Protein Science (1993), *2*, 626–639. Cambridge University Press. Printed in the USA.

Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipoylated H-protein from the pea leaf glycine cleavage system: A new automated method for the prediction of protein tertiary structure

SIMON M. BROCKLEHURST AND RICHARD N. PERHAM
Cambridge Centre for Molecular Recognition, Department of Biochemistry,
University of Cambridge, Cambridge CB2 1QW, United Kingdom

(RECEIVED September 8, 1992; REVISED MANUSCRIPT RECEIVED November 27, 1992)

Abstract

A new, automated, knowledge-based method for the construction of three-dimensional models of proteins is described. Geometric restraints on target structures are calculated from a consideration of homologous template structures and the wider knowledge base of unrelated protein structures. Three-dimensional structures are calculated from initial partly folded states by high-temperature molecular dynamics simulations followed by slow cooling of the system (simulated annealing) using nonphysical potentials. Three-dimensional models for the biotinylated domain from the pyruvate carboxylase of yeast and the lipoylated H-protein from the glycine cleavage system of pea leaf were constructed, based on the known structures of two lipoylated domains of 2-oxo acid dehydrogenase multienzyme complexes. Despite their weak sequence similarity, the three proteins are predicted to have similar three-dimensional structures, representative of a new protein module. Implications for the mechanisms of posttranslational modification of these proteins and their catalytic function are discussed.

Keywords: biotinyl domain; H-protein; lipoyl domain; modeling; simulated annealing; three-dimensional structure prediction

An emerging common feature of multifunctional polypeptide chains is their construction from independently folded protein domains joined by linker sequences of various lengths and degrees of conformational flexibility. The dihydrolipoamide acyltransferase (E2) chains of 2-oxo acid dehydrogenase multienzyme complexes are excellent examples (Reed & Hackert, 1990; Perham, 1991): they comprise, from the N-terminus, up to three lipoyl domains, a peripheral subunit-binding domain, and a larger core-forming acyltransferase domain, all linked together by long (25–30-residue) segments of polypeptide chain rich in alanine, proline, and charged/hydrophilic amino acids. These molecules have proved to be resistant to crystallization, perhaps because of the flexibility of the interdomain linkers, but their domain-and-linker arrangement has opened the way to a process of cumulative structural analysis. Thus, the structures of the lipoyl domain (approx. 80 residues) of the pyruvate dehydrogenase (PDH) complex of *Bacillus stearothermophilus* (Dardel et al., 1991, 1993) and *Escherichia coli* (J.D.F. Green, E.D. Laue, & R.N. Perham, unpubl.) and the peripheral subunit-binding domain (approx. 40 residues) of the 2-oxoglutarate dehydrogenase complex of *E. coli* (Robien et al., 1992) and the PDH complex of *B. stearothermophilus* (Kalia et al., 1993) have been determined by means of nuclear magnetic resonance (NMR) spectroscopy. The structure of the acetyltransferase domain as the assembled octahedral core of the PDH complex of *Azotobacter vinelandii* has been determined by means of X-ray crystallography (Mattevi et al., 1992), and the flexibility of the

Reprint requests to: Richard N. Perham, Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, United Kingdom.

interdomain polypeptide linkers in (Texter et al., 1988) and out of (Radford et al., 1989) the *E. coli* PDH complex has also been examined using directed mutagenesis and NMR spectroscopic techniques.

In the 2-oxo acid dehydrogenase complexes, the lipoic acid cofactor is attached in amide linkage to the N^6-amino group of a specific lysine residue in the lipoyl domain. The biotin cofactor in biotin-dependent carboxylases is similarly attached to a lysine residue in an amino acid sequence that is widely conserved among such enzymes (Samols et al., 1988). In yeast pyruvate carboxylase (PC), the biotinylated region appears to form a protein domain of about 70 amino acid residues that is located at the C-terminal end of each of the four component polypeptide chains, and it has been noted that there is some sequence similarity between this biotinyl domain and the lipoyl domain of 2-oxo acid dehydrogenase complexes (Lim et al., 1988). Any potential resemblance is emphasized by the similarity of the interaction of avidin with the biotinyl domains of carboxylases and the lipoyl domains of PDH complexes (Hale et al., 1992).

Another multienzyme complex containing a lipoylated protein is the glycine cleavage system. In this instance, the aminomethyl group derived from the decarboxylated glycine is transferred to the lipoyl-lysine residue of the H-protein (Hiraga & Kikuchi, 1980). The H-protein is about 120 amino acids in length (Fujiwara et al., 1986) and, as for the biotinyl domain, there is some evidence of sequence similarity to the lipoyl domain (Fujiwara et al., 1991). The H-protein from the glycine cleavage system of pea leaf has been crystallized, but no three-dimensional structure is yet available (Sieker et al., 1991).

The determination of the structures of the lipoyl domains of the *B. stearothermophilus* (Dardel et al., 1991, 1993) and *E. coli* (J.D.F. Green, E.D. Laue, & R.N. Perham, unpubl.) PDH complexes has now enabled us to make predictions about the structures of the other two proteins. By means of a largely automated procedure, we use the structures of the two lipoyl domains to identify potential key residues in the folding of the domains and project this information onto the one-dimensional sequences of the biotinyl domain from yeast PC and the H-protein from the pea leaf glycine cleavage system. The conservation of key residues indicates that there is likely to be considerable structural similarity between these proteins. We show how geometric restraints on the unknown, target three-dimensional structures may then be generated from a consideration of the structures of the template lipoyl domains. Dihedral angles, inter-C_α distances, and hydrogen bond distances and angles are among the structural features that are restrained. Model structures that satisfy these restraints are then calculated by using high-temperature molecular dynamics simulations followed by slow cooling using nonphysical potentials. The quality of the models and the potential errors are also discussed.

Results and discussion

Structure-based alignment strategy

The alignment of the templates and target is one of the most crucial stages in the modeling process. If the alignment is incorrect, then the restraints on the target will introduce errors into the model structures. Current automated structure-based alignment algorithms are not yet sufficiently robust to produce alignments that can be used without careful manual checking of the results. Often, subsequent manual adjustment is required. For this reason, we have developed automated coordinate analysis algorithms that facilitate systematic manual alignments (manual alignments are, of course, subjective if made from visual inspection of structures). Our alignment approach proceeds hierarchically, first with alignment of structural motifs (helices, strands, turns), then proceeding to alignment of hydrogen bond donors and acceptors and main chain dihedral angles. Our approach thus incorporates a flexible and general definition of topological equivalence (Sali & Blundell, 1990), allowing distantly related structures to be aligned. It is worth noting, however, that a more conventional structure alignment based on least-squares superposition and comparison of relative C_α positions would give the same result for the lipoyl domain template structures.

The information content of the hydrogen bond and van der Waals restraints

The restraint-based modeling procedure presented here uses restraints on main chain-main chain hydrogen bonds and close attractive van der Waals contacts to define the protein fold. Hydrogen bonds are well conserved in related protein structures, but van der Waals interactions are poorly conserved. There is, therefore, a requirement for new potentials and efficient sampling methods to search for correct van der Waals interactions, which could then be used as restraints in the present approach. We are testing such methods at present. To alleviate partially the lack of van der Waals restraints on the models generated by our current procedures, we have used minimal sets of restraints that do not directly relate to interactions that stabilize the folded state: dihedral angle restraints and local groups of inter-C_α distance restraints. The price to be paid for using these minimal sets of restraints is that we preclude accurate target structure prediction if the template structures do not cluster around the target.

An important question to ask is whether for van der Waals contacts the search algorithms that we are developing are likely to obviate the need for restrictive dihedral angle and multiple atom–atom distance restraints, i.e., we need to know if correct van der Waals and main chain–main chain hydrogen bond restraints alone provide for a well-defined and accurate structure. Previous restraint-based methods (Sali et al., 1990; Havel & Snow, 1991) have relied heavily on the multitude of interatomic distance restraints that are available from protein structures (e.g., the distances from one C_α to every other C_α in a structure could be used as restraints), most of which are not related to stabilizing physicochemical interactions.

In order to investigate whether we may be able accurately to predict target structures when the templates do not cluster around the target, we performed the following investigation using the structure of the peripheral subunit-binding domain of the dihydrolipoamide acetyltransferase chain from the PDH multienzyme complex of *B. stearothermophilus* (Kalia et al., 1993). Analysis of the structure, which is known to high resolution, suggests that a number of side chain-main chain and side chain-side chain interactions, in addition to the interactions that we represent by restraints, are important in stabilizing the fold. Reconstruction of this protein from our restraints should thus highlight deficiencies in our method, since interactions other than the ones we are attempting to include are obviously involved in stabilizing the structure.

