Molecular Modeling With Substructure Libraries Derived From Known Protein Structures

Barry C. Finzel, S. Kimatian, D. H. Ohlendorf,
J.J. Wendoloski, M. Levitt*, and F.R. Salemme

Central Research & Development Department
Du Pont Experimental Station ES228/320
Wilmington, DE 19880-0228

* Department of Cell Biology
Stanford University
Stanford, CA 94305

ABSTRACT

Many studies have illustrated that proteins are hierarchical assemblies of localized substructures. Here we describe the organization and use of an interactive computer program that allows the graphical construction of protein models using fragments extracted from known x-ray crystal structures. An initial α-Carbon conformational template is used to screen and retrieve matching polypeptide fragments from a library of known protein structures. Fragments are evaluated either graphically or by RMS fit with the template, and appropriately incorporated into the developing molecular model. Amino acid side-chain conformations can also be extracted from known structures or from a library of standard rotamer conformations. A flexible method for specification of target geometry enables identification of substructures that conform to a variety of structural or amino acid sequence contexts. The method has been used to aid model building from crystallographic electron density maps, for homology model building, and in structure analysis applications relevant to protein engineering.

INTRODUCTION

Structural studies of proteins have shown them to be organized on the lines of a relatively limited number of structural motifs. The recognition of these motifs has generally followed an understanding of the hierarchical organization of protein structure. The secondary structural elements that predominate in fibrous proteins were understood first. Later, when more globular protein structures became known, longer range structural motifs were recognized and classified (e.g. Richardson,1981;Weber, et al.,1980; Richmond et al.,1978; Chothia et al.,1977). In practical terms these observations had relatively little impact on detailed molecular modeling tasks, such as required to build protein structures into x-ray electron density maps. Fundamentally, this reflects the relatively limited utility of knowing general features of protein structure when the task at hand involves construction of a unique and detailed molecular model that must conform to experimental data. However, the combination of interactive computer graphics, and an extensive substructure library derived from well resolved protein crystal structures, now make it possible to use detailed information about structural precedents in building models of new proteins.

Traditional implementations of computer graphics for protein modelling have concentrated on emulating the Kendrew models used originally to construct physical models of proteins. Molecular manipulations have been largely limited to the molecular degrees of freedom such as rotations about dihedral angles. Graphics programs popular with protein crystallographers, such as FRODO (Jones; 1985, 1978), have provided additional flexibility, and allow specific atoms or groups of atoms to be disconnected and relocated independent of the rest of the molecule, thereby making it easier to fit structural elements to an electron density map. Standard bond lengths, angles and torsion angles can then be restored by application of regularization procedures (Hermans et al., 1974; Jones et al., 1984). This is a powerful approach, but relies heavily on the model builder's knowledge of protein structure and conformation, since the regularization procedure cannot readily overcome barriers between local minima.

An alternative to conventional graphic modelling involves the utilization of fragments from known protein structures. Jones et al.(1986) have recently demonstrated that the entire backbone of retinol binding protein can be assembled from 22 fragments between 4 and 16 residues in length. The present work amplifies and extends this pioneering application and illustrates how the simultaneous utilization of both a backbone structural fragment and a side-chain rotamer library can substantially aid in the rapid construction of protein models from electron density maps, and additionally provide a powerful tool for structural analysis and protein engineering.

Figure 1 schematically illustrates the architecture of FRAGLE (FRAGment Locate and Exchange), a comprehensive interactive program implemented in this laboratory to examine the utility of modeling protein structures from fragments. It illustrates how commands invoked by the user direct atomic coordinate data from a library of known protein structures, through a variety of screening procedures, to eventual incorporation in a new molecular model. The design fulfills two critical requirements of a useful model building tool: 1) flexibility for use in a wide variety of applications, and 2) enough speed to allow interactive use. The following paragraphs describe specific details of the implementation.

Database Definition.

In order to allow rapid and flexible search of a large number of protein structures, structural information from the Protein Data Bank (Bernstein et al.,1977) is condensed to a more readily accessible form. The residue-indexed binary file format utilized by FRODO (DSN2) is convenient for rapid retrieval of selected residue ranges, so Data Bank structures are cast in this format. Much greater efficiency in searching can be realized by precomputing distances between all Cα atoms in each polypeptide chain, which can then be used for preliminary evaluations of chain geometry (Jones et al.,1985). The interatomic distances d(ij), together with amino acid sequence information and pointers to more complete atomic coordinate data (PDB DSN2), are written into a Compacted Library of Known Protein Structures (Figure 1). To speed all comparisons to structures in the library, the distances are stored and manipulated as integers (e.g. Å*10). The selection of structures to be included is determined upon the basis of structure resolution and refinement criteria. Most characterized elements of protein structure are well represented in a relatively small number of protein structures that have been determined at high resolution (>1.8Å) and refined to R-values in the middle teens. For many model building applications, a library with only these few structures will suffice. Other applications, such as investigations involving structure/sequence correlation or large structural motifs, often benefit from use of a larger database. For this reason a library specification mechanism has been implemented which makes it possible to readily change from one Compacted Library to another. Since new structural data is constantly being obtained, appropriate tools for managing the library have been developed. Table 1 gives a representative list of structures included in a typical working library.

