PROTEINS: Structure, Function, and Genetics 1:146–163 (1986)

An Algorithm for Determining the Conformation of Polypeptide Segments in Proteins by Systematic Search

J. Moult and M.N.G. James
Medical Research Council of Canada Group in Protein Structure and Function, Biochemistry Department, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

ABSTRACT **The feasibility of determining the conformation of segments of a polypeptide chain up to six residues in length in globular proteins by means of a systematic search through the possible conformations has been investigated. Trial conformations are generated by using representative sets of ϕ, ψ, and χ angles that have been derived from an examination of the distributions of these angles in refined protein structures. A set of filters based on simple rules that protein structures obey is used to reduce the number of conformations to a manageable total. The most important filters are the maintenance of chain integrity and the avoidance of too-short van der Waals contacts with the rest of the protein and with other portions of the segment under construction. The procedure is intended to be used with approximate models so that allowance is made throughout for errors in the rest of the structure. All possible main chains are first constructed and then all possible side-chain conformations are built onto each of these. The electrostatic energy, including a solvent screening term, and the exposed hyrophobic area are evaluated for each accepted conformation. The method has been tested on two segments of chain in the trypsin like enzyme from *Streptomyces griseus*. It is found that there is a wide spread of energies among the accepted conformations, and the lowest energy ones have satisfactorily small root mean square deviations from the X-ray structure.**

Key words: protein conformation, energy calculations, protein modeling, loop conformation, surface area, homologous proteins

INTRODUCTION

This paper describes a method of determining the conformation of short regions of polypeptide chain in globular proteins, assuming an approximate model for the rest of the molecule. The method was developed to assist in the building of models of proteins of unknown structure but for which the amino acid sequence is homolgous to one or more proteins of known structure. It may also be applied to the prediction of effects on structure of single amino acid replacements, such as those introduced by site directed mutagenesis experiments.

For a number of years now, models of proteins of known sequence but unknown structure have been based upon the structures of proteins related to them by sequence homology: for example, construction of bovine trypsin from α-chymotrypsin[1]; α-lytic protease from elastase [2]; α-lactalbumin from hen egg white lysozyme[3]; thrombin and other blood factors from serine proteases[4]; and renin from a fungal aspartic protease.[5] The recent enormous increase in the number of sequences available, together with a steadily increasing number of structures determined by X-ray crystallography, has resulted in many more candidates for such models, especially those for use in drug design activities.

Although a number of models have been produced in this manner, there are so far few examples of the subsequent critical checking of the predicted structure against that determined from X-ray crystallography.[6,7] One of these is the case of the structure of *Streptomyces griseus* trypsin (SGT), which was modeled on the basis of the known structure of bovine trypsin (BT).[8,9] Jurášek et al.[8] carefully built a model of SGT from Watson-Kendrew wire model parts, using as a basis the preliminary structure of BT provided by R.M. Stroud. The model of Greer[9] was an outline only and did not include a set of derived atomic coordinates. The structure of SGT has now been solved, refined at 1.7 Å resolution to an R-factor of 16%, and compared with the Jurášek et al. model.[7,10]

SGT has 33% sequence identity with BT, and the model exploited this relationship by assuming that the main chain followed the same course in regions of homologous sequence. The comparison of the X-ray structure of SGT with that of BT showed this was a reasonable approximation: 122 of the 223 residues have α-carbon atom positions within 1 Å after a rigid body superposition of the two structures, and a further 63 α-carbon atoms are within 1.9 Å. These values are in agreement with an analysis of the relationship between structural and sequence differ-

Received June 24, 1986; accepted July 31, 1986.

Address reprint requests to Dr. Michael N.G. James, Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada.

ences for homolgous proteins.[11] Superposition of fragments of the structures produces significantly smaller differences,[10] indicating that local structure is more conserved than global structure.

The Jurášek et al. model had two types of defect. One stretch of about 25 residues with low sequence homology to BT was seriously misaligned, leading to an erroneous prediction of this segment of structure. Given an awareness of the possibility of such sequence misalignments one can consider the small number of alternatives. In this paper, we assume that a correct sequence alignment has been made. More seriously, the four regions of 3 to 6 residues in length, where there are insertions and deletions in one structure relative to the other, were built incorrectly. We do not believe that this reflects incompetence on the part of the modellers, but rather reflects the large number of possible structures which must be examined in building even short pieces of chain. It will become clear below that there are in general many more than can be investigated by a human operator. In the case of SGT these regions represent segments of chain with limited functional relevance. More usually though, they will be just those parts of the structure leading to a different function (consider the likely origins of the differences in pH profile and substrate specificity of renin compared with other aspartic proteases of known structure, for instance).

Several possible approaches to the modeling of such short stretches of chain are possible. The first, and most widely used, is simply to examine the structure on a computer graphics system and to use the knowledge of the operator to help in identifying the correct conformation. The evidence from the SGT work is that this approach is unlikely to work and is open to the objections that there are no objective criteria for success and that the results are not reproducible. A second approach is to examine a data base of known structures and to extract the subset fitting the end points of the region under consideration and containing the correct number of residues.[12,13] This appears to be a useful approach for suggesting possible courses the chain might take, thereby providing starting points for more detailed modeling. A third approach is to take an inital model produced by hand and subject it to a molecular dynamics simulation. Although this technique will produce energetically more respectable coordinates—i.e., obvious clashes and bad geometry will be corrected—it is unlikely to converge successfully unless the starting conformation is very close to the correct one. Since the ends of the segment are covalently bonded to the bulk of the protein, and there are always close interactions with this portion, the energy barriers to be overcome in making any significant conformational changes are far too high. The use of "soft atoms"[14,15] might provide a means of overcoming this obstacle. A fourth approach is the use of distance geometry algorithms[16] to build a number of possible structures for the chain.[17] This approach is still largely unexplored.

None of these methods provides a satisfactory solution to the problem of modeling short segments of chain at the present time. We examine here a fifth approach which provides a practical approach at this stage of development and which has considerable scope for extension. We show that it is feasible to systematically search the set of possible conformations of short segments of chain, using filters to keep the number of conformations manageable, and to select from amongst those generated one close to the correct structure. A related technique has been used for peptides, with fixed orientations of the side chains.[18,19] Here we include all degrees of rotational freedom about single bonds and make sure of information about the rest of the structure to restrict the number of possiblities. An important consideration in the development of this algorithm has been its usefulness in situations where there are substantial errors in the rest of the model.

METHODS

Properties of a Systematic Search of Conformations

Is a systematic search of the possible conformations of short stretches of chain possible? The many different conformations a protein-length piece of polypeptide chain can adopt have long been recognized.[20] How many conformations must be considered for short lengths of chain for our present purposes? This is determined by the accuracy required in the final result. Our objective here is to obtain amongst the generated structures one which is sufficiently close to the true one that it can be recognized as such by a suitable evaluation of its energy and comparison of this quantity with the energies of all other conformations. Our experience suggests that this requirement implies a maximum acceptable error in position of approximately 1.5 Å for any atom, with a root mean square (RMS) error of around 1 Å. Thus we must sample the conformations in such a way as to be assured of including in the set at least one within that range of the true structure.

The simplest approach to such sampling, and the one adopted here, is to generate different conformations by using the dihedral angles around the single bonds in the polypeptide backbone (ϕ and ψ) and side chains (χ_n) with standard geometry for the bond lengths, bond angles, and partial double bond torsion angles (a so-called rigid geometry description). The density of sampling in this angle space does not map in a straightforward manner onto the Cartesian space, because of coupling between successive angles along a chain.

For lengths of up to six residues we have found that the desired degree of Cartesian space sampling can be obtained by using a set of up to at most 11 ϕ, ψ pairs for the main chain of each residue. Examination

of the distribution of side-chain rotation angles indicates that the staggered orientations provide sufficiently fine sampling. This usually results in three different values per torsion angle, and, depending on residue type, between zero and four angles per side chain. This level of sampling produces a very large number of possible conformations for even a small number of residues: If we assume an average of two torsion angles per side chain, there will be 99 possible conformations per residue, or about 10^{10} for 5 residues. Clearly, such large numbers of conformations cannot be examined. The number of conformations needed would be even larger if it were not for the strong preferences for particular values of the angles found in protein structures. The success of the method depends on identifying such rules that known structures obey. These rules are then used to contain the number of conformations considered within reasonable bounds.

Rules and Filters

The large number of conformations possible is of course the result of the combinatorial explosion often encountered in search procedures. The solution to this problem is, as always, to prune branches from the tree of possibilities. Pruning is done by the use of filters that are devised from rules the system under consideration is believed to obey (for a general account of such procedures, see reference 21). At present, very simple filtering procedures are used for this application.

Useful rules are not restricted to any particular type: for instance, they need not necessarily be based directly on the relative energies of different conformations of the loop but may introduce statistical data as well. This type of heuristic approach in predicting protein structure was first suggested by Robson[22] in the context of secondary structure prediction, and Cohen et al.[23] have used such general filters to eliminate possible secondary structure element arrangements. For the current problem, we identify the following rules as useful:

Rule 1

Departures from standard geometry of bond lengths, bond angles, and torsion angles about nonsingle bonds are small. This is well established from small peptide structures and high-resolution refined protein structures.

Rule 2

Main-chain torsion angles about single bonds (ϕ and ψ) lie in well-defined ranges, which can be determined from refined protein structures.

Rule 3

Side-chain torsion angles adopt values close to staggered conformations for most residues.

Rule 4

There are no general nonbonded contacts between atoms substantially shorter than the sum of their van der Waals radii. The energy cost of such contacts is too high.

Rule 5

The conformation with the lowest electrostatic energy, taking into account solvent effects, is very close to the correct one. This assertion is supported by the outcome of this work.

Rule 6

The total exposed hydrophobic area is small.[24]

Two more rules which have not so far been incorporated in the algorithm are:

Rule 7

Packing is efficient, with few water-sized cavities.[25,26]

Rule 8

There are very few internal polar or charged groups not interacting favorably with other such groups. Although this is generally true, this effect has so far not been adequately analyzed.