To save computer time, we used loose main chain dihedral angle restraints (allowing each dihedral angle 80° of flexibility), which serve to increase the ratio of converged to nonconverged structures, but which do not contain any information that is not already contained in the other restraints, provided that the calculated structures are well defined at a particular residue position. We constructed a list of restraints on the structure, defining all the main chain-main chain hydrogen bonds and all the pairwise van der Waals interactions (distance restraints on the positions of only C_α and C_β atoms were used). No restraints on side chain dihedral angles were used: if the backbone can be accurately defined by our restraints, then most of the side chain conformations follow (Desmet et al., 1992; Holm & Sander, 1992). Ten structures were calculated from different starting conformations. The root mean square deviation (rmsd) from the mean structure was 0.47 Å. The main chains for five predicted structures, superposed on the experimentally determined structure (Kalia et al., 1993), are shown in Figure 1. The least well defined region of the main chain (residues 24–32) is exactly the region that was expected to be stabilized by multiple hydrogen bonds from backbone amide protons to the buried hydrophilic groups of the side chains of D34 and T24. The existence of these hydrogen bonds in homologous structures could have been predicted, since hydrogen bonds involving buried hydrophilic residues are extremely well conserved during evolution. It is clear then that our approach to modeling protein structures by using restraints related only to interactions that stabilize the folded state shows promise: its ultimate success will depend on how well the van der Waals contacts can be predicted.

It should be noted that the potentials used in the present work are dominated by the energy penalty for restraint violation, and the structure is largely defined by the restraints. Inclusion of water molecules in the system is not necessary for construction of a "good" structure. In most cases, surface side chains, for which time-averaged models are appropriate, appear to be involved in intramolecular protein–protein interactions; for example, side chain–main chain hydrogen bonds are common in β-turns and at caps of helices. Such features are well conserved in homologous structures and thus can be inferred from the template structures if they are of sufficient quality.

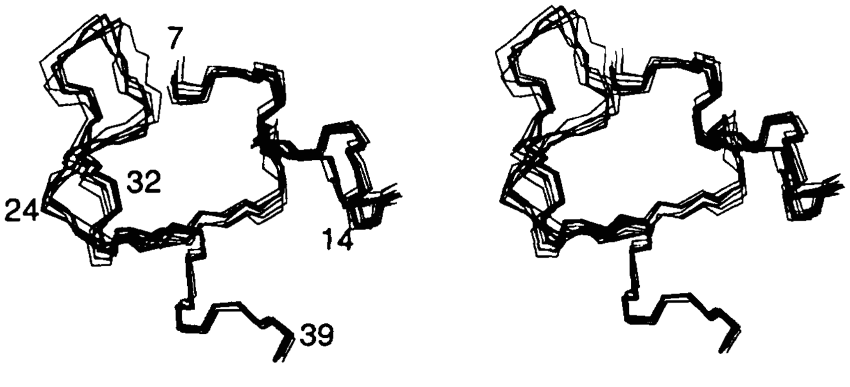

Fig. 1. A test case illustrating the potential of reconstructing protein structures by using restraints based solely on interactions stabilizing the folded state. Restraints on only main chain-main chain hydrogen bonds and close attractive van der Waals interactions were used. Illustrated are five calculated structures (thin lines) superposed on the experimentally determined structure (thick line) of the peripheral subunit binding domain of dihydrolipoamide acetyltransferase from the pyruvate dehydrogenase multienzyme complex of *B. stearothermophilus* (Kalia et al., 1993). The figure shows that the α-helices (residues 7–14 and 32–39) and the 3_{10}-helix (residues 17–21) are extremely well defined by these restraints. The region comprising residues 24–32 is rather less well defined. This was expected because analysis of the structure suggested that side chain-main chain hydrogen bonds are crucial in determining the conformation of this loop. This figure and the other protein structure figures were created by use of the program MOLSCRIPT version 1.2 (Kraulis, 1991).

```
                    5         10        15        20        25        30        35
+ve phi                     -     -                           -
B.st E2p    A F E F K L P D I G E G I H E G E I V K W F V K P G D E V N E D D V L C E V Q
Contacts      3   5 2 1 1   1                 4   1 7 3 3 1 2       2         1 3     2
Contacts    1 1   3 1 3 2   3             2   5     4 1 2   1     1 6   1       5 7 1 6
E.coli E2p  A I E I K V P D I G - - A D E V E I T E I L V K V G D K V E A E Q S L I T V E
+ve phi                       E                               G           -


            40        45        50        55        60        65        70        75
+ve phi                               E                   -           -
B.st E2p    N D K A V V E I P S P V K G K V L E I L V P E G T V A T V G Q T L I T L D
Contacts          1   1   2 1   2 4     1 4 2   4 1 4 3     1   1   2       6 4 1 3
Contacts          1   2   4 1 3 1 5 1   1 2     2   1   1       4         2 5 5 2 6
E.coli E2p  G D Q A S M E V P A P F A G V V K E L K V N V G D K V K T G S L I M I F E
+ve phi                               E                   -           G
```

Fig. 2. Structure-based alignment of the amino acid sequences of the *B. stearothermophilus* and *E. coli* lipoyl domains. The sequences are taken from Borges et al. (1990) and Ali and Guest (1990). It should be noted that the *E. coli* lipoyl domain is in fact an inactive mutant; the lipoyl lysine residue (underlined) has been replaced by a glutamine. It is also a hybrid construct, with the first half of the sequence being that of the first lipoyl domain of the *E. coli* E2 chain, and the second half being that of the third lipoyl domain (Ali & Guest, 1990). Nonpolar contact numbers are indicated. The main chain conformational identifiers of residues adopting positive ϕ are indicated in the top and bottom lines of the alignment. Key residues identified from the structural analysis are indicated by the sans serif typeface.

Other surface side chains are assumed to be disordered in solution. Good evidence for these assumptions may be seen from correlation of measurements of $^3J_{\alpha\beta}$ coupling constants with $^2H/^1H$ exchange NMR data (Kalia et al., 1993; S.M. Brocklehurst, Y.N. Kalia, & R.N. Perham, unpubl.).

The minimal groups of restraints that are not related to stabilizing interactions are sufficient to ensure that the target structures are as compact as the template structures, while at the same time allowing secondary structural motifs to form from the restraints relating to interactions. Analysis of the secondary structure formed in the calculated structure thus gives a measure of how well the simulation has performed, since we know what the secondary structural motifs should be from the template/target alignment.

Template/target alignments

The structure-based alignment of the sequences of the lipoyl domains of the *B. stearothermophilus* and *E. coli* PDH complexes is shown in Figure 2, with the nonpolar contact numbers of relevant residues shown. Since no information concerning nonpolar contacts was included in the alignment procedure, evidence for the requirement for a large nonpolar group (V, I, L, Y, F, W) at a particular residue position can be obtained from analysis of this alignment (see Theory and methods). The alignment of the lipoyl domain sequences with those of the yeast PC biotinyl domain and the pea leaf H-protein is shown in Figure 3. From this, it can be seen that the target structures have less than 15% sequence identity with the templates.

Residues making multiple nonpolar contacts in the lipoyl domain structures, together with the corresponding residues from the sequences of the biotinyl domain and H-protein, are shown in Table 1. At almost every position where there appears to be a folding requirement for a large nonpolar side chain in the lipoyl domains, there is a similar residue in the biotinyl domain and H-protein. We have found the contact number (which is an indication of the number of residues whose side chains are making van der Waals contacts with the side chain of a given residue) to be a good measure of the requirement for a hydrophobic residue at a particular position in the sequence. Although correlated with solvent accessibility, it has the advantage of representing an attractive force that is likely

```
                    5         10        15        20        25        30        35
B.st E2p    A F E F K L P D I G E G I H E G E I V K W F V K P G D E V N E D D V L C E V Q
E.coli E2p  A I E I K V P D I G - - A D E V E I T E I L V K V G D K V E A E Q S L I T V E

pea leaf H  V A T I G I T D H A Q D H L G E V V F V E L P E P G V S V T K G K G F G A V E
yeast PC    H I G A P M A - - - - - - - - G V I V E V K V H K G S L I K K G Q P V A V L S
              β β β β β                     β β β β β β β       β β β         β β β β β β


            40        45        50        55        60        65        70        75
B.st E2p    N D K A V V E I P S P V K G K V L E I L V P E G T V A T V - G Q T L I T L D
E.coli E2p  G D Q A S M E V P A P F A G V V K E L K V N V G D K V K T - G S L I M I F E

pea leaf H  S V K A T S D V N S P I S G E V I E V - - N T G L T G K P - G L I N S S P Y
yeast PC    A M K M E M I I S S P S D G Q V K E V F V S D G E N V D S S D L L V L L E D
                    β β β β β         β β β β β β           β β β           β β β β β β
```

Fig. 3. Alignment of amino acid sequences of template lipoyl domains and target proteins. The biotinyl domain is that of yeast pyruvate carboxylase (Lim et al., 1988) and the lipoylated H-protein is from the pea leaf glycine cleavage system (Kim & Oliver, 1990). The sequences of the lipoyl domains begin at the first structured residues of the domains. The aligned sequence of the H-protein begins at residue 56 in the native sequence, that of the carboxylase at position 1102. The underlining at position 42 indicates the posttranslationally modified lysine residues. Proposed key residues are indicated by the sans serif typeface. The consensus secondary structure of the lipoyl domain templates is indicated under the alignment: β = β-strand.