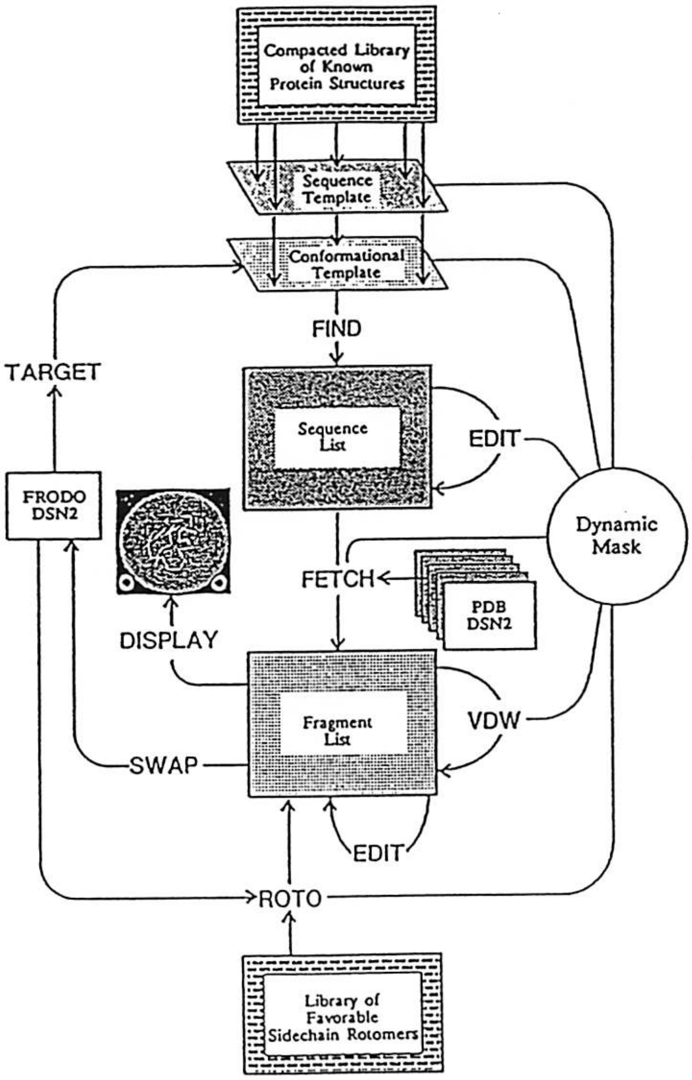

Figure 1. Schematic illustration of FRAGLE program architecture. Squares denote specific data structures; heavy lines and arrows illustrate data flow and the command under whose control the transfer is made. Other constructions denote logical conditions which influence data manipulation. The Sequence and Conformational Templates define chacteristics of protein fragments allowed to cascade from the Library of Known Structures to the Sequence List during execution of a FIND command. FETCH causes complete atomic coordinate information for fragments to be loaded from disk files (PDB DSN2) to memory (the Fragment List), from which they may be displayed, or selected to replace the original target. Sidechains conformations may be assembled from the Library of rotamers with the ROTO command. Many operations are influenced by the status of the Dynamic Mask; a logical construction to enable interactive user control of fragment manipulation (See text).

Table 1 : Fragment Structure Library
(46 Structures)

