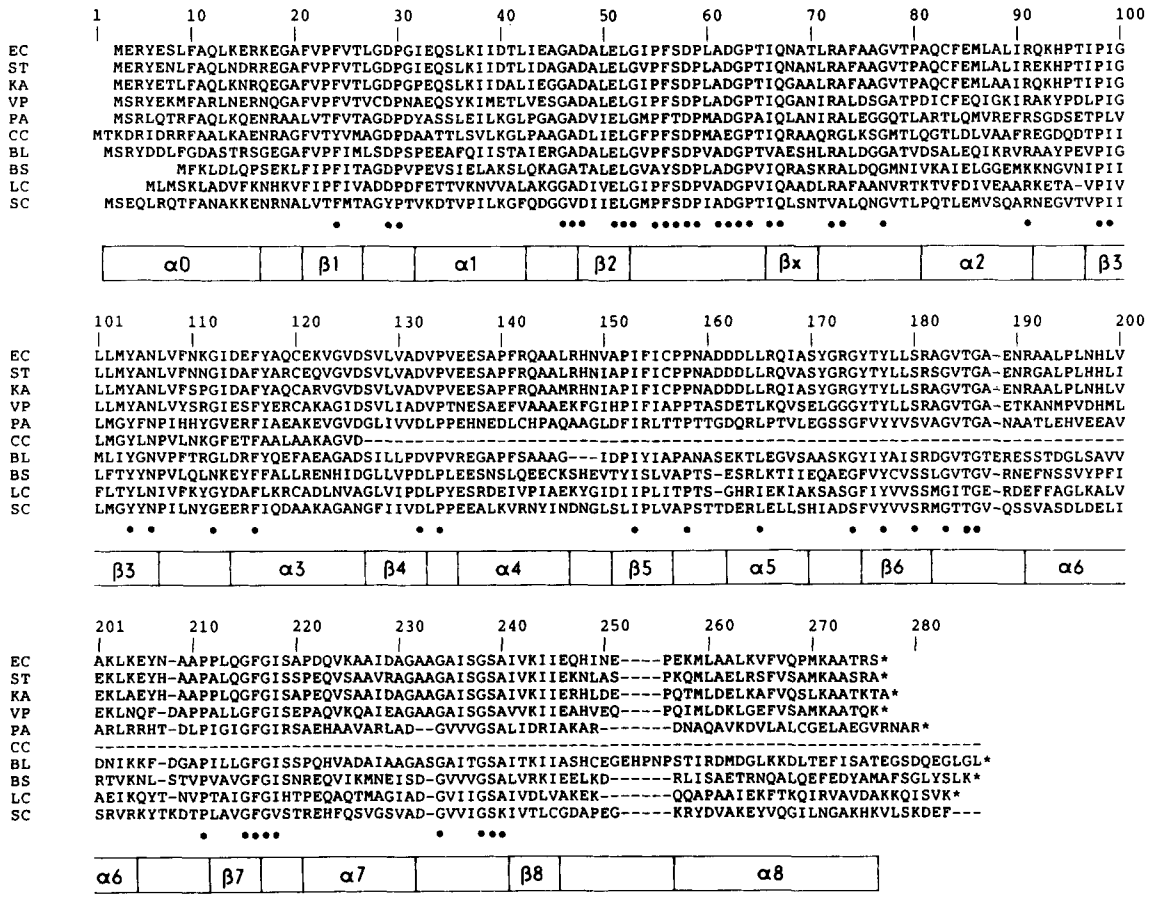

Fig. 1. Amino acid sequence alignment and predicted secondary structure for the α subunit of tryptophan synthase: correlation between sites accepting amino acid insertions and predicted loop segments. The sequences are aligned to maximize similarities. Single-letter abbreviations are used for amino acid residues, with hyphens for gaps, and asterisks at the location of stop codons. The following abbreviations are used for the sources of α subunits: (BL) *B. lactofermentum;* (BS) *B. subtilis;* (CC) *C. crescentus;* (EC) *E. coli;* (KA) *K. aerogenes;* (LC) *L. casei;* (PA) *P. aeruginosa;* (SC) *S. cerevisiae,* N-terminal α-domain; (ST) *S. typhimurium;* (VP) *V. parahemolyticus.* Numbering is from M1 of *C. crescentus* to L282 of *B. lactofermentum,* counting gaps as pseudoresidues. (●) Positions where all or all but one of the residues are identical. Below the sequences is shown the joint secondary structure assignment (JNT) from Table I: (αi) i th α helix, (βi) i th β strand, empty boxes are loops between α and β elements.

more than one sequence. These sites appear to accept insertions or deletions and therefore to independently mark the position of surface loops.

Table I gives the limits of the consensus structure resulting from the two secondary structure prediction methods (cf. Figs. 4 and 5). The CHF method predicts only three different states (α, β, and turn), but, as discussed by Taylor and Thornton,[23] the GOR method as used here does not distinguish between predicted coils and reverse β turns. Therefore we designate the third state as "loop" to emphasize the role of these regions as connectors with unspecified secondary structure.

A joint prediction involving a superposition of the CHF and GOR consensus predictions can only shorten the predicted α helices and β strands when applied conservatively. Moreover, there is no rational means of resolving contradictory predictions. Therefore, we sought to resolve ambiguities and strengthen the predictions by incorporating additional information into the predictive scheme.

Studies with the bifunctional enzyme PRA isomerase:IGP synthase have shown that two properties of amino acid residues correlate with β strands and surface loops. The first is the hydropathy index, [26] which reflects the preference of a given residue for the hydrophobic interior of proteins.[1,23] The second is the relative chain flexibility, which correlates with loops on the protein surface.[27] Whereas the hydropathy and flexibility profiles of individual sequences of both PRA isomerase and IGP synthase are difficult to interpret, the average values at each homologous se-

TABLE I. Predicted Secondary Structure Elements of the α Subunit of Tryptophan Synthase*