The most useful rules are those which are effectively 100% obeyed: only then can they be used in one step for pruning. Less-powerful rules can be used to accumulate probabilities which eventually lead to pruning, but this level of sophistication has not been incorporated at present. Each utilized rule leads to one or more filters. Table I summarizes the procedure followed in the algorithm and indicates how rules and filters are associated. The following filters have been used:

(a) Loop closure (rule 1). Very few of the possible chains that can be built from a given starting point will come close to the desired end point, and of those that do, few will have a suitable geometry for linking to it. Because of errors in the model of the rest of the protein, and the finite sampling of dihedral angles, we cannot require exact closure or very good geometry at the link. Nevertheless, this is a very effective filter. It is employed here by splitting the loop to be built into two halves and building all possible main-chain conformations for these half lengths from their respective roots. This method has the advantage of drastically reducing the number of conformations to be built.

The positions of all end residues on one-half are compared with those of the end residues on the other half. Only pairs with ends for which the atoms which should superimpose (the imino nitrogen, N) are closer together than some distance D_{cl}, and for which the virtual bond between the Cα atoms of the last residues on each side is between specified limits B_{min} and B_{max} are selected to form complete loops. The geome-

TABLE I. Flow Chart of the Systematic Search Alogorithm*

Stage	Operation	Rule
Setup	Input segment sequence, coordinates of rest of structure, residue coordinate library.	
Build main-chain halves	Add main chain units	Standard geometry (1)
	Apply ϕ ψ rotations (b)	Restricted torsion angles (2)
	Closure of loop possible?	Standard geometry (1)
	Clashes with rest of protein? (a)	VW† contacts (4)
Form complete main chains	Find pairs of half chains to join (a)	Standard geometry (1)
	Standardize link geometry	Standard geometry (1)
Add side chains	Dock parts of side chains	Standard geometry (1)
	Apply x rotations (b)	Staggered side chain conformations (3)
	Clashes with rest of protein? (d)	VW contacts (4)
	Clashes with main chain? (d)	VW contacts (4)
	Clashes with side chains? (d)	VW contacts (4)
Choose a conformation	Evaluate energies of electrostatically distinct conformations (e)	Low electrostatic energy (5)
	Evaluate hydrophobic areas (f)	Small exposed hydrophobic area (6)

*Numbers and letters in parentheses indicate the rules and filters discussed in the text.
†VW stands for van der Waals.

try of each complete main-chain conformation formed by such pairs is then adjusted by an energy minimization procedure to obtain bond lengths, bond angles, and peptide planes throughout that are within acceptable deviations from the standard values.

Because of the variable effect of the different main-chain angles on the overall conformation, and the constraints of chain joining, some of the resulting conformations are very similar. The set is reduced to those for which at least one atom differs in position from the equivalent atom in all other sets by a specified amount δ_{uniq}, usually 0.8Å.

(b) Restricted set of torsion angles (rules 2 and 3). Plots of the distribution of main chain (for instance reference 27) and side-chain torsion angles[28] found in protein structures show that extensive regions of the conformational angle spaces are sparsely populated. To provide a data base for selecting angles for loop construction, a compilation of these distributions was made for 13 protein structures which have been determined at 2Å or higher resolution and refined to R-factors better than 20% (Herzberg, Moult, Sielecki, and James, unpublished).

Figure 1 shows this distribution of the ϕ, ψ angles for all residues other than glycine and proline. As others have pointed out,[31] the use of highly refined structures results in a considerable sharpening of the distribution, with many areas unpopulated, compared to the one obtained with unrefined structures. There is, as expected, a broad population in the α helix $(-63,-40)$ and β sheet $(-90,120)$ conformation regions, and a sparser occupation of the bridge between them. No residues with β-branched side chains fall in this latter region. There is also a population of the left-handed α helix (55,40) region, and this conformation must be included in the building set of angles.

A set of 11 pairs of ϕ, ψ angles (listed in the caption of Fig. 1) was compiled from this distribution such that most observed points are within 20° of a building point, and very few points are more than 30° away. This set is suitable for use in building non-β-branched residues other than glycine and proline. Three pairs are sufficient for proline and a total of 20 are necessary for glycine. β-branched residues can be adequately represented by nine pairs. Amongst the few outlying points are four very far away from the allowed regions. One of these (at $81,-75$) is from a protein solved in this laboratory (*Streptomyces griseus* protease A, SGPA), two (at $50,-120$; $73,172$) from actinidin, and one (at $-115,119$) from insulin. Baker[32] comments that these particular residues in actinidin are in surface loops and do not appear particularly constrained. The SGPA example is for an asparagine residue which is part of a tight turn and which lies in a very clear, strong piece of density in the electron density map. We are currently conducting calculations to investigate the nature of the strain apparently associated with this conformation. Some of the scatter around the populated regions probably comes from the effect of coordinate errors. For such refined proteins, the X-ray analyses indicate that the coordinate errors are around 0.2 Å RMS.[33,34] This value is

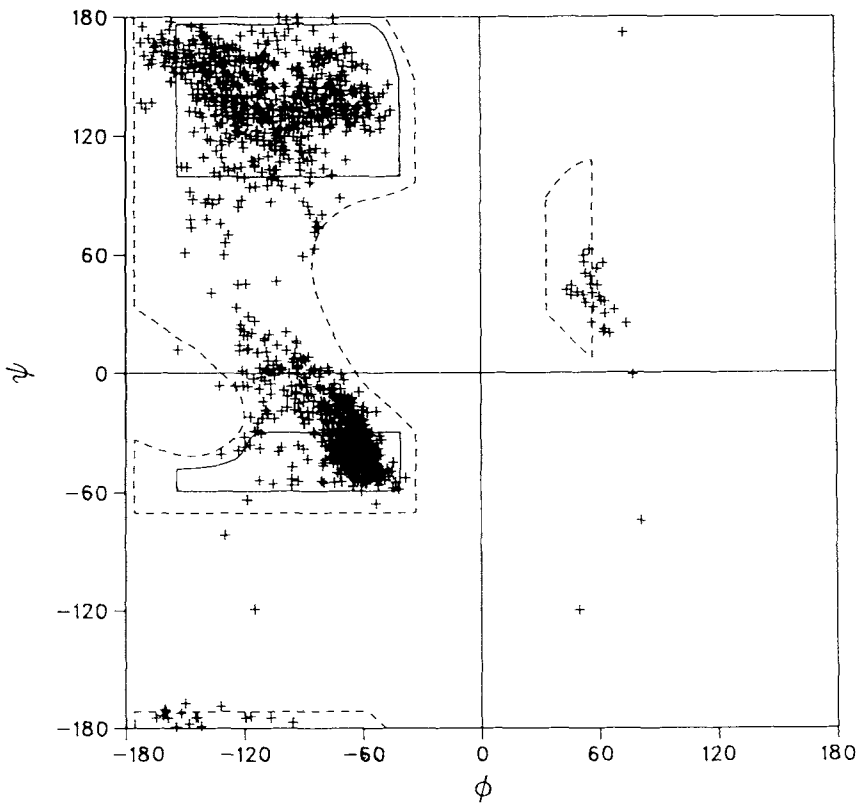


Fig. 1. Distribution of observed ϕ-ψ values for 13 refined protein structures. Glycine and proline residues are omitted. To help eliminate any less-reliable observations, residues are also omitted if one or more of the four atoms defining each torsion angle has a temperature factor greater than 20 Å^2. The continuous lines enclose areas that are fully allowed conformational regions and the broken lines show areas of acceptable van der Waals contact, for a $\tau(C\alpha)$ of 115°.[29] The left-handed α region is slightly misplaced. More sophisticated energy calculations correct this. With this proviso, all but four of the observed points fall within or just outside the outer limit. This tight clustering allows all populated regions to be represented by 11 points. The 11 points used in this work were; −160,160; −120,150; −120,110; −100,10; −90, −30; −80,130; −80,170; −80,70; −80,10; −70,−30, and 60,40. Criteria for inclusion of proteins were resolution 2.0 Å or better and an R-factor of less than 20%. For very similar structures, only one representative was used. The proteins are actinidin (2ACT); crambin (1CRN); insulin (1INS); lysozyme (1LZ1); myoglobin (1MBO); phospholipase (1BP2); hemerithrin (1HMQ); penicillopepsin (2APP); troponin C (Herzberg and James, in preparation); α-lytic protease (2ALP); protease A (2SGA); ovomucoid inhibitor (part 3SGB); and *Streptomyces griseus* trypsin (SGT).[10] Codes refer to entries in the Brookhaven data bank.[30] Details and references for the structures may be found there.

supported by comparison of identical structures solved under different conditions.[11] It implies errors in torsion angles of approximately 10° RMS.

Figure 2 shows sample distributions for side-chain torsion angles in this data base. Again, these distributions are very much sharper than those for unrefined proteins.[28] For the first two torsion angles along the chain (χ_1 and χ_2) the populated regions are around the staggered values (−60, 60, and 180°). Some combinations of staggered values are not observed (i.e., χ_1=60, χ_2=−60 for leucine), and other combinations occur only rarely. For the second pair of torsion angles (χ_3 and χ_4), the distributions are somewhat more scattered, although they still tend to be concentrated around the staggered positions.

On the basis of these distributions, side-chain angles were represented by the staggered values. For the χ_1 and χ_2 values this is a good approximation—for the χ_3 and χ_4 values, less so. However, errors in position from this sparse sampling of the end torsion angles of side chains will be relatively small because of the short distances the affected atoms subtend to these rotation axes.