ID	No. Res.	Name	PDB File
P450CAMA	405	CYTOCHROME P450 (10-250)	PDB$:NATCAM108
BPN'7113	275	SUBTILISIN BPN' (GENEX)	PDB$:SUBT80.DN2
CCP	29	CYTOCHROME C PEROXIDASE	PDB$:1CYP.DN2
CRAMBIN	46	CRAMBIN (1.5A STRUCTURE)	PDB$:1CRN.DN2
LYSOZYME	130	HUMAN LYSOZYME	PDB$:1LZ1.DN2
LC.DHFR	162	DIHYDROFOLATE REDUCTASE	PDB$:3DFR.DN2
CPEPTDAS	307	CARBOXYPEPTIDASE-A	PDB$:5CPA.DN2
G-PEROXY	185	GLUTATHIONE PEROXIDASE	PDB$:1GP1.DN2
TUNACYTC	103	CYTOCHROME C (REDUCED)	PDB$:4CYT.DN2
MYOGLOBN	153	DEOXY-MYOGLOBIN	PDB$:1MBD.DN2
HEMRTHRN	113	MET-HEMERYTRHIN	PDB$:1HMQ.DN2
PENPEPSN	323	PENICILLOPEPSIN	PDB$:2APP.DN2
SG.PRO-A	181	STREPT GRES. PROTEASE-A	PDB$:2SGA.DN2
WG AGGLU	170	AGULUTININ WHEAT GERM	PDB$:3WGA.DN2
A-L PROT	198	ALPHA-LYTIC PROTEASE	PDB$:2ALP.DN2
AZURIN	130	AZURIN	PDB$:2AZA.DN2
CHYM PI	131	A-CHYMOTRYPSIN PT.I	PDB$:5CHA.DN2
CHYM PII	97	A-CHYMOTRYPSIN PT.II	PDB$:5CHA.DN2
CIT SYN	437	CITRATE SYNTHASE	PDB$:1CTS.DN2
MOLI-CC'	127	CYTOCHROME C'	PDB$:2CCY.DN2
CC3	107	CYTOCHROME C3	PDB$:2CDV.DN2
ERYTHROC	136	ERYTHROCRUORIN	PDB$:1ECA.DN2
CYC RICE	111	CYTOCHROME C (RICE)	PDB$:1CCR.DN2
HEME A	141	HEMOGLOBIN (A-SUBUNIT)	PDB$:2HHB.DN2
HEME B	146	HEMOGLOBIN (B-SUBUNIT)	PDB$:2HHB.DN2
IMMUNO	114	IMMUNOGLOBULIN	PDB$:2RHE.DN2
LYZ T4	164	LYSOZYME (T4 PHAGE)	PDB$:2LZM.DN2
R PROT	68	L7/L12 50S RIBOSOMAL PRO	PDB$:1CTF.DN2
PLASTOC	99	PLASTOCYANIN	PDB$:5PCY.DN2
RNASE-X	124	RIBONUCLEASE-X (GENEX)	PDB$:1RSM.DN2
TRYPSIN	223	TRYP/P-AMID-PHEN-PYRU	PDB$:1TPP.DN2
TR INHIB	58	TRYPSIN INHIBITOR	PDB$:4PTI.DN2
PARAVALB	108	CARP PARVALBUMIN	PDB$:1CPV.DN2
ACD PROT	330	ACID PROTEINASE	PDB$:4APE.DN2
C ANHYD	256	CARBONIC ANHYDRASE	PDB$:2CAB.DN2
GII CRYS	174	GAMA II CRYSTALLIN	PDB$:1GCR.DN2
DHFR ECO	159	E-COLI DHFR	PDB$:4DFR.DN2
ERABUTOX	62	ERABUTOXIN B	PDB$:2EBX.DN2
FLAVODOX	138	FLAVODOXIN	PDB$:4FXN.DN2
KALKRN A	80	KALLIKREIN A (A16 A95)	PDB$:2PKA.DN2
KALKRN B	152	KALLIKREIN A (B95 B246)	PDB$:2PKA.DN2
LYZ TRIC	129	LYSOZYME TRICLINIC	PDB$:1LZT.DN2
OVO	56	OVOMUCOID 3RD DOMAIN	PDB$:2OVO.DN2
SGPB/E	185	PROTEINASE B (STREP)ENZ	PDB$:3SGB.DN2
SGPB/I	50	PROTEINASE B(STREP)INHIB	PDB$:3SGB.DN2
RMCP II	224	RAT MAST CELL PROTEINASE	PDB$:3RP2.DN2

Target Specification.

The first step in a search is to define characteristics of desired structural elements. A 'target' is defined as a residue range (or ranges) of partial atomic coordinates from the current molecular model which may potentially be replaced by fragments selected from the library of known structures. The target specification establishes the length of fragments to be considered and defines (through predetermined α-Carbon positions) an approximate geometry of acceptable fragments (the Conformational Template; Figure 1). Many structural units of protein structure may not be represented by a single residue range. The adjacent strands of a twisted β-sheet, for example, have a well defined structure independent of the length of the intervening loops or the relative position of the segments in the overall amino acid sequence. The target specification allows multiple ranges to be selected to accommodate these situations. An important application of fragment fitting is the modelling of incomplete structures, where the target is not entirely defined beforehand. As outlined below, the approach followed in modeling such regions will differ depending on the target specification required in a particular application.

Target Sequence	[YNQLSGTF]
Mask Status	[---------S-----l]
	22 29

Sequence Selection of Specific Amino Acid by Class:

Key	Class	Allowed Amino Acid
' '	Any	ACDEFGHIKLMPQRSTVWY
'b'	Beta	CFOVWY
'h'	Helical	AEHKLMQR
't'	Turn	DGNPST
'l'	Large	FHMWY
's'	Small	AGS
'y'	CB Branched	ITV
'n'	Nonpolar	ACFILMVWY
'p'	Polar	DEHKN QRST
'+'	Positive	HKR
'-'	Negative	DE

Figure 2. Dynamic Mask Sequence Designation. The library of x-ray structures may be searched using a flexible amino acid sequence and property mask. The mask illustrated confines the search to fragments whose backbone α–Carbons fit the target to some prespecified tolerance, and which also incorporate a serine residue in the position corresponding to residue 26 of the target and a "large" residue at position 29.

