Extracting Information on Folding from the Amino Acid Sequence: Consensus Regions with Preferred Conformation in Homologous Proteins

Marianne J. Rooman‡ and Shoshana J. Wodak*

Unité de Conformation des Macromolécules Biologiques, Université Libre de Bruxelles, CP 160/16, Avenue P. Hêger, 1050 Brussels, Belgium

Received January 17, 1992; Revised Manuscript Received June 12, 1992

ABSTRACT: It is investigated whether protein segments predicted to have a well-defined conformational preference in the absence of tertiary interactions are conserved in families of homologous proteins. The prediction method follows the procedures of Rooman, M., Kocher, J.-P., and Wodak, S. (preceding paper in this issue). It uses a knowledge-based force field that incorporates only local interactions along the sequence and identifies segments whose lowest energy structure displays a sizable energy gap relative to other computed conformations. In 13 of the protein families and subfamilies considered that are sufficiently homologous to have similar 3D structures, at least one region is consistently predicted as having the same preferred conformation in virtually all family members. These regions are between 4 and 26 residues long. They are often located at chain ends and correspond primarily to segments of secondary structure heavily involved in interactions with the rest of the protein, suggesting that they could act as nuclei around which other parts of the structure would assemble. Experimental data on early folding intermediates or on protein fragments with appreciable structure in aqueous solution are available for more than half of the protein families. Comparison of our results with these data is quite favorable. They reveal that each of the experimentally identified early formed, or independently stable, substructures harbors at least one of the segments consistently predicted as having a preferred conformation by our procedure. The implications of our findings for the conservation of folding pathways in homologous proteins are discussed.

The detailed mechanism whereby information contained in the amino acid sequence leads the protein into its folded functional state is not fully understood. Possibly in part for this reason, attempts to deduce the tertiary conformation from the amino acid sequence are still without a satisfactory outcome (Levitt, 1991).

The very large number of possible conformations accessible to a polypeptide chain of even moderate length makes exhaustive surveys of conformational space in folding simulations impractical. It also excludes the possibility that natural folding proceeds via random sampling (Levinthal, 1968). Plausibly therfore, folding should follow one or more well-defined pathways, and several hypotheses have been made about their nature. A central idea of these hypotheses has been the existence of nucleation or initiation sites for folding composed of short polypeptide segments (Wetlaufer, 1973). Such segments would form their native structure early during folding and lead to the final folded state either through a diffusion collision process (Karplus & Weaver, 1976; Bashford et al., 1988) or through propagation (Wetlaufer, 1973).

Relying on the views that the hydrophobic effect provides a dominant contribution to the stability of the native structure (Kauzmann, 1959; Chothia & Janin, 1975; Dill, 1990) and that this structure preserves a record of the folding process (Richardson, 1981), attempts have been made to define folding pathways from analyses of known structures. In particular, nucleation sites and stable domains, as well as hierarchic schemes for their assembly, were defined on the basis of geometric and packing considerations (Crippen, 1978; Lesk & Rose, 1981), contact densities (Montelione & Scheraga, 1989), and buried surface areas (Rashin, 1981, 1984; Chiche et al., 1990). More elaborate approaches which incorporate in addition rough estimates of chain entropy contributions (Moult & Unger, 1991), and computer simulations using lattice models (Covell & Jernigan, 1990; Skolnick & Kolinski, 1989), have also provided valuable insight.

The most concrete promise to advance our understanding of the folding process comes from recent progress in experimental studies of folding intermediates (Dobson, 1991; Kim & Baldwin, 1990) and peptide models corresponding to protein substructures of various sizes (Oas & Kim, 1988; Staley & Kim, 1990; Dyson & Wright, 1991). Clear evidence is emerging that certain regions of the tertiary structure can adopt their native fold ahead of others, in the absence of interactions with other parts of the structure which form later. It is thus likely that their folding is dominated by local sequence features and that they play an important role in defining the folding pathway (Wright et al., 1988).

Our theoretical analyses support this view. In agreement with earlier suggestions, we have shown that secondary structure prediction methods, known to emphasize contributions from local interactions, perform much better than average in such early folding regions and on peptides shown to adopt relatively well-defined conformations in solution (Rooman & Wodak, 1991). This was further confirmed using a more powerful prediction procedure. This procedure, though still restricted to contributions from local interactions along the sequence, uses a more detailed three-dimensional representation of the protein backbone (Rooman et al., 1991). It has the important advantage of providing a very efficient means of computing any number of lowest energy conformations without performing conformational searches, which also allows one to evaluate the relative preference of computed structures from the corresponding energy values.

Applying our procedure to 69 highly resolved and well-refined protein structures [see preceding paper, Rooman et al. (1992)], we find that the predictions are most successful

‡ Chargée de Recherches at the Fonds National Belge de la Recherche Scientifique.

0006-2960/92/0431-10239$03.00/0

Table I: Families and Subfamilies of Homologous Proteins[a]

family	*N*	representative protein	PDB code	*M* (%)	SD
cytochrome *c*[a]	77	rice cytochrome *c*	1CCR	64	14
lysozyme[b]	16	human lysozyme	1LZ1	62	11
plastocyanin[b]	18	poplar leaves plasocyanin	1PCY	72	11
adenylate kinase[b]	11	porcine adenylate kinase	3ADK	56	27
thermolysin[b]	5	*Bacillus thermoproteolyticus* thermolysin	3TLN	63	17
mammalian ribonucleases[b]	38	bovine ribonuclease A	7RSA	81	9
alcohol dehydrogenase[b]	19	horse alcohol dehydrogenase	8ADH	80	14
globin[a]	481	human hemoglobin α-chain	4HHB	49	23
hemoglobin α-chain[a]	183	human hemoglobin α-chain	4HHB	74	15
hemoglobin β-chain[a]	171	human hemoglobin β-chain	4HHB	75	13
vertebrate myoglobins[a]	67	sperm whale myoglobin	1MBD	79	14
lamprey & hagfish globin[a]	5	sea lamprey hemoglobin	2LHB	79	23
legume hemoglobin[a]	13	yellow lupin leghemoglobin	2LH4	54	10
insect globin[a]	12	*Chironomous thummi thummi* hemoglobin	1ECD	50	13

[a] Column 1 lists the family name, and column 2, the number of aligned sequences (*N*). Alignments *a* are derived by the method of Bashford et al. (1987) and were obtained from Chothia (private communication). Alignments *b* are derived by the method of Lüthy et al. (1991) and were obtained from Eisenberg (private communication). The representative tertiary structure of the family is given by its full name in column 3 and by its PDB code (Bernstein et al., 1977) in column 4. The last two columns indicate respectively the average sequence identity (*M*) and the standard deviation (SD), between any sequence of the family and the sequence of the representative protein (given in %). These values are directly computed from the alignments.

for segments of the polypeptide whose lowest energy conformation displays a sizable energy gap relative to other computed structures. Such segments have well-defined conformational preferences in the absence of interactions with other parts and could play an important role in folding, by restricting the conformational freedom of the polypeptide, and by providing starting structures around which other parts can assemble. Preliminary analysis has shown indeed that they correlate well with early folding sites (Rooman et al., 1991, 1992; Kang et al., 1992). Reliable methods for predicting such segments from the amino acid sequence could thus provide valuable insight into the folding process.

Here, we extend our analysis to families of evolutionary related proteins. Such proteins may display common features at the sequence, functional, and/or structural level. Conservation of the tertiary fold seems to be of particular significance, since it is observed in proteins with very low sequence identity (Farber & Petsko, 1990; Kabsch et al., 1990). This conservation is not stringently correlated to stability properties, which vary to some extent, as a function of physiological requirements. But not much is known about how conservation of 3D structure is related to the folding process. In particular, to what extent do similar folds, which may differ significantly in detailed atomic structure as their sequences diverge (Chothia & Lesk, 1986), conserve their folding pathways? There are unfortunately very few detailed experimental data on the (un)folding properties of homologous proteins. Those are chiefly limited to highly homologous protease inhibitors and cytochrome *c*'s [see Jaenicke (1987) for review], or to closely related variants produced by chemical (Zuniga & Nall, 1983) and site-directed mutagenesis (Alber, 1989), and suggest that at least in closely related proteins essential features of the folding pathway are conserved.

To address these questions, we set out to identify segments of the polypeptide, ranging in length from 5 to 15 residues, that adopt well-defined conformations in the absence of tertiary interactions in 14 groups of evolutionarily related proteins, using a total of 1116 aligned sequences. We then investigate the pattern of conservation in the predictions of these segments within members of the same family. It is proposed that segments consistently predicted as having a preferred conformation in more than 80% of the proteins in a family represent regions that define crucial features of the folding pathway. On the other hand, weakly conserved predictions of segments with well-defined conformation may reflect the existence of variations in the folding pathway.