(c) Van der Waals clashes with the bulk of the protein (rule 4). The atoms in the growing polypeptide chain should not clash unacceptably with those forming the fixed part of the model. Since the model is imperfect, and because of the finite sampling of the dihedral angles, it is not appropriate to apply criteria of close contacts between atoms based directly on the sum of their van der Waals radii. The use of smaller van der Waals radii to allow for these errors is also not a useful approach. Realistic errors of 1.0–1.5 Å would then permit very deep interpenetrations. To deal with

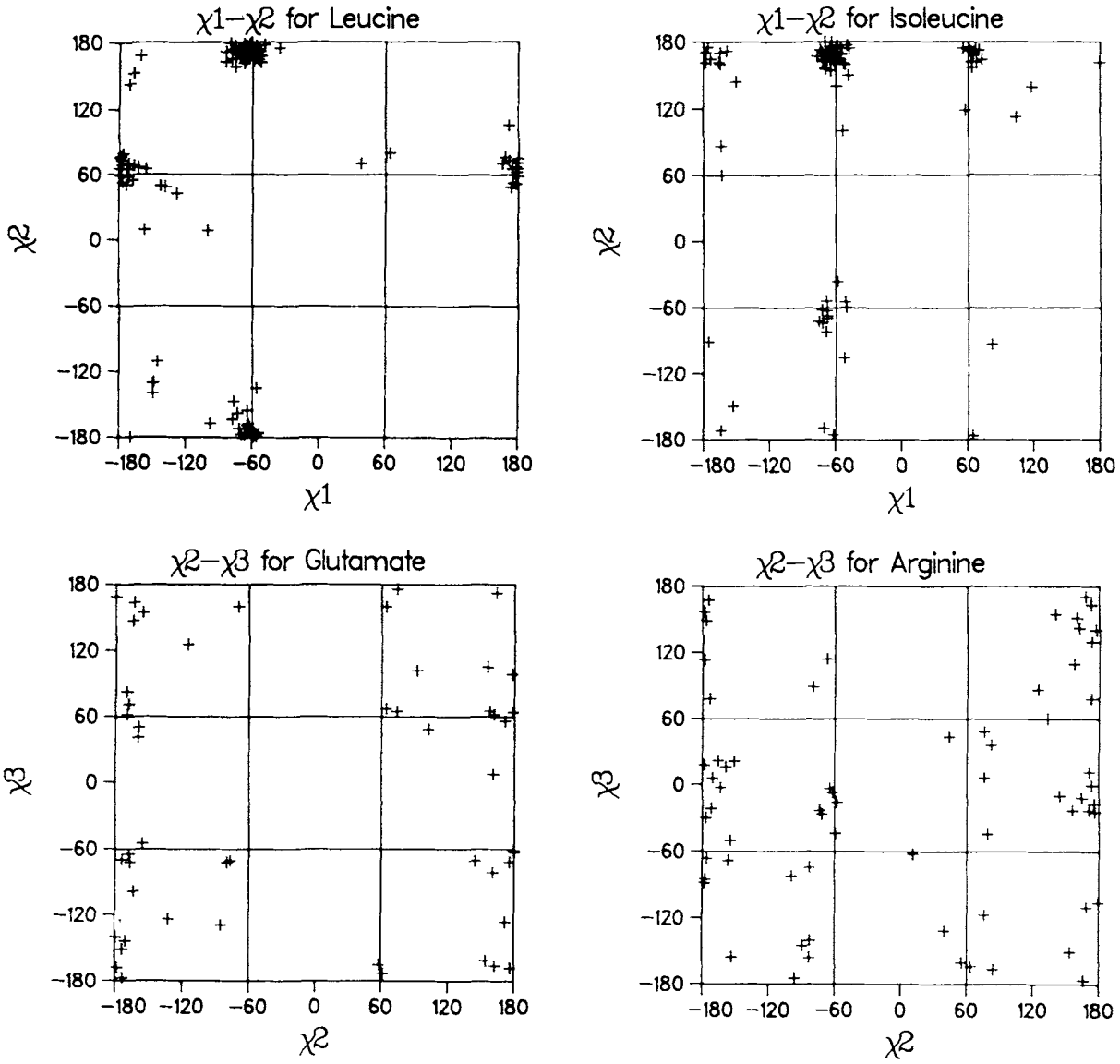


Fig. 2. Representative plots of the distribution of side-chain torsion angles for refined proteins. The proteins used to compile these are the same as those for Figure 1, and the same temperature factor selection criterion was used. For χ_1 and χ_2 there is a strong tendency to cluster around the values representing fully staggered conformations. These are at −60, 60, and 180°. Some combinations (i.e., 60, −60 for leucine) are not observed. The clustering is less strong for χ_3 and χ_4 values, though still evident. On the basis of this type of data, side chains were constructed by using the observed staggered angles for each χ stage of all side chains.

these circumstances, a "clash grid" is used. At each point, the distance to the nearest grid point where a standard atomic probe (radius 1.5 Å) can be positioned without making short contacts to the rest of the model is stored. Thus, if a point can itself accept a probe with no such clash, the grid value is zero. If the nearest such position is, for example, 1.5 Å away, a value of 1.5 Å is stored. Figure 3 shows an example of part of a grid in the region of one of the segments of SGT subjected to the systematic search in this work. The grid technique produces results similar to those produced with the use of small van der Waals radii in surface regions but allows only appropriately limited penetration into crevices in the underlying protein structure.

During the loop-building process, the value of the grid point closest to each atom of the growing chain is checked. If this value is greater than the maximum error believed to be associated with the positioning of this atom in relation to the rest of the molecule (ϵ_{pos}), the conformation is rejected. The appropriate value of ϵ_{pos} depends on the particular situation. In order to

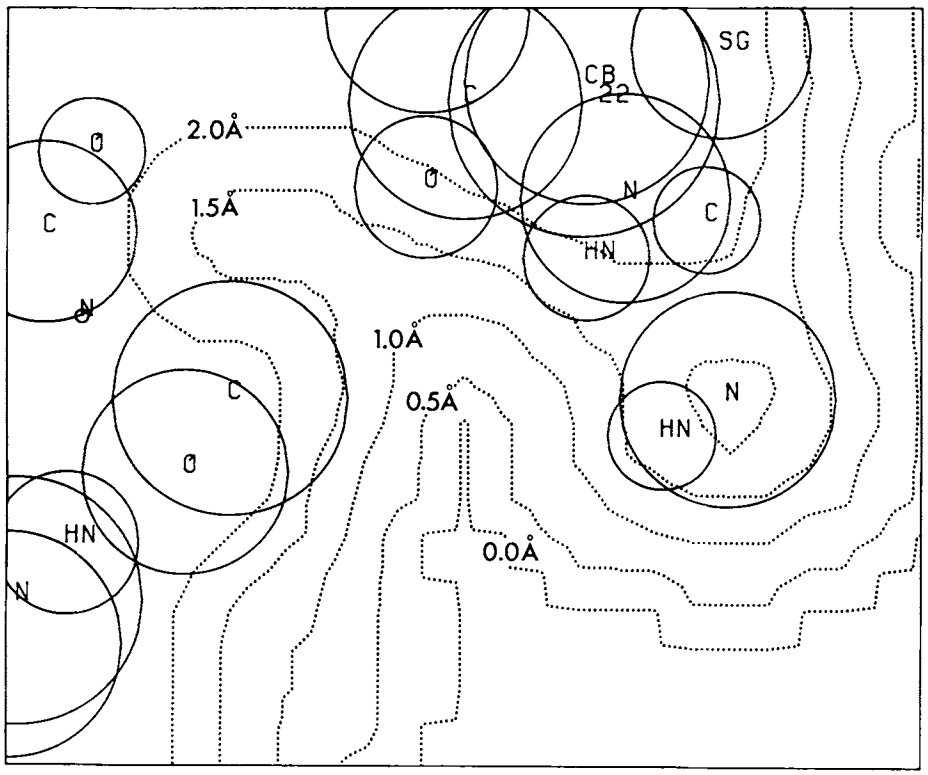


Fig. 3. Section of a clash grid used to test for unacceptably short van der Waals interactions between a segment and the rest of the protein. Circles represent the intersection of the van der Waals spheres of the atoms of the protein (other than the segment under construction) with this section of the grid. The contour line furthest from the protein surface (0.0 Å) shows the edge of the region accessible to the center of a spherical probe of radius 1.5 Å, without making short van der Waals contacts with atoms of the protein. This contour thus defines a surface equivalent to the solvent accessible surface of Lee and Richards, [35] and is the limit of the region where atoms of the segment under construction can be placed, if no allowance is made for errors. The points of the grid between this contour and the next one are within 0.5 Å of a probe-accessible point, and those to the next closet in contour, 1.0 Å, and so on. Thus, if it is assumed that errors in the bulk of the protein and in positioning atoms of the segment may be up to 1.5 Å, segment atoms may be positioned at any point outside of the contour at that level.

simulate a typical homologous modeling problem such as the building of SGT from BT, a value of 1.5 Å was used here, together with a 0.5 Å grid spacing. Such a large value allows many more conformations to be accepted than would be the case if errors in the bulk of the model were ignored.

(d) Side-chain contacts (rule 4). All acceptable main-chain loop conformations are first generated and then all acceptable side-chain conformation combinations are built into each of these. As individual side chains are built, a check is made for clashes with the bulk of the protein by using the clash grid in the same way as for the main-chain atoms. The distance to all main-chain atoms of the loop is also checked for van der Waals clashes. Errors in the model for the bulk of the protein do not affect these latter interactions, so that it is sufficient to use atom-atom distances with small van der Waals radii for this check. Contacts less than $D_{con}=2.5$ Å were considered unacceptable in this work. Side-chain–side-chain interatomic contacts within the segment are also checked in the same manner. This check requires all combinations of side-chain conformations along a particular main-chain loop conformation to be examined and is therefore performed last.

(e) Total electrostatic energy (rule 5). Application of the above filters will still allow many conformations to be generated for most loops. From these, a conformation must be selected which is close to the true one. In principle, the relative free energies of all the conformations should be evaluated. The commonly used approximation is made here of treating the enthalpy component (the potential energy) as representative of the total free energy. A screening term is included in the potential energy to allow for the effect of solvent. The relationship between potential energy and free energy may break down in this type of application because of the significance of the contribution to the entropy resulting from variations in solvent ordering associated with the exposure to solvent of nonpolar portions of the protein (the hydrophobic effect[36]). This is treated separately below.

The van der Waals energy of a conformation cannot be included in the potential energy because of the very short nonbonded contacts which have been allowed through filters c and d. The electrostatic energy is not so sensitive to errors in the coordinates and can be used to discriminate among conformations.

Different conformations of the hydrophobic side chains will only affect the electrostatic energy at a very second-order level through their possible effect on the solvent screening contribution. The degrees of freedom of these side chains may therefore be ignored in calculating the electostatic energy. For every accepted conformation of the main chain, and polar and charged side chains, a single accepted conformation is sought for the hydrophobic side chains, and, if found, it is included in the calculation of the electrostatic energy. Once a best conformation has been selected on the basis of the energy, all possible conformations of the hydrophobic side chains compatible with this arrangement of the main chain and polar and charged side chains are generated on it. The best overall conformation may now be selected from these on the basis of exposed hydrophobic area.