Masking.

In order to improve the adaptability of the target specification toward accommodating the specific needs of the user, two logical constructions are defined that collectively constitute a 'Dynamic Mask' (Figure 1). The first is a vector (a(n)) of logical quantities (where n is the number of residues in the target) that flag whether or not a particular residue is 'active' or 'inactive' in the context of specific operations defined below. The second is a matrix (20 by n) of logical quantities which flag the activity or inactivity of each of the twenty amino acids at each target residue position. The latter mask is used primarily as a means of specifying amino acid sequence or residue type (eg. polar, charged, etc.) requirements for a given fragment match to a target. The status of the Dynamic Mask is constantly indicated in a menu-like display and can be adjusted at any time (Figure 2).

Searching The Database of Known Structures.

All potential protein fragments from the library of known structures are screened for compatibility with the input target. Incompatible fragments are eliminated as efficiently as possible. A potential fragment is first tested against the sequence requirements given by the user in the specification of the Dynamic Mask, where any discrepancy between target and fragment sequences causes immediate rejection.

To screen out fragments of inappropriate geometry, individual elements of a triangular matrix of inter-Cα distances characteristic of the target (dt(i,j)) are compared with corresponding elements precomputed for the potential fragment (df(i,j)). Any difference |dt(i,j)-df(i,j)| greater than a user selected tolerance (e.g. 1.0 Å) causes rejection of the fragment. The search through all library structures can be made efficient by comparing interatomic distances d(1,n) first, then backwards to d(1,2), since distances between α-Carbon atoms show more variability as the number of residues between them increases. Distance correspondence between Cα's is only required at active residue positions, as defined by the status of the Dynamic Mask. Any distance d(i,j), where i or j is an index to an inactive sequence position is ignored. This makes it possible to search for a fragment loop with particular endpoint geometry, while making no requirements on the intervening structure except the number of residues.

For targets involving more than one sequentially connected residue range, the above procedure is first used to find fragments compatible with the first segment. The characteristic relationships between these residues and residues of other target segments are then investigated to identify additional fragment segments over the entire length of the polypeptide.

The result of a search is a list of sequences which comply within the requested tolerance. Since the gathering and superposition of complete atomic coordinate data is relatively slow, it is useful to be able to rank these sequences prior to subsequent processing. For each fragment recovered, a mean square deviation (Δ[d(i,j)]) from the target is obtained during comparison of interatomic distance matrices, and a numerical sequence homology score (Dayhoff et al., 1978) derived from comparison of the target and fragment amino acid content. An editing facility has been implemented to enable sorting of the sequence list using either of these criteria or selection of specific entries for preservation or deletion.

Superposition.

The superposition of the target and fragments is accomplished by the method of (Kabsch ,1978) based on a list of atom correspondences. Since the RMS difference in position of individual atoms after transformation is often the most quantitative gauge of the quality of the fit of a fragment to the target, it is important to be able to adjust which atoms will be included in this mean. Whole residues can be excluded from a correspondence list by inactivating the residue position in the Dynamic Mask. Further specificity can be achieved to defining a set of atom names (e.g. N CA C O) that will be forced to correspond. Unnamed atoms are ignored.

Fragment Disposition.

Once fragments have been identified and oriented coincident with the target, fragments may be displayed for evaluation on the graphics screen, or subjected to screening under van der Waal's restraints. The user may eventually select a fragment to replace complete atomic coordinate information of the target. As in the case of superposition, precise control over the disposition of each atom is possible during the replacement process.

Side-Chain Modeling.

Surveys examining the occurence of amino acid side-chain conformations in known protein crystal structures have demonstrated a preference for a limited number of low energy conformations (Janin et al., 1978; Bhat et al., 1979; James et al., 1983). A more recent survey of highly refined crystal structures (Ponder et al., 1987) has confirmed the earlier findings, and concluded that deviations from ideal stereochemistry are more infrequent than previously thought. We have implemented a procedure which allow all common conformations of a side-chain to be placed on a backbone. (The backbone atomic positions must be already defined). Side-chains are extracted from a library of favorable rotamers (Ponder et al.; 1987) and placed upon the target backbone. Rotamers may be manipulated exactly as fragments extracted from known structures, displayed for evaluation on the graphics system and selected to replace the target model.

Implementation.

The fragment manipulation capabilities of FRAGLE have been implemented as a module within FRODO (Jones, 1985). In this way, all the functionalities (such as electron density display, model manipulation and geometry regularization) indispensable to crystallographic applications are retained. Our implementation utilizes a version of FRODO specific to Evans and Sutherland PS300 series graphics devices (Pflugrath et al.,1984), although it may be easily modified for any hardware to which FRODO has been adapted.