Table 1. *Comparison of key hydrophobic residues from the template lipoyl domain structures with residues at equivalent positions in the sequences of the target proteins*[a]

E. coli		*B. st.*			
Residue	Contact no.	Residue	Contact no.	yPC	plH
I4	3	F4	5	A	I
I16	5	I18	4	I	V
I19	4	W21	7	V	E
V21	2	V23	3	V	P
V27	6	V29	2	I	V
L33	5	L35	3	V	F
V36	6	V38	2	L	V
V45	4	I47	2	I	V
F49	5	V51	4	S	I
V53	2	V55	4	V	V
L56	2	I58	4	V	I
I70	5	L72	6	V	N
M71	5	I73	4	V	S
F73	6	L75	3	E	P

[a] The template lipoyl domain structures are from *Escherichia coli* and *Bacillus stearothermophilus* (*B. st.*); the target proteins are the biotinylated domain of yeast pyruvate carboxylase (yPC) and the lipoylated H-protein of the pea leaf glycine cleavage system (plH). The contact numbers indicate the number of residues making close attractive nonpolar contacts with a given residue side chain.

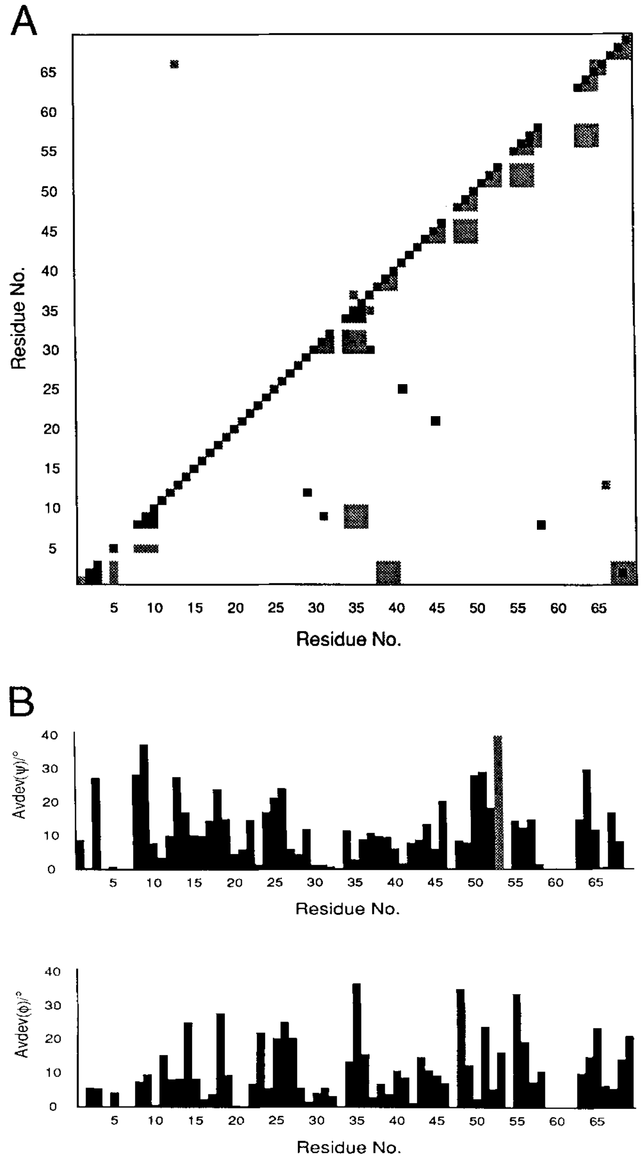

Fig. 4. A: A schematic illustration of the restraints on structurally conserved regions (SCRs) used to calculate the structure of the biotinyl domain from yeast pyruvate carboxylase. The gray scale indicates the number of restraints between a particular pair of residues: the darker the square, the greater the number of restraints. Squares below the diagonal indicate main chain–main chain restraints, those above indicate side chain–side chain restraints. Gaps in the main diagonal indicate variable regions (VRs). **B:** Histogram showing the average deviation of the main chain dihedral angles ϕ and ψ between topologically equivalent residues in the templates, for the SCRs in the biotinyl domain. The gray bar indicates a value >40°.

to stabilize the folded state of the protein. It would seem appropriate to use solvent accessibility when considering protein–protein interactions that are required when groups are not solvated, for example, hydrogen bonds and salt-bridges.

Additional evidence that these proteins may all adopt similar folds is provided by the residues adopting positive ϕ conformations in the template structures. There are glycines at the equivalent positions in the target structures in all cases bar one, where an aspartate is present (D70 for yeast PC in Fig. 3); this is a common substitute for glycine in turns (Wilmot & Thornton, 1990).

Three-dimensional model of the biotinyl domain of yeast pyruvate carboxylase

From the alignment of the template sequences and structures, we define the following regions of the yeast PC sequence to be structurally conserved regions (SCRs): 1-3 (1-3), 5 (5), 16-40 (8-32), 42-54 (34-46), 56-61 (48-53), 63-66 (55-58), and 71-77 (63-69). The numbers in parentheses indicate absolute residue number in the protein sequence; those outside parentheses are according to the numbering in Figure 3. The restraints on the SCRs using absolute residue numbering are illustrated in Figure 4A; the average deviations of the main chain dihedral angles of the SCRs of the template are shown in Figure 4B. There is a much lower number of restraints on the system compared with the number of restraints obtained by analysis of NMR spectroscopic data, as emphasized by equivalent diagonal plots representing restraints on NMR structures (e.g., see Metzler et al., 1992). A schematic view of the biotinyl domain indicating the secondary structural motifs is shown in Figure 5. The β-structure formed (residues 2, 15-23, 26-29, 35-39, 53-55, 63-66, and 76 [Fig. 3 numbering]) falls largely within the expected regions. It should

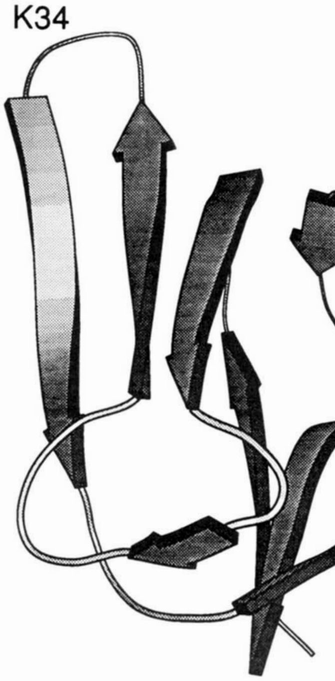

Fig. 5. Schematic drawing of the predicted three-dimensional structure of the biotinyl domain from yeast pyruvate carboxylase. β-Strands are indicated by arrows.

be noted that the definition of secondary structure used here was a conservative one based solely on hydrogen bonding patterns, hence the presence of single residue β-strands (β-bridges). The backbone of the domain with hydrophilic and hydrophobic side chains included is shown in Figure 6A and B, respectively, and in Kinemage 1. The positions of most of the side chains are in good agreement with general features of protein structures: the hydrophilic side chains point out into solution and the interior of the protein consists of nonpolar side chains forming a core. There is an "openness" to the structure that is inherited from the template structures and may be characteristic of this family of structures as a whole. There are five β-turns in the structure as classified by the program TURNPIN: H16–S19 Type II, K22–Q25 Type IV, A32–M35 Type IV, S53–E56 Type IV (near Type II, i.e., this turn falls just outside the region describing a Type II turn), and S60–L63 Type IV (absolute numbering). These are as expected.

The biotinyl prosthetic group is in amide linkage with K34 (Lim et al., 1988) and is thus located in a β-turn. The nonpolar side chains are not as well packed in the model as in the template structures, as evidenced by the nonpolar contact numbers for the nonpolar residues in the model, which are lower than those of the templates (Fig. 7).