The method used to calculate the electrostatic energy is described in the section on computational details. Ideally, energy minimization should be employed at this stage to optimize all interactions in a particular conformation. However, this is too time consuming a procedure to use where the number of conformations is large. It turns out that this step is not necessary in order to select a conformation close to the most correct one.

It is desirable to include some model of the effect of the surrounding solvent on the electrostatic behavior of the system, since this is likely to be a major factor in distinguishing correct from incorrect structures.[37] We have used the method of image charges for this purpose.[38,39]

In spite of the various approximations involved, the energy obtained from these procedures has proven adequate for identifying a conformation close to the true one. The reasons for this are addressed in the Discussion.

(f) Exposed hydrophobic area (rule 6). The contribution from the amount of solvent-exposed non-polar surface to the free energy may in principle vary significantly amongst the possible loop structures. This quantity was defined as the sum of the areas of atoms forming parts of hydrophobic groups exposed to solvent for each conformation. The area thus obtained is proportional to the unfavorable contribution to the free energy from the hydrophobic effect.[40] For the cases examined, the variation in this quantity amongst the accepted conformations is small (100–200 Å^2), and it has not been considered together with the electrostatic energy. However, it does provide a means of choosing the best conformation from any set in which only hydrophobic side-chain conformations vary.

Computational details

An unusual feature of the procedure is that a large number of different conformations are held in store at one time: all the main-chain conformations in the first step and all the side-chain conformations on a particular main chain in the second. This necessitated the development of appropriate data structures. The algorithm has been implemented in FORTRAN. Although many of the data structures involved are not very convenient in this language, considerations of efficiency of execution leave little choice.

All coordinates for the building blocks of the structure were taken from the library associated with the Hendrickson-Konnert refinement program.[41] These had polar group hydrogens built onto them in standard positions, with staggered orientations for those on the Oγ atoms of serine and threonine. Hydrogen atoms bonded to the phenoxy oxygen atom of tyrosine residues were positioned in the plane of the ring. These three types of hydrogen atom give rise to additional rotational degrees of freedom in the systematic search of side-chain conformations.

A standard main chain-building unit was established from these coordinates. It consists of a central residue of N, HN, Cα, C′, and O′, as well as Cβ when present and Cγ and Cδ for prolines. This central residue was flanked by additional atoms for orientation with respect to other residues.

For each newly positioned main-chain unit, ϕ and ψ rotations were applied, using the standard set of angles (Fig. 1), and the positioned atoms were tested for clashes with the bulk of the protein (filter c). Van der Waals radii for all clash tests were taken from reference 42. It is not necessary to complete all half loops, since it can often be established that the conformation of a partially built section is such that closure has become impossible. To determine this, a check was made on the distance from the current end atom of the chain to the last atom on the other root of the segment. If this distance was larger than D_R, where D_R is the number of residues which are to be built to bridge this gap times 3.8 Å, that conformation was abandoned.

Accepted conformations of each main chain unit were added to the data structure, so that a tree of conformations was created, with nodes before each residue, and up to as many branches as conformation generating ϕ, ψ pairs from each node. When the tree for each half of the loop (rooted on the N or C side) had been built, complete loop conformations were generated by selecting N and C halves between which the geometry of the link was within acceptable limits (D_{cl} and B_{min}, B_{max}). Each such pair of selected half chains was subjected to ten steps of energy minimization by using the GROMOS package,[43] not including any nonbonded (van der Waals or electrostatic) interactions. The main chain of the root residues on both sides is also included in the minimization, but

the atoms in these root regions are constrained to remain very close to their starting positions. This procedure produces values of all bond lengths and bond angles within acceptable deviations from the standard values for the whole loop and maintains connectivity to the rest of the protein.

In the second stage of the process, side chains were added to each main-chain conformation in turn. Each side chain was built up in stages, adding first those atoms whose positions are affected only by changes in the first χ angle and not the subsequent ones, then those affected only by the second, and so on. Docking of each section of side chain was performed by overlapping the three atoms preceding it in the standard coordinates with the equivalent atoms already in position. The bookkeeping of the particular atoms involved and the changing of torsion angles was done with code based on the program MUTATE written by Randy Read. As each stage was built, the coordinates of its atoms were checked for clashes with the bulk of the protein and for clashes with main-chain atoms of the segment (the first part of filter c).

Accepted side-chain units were stored in a manner similar to that in which the main-chain conformations were stored. A tree of conformations is generated for each side chain, with as many nodes as side-chain torsion angles, and at each node, up to as many branches as the number of χ values allowed for the corresponding angle.

When all the individual conformations of all the side chains on the current main-chain conformations had been considered, their combinations were formed. It is at this stage that the full effects of the number of combinations possible are felt. The data structure was therefore designed to apply the last filter used during conformation generation efficiently. This is the third part of filter c, checks for clashes between side-chain atoms. In practice, this was found to be a disappointingly weak constraint, so the order of its application is not really important.

Finally, the electrostatic energy and relative hydrophobic area of each conformation were evaluated. This is a relatively computationally intense step in the procedure, partially as a consequence of the inclusion of a solvent model. The basis for the electrostatic energy evaluation used here is the usual sum of pairwise interactions between atoms,

$$U = \Sigma_{ij} q_i q_j / r_{ij}$$

where q_i, q_j are partial atomic charges on the atoms of polar and charged groups, the i set comprises the atoms of the segment, the j set comprises all other atoms (the other atoms of the segment and the atoms of the rest of the protein), and r_{ij} is the distance between atoms i and j. Interactions were included between pairs of electrostatic groups with centers closer than 5 Å. This rather small cutoff distance has proved sufficient for the cases so far examined. Group centers x_j were calculated as

$$x_j = \Sigma_n |q_n| x_n / \Sigma |q_n|$$

where the sums are over the partial charges (q_n) of the atoms composing a group. Groups were defined as collections of atoms with integral or zero total charge.

Partial charges used were those obtained from the fitting of the geometry and sublimation energies of crystalline amides and carboxylic acids[44–46], augmented for the charged groups (A.T. Hagler, personal communication).

Polar hydrogens (those covalently bonded to oxygen and nitrogen atoms) are included explicitly in this version of the force field. Nonpolar hydrogens (those bonded to carbon and sulphur) are not. All carbon and sulphur atoms, other than carbonyl carbons and methylenes adjacent to carboxylate and amino groups, have zero electronic charge. These zero charge atoms form the set considered to be hydrophobic. The relative hydrophobic area of a particular segment conformation is defined as the sum of the solvent accessibilities of this class of atom in the segment, minus the reduction in this sum for atoms on the rest of the protein resulting from the presence of the segment in the current conformation. Solvent accessibilities were calculated by using the method of Lee and Richards.[35]

Solvent effects were treated by using the method of image charges.[38] For every real charge an appropriate image charge is introduced and included in the energy sum above. The role of solvent in the electrostatic properties of proteins may be considered as consisting of two parts.[38,47] One part is a reduction in the effective interaction energy between two groups near to the protein surface. The reduction is brought about by the alignment of water dipoles in such a way as to reduce the net potential at a distance from each group. This is the solvent screening effect. The second part is a favorable interaction between each group near the surface and the surrounding water dipoles, brought about by an alignment of these with the field from the group. This gives rise to the solvation energy of the groups. In the image charge approximation, these two terms are separated. The first term is represented by the interaction of the image charges of each group with the real and image charges of the other groups, and the second by interaction of the real charges with their own images. The screening effect calculated by using image charges is the same to a very good approximation as the work potential calculated by using the Kirkwood Tanford method[48,49] at zero ionic strength. The solvation energy produced by the image charge method is relatively unreliable and is not included here.

Details of the implementation of the image charge algorithm have been published elsewhere.[39] Briefly, the protein is treated as locally spherical and the solvent as a continuous high dielectric medium. The definition of the image charges requires calculation of the depth of the groups interacting below the locally defined protein surface. This surface is defined

in terms of atomic solvent accessibilities. Thus it is necessary to update the surface accessibilities upon which these depths depend for each conformation. Depths of groups below the surface were calculated by the method described.[39] A protein sphere radius of 15 Å and an additional buffer shell of 3-Å thickness were used.

An additional correction term is required to allow for the approximate nature of the coordinates, which can result in charged groups becoming unrealistically close. For the groups with net charge, this leads to a serious overestimate of interaction energies, as a consequence of the 1/r dependence of energy on distance. Interaction energies involving such charges for which the group centres are closer than permitted were rescaled by the ratio of the actual distance to the allowed distance (3.9 Å).

RESULTS

We describe here the application of the algorithm to the determination of the conformation of two of the regions in SGT where the model builders were unsuccessful. The X-ray structure of SGT was used to represent the rest of the protein, providing a more accurate environment than would often be available in practice. However, values of the parameters were chosen suitable for a model with an accuracy of around 1 Å RMS, so that the exercise does simulate a typical model-building situation fairly closely.

Structure of the Modeled Regions

Figure 4 shows the structures in SGT and BT for the segments of chain modeled. Segment 1 is a region where there are 3 residues fewer in SGT than BT. The aligned sequences in this area[7] are

SGT:	ARG	LEU	SER	MET	–	–	–	GLY	CYS
BT:	SER	LEU	ASN	SER	GLY	TYR	HIS	PHE	CYS
	32	33	34	35	38	39	40	41	42

(Sequence numbers are based on an alignment of SGT with α-chymotrypsin[50]. The main chain and side chains of LEU 33, SER 34, and MET 35, together with the preceding ARG 32 side chain, were modeled by the systemic search procedure. The segment contains two side chains (LEU 33 and MET 35) which are considered hydrophobic. The conformational variability of these may therefore be treated separately from the rest of the segment. The extra residues in BT form a bulge out from the surface, while in SGT the chain runs along the protein surface. The leucine side chain fits into a hydrophobic pocket (Fig. 4c), and the side chain of the arginine residue lies in a hydrophobic groove, with the guanidinium group forming favorable electrostatic interactions with a glutamine side chain (GLN 151) and the carbonyl group of GLY 193. The site of this guanidinium group is occupied by the side chain of HIS 40 in BT. Other important features are main-chain to main-chain hydrogen-bond-type favorable electrostatic interactions from SER 34 HN to THR 65 C=O, LEU 33 HN to CYS 42 C=O, and LEU 33 C=O to GLY 41 NH. The loop is well ordered in SGT (B factors in the range 8–18 Å^2) and there are few contacts with neighboring molecules. The model builders[8] positioned the main chain partially in the groove that is occupied by the arginine 32 side chain.