APPLICATIONS

Electron Density Map Fitting.

FRAGLE can be used very effectively, in conjunction with model building capabilities incorporated in FRODO, as an aid in electron density map fitting. A recent application of FRAGLE in molecular model construction was the structure determination of protocatechuate 3,4 dioxygenase (EC 1.13.1.3) (PCDase), an enzyme from *Pseudomonas aeruginosa* that catalyses the the oxygenolytic cleavage of protocatechuic acid (3,4 dihydroxybenzoic acid) into β-carboxy-cis,cis-muconic acid. The holoenzyme is a 587000 dalton dodecamer of protamers arranged with 23 local symmetry. Each protamer is composed of 2 polypeptide chains containing a total of 440 amino acids.

The structure of PCD was solved using a combination of multiple isomorphous replacement and non-crystallographic symmetry averaging methods to produce a final, 6-fold symmetry averaged, map at 2.8 Å resolution (Ohlendorf et al., 1988). This map was completely interpreted without the use of a minimap using FRODO/FRAGLE on an Evans and Sutherland PS330 graphics system. The procedure started by positioning a string of points representing α-Carbons positions separated by 3.8 Å throughout the electron density map. The string was oriented so that the α-Carbons 1 and 2 were in their proper locations in the density. The bond between α-Carbons 1 and 2 was then broken, and the string rotated about α-Carbon 2 to position α-Carbon 3. This process was repeated to determine successive α-Carbon atom positions until the density ended or became ambiguous. In this manner over 90 percent of the α-Carbons were placed in the map.

In the next stage, the α-Carbons were replaced with complete polypeptide backbone by searching the library of refined protein structures (Table 1) for α-Carbon segments which gave the best RMS fits to the α-Carbon string originally built into the electron density map. Usually this search involved 5 to 9 residue segments. The best fitting structures were visually inspected and the one which optimally fit β-Carbons and carbonyl oxygens into the map density was incorporated into the model. In adding successive segments to the growing model, the terminal residues were generally not incorporated, since the ends had a tendency to fray due to lack of constraints on the course of chain. Finally, side chains were positioned in the density and added to the structure. FRAGLE displayed the most common rotamers for each residue as defined from the rotamer library (Ponder et al., 1987). Generally, a single rotamer conformation could be uniquely selected that best fit the electron density and was subsequently incorporated into the model.

PCDase contains large amounts of regular secondary structure, primarily in the form of β sheet. In these regions relatively long fragments (up to 13 residues) could be built as a single segment. Turns generally required shorter segments and often fit the density less well. Since turns generally have higher temperature factors, their densities are generally weaker. Also, some turns were not well represented in the data base. Nevertheless, it was possible using FRAGLE to routinely build 40 residues per day. The process is outlined in Figure 3.

In addition to speed in model construction, the use of the structural data base incorporated in FRAGLE provides important conformational information that may not be readily evident from initial inspection of the electron density map. For example, if a β strand is being built, and the 12 best examples all have a carbonyl group in a particular position, one is willing to accept it even though the map might be ambiguous about its placement. Similar situations also occur frequently with side chain placements.

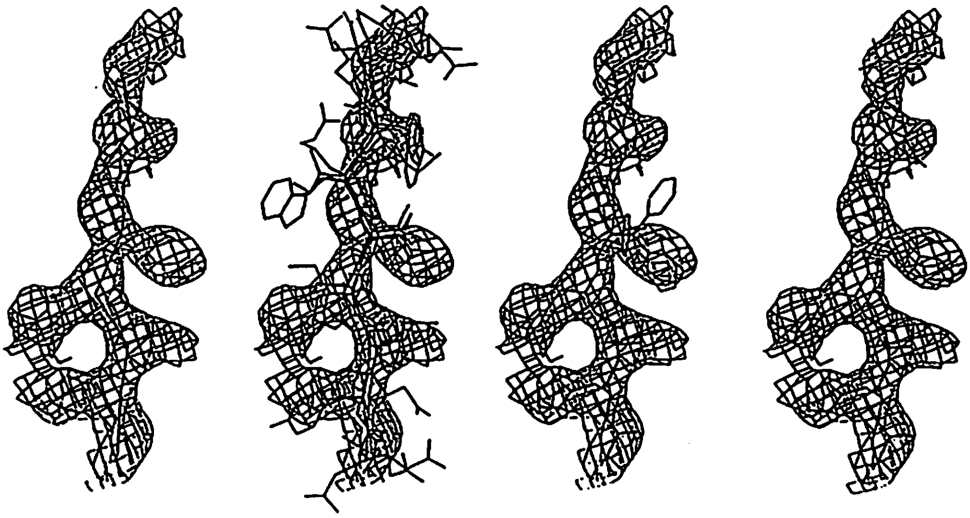

Figure 3. The panels illustrate the isolated electron density for a β strand in the 2.8 Å resolution x-ray map of PCDase. Successive panels show map with 3.8 Å αC string, superimposed polypeptide backbone structures retrieved from the database, the selected polypeptide backbone with multiple side chain rotamers displayed, and the final model structure.

