Reprinted from *J. Mol. Biol.* (1991) **218**, 183–194

Database Algorithm for Generating Protein Backbone and Side-chain Co-ordinates from a C^{α} Trace

Application to Model Building and Detection of Co-ordinate Errors

Liisa Holm and Chris Sander

J. Mol. Biol. (1991) 218, 183–194

Database Algorithm for Generating Protein Backbone and Side-chain Co-ordinates from a C^{α} Trace

Application to Model Building and Detection of Co-ordinate Errors

Liisa Holm and Chris Sander

EMBL, Meyerhofstrasse 1, D-6900 Heidelberg, F.R.G.

(Received 21 February 1990; accepted 30 October 1990)

The problem of constructing all-atom model co-ordinates of a protein from an outline of the polypeptide chain is encountered in protein structure determination by crystallography or nuclear magnetic resonance spectroscopy, in model building by homology and in protein design. Here, we present an automatic procedure for generating full protein co-ordinates (backbone and, optionally, side-chains) given the C^{α} trace and amino acid sequence. To construct backbones, a protein structure database is first scanned for fragments that locally fit the chain trace according to distance criteria. A best path algorithm then sifts through these segments and selects an optimal path with minimal mismatch at fragment joints. In blind tests, using fully known protein structures, backbones (C^{α}, C, N, O) can be reconstructed with a reliability of 0·4 to 0·6 Å root-mean-square position deviation and not more than 0 to 5% peptide flips. This accuracy is sufficient to identify possible errors in protein co-ordinate sets. To construct full co-ordinates, side-chains are added from a library of frequently occurring rotamers using a simple and fast Monte Carlo procedure with simulated annealing. In tests on X-ray structures determined at better than 2·5 Å resolution, the positions of side-chain atoms in the protein core (less than 20% relative accessibility) have an accuracy of 1·6 Å (r.m.s. deviation) and 70% of χ^1 angles are within 30° of the X-ray structure. The computer program MaxSprout is available on request.

1. Introduction

Crystallographers and those who analyze their protein co-ordinates have known for some time that local structure motifs occur time and again in newly solved protein structures. The work of Jones and colleagues (Jones & Thirup, 1986) has led to widespread use of loop building procedures where substructures are selected from the database of known structures. Typically, loops of a specified length most similar to user-specified intra-loop distances are extracted from the database of known structures *via* a rapid search of diagonal distance plots. Such loops are then used constructively to fill a gap in a protein model under construction. Here we present a fully automatic database procedure of going from a C^{α} trace to a full set of co-ordinates, which may be useful in a wide range of model-building problems.

Existing software tools for building loops from fragments, such as a recent version of FRODO (Jones & Thirup, 1986), usually require user interaction and judgement in loop patching (e.g. see Reid & Thornton, 1989). In addition, the problem of annealing the joints between fragments generally requires a subsequent independent refinement procedure, and difficulties in closing gaps do not automatically lead to a new choice of database fragment. For example, the automatic backbone construction procedure of Claessens *et al.* (1989) tries to identify single continuous fragments that fit best to a given piece of the C^{α} trace but makes no checks on individual joints. Our method for constructing backbones selects the best parts from many fragments and optimizes the overall choice by smooth joining of the parts. Thus, the reconstructed backbones should have backbone dihedral angles in the allowed range also across fragment joints.

For side-chain construction in proteins, one of the more successful approaches has been to copy their conformation from a homologous template (Blundell *et al.*, 1987; Summers & Karplus, 1989). Given merely a chain trace, the generation of correct side-chains is much more difficult than building backbone co-ordinates, as the geometrical constraints are less localized. One can exploit two observations: (1) the statistical distribution of side-chain rotamers in known structures is rather sharp (e.g. see Ponder & Richards, 1987; Tuffery *et al.* (1991); and (2) atoms in the protein interior are

0022-2836/91/050183-12 $03.00/0

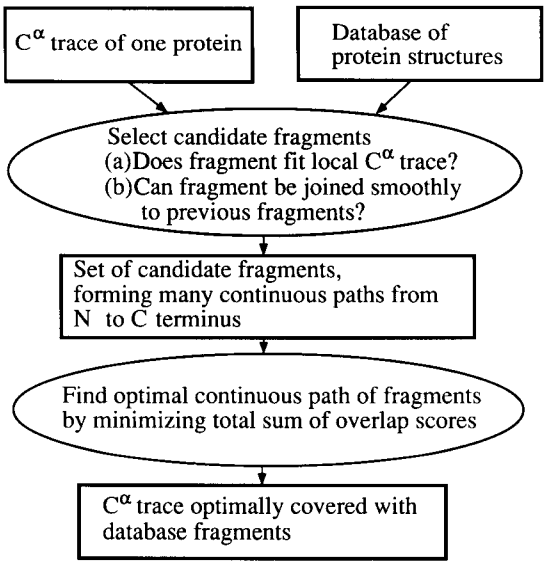

Figure 1. Backbone construction. Given C^{α} co-ordinates, the program screens a database of protein structures with full backbone co-ordinates and looks for fragments with matching C^{α}–C^{α} distances. The selected fragments are candidates for a complete covering of the C^{α} trace, and there are more candidate segments than minimally required. The overlap of 2 successive fragment candidates is evaluated by a function called the overlap score. The optimal succession of overlapping fragments (path) is found using a dynamic programming algorithm.

closely packed and do not overlap. The main difficulties are the strongly co-operative character of packing and the correctness of potential energy functions used to evaluate packing arrangements. We approach the side-chain optimization problem by a Monte Carlo procedure (Metropolis *et al.*, 1953) in the subspace of frequently observed rotamers with simple potential energy functions.

2. Methods

Backbone construction is in two steps: (1) screen the database for matching fragments and select candidates for a complete covering of the chain; (2) find the optimal path of fragments covering the C^{α} trace (Fig. 1). Fragments that locally match the given C^{α} trace are retrieved from the structure database using distance criteria and ordered locally at each sequence position in terms of decreasing C^{α} trace mismatch. There are more candidate segments than minimally required. Only a few fragments are tested in regions with standard structure and more in regions of irregular structure. The way in which fragments are joined into a continuous chain depends on a function that evaluates the overlap of 2 successive fragments, called the overlap score. The optimal path through the selection of fragments is found using a dynamic algorithm, which minimizes the sum of overlap scores.

Side-chain construction also has two steps: (1) given the backbone co-ordinates, generate sets of plausible side-chain co-ordinates using a rotamer library and calculate a table of all rotamer–rotamer interaction energies; (2) minimize the intramolecular energy by optimized choice of rotamers. As energy minimization is incapable of leaving local minima, and molecular dynamics or systematic searches are very time-consuming, a random search method is employed for optimization. The key idea of the Monte Carlo Metropolis algorithm is iterative improvement in which controlled uphill steps can be incorporated occasionally in the exploration of multiple energy minima of complex systems. The higher the temperature of the system, the more uphill steps are accepted. In simulated annealing, the temperature is gradually lowered until the system settles into a local minimum (Kirkpatrick *et al.*, 1983). The conformational subspace of discrete side-chain rotamer states is chosen for calculational efficiency and conceptual simplicity. With precalculated rotamer–rotamer interaction energies, the calculation of the total energy of the system at each Monte Carlo step is reduced to a table look-up and a few additions and subtractions. This results in a side-chain construction procedure that is extremely efficient, at the cost of limited accuracy in side-chain conformations.