The results are analyzed in the context of the available information on the three-dimensional fold of these proteins, and detailed comparisons are performed, whenever possible, with experimental data on folding intermediates or on protein fragments shown to have a well-defined structure on their own. This is seen to lead to a coherent picture, which may have important implications for our understanding of the folding process, and for improving our capacity to generate folded structures from sequence.

MATERIALS AND METHODS

(*1*) *Homologous Proteins.* Eight families of homologous proteins are analyzed using two sets of sequence alignments, respectively: set 1, derived by the method of Bashford et al. (1987), and set 2, derived by that of Lüthy et al. (1991). Each protein family contains a member from a dataset of 69 highly resolved ($\leq$2.5 Å) and well-refined protein structures with low sequence homology, which are used to derive our knowledge-based force field (Rooman et al., 1991, 1992). These proteins are listed in Table I of the preceding paper (Rooman et al., 1992). The protein of known structure that is a member of both our dataset and the considered family is taken as the representative tertiary structure of the family.

Various characteristics of each family, such as the number of aligned sequences, the average sequence identity among family members, and the representative tertiary structure, are summarized in Table I. The large number and diversity of the globins required further subdividing this family into six subclasses (Bashford et al., 1987). With this subdivision, the sequence identities within the considered families and subclasses range between 50% and 81%. Proteins with this level of sequence homology are expected to adopt sufficiently similar tertiary folds (Chothia & Lesk, 1986) to warrant using a single representative tertiary structure per family or subclass.

The cytochrome *c*'s and the globins were analyzed using alignments from set 1, while for the other families, alignments from set 2 were used. To check the robustness of our procedures, the cytochromes *c*'s, present with 77 and 85 sequences in alignment sets 1 and 2, respectively, were analyzed twice and found to yield identical conclusions.

(*2*) *Prediction of Polypeptide Segments with Preferred Conformation.* The algorithm for detecting protein segments that adopt a well-defined conformation in the absence of tertiary interactions is identical to the prediction procedure presented in the preceding paper (Rooman et al., 1992) and

is entirely based on the principles and methods described previously (Rooman et al., 1991). In this procedure, backbone conformations are described as combinations of only 7 structural states: A, C, B, P, G, E, and O (one per residue). These represent allowed conformational states of the isolated dipeptide and are characterized by single values of the dihedral angles ϕ, ψ, and ω [see Figure 1 of Rooman et al. (1992) for definition]. Side chain degrees of freedom are not considered.

Using this backbone description, a potential of mean force is derived from the dataset of 69 known protein structures. This potential determines the preference of a residue at position i for any of the allowed structural states, by considering the statistical influences of single residues and residue pairs removed from i by at most 8 positions along the sequence. It thus takes into account only local interactions along the chain and ignores interactions between residues far removed along the sequence but close in space. Given a segment of sequence, ranging here in length from 5 to 15 residues, the prediction algorithm computes a series of lowest energy conformations, ranked by order of increasing energies. This computation is very fast, because conformations of individual residues are considered as independent. It can thus be applied systematically to a large number of protein sequences, something that is not feasible with methods requiring extensive conformational surveys.

From the ranked list of lowest energy conformations, the difference, or gap, in energy between the minimum energy conformation and all the others can be evaluated. An energy gap of 0.5 kcal/mol or more, between the lowest energy structure and the next one in the rank that has a significantly different fold, is taken here to define polypeptide segments that are likely to adopt well-defined preferred conformations [see Rooman et al. (1992) for details].

The algorithm described above is applied independently to all the sequences of a given protein family. A window of fixed length, corresponding to the segment length one wishes to consider, is moved along each sequence, one residue at a time, and predictions are performed for the segment delimited by the window. Since we are interested in predicting regions of the intact polypeptide that are likely to fold before tertiary interactions are established, the influence of residues along the sequence flanking each window is also included in the predictions. The predicted states at individual positions in the different sequences of a given family are aligned according to the sequence alignment and compared to the observed conformations in the representative structure. To avoid biasing the results, the representative tertiary structures of each family are in turn removed from the dataset of 69 protein crystal structures when deriving the potentials of mean force (Efron, 1982).

Of particular interest are regions along the sequence that are consistently predicted as having well-defined conformational preferences in all, or nearly all, members of a protein family. Such regions, hereafter referred to as consensus structured regions (CSR), are defined as follows: (1) They must comprise at least 4 consecutive residues predicted as part of segments with well-defined conformations in more than 80% of the proteins in a family. Since, in general, the number of proteins contributing the prediction at each position is different, a range of percentages is associated with each consensus region. (2) For each residue in the consensus region, the predicted structural states must be the same, in no less than 98% of all overlapping segments, in all family members. At this stage, only 3 structural states, (A–C), (B–P), and (G–E–O), are considered, since A and C may be regarded as local distortions of the helical structure, and B and P, of the extended one. When only one region satisfies the above criteria in a given protein family, regions identified with a consensus somewhat lower than 80% are discussed.

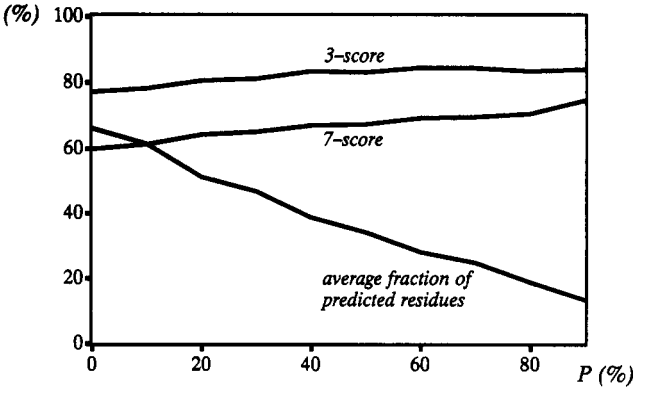

FIGURE 1: Prediction scores on 3 and 7 structural states (denoted 3- and 7-scores, respectively) and the average fraction of predicted residues, as a function of P, the minimum fraction of proteins in a family for which segments with preferred conformation map onto a given residue. All quantities are given in %, with the smallest value of P being $100/N$, where N is the number of proteins in the family.

Information on homologous sequences can also be used to improve the prediction accuracy for their common structure (Zvelebil et al., 1987; Levin & Garnier, 1988). This is achieved here by combining predictions of segments from the different sequences in the family using the following protocol. Predictions are recorded only when segments from more than a minimum fraction, P, of the family members map onto a given sequence position, and when the computed structural state at this position is the same (in the three-state classification) in no less than 98% of all the overlapping segments. The predicted structure of each recorded residue is taken to be the structural state, in the seven-state classification, that occurs most often in all the identified segments mapping onto the residue. The *prediction score* is then defined as the percent of correctly predicted structural states (in either the 3- or 7-state classification), in the representative protein of a given family, ignoring unpredicted residues. The dependence of this score on the minimum fraction, P, of family members contributing overlapping segments is analyzed.

RESULTS

Using our procedure, segments with well-defined conformational preferences were identified in a total of 665 protein sequences, grouped into 8 evolutionarily related families, or 13 subfamilies when the globin subdivision is considered (Table I).

(*1*) *Predictions of the Protein Backbone Structure.* Combining predictions from segments with well-defined conformations identified in the different sequences of each family yields a prediction for the backbone structure of the representative protein (section 2 of Materials and Methods). Figure 1 depicts how the average score of this prediction varies in our sequence sample as a function of P, the minimum fraction of family members contributing overlapping segments. Increasing P from just above 0% up to 80%, raises the score from 60% to 74% considering 7 structural states (or from 77% to 84% for 3 states). Not unexpectedly, however, this increase in accuracy is accompanied by a sharp decrease (66% to 13%) in the fraction of predicted residues. By comparison, predictions by the same method on individual sequences reach scores of 53% on 7 states (70% on 3 states) with 60% of the residues being predicted (Rooman et al., 1992). Though these