As a result of incorporating fundamentally correct stereochemistry during model building, the final structures produced with FRAGLE have excellent geometry, a result which saves considerable subsequent effort in crystallographic refinement. In the case of PCDase, the model built with FRAGLE produced an initial R value of 0.38 for for 20,616 atoms and 59,057 data to 3.0 Å resolution (Ohlendorf et al., 1988). Additional examples of the use of FRAGLE to build crystal structures can be found in Weber et al., (1989).

Molecular Analysis.

Comparison of a collection of protein fragments with similar polypeptide conformation frequently reveals other structural characteristics common to the ensemble which had gone unrecognized upon examination of structures individually. Often these characteristics are simple or predictable (such as the requirement for a specific amino acid at a position in a tight turn, or the preference for a particular conformation of threonine side-chains in an extended α-helix), but more complicated generalities may also be revealed. In many cases, these localized motifs provide important information potentially relevent to engineering protein structures (Richardson et al., 1988; Blundell et al., 1987). FRAGLE has been extensively used to search for or verify such patterns, two examples of which are given below and illustrated in Figure 4.

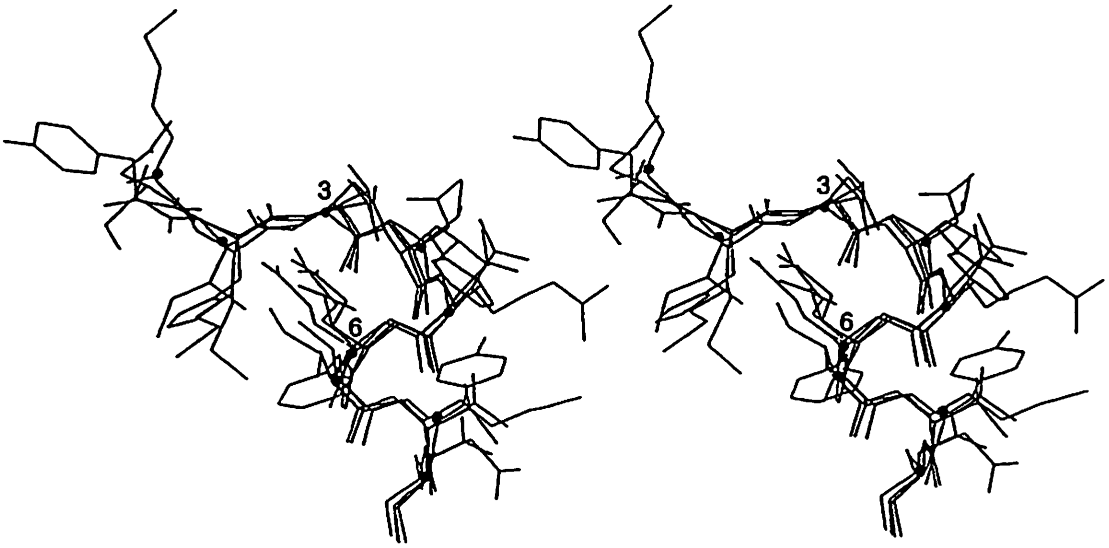

Figure 4. Stereoscopic representation of four similar protein fragments extracted from the library of known structures. The four (and the amino acid sequence) are from 1) Cytochrome P450 residues 190-198 (SMTFAEAKE); 2) Cytochrome c Peroxidase residues 162-170 (NMNDREVVA); 3) Carboxypeptidase residues 12-20 (YHTLDEIYD); and 4) Parvalbumin residues 96-104 (KIGVDEFTA). All are superimposed on the C-α backbone of Calmodulin residues 99-107 (FISAAELRH) (shown as dots) used as a target conformational template.

In Figure 4, four structurally homologous protein fragments are displayed. The four were selected as the only examples from a library of 40 known protein structures which meet a stringent (0.7 Å) tolerance in comparison to a nine residue section of the calmodulin α-Carbon backbone (Babu et. al., 1985). The polypeptide conformation adopted by all these fragments consists of an extended β conformation that loops immediately into an α helix. Results of the search demonstrate a preference for short oxygen-containing side-chains (serine, threonine or asparagine) at the third residue position, where a hydrogen bond can be accepted from the amide at position 6. Examination of the fragment sequences also reveals a strong preference for glutamic acid at position 6 in helical N termini of this geometry, possibly because of potential interactions between the side-chain carboxylate and the amide of residue 3. The existence of these specific interactions in a variety of otherwise unrelated protein structures illustrates the importance of these "capping" interactions in the stabilization of this conformation, and a possible mechanism by which both serine and glutamic acid contribute to the initiation of helix formation in globular proteins (Robson et al.,1972, Richardson et al., 1988).