Segment 2 is a 5-residue insertion in SGT relative to BT:

SGT:	ARG	LYS	ASP	ASN	ALA	ASP	GLU	TRP	ILE
BT:	CYS	SER	–	–	–	–	–	GLY	LYS
	201	202	203	204	204A	205	206	207	208

The systematic search was performed on the conformation of the five inserted residues. This segment is highly charged, with three negative charges in the loop itself, interacting with two lysine residues in the rest of the molecule. It is the reverse situation to segment 1 in the sense that the extra residues form a bulge on the surface of SGT whereas in BT there is a single surface residue. For segment 2, both Jurášek et al.[8] and Greer[9] noticed a sequence similarity in SGT with the corresponding region in bovine chymotrypsin, and the X-ray structures for this region were found to be similar. Jurášek et al., however, concluded that this was probably the calcium binding site in SGT, because of the concentration of negative charges. The X-ray structure shows the calcium site to be elsewhere and the negative charges to be stabilized by the surrounding positively charged residues.

The X-ray structure shows a number of electrostatic interactions to be significant in stabilizing the observed conformation. Most important are those involving the charged groups, particularly between the carboxyl group of ASP 203 and the amino group of LYS 122, LYS 202 C=O, ALA 204A NH, and C=O, and GLU 206 NH. The carboxyl group of ASP 205 and the carboxyl group of GLU 206 both make favorable interactions with the amino group of LYS 202. There is one main-chain hydrogen-bond-type interaction within the segment, from ASP 203 NH to GLU 206 C=O. The ϕ, ψ angles of ASP 203, ALA 204A, and GLU 206 are such that the NH and C=O groups within these residues have significant interaction energies. There are unfavorable interactions between the carboxyl group of ASP 203 and GLU 184 C=O and between the carboxyl group of ASP 205 and the NH group of the same residue. B factors are in the range 20–30 Å^2, with high values for the side chains of the adjacent LYS 202, ASN 204, ASP 205, and GLU 206. These values for the temperature factors imply errors in the crystallographic coordinates of 0.15 to 0.2 Å.[34]

Generation of Main-Chain Conformations

The first part of Table II summarizes the behavior of the algorithm for the building of possible main-

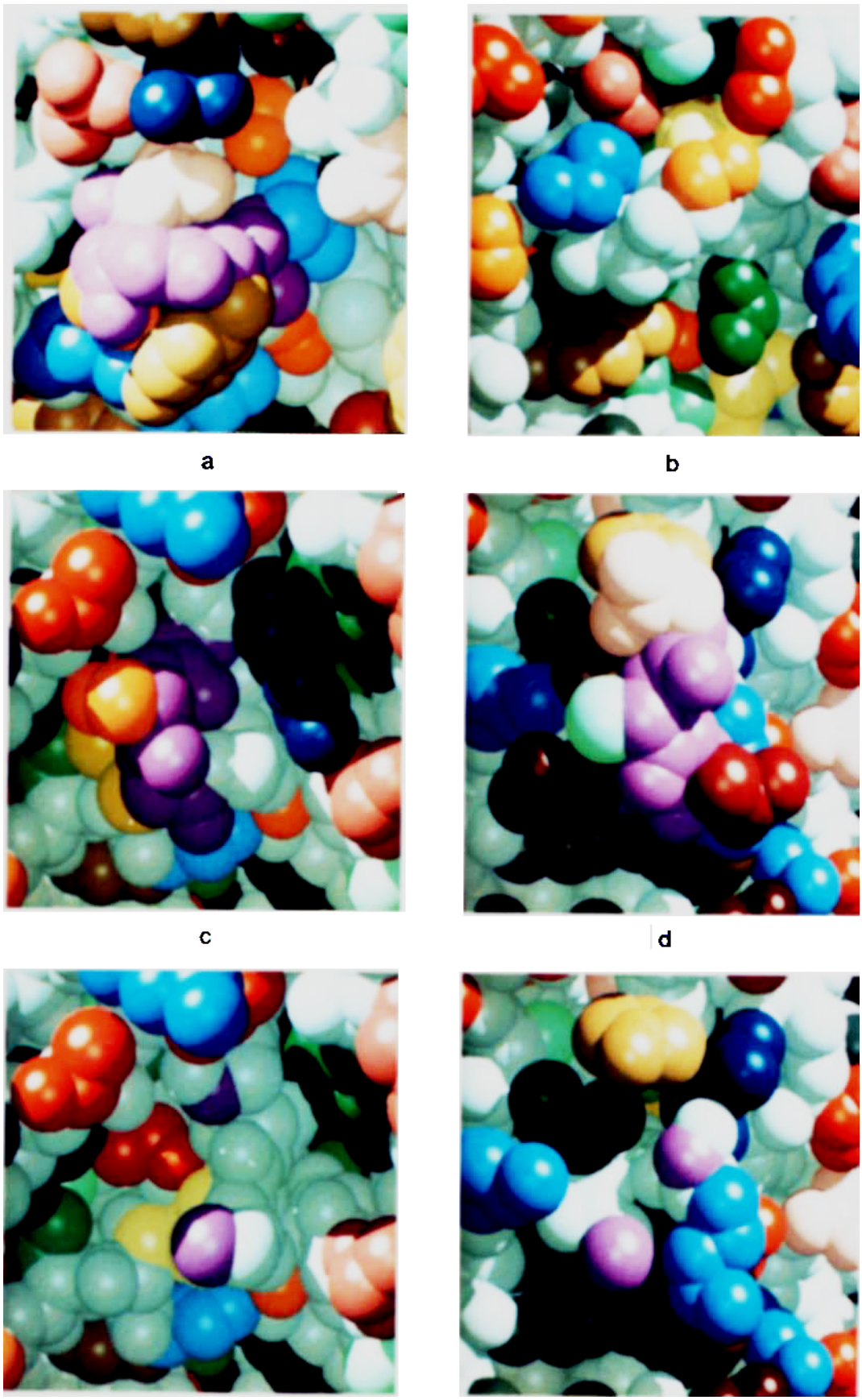


Figure 4.

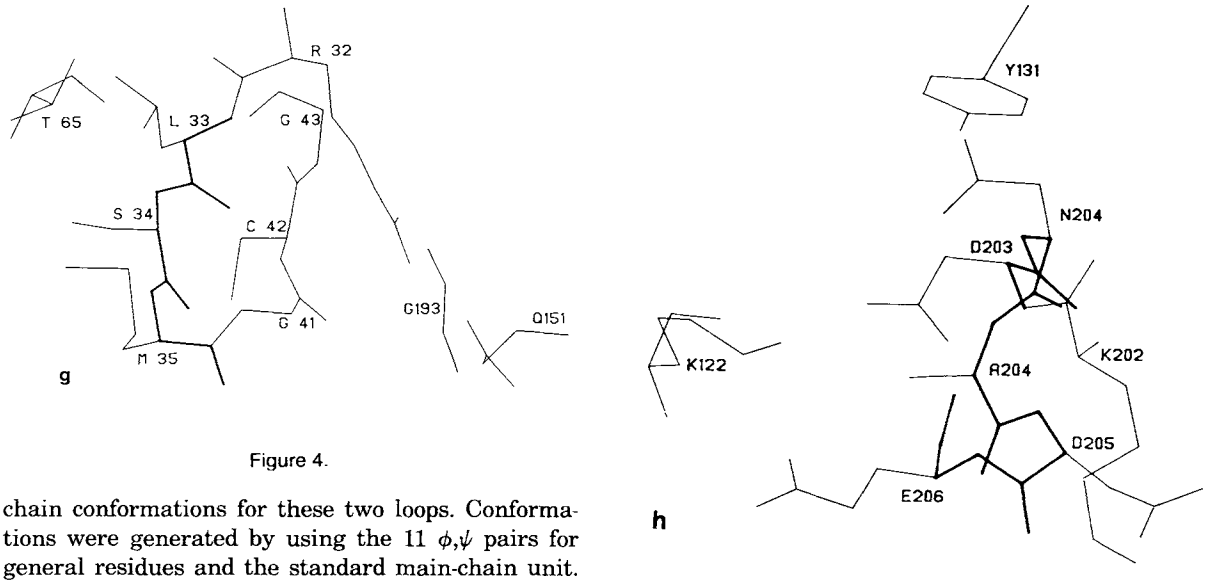


Figure 4.

chain conformations for these two loops. Conformations were generated by using the 11 ϕ,ψ pairs for general residues and the standard main-chain unit. The shorter of the two loops has relatively few possible main-chain conformations when defined in this way (1,331), and the second loop, with 5 residues, has rather more: 1.6×10^5.

For building, segment 1 is split into 1 residue on the N-terminal side and 2 on the C-terminal, and segment 2 into 2 on the N-end residues and 3 on the C-end. In the building process, the number of C ends of these two loops is reduced to around 15% of the total possible by application of the clash filter and termination of growth because of chain direction making loop closure impossible. (The clash filter was used with a tolerance of $\epsilon_{pos}=1.5$ Å for clashes with the rest of the protein). The number of total possible loops at this stage is the product of the two sets of half loops and is still quite large for the second case: 20,790. However, of these possible pairing, only 182 have the end atoms within $D_{cl}=1.6$ Å of each other and Cα-Cα virtual bonds of between $B_{min}=2.8$ Å and $B_{max}=4.8$ Å, the criteria for loop closure used in these runs.