(a) *Data sets*

Fragments for constructing the backbone were retrieved from a protein structure database containing 34 high-resolution proteins with a total of 4759 residues (see Table 2). The method was calibrated and evaluated on a separate set of 20 proteins (see Table 1), which were selected from the Protein Data Bank (PDB, Bernstein *et al.*, 1977) so that they represent different resolutions and have less than 25% sequence identity with any protein in the fragment database.

(b) *Backbone construction*

(i) *Selection of matching fragments*

Database fragments similar to fragments of the given C^{α} trace are located by comparing precalculated C^{α}–C^{α} distance tables. The minimum fragment length and the maximum allowed deviation of C^{α}–C^{α} distances are adjustable parameters. A fragment match in the database is extended in both chain directions as long as the C^{α}–C^{α} distance deviations are less than the specified maximum. Chain chirality (dihedral angle of 4 C^{α} atoms) is stored with the C^{α}–C^{α} distances and is checked to reject mirror structures.

Two modes of searching are used. Certain substructures, such as the α helix, type I turn and extended strand, occur in many identical copies in the database. First, the algorithm tries to locate such pieces in the C^{α} trace using a quick search with each residue in turn as start residue. This search uses a distance deviation cutoff of 0·2 Å (1 Å = 0·1 nm) and stops after one fragment has been found. The 2nd search mode locates useful fragments using a distance deviation cutoff of 0·8 Å. Fragments that pass the distance deviation filter are sorted by r.m.s.† distance deviation. The optimal C^{α} superposition onto the C^{α} trace is calculated for the best 50 fragments using the algorithm by Kabsch (1978) and the fragments are sorted by r.m.s. C^{α} deviation. Fragments found in the quick search and selected fragments from the sorted lists obtained using the 2nd mode of searching are used to select candidate paths that cover the chain.

(ii) *Evaluation of fragment joints*

Two fragments can be joined if they have an overlapping peptide unit. The location of joints along the frag-

† Abbreviations used: r.m.s., root-mean-square; c.p.u., central processor unit.

ments is not determined beforehand: all possible joining points between two selected fragments are considered by the algorithm. The overlap score is defined as the r.m.s. distance between equivalent C^{α}, C, N, O, C^{α} atoms. The user may control the stringency of joints through a parameter that imposes an upper limit on the mismatch encountered at any particular joint. In the stringent mode, the upper limit is set to a small value (here 0·4 Å), meaning that no joint may have a mismatch of more than this value. In the permissive mode, the upper limit is set to 1000 Å, i.e. it is not applied. In either case, the sum of overlap scores is minimized in the subsequent determination of the optimal path.

(iii) *Selection of candidate paths*

For technical reasons the algorithm makes a compromise between (1) reducing the number of candidate fragments per residue in order to reduce combinatorial complexity and (2) increasing the number of candidate fragments/residue so that optimal joints can be formed; minimally, joints must satisfy the constraint imposed by the local upper limit on overlap mismatch. The following steps are repeated for each residue going from N to C terminus: (1) check if the residue is covered by at least one fragment that has acceptable joints all the way back to the N terminus of the chain; (2) if not, take additional fragments from the sorted reservoir list of matching database fragments and repeat (1); (3) if the reservoir of fragments is exhausted, a chain break is reported and a new chain started from the residue following the gap.

(iv) *Global optimization*

Once the chain has been covered with overlapping fragments, only fragments that are part of continuous paths from the C terminus back to the N terminus are retained. The best succession of overlapping fragments (path) is found by an inductive (dynamic) algorithm, analogous to algorithms used in sequence alignment (Needleman & Wunsch, 1970; Smith & Waterman, 1981). The global optimum is the path that minimizes the sum of overlap scores. The repetitive step in the algorithm is as follows: assume the best path has been found starting at the beginning of the chain and ending at a particular residue k, for all $k < n$. To determine the best path ending at n, a local choice is made between all combinations of a path ending at k plus a new fragment overlapping at residue k and ending at residue n. The best path to residue n is the combination that has the best combined score (sum of all overlap scores in the path ending at n). This procedure is repeated for $n+1$, etc., up to the C terminus. The algorithm is guaranteed to find the global optimum for an additive global error function, such as the sum of local overlap quality scores.

The result is a full set of backbone co-ordinates. Backbone co-ordinates (C^{α}, C, N, O) are taken from the fragments that were superposed on the C^{α} trace. The backbone of triose phosphate isomerase (1TIM) with 247 residues, for example, was built in 2 min 23 s c.p.u. time on a VAX8650 computer. Main-chain hydrogen bond and secondary structure assignments based on backbone co-ordinates were generated using the program DSSP (Kabsch & Sander, 1983).

(c) *Construction of side-chains*

(i) *Rotamer library*

Several rotamer libraries were used and results with that of Tuffery *et al.* (1991) with a few additional rotamers for aromatic residues are reported. The total number of rotamers was 104 for 19 non-glycine amino acids. Proline and alanine have only 1 rotamer. The positions of L-amino acid C^{β} atoms are constructed geometrically from the positions of N, C^{α} and C atoms using standard bond lengths and angles. Side-chain rotamers for residues other than Gly and Ala are superposed (Kabsch, 1978) onto N-C^{α}-C^{β}-C with weights 1–10–10–1.

(ii) *Interatomic potential energy*

Atom–atom distances in the fragment database (Table 2) were calculated and studied as histograms, which showed that practically no contacts shorter than 3·5 Å are observed except between O and N atoms, which form hydrogen bonds at a distance of 2·9 to 3·0 Å and cysteine sulfur atoms, which form disulphide groups at 2·0 Å. Based on this, atom–atom 6–9-potentials with minimum at 2·0 Å (cysteine S–S pairs), 3·0 Å (N–O pairs) and 4·0 Å (all other atom pairs) were used, truncated (the value at the point of truncation used down to zero distance) at 1·25 Å (S–S), 1·75 Å (N–O) or 2·25 Å (all other pairs). The total energy is calculated by summing pairwise atomic energies for distances up to 6·0 Å. Rotamer–rotamer energies with a list of which residues are in contact are precalculated and stored in tables.

(iii) *Monte Carlo optimization of rotamer choice*

The present implementation starts by constructing a random configuration of rotamers, followed by a process of rearrangement of rotamers. The algorithm works by trying a randomly selected new rotamer at a randomly selected residue. For each rearrangement, the program computes the total energy and decides whether to accept the rearrangement. If the energy of the new configuration is E' and that of the previous configuration was E, then the new configuration is accepted with a probability:

$$p = 1/(1+\exp(-\alpha(E-E')).$$

The decision is made by comparing p to a random real number in the interval (0, 1). The inverse temperature, α, is initially set to zero. After each rearrangement, the inverse temperature is increased by an amount δ. When α is small (as it is initially), the probability of accepting a rearrangement is nearly 0·5 and independent of the energy difference. This means that in the initial period the rotamer configuration is rearranged essentially at random. As α is increased in each iteration step, the bias towards lower energies continually increases. The search terminates when α reaches a preset maximum value, e.g. 999, or there is no drop in energy for 500 steps.

The best choice of cooling parameters depends on the energy differences and complexity of the system. For each protein, several (3500) simulations are done using predefined strategies. The program returns the lowest-energy configuration found. The random search for low-energy minima is enhanced by applying weights to residue and rotamer selection probabilities. Residues with high energies are changed more often than residues with low inter-residue energy. These weights are updated every 100 steps during each simulation. The choice of rotamers is also biased by a simple weighting scheme. Initially, all rotamers at a given residue position are tried with equal probability, but after each 500 completed optimization runs the program calculates new weights that equal the frequency of each particular rotamer in the last 500 optimized configurations.