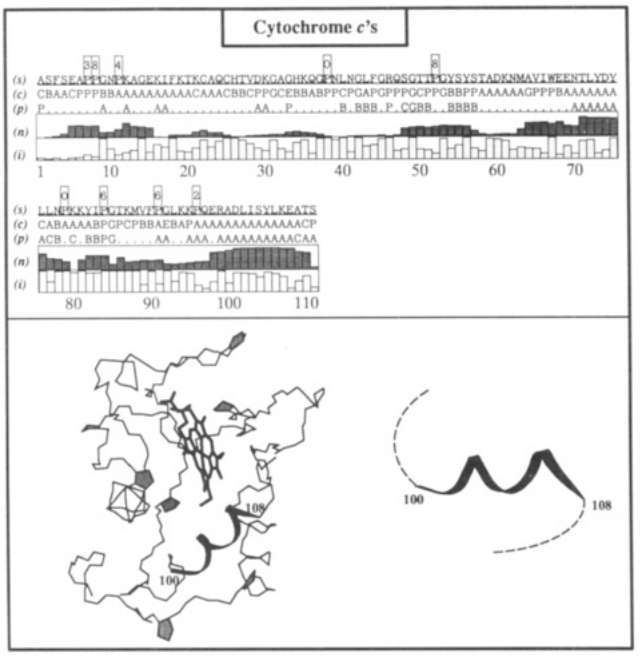

FIGURE 2: Prediction of consensus structured regions (CSR's) in the cytochrome *c* family, with rice cytochrome *c* (1CCR) as representative. The heme is indicated in bold lines. All predictions are performed taking the contributions from flanking residues into account. In rows marked by (s), the amino acid sequence of the representative protein is given in the one-letter code (underlined). The prolines are framed, and the fraction of their surface area accessible to solvent in the native structure is indicated above. The latter is defined as the ratio of the accessible surface area obtained by the algorithm SurVol (Alard, 1991) and the maximum accessible surface of prolines as calculated by Rose (1985) (146.8 Å). This ratio is then multiplied by 10, to obtain a one-digit value. A value of 9 is assigned to prolines completely accessible to solvent. The rows marked by (c) display the structural state of each residue observed in the representative protein. "X" indicates that the residue ϕ–ψ–ω angles are undetermined, due to incomplete atomic coordinates, and "Y" that the ϕ–ψ–ω values fall outside the 7 allowed regions. The rows marked by (p) display the predicted structural states when all segments mapping onto a given residue yield identical predictions in all proteins of the family (considering only 3 structural states), up to 2% allowed discrepancies (section 2 of Materials and Methods). The structural state indicated is the one that occurs most often among the 7 states. For unpredicted residues, a dot is indicated instead. The charts marked by (n) display, for each residue, the percent of proteins of the family in which at least one segment with preferred conformation maps onto that residue. The charts marked by (i) indicate the averaged amino acid conservation of each sequence position over all pairs of proteins of the family. Gaps in the alignment are considered as nonconserved residues. The positions along the sequence appear underneath at 10-residue intervals, using the PDB numbering (Bernstein et al., 1977) of the representative protein structure of the family. In the lower half of the figure, the observed three-dimensional backbone structure of the family representative is drawn using BRUGEL (Delhaise et al., 1985). Proline side chains are included and are shown in gray. The CSR's are depicted as ribbons, with the PDB sequence number of the first and last residues indicated. In the right-hand lower corner, the predicted conformations of these regions are drawn in ribbon-like representation. These conformations have been generated from the predicted seven state structure assignments for each residue in the segment, using average bond distance and angles.

results are based only on 13 families and may thus not be entirely representative, they clearly indicate that combining information from homologous sequences improves the predictions, in agreement with previous conclusions (Zvelebil et al., 1987; Levin & Garnier, 1988).

(*2*) *Consensus Structured Regions*. Results on the identification of consensus structured regions (CSR's) in the considered protein families are summarized in Figures 2–9. As described in section 2 of Materials and Methods, CSR's are defined by 4 or more consecutive residues, which have a predicted structure in rows (p) of Figures 2–8 and 9a and for which the protein fraction in charts (n) exceeds 80%. The predicted and observed conformations of these CSR's are displayed in the lower half of Figures 2–8 and 9b, on the right- and left-hand sides, respectively.

To investigate the question of sequence conservation in CSR's, the percent of residues conserved among the proteins of the family at each position along the sequence was compiled from the available alignment and is given in charts (i) of Figures 2–8 and 9a. Analysis of the results clearly indicates that there is no obvious correlation between CSR's and sequence conservation. The CSR's display on the average the same level of sequence identity as the proteins in the family and therefore appear to be just as tolerant to sequence variations as the overall three-dimensional fold. They are thus more tolerant than regions involved in function, known to display higher than average sequence conservation.

In the following, results obtained for individual protein families are described. Whenever possible, comparisons are made with information available from experiments on the structure of early folding intermediates or on protein substructures that appear to be independently stable. With such information being available for only a few proteins, the bulk of our results should be considered as genuine predictions, to be confronted with experiments to come.

(*2.a*) *Cytochrome c's*. In this family, only 1 CSR is identified in 82–92% of the proteins (Figure 2). It comprises residues 100–108, corresponding to part of the C-terminal helix. Its predicted conformation is α-helical (A), in agreement with the structure observed in the native protein.

A second somewhat less consistently predicted region (in only 78–83% of the proteins, and thus not quite making it by our criteria) comprises residues 71–75. This region corresponds to the central helix and is also well predicted.

The two regions, taken together, represent only 14% of the cytochrome *c* polypeptide sequence. Although, in the remaining 86%, segments with well-defined conformational preferences are still identified in family members, they tend either to occur in different positions or to have different structures.

Cytochrome *c* is one of the few protein families for which experimental data on early folding intermediates is available. Stopped-flow hydrogen exchange NMR experiments on horse cytochrome *c* refolding (Roder et al., 1988) are consistent with the formation of the N- and C-terminal helices in the millisecond time scale. Stopped-flow CD spectroscopy, performed on the same system under somewhat different conditions (Kuwajima et al., 1987), suggests that, during similar time scales, essentially all helical regions are formed. Those include in addition to the terminal helices, two centrally located helices, one of which is helix [71–75].

It is rather encouraging that the only CSR identified by our procedure maps onto the C-terminal helix shown to be part of an early folding region. The fact that this region is predicted as having a well-defined helical conformation in the major fraction of the considered cytochrome *c*'s could indicate that it plays an important role in their folding. Our predictions and the experiments also seem to agree on the more ambiguous behavior of the centrally located helix [71–75]: it is less consistently identified by our procedure and appears to form early under certain experimental conditions and not others.

We see on the other hand (Figure 2) that the N-terminal helix, also observed to be part of an early folding region, is definitely not a CSR by our criteria. As argued earlier (Roder et al., 1988; Rooman & Wodak, 1991; Rooman et al., 1991),

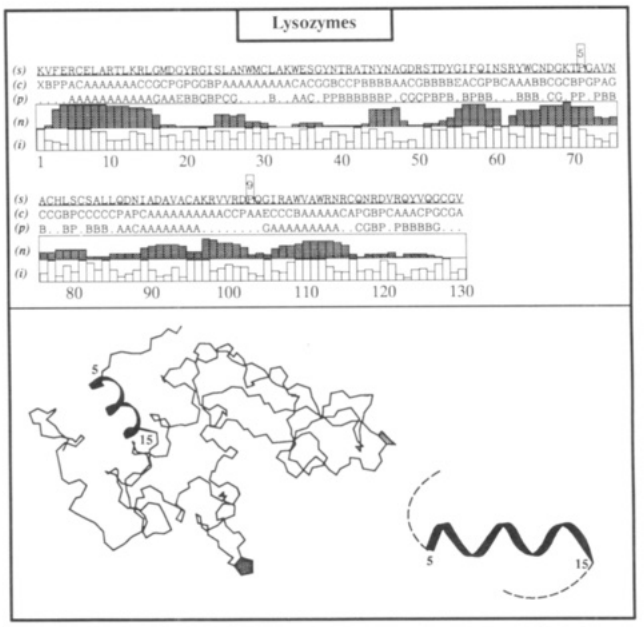

FIGURE 3: Prediction of consensus structured regions (CSR's) in the lysozyme family, with human lysozyme (1LZ1) as representative. See Figure 2 caption for details.

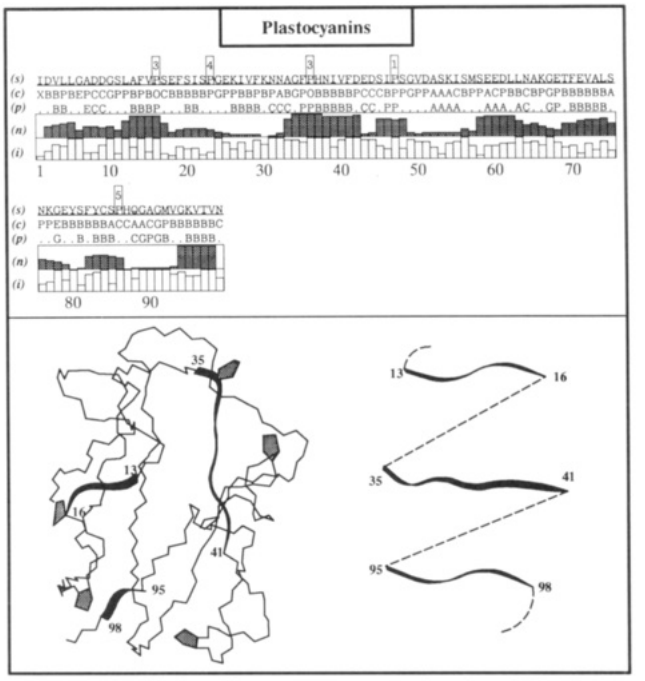

FIGURE 4: Prediction of consensus structured regions (CSR's) in the plastocyanin family, with poplar leaves plastocyanin (1PCY) as representative. See Figure 2 caption for details.

this may suggest that this helix is not sufficiently stable on its own, but that it forms upon docking to the C-terminal helix, and/or upon interactions with the heme group, with which both terminal helices are in contact in the native structure.