The 17 segment 1 and 182 segment 2 complete main-chain structures were subjected to ten steps of energy minimization in the GROMOS EM program[43] excluding all nonbond (electrostatic and van der Waals) interaction terms. This operation was sufficient to produce acceptable bond lengths and bond angles in all cases. The process tends to make the loops more similar in conformation, and cross comparison of all atoms in the loops showed nine out of the 17 for segment 1 and 84 out of the 182 for segment 2 to be unique. The criterion used to determine whether or not a segment conformation was unique was that at least one atom in it must differ in position by

Fig. 4. Space-filling representation of the loops modeled. These pictures were produced by using the program RASTER3D written by David Bacon. The left column shows segment 1, and the right, segment 2. The top frame in each column (a and b) is the structure in bovine trypsin; the middle frame (c and d), that in SGT; and the bottom frame (e and f), SGT with the modeled region removed. Backbone of the modeled region in SGT and the corresponding residues in BT are shown in purple. In the two frames with the segment atoms removed, the atoms immediately adjacent to the segment main chain region are colored purple. Other backbone atoms are off-white. Aliphatic side chains are green, aromatics brown, polar pink, and orange. Negative-charged residues are red, and positively charged ones are blue. The darker the color, the longer the side chain. The line drawings on the facing page (4g and 4h) provide a key for the color pictures and show the residues of the segments and the residues forming the most significant electrostatic interactions with these. The main chains of the segments are in thick lines and correspond to the purple regions in frames c and d. Segment 1 is a 3-residue deletion in SGT relative to bovine trypsin (BT). The extra residues in BT form a bulge on the surface of the molecule while in SGT the chain runs approximately straight, packing tightly against the underlying surface. The modeled region was composed of the following residues: The side chain of ARG (R) 32, which lies in a groove in the protein surface, with its hydrophobic portion partly packed against a neighboring aliphatic side chain; LEU (L) 33, the side chain of which packs tightly into a pocket in the underlying protein surface: this pocket is also hydrophobic (the orange residue in it is a threonine, with the side-chain methyl group packing against the leucine side chain); SER (S) 33, the side chain of which is solvated; and MET (M) 35, the first two atoms of the side chain of which can just be seen to the left of the main chain at the lower end of the purple stretch. Segment 2 is a 5-residue insertion in SGT relative to BT. The extra residues form a twisted loop with its mean plane approximately perpendicular to the local underlying protein surface. The ends of the loop are close together. Aspartic acid D203 makes a salt bridge with the side chain of lysine K122. The light green single atom in the middle of the purple region is the side-chain of alanine A204A. The second aspartic acid, D205, makes a salt bridge with K202, and the glutamic acid, E206, also makes a rather weak salt bridge with K122. There is some general burying of parts of hydrophobic side chains (Y131 and a valine residue near the top of the pictures, and a tryptophan at the bottom).

δ_{uniq}=0.8 Å from the positions in other conformations.

For segment 1, all the main-chain conformations generated are quite close to the x-ray structure, with RMS differences from this ranging from 0.39 to 0.60 Å. For segment 2, the spread of agreement is much larger, from 0.7 Å to 2.5 Å RMS. The lowest RMS structures are thus sufficiently close to the x-ray structure to support the adequacy of the density of the ϕ-ψ space sampling used in generating the conformations.

Addition of Side Chains

The second part of Table II gives the data for the generation of side-chain conformations onto each of the main-chain possibilities for each segment. The total number of different possible conformations is now very high in both cases. However, the number of main-chain conformations has already been drastically reduced by application of the initial filters. A further reduction now takes place, in that only seven of the nine main-chain conformations for segment 1 and 12 of the 84 for segment 2 can accept any complete sets of side chains at all. This is because for the others, no set of side-chain conformations was found which did not clash with the bulk of the protein or with the main chain of the segment. The number of times each of the three side-chain filters was applied is given in the table. Clashes with the main chain and other side chains were defined as nonhydrogen atom contacts less than D_{con}=2.5 Å, and a clash with the protein was taken to be a nonhydrogen atom positioned near a clash grid point more than ϵ_{pos}=1.1 Å from an acceptable position. A somewhat surprising result here is how little filtering was done by the side-chain to side-chain distance check.

The total number of conformations that could be accepted on the seven main-chain conformations of segment 1 is rather high: 30,198. Partitioning out from these the conformational variability of the two hydrophobic side chains is thus important. There are 393 conformations of the main chain and hydrophilic side chains which will accept at least one conformation of the two hydrophobic side chains. These conformations contain between five and ten conformations of the possible 81 for the side chain of ARG 32, and between six and nine of the nine possible for SER 34. The electrostatic energy calculation was carried out on just this set of 393. The conformation of lowest energy from these could accept 85 conformations of the hydrophobic side chains (five of the possible nine for LEU 33 and 17 of the possible 27 for MET 35), and the relative hydrophobic areas of these were evaluated.

For segment 2, the total number of conformations generated is much smaller at 1,421. There are no hydrophobic side chains, and so only the electrostatic energy needs to be evaluated. Most conformations are found for the longest side chain, that of the glutamic acid, with between 11 and 13 of the possible 18 accepted. Aspartic acid 203 has between one and two accepted conformations and ASP 205 has between two and five; and asparagine 204 has between one and four.

Selection of the Most-Correct Structures

The remaining task is to select from amongst the generated conformations a single structure for each loop that is, ideally, the one closest in structure to the X-ray result. (For these segments, with few interactions with neighboring molecules in the crystal, we take the X-ray structure to represent the structure of the isolated molecule.) For this purpose, the electrostatic interaction energy within each loop conformation and with the rest of the structure was calculated by using the force field and solvent model already described.

Figure 5a shows the distribution of electrostatic energies versus RMS deviation from the X-ray structure for the 393 electrostatically distinct conformations of segment 1. In general, there is an overall correlation between these quantities, and the conformation with lowest energy is very close to being the lowest in RMS: 1.4 Å with 1.3 Å the lowest possible. Figure 5b shows the variation in exposed hydrophobic area for the 85 conformations of the two hydrophobic side chains which were accepted on the conformation with lowest electrostatic energy. Here the conformation with the smallest exposed hydrophobic area is the one with the lowest RMS deviation to the X-ray structure of 0.59 Å. Figure 5c shows the electrostatic energy distribution for segment 2. Again, the lowest energy structure has close to the lowest RMS value: an RMS of 1.25 Å compared with a best possible of 1.17 Å.

DISCUSSION

We have shown that it is possible to generate sets of structures for loops up to 5 residues long in a systematic way, such that one or more of them will be recognizable as being close to the correct conformation. In doing this, utilization of the approximate structure of the protein forming the environment of the loop has been essential in restricting the number of conformations to a manageable number. From other data[51] one expects the conformations of such short lengths of polypeptide chain to be highly dependent on the immediate environment.

It may be argued that a 1 Å RMS deviation from the correct structure is not sufficiently accurate for, say, drug design purposes. This is true, but one would hope that from such a starting point energy minimization or molecular dynamics would lead to a more accurate result. Molecular dynamics with a full solvent representation can produce a protein model with 1-Å RMS from the corresponding X-ray crystal, start-

TABLE II. Generation of Segment Conformations*

Sequence	Seg. 1 (RLSM)†	Seg. 2 (DNADE)†
Main-Chain Construction		
Total possible main chain conformations	1,331	1.6×10^5
Possible N-end main-chain conformations	11	121
Possible C-end main-chain conformations	121	1,331
Clash grid filter applications	0	133
Growth stopped — closure impossible	92	690
Generated N-end conformations	11	110
Generated C-end conformations	18	189
Generated complete main chains	17	182
Unique complete main chains	9	84
Furthest from X-ray (RMS Å)	0.6	2.5
Closest to X-ray (RMS Å)	0.4	0.7
Complete segments		
Total possible conformations	2.4×10^8	6.2×10^8
Clash grid filter applications	93	188
Side chain to main-chain clashes	26	307
Side chain to side-chain clashes	5	55
Main chains accepting side chains	7	12
Total conformations built (hydrophilic)	393	1,421
Total conformations built (hydrophobic)	85	—
Furthest from X-ray (RMS Å)	3.5	3.9
Closest to X-ray (RMS Å)	0.48	1.16

*RMS, root mean square.
†The one letter code is used for these segments: ARG LEU SER MET and ASP ASN ALA ASP GLU.

ing from that structure.[52] Clearly, some improvement in this performance is desirable. However, this is a separate problem from the one of selecting a conformation sufficiently close to the correct one that it can be used as a starting point for a molecular dynamics calculation.

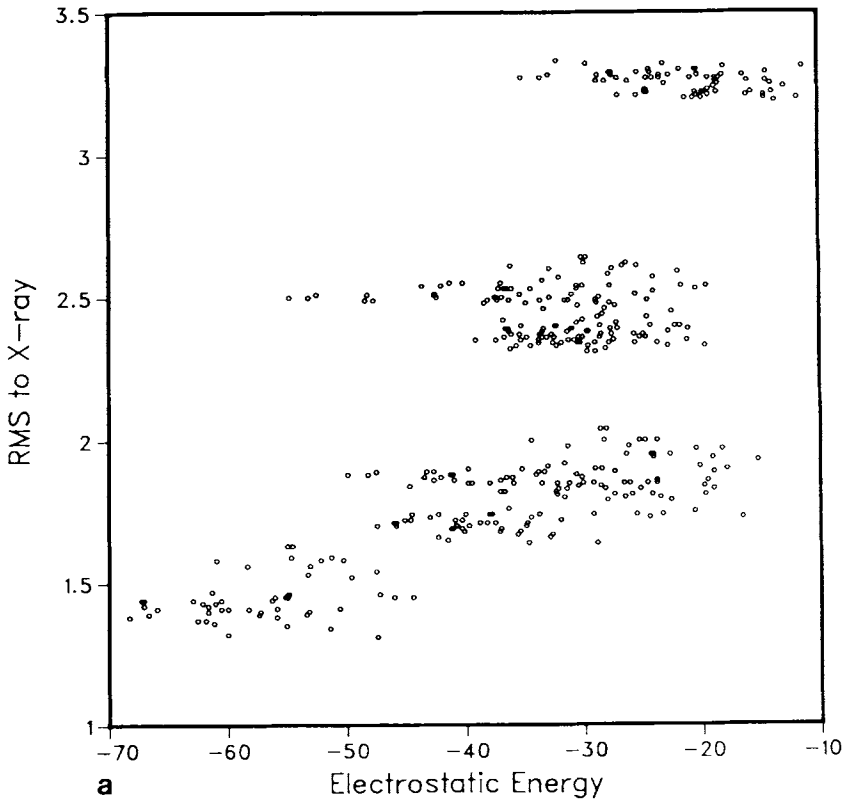
A more serious problem is the possibility of incorrect solutions. One of the great strengths of crystallography compared with computer modeling at its present stage of development is the ability in the former discipline to objectively determine whether a proposed piece of structure is correct or not, by the use of statistical tests against the observed data. No absolutes such as an R-factor or a difference map have been developed for modeling, and it has been established that potential energy alone, excluding solvent effects, is not adequate.[37]

Here we have included a solvent model in addition to the more usual potential energy terms, and discrimination is straightforward, without any attempt to tune the potential. Although a number of cases need to be explored before the generality of this result is established, these results do suggest that the conformations adopted by short pieces of chain have energies very distinct from most of the other possibilities. To some extent this must be the case in order to provide the energy difference needed to freeze out one conformation from many—that is, to compensate for the entropy term representing a large choice of states. But this term is not large enough to require the large spread of energies found here (see for instance reference 53 for a discussion of this factor). The correct structure appears to be highly organized to obtain a better electrostatic energy than can be approached by any of the alternatives. If this is the case, it may be much easier to identify correct structures than thermodynamic arguments would suggest.