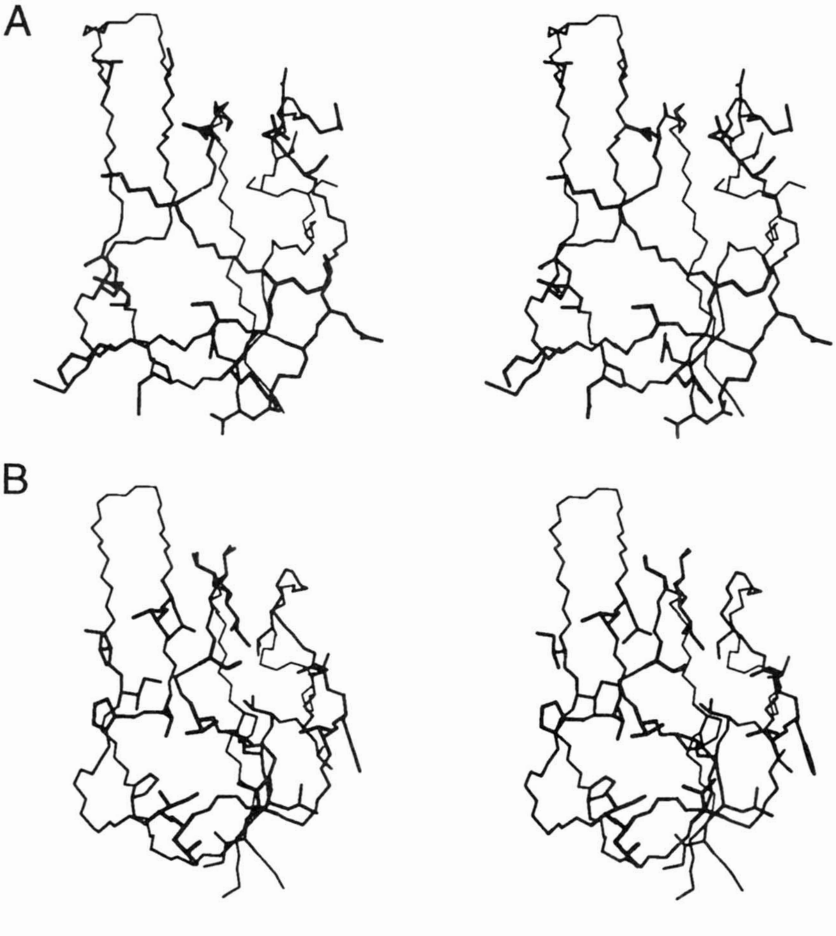

Fig. 6. Stereo representations of the biotinyl domain of yeast pyruvate carboxylase showing hydrophilic side chains (**A**) and hydrophobic side chains (**B**). The orientations are the same as in Figure 5.

```
                  5         10        15        20        25        30        35        40
Contacts  1     1 1         5     3 2 1           1 1 1     2 3 3 2 2               2 2
Yeast PC  H I G A P M A G V I V E V K V H K G S L I K K G Q P V A V L S A M K M E M I I S

                  45        50        55        60        65
Contacts              2     1 2 3           3         2 2 2 2 2
Yeast PC  S P S D G Q V K E V F V S D G E N V D S S G L L V L L E D
```

Fig. 7. The sequence of the biotinyl domain from yeast pyruvate carboxylase showing the nonpolar contact numbers calculated from the predicted three-dimensional structure.

Given the low level of sequence identity between the target and template structures, it should be noted that errors in the main chain as well as in the side chain coordinates are likely to contribute to less than optimal side chain packing. It may be that some relative movement of the β-sheets is required, which would be consistent with observations on other β protein families, where such shifts occur between distantly related members. For this reason we have not optimized the side chain packing (hence the low values of contact numbers), since correct prediction of side chain conformation falls sharply with increasing errors in the main chain (Holm & Sander, 1992).

Three-dimensional model of the H-protein of the pea leaf glycine cleavage system

We designate the following regions of the pea leaf H-protein sequence as SCRs: 1-3 (1-3), 5 (5), 15-38 (15-38), 42-55 (42-55), 64-67 (62-65), and 71-77 (68-74). Again, the numbers in parentheses are absolute residue numbers; those outside are the numbering in Figure 3. The restraints on the SCRs are illustrated in Figure 8A, and the average deviations of the SCR template main chain dihedral angles are shown in Figure 8B.

A schematic view of the three-dimensional structure of the H-protein is shown in Figure 9. Most of the hydrophilic residues point out into solution, and a nonpolar core is formed (Fig. 10A,B; Kinemage 2). As with the biotinyl domain, the packing of the nonpolar side chains in the core is not as good as that of the templates (Fig. 11), so the same caveats apply to this model as to that of the biotinyl domain. The β-structure formed (residues 2, 15-21, 29, 37-39, 45, 53-54, 66, and 76-77 [Fig. 3 numbering]) falls largely within the expected regions. There are β-turns as follows: E24-V27 Type II, T30-K33 Type IV, S40-A43 Type IV, N59-L62 Type IV, and K65-L68 Type IV (near Type II). The turns are as expected. Again there is some openness to the structure. The lipoylated lysine residue (K42) (Fujiwara et al., 1986; Kim & Oliver, 1990) is situated in the β-turn defined by residues 40-43.

We have presented a model for approximately 80 of the 120 residues making up the H-protein. Examination of Table 1 and Figure 2 reveals that two consecutive residues near the C-terminal end of our model of the H-protein that are expected to be hydrophobic are, in fact, hydrophilic (asparagine and serine). This difference may indicate that our model is in error around the region in space

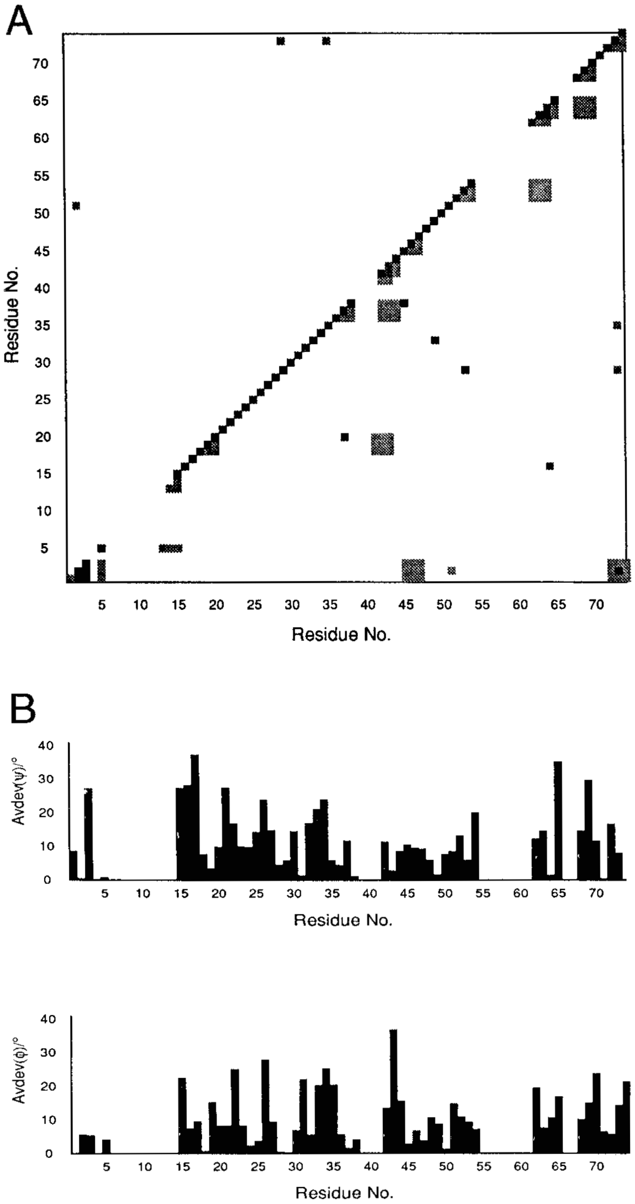

Fig. 8. **A:** Schematic illustration of the restraints on SCRs used to calculate the structure of the lipoylated H-protein from the pea leaf glycine cleavage system. The gray scale indicates the number of restraints between a particular pair of residues: the darker the square, the greater the number of restraints. Squares below the diagonal indicate main chain–main chain restraints, those above indicate side chain–side chain restraints. Gaps in the main diagonal indicate VRs. **B:** Histogram showing the average deviation of the main chain dihedral angles ϕ and ψ between topologically equivalent residues in the templates, for the SCRs in the H-protein.