Another example typifies the sort of fragment analysis which has potential utility in prediction of amino acid side-chain conformation. Surveys of side-chain conformations, such as that of (Ponder et al., 1987), have shown higher frequency of occurrence for some conformations than others. However, the conformation adopted at a particular position in a protein structure is highly dependent upon the local conformation of the polypeptide chain and other contextual factors. Serine, for example, can adopt three common side-chain conformations: + (chi=+60°), - (chi=-60°), and t (chi=180°), of which the + conformation occurs most frequently. For fragments in an extended β conformation, the - conformation occurs with higher frequency (41%, vs. 28% + and 32% t). If only fragments from extended parallel β sheets are selected (Figure 5), it can be shown that only the - conformation is observed, presumably because of packing restrictions imposed by the close proximity of neighboring strands. The same conclusion cannot be drawn for serine in anti-parallel β sheets, where other serine side-chain conformations are frequently observed. While this is a very specific example, analyses of side-chain conformation frequencies among fragments of homologous structure can often provide convincing evidence to support the selection of one conformation over others.

Structure Generation From Backbone Coordinates and Homology Building.

In many cases, it is necessary or desirable to generate an essentially complete molecular model when only α-Carbon coordinates may be available in a database. Alternatively, it may be that coordinates are available for one species of a protein, but not those of a related, homologous species. In both cases it is possible to use the fragment replacement capabilities of FRAGLE to extend the input structure.

In one such study (Weber et. al., 1990), the objective was to study the consequences of alterations in surface charge residues on the functional properties of plant calmodulin mutants that were highly homologous to the vertebrate protein. When this study was begun, only α–Carbon coordinates of the protein were available, so that it was necessary to reconstruct the entire structure from these partial structural data. This case was felt to be particularly favorable for two reasons: 1) Calmodulin is composed predominantly of α helices, fragments which are very well represented in the structural database. 2) The objective involved the computation of molecular electrostatic fields, which are relatively insensitive to small errors in side-chain placement (compared, say, to the precision required to accurately reconstruct an enzyme active site.)

In this example, α-Carbon coordinates from rat testis calmodulin (Babu et al, 1985) were extended using the fragment fitting strategies outlined above to produce a complete polypeptide backbone structure. Indeed, experience has now shown that location of α–Carbon positions is nearly always sufficient to correctly orient the polypeptide backbone planes within a few tens of degrees. Since this also correctly orients the Cα-Cβ bond, the only remaining variables to be determined involve selection of the proper side chain rotamers. As illustrated above and described also in (McGregor et al., 1987), side chain rotamers can in many cases be predicted owing to their incorporation in a particular secondary structure. Thus, side chains were added to calmodulin by selecting 3 to 5 residue fragments whose backbone matched the input α-Carbon target, using a mask that required amino acid identity in the side chain position to be substituted. Unfavorable steric interactions that occurred in the building process were relieved either by selection of an alternative allowed side chain rotamer, or by energy minimization. This process produces a very well packed protein structure with solvent exposed charge groups and virtually all essential features of the calcium binding site of the related protein parvalbumin (not included in the structure library used for this study) properly regenerated. These mutants proteins are now being examined crystallographically, and it will be of some interest to examine the accuracy of the prediction.