The solvent model remains the most unsatisfactory aspect of the potential, however, and it is desirable to use a more realistic treatment than this type of continuous dielectric model. In this respect, the orientable dipole models[54] appear to be the most promising alternatives.

At the current stage of development, the technique may be applied to protein model-building problems involving the structure of up to 7 residues (depending on the number of side-chain degrees of freedom) if the surrounding protein structure is known to about 1- or 1.5- Å accuracy. This fits the requirements of typical homologous modeling situations such as SGT from BT, with more than 30% identity of sequence. It can also be used for modeling the effect of single amino acid substitutions, under the assumption that the large-scale perturbations of structure will be limited to a few residues.

A large number of possible extensions of the technique are now apparent. An alternative procedure for closing such loops is the Go and Scheraga algorithm[55]


3.5
3
2.5
2
1.5
1
RMS to X-ray
−70
−60
−50
−40
−30
−20
−10
a
Electrostatic Energy

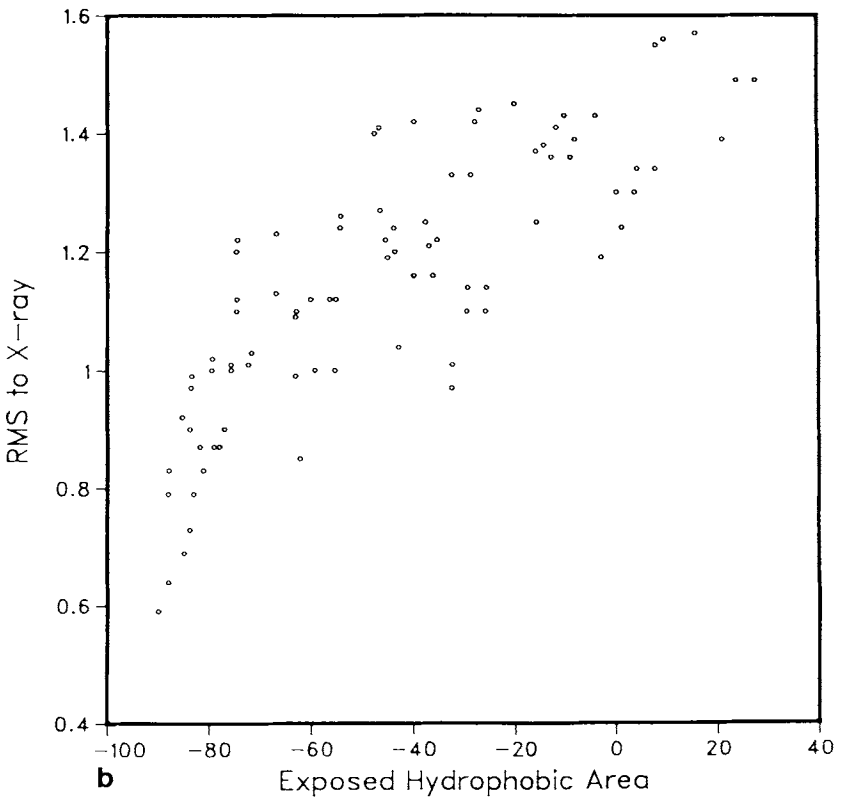

1.6
1.4
1.2
1
0.8
0.6
0.4
RMS to X-ray
−100
−80
−60
−40
−20
0
20
40
b
Exposed Hydrophobic Area

Figure 5.

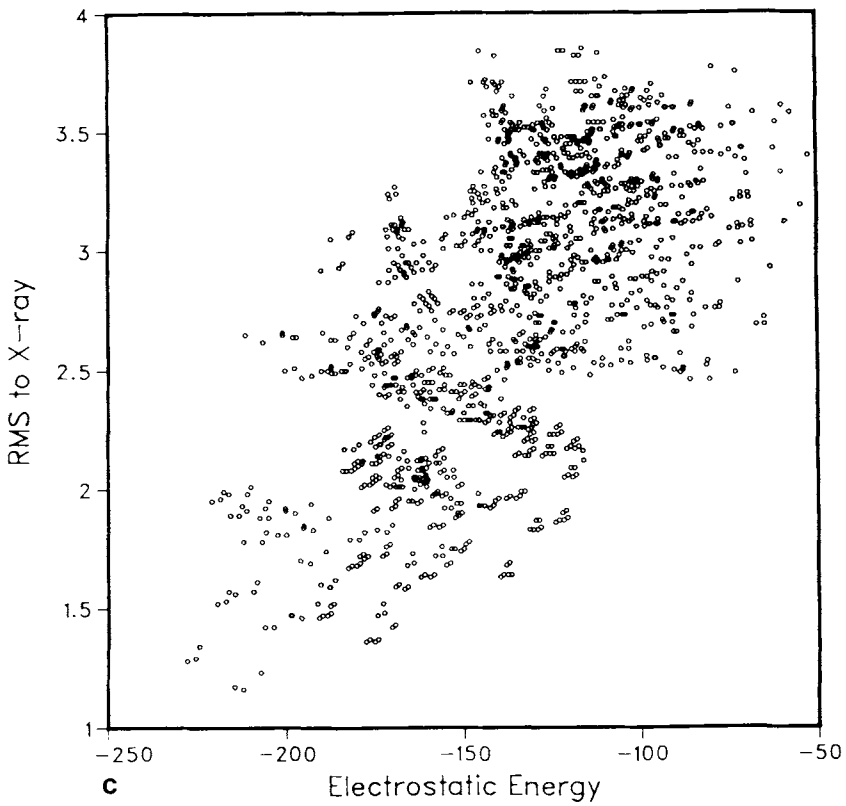


Figure 5.

which finds analytical solutions for the ϕ and ψ angles for a stretch of 3 residues, given the position and orientation of each end. This is an attractive alternative at first sight, since it requires combinations for 3 less residues to be included in the systematic search. However, it is not clear how one would allow adequately for errors that affect the position of the root residues when this procedure is used. In practice, the generation of main-chain conformations is not the limiting factor in the utility of the method, so that greater efficiency in this step is not particularly useful. Additionally, tests of the Go and Scheraga procedure, removing 3 residues at a time from SGT and building them back, failed to find any closure in approximately 10% of the cases (Israel and Moult, unpublished). It seems that restriction to rigid geometry combined with an exact solution requirement is too demanding. However, this type of procedure may be useful in another way here. The present sampling of ϕ-ψ space used in generating the main-chain conformations results in the possible accumulation of significant errors when more than 3 residues are built onto either half of the loop. Ignoring the other degrees of main-chain freedom, particularly bond angles and the distortion of the peptide planes, also causes significant errors over longer stretches. It is desirable to escape from an angle description altogether, since what is ideally required is a systematic sampling of

Fig. 5. Scatter plots showing the relationship of electrostatic energy and exposed hydrophobic area to the root mean square (RMS) deviation from the X-ray structure for the generated segment conformations. Each point represents one conformation. Energies are in kcal/mol, areas in $Å^2$, and RMS values are in Å and are for all nonhydrogen atoms of both main chain and side chains. a. Segment 1. Electrostatically distinct conformations. (The conformations of the side chain of ARG 32, main chain of LEU 33, all of SER 34, and main chain of MET 35 are varied;) 393 conformations are shown, built up from seven different main-chain conformations. For each conformation, a single conformation of the side chains of LEU 33 and MET 35 was included. There is a clear overall correlation between RMS and energy, and the lowest energy structure has almost the best agreement with the X-ray result. b. Segment 1. Hydrophobic side-chain conformations. (The conformation of the side chains of LEU 33 and MET 35 are varied on the lowest energy conformation from a). The relative hydrophobic area for each of the 85 conformations was calculated as the sum of the solvent accessible areas of the hydrophobic atoms of the segment, minus the sum of the loss of these areas on the rest of the protein caused by the presence of the segment. Areas are defined in the manner of Lee and Richards.[35] The conformation with the smallest hydrophobic surface area is the one with the smallest RMS deviation from the X-ray structure. c. Segment 2. (residues ASP 203, ASN 204, ALA 204A, ASP 205, and GLU 206); 1,421 conformations are shown, built up from 12 different main-chain conformations. Again, there is a general correlation between energy and RMS, and the lowest energy structure has almost the best agreement with the X-ray result.

Cartesian space. This could be achieved by dividing the space that a segment may occupy into cubes and considering the possible main-chain constructions that could lead to occupation of each of these by particular atoms. A modification of the Go and Scheraga algorithm[55] solving for the values of three dihedral angles that cause the chain to pass close to a particular point, would be suitable for this.

Of the possible rules which may be used to devise filters, we have relied heavily on the one which rejects possible conformations on the basis of van der Waals clashes. There is a good reason for this: it is a rule which has close to 100% applicability because of the very high energy cost implied by its violation. In the present examples, we have not applied this filter very carefully, using a universal atom size to determine allowed conformations either directly or indirectly. The spread of atomic radii[42] is enough to have a significant impact on the number of allowed conformations, even in the presence of errors from the rest of the structure, so that some tightening of the application of the filter would be helpful.

The next easiest filter to implement is an electrostatic one, which utilizes the principle that there are few bad electrostatic interactions in proteins. Its application does require some care. For example, there are sometimes bad dipole-dipole interactions which are compensated by stronger favorable dipole-charge interactions. These can be anticipated though. The large spread of energies for the generated conformations (Fig. 5) suggests that such a filter ought to be very effective.