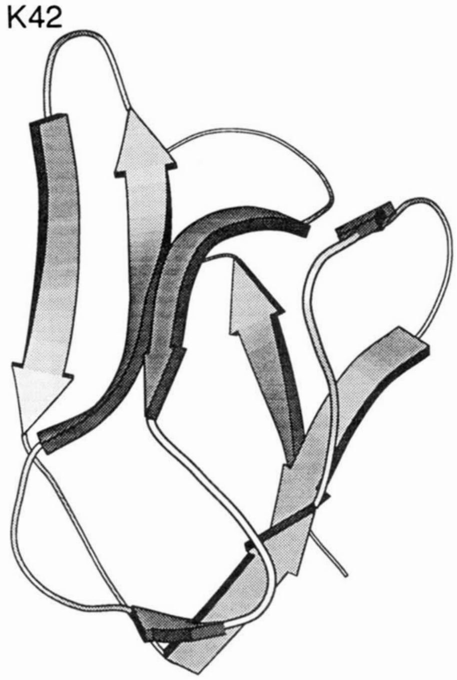

Fig. 9. Schematic drawing of the predicted three-dimensional structure of the H-protein from the glycine cleavage system of pea leaf. β-Strands are indicated by arrows.

surrounding these nonconservative substitutions. This region includes both the N- and C-termini of our model. Perhaps coincidentally, this is precisely the region where the unmodeled sequence is likely to be found in the structure, because both termini are extended by about 20 residues in the full protein. We cannot, however, make any detailed predictions as to the nature of the extra structure, nor its influence on our model.

General comments on the models

The structures of the models presented here indicate that both the H-protein and biotinyl domain are all-β proteins, at least for the regions that are modeled. This is in contrast with previous structure predictions (see below), which indicated that some α-helix was present. Support for our models at the secondary structural level is provided indirectly from attempts to predict the secondary structure of the lipoyl domains from their sequences: these domains are also predicted to have α-helical content, although somewhat less strongly than are H-proteins (S.M. Brocklehurst & R.N. Perham, unpubl.; see also Spencer et al., 1984). This indicates, perhaps, that some common pattern in the sequences of all three families is

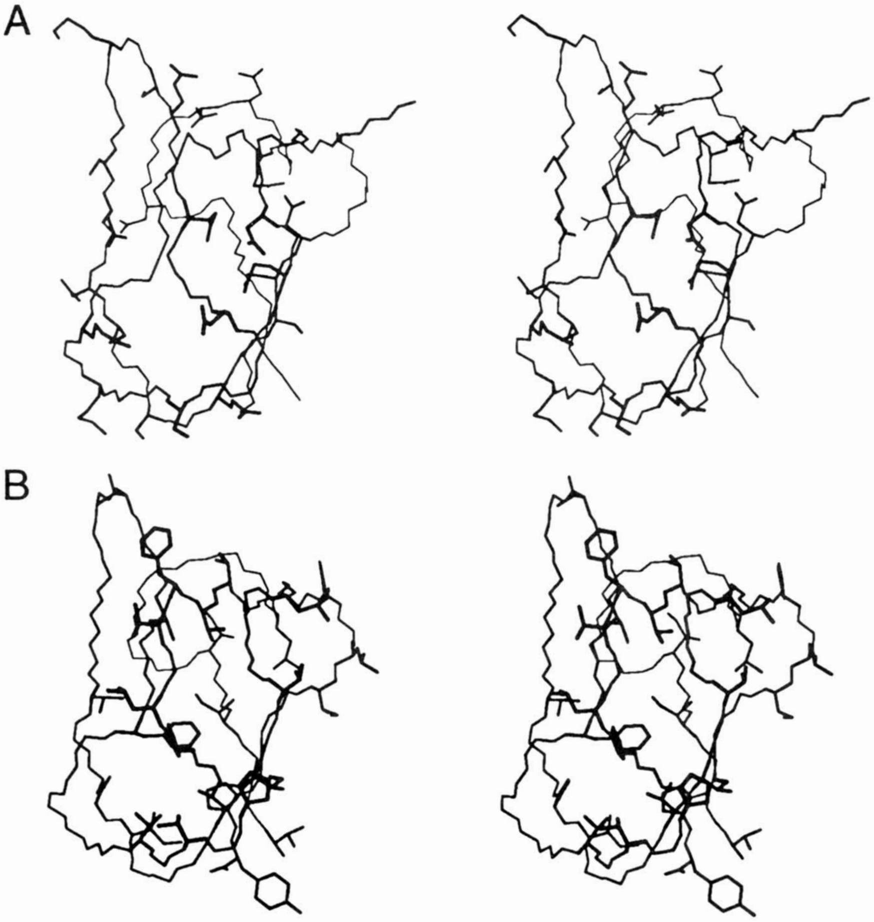

Fig. 10. Stereo representations of the H-protein from the pea leaf glycine cleavage system showing hydrophilic side chains (**A**) and hydrophobic side chains (**B**). The orientations are the same as in Figure 9.

```
                  5         10        15        20        25        30        35        40
Contacts  1 1                     2 2         2 2   1 2   1       2 1 1   1   2   2
Pea H     V A T I G I T D H A Q D H L G E V V F V E L P E P G V S V T K G K G F G A V E S

                  45        50        55        60        65        70
Contacts  1 1   2             1       3 1   2   2   2     1 2     1       3 1
Pea H     V K A T S D V N S P I S G E V I E V N T G L T G K P G L I N S S P Y
```

Fig. 11. The sequence of the H-protein from the pea leaf glycine cleavage system showing the nonpolar contact numbers calculated from the predicted three-dimensional structure.

erroneously detected by secondary structural prediction techniques.

At the tertiary structural level, the predicted strand topologies of the models are, of course, inherited from the template lipoyl domains. The topology is likely to be correct, provided that the conserved pattern of turn-forming and hydrophobic residues identified here uniquely determines the fold of the proteins, as we have postulated. Pairwise comparisons of core residues of the models and the template lipoyl domains give a measure of the influence of the template structures on the targets. The rmsd between the two lipoyl domains is 2.5 Å. The rmsd between the H-protein and the *B. stearothermophilus* lipoyl domain is 1.9 Å, and between the H-protein and *E. coli* lipoyl domain is 2.9 Å. The rmsd between the biotinyl domain and the *B. stearothermophilus* lipoyl domain is 1.7 Å, and between the biotinyl domain and the *E. coli* lipoyl domain is 1.4 Å. The model of the H-protein probably contains more errors than that of the biotinyl domain; some hydrophilic residues point erroneously into the interior of the protein. This is a consequence of the template/target alignment. The openness of the models and the template structures could be reduced significantly by bringing the edges of the open face of the "barrel" closer together by about 1-2 Å. This would be a particularly straightforward operation with our approach, because we could specify this requirement as an additional restraint. In the present work, however, our intention was to predict only the global topology of the target proteins.

Comparison of posttranslationally modified hairpin loops

It is interesting to compare the sequences of the lipoyl domains, biotinyl domains, and H-proteins in the region around the lysine residue to which the prosthetic group is attached, because it may be that other residues important to the process of posttranslational modification and to the function of these proteins are located here. Secondary structure prediction has suggested that in the H-protein this region is α-helical (Fujiwara et al., 1991). However, our models (and, of course, the template/target alignment) of the H-protein and the biotinyl domain show this region forming a β-hairpin loop containing a single β-turn in which the modified lysine residue is situated, as it is in the lipoyl domain of 2-oxo acid dehydrogenase complexes.

If our predicted conservation of structure among the three proteins proves to be real, there have to be superficial structural features that enable the lipoylating enzyme(s) to select the apo-lipoyl domain for lipoylation and the apo-biotinyl domain for biotinylation. In this context, patterns of sequence variability at the domain surface will be of particular interest. For example, the residue immediately preceding the modified lysine residue is conserved within a family but is different across the three families. In the lipoyl domains of 2-oxo acid dehydrogenase complexes it is an aspartate (Perham, 1991), in H-proteins it is a valine (Kim & Oliver, 1990), and in the biotinyl domains of carboxylases it is a methionine (Lim et al., 1988). The implication is that this residue may not be important for the protein fold; indeed, given its predicted position at the protein surface in a β-turn, this would not be unexpected. The conservation of such a residue within a family may, therefore, be important either as a recognition signal for the posttranslational modification of the relevant lysine residue or for the catalytic function. It is interesting that mutation of this residue in a biotinyl domain (Samols et al., 1988) indicates that it is not important for the posttranslational modification, but does affect the transfer of the carboxyl group to the biotin prosthetic group.