Limits				Desig-	
GOR	CHF	JNT	Length	nation	Comments
1–16	2–16	1–16	(14)	α0	Length given for *E. coli* sequence
17–20	17–19	17–20	4	t01	
21–26	20–27	21–26	6	β1	Irregular because of partially conserved P23
27–31	28–32	27–31	5	t11	
32–50	33–54	32–42	11	α1	Beginning of uncertain region
		43–47	5	t12	42–45: similar GOR probabilities for α, coil
				(α)	44–54: uncertain region
51–52		48–52	5	β2	Good correlation with hydropathy and flexibility
53–65	57–64	53–65	13	t2x	
66–70	66–69	66–70	5	βx	66–75: similar GOR probabilities for α, β, and coil
	70–74			(α)	End of uncertain region
71–80	77	71–80	10	tx2	
	78–84			(β)	
81–91	85–93	81–91	11	α2	
92–96	94–98	92–96	5	t23	Insertion site 1 at 96
97–109	99–109	97–106	10	β3	Irregular because of conserved P98
110–113	110–112	107–113	7	t33	
114–126	113–123	114–126	13	α3	
	124–126		0	t34	
127–131	127–131	127–132	6	β4	
132–135		133–135	3	t44	Good correlation with hydropathy and flexibility
136–147	132–145	136–146	11	α4	
148–150	147–148	147–150	4	t45	Insertion site 2 at 147–149
151–156	149–156	151–156	6	β5	Irregular because of partially conserved P152
157–161	157–162	157–161	5	t55	Insertion site 3 at 161
162–169	163–169	162–169	8	α5	
170–174	170–174	170–174	5	t56	
175–181	175–179	175–181	7	β6	
182–190	180–186	182–190	9	t66	Insertion site 4 at 188
191–204	187–201	191–204	14	α6	
205–211	202–211	205–211	7	t67	Insertion site 5 at 207, 208
212–214	212–216	212–216	5	β7	
215–221		217–220	4	t77	
222–234	218–230	221–231	11	α7	
	231–239	232–240	9	t78	Insertion site 6 at 232, 233; unusually low flexibility and high hydropathy values
235–238	240–245	241–245	5	β8	235–245: similar GOR probabilities for α and β , good correlation with hydropathy and flexibility
	246–248			(α)	
	249–256	246–256	11	t88	Insertion site 7 at 249–256
256–276	258–end	257–end	(20)	α8	Length given for *E. coli* sequence

*Predictions of individual sequences of Figure 1 by the methods of Garnier et al.[1] and Chou and Fasman [24, 25] are averaged at homologous sequence positions to obtain the consensus predictions of columns 1 and 2 (GOR and CHF, respectively). Information from the averaged hydropathy and flexibility profiles of Figure 2 is combined with the consensus predictions to derive a joint prediction in column 3 (JNT). Numbering is as in Figure 1 with limits given in residue numbers. The length of the joint prediction is in numbers of residues.

quence position exhibit much improved signal-to-noise ratios (T. Niermann and K. Kirschner, unpublished results).

Figure 2 shows a plot of the average values of both hydropathy and chain flexibility for the aligned α subunits. The maximally attainable extremes of both properties were arbitrarily set at ± 10 units. Where the amplitudes are large they generally vary in a reciprocal fashion. Moreover, the maximum of hydropathy and minimum of chain flexibility correlate

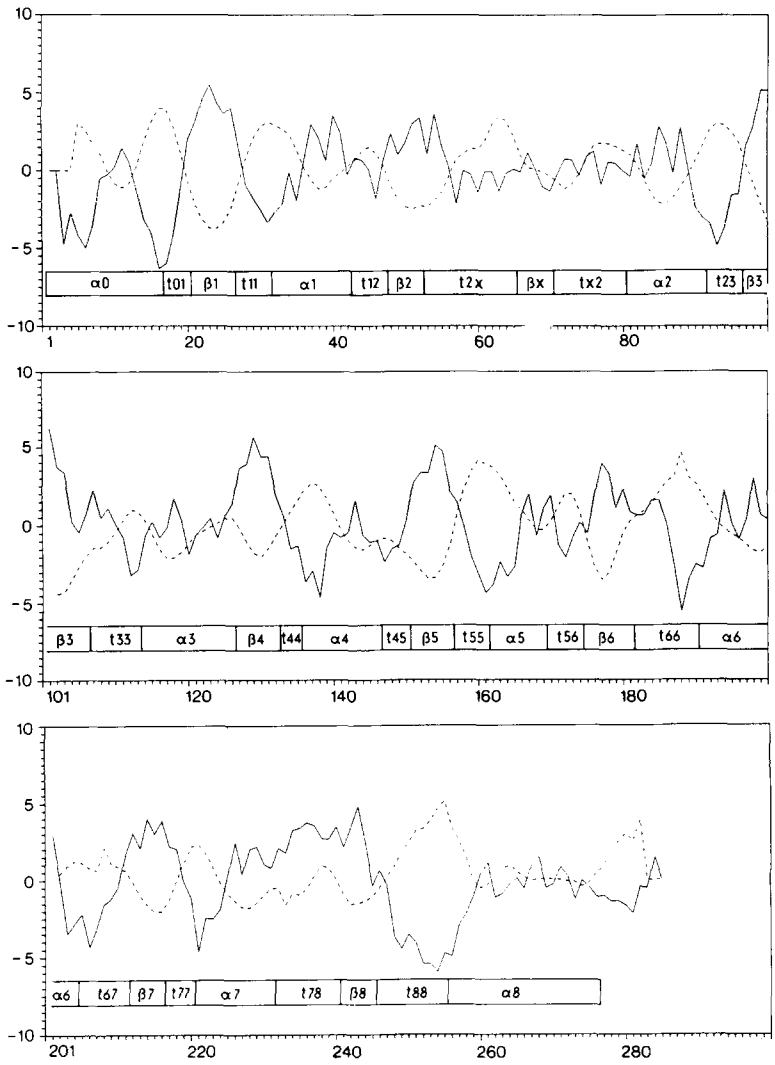

Fig. 2. Average hydropathy and flexibility profiles of the α subunit of tryptophan synthase: correlation between high hydropathy, low flexibility, and predicted β stands, and between low hydropathy, great flexibility, and predicted loop segments. The values of the hydropathy index (——) and chain flexibility (- - -) at the homologous sequence positions of Figure 1 were averaged as described under Materials and Methods. Below the profiles is shown the joint secondary structure assignment (JNT) from Table I: (αi) i th α helix; (βi) i th β strand; (tij) loop between αi and βj, (tii) loop between βi and αi.

closely with all of the predicted β strands. By contrast, minimal hydropathy and maximal flexibility values correlate with most of the predicted loop segments. Finally, the amplitudes of both hydropathy and chain flexibility are comparatively small and variable for predicted α helices.

The consensus structure predictions and both the hydropathy and flexibility profiles were used to generate a joint prediction as follows (Table I). In general, the GOR and CHF consensus predictions agree well. However, the GOR method was deemed more reliable than the CHF method because it is unambiguous and scores significantly better in predicting the secondary structure of α/β proteins.[1,28–30] Where the two methods yield contradictory predictions, e.g., between residues 48 and 52, the strongly correlated hydropathy and flexibility values were used to assign the probable secondary structure, in this particular case, β2. In other ambiguous regions, e.g., residues 232–240 and 246–256, the CHF prediction of loops is supported by the presence of insertion sites within their bounds. Individual segments predicted to be α helical frequently show the repetitive pattern expected for amphipathic helices: hydrophobic residues occurring at positions i, i+3, and i+4.[1,23] The averaging of hydropathy values at homologous sequence positions enhances these patterns for some of the predicted α helices (e.g., α1, α2, α3, and α6) but attenuates them for others (e.g., α0, α4, α5, α7, and α8).