Even for a protein as large as triose phosphate

Table 1

Proteins used to calibrate the method. These proteins constitute the data points in Figures 2, 3 and 5

Code	Resol. (Å)	Length	Code	Resol. (Å)	Length	Code	Resol. (Å)	Length
4RXN	1·2	54	2LYZ	2·0	129	1CTS	2·7	437
2PRK	1·5	279	3ADK	2·1	195	3HVP	2·8	99
1GCR	1·6	174	1CRO	2·2	66	2TMV	2·9	157
2WRP	1·7	104	1PFK	2·4	320	1PYP	3·0	281
1UBQ	1·8	76	156B	2·5	103	1CN1	3·2	237
3WGA	1·8	171	1CY3	2·5	118			

Code, Protein Data Bank dataset identifier; resol., nominal crystallographic resolution (Å); length, number of amino acid residues. Complete references to the original crystallographic work can be found in the headers of Protein Data Bank datasets; they are omitted here for space reasons.

isomerase (1TIM, 247 residues), side-chains were built in less than 16 min c.p.u. time on a VAX8650 computer.

3. Results and Discussion

(a) *Quality of model backbone co-ordinates*

The quality of the method was assessed in a blind test by reconstructing the backbones of a set of fully known protein structures (Table 1) from their C^{α} co-ordinates. Fragments were retrieved from a separate database of non-homologous proteins (Table 2). Two series of models were built, one enforcing good joints (stringent mode) and the other allowing joints even with bad mismatches (permissive mode). Model quality was assessed by the number of chain gaps (in stringent mode), the positional r.m.s. deviation of backbone atoms, the number of peptide flips and the number of backbone–backbone hydrobgen bonds.

(i) *Chain gaps*

Comparison between model co-ordinates and the original X-ray co-ordinates shows that the backbone of proteins can be well represented by database fragments. Even in stringent mode, relatively few fragment testing cycles were required to build the backbone: 13% of all residues could be reconstructed using only fragments retrieved in the quick search, and 61% of the residues were reconstructed using the top five fragments from the candidate list. Gaps, or failures of the algorithm to build a continuous chain, may occur in stringent mode but are very rare in higher-resolution structures. Two or more gaps occurred only in structures determined at lower than 2·5 Å resolution.

(ii) *Positional r.m.s. deviation of main-chain atoms*

The r.m.s. deviation of atom positions between model and X-ray structure ranges from 0·1 to 0·2 Å for C^{α} atoms and from 0·4 to 0·6 Å for all backbone atoms in structures determined at better than 2·5 Å resolution (Fig. 2(a)). The accuracy of the position of C^{β} atoms is related to backbone accuracy, and the r.m.s. error in the whole test set is an average 0·5 Å, again lower for high-resolution proteins and higher for X-ray structures determined at lower resolution.

(iii) *Peptide flips*

Enforcing good fragment joints (stringent mode) leads to a better fit to the C^{α} trace and main chain (Fig. 2(a)), and to better geometry in terms of flips (Fig. 2(b)) and hydrogen-bonding patterns (Fig. 2(c)). The number of wrongly oriented peptide units (flips, defined as the X-ray peptide carbonyl carbon atom and X-ray and model oxygen atoms forming an angle wider than 90°) is consistently very low in stringent mode. It is lower than five per 100 residues for structures resolved at better than 3·0 Å if one outlier (1CY3, preliminary crystallographic model) is excluded (Fig. 2(b)). Most flips

Table 2

Proteins used to generate a database of fragments

Code	Resol. (Å)	Length	Code	Resol. (Å)	Length	Code	Resol. (Å)	Length
5PTI	1·0	57	1PCY	1·6	99	3FXN	1·9	138
1PPT	1·37	36	3TLN	1·6	280	1FB4	1·9	216
1NXB	1·38	42	1MBO	1·6	153	1LH1	2·0	153
1ECD	1·4	136	351C	1·6	82	1HIP	2·0	85
1INS	1·5	21	1BP2	1·7	123	2B5C	2·0	85
2OVO	1·5	56	2ACT	1·7	218	1CAC	2·0	256
1CRN	1·5	46	3DFR	1·7	162	1REI	2·0	107
4CYT	1·5	103	2APP	1·8	323	2C2C	2·0	112
2SGA	1·5	180	1SN3	1·8	62	1FDX	2·0	54
2SNS	1·5	141	2PAB	1·8	114	1RHD	2·5	293
5CPA	1·54	307	2AZA	1·8	129			
2PTN	1·55	123	1CPV	1·85	108			

The notation is the same as in Table 1.