Comparison of loops close in space to the modified lysine hairpin

For lipoic acid to act as an effective substrate for the first component enzyme (E1, 2-oxo acid dehydrogenase [lipoamide]) of a 2-oxo acid dehydrogenase complex, the lipoyl group must be attached to the lipoyl domain. Not just is k_{cat}/K_m raised in value by a factor of 10^4, but the lipoyl domain confers specificity on the pendant lipoyl group for reductive acylation only by the E1 component of the parent 2-oxo acid dehydrogenase complex (Graham et al., 1989). This offers a most effective means of substrate channeling (Perham, 1991), but the structural basis remains unclear. What does seem obvious is that there is a process of molecular recognition whereby only the relevant lipoyl domain is selected for productive interaction with a given E1, a process that again must rely on distinguishing surface features of the domain.

The protein surface in the vicinity of the posttranslationally modified lysine residue in the lipoyl domain, biotinyl domain, and H-protein includes, in addition to the lysine hairpin itself, residues from a loop near the N-terminal end of the polypeptide chain. Across the family of 2-oxo acid dehydrogenase complexes, there is considerable variability in the sequence of this loop. In addition, there are frequent single- and double-residue insertions and deletions across species, suggesting that this region may be important in the function of the domains. Interestingly, the sequences of H-proteins from pea leaf (Kim & Oliver, 1990) and chicken (Fujiwara et al., 1986) in this region show no common features either with each other or with the E2 lipoyl domains, although all are lipoylated, and there is a deletion of eight residues in this region in the biotinyl domain (Fig. 3). The absence of this region in biotinyl domains is paralleled by less stringent substrate requirements compared with E2 lipoyl domains, because free biotin can function as a substrate for carboxylation, although attachment to a biotinyl domain improves its effectiveness (Samols et al., 1988). Whether or not this loop has anything to do with function remains to be determined.

Conclusion

We have initiated the development of an automated, restraint-based approach to the modeling of protein tertiary structure. Our aim has been to construct models of proteins by using information directly related to interactions that stabilize the folded state. In this work, we have considered only conserved main chain–main chain hydrogen bonds and close attractive nonpolar contacts. In combination with dihedral angle and inter-C_α distance information, models of proteins with realistic structure may be constructed by our method. It is important to realize, however, that hydrogen bonds in protein structures are much more conserved than are van der Waals contacts. This is in accordance with the observation that equivalent helices may rotate and translate relative to each other in homologous proteins but that the movement of β-strands is more restricted and occurs at the sheet level, where sliding may occur (although more local conformational changes can occur within strands than within helices). In order properly to take account of secondary structural shifts, we will need separate simulations utilizing novel hydrophobic potentials. We are currently developing such procedures.

As it stands, our method is equivalent to a number of other procedures in which good models may be produced where the templates cluster around the target structures. An important advantage of our method over previously published restraint-based modeling methods is that we place minimal reliance on restraints that do not represent interactions that stabilize the folded state of the target. This allows for the possibility of inclusion of experimentally determined restraints, e.g., from incomplete or uninterpretable NMR or X-ray data, without the system being dominated by a multitude of theoretical interatomic distance restraints. Our de novo loop-building method has the property of producing loops that have no bad steric clashes with the rest of the protein, and whose main chain conformation falls within allowed regions of ϕ/ψ space. Where the quality of the template structures allows, we may further restrict the conformation of the loop if the role of "key" residues can be pinpointed; for example, side chain–main chain hydrogen bonds in β-turns or hydrophobic interactions with the core may be important in some instances.

Structure–function studies in this laboratory are elucidating the role of particular residues in the lipoyl domains (N.G. Wallis & R.N. Perham, unpubl.), and the models of the lipoylated H-protein and the biotinylated domain from pyruvate carboxylase will facilitate a comparison of the mechanisms by which these molecules function and are posttranslationally modified. The overall folds of the models have been inherited, to a large extent, from the template lipoyl domains. It is interesting to note that despite the apparent openness of the fold, resulting in some of the hydrophobic core being exposed to solvent, some regions of this fold appear to be highly stable. This is reflected in the resistance of the lipoyl domain to proteolysis (Perham [1991] and references therein) and in the presence of some very slowly exchanging protons in the structures (Dardel et al., 1993).

The models are necessarily speculative given the low (less than 15%) sequence identity with the template lipoyl domain structures, but the pattern of key residues identified here strongly suggests that the lipoyl domain, biotinyl domain, and H-protein are all based on a common structural motif, a new protein module. Given the weak sequence similarity between the template and target structures, it must be emphasized that our models are approximate and it may be expected that some secondary and supersecondary structural drifts may occur in the biotinyl domain and H-protein when compared with the lipoyl domain. Further speculation must await the determination of the structure of an H-protein and a biotinyl domain.

Theory and methods

Structural analysis

The existence of particular interatomic physicochemical interactions in a molecule may be inferred from an analysis of its three-dimensional atomic coordinates. We consider here main chain–main chain hydrogen bonds and side chain–side chain nonpolar contacts, because these seem likely to be important in determining three-dimensional structures of proteins.

1. Hydrogen bonds

We use the following geometric criteria to delineate the existence of an NH . . . O=C hydrogen bond:

$$r(H\ldots O) < 2.5\ \text{Å}$$

$$ang(H\ldots O{=}C) > 90°$$

These criteria reasonably well reflect the "kidney-bean" shaped electron density around carbonyl oxygen atoms and allow bifurcated hydrogen bonds to be identified reliably. We calculate the positions of protons where these are not available directly.

2. Side chain–side chain nonpolar contacts and contact number

We represent each side chain carbon atom together with its covalently bonded proton(s) in the amino acid residues A, F, I, L, V, W, Y, M, H, K, P, and T as a sphere of diameter 4.5 Å. Where two such spheres from different residues interpenetrate, a nonpolar contact is inferred. Where the side chains contain other atoms or groups, e.g., an oxygen in T or a sulfur in M, these are not considered, because they may be involved in separate interactions that stabilize the tertiary structure. We define a nonpolar contact number for a residue that is simply the number of nonpolar contacts that a residue makes (only one contact per residue pair is counted). Our contact number is directly correlated with core side chain definition in a structure and thus may be used as a measure of the quality of the model.

Model building overview

The starting point for building a model is, of course, the alignment of the templates and target sequences. We perform the structure-based alignments manually but in a systematic way by using a subset of the structural information available from the templates. Usually information pertaining to hydrogen bond geometries, general definitions of secondary structural motifs (α-, 3_{10}-, π-helices, β-strands, β-turns) and supersecondary structural motifs (β-hairpin loops), and detailed main chain conformation is used simultaneously (Sali & Blundell, 1990). The sequence of the target is aligned to the templates in such a way as to minimize gaps in the secondary structural motifs. In the alignment of the sequences presented here, we choose to leave out atomic-packing information in the alignment procedure in order not to bias information regarding the position of "key" hydrophobic residues.

From the alignment we delineate SCRs as those regions where the average deviations of the main chain dihedral angles (ϕ,ψ) differ by not more than a particular value (usually 40°). The variable regions (VRs) are those regions not defined as SCRs, and most often are loops joining secondary structural motifs. These definitions correlate with SCR definitions based on rigid body superposition of structures where no shifts in the relative positions of elements of structure occur. The delineation of SCRs based on dihedral angles is suitable for defining SCRs in protein families where such secondary structural shifts do occur. There is a loss of sensitivity in our approach compared with methods that make use of spatial superposition (Sutcliffe et al., 1987) in rare cases where large, complementary changes in main chain dihedral angles result in spatially equivalent positions across a family. Geometric restraints on a target structure are calculated from a consideration of the template structures and the target sequence. The above procedures have been incorporated into a new computer program, NAOMI, which combines molecular modeling, structure analysis, and a protein-specific database management system. Three-dimensional models of target structures are constructed from the calculated restraints by using a methodology that is essentially the same as that commonly used in deducing structures from multidimensional NMR spectroscopic data, based on the technique introduced by Nilges and coworkers (1988a,b), i.e., a combination of high-temperature molecular dynamics and slow cooling simulations using nonphysical potentials (cf. Sali et al. [1990] and Havel & Snow [1991]). We use the program XPLOR version 2.1 (Brünger, 1990) for the simulated annealing calculations.

Template/target alignment

The templates are aligned by maximizing the similarities of certain structural features and interresidue relationships. These are described below.

1. Main chain conformation

The main chain conformation of an amino acid residue may be described by the region of (ϕ,ψ) space it occupies. We use the nomenclature of Wilmot and Thornton (1990), which is based on that of Effimov (1986). Briefly, we assign each residue a conformational identifier; either α_R (right-handed helix), β (beta strand), β_p (conformation adopted by prolines), α_L (left-handed helix), γ_L and ε (conformations adopted primarily by glycines), and - (other).

2. Hydrogen bonds

These are as delineated as described above.