The estimated length of each element of secondary

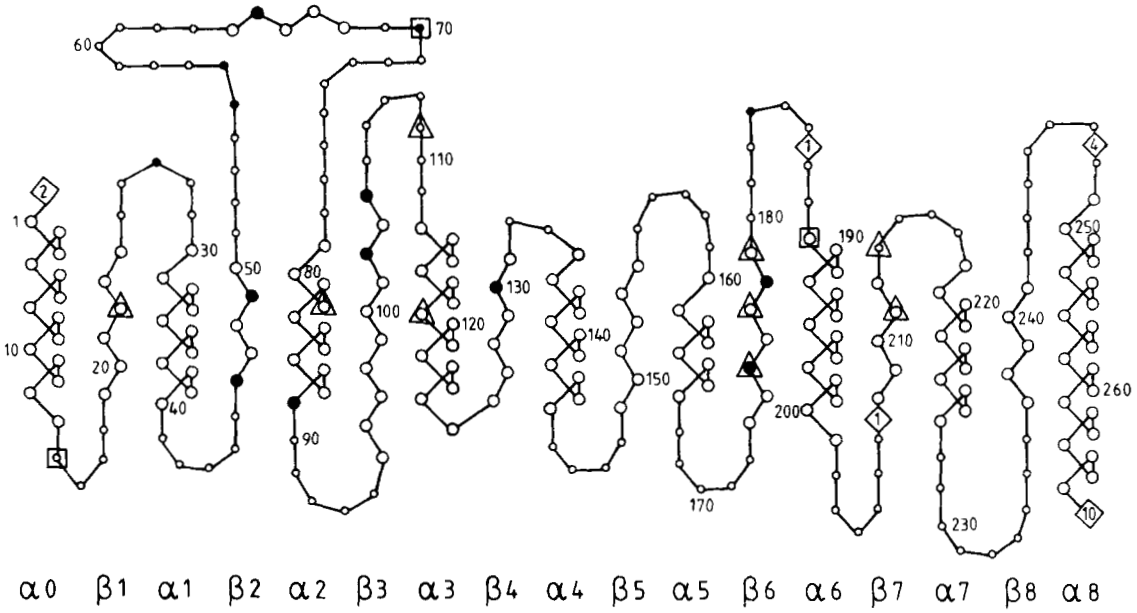

Fig. 3. Possible catalytic and interacting residues in the α subunit of tryptophan synthase: their distribution is compatible with an eightfold α/β "TIM barrel." The β strands of the joint prediction (Figs. 1, 2, Table I) are arranged sequentially in parallel from left to right with their carboxyl-termini facing upward. The connecting α helices are interspersed. Numbering is that of the *E. coli* α subunit (15). (○) Residues belonging to predicted secondary structure elements. (◦) Residues in loops. (●,•) Conserved polar residues, allowing for 1 mismatch (Fig. 1). (△) Interacting residues as described in the text. (□) Cleavage sites for trypsin. (◇) Sites that accept insertions or deletions of n residues with respect to the *E. coli* sequence.

structure is given in Table I and the elements are designated αi, βi, tii, and tij to identifiy α helices, β strands, and loops between either β and α or α and β, respectively. The joint prediction is also included in Figures 1 and 2 for comparison.

DISCUSSION

The predominant secondary structure prediction at homologous positions of the aligned amino acid sequences resolves ambiguities in the predictions of some of the individual sequences (see Figs. 4–6). The potential of this approach is well known,[1] but, as far as we are aware, has been applied only rarely.[31,32]

The novel aspect of our approach is the use of chain flexibility,[27] in conjunction with the widely used hydropathy index,[26] as a complementary, reciprocal indicator of polypeptide segment position with respect to the hydrophobic interior of the protein. The averaging of hydropathy and flexibility values at homologous positions is a new approach to enhancing the weak patterns of the individual sequences (T. Niermann and K. Kirschner, unpublished results). Moreover, a correlation between segments accepting amino acid insertions and flexible loops increases confidence in the reliability of the joint prediction. However, the true border between α helices, β strands, and loops may very somewhat from one individual sequence to another, since the predictions are based on averaged properties. Nine α helices and eight β strands are clearly predicted (Table I). They follow an alternating pattern of the "β-loop-α-loop" motif, classifying the α subunit as an α/β protein.[33] This motif is interrupted only between residues 60 and 90, a region which contains a short, extra β strand, βx. The region between residues 45 and 73 contains the most highly conserved sequence (19 of 29 residues) and probably has a well-defined but irregular structure.[34]

The eightfold repeat of the "β-loop-α-loop" motif predicted for the α subunit is consistent with the highly symmetrical folding topology of the "TIM barrell."[5–7,35] It consists of a central, twisted hyperboloid[36] or "barrel" of 8 parallel β strands with 8 α helices on the outside connecting the β strands. Figure 3 is a two-dimensional representation of the TIM barrel predicted for the *E. coli* α subunit. It is obtained by rolling out the internal cylinder of the β strands, with their respective carboxyl-termini pointing upward. Their orientation remains parallel to the barrel axis, which lies in the plane of the paper. The connecting α helices are interspersed and thus have an antiparallel orientation with respect to the β strands. The position of the β strands and α helices relative to the putative barrel equator was arbitratily assumed to by symmetrical. The rhombuses indicate sites at which homologous sequences of the α subunit