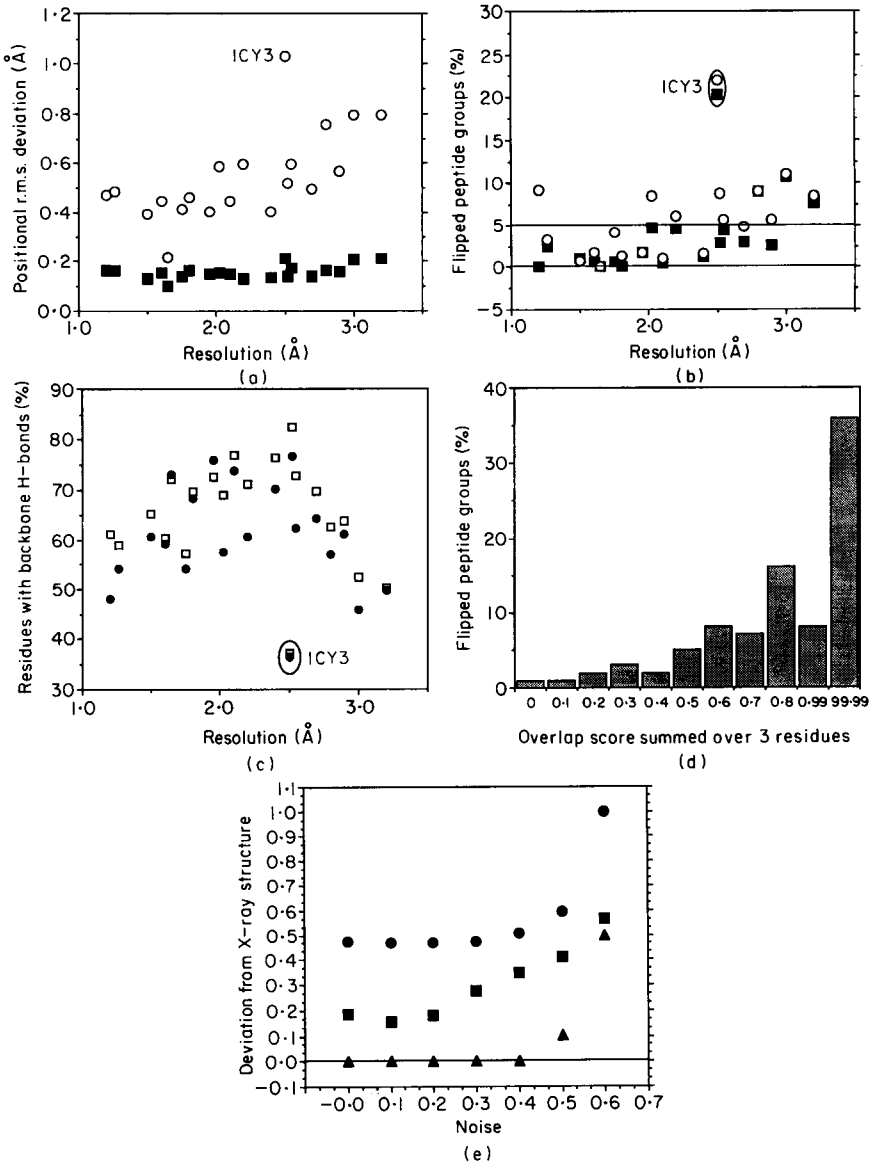

Figure 2. Quality check of backbone construction. (a) In terms of atom positions. Positional r.m.s. deviation between equivalent atoms in the X-ray and model structures for proteins in the test dataset. The success of the construction procedure does not depend much on the resolution of the crystallographic structure determination. Main-chain atoms include N, C^{α}, C, O. The average improvement from using stringent joint tolerance (enforcing good fragment joints) relative to permissive mode of backbone construction (covering the chain with fragments but accepting all joints regardless of the quality of the overlap) is 0·06 Å in C^{α} r.m.s. and 0·08 Å in main-chain r.m.s. deviation; (■) C^{α}; (○) main chain. (b) In terms of peptide group orientation. Comparing model to X-ray structure a peptide group is considered as flipped if the relative orientation of the C–O vector in equivalent peptide units differs by more than 90° (the first and last residue in the chain are not considered). Most of the flipped peptide groups are in loop regions, a few in β-sheets. The average benefit from using stringent joint tolerance (■) is 2·0 percentage points relative to permissive mode (○). (c) In terms of backbone hydrogen bonding. In many cases, the constructed backbone hydrogen bond structure is sufficient for extraction of secondary structure features. Models built in permissive mode (not shown) have, on average, 6·7 percentage points and stringent models (●) 4·6 percentage points fewer backbone hydrogen bonds than the X-ray structure (□). The preliminary crystallographic data set 1CY3 of cytochrome *c*3 is an outlier in graphs (a) to (c) (see also Fig. 4). (d) Flip probability increases with overlap scores. The variation of overlap scores along the chain indicates where the probability of finding flips is highest in reconstructed backbones. Higher overlap scores increase the probability of peptide flips. Residues next to gaps have undefined joint scores and are shown in the right-most column. (e) Noise test. Random shifts were introduced to the C^{α} co-ordinates of ubiquitin (1UBQ) and the backbone was reconstructed. The quality of reconstructed backbone co-ordinates starts to deteriorate drastically around a noise level of 0·5 Å. (■) C^{α} r.m.s. (Å); (●) main-chain r.m.s. (Å); (▲) number of gaps/10.

occur in loops, very few in β-sheet or near helix ends. No flip is observed in the middle of an α-helix. The overlap score along the sequence of a reconstructed model appears to have some predictive value with respect to flips. The probability of finding a flip increases with increasing overlap score (Fig. 2(d)). Shorter fragments are used at higher stringency for fragment joints (average progression/fragment used in the optimal path is 1·7 residues in stringent mode and 2·0 residues in permissive mode), but the quality of models increases as the algorithm is able to compose peptide conformations that may be absent from the database by combining fragments of existing ones.

(iv) *Response to inaccurate C^α trace*

It is of practical interest to investigate to what extent inaccuracies in the input C^α trace influence the quality of reconstructed backbones. Random shifts were introduced to the C^α co-ordinates of fully known protein structures, after which the backbones were reconstructed in stringent mode. Figure 2(e) shows the response to positional noise for ubiquitin (1UBQ, Vijay-Kumar *et al.*, 1987). Similar curves were obtained for other proteins (data not shown). The procedure was able to generate backbones without any gaps for random C^α shifts of up to 0·4 Å. Above this noise level, the number of gaps starts to increase rapidly as a consequence of two effects. First, since the fragment search is based on similarity of C^α–C^α distances, fewer fragments are found. Second, since the fragments are fitted to the given C^α trace, joining gets more difficult. Iterating the procedure on the noisiest C^α traces removed some of the gaps, but some remained. One might suppose that errors in a very approximate C^α trace can be smoothed out by increasing the minimal fragment length and the maximal allowed distance deviation in the database search. However, changing the parameters in this direction results in more gaps (in stringent mode), more flips and higher C^α and main-chain r.m.s. for high-resolution structures. We conclude that backbone reconstruction by this method in stringent mode is robust up to positional noise of 0·4 to 0·5 Å in C^α positions.

(v) *Extension of the database of secondary structure assignments*

The total number of backbone–backbone hydrogen bonds, an indicator of well-formed secondary structure in globular proteins, is not much lower in the reconstructed models than for the original structures (Fig. 2(c)), so the model backbone co-ordinates generated from incomplete information are of sufficient quality to be used for automatic extraction of secondary structure features based on backbone hydrogen-bonding patterns by the program DSSP (Kabsch & Sander, 1983). Moreover, in low-resolution proteins, more H (helix) and E (strand) states are assigned in the model co-ordinates than in the X-ray co-ordinates, indicating that the model co-ordinates partly anticipate the result of further refinement. For example, the reconstructed model of concanavalin A starting from a low-resolution structure (1CN1) has an increased number of hydrogen bonds in β-strands (Table 3), some of which also appear in the refined structures. As a result, we can now use automatically determined secondary structure assignments of reasonable quality for more than 30 entries in the Protein Data Base for which only the C^α co-ordinates have been deposited (Table 4). These will be available as *nxxx.dssp_ca* files (*nxxx* is the dataset identifier) on the EMBL network file server. Send e-mail message "help" to NETSERV@EMBL.bitnet for more information.

Table 3

Reconstructed model of a low-resolution structure of concanavalin A anticipates the result of further refinement

Model	Resol. (Å)	*n*(E)	*n*(H)	*n*(hb)
1CN1 (X-ray)	3·2	65	0	119
Model rebuilt from 1CN1 C^α trace		78	4	126
3CNA (X-ray)	2·4	96	0	132
2CNA (X-ray)	2·0	103	4	146

The model rebuilt from the C^α trace of the low-resolution structure of concanavalin A, data set 1CN1, has a higher number of H bonds in well-formed secondary structure (β-strands) than the original X-ray structure. Dataset 1CN1 is the structure of demetallized concanavalin A. The higher-resolution datasets 2CNA and 3CNA include metal ions. The protein has 237 amino acid residues.

Resol., nominal resolution; *n*(E), number of residues in β-strand structure; *n*(H), number of residues in α-helical structure; *n*(hb), number of backbone hydrogen bonds.

(b) *Detecting and correcting errors in protein models*

With the continually increasing interest in model building of proteins, either from scratch (e.g. see Sander, 1987; DeGrado *et al.*, 1989; Richardson & Richardson, 1989) or by homology, and with occasional errors in crystallographic structures (Branden & Jones, 1990) it is important to be able to assess the accuracy of models. Statistical comparisons can be useful in identifying improbable protein structures. Figure 2 suggests that backbones rebuilt with the database algorithm that have more than 10% flips, more than two chain breaks per protein (in stringent mode), higher than 0·3 Å C^α r.m.s. deviation, higher than 0·8 Å main-chain r.m.s. deviation relative to the original model and/or fewer than 40% backbone residues forming hydrogen bonds are models of much lower quality than those of high-resolution proteins. The criteria listed above reflect the geometry of covalent bonds, restrictions in backbone dihedral angles due to clashes between amino acid side-chain and main-chain atoms, and the fact that all possible hydrogen bonds appear to be formed in correctly folded proteins.