3. Secondary structural motifs

Secondary structural motifs are delineated according to a new algorithm that considers hydrogen bonds and three conformational properties of the polypeptide chain. These are the main chain dihedral angles, ϕ and ψ, and the angle defined by the three atoms $C_{\alpha i-2}$, $C_{\alpha i}$, and $C_{\alpha i+2}$. The algorithm is equally well suited for use with both high and low resolution structures.

4. Supersecondary structural motifs

At present we consider only β-hairpin loops (Sibanda et al., 1989), which are classified by a modified form of the algorithm used in the program TURNPIN developed by Y.J. Edwards and S.M. Brocklehurst (unpubl.). The algorithm facilitates rapid pattern recognition of hydrogen bond and secondary structural motifs, including recognition of distortions, e.g., β-bulges.

Calculation of restraints

1. General form of restraints

The restraints on the target structure are calculated by generation of a mean (μ) and a mean absolute deviation (hereafter referred to as average deviation, AvDev) of a geometrical feature that is present in the template structures. They are of the form

$$\mu \pm \zeta . \text{AvDev},$$

where ζ is an empirical weight function. It should be noted that these are restraints and not constraints on the structure, so small violations are acceptable and expected in the calculated structures. In the dynamics simulations, the potential energy function rises smoothly but steeply where the restraints are violated.

2. Dihedral angle restraints

We calculate restraints for the main chain dihedral angles (ϕ, ψ) for those residues in SCRs. The restraints on ϕ are calculated, viz.:

$$\bar{\phi}^h_{s,r} = \sum_{\text{SCRs}}^{\text{S}} \sum_{\text{residues}}^{\text{R}} \frac{1}{H} \sum_{\text{homologues}}^{H} \phi^h_{s,r}$$

$$\text{AvDev}(\phi^h_{s,r}) = \sum_{\text{SCRs}}^{\text{S}} \sum_{\text{residues}}^{\text{R}} \frac{1}{H} \sum_{\text{homologues}}^{\text{H}} |\phi^h_{s,r} - \bar{\phi}^h_{s,r}|,$$

where S is the number of SCRs, R is the number of residues in a given SCR, and H is the number of homologous structures used to produce one model ($H > 1$). Restraints on ψ are calculated analogously.

The periodic nature of dihedral angles is taken into account when calculating these statistics, so that for example the mean of $-170°$ and $+170°$ is $180°$, and not $0°$. In the present work, we set ζ to 1.

3. Hydrogen bond restraints

Only main chain-main chain hydrogen bonds that are completely conserved across all the templates are included in the present procedure. We calculate hydrogen bond restraints by calculating statistics on the interatomic distances r(N...O) and r(N...C). The distance r(N...C) is correlated with the angle ang(NOC); angles (except dihedral angles) cannot be used as restraints in XPLOR. The equations for generating these restraints are analogous to those for the calculation of dihedral angle restraints (see above). We have found the use of both distance and "angle" information essential to produce structures with good geometry in the absence of C_α-C_α restraints.

4. Inter-SCR endpoint restraints

Secondary structural motifs often shift relative to one another in distantly related proteins. In general, we allow the elements of conserved secondary structure to make rotational and translational shifts to a greater extent than the secondary motifs in the templates shift relative to one another. Simultaneously, the secondary structural elements are permitted distortions, where appropriate, to make required interactions. Thus the secondary structural elements are not represented as rigid bodies. We do this by calculating means and average deviations between atoms near the endpoints of SCRs.

Typically, one, two, or three residues at the endpoints of two SCRs connected by a loop are used. The structural features used to calculate the endpoint restraints are shown in Figure 12. The weights are set to 1. In general, it seems desirable to use minimal numbers of restraints that do not represent interactions stabilizing the folded state of the target, but additional such sets of restraints between any regions of structure can be included to improve the convergence of the simulations and to force the global conformation of the main chain to resemble closely that of the template structures. This makes the method

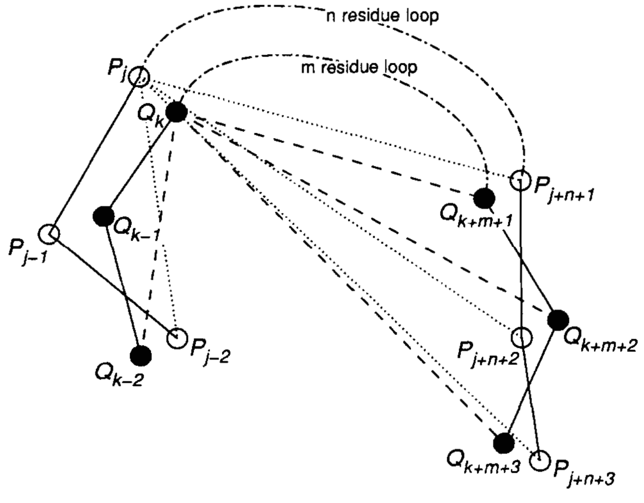

Fig. 12. Illustration of some of the distances used in the calculation of inter-C_α restraints. The C_α atoms of two segments, each of three residues, of polypeptide chain are shown for two template structures P and Q. The C_α's of residues of structure P are represented by open circles, and of structure Q by shaded circles. Virtual bonds are shown by solid lines. Loops of n and m residues, respectively, join the segments of P and Q illustrated. Mean and average deviations of distances between pairs of atoms are calculated for equivalent pairs of atoms in the templates. Distances for one pair of topologically equivalent residues are indicated by broken lines joining the C_α atoms. A nonredundant list of statistics on such distances is constructed by considering each topologically equivalent residue position; the remaining lines are omitted for clarity.

similar to other modeling approaches where the results are good if the template structures cluster around the target.

It should be possible to omit these restraints entirely, however, if limited experimental information is available, for example, from incomplete NMR or X-ray data. Restraint-based modeling methods are particularly well suited for incorporation of such experimental information. It may be that such unification of experimental and theoretical information will become important for rapid but accurate structure determination in the future.

5. Variable region restraints

Each residue in a VR has dihedral angle restraints based on general features of protein folds, which exclude regions of (ϕ, ψ) space that are rarely occupied. Proline residues are particularly restricted in the conformations they may adopt, even taking both *cis* and *trans* isomers into account. The residue immediately preceding proline is usually restricted also. In most cases we also restrict amino acids to negative ϕ values, except for G, N, D, and S, which show a preference for positive ϕ and are abundant in β-turns (Wilmot & Thornton, 1990). These restrictions are necessary since the potentials used here allow the peptide planes to "flip". In addition, the use of dihedral angle restraints in the system improves the convergence properties of the annealing protocol.

6. Side chain restraints

In the present work we place restraints on the atomic positions of C_β's in side chains involved in conserved van der Waals interactions. In addition, we transfer $\chi 1$ dihedral angles from templates to target in a manner similar to that of Summers and Karplus (1991), except that the dihedral angles are represented as restraints calculated analogously to the main chain dihedral restraints (see above).

Structure calculations

Initial partly folded structures are calculated that possess mean values of the main chain dihedral angles for the SCRs from the templates. The conformations of VR residues are initially extended but randomized to provide different starting structures for the simulated annealing calculations. The simulated annealing protocol consists of four phases. In the first phase, 5 ps of molecular dynamics (a time step of 2 fs is facilitated by using the SHAKE algorithm) is performed using a soft potential that allows atoms to pass through one another. In the second phase, the van der Waals radii are increased over a further 3 ps of dynamics. In the third phase, after switching to a square-well potential, the system is cooled from 1,000 K to 300 K in 25 K steps, 0.3 ps of simulation being performed at each intermediate temperature. Finally, 200 steps of Powell minimization are performed to regularize the geometry of the molecule. The computer program XPLOR version 2.1 (Brünger, 1990) was used for the molecular dynamics simulations. The potentials used in the present work are described in the XPLOR manual using default values for all force constants.

Models

Three-dimensional models of the biotinylated domain of yeast PC and the lipoylated pea leaf H-protein were constructed using the procedures described above, based on the structures of the lipoyl domains of the *B. stearothermophilus* (Dardel et al., 1991, 1993) and *E. coli* (J.D.F. Green, E.D. Laue, & R.N. Perham, unpubl.) PDH multienzyme complexes.

Computational resources

The program NAOMI was written in ANSI C and run on a Microvax 3100. XPLOR was run on a Meiko Computing Surface with 20 transputers. The calculated structures were examined on an SGI 4D/70GT workstation by using the program QUANTA release 3.2 and on an Evans and Sutherland PS390 driven by a Microvax 3100 by using the program HYDRA (Hubbard, 1986). Each model took approximately 18 h CPU time to construct.