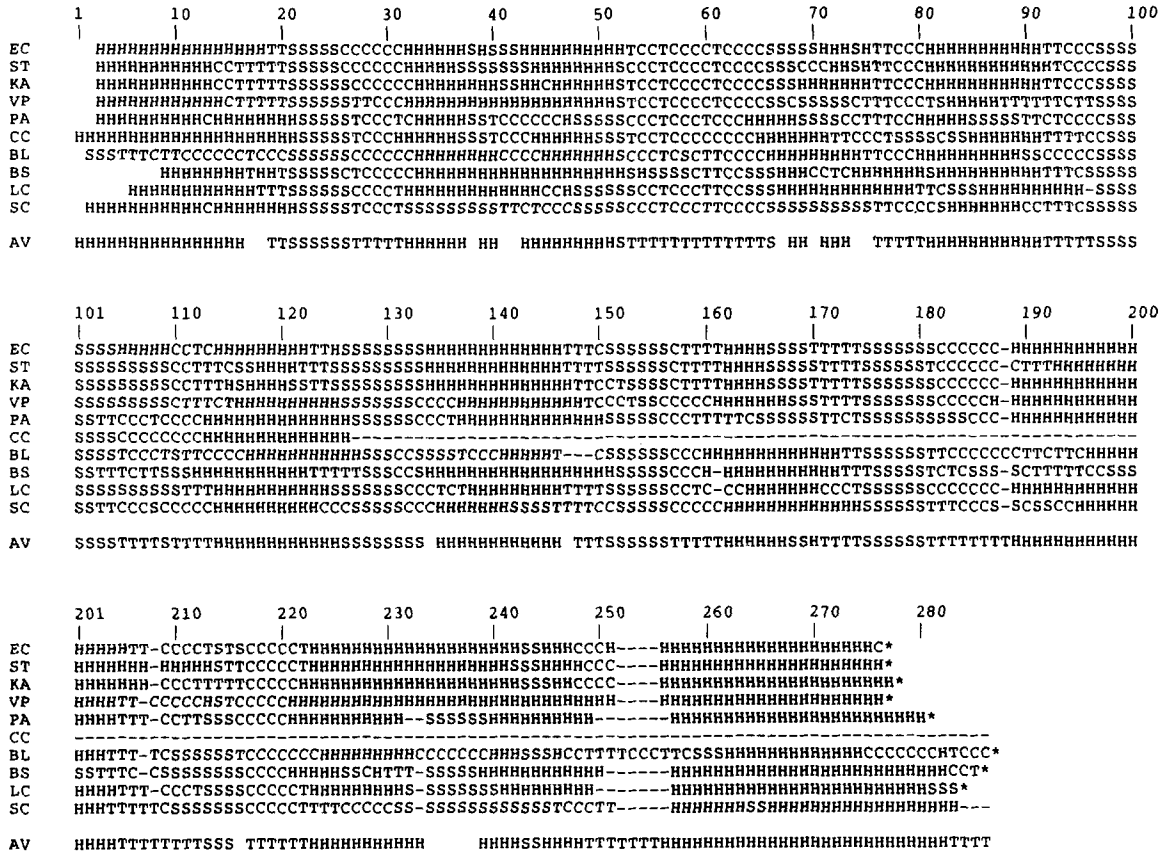

Fig. 4. Secondary structure prediction of individual sequences by the method of Garnier et al.[1] Numbering and arrangement of sequences as in Figure 1. H = α helix, S = β strand, C = coil, T = β reverse turn. AV = consensus prediction (T for "loop" and blanks for ambiguous positions).

accept insertions relative to the *E. coli* sequence (giving at each site the maximum number of extra residues shown in Fig. 1). It is likely that $\alpha 0$ is the extra appendage, because it is very short in the *B. subtilis* α subunit (Fig. 1), and more flexible than $\alpha 8$ (Fig. 2).

The following comparison of the predicted secondary structure of the α subunit to the known structures of the PRA isomerase and IGP synthase domains supports this working model. First, the lengths of the three polypeptide chains are similar (Table II). An extra N-terminal α helix ($\alpha 0$) is found in both IGP synthase (255 amino acids) and the α subunit (268 amino acids). It is missing in PRA isomerase (197 amino acids), which is a pure α/β barrel. Second, the average lengths of the predicted α and β elements of the α subunit are very similar to those of the other other two bona fide α/β barrel proteins, and to the average lengths from 14 other $\beta/\alpha/\beta$ proteins of known three-dimensional structure.[23] Moreover, as expected from cylindrical symmetry, the lengths of the α and β segments vary randomly about the mean.

Third, because the most highly conserved segment of the α subunit of *E. coli* (residues 49–65) lies between $\beta 2$ and $\alpha 2$, the active site appears to be located at the carboxyl-termini of the β strands. The active sites of triosephosphate isomerase,[6] ribulose-1,5-bisphosphate carboxylase/oxygenase,[7] glycolate oxidase,[6] PRA isomerase, and IGP synthase [5] have been found at this end of the α/β barrel. As expected, most of the additional conserved acidic or hydrogen-bonding residues are located at either the carboxyl-termini of β strands or in the adjacent turns between β strands and α helices: ($\beta 2$) E49; (t2x) S55, D56, D60; (βx) Q65; ($\beta 3$) Y102, N104, ($\beta 4$) D130, ($\beta 6$) S178; and (t66) T183 (Fig. 3). These residues are therefore good candidates *a priori* for a role in catalysis. Moreover, the active site region of the α subunit must make intimate contact with the β subunit to generate the