For example, comparing reconstructed models and original crystallographic structures, datasets 2FD1 (ferrodoxin), 1CY3 (cytochrome *c*3), and

Table 4

Extension of the Dictionary of Secondary Structure of Proteins (Kabsch & Sander, 1983) by completion of C^{α}-only datasets

Code	Resol. (Å)	Length	%H	%E	%hb	Protein – Year of deposition – Refined *y*/*n* (*R*-value)
2RUB	1·7	827	39	12	63	Rubisco – 1989 – *y* (0·21)
1SIC	2	275	20	7	56	Subtilisin BPN' – 1984 – *y* (0·34)
1CRO	2·2	66	41	12	61	cro repressor – 1987 *y*
2P21	2·2	166	36	24	69	C-H-ras P21 catalytic domain – 1989 *y* (0·188)
2ENL	2·25	436	38	13	69	Enolase – 1989 – *y* (0·155)
1XIA	2·3	393	42	8	67	D-Xylose isomerase – 1988 – *n*
1MCG	2·3	216	5	31	45	Lambda-type Bence–Jones MCG – 1978 – *y* (0·37)
1BLM	2·5	257	31	12	62	β-Lactamase – 1987 – *y* (0·284)
1CBP	2·5	86	7	28	57	Cucumber basic protein – 1988 – *y*
1PYK†	2·6	432	22	2	41	Pyruvate kinase – 1980 – *n*
1EFM	2·7	158	23	6	50	Elongation factor Tu – 1987 – *n*
1MON	2·75	94	15	28	50	Monellin – 1989 – *y* (0·22)
1SRX	2·8	108	20	6	44	Thioredoxin – 1976 – *n*
1AAT	2·8	107	15	0	33	Cytosolic aspartate aminotransferase – 1982 – *n*
1DPI	2·8	546	42	8	63	DNA polymerase Klenow fragment – 1987 – *y* (0·20)
1LZ2	2·8	129	23	12	53	Lysozyme (turkey – 1981 – *n*
1CPB	2·8	82	21	18	48	Carboxypeptidase B – 1976 – *y* (0·33)
1PTE‡	2·8	340	7	1	31	Carboxypeptidase/transpeptidase – 1985 – *n*
2HVP	3	94	0	15	36	HIV-1 protease – 1989 – *y* (0·37)
1TS1	3	321	45	7	65	Tyrosyl tRNA synthetase – 1982 – *y*
1GBP	3	292	20	0	47	Galactose-binding protein – 1983 – *n*
1PEP	3	327	0	0	28	Pepsin – 1978 – *n*
1LRP	3·2	89	52	0	56	Lambda repressor (N-domain) – 1987 – *n*
1THI§	3·2	207	0	4	30	Thaumatin – 1989 – *n*
2XIA	3·5	382	43	7	64	D-Xylose isomerase – 1988 – *n*
1PGI	3·5	514	32	4	52	Glucose-6-phosphate isomerase – 1977 – *n*
1KGA	3·5	173	17	0	28	KDPG aldolase – 1978 – *n*
1HRB	5·5	113	58	0	64	Hemerythrin B – 1976 – *n*
1HR3	5·5	118	64	0	70	Hemerythrin – 1983 – *n*
1LZH	6	129	31	8	64	Lysozyme (hen) – 1981 – *n*
2LZH	6	129	35	8	66	Lysozyme (hen) – 1981 – *n*
5PFK	7	319	41	17	70	Phosphofructokinase – 1988 – *y*
2TMA	15·0	284	96	0	96	Tropomyosin – 1987

Secondary structure was calculated from the backbone co-ordinates of models built with permissive joint tolerance. Five structures, 1PEP, 1PTE, 1KGA, 1AAT and 1THI, have more than 10 percentage points fewer backbone hydrogen-bonded residues than any protein in the test set of fully known proteins (excluding 1CY3, which is a preliminary dataset, from the comparison), indicating difficulties in model construction due to co-ordinate inaccuracies. Gaps or poor hydrogen bonding may indicate incomplete refinement or errors in chain tracing.

Code, Resol. and Length, as for Table 1.

%H, percentage of residues in α-helical structure; %E, percentage of residues in β-strands; %hb, number of backbone hydrogen bonds/100 residues (Kabsch & Sander, 1983).

† Chain tracing is uncertain for residues 67 to 158.

‡ Tentative model with uncertain chain tracing.

§ Preliminary dataset.

2/4ATC (aspartate transcarbamoylase) have an unusually large number of peptide flips, an unusually low number of backbone hydrogen bonds or an unusually large number of gaps, indicative of inaccuracies in the co-ordinates. The by now classical example of a structure with incorrect chain trace, data set 2FD1 of ferrodoxin (Ghosh *et al.*, 1982, corrected in 4FD1, C. D. Stout, 1989), is exceptional with 11 chain breaks and 35 flipped peptides in a chain of 106 residues (Fig. 3). The incorrect chain trace (switch of 2 β-strands) in the regulatory subunit of aspartate transcarbamoylase (2ATC, Ke *et al.*, 1984; corrected in 7ATC, Kim *et al.*, 1987) leads to 32 peptide flips out of 152 residues (21%, to be compared with, at most, 5% in Fig. 2(b)) and three gaps. Cytochrome *c*3 (1CY3, Pierrot *et al.*, 1982) reconstructed from the experimental C^{α} trace also has an unusually high fraction of flipped peptides compared to the original model (20%). The conventional (ϕ, ψ) plots for X-ray structure and model (Fig. 4) show that the reconstructed model has fewer points in sterically unfavorable regions (Ramachandran *et al.*, 1963) and is possibly an improvement over the experimental structure. However, there is suspiciously little (36%, Fig. 2(c)) hydrogen-bonded structure even in the "better" reconstructed model, so we expect that the chain trace will undergo changes upon further refinement of this preliminary structure.

The algorithm for building backbones could be applied in crystallography to decide the orientation of peptide groups, especially in regions of low electron density. An example of correctly predicted flips