(*2.b*) *Lysozymes*. Proteins in this family correspond to the *c*-type lysozyme subgroup, with human lysozyme as the representative 3D structure. As shown in Figure 3, only one CSR is identified in this family, with 81–100% of the considered proteins contributing to its prediction. This region comprises residues 5–15 of the N-terminal helix and is part of an α-helical domain that contains the N- and C-terminal portions of the polypeptide. The predicted and observed structures of this CSR are in good agreement.

A second consistently predicted region, albeit with a somewhat lower consensus (in 75–94% of the considered sequences) and therefore not depicted in Figure 3, comprises residues 56–59. This segment, predicted as extended (B/P) by our procedure, belongs to a loop in the central β-domain of the native structure.

NMR-monitored hydrogen exchange of hen lysozyme folding (Miranker et al., 1991) indicates that the α-helical noncontiguous domain is formed on a 10-ms time scale, ahead of other regions. The central β-sheet domain appears to form later as its amide protons show little protection over the same time scale. However, 3 residues located in highly exchangeable loops of this domain, Trp 63, Lys 64, and Ile 78 (corresponding to Trp 64, Cys 65, and Leu 79 in human lysozyme), remain significantly protected, which was taken to suggest the presence of residual nonnative structure early during folding.

Our predictions fit well with the above description. The N-terminal helix [5–15], predicted as having a preferred helical conformation in nearly all the C-type lysozymes, belongs to the early folding helical domain of these proteins, where it interacts with 3 out of the remaining 4 helices: [24–36], [89–100], and [121–126]. The tendency of this helix to form in the absence of tertiary interactions could thus be important in initiating the folding of the entire domain. However, it is in general not the only helix with preferred conformation in this domain. Other helices with preferred conformations are detected, but they are not the same in different members of the family (see, for example, helix [89–100] and the tail of the C-terminal helix [109–11], identified in human lysozyme, but not in many other lysozymes).

Unfortunately, the folding experiments provide no direct information on the nascent conformation of the second consistently predicted region [56–59], for which predicted and observed structures differ, but suggest that residues Trp 64, Cys 65, and Leu 79, in the immediate vicinity of this region, could adopt a nonnative conformation early during folding. Experiments and predictions thus support the contention that parts of the lysozyme β-sheet domain may be misfolded at such times.

(*2.c*) *Plastocyanins*. Three CSR's are identified in this family (see Figure 4). They correspond to segments [95–98], [35–41], and [13–16].

The first one, [95–98], identified in all the considered sequences, corresponds to the tail of the C-terminal β-strand. Its predicted and observed structures agree, considering as equivalent the two extended conformational states B and P. Inspection of the native structure shows that this C-terminal strand is centrally located within a β-sheet, forming extensive H-bonds with neighboring β-strands.

The second CSR, also identified with high consensus (in 89–100% of the sequences), corresponds to an extended region whose last residues (40–41) are part of a β-sheet. Its structure is correctly predicted as extended (B/P), except for the cis conformation of Pro 36 that forms a kink in the native structure, which we do not predict. Our failure here is not too surprising, since our database does not contain enough examples of *cis*-prolines to yield reliable knowledge-based potentials for this conformation.

The third CSR, identified in 94% of the family members, is also correctly predicted as extended (B/P), except, once more, for the *cis*-proline at position 16. In the crystal structure, part of this region (residues 14–15) is a short β-strand that forms H-bonds with the N-terminal β-strand [4–5].

A recent NMR study concludes on the predominance of β-conformations in a series of peptides derived from the entire French bean plastocyanin sequence (Dyson et al., 1992). In

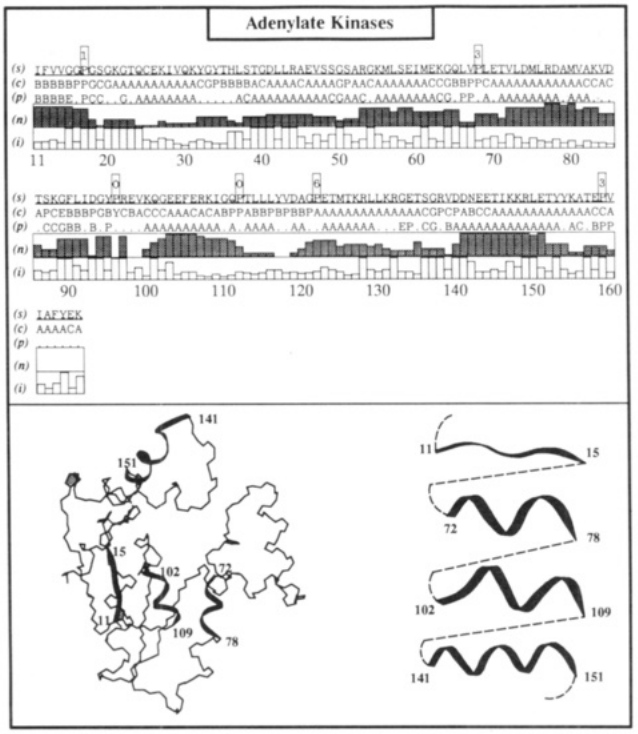

FIGURE 5: Prediction of consensus structured regions (CSR's) in the adenylate kinase family, with porcine adenylate kinase (3ADK) as representative. Note that the first 10 and the last 28 residues of 3ADK are not involved in the alignment and are therefore omitted from the figure. See Figure 2 caption for details.

particular, the peptides corresponding to the extended CSR's identified in our study seem to be devoid of helix or turn tendencies.

(*2.d*) *Adenylate Kinases*. In the adenylate kinases, 4 CSR's comply with our criteria (Figure 5). These are respectively [141–151] (100%), [72–78] (82–100%), [102–109] (82–100%), and [11–15] (94%), with segment limits given between square brackets and the fraction of family members contributing to the predictions given in parenthesis.

The first CSR is located in the second structural domain of porcine adenylate kinase (3ADK). It corresponds to the beginning of the C-terminal helix, which is in contact with helix [123–135] in the 3D structure. The predicted preferred structure is helical throughout, though Asn 142 adopts an extended (B) conformation in the representative structure.

The following two segments [72–78] and [102–109], predicted and observed to be fully helical, are located in the first structural domain of the protein, where they are in contact with each other, as well as with other helices not identified as CSR's by our procedure. Helix [72–78] interacts with helix [41–49], while helix [102–109] interacts with the N-terminal helix of 3ADK, not considered in the predictions (see legend of Figure 5).

The fourth CSR is the N-terminal β-strand [11–15], whose preferred conformation is correctly predicted as extended (B/P), except for residue 15 whose computed conformation is loop (G/E). This strand is located in the middle of the sheet in the second domain, at the interface with the first domain. Interestingly, a synthetic segment corresponding to the first 45 residues of adenylate kinase, including the [11–15] CSR, was shown to bind ATP with the same affinity as the intact enzyme, and to be appreciably structuredon its own (Fry et al., 1988).

(*2.e*) *Thermolysins*. Four CSR's are identified in all 5 considered thermolysins. They are [102–106], [129–132],

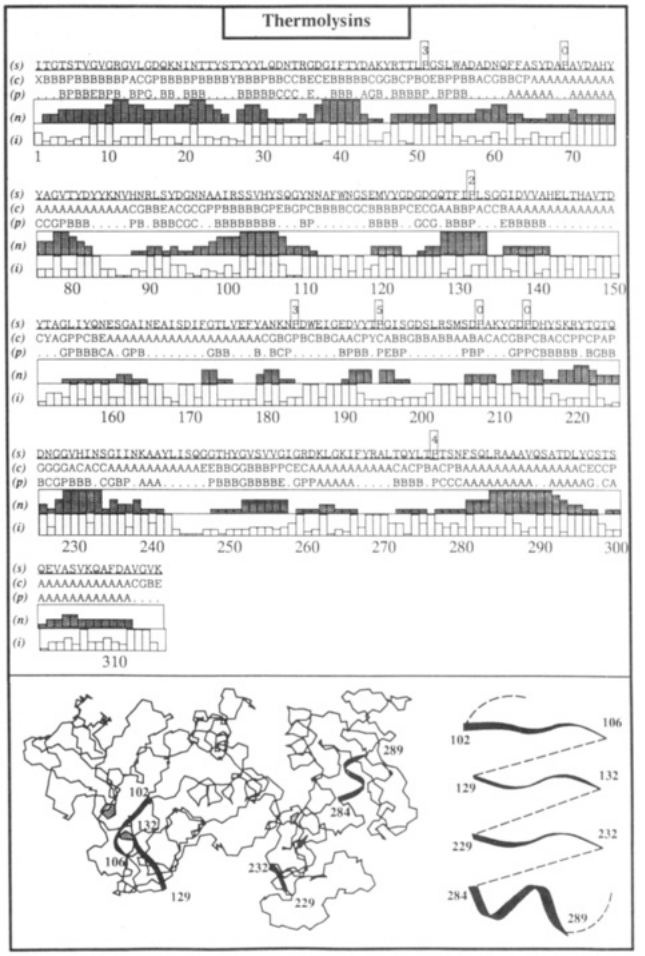

FIGURE 6: Prediction of consensus structured regions (CSR's) in the thermolysin family, with *Bacillus thermoproteolyticus* thermolysin (3TLN) as representative. See Figure 2 caption for details.