H:	138	72	135	136	146	179	194	196	194	181	171	160	135	122	91	86	54	45	20	19
S:	-52	0	-26	-61	-73	-85	-82	-74	-52	-54	-63	-92	-101	-119	-117	-124	-125	-78	-55	-5
T:	-81	-38	-104	-70	-33	-53	-54	-73	-87	-83	-85	-64	-26	2	19	64	93	65	62	57
C:	-68	-16	-9	-50	-70	-69	-72	-57	-72	-82	-47	3	44	62	59	61	71	36	46	50
1.	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	T	T	T	T
H:	5	-31	-15	-50	-73	-102	-144	-187	-182	-96	-39	62	106	127	131	125	132	116	108	84
S:	69	128	187	201	140	60	-36	-99	-132	-139	-136	-89	-41	-3	32	34	61	68	67	68
T:	-37	-145	-49	18	-17	23	92	87	-5	54	63	-27	-47	-61	-56	-63	-39	-43	-48	-44
C:	14	-8	-31	-44	-22	27	91	137	152	148	126	24	-30	-53	-25	-30	-41	-23	-17	-18
21.	S	S	S	S	S	S	T	C	C	C	C	H	H	H	H	H	H	H	H	H
H:	56	50	57	52	63	64	113	124	142	119	67	3	-75	-100	-68	-105	-116	-156	-107	-103
S:	67	42	-21	-71	-104	-120	-71	-32	24	68	99	60	17	-11	-14	-8	12	21	3	-54
T:	-57	-90	-39	10	2	-21	-75	-110	-135	-128	-76	-7	36	-45	55	135	7	-70	70	156
C:	3	17	21	33	45	37	24	-20	-41	-60	-86	-43	33	65	95	80	64	56	100	149
41.	S	H	H	H	H	H	H	H	H	H	S	S	T	C	C	T	C	C	C	T
H:	-66	-72	-105	-8	31	64	76	64	66	58	73	83	90	75	47	-2	-54	-51	-41	13
S:	-88	-73	-33	13	44	70	84	75	75	68	59	69	52	54	15	-17	-44	-31	-21	-10
T:	135	72	-55	21	27	-63	-70	-40	-27	-34	-18	23	35	58	78	77	95	32	-58	-28
C:	117	96	87	79	69	9	-26	-3	13	42	45	42	53	67	71	73	74	75	74	64
61.	T	C	C	C	C	S	S	S	S	S	H	H	H	H	T	T	T	C	C	C
H:	51	77	119	147	187	214	215	207	199	166	114	33	-44	-112	-111	-144	-130	-54	-26	-7
S:	3	20	39	51	29	33	4	-31	-44	-58	-70	-76	-77	-67	-42	-6	52	112	161	197
T:	7	-42	-69	-95	-93	-88	-83	-52	-36	2	46	104	105	58	91	39	-159	-40	5	-73
C:	28	-22	-41	-46	-71	-62	-47	-21	-4	8	21	58	91	110	103	83	13	-20	-31	-58
81.	H	H	H	H	H	H	H	H	H	H	H	T	T	C	C	C	S	S	S	S
H:	-6	-20	-41	-64	-84	-92	-59	-51	-63	-82	-72	-62	1	49	92	127	164	184	187	172
S:	188	164	114	103	61	41	55	60	51	38	-5	-19	-44	-43	-16	19	3	-37	-70	-108
T:	-56	-16	25	45	23	-31	-14	23	23	38	58	84	59	8	-11	-40	-72	-63	-46	-22
C:	-25	4	22	19	32	41	22	30	46	77	72	72	60	20	-16	-74	-70	-42	-50	-39
101.	S	S	S	S	S	C	S	S	S	C	C	T	C	H	H	H	H	H	H	H
H:	147	121	91	53	52	27	-18	-35	-50	-80	-76	-92	-97	-16	44	81	103	115	142	194
S:	-93	-82	-85	-90	-36	9	67	130	155	120	87	34	11	-26	-70	-95	-115	-127	-118	-110
T:	-10	8	25	30	-2	-17	-31	-56	-102	-114	-40	-48	-213	-80	15	-17	-11	-8	-46	-5
C:	-34	-21	17	18	-4	-17	-17	-41	-38	-23	-3	36	46	60	64	14	-24	5	10	12
121.	H	H	H	H	H	H	S	S	S	S	S	C	C	C	C	H	H	H	H	H
H:	211	222	216	199	160	114	84	40	-9	-24	-59	-77	-112	-122	-168	-203	-196	-138	-90	-30
S:	-87	-63	-54	-45	-48	-68	-69	-49	-36	24	77	141	190	192	160	103	21	-53	-115	-134
T:	-20	-88	-82	-35	-9	22	68	80	99	46	-11	48	40	31	42	-55	-46	129	134	53
C:	-18	-45	-45	-24	4	17	34	47	54	47	41	31	16	3	38	70	121	149	146	90
141.	H	H	H	H	H	H	H	T	T	C	S	S	S	S	S	C	C	C	C	C
H:	14	63	96	118	122	126	101	81	31	-21	-64	-94	-109	-133	-117	-111	-111	-109	-109	-107
S:	-126	-90	-37	8	46	67	77	76	59	39	7	-25	-25	10	97	152	192	211	176	121
T:	56	27	-8	-55	-65	-37	-36	-47	-24	29	83	132	137	99	40	-12	-21	-21	8	46
C:	58	24	-12	-18	-15	-29	-27	-5	31	47	78	104	83	56	5	-5	-21	-11	13	71
161.	C	H	H	H	H	H	H	H	S	C	T	T	T	T	S	S	S	S	S	S
H:	-107	-110	-130	-105	-97	-90	-47	-57	3	27	52	71	82	90	122	151	167	191	210	229
S:	85	42	20	21	12	-5	2	0	-20	-46	-40	-71	-65	-83	-99	-65	-63	-49	-22	-15
T:	81	74	72	30	-21	-22	-36	-22	-48	-1	14	21	-9	-88	-43	-61	-110	-128	-139	-141
C:	58	76	93	114	95	85	39	71	17	39	29	20	35	26	13	-8	-39	-61	-84	-93
181.	S	C	C	C	C	C	C	C	C	C	H	H	H	H	H	H	H	H	H	H
H:	221	186	147	95	40	-19	-49	-133	-106	-108	-54	-25	-15	-31	-94	-99	-142	-118	-106	-71
S:	-21	-16	-12	-22	-39	-29	-34	-21	-14	25	41	51	88	94	84	95	45	31	1	-50
T:	-96	-45	-14	32	62	84	82	80	-18	-138	-23	41	1	18	85	101	84	82	11	-43
C:	-82	-62	-34	-9	15	42	54	62	45	44	47	44	14	15	67	93	136	163	168	134
201.	H	H	H	H	T	T	T	T	C	C	C	S	S	S	T	T	C	C	C	C
H:	13	56	92	103	124	144	151	144	132	107	72	80	96	14	30	31	26	-3	31	64
S:	-79	-95	-95	-37	28	79	82	52	3	-18	-22	-18	-41	12	47	62	34	6	-1	10
T:	27	31	-29	-52	-57	-68	-71	-65	-44	-35	-28	-37	-71	-43	-45	-63	-49	-32	-49	-93
C:	87	30	-11	-10	-7	-23	-17	-5	-1	9	12	1	23	-3	11	7	21	-9	-3	-21
221.	C	H	H	H	H	H	H	H	H	H	H	H	H	H	S	S	S	S	H	H
H:	77	87	98	113	115	106	94	82	74	81	46	-114	-167	-151	-153	119	136	158	166	196
S:	42	58	98	105	76	-12	-81	-110	-141	-172	-184	-145	-170	-198	-191	-165	-101	-71	-25	-7
T:	-106	-116	-90	-108	-108	-71	-23	-21	20	29	-17	173	87	151	97	45	36	-66	-109	-103
C:	-26	-47	-56	-83	-69	-47	1	26	32	37	60	131	186	213	191	48	4	-34	-43	-43
241.	H	H	S	H	H	H	H	H	H	H	C	T	C	C	C	H	H	H	H	H
H:	196	191	187	174	185	186	179	168	177	170	182	201	213	197	176	155	140	116	98	85
S:	-8	-28	-40	-45	-46	-37	-34	-26	-43	-62	-69	-94	-95	-83	-77	-83	-85	-105	-96	-86
T:	-65	-67	-59	-38	-50	-72	-91	-98	-88	-74	-83	-75	-68	-75	-57	-42	-48	-34	-14	-13
C:	-37	-36	-36	-43	-44	-63	-46	-35	-32	-28	-46	-31	-20	-31	-33	-15	-20	-18	6	-12
261.	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
H:	81	68	57	15	-1	72														
S:	-58	-53	-36	-36	16	42														
T:	-30	-23	8	42	73	74														
C:	-6	-26	-7	62	69	198														
281.	H	H	H	C	T	C														

Fig. 5. Averaged secondary structure prediction by the method of Garnier et al. [1] Numbering and arrangement of sequences as in Figure 1. H = α helix, S = β strand, C = coil, T = β reverse turn, AV = consensus prediction (cf. column GOR in Table I).

composite active site.[8] This requirement rationalizes the presence of many additional highly conserved but noncatalytic amino acid residues in the loops joining the carboxyl-termini of β strands to the subsequent α helices (Fig. 1).