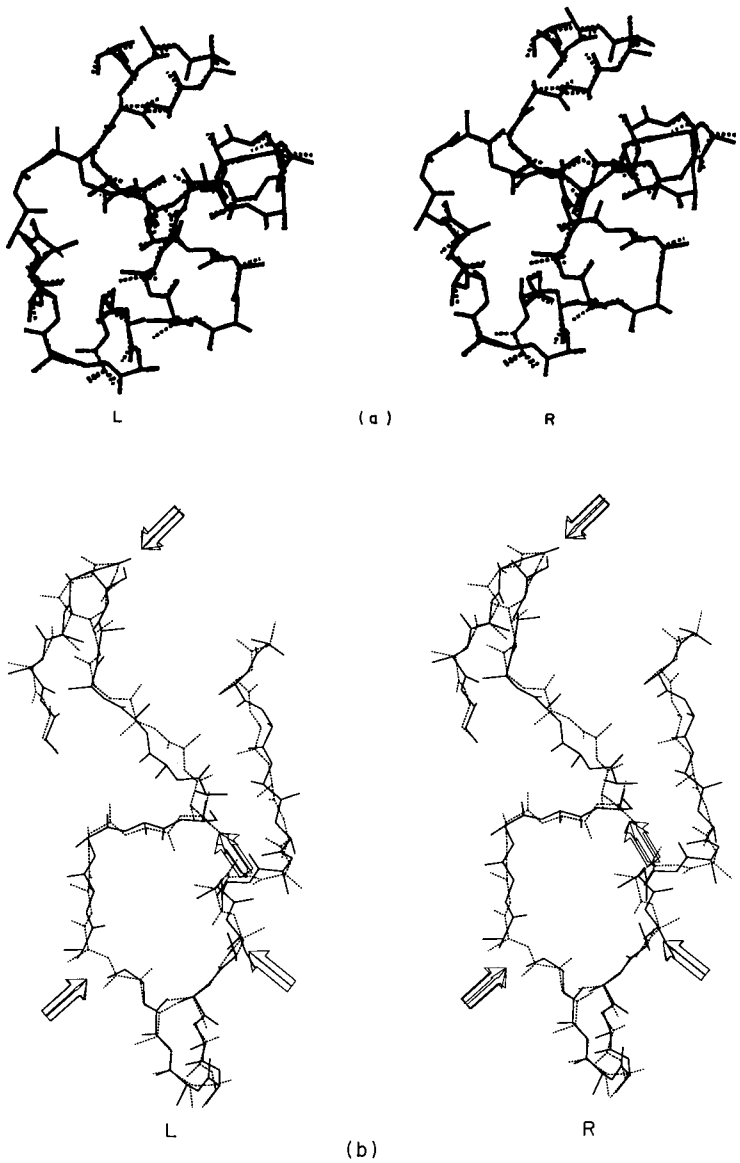

Figure 3. Examples of (a) successful (3WGA, residues 1 to 45) and (b) unsuccessful (2FD1, residues 1 to 32) backbone reconstruction. The backbone of wheat-germ agglutinin (3WGA) was reconstructed in stringent mode from its C^{α} trace with a C^{α} r.m.s. of 0·14 Å, backbone r.m.s. of 0·42 Å and 1 flip at residue 32. The chain trace of ferredoxin dataset 2FD1 is incompatible with backbone geometries found in the fragment database. The reconstructed model has a C^{α} r.m.s. of 0·53 Å, backbone r.m.s. of 1·36 Å, 35 flips and 11 gaps, of which four are marked by arrows in (b). The reconstructed models are shown as continuous lines and the X-ray structures as broken lines. Drawn using WHAT IF by G. Vriend.

comes from immunoglobulin structures: 14 flips compared to the X-ray co-ordinates are predicted in the reconstructed model of the variable domains of the IgA Fab fragment J539 (1FBJ, Suh *et al.*, 1986) at lower resolution (2·6 Å). At 1·9 Å resolution (2FBJ, T. N. Bhat, E. A. Padlan & D. R. Davies, unpublished results) the reconstructed model has nine flips with the balance (5 peptide groups) changed relative to the earlier structure (1FBJ) precisely as predicted. The chain tracing clearly has improved with higher resolution: the reconstructed 1FBJ model had two gaps, while the refined model (2FBJ) had none. Alwyn Jones (personal communication) has implemented a similar and independently developed correction mode using database fragments as a crystallographic tool in the program "O". Wolfgang Kabsch (Kabsch *et al.*, 1990), also independently, has written a procedure for constructing full co-ordinates based on dynamic programming and applied it to crystallographic model building.

(c) *Quality of model side-chain co-ordinates*

The quality of side-chain construction also was assessed in a blind test by adding side-chains to models (Table 1) for which the backbone had been reconstructed in stringent mode. The full models were then compared to the original X-ray structures. The results are comparable to starting from the backbone co-ordinates of the X-ray structures (data not shown).

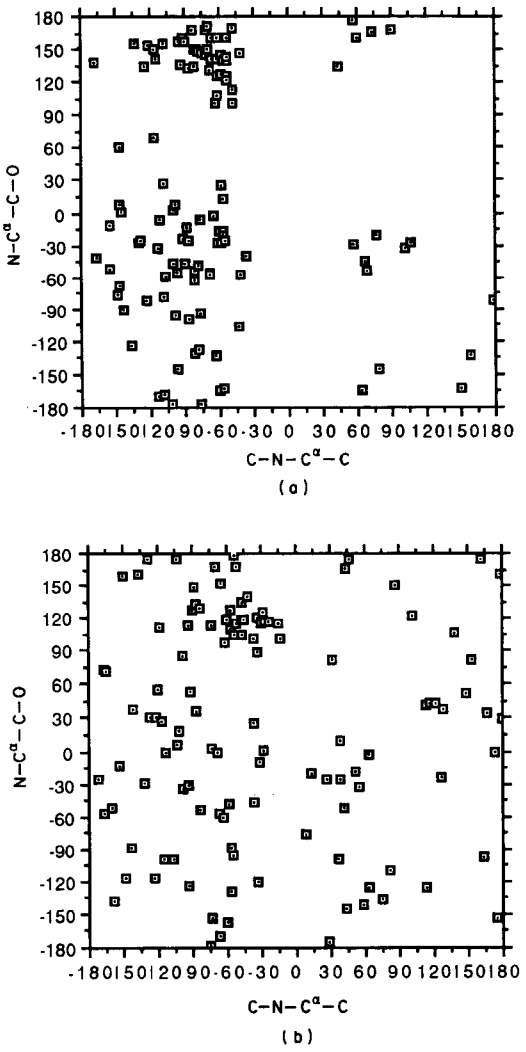

Figure 4. Potential errors in protein co-ordinates. Contrast between the distributions of backbone ϕ, ψ angles (a) for the reconstructed model and (b) the original dataset for cytochrome *c*3 (1CY3) points to possible inaccuracies in the original dataset in agreement with the unusual quality indices (Fig. 3) for this structure. The reconstructed dataset may be a reasonable 1st attempt at correction. Note the atypical distribution outside of the "allowed" regions in the original dataset. The authors of the structure of this multihaem cytochrome *c*3 (Pierrot *et al.*, 1982) did remark on the incomplete nature of the dataset: preliminary refinement to an *R*-value of 0·34 for 2085 reflections between 3·0 and 6·0 Å resolution; the ω angles (peptide unit torsion) flagged as deviating significantly from the expected value for residues 27-28, 61 to 63, 71-72, 83-84, and 106-107 (see comments in the Protein Data Bank file).

Side-chain conformations of high-resolution proteins can be modelled very well by rotamers in staggered conformations (Fig. 5(a)), while unusual side-chain conformations can occur in low-resolution structures. Given the restriction to the subspace of rotamers, the best possible value for side-chain atom r.m.s. (rotamers closest to X-ray structure) is on the average 1·0 Å, while the worst possible is on the average 4·0 Å. The reconstructed models average 2·21 Å r.m.s. for side-chain atoms in proteins of better than 2·5 Å resolution. Side-chain construction is more successful in the protein interior (relative side-chain solvent accessibility less than 20%, corresponding to 40% of total residues), with the average side-chain atom r.m.s. deviation for buried residues reduced to 1·56 Å (Fig. 5(b)). As another measure, χ^1 side-chain dihedral angles are correct within 30° for an average of 70·2% of core residues in proteins determined at better than 2·5 Å resolution (Fig. 5(c)). Visual comparison of several models and X-ray structures gives the impression that side-chain positions agree remarkably well.

(i) *Limitations of the simplified Monte Carlo method*

The main limitation comes from working in the subspace of discrete rotamers, i.e. small adjustments in dihedral angles are not possible. In some cases, the rotamer closest to the X-ray structure has a rather high energy due to clashes with the main-chain or other side-chains, especially in the interior. Also, a small error in the position of the C^{β} atom or in the χ^1 or χ^2 angle of the large aromatic residues can significantly affect neighboring residues, which may be prevented from adopting the correct rotamer conformation. A second limitation is in the lack of a solvent term (e.g., see Colonna-Cesari & Sander, 1990) in the energy function, so that rather weak criteria apply to the flexible surface side-chains, and holes can be left in the core. A more complete exploration of side-chain conformation would require a more refined potential energy function as well as setting free many more side-chain and main-chain degrees of freedom, i.e. fully relaxing atomic co-ordinates. Given the limitations mentioned above, our simplified and very rapid approach works surprisingly well.