[229–232], and [284–289] (see Figure 6). Given the small number of sequences used in the computations, these results must however be considered as less reliable than those obtained for more populated families.

The first two CSR's, predicted as fully extended (B/P), are part of the middle $\alpha + \beta$ structural domain of thermolysin. In the 3D structure, segment [102–106] is indeed extended, but only up to residue 104, and corresponds to a central β-strand, while segment [129–132] is in a loop.

The two other segments, [229–232] and [284–289], are in the large C-terminal structural domain. Segment [284–289] is predicted as α-helical, in agreement with the observed structure, and corresponds to the central helix of the domain. Segment [229–232] is consistently predicted as extended (B/P), but is a loop in the crystal structure. It is noteworthy that our procedure detects no CSR's in the two long stretches 1–102 and 132–229, the former containing the complete N-terminal domain.

Several segments of thermolysin, isolated by limited proteolysis (Fontana, 1988, 1989; Vita et al., 1989), have been characterized by CD spectroscopy and by reactivity with specific antibodies. These studies indicate that the isolated C-terminal segment [255–316] retains its native structure and is highly stable, while smaller segments from the same region, respectively, [296–307], [308–316], [296–302], and [303–316], are, on the contrary, unable to acquire their native helical conformation in aqueous solution.

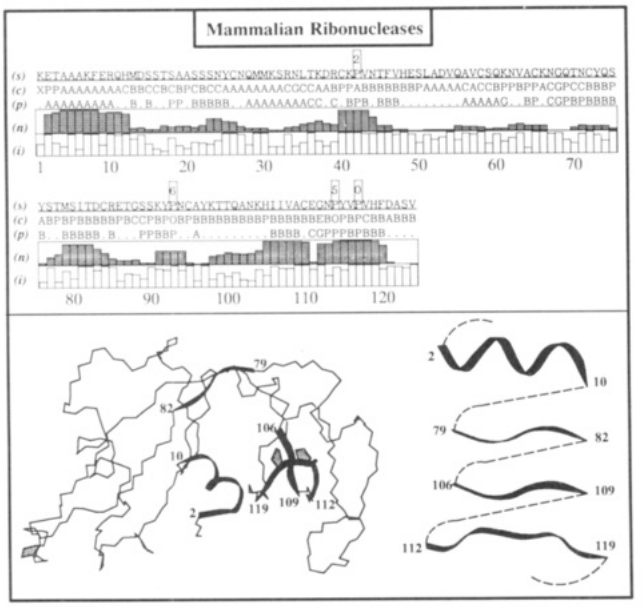

FIGURE 7: Prediction of consensus structured regions (CSR's) in the mammalian ribonuclease A family, with bovine ribonuclease A (7RSA) as representative. See Figure 2 caption for details.

We see that no CSR's are identified in these smaller unstructured segments, while the only α-helical CSR identified by our procedure (segment [284–289]) is located in the center of the independently structured C-terminal segment. This suggests that the [284–289] helix may form ahead of other regions in the C-terminal domain, possibly acting as the core around which other helices would coalesce.

(*2.f*) *Mammalian Ribonucleases*. In the mammalian ribonucleases a total of 4 CSR's are detected (Figure 7). These are [2–10] (82–100%), [112–119] (87–100%), [72–82] (90–97%), and [106–109] (90–97%), with segment limits given in square brackets, and the protein fraction contributing to the predictions given in parentheses.

The first 2 CSR's are located at the N- and C-termini, respectively. Residues [2–10] are predicted as helical, though the helix starts only at residue 4 in the 3D structure (bovine RNase-A). This helix makes extensive interactions with the rest of the structure. The C-terminal segment [112–119] is predicted as extended (B/P), except for residue 112, predicted as loop (G/E). The observed structure in this region is somewhat different, with a cis conformation at Pro 114 and a 3_{10} conformation (C) at Val 118, which forms in fact a β-bulge. The other 2 CSR's correspond to β-strands in the 3D structure. Their structure is predicted as extended (B/P) in agreement with the observations. These strands are moreover centrally located within the large bent sheet of RNase-A, with each strand making H-bonds with 2 flanking strands.

A variety of experimental data are available on the folding of RNase-A, and on the stability of its segments and variants. CD and NMR measurements (Brown & Klee, 1971; Bierzynski et al., 1982; Shoemaker et al., 1985) indicate that an isolated fragment, comprising the 13 N-terminal residues, shows significant helical propensity in aqueous solution. Another interesting series of experiments show that cleaving respectively 1–6 residues from the C-terminus of RNase-A results in a completely inactive and unstable protein (Anfinsen, 1956; Taniuchi, 1970; Schülke, 1984; Sela et al., 1957; Potts et al., 1964; Andria & Taniuchi, 1978; Teschner, 1987). Finally, NMR studies on the bovine enzyme (Udgaonkar & Baldwin, 1988) indicate that the β-sheet is formed during the first 1.5 s of folding.

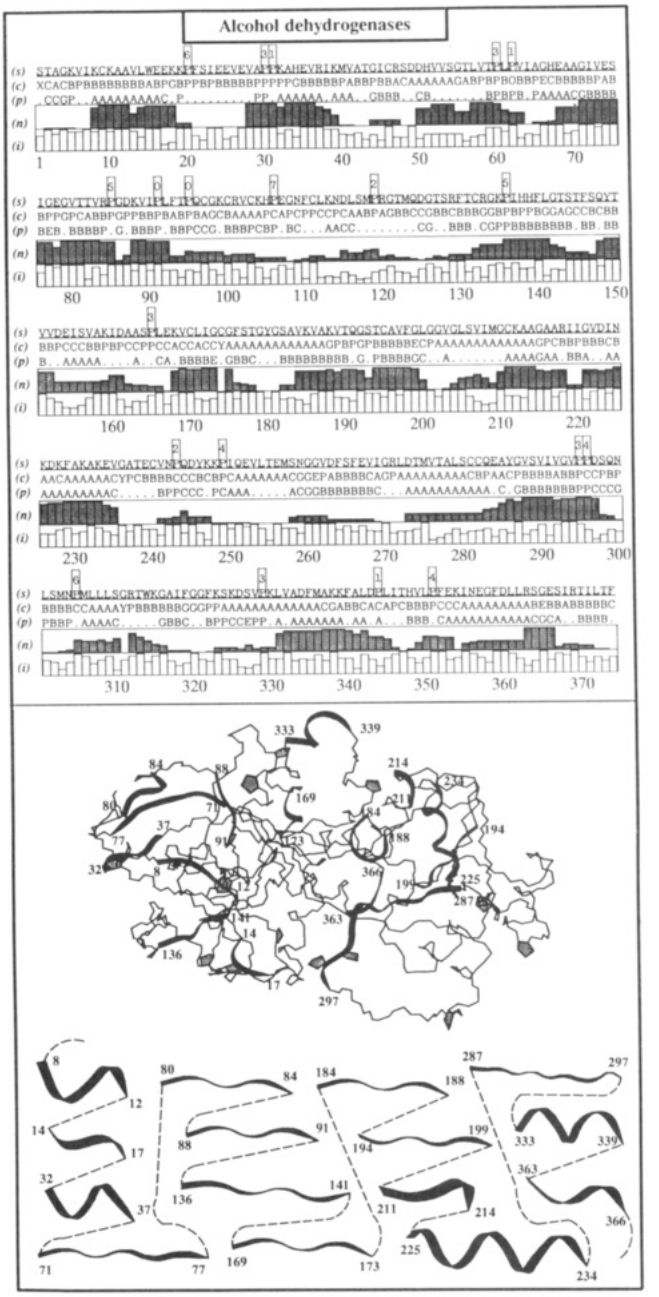

FIGURE 8: Prediction of consensus structured regions (CSR's) in the alcohol dehydrogenase family, with horse alcohol dehydrogenase (8ADH) as representative. See Figure 2 caption for details.