The hydropathy maxima (flexibility minima) are generally centered within the limits of the β strands (cf. Fig. 2). Whereas only 4 out of 8 loops tij, which connect the carboxyl-termini of α helices with subsequent β strands, are flexible and have flexibility max-

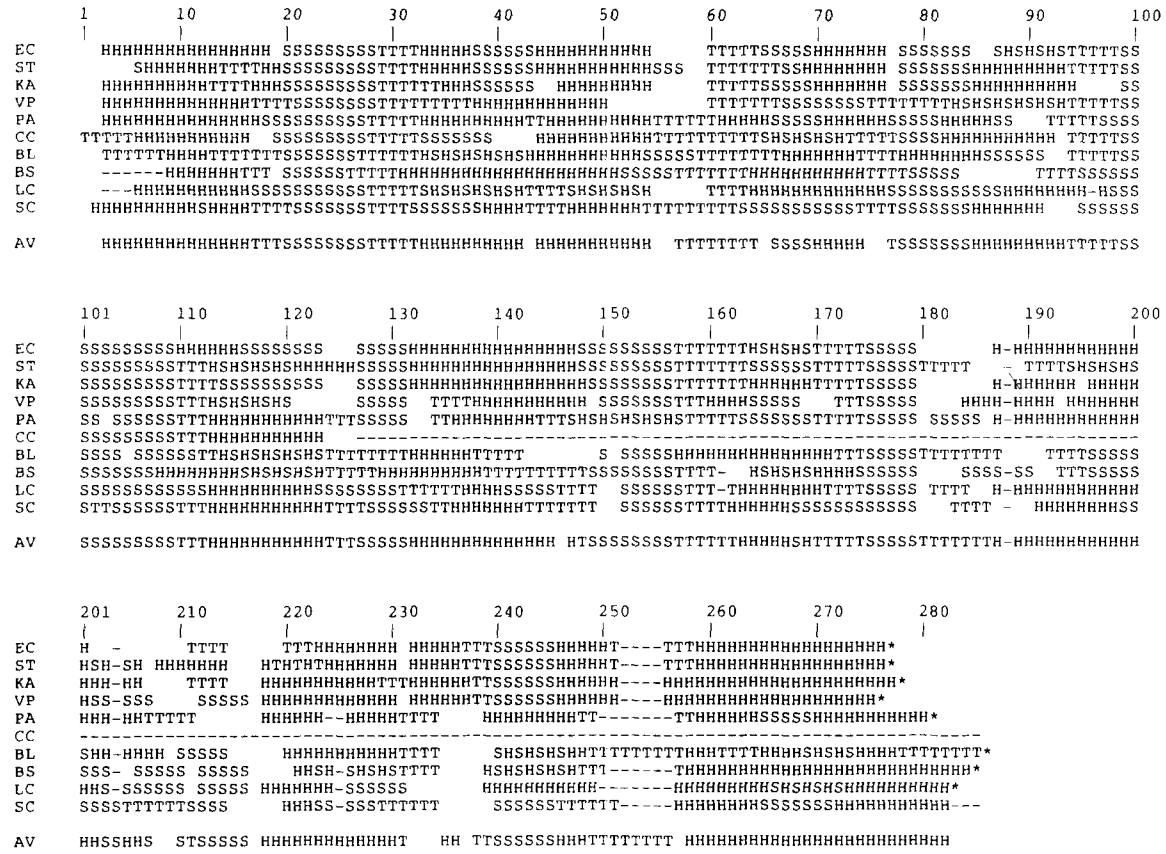

Fig. 6. Secondary structure prediction of individual sequences by the method of Chou and Fasman. [24,25] Numbering and arrangement of sequences as in Figure 1. H = α helix, S = β strand, T = β turn, blank space = no prediction, AV = consensus prediction (T for "loop" and blanks for ambiguous positions, cf. column CHF in Table I).

TABLE II. Comparison of Secondary Structure Composition and Average Lengths of Three Eight-fold α/β Barrels

Protein	IGP synthase*	PRA isomerase*	TSase α†	Average‡
Residue limits	1–255	256–452	1–268	—
Total amino acids	255	197	268	—
%α	40	43	43	42
%β	22	29	21	21
Average length α (residues)	11.2	10.5	12.7	11.9
Average length β (residues)	7.1	7.3	6.1	5.4

*Known structure.[5]
†Joint secondary structure prediction (Table I).
‡Observed average compositions and lengths from 14 $\beta\alpha\beta$ proteins.[23]

ima (hydropathy minima) centered within their limits, most of the loops tii are flexible. Moreover the maxima of flexibility (minima of hydropathy) are generally located either at the junction between tii and αi or within the N-terminus of αi. These flexible loops may have an important role in the local conformation changes that are induced by substrates,[37] in the gross conformation changes induced by binding of the β_2 subunit,[38] or both.

Independent evidence from the effects of proteolysis, chemical modification, and mutagenesis on the two functions of the α subunit (catalysis of IGP aldolysis and activation of the β_2 subunit) is also consistent with an α/β barrel structure for the α subunit of

tryptophan synthase. The major and minor sites of proteolysis of the *E. coli* protein by trypsin[39] (E.W. Miles, personal communication; see squares in Fig. 3) are either in external loops or at the end of a secondary structural segment: (t01) K15, (tx2) R70, and (α6) R188. Since R188 is accessible to trypsin even in the $\alpha_2\beta_2$ complex it cannot be completely buried in the contact region between the α and β protomers.

Indolepropanol phosphate bound to the α subunit of *E. coli* typtophan synthase protects R179, which is located at the carboxyl-terminus of β6, from chemical modification.[40] Although R179 is not conserved, it is possible that the strictly conserved S178 is responsible for binding the substrate's phosphate residue and that bound indolepropanol phosphate prevents the reaction of R179 with phenylglyoxal. A bifunctional sulfhydryl reagent crosslinks C81 with C118.[41] The triangles at positions 81 and 118 in Figure 3 show that these residues would be spatially juxtaposed in the adjacent external α helices α2 and α3. Finally, the β_2 subunit protects K109, which lies in a loop between β3 and α3, from chemical modification,[42] suggesting that its ϵ-amino group is buried within the contact region between the α and β_2 subunits.

The analysis of Trp^- mutants can suggest a catalytic role for certain conserved amino acid residues. For example, all of the missense mutants at position 49 (Glu in the wild type) are inactive in the IGP aldolysis reaction but can still activate the β_2 subunit.[8,43] Three positions have been implicated by mutational analysis as sites where modification drastically alters subunit interaction but does not affect the IGP aldolysis activity of the *E. coli* α subunit.[11] All appear to be highly conserved during evolution. They are P28 and S33, located near the amino-terminus of α1, and G51 at the carboxyl-end of the β2 strand. These positions are all compatible with the assumed location of the combining site of the α subunit for the β_2 subunit near the carboxyl-termini of the β strands.

The properties of single and double mutants indicate that a structural interaction must occur between F22 (β1) and G211 (β7), Y175 (β6) and G211 (β7), and L177 (β6) and G213 (β7).[8,44] The triangles in Figure 3 show how the interaction of these residues can be rationalized by assuming an eightfold α/β barrel structure. For example, in the intact β-barrel strand β1, which contains F22, is the next nearest neighboring strand to β7, which contains G211. Residue pairs 175/211 and 177/213 occupy equivalent positions on neighboring β strands, whose relative position with respect to the equator of the barrel is, of course, unknown. It is clear that such interactions can occur between hydrophobic residues on β strands that make up the central core of an α/β barrel.[5–7, 35]

The only evidence against the α subunit having an eight-fold α/β barrel structure seems to be the two-step reversible unfolding of the molecule.[43,45,46] The two tryptic fragments,[37] which correspond to two autonomously folding structural domains of the *E. coli* α subunit (α-1: residues 1–188, and α-2: residues 189–268), seem to be responsible for the two-step unfolding behavior, α-2 unfolding first followed by α-1.[45] These domains have even been assumed to be approximately spherical in a model of the structure of tryptophan synthase at low resolution. Although this model interprets successfully both fluorescence and hydrodynamic data,[47] both X-ray[13] and neutron[14] small-angle scattering data are compatible with a compact form of the α subunit.