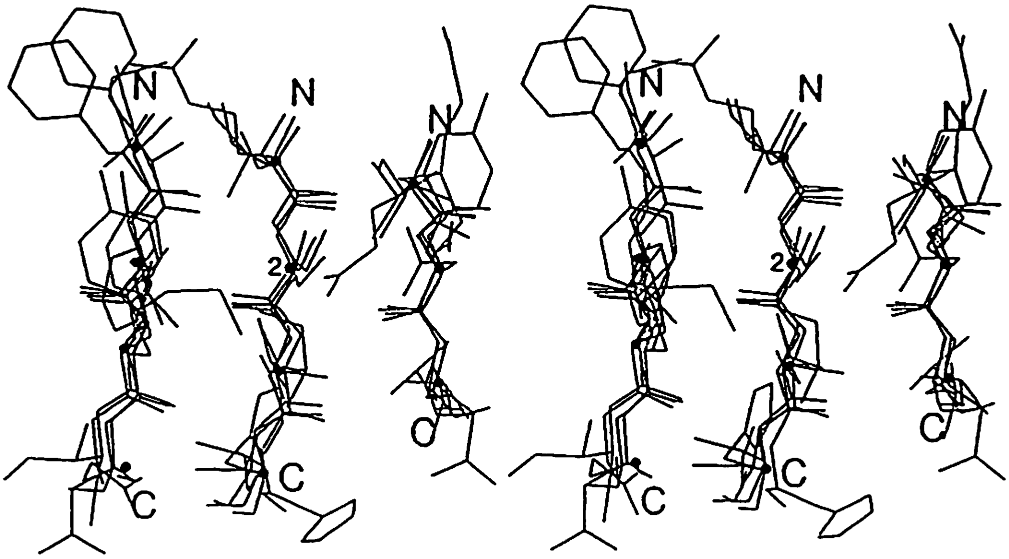

Figure 5. Stereoscopic illustration of the conformation of all occurences of three contiguous parallel β strands from the library of known structures which have a serine at amino acid position 2 of the central strand. Selected from four different protein structures, all have the same conformation for the serine sidechain. The target geometry was specified using α-Carbon positions from Subtilisin BPN' (residues 28-30, 121-124 and 148-151), which utilizes an isoleucine at this position (residue 122).

References.

Babu, Y.S., Sack, J.S., Greenhough, T.J., Bugg, C.E., Means, A.R. and Cook, W. J., Three-dimensional Structure of Calmodulin. Nature, 315, 37-40 (1985).

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. J. Mol. Biol., 112, 535-542 (1977).

Bhat, T. N., Sasisekharan, V. and Vijayan, M., An Analysis of Side-chain Conformations in Protein Structures. Int. J. Pept. Res.,13, 170-184 (1979).

Blundell, T. L., Sibanda, B. L., Sternberg, M.J.E. & Thornton, J. M., Knowledge-based Prediction of Protein Structures and the Design of Novel Molecules. Nature, 26, 347-352 (1987).

Chothia, C., Levitt, M., and Richardson, D. Structure of Proteins: Packing of α-Helices and Pleated Sheets, Proc. Natl. Acad. Sci. USA 74, 4130-4134 (1977).

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C., A Model of Evolutionary Change in Protein Structures. Atlas of Protein Sequence and Structure. (1978)Vol. 5 Suppl.3 (M. O. Dayhoff, ed.). pp. 345-352 (1987).

Hermans, J. & McQueen, J.E., Computer Manipulation of Macromolecules with-the Method of Local Change. Acta Crystallogr., A30, 730 (1974).

James, M.N.G. and Sielecki, A.R. Structure and Refinement of Penicillopepsin at 1.8 Å Resolution J. Mol. Biol., 183, 299-361 (1983) .

Janin, J., Wodak, S., Levitt, M. and Maigret, B. Conformation of Amino Acid Side-chains in Proteins. J. Mol. Biol., 125, 357-386 (1978).

Jones, T.A., Interactive Computer Graphics: FRODO Meth. Enzymol., 115,157-171 (1985).

Jones, T.A. & Liljas, L., Crystallographic Refinement of Macromolecules having Non-crystallographic Symmetry. Acta Crystallogr., A40, 50 (1984).

Jones, T.A. & Thirup, S., Using Known Substructures in Protein Model Building and Crystallography EMBO J., 5, 819-822 (1986).

Kabsch, W., A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. Acta Crystallogr., A34, 827-828 (1978).

McGregor, M., Islam, S. A., and Sternberg, M. J. E., Analysis of the Relationship Between Side-chain Conformation and Secondary Structure in Globular Proteins. J. Mol. Biol. 198, 295-310 (1987).

Pflugrath, J.W., Saper, M.A. and Quiocho, F.A. (1984) in "Methods and Applications in Crystallographic Computing" (S. Hall & T. Ashida, eds.) Oxford Univ. Press, London, pp 404.

Ponder, J.W. & Richards, F.M., Tertiary Templates for Proteins. J. Mol. Biol., 193, 775-791 (1987).

Richardson, J.S. The Anatomy and Taxonomy of Protein Structure. Adv. in Protein Chem., 134, 167-338 (1981).

Richardson, J.S., and Richardson, D.C., Amino Acid Preferences for Specific Locations at the Ends of α-Helices. Science, 240, 1648-1652 (1988).

Richmond, T.J and Richards, F.M. Packing of α-Helices: Geometrical Constraints and Contact Areas, J. Mol. Biol. 119, 537-555 (1978).

Robson, B. and Pain, R.H., Directional Information Transfer in Protein Helices. Nature, 238, 107-108 (1972).

Weber, P. C.,Lukas, T. J., Craig, T. A., Wilson, E,. King, M. M., Kwiatkowski, A. P. and Watterson, D. M., Computational and Site-Specific Mutagenesis

Analyses of the Asymmetric Charge Distribution on Calmodulin. (in press) PROTEINS:Structure, Function and Genetics (1990).

Weber, P. C. and Salemme, F. R., Structural and Functional Diversity in 4-α-Helical Proteins. Nature 5777, 82-84 (1980).

Weber, P. C., Ohlendorf, D.H., Wendoloski, J. J., and Salemme, F.R. Structural Origins of High-Affinity Biotin Binding to Streptavidin. Science 243, 85-88 (1989).