The observed helical propensity of the N-terminus agrees well with our results and was discussed previously (Rooman & Wodak, 1991; Rooman et al., 1991). The observed destabilizing effect of trimming the C-terminus is also in complete accord with predicting near it a CSR which we take as playing an important role in folding. The experimental observations on sheet formation are unfortunately not detailed enough to permit direct comparison with our predictions for strands [72–82] and [106–109]. Our analysis suggests that these strands could form the initial building blocks onto which other strands assemble to form the sheet. Further experiments combining mutagenesis with detailed analyses of folding intermediates are however needed to validate this hypothesis.

(*2.g*) *Alcohol Dehydrogenase*. As many as 15 CSR's are identified in the 19 considered alcohol dehydrogenases (Figure 8). For 10 of these, [8–12], [14–17], [32–37], [71–77], [80–84], [169–173], [184–188], [287–297], [333–339], and [363–366], the predictions agree only partially or not at all with the

observed structure and will therefore not be discussed, except to mention that 8 among them are located in the catalytic domain of the subunit. For the remaining 5 consensus regions, the observed and predicted structures agree. Three of these, [194–199], [211–214], and [225–234], are part of the nucleotide binding domain, and two, [88–91] and [136–141], are in the catalytic domain.

Two out of the 3 CSR's in the nucleotide binding domain, [211–214] and [225–234], are predicted and observed to be α-helical, with respectively 95% and 90–100% of the 19 sequences contributing to the predictions. The third region, [194–199], predicted as extended (B/P) in 90–95% of the sequences, is a central β-strand except for residue 199 which adopts a loop (G/E) conformation. It is remarkable that all three CSR's map onto the $\beta/\alpha/\beta/\alpha$ motif that binds nucleotide. Segment [194–199] is the first strand of the motif, segment [211–214] is the C-terminus of the first helix, and segment [225–234], the second helix. The latter packs against the first β/α loop which binds the phosphate moiety.

The CSR's [88–91] and [136–141] identified in 95–100% and 90% of the sequences, respectively, occur in the catalytic domain that mostly contains β-structure. Both are predicted and observed as extended (B/P), except Leu 141, observed to be in loop (G) conformation. The first of the 2 segments is a middle β-strand, while the second corresponds to an extended coil following the short β-strand [135–136].

Our findings that crucial portions of the nucleotide binding domain are consistently predicted as preferring their native conformation, while the catalytic domain contains most of the regions predicted to prefer nonnative structures, can be taken to suggest that, early during folding, the former would acquire its native fold, while the latter would adopt essentially a nonnative conformation. Since subunit contacts involve primarily the nucleotide binding domains, protein association would also occur early. Interactions with the nascent dimer would then mediate the folding of the catalytic domains, which make extensive contacts with the nucleotide binding domains from both subunits. Folding experiments on this protein should be able to test this hypothesis.

(*2.h*) *Globins*. The globins are the largest and most diverse family analyzed in the present study. In the 481 sequences considered, only very few residues are strictly conserved. Among those are the proximal and distal histidines, involved respectively in heme ligation and ligand binding (Bashford et al., 1987). In this family, only one CSR [A83–A86] is identified in 85–88% of the sequences, representing a relatively low fraction, but quite a large number of members: 423. It may be significant that this segment is adjacent to A87, the heme ligating histidine.

Comparison of charts similar to those of Figures 2–8 (not shown) with those obtained for other families indicates clearly that predictions of segments with well-defined conformations are less well conserved across the globins where not a single CSR is identified in all members, a situation rarely encountered in other families. Taking this to suggest that the 481 globins taken together may also be more diverse with regard to their folding pathways, our prediction procedure was applied individually to each of the 6 globin subclasses. The results, summarized in Figure 9, are described below.

(*2.h.i*) *Hemoglobin α-Chains*. Three CSR's are identified in this subclass: [A83–A89], [A25–A29], and [A42–A45]. The first region predicted with the highest consensus (in 81–98% of the proteins) to adopt its native conformation is located in the F-helix (Figure 9) and contains the proximal histidine. This helix is in contact with the heme group and with the H-helix in the 3D structure. The second well-predicted CSR identified in 86–92% of the proteins is in the B-helix, which, likewise, makes extensive contacts with other helices. The third CSR, identified in a smaller protein fraction (80–85%), belongs to the C-helix, but is predicted as fully extended (B/P), hence to adopt a nonnative structure. These results suggest that, during the early folding stages of the hemoglobin α-chains, the F- and B-helices would form their native structure, while the C-helix would be misfolded.

(*2.h.ii*) *Hemoglobin β-Chains*. In the somewhat longer β-chains, two CSR's, [B125–B133] and [B138–B141], are identified in 83–100% and 82–88% of the proteins, respectively. Both are in the long H-helix, and their structure is correctly predicted as α-helical. In the folded polypeptide, the corresponding segments are in contact with the A-, E-, F-, and G-helices in the same subunit, and the [B125–B133] CSR is in contact with the B-helix of a neighboring α subunit across the $\alpha 1/\beta 1$ interface. Our results thus suggest that, while the intrinsic helical propensity and early formation of the F- and B-helices may be crucial for the correct folding of the α-chain, it may be less so for folding the β-chain. Rather, correct folding of this somewhat longer polypeptide would require early structuring the H-helix at the C-terminus. Furthermore, the fact that the latter interacts with the B-helix CSR of the α-chain in the tetramer suggests that subunit association may occur at the earlier stages of folding.

(*2.h.iii*) *Vertebrate Myoglobins*. Our procedure identifies 4 CSR's for this family (Figure 9). The first, identified in 85–100% of the sequences, comprises residues [125–150]. It is the largest CSR identified so far (26 consecutive residues) and corresponds to the entire H-helix. Its predicted and native structures agree fully, including the loop conformation (G) of the last residue. The other 3 CSR's are as follows: [89–95], [51–63], [39–48], identified in 90–97%, 84–99%, and 81–96% of the proteins, respectively. They correspond to the F-helix containing the proximal histidine, the C-helix, and the D–E loop. For the F-helix, predicted and observed structures agree well, while those for the D–E loop and C-helix, do not. The latter two are predicted as fully α-helical, though Phe 43 in the distorted C-helix is observed in extended (B/P) conformation. As in the hemoglobins, the F- and H-helices are in contact with each other and with other helices in the structure (A, E, and G). This is not the case for the solvent-exposed C-helix and the D–E loop.

The myoglobins are the only globin subfamily for which experimental data on structures of folding intermediates and isolated segments are available. NMR hydrogen exchange studies of apomyoglobin folding intermediates (Hughson et al., 1990) suggest that the A-, G-, and H-helices fold before the B- and E-helices. In addition, peptides with sequences corresponding to the H-helix are found to be substantially helical in water (Waltho et al., 1989), but experiments on peptides corresponding to the G-helix are less conclusive (Hughson et al., 1990). These data are consistent with the H-helix acting as the core around which other helices assemble, but they provide little information to what extent the structures of other helices are present beforehand or are induced upon contact with it.

The experiments thus agree well with our predictions on the intrinsic helical propensity of the H-helix. In further agreement with the experiments, we identify no CSR's in the B- and E-helices, except for the D–E turn for which a preferred nonnative helical conformation is predicted in the absence of tertiary interactions. No data are available on the fate of the