(d) *Comparison with other methods*

Purisima & Scheraga (1984) have derived an analytical solution to the problem of selecting ϕ and ψ angles given fixed positions of C^{α} atoms. Database fragment methods are less sensitive to deviations from strict standard geometry in the original co-ordinates than their approach. The application of dynamic programming to joint optimization used here appears to improve main-chain construction compared to the method of Claessens *et al.* (1989) (Table 5, section A.1). Tuffery *et al.* (1991) approach the problem of adding side-chains to the backbone in a way rather similar to ours, but include electrostatic and directional hydrogen-bonding terms in the energy function and use a more complex Monte Carlo algorithm, and obtain, on the average, 0·07 Å better r.m.s. deviations (Table 5, section A.2).

Summers & Karplus (1989) report a case study of building side-chains for rhizopuspepsin using structural information on the homologous penicillopepsin. Unfortunately, this method has not yet been automated. As co-ordinates for rhizopuspepsin are not generally available, we tested our Monte Carlo

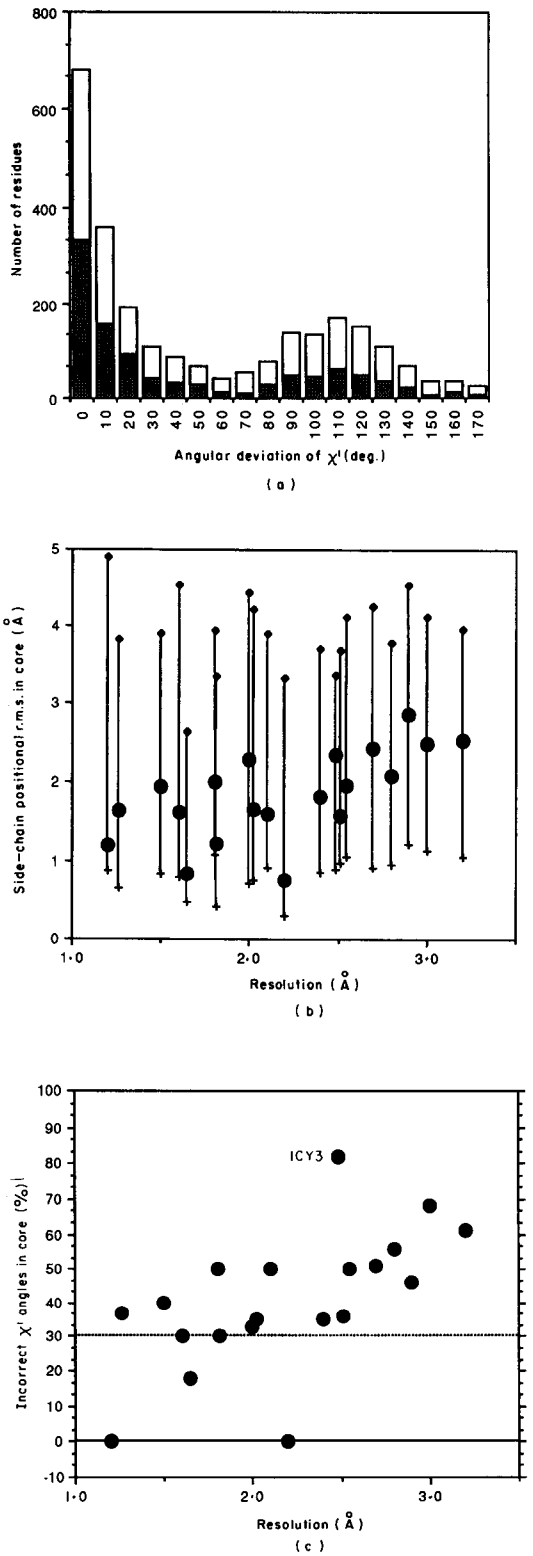

Figure 5. Quality of side-chain construction in terms of dihedral angle deviations and positional r.m.s. deviation. Residues were classified as buried or exposed by comparing the solvent-accessible surface area in the X-ray structure to the surface in an extended peptide Gly-Gly-X-Gly-Gly (values from Baumann *et al.*, 1989); 20% was used as the threshold. (a) Distribution of χ^1 angle deviations in the reconstructed models (Table 1) relative to X-ray structures. χ^1 angles are modelled within 30° in 67% of core residues and in 54% of exposed residues. (□) Exposed; (▩) core. (b) r.m.s. position deviation calculated for side-chain atoms (●); in proteins of better than 2·5 Å resolution the average is 2·21 Å for all residues and 1·56 Å in the core. (+) Best possible; (◆) worst possible. Using discrete rotamers limits the possible range of values to an average of between 1·0 Å (minimum) and 4·0 Å (maximum). (c) Percentage of incorrect χ^1 angles (deviation larger than 30° relative to X-ray structure) in the core. An average of 70·2% are correct in proteins determined at better than 2·5 Å resolution.

procedure on penicillopepsin (Protein Data Base dataset 2APP). Fixing side-chains that are identical in sequence (35% sequence identity) between the two proteins improved the prediction of χ^1 angles by one-sixth to 84% compared to making no use of homology information, which gave 72% correct χ^1 angles (within 40° of X-ray structure). Steps 2 and 3 of Summers & Karplus (1989), which involve a sophisticated set of homology-derived rules to transfer χ angles combined with rigid rotation mapping of van der Waals energy, reach 86% correct χ^1 angles (Table 5, section B.1). Up to this stage, MaxSprout appears to perform comparably. In a subsequent energy refinement step, Summers & Karplus (1989) further improve their model to 92% correct χ^1 angles.

A final comparison can be made to the exercise of Reid & Thornton (1989), who rebuilt flavodoxin from its C^α co-ordinates. We removed the dataset for flavodoxin (3FXN) from the fragment database and built full co-ordinates based on the C^α trace. In terms of backbone geometry, as well as side-chain positions, the automatic database algorithm gave better results than the careful manual approach of Reid & Thornton (Table 5, section B.2).

(e) *Overall assessment of the method*

For practical applications it is important to have an estimate for the expected accuracy of a constructed model. Based on the noise test, we know that up to about 0·5 Å inaccuracies in C^α positions are tolerated by the backbone construction procedure using stringent mode, whereas inadequate models are obtained for traces with inaccuracies above 0·5 Å. If there are no or very few gaps in a reconstructed model, the calibration shows that the probable accuracy of backbone co-ordinates lies between 0·4 and 0·6 Å. The occurrence of many gaps in a reconstructed backbone model should alert the user not to trust the results and to attempt to obtain more accurate C^α positions. In many cases, the accuracy of reconstructed backbones is an excellent starting point for the addition of side-chains. Apparently, a considerable amount of information about main-chain and C^β geometry is contained in the C^α trace alone. Occasional flips have little influence on side-chain construction: anyone equipped with a set of plastic molecular models can easily verify the fact that the orientation of C^β atoms is not much altered by flipping peptide units by 180°.

Table 5

Comparison of MaxSprout and other methods for building backbones and side-chains

A. *Automated methods applied to more than one protein*

A.1 *Building backbone from fragments, given C^{α} co-ordinates*

		Claessens *et al.* (1989)		MaxSprout		
Protein	Length	Main-chain r.m.s. (Å)	Flips	Main-chain r.m.s. (Å)	Flips	Gaps
5cpa	307	0·55	15	0·48	8	2
1tim	247	0·56	12	0·59	11	0
2cts	437	0·62	Not reported	0·45	10	1

None of the reconstructed proteins was present in the fragment database used for building them (i.e. 5cpa was removed from the fragment database in rebuilding 5CPA). r.m.s., root-mean-square deviation. Note the slightly better performance of MaxSprout.