Carbonic anhydrase is monomeric and also consists of a single structural and functional domain.[48] Its unfolding transition monitored by optical rotatory dispersion is broad and slightly asymmetric.[49] However, kinetic studies of unfolding and refolding have identified two partially folded intermediates.[50] Therefore, it is not safe to conclude that proteins consisting of single domains must unfold in a one-step manner, i.e., without passing through partially unfolded intermediates.

Improved methods of multiple sequence alignment,[51,52] more sophisticated prediction algorithms,[30,53,54] and the analysis of even larger sets of homologous sequences will probably improve the reliability of prediction further. The more reliable approach to the prediction of secondary structure presented here should be generally useful for interpreting low-resolution electron density data, however, and for discerning possible catalytic residues.

While this paper was in preparation, E.W. Miles and D.R. Davies informed us personally that, on the basis of preliminary X-ray crystallographic analysis, the α subunit of tryptophan synthase from *S. typhimurium* does have an eight-fold α/β barrel topology with and extra N-terminal α helix.

ACKNOWLEDGMENTS

We thank F. Imamoto, C. Ross, and M. Winkler for providing unpublished data on protein sequences and D.R. Davies and E.W. Miles for information on the overall folding topology of the *S. typhimurium* tryptophan synthase complex prior to publication.

Part of this work was supported by grant 3.255-1.85 of the Swiss National Science Foundation (K.K.) and NIH grant AI20279 (I.P.C.). Computer resources used to carry out the Chou-Fasman analyses were provided by the NIH sponsored BIONET National Computer Resource for Molecular Biology (grant 1 U41 RR-01685-03).

NOTE ADDED IN PROOF

The predicted assignments of major secondary structural elements to the amino acid sequence of the α subunit agree very well with the preliminary results of x-ray crystallography. The sequences of all 9 α helices and of 6 of the 8 β strands assigned by the two methods share common overlapping residues. The assignments are essentially identical in several cases.

The discrepant assignments are in $\beta 7$ and $\beta 8$ (C.C. Hyde, E.W. Miles and D.R. Davies, personal communication).