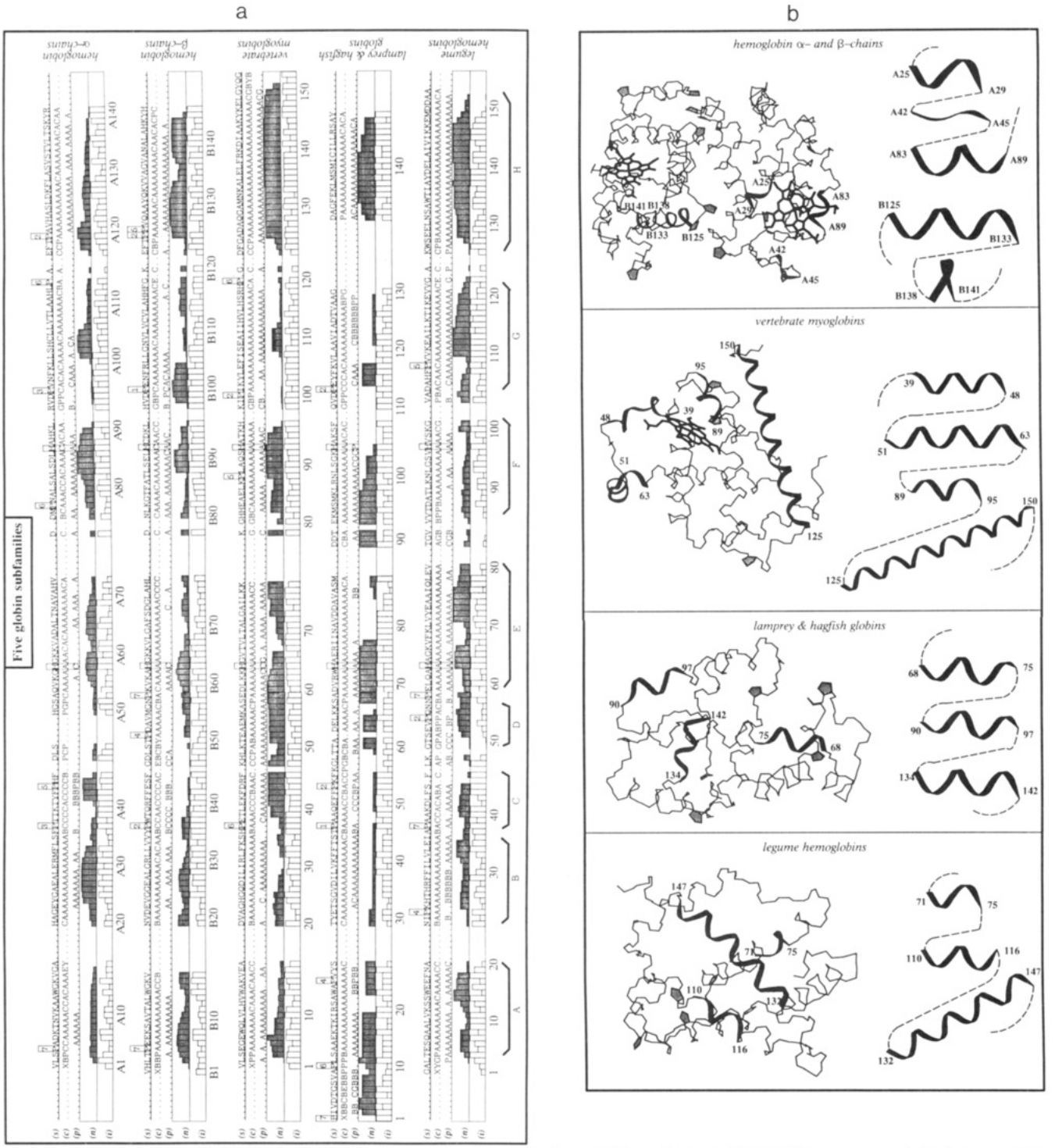

FIGURE 9: Prediction results for five individuals globin subfamilies. The hemoglobin α-chains (with 4HHBα as representative), hemoglobin β-chain (4HHBβ), vertebrate myoglobins (1MBD), lamprey and hagfish globins (2LHB), and legume hemoglobins (2LH4). No consensus structured regions (CSR's) were identified for the insect globins. (a) Sequences as well as observed and predicted structures are aligned according to Bashford et al. (1987). The approximate boundaries of the 8 globin helices, conventionally denoted A–H, are given. The distal and the proximal histidines are framed. See legend of Figure 2 for all further details. (b) Drawings of predicted and observed structures of the CSR's identified in the 5 globin subfamilies of panel a. Heme coordinates, whenever present in the Protein Databank entry (Bernstein et al., 1977), are drawn in bold lines. See legend of Figure 2 for further details.

F- and C-helices during folding, to allow comparison with our predictions.

(*2.h.iv*) *Other Globins*. For the lamprey and hagfish globins, only 5 sequences were considered, yielding less reliable results, with 3 identified CSR's in all 5 members of the family (Figure 9). These are [68–75], [90–97], and [137–142]. They map onto portions of the F-, E-, and H-helices, respectively, and their native structure is well predicted. The F-helix CSR does not contain the proximal histidine, but the E-helix CSR includes the distal histidine.

In the legume hemoglobins, also 3 CSR's are identified (Figure 9). These are respectively [71–75], [110–116], and [132–147]. The first corresponds to the C-terminus of the E-helix and is identified in all members. The other 2, identified in 85–100% of the proteins, are in the middle of the G- and H-helices, respectively. The E-helix CSR does not include the distal histidine. Predicted and observed structures agree in all three regions.

No CSR's were detected for the 13 midge larvae insect globins, which, with an average of 50% sequence identity, are the most diverse of the 6 globin subfamilies considered here.

DISCUSSION

Our results show that, except for the insect globins, all the families and subfamilies considered here have one or more CSR's. Considering that CSR's correspond to segments that consistently form their structure independently of other regions, in all members of the same family, we suggest that they define crucial features of the folding pathway(s), which are shared by the family members. We see however that a minumum level of sequence conservation, and consequently most probably also of structural similarity, is necessary for this to hold. This is not due to a trivial correlation between CSR's and local sequence conservation, since CSR's are seen to be just as tolerant to sequence changes as the corresponding complete 3D structures. Hence, the absence of CSR's in the insect globins, whose average sequence identity (50%) is lower than in other analyzed families, may indicate that they are diverse enough for their folding pathways to differ. This may apply to some extent to all the globins taken together or even to the cytochrome *c*'s, where the only identified CSR is not shared by all family members. Such families are likely to contain subclasses with different CSR's (we have seen this in the globins) and hence possibly also with alternative folding pathways. Following a similar reasoning, regions predicted to have a preferred conformation only in some family members and not in others—the weakly conserved structured regions—could reveal the existence of variations in the main folding pathway among family members, possibly reflecting adjustments to different physiological constraints.

Further insight into the possible roles of CSR's is provided by several clear trends that emerge from our results. We see that CSR's are frequently (in 9 out of the 13 sequence groups) located at chain ends, and somewhat more so, at the C- than the N-termini. Given that chain ends are in close spatial proximity in most of the representative tertiary structures of our families, in agreement with the observations of Barlow and Thornton (1983), the suggestion of these authors that such proximity reflects folding requirements is further supported by our contention that chain ends would tend to form their native structure ahead of other regions.

Another interesting trend is that most of the correctly predicted β-strand and α-helical CSR's make extensive contacts with the rest of the structure. A majority of the β-strand CSR's are buried, and in the middle of sheets. α-Helical CSR's are nearly always well packed against other secondary structure elements, and/or against the heme (in the globins and cytochrome *c*'s). These observations are similar to those made earlier for protein segments containing highly predictive amino acid sequence patterns (Rooman et al., 1990). They suggest that the CSR's could act as nuclei, around which other parts of the structure assemble. This is all the more remarkable given that CSR's were not selected on the basis of hydrophobicity, buried surface area, or position in the 3D structure. Rather, it follows from their tendency to have well-defined structures on their own, which has the effect to minimize the entropy loss upon transfer into the constrained protein matrix.

Coil and loop regions in the native structure generally do not contain CSR's. This is consistent with the inherent flexibility of these regions in the folded state, which in turn translates into their larger sequence and structure variability in related proteins. Chain entropy lost upon folding should be smaller in these regions than for chain segments incorporated into the protein core. Positioning CSR's in such regions should therefore be less favorable entropically. Two exceptions, representing somewhat special cases are however found. They are the CSR's [136–141] and [35–41] of alcohol dehydrogenase and plastocyanin, respectively. Both CSR's, predicted as extended, form an extended coil that connects two β-strands in the native structure. They are moreover buried and make extensive contacts with the rest of the protein. In plastocyanin, it literally bridges the two parallel β-sheets and could therefore play a role in achieving this particular sheet arrangement.

A different category of CSR's corresponds to that where the predicted and observed structures do not agree. A majority are loops in the folded protein, but predicted either as extended structure (B/P), or as α-helical (A/C). Only in 4 instances do we find a CSR to be helical in the folded protein, but predicted as extended, or vice versa. This general category of CSR's may be interpreted in two ways. Either they play an important role in the folding process, which would thus require nonnative substructures to be present early during folding, probably to avoid nonproductive pathways. Or they may be conserved as nonnative local substructures among members of a family for reasons unrelated to folding. Better understanding of the folding process is needed to settle these questions.

Not too surprisingly, CSR's do not seem to map onto sites involved in function (ligand binding and catalysis). Though these sites correspond to highly conserved regions of the sequence, they occur most often in loops, shown to seldom contain CSR's. Regions involved in binding coenzyme or prosthetic groups may belong to a different category, since those groups are often intimately involved in maintaining the tertiary structure. We see indeed that in the cytochrome *c*'s the C-terminal helical CSR does not include the heme ligands, but forms noncovalent contacts with the heme. In the hemoglobin α-chains and the myoglobins, one of the CSR's (the F-helix) is adjacent to or contains the proximal histidine, while in the lamprey and hagfish group, only the distal histidine is part of CSR's. In all these cases, the CSR's make nonbonded contacts with the heme. In the alcohol dehydrogenases, three CSR's map onto the $\beta/\alpha/\beta/\alpha$ unit that binds NAD.

Overall, CSR's could thus exert positive (nucleation) or negative (shunting nonproductive pathways) "control" on folding. Being particularly robust substructures, they should be detectable experimentally, in the unfolded state, and early during folding, as well as in isolated protein fragments. The favorable correlation between our predicted CSR's with available experimental data on folding supports this view. But more detailed information on structures present at the various physical states of the polypeptide is needed to gain further insight.