A.2 *Optimization of rotamer choice in adding side-chains to given backbone co-ordinates*

		Tuffery *et al.* (1991)			MaxSprout		
Protein	Length	sc r.m.s.	Core sc r.m.s.	Opt. E	sc r.m.s.	Core sc r.m.s.	Opt. E
1CPV	109	1·93	1·71	1303	1·65	2·05	−108
1CRN	46	1·46	1·95	−14	2·12	2·29	−303
1CTF	74	1·51	1·07	−36	1·85	1·29	−315
1FDX	54	1·73	1·35	3158	2·20	2·16	82
1LZ1	130	2·15	1·31	2986	1·68	1·19	−849
1MLT	26	1·86	1·07	481	1·63	1·37	−78
1NXB	62	2·11	2·03	50,330	2·05	2·02	536
1PPT	36	1·43	0·94	−109	1·91	1·37	−111
1SN3	65	2·07	2·04	385	2·36	2·00	−275
1UBQ	76	1·71	1·27	−195	1·91	0·80	−478
2OVO	56	1·64	1·10	−62	1·79	1·05	−191
4LYZ	129	1·85	1·52	14,140	1·72	1·43	−344
4PTI	58	2·05	2·04	244	1·92	1·54	−217
5CYT	104	2·25	2·12	1981	1·99	1·87	−496
Average		1·84	1·54	5328	1·91	1·60	−225

sc, side-chain; core sc, side-chains in core (less than 25% relative solvent accessibility); r.m.s., root-mean-square deviation; Opt. E, optimized energy. The energy function used in MaxSprout is much simpler than that employed by Tuffery *et al.* (1991) and hence the optimized energy is comparable only within one and the same method (arbitrary units). Note the slightly better performance of Tuffery *et al.* in terms of average side-chain r.m.s. deviation.

B. *Methods not yet automated and applied to only one test case*

B.1 *Building side-chains in homology modelling: C-terminal lobe of rhizopuspepsin/penicillopepsin (residues 181 to 323)*

Summers & Karplus (1989): rhizopuspepsin† % correct within 40°			MaxSprout: penicillopepsin‡		
	χ^1	χ^2		χ	χ^1 & χ^2
(1) χ angle transfer rules for conservative substitutions	58	44	(A) Monte Carlo with all side-chains free	72	59
(2)+(3) Rigid rotation maps +χ rules for ambiguous cases	86	75	(B) Monte Carlo with 35% fixed side-chains	84	73
(4) Energy refinement	92	81			

† Side-chains were added to the backbone of rhizopuspepsin using information on dihedral angles in the homologous protein penicillopepsin (35% identical residues).

‡ MaxSprout was used to add side-chains to the backbone of penicillopepsin (available in the Protein Data Bank as entry 2APP). Model (A) is an optimization of rotamers given only the backbone co-ordinates. The starting structure of model (B) also included side-chain co-ordinates for the residues that are identical between rhizopuspepsin and penicillopepsin (35% identical residues) and these were kept fixed during the optimization.

Note the better performance of step (4) of Summers & Karplus (1989).

Table 5. *(continued)*

B.2 *Rebuilding full co-ordinates from C^{α} co-ordinates: flavodoxin (3FXN, 138 residues)*

	Reid & Thornton (1989)	MaxSprout
Backbone r.m.s. (Å)	0·57	0·48
All atoms r.m.s. (Å)	1·73	1·57
Side-chain atoms r.m.s. (Å)	2·41	2·19
Peptide flips	41–42, 57–58, 74–76, 89–92, 117–121	At 75, 89, 111, 121
χ^1 correct within 20° (%)	40	44
Both χ^1 and χ^2 within 20° (%)	17	25

Note the slightly better performance of MaxSprout.

In our current procedure, the accuracy of side-chain co-ordinates is limited because of the simplicity of the rigid rotamer representation and the complexity of global optimization. The limited rotamer library is, in principle, capable of producing models with a side-chain atom r.m.s. deviation of about 1·0 Å. Our Monte Carlo method with simple energy functions does not achieve the theoretically possible optimum, but it is a quick, simple, fully automatic, and general way of generating a reasonable starting point for further refinement.

In summary, we have demonstrated that the database algorithm may be useful at certain stages in experimental structure determination or model building by homology, e.g. by generating a set of full co-ordinates from incomplete information as a starting point for further refinement. Also, it may be useful as the second step in *de novo* protein design, after constructing an outline of the backbone. Finally, by restricting its application to parts of a chain, it can be used in any situation where the database search for single continuous loops (Jones & Thirup, 1986) is considered inadequate, e.g. in crystallographic model building. The program MaxSprout, written in Pascal with subroutines in FORTRAN, is available from the authors on request (academic or other licence agreement).

Support (to L.H.) by the Academy of Finland (S.R.C) and an EMBO short-term fellowship is gratefully acknowledged.

References

Baumann, G., Froemmel, C. & Sander, C. (1989). *Protein Eng.* **2**, 329–334.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. & Thornton, J. M. (1987). *Nature (London)*, **326**, 347–352.

Branden, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.

Claessens, M., Van Cutsem, E., Lasters, I. & Wodak, S. (1989). *Protein Eng.* **2**, 335–345.

Colonna-Cesari, F. & Sander, C. (1990). *Biophys. J.* **57**, 1103–1107.

DeGrado, W. F., Wasserman, Z. R. & Lear, J. D. (1989). *Science*, **243**, 622–628.

Ghosh, D., O'Donnell, S., Furey, W., Robbins, A. H. & Stout, C. D. (1982). *J. Mol. Biol.* **158**, 73.

Jones, T. A. (1985). *Methods Enzymol.* **115**, 157–171.

Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.

Kabsch, W. (1978). *Acta Crystallogr. sect. A*, **34**, 827–828.

Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F. & Holmes, K. C. (1990). *Nature (London)*, **347**, 37–44.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Ke, H. M., Honzatko, R. B. & Lipscomb, W. N. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 4037.

Kim, K. H., Pan, Z. X., Honzatko, R. B., Ke, H. M. & Lipscomb, W. N. (1987). *J. Mol. Biol.* **196**, 853–875.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). *Science*, **220**, 671–680.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087.

Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.

Pierrot, M., Haser, R., Frey, M., Payan, F. & Astier, J.-P. (1982). *J. Biol. Chem.* **257**, 14341.

Ponder, J. W. & Richards, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.

Purisima, E. O. & Scheraga, H. A. (1984). *Biopolymers*, **23**, 1207–1224.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95–99.

Reid, L. S. & Thornton, J. M. (1989). *Proteins*, **5**, 170–182.

Richardson, J. S. & Richardson, D. C. (1989). *Trends Biochem. Sci.* **14**, 304–309.

Sander, C. (1987). Editor of *Protein Design*, EMBL BIOcomputing Technical Document 1.

Smith, T. F. & Waterman, M. S. (1981). *J. Mol. Biol.* **147**, 195–197.

Stout, C. D. (1989). *J. Mol. Biol.* **205**, 545–555.

Suh, S. W., Bhat, T. N., Navia, M. A., Cohen, G. H., Rao, D. N., Rudikoff, S. & Davies, D. R. (1986). *Proteins*, **1**, 74–80.

Summers, N. L. & Karplus, M. (1989). *J. Mol. Biol.* **210**, 785–811.

Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). *J. Biomol. Struct. Dynam.* In the press.

Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. (1987). *J. Mol. Biol.* **194**, 531–544.

Edited by B. W. Matthews