REFERENCES

1. Garnier J., Osguthorpe, D.J. Robson, B.: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
2. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
3. McQuade, J.F. III, Creighton, T.E: Purification and comparison of the N-(5′-phosphoribosyl)anthranilic acid isomerase/indoleglycerol phosphate synthetase of tryptophan biosynthesis from three species of Enterobacteriaceae. Eur. J. Biochem. 16:199–207, 1970.
4. Krischner, K., Szadkowski, H., Henschen, A., Lottspeich, F.: Limited proteolysis of N-(5′-phosphoribosyl)anthranilate isomerase:indoleglycerol phosphate synthase from *Escherichia coli* yields two different, enzymatically active, functional domains. J. Mol. Biol. 143:395–409, 1980.
5. Priestle, J.P., Grütter, M.G., White, J.L., Vincent, M.G. Kania, M., Wilson, E., Jardetzky, T.S., Kirschner, K., Jansonius, J.: Three-dimensional structure of the bifunctional enzyme phosphoribosyl anthranilate isomerase:indoleglycerol phosphate synthase from *Escherichia coli*. Proc. Natl. Acad. Sci. USA (in press).
6. Muirhead, H.: Triosephosphate isomerase, pyruvate kinase and other α/β barrel enzymes. Trends Biochem. Sci. 8:326–329, 1983.
7. Schneider, G., Lindqvist, Y., Bränden, C.-I., and Lorimer, G.: Three-dimensional structure of ribulose-1, 5-bisphosphate carboxylase/oxygenase from *Rhodospirillum rubrum* at 2.9 A resolution. EMBO J. 5:3409–3415, 1986.
8. Yanofsky, C., Crawford, I.P.: Tryptophan synthetase. In: "The Enzymes," 3rd Ed., Vol. 7. Boyer, P.D., ed. New York: Academic Press, Inc., 1972:1–37; Miles, E.W.: Tryptophan synthase: Structure, function and subunit interaction. Adv. Enzymol. 49:127–186, 1979.
9. Yanofsky, C., Crawford, I.P.: The tryptophan operon. In: "*Escherichia coli* and *Salmonella typhimurium:* Cellular and Molecular Biology," Vol. 2, Chapter 90. Washington, D.C.: ASM Publications, 1987: 1453–1472.
10. Hütter, R., Niederberger, P., DeMoss, J.A.: Tryptophan biosynthetic genes in eukaryotic microorganisms. Annu. Rev. Microbiol. 40:55–77, 1986.
11. Milton, D.L., Napier, M.L., Meyers, R.M., Hardman, J.K.: *In vitro* mutagenesis and over-expression of *E. coli trpA* and *trpAB* genes and partial characterization of the resultant tryptophan synthase mutant α-subunits. J. Biol. Chem. 261:16604–16615, 1986.
12. Ahmed, S.A., Miles, E.W., and Davies, D.R.: Crystallization and preliminary X-ray crystallographic data of the tryptophan synthase $\alpha_2\beta_2$ complex from *Salmonella typhimurium*. J. Biol. Chem. 260:3716–3718. 1985.
13. Wilhelm, P., Pilz, I., Lane, A.N., Kirschner, K.: Small-angle X-ray scattering studies of tryptophan synthase from *Escherichia coli* and its α and β_2 subunits. Eur. J. Biochem. 129:51–56, 1982.
14. Ibel, K., May, R.P., Kirschner, K., Lane, A.N., Szadkowski, H., Dauvergne, M.-T., Zulauf, M.: The domain structure of tryptophan synthase, a neutron scattering study. Eur. J. Biochem. 151:505–514, 1985.
15. Nichols, B.P., Yanofsky, C.: Nucleotide sequences of *trpA* of *Salmonella typhimurium* and *Escherichia coli:* An evolutionary comparison. Proc. Natl. Sci. USA 76:5244–5248, 1979.
16. Hadero, A., Crawford, I.P.: Nucleotide sequence of the genes of tryptophan synthase from *Pseudomonas aeruginosa*. Mol. Biol. Evol. 3:191–204, 1986.
17. Henner, D.J., Band, L., Shimotsu, H.: Nucleotide sequence of the *Bacillus subtilis* tryptophan operon. Gene 34:169–177, 1985.
18. Zalkin, H., Yanofsky, C.: Yeast TRP5: Structure, function, regulation. J. Biol. Chem. 257:1491–1500, 1982.
19. Nichols, B.P., Blumenberg, M., Yanofsky, C.: Comparison of the nucleotide sequence of *trpA* and sequences immediately beyond the *trp* operon of *Klebsiella aerogenes, Salmonella typhimurium* and *Escherichia coli*. Nucleic Acids Res. 9:1743–1755, 1981.
20. Matsui, K., Sano, K., Ohtsubo, E.: Complete nucleotide and deduced amino acid sequences of the *Brevibacterium lactofermentum* tryptophan operon. Nucleic Acids Res. 14:10113–10114, 1986.
21. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453, 1970.
22. Dayhoff, M.O., Barker, W.C., Hunt, L.T.: Establishing homologies in protein sequences. Methods Enzymol. 91:524–545, 1983.
23. Taylor, W.R., Thornton, J.M.: Recognition of super-secondary structure in proteins. J. Mol. Biol. 173:487–514, 1984.
24. Chou, P.Y., Fasman, G.D.: Prediction of protein conformation. Biochemistry 13:222–245, 1974.
25. Chou, P.Y., Fasman, G.D. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. 47:54–148, 1978.
26. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105–132, 1982.
27. Karplus, P.A., Schulz, G.E.: Prediction of chain flexibility in proteins. Naturwissenschaften 72:212–213, 1985.
28. Kabsch, W., Sander, W.: How good are predictions of protein secondary structure? FEBS Lett. 155:179–183, 1983.
29. Nishikawa, K.: Assessment of secondary structure prediction of proteins: Comparison of computerized Chou-Fasman method with others. Biochim. Biophys. Acta 748:285–299, 1983.
30. Sweet, R.M.: Evolutionary similarity among peptide segments is a basis for prediction of protein folding. Biopolymers 25:1565–1577, 1986.
31. Lenstra, J.A., Hofsteenge, J., Beintema, J.J.: Invariant features of the structure of pancreatic ribonuclease: A test of different predictive methods. J. Mol. Biol. 109:185–190, 1977.
32. Duncan, T.M., Parsonage, D., Senior, A.E.: Structure of the nucleotide-binding domain in the β-subunit of *Escherichia coli* F_1-ATPase. FEBS Lett. 208:1–6, 1986.
33. Levitt, M., Cothia, C.: Structural patterns in globular proteins. Nature 261:552–558, 1976.
34. Leszczynski, J.F., Rose, G.D.: Loops in globular proteins: A novel category of secondary structure. Science 234:849–855, 1986.
35. Lindqvist, Y., Bränden, C.-I.: Structure of glycolate oxidase from spinach. Proc. Natl. Acad. Sci. USA 82:6855–6859, 1985.
36. Novotny, J., Bruccoleri, R.E., Newell, J.: Twisted hyperboloid (strophoid) as a model of β-barrels in proteins. J. Mol. Biol. 177:567–573, 1984.
37. Weischet, W.O., Kirschner, K.: Steady-state kinetic study of the synthesis of indoleglycerol phosphate catalyzed by the α subunit of tryptophan synthase from *Escherichia coli:* Comparison with the $\alpha_2\beta_2$ complex. Eur. J. Biochem. 65:375–385, 1976.
38. Lane, A.N., Paul, C.H., Kirschner, K.: The mechanism of self-assembly of the multienzyme complex tryptophan synthase from *Escherichia coli*. EMBO J. 3:279–287, 1984.
39. Higgins, W., Fairwell, T., Miles, E.W.: An active proteolytic derivative of the α subunit of tryptophan synthase: Identification of the site of cleavage and characterization of the fragments. Biochemistry 18:4827–4835, 1979.
40. Eun, H.-M., Miles, E.W. The reaction of phenylglyoxal with arginines and cysteines in the α subunit of tryptophan synthase. Biochemistry 23:6484–6491, 1984.
41. Freedberg, W.B., Hardman, J.K.: Structural and functional roles of the cysteine residues in the α subunit of *Escherichia coli* tryptophan synthetase. III. Studies with the bifunctional sulfhydryl reagent bismaleimidomethyl ether. J. Biol. Chem. 246:1439–1448, 1971.
42. Miles, E.W., Fairwell, T.: The tryptophan synthase $\alpha_2\beta_2$ complex: A comparison of the reactivity of amino groups in the α and β_2 subunits and in the complex by differential labelling studies. Arch. Biochem. Biophys. 230:430–439, 1984.
43. Ogasahara, K., Tsunasawa, S., Soda, Y., Yutani, K., Sugino, Y.: Effect of single amino acid substitutions on the protease susceptibility of tryptophan synthase α subunit. Eur. J. Biochem. 150:17–21, 1985.
44. Murgola, E.J., Hijazi, K.A.: Selection for new codons corresponding to position 234 of the tryptophan synthetase α chain of *Escherichia coli*. Mol. Gen. Genet. 191:132–137, 1983.

45. Miles, E.W., Yutani, K., Ogasahara, K.: Guanidine hydrochloride induced unfolding of the α subunit of tryptophan synthase and of the two α proteolytic fragments. Evidence for stepwise unfolding of the two α domains. Biochemistry 21:2586–2592, 1982.
46. Beasty, A.M., Matthews, C.R.: Characterization of an early intermediate in the folding of the α subunit of tryptophan synthase by hydrogen exchange measurements. Biochemistry 24:3547–3553, 1985.
47. Lane, A.N., Kirschner, K.: The quaternary structure of tryptophan synthase from *Escherichia coli*: Fluorescence and hydrodynamic studies. Eur. J. Biochem. 129:675–684, 1983.
48. Liljas, A., Kannan, K.K., Bergsten, P.-C., Waara, I., Fridborg, K., Strandberg, B., Carlblom, U., Järup, L., Lövgrens, S., Petef, M.: Crystal structure of human carbonic anhydrase C. Nature New Biology 235:131–137, 1972.
49. Wong, K.-P., Tanford, C.: Denaturation of bovein carbonic anhydrase by guanidine hydrochloride. J. Biol. Chem. 248:8517–8523.
50. McCoy, L.F., Jr., Rowe, E.S., Wong, K.-P.: Multiparameter kinetic study on the unfolding and refolding of bovine carbonic anhydrase B. Biochemistry 19:4738–4743, 1980.
51. Bacon, D.J., Anderson, W.F.: Multiple sequence alignment. J. Mol. Biol. 191:153–161, 1986.
52. Johnson, M.D., Doolittle, R.F.: A method for the simultaneous alignment of three or more amino acid sequences. J. Mol. Evol. 23:267–278, 1986.
53. Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J.: Turn prediction in proteins using a pattern-matching approach. Biochemistry 25:266–275, 1986.
54. Taylor, W.R.: Identification of protein sequence homology by consensus template alignment. J. Mol. Biol. 188:233–258, 1986.