One factor may offset the proposed link between CSR's and substructures detected early during folding. It is the possible effect of proline isomerization, which our procedure cannot take into account. Though it seems fairly clear that proline isomerization occurs on much longer time scales than those associated with the formation of early folding intermediates [Brandts et al., 1975; for review, see Kim and Baldwin (1990)], its potential interference with their formation has been invoked (Roder et al., 1988). To help address this question as more experimental data on folding become available for the proteins analyzed here, the position of the prolines and their solvent-accessible surface areas in all the representative 3D structures are emphasized in the drawings and charts of Figures 2–9.

To this overall coherent picture, several cautionary notes must be added. The quality of our knowledge-based force field is no doubt affected by the limited size of the protein structure database (Rooman & Wodak, 1988). In addition, our postulate of considering conformations of individual residues as independent from one another may be a serious approximation, though it should be of lesser consequence as long as we deal with identifying conformations with sizable energy gaps (Rooman et al., 1991), as we did here. The accuracy our method is furthermore improved by considering consensus predictions in a large number of homologus sequences.

ACKNOWLEDGMENT

We are grateful to J. Richelle for useful discussions and for use of the SESAM database (Huysmans et al., 1991). We also thank J.-P. Kocher for fruitful discussions and C. Chothia and D. Eisenberg for the generous gifts of sequence alignments.

REFERENCES

Alard, P. (1991) Ph.D. Thesis, Université Libre de Bruxelles.
Alber, T. (1989) *Annu. Rev. Biochem. 58*, 765–798.
Andria, G. & Taniuchi, H. (1978) *J. Biol. Chem. 253*, 2262–2270.
Anfinsen, C. B. (1956) *J. Biol. Chem. 221*, 405–412.
Barlow, D. J., & Thornton, J. M. (1983) *J. Mol. Biol. 168*, 867–885.
Bashford, D., Chothia, C., & Lesk, A. M. (1987) *J. Mol. Biol. 196*, 199–216.
Bashford, D., Cohen, F. E., Karplus, M., Kuntz, I. D., & Weaver, D. L. (1988) *Proteins 4*, 211–227.
Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanoushi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.
Bierzynski, A., Kim, P., & Baldwin, R. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 2470–2474.
Brandts, J. F., Halvorson, H. R., & Brennan, M. (1975) *Biochemistry 14*, 4953–4963.
Brown, J., & Klee, W. (1971) *Biochemistry 10*, 470–476.
Chiche, L., Gregoret, L. M., Cohen, F. E., & Kollman, P. A. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 3240–3243.
Chothia, C., & Janin, J. (1975) *Nature 256*, 705–708.
Chothia, C., & Lesk, A. M. (1986) *EMBO J. 5*, 823–826.
Covell, D. G., & Jernigan, R. L. (1990) *Biochemistry 29*, 3287–3294.
Crippen, G. N. (1978) *J. Mol. Biol. 126*, 315–332.
Delhaise, P., Van Belle, D., Bardiaux, M., & Wodak, S. (1985) *J. Mol. Graph. 3*, 116–119.
Dill, K. A. (1990) *Biochemistry 29*, 7133–7155.
Dobson, C. M. (1991) *Curr. Opin. Struct. Biol. 1*, 22–27.
Dyson, H. J., & Wright, P. E. (1991) *Annu. Rev. Biophys. Biophys. Chem. 20*, 519–538.
Dyson, H. J., Sayre, J. R., Merutka, G., Shin, H. C., Lerner, R. A., & Wright, P. E. (1992) *J. Mol. Biol. 226*, 819–835.
Efron, B. (1982) *The Jack Knife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia.
Farber, G. K., & Petsko, G. A. (1990) *Trends Biochem. Sci. 15*, 228–234.
Fontana, A. (1988) *Biophys. Chem. 29*, 181–193.
Fontana, A. (1989) *Proceedings of the 11th American Peptide Symposium*, La Jolla, CA.
Fry, D. C., Byler, D. M., Susi, H., Brown, E. M., Kuby, S. A., & Mildvan, A. S. (1988) *Biochemistry 27*, 3588–3598.
Hughson, F., Wright, P., & Baldwin, R. (1990) *Science 249*, 1544–1548.
Huysmans, M., Richelle, J., & Wodak, S. J. (1991) *Proteins 11*, 59–76.
Jaenicke, R. (1987) *Prog. Biophys. Mol. Biol. 49*, 117–237.
Kabsch, W., Mannherz, H. G., Suck D., Pai, E. F., & Holmes, K. C. (1990) *Nature 347*, 37–44.
Kang, H. S., Kurochkina, N. A., & Lee, B. K. (1992) *J. Mol. Biol.* (in press).
Karplus, M., & Weaver, D. L. (1976) *Nature 260*, 404–406.
Kauzmann, W. (1959) *Adv. Protein Chem. 14*, 1–63.
Kim, P. S., & Baldwin, R. L. (1990) *Annu. Rev. Biochem. 59*, 631–660.
Kuwajima, K., Yamaya, H., Miwa, S., Sugai, S., & Nagamura, T. (1987) *FEBS Lett. 221*, 115–118.
Lesk, A. M., & Rose, G. D. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 4304–4308.
Levin, J. M., & Garnier, J. (1988) *Biochim. Biophys. Acta 955*, 283–295.
Levinthal, C. (1966) *Sci. Am. 214*, 42–52.
Levitt, M. (1991) *Curr. Opin. Struct. Biol. 1*, 224–229.
Lüthy, R., McLachlan, A., & Eisenberg, D. (1991) *Proteins 10*, 229–239.
Miranker, A., Radford, S. E., Karplus, M., & Dobson, C. M. (1991) *Nature 349*, 633–636.
Montelione, G. T., & Scheraga, H. A. (1989) *Acc. Chem. Res. 22*, 70–76.
Moult, J., & Unger, R. (1991) *Biochemistry 30*, 3816–3824.
Oas, T. G., & Kim, P. S. (1988) *Nature 336*, 42–48.
Potts, J. T., Jr., Young, D. M., Anfinsen, C. B., & Sandoval, A. (1964) *J. Biol. Chem. 239*, 3781–3786.
Rashin, A. A. (1981) *Nature 291*, 85–87.
Rashin, A. A. (1984) *Biopolymers 23*, 1605–1620.
Richardson, J. W. (1981) *Adv. Protein Chem. 34*, 167–339.
Roder, H., Elöve, G., & Englander, W. (1988) *Nature 335*, 700–704.
Rooman, M. J., & Wodak, S. J. (1988) *Nature 335*, 45–49.
Rooman, M. J., & Wodak, S. J. (1991) *Proteins 9*, 69–78.
Rooman, M. J., Rodriguez, J., & Wodak, S. J. (1990) *J. Mol. Biol. 213*, 337–350.
Rooman, M. J., Kocher, J.-P. A., & Wodak, S. J. (1991) *J. Mol. Biol. 221*, 961–979.
Rooman, M. J., Kocher, J.-P. A., & Wodak, S. J. (1992) *Biochemistry* (preceding paper in this issue).
Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., & Zehfus, M. H. (1985) *Science 229*, 834–838.
Schülke, N. (1984) Thesis, Regensburg University.
Sela, M., Anfinsen, C. B., & Harrington, W. F. (1957) *Biochim. Biophys. Acta 26*, 502–512.
Shoemaker, K., Kim, P., Brems, D., Marqusee, S., York, E., Chaiken, I., Stewart, J., & Baldwin, R. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 2349–2353.
Skolnick, J., & Kolinski, A. (1989) *J. Mol. Biol. 212*, 787–817.
Staley, J. P., & Kim. P. S. (1990) *Nature 344*, 685–688.
Taniuchi, H. (1970) *J. Biol. Chem. 245*, 5459–5468.
Teschner, W. (1987) Thesis, Regensburg University.
Udgaonkar, J., & Baldwin, R. (1988) *Nature 335*, 694–699.
Vita, C., Fontana, A., & Jaenicke, R. (1989) *Eur. J. Biochem. 183*, 513–518.
Waltho, J. P., Feher, R. A., Lerner, R. A., & Wright, P. E. (1989) *FEBS Lett. 250*, 400–404.
Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 697–701.
Wright, P., Dyson, J., & Lerner, R. (1988) *Biochemistry 27*, 7167–7175.
Zuniga, E. H., & Nall, B. T. (1983) *Biochemistry 22*, 1430–1437.
Zvelebil, M. J., Barton, G. J., Taylor, W. R., & Sternberg, M. J. E. (1987) *J. Mol. Biol. 195*, 957–961.

Registry No. cytochrome *c*, 9007-43-6; lysozyme, 9001-63-2; adenylate kinase, 9013-02-9; thermolysin, 9073-78-3; ribonuclease, 9001-99-4; alcohol dehydrogenase, 9031-72-5.