Acknowledgments

This work was supported by the Science and Engineering Research Council. We are grateful to the Protein Engineering Initiative of SERC for the award of an earmarked Research Studentship to S.M.B. We thank Drs. J.D.F. Green and F. Dardel for agreeing to make coordinates available before publication and members of the Cambridge Centre for Molecular Recognition, especially Drs. Y.N. Kalia, P.J. Kraulis, and A.R.C. Raine, for stimulating discussions. We thank Dr. E.D. Laue for the use of the Meiko Computing Surface and Dr. J.C.J. Barna for assistance with the use of various computers.

References

Ali, S.T. & Guest, J.R. (1990). Isolation and characterization of lipoylated and unlipoylated domains of the E2p subunit of the pyruvate dehydrogenase complex of *Escherichia coli*. *Biochem. J. 271*, 139-145.

Borges, A., Hawkins, C.F., Packman, L.C., & Perham, R.N. (1990). Cloning and sequence analysis of the genes encoding the dihydrolipoamide acetyltransferase and dihydrolipoamide dehydrogenase components of the pyruvate dehydrogenase multienzyme complex of *Bacillus stearothermophilus*. *Eur. J. Biochem. 194*, 95-102.

Brünger, A.T. (1990). *XPLOR Manual Version 2.1.* Yale University, New Haven, Connecticut.

Dardel, F., Davis, A.L., Laue, E.D., & Perham, R.N. (1993). The three-dimensional structure of the lipoyl domain from *Bacillus stearothermophilus* pyruvate dehydrogenase multienzyme complex. *J. Mol. Biol. 229*, 1037-1048.

Dardel, F., Laue, E.D., & Perham, R.N. (1991). Sequence-specific ^{1}H NMR assignment and secondary structure of the lipoyl domain of the *Bacillus stearothermophilus* pyruvate dehydrogenase multienzyme complex. *Eur. J. Biochem. 201*, 203-209.

Desmet, J., de Maeyer, M., Hayes, B., & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side chain positioning. *Nature 356*, 539-542.

Effimov, A.V. (1986). Standard conformations of a polypeptide chain in irregular regions of proteins. *Mol. Biol. (Moscow) 20*, 250-260.

Fujiwara, K., Okamura-Ikeda, K., & Motokawa, Y. (1986). Chicken liver H-protein, a component of the glycine cleavage system – Amino acid sequence and identification of the N-epsilon lipoyl lysine residue. *J. Biol. Chem. 261*, 8836-8841.

Fujiwara, K., Okamura-Ikeda, K., & Motokawa, Y. (1991). Lipoylation of H-protein of the glycine cleavage system. *FEBS Lett. 293*, 115-118.

Graham, L.D., Packman, L.C., & Perham, R.N. (1989). Kinetics and specificity of reductive acetylation of lipoyl domains from 2-oxo acid dehydrogenase multienzyme complexes. *Biochemistry 28*, 1574-1581.

Hale, G., Wallis, N.G., & Perham, R.N. (1992). Interaction of avidin with the lipoyl domains in the pyruvate dehydrogenase multienzyme complex: Three-dimensional location and similarity to biotinyl domains in carboxylases. *Proc. R. Soc. Lond. B 248*, 247-253.

Havel, T.F. & Snow, M.E. (1991). A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol. 217*, 1-7.

Hiraga, K. & Kikuchi, G. (1980). The mitochondrial glycine cleavage system. Functional association of glycine decarboxylase and aminomethyl carrier proteins. *J. Biol. Chem. 255*, 11671-11676.

Holm, L. & Sander, C. (1992). Fast and simple Monte-Carlo algorithm for side-chain optimization in proteins – Application to model building by homology. *Proteins Struct. Funct. Genet. 14*, 213-223.

Hubbard, R.E. (1986). HYDRA: Current and future developments. In *Computer Graphics and Molecular Modelling* (Fletterick, R. & Zoller, M., Eds.), pp. 9-12. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Kalia, Y.N., Brocklehurst, S.M., Hipps, D.S., Appella, E., Sakaguchi, K., & Perham, R.N. (1993). The high resolution structure of the peripheral subunit-binding domain of dihydrolipoamide acetyltransferase from the pyruvate dehydrogenase multienzyme complex of *Bacillus stearothermophilus*. *J. Mol. Biol.*, in press.

Kim, Y.H. & Oliver, D.J. (1990). Molecular cloning, transcriptional characterization, and sequencing of cDNA encoding the H-protein of the mitochondrial glycine decarboxylase complex in peas. *J. Biol. Chem. 265*, 848-853.

Kraulis, P.J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr. 24*, 946-950.

Lim, F., Morris, C.P., Occhiodoro, F., & Wallace, J.C. (1988). Sequence and domain structure of yeast pyruvate carboxylase. *J. Biol. Chem. 263*, 11493-11497.

Mattevi, A., Obmolova, G., Schulze, E., Kalk, K.H., Westphal, A.H., Kok, A., & Hol, W.G. (1992). Atomic structure of the cubic core of the pyruvate dehydrogenase multienzyme complex. *Science 255*, 1544-1550.

Metzler, W.J., Valentine, K., Roebber, M., Marsh, D.G., & Mueller, L. (1992). Proton resonance assignments and three-dimensional solution structure of the ragweed allergen *Amb a* V by nuclear magnetic resonance spectroscopy. *Biochemistry 31*, 8697-8705.

Nilges, M., Clore, G.M., & Gronenborn, A.M. (1988a). Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Lett. 229*, 317-324.

Nilges, M., Clore, G.M., & Gronenborn, A.M. (1988b). Determination of three-dimensional structures of proteins from inter-proton distance data by dynamical simulated annealing from a random array of atoms. *FEBS Lett. 239*, 129-136.

Perham, R.N. (1991). Domains, motifs and linkers in 2-oxo acid dehydrogenase multienzyme complexes – A paradigm in the design of a multifunctional protein. *Biochemistry 20*, 8501-8512.

Radford, S.E., Laue, E.D., Perham, R.N., Martin, S.R., & Appella, E. (1989). Conformational flexibility and folding of synthetic peptides representing an interdomain segment of polypeptide chain in the pyruvate dehydrogenase multienzyme complex of *Escherichia coli*. *J. Biol. Chem. 264*, 767-775.

Reed, L.J. & Hackert, M.L. (1990). Structure-function relationships in dihydrolipoamide acyltransferases. *J. Biol. Chem. 265*, 8971-8974.

Robien, M.A., Clore, G.M., Omichinski, J.G., Perham, R.N., Appella, E., Sakaguchi, K., & Gronenborn, A.M. (1992). Three-dimensional solution structure of the E3-binding domain of the dihydrolipoamide succinyltransferase core from the 2-oxoglutarate dehydrogenase multienzyme complex of *Escherichia coli*. *Biochemistry 31*, 3463-3471.

Sali, A. & Blundell, T.L. (1990). Definition of general topological equivalence in protein structures. *J. Mol. Biol. 212*, 403-428.

Sali, A., Overington, J.P., Johnson, M.S., & Blundell, T.L. (1990). From comparisons of protein sequences and structures to protein modelling and design. *Trends Biochem. Sci. 15*, 235-239.

Samols, D., Thornton, C.G., Murtif, V.L., Kumar, G.K., Haase, F.C., & Wood, H.G. (1988). Evolutionary conservation among biotin enzymes. *J. Biol. Chem. 263*, 6461-6464.

Sibanda, B.L., Blundell, T.L., & Thornton, J.M. (1989). Conformation of beta hairpins in protein structures – A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol. 206*, 759-777.

Sieker, L., Cohen-Addad, C., Neuberger, M., & Douce, R. (1991). Crystallographic data for H-protein from the glycine decarboxylase complex. *J. Mol. Biol. 220*, 223-224.

Spencer, M.E., Darlison, M.G., Stephens, P.E., Duckenfield, I.K., & Guest, J.R. (1984). Nucleotide sequence of the *suc*B gene encoding the dihydrolipoamide succinyl transferase of *Escherichia coli* K12 and homology to the corresponding acetyl transferase. *Eur. J. Biochem. 141*, 361-374.

Summers, N.L. & Karplus, M. (1991). Modelling of side chains, loops, and insertions in proteins. *Methods Enzymol. 202*, 156-204.

Sutcliffe, M.J., Haneef, I., Carney, D., & Blundell, T.L. (1987). Knowledge-based modelling of homologous proteins. 1. Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng. 1*, 385-392.

Texter, F.L., Radford, S.E., Laue, E.D., Perham, R.N., Miles, J.S., & Guest, J.R. (1988). Site-directed mutagenesis and ^{1}H-NMR spectroscopy of an interdomain segment in the pyruvate dehydrogenase multienzyme complex of *Escherichia coli*. *Biochemistry 27*, 289-296.

Wilmot, C.J. & Thornton, J.M. (1990). β-Turns and their distortions: A proposed new nomenclature. *Protein Eng. 3*, 479-493.