There is a sense in which this algorithm may be regarded as an expert system for the prediction of protein conformation. Rules are identified which known protein structures obey, and these rules are incorporated in a knowledge base which is operated on by an inference engine to generate possible structures. This type of approach has been successful in a number of areas (see for instance reference 56). Although we have found this viewpoint useful in considering generalizations of the present algorithm there are a number of respects in which it is not an accurate description of the procedures used here. At the level of algorithm implementation, this one obeys almost none of the architectural rules of expert systems, and the knowledge base used here is very small. There are less than ten rules utilized, whereas thousands are more typical.

One of the principles of expert systems is very relevant, however, and that is that one should be able to interrogate the system as to the basis for its conclusions. We have presented the tests as blind, in the sense that one simply reads in the coordinates, presses the button, and accepts the program's conclusion as to which is the best structure. A good deal of monitoring information is provided on the output, which allows one to see why any particular conformation was rejected. This could be made more intelligible by displaying in real time on a graphics system the growing segments and by highlighting filtering interactions. In this manner the program may be made interactive, stopping at some point and resetting a parameter to see how it affects the outcome. This is important in the extension to larger simulated regions, since in order to restrict the number of structures it may be necessary to be unrealistic about the accuracy of the rest of the coordinates. If it is then seen that a particular clash is responsible for eliminating an otherwise promising developing conformation, this could be relaxed. Such use is, however, a departure from the spirit of the work, which is to aim at making a fully reliable conformation predictor.

It is hard to estimate at this stage what the ultimate limitations on this type of approach may be. The number of conformations which can be generated increases dramatically with increasing chain size, but on the other hand new filters can have an equally dramatic effect on reducing the number propagated. We are optimistic that we are only at the beginning of its development.

ACKNOWLEDGMENTS

We thank Randy Read and David Bacon for many stimulating discussions and for the use of the programs MUTATE and RASTER3D; Wilfred van Gunsteren and Herman Berendsen for the provision of the GROMOS package, and the University of Alberta Computer Center for computational facilities. Some of the impetus for the project came from a CECAM discussion meeting on drug and vaccine design. This work was supported by the Canadian Medical Research Council and by an Alberta Heritage Fund for Medical Research Visiting Scientist award to J.M.

REFERENCES

1. Hartley, B.S. Homologies in serine proteases. Philos Trans. R. Soc. Lond. [Biol.]257:77–87, 1970.
2. McLachlan, A.D., Shotton, D.M. Structural similarities between alpha-lytic protease of myxobacter-495 and elastase. Nature 229:202–205, 1971.
3. Warme, P.K., Momany, F.A., Rumball, S.V., Tuttle, R.W., Scheraga, H.A. Computation of structures of homologous proteins. Biochemistry 13:768–782, 1974.
4. Furie, B., Bing, D.H., Feldmann, R.J., Robison, D.J., Burnier, J.P., Furie, B.C. Computer-generated models of blood coagulation factor Xa, factor IXa, and thrombin based upon structural homology with other serine proteases. J. Biol. Chem. 257:3875–3882, 1982.
5. Blundell, T., Sibanda, B.L., Pearl, L. 3-dimensional structure, specificity and catalytic mechanism of renin. Nature 304:273–275, 1983.
6. Delbaere, L.T.J., Brayer, G.D., James, M.N.G. Comparison of the predicted model of alpha-lytic protease with the x-ray structure. Nature 279:165–168, 1971.
7. Read, R.J., Brayer, G.D., Jurášek, L., James, M.N.G. Critical evaluation of comparative model building of *Streptomyces griseus* trypsin. Biochemistry 23:6570–6575, 1984.
8. Jurášek, L., Olafson, R.W., Johnson, P., Smillie, L.B. Relationships between the structures and activities of some microbial serine proteases. I. Purification, enzymic properties and primary structures of *Streptomyces griseus* proteases A, B and trypsin. Miami Winter Symp. 11:93–123, 1976.
9. Greer, J. Comparative model building of the mammalian serine proteases. J. Mol. Biol. 153:1027–1042, 1981.

10. Read, J.R., X-ray Crystallography of Serine Proteases. Ph.D. Thesis, Univ. Alberta, 1986.
11. Chothia, C., Lesk, A.M. The relationship between the divergence of sequence and structure in proteins. EMBO J. 5:819–822, 1986.
12. Sibanda, B.L., Thornton, J.M. Beta-hairpin families in globular proteins. Nature 316:170–174, 1985.
13. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5:823–826, 1986.
14. Robson, B. The prediction of molecular conformation. Biochem. Soc. Trans. 10:297–298, 1982.
15. Levitt, M. Protein folding by restrained energy minimization. J. Mol. Biol. 170:723–764, 1983.
16. Harvel, T.F., Kuntz, I.D., Crippen, G.M. The theory and practice of distance geometry. Bull. Math. Biol. 45:665–720, 1983.
17. O'neil, K.T., DeGrado, W.F. A predicted structure for calmodulin suggests an electrostatic basis for its function. Proc. Natl. Acad. Sci. U.S.A. 82:4954–4958, 1985.
18. Madison, V. Cyclic peptides revisited. Biopolymers 24:97–103, 1985.
19. Vásquez, M., Scheraga, H.A. Use of buildup and energy minimization procedures to compute low energy structures of the backbone of enkephalin. Biopolymers 24:1437–1447, 1985.
20. Levinthal, C. Are there pathways in protein folding? J. Chim. Phys. 65:44–45, 1968.
21. Reingold, E.M., Nievergelt, J., Deo, N. Exhaustive search. In: Combinatorial Algorithms: Theory and Practice. New Jersey: Prentice Hall 1977:134–148.
22. Robson, B. Analysis of the code relating sequence to conformation in globular proteins—Theory and application of expected information. Biochem. J. 141:853–867, 1974.
23. Cohen, F.E., Sternberg, M.J., Taylor, W.R. Analysis and prediction of protein beta-sheet structures by a combinatorial approach. Nature 285:378–382, 1980.
24. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838, 1985.
25. Richards, F.M. The interpretation of protein structures: Total volume, group volume distributions and packing density. J. Mol. Biol. 82:1–14, 1974.
26. Chothia, C. Structural invariants in protein folding. Nature 254:304–308, 1975.
27. Burgess, A.W., Ponnuswamy, P.K., Scheraga, H.A. Conformation of amino acid residues and prediction of backbone topography in proteins. Isr. J. Chem. 12:239–286, 1974.
28. Janin, J., Wodak, S., Levitt, M., Maigret, B. Conformation of amino acid side-chains in proteins. J. Mol. Biol. 125:357–386, 1978.
29. Ramakrishnan, C., Ramachandran, G.N. Stereochemical criteria for polypeptide and protein conformations: 2. Allowed conformations for a pair of peptide units. Biophys. J. 5:909–933, 1965.
30. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. Computer based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
31. Baker, E.N., Dodson, E.J. Crystallographic refinement of the structure of actinidin at 1.7 Å resolution by fast fourier least squares methods. Acta Cryst. A36:559–572, 1980.
32. Baker, E.N. Structure of actinidin after refinement at 1.7 Å resolution. J. Mol. Biol. 141:441–484, 1980.
33. Chambers, J.L., Stroud, R.M. Accuracy of refined protein structures—Comparison of 2 independently refined models of bovine trypsin. Acta Cryst. B35:1861–1874, 1979.
34. Read, R.J., Fujinaga, M., Sielecki, A.R., James, M.N.G. Structure of the complex of *Streptomyces griseus* protease B and the third domain of the turkey ovomucoid inhibitor at 1.8 Å resolution. Biochemistry 22:4420–4433, 1983.
35. Lee, B., Richards, F.M. Interpretation of protein structure-Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
36. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Protein Chem. 14:1–63, 1959.
37. Novotný, J., Bruccoleri, R., Karplus, M. An analysis of incorrectly folded models—Implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.
38. Friedman, H.L. Image approximation to the reaction field. Mol. Phys. 29:1533–1543, 1975.
39. Moult, J., Sussman, F., James, M.N.G. Electron density calculations as an extension of protein structure refinement-*Streptomyces griseus* protease A at 1.5 Å resolution. J. Mol. Biol. 182:555–566, 1985.
40. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.
41. Sielecki, A.R., Hendrickson, W.A., Broughton, C.G., Delbaere, L.T.J., Brayer, G.D., James, M.N.G. Protein structure refinement: *Streptomyces griseus* serine protease A at 1.8 Å resolution. J. Mol. Biol. 134:781–804, 1979.
42. Chothia, C. Nature of accessible and buried surfaces in proteins. J. Mol. Biol. 105:1–14, 1976.
43. GROMOS is distributed by W.F. van Gunsteren and H.J.C. Berendsen, Laboratory of Physical Chemistry, Univ. Groningen, Nijenborgh 16, 9747 AG Groningen, the Netherlands.
44. Hagler, A.T., Huler, E.H., Lifson, S. Energy functions for peptides and proteins—I. Derivation of a consistent force field including the hydrogen bond from amide crystals. J. Am. Chem. Soc. 96:5319–5327, 1974.
45. Lifson, S., Hagler, A.T., Dauber, P. Consistent force field studies of intermolecular forces in hydrogen bonded crystals—I. Carboxylic acids, amides and the C=O. . .H-hydrogen bonds. J. Am. Chem. Soc. 101:5111–5121, 1979.
46. Hagler, A.T., Lifson, S., Dauber, P. Consistent force field studies of intermolecular forces in hydrogen bonded crystals—II. A benchmark for the objective comparison of alternative force fields. J. Am. Chem. Soc. 101:5122–5130, 1979.
47. Gilson, M.K., Rashin, A., Fine, R., Honig, B. On the calculation of electrostatic interactions in proteins. J. Mol. Biol. 183:503–516, 1985.
48. Tanford, C., Kirkwood, J.G. Theory of protein titration—I. General equations for impenetrable spheres. J. Am. Chem. Soc. 20:5333–5339, 1957.
49. Matthew, J.B., Hanania, G.I.H., Gurd, F.R.N. Electrostatic effects in hemoglobin—Hydrogen ion equilibria in human deoxyhemoglobin and oxyhemoglobin. Biochemistry 18:1919–1928, 1979.
50. Birktoff, J.J., Blow, D.M. Structure of crystalline alpha-chymotrypsin- 5. Atomic structure of tosyl alpha-chymotrypsin at 2 Å resolution. J. Mol. Biol. 68:187–240, 1972.
51. Kabsch, W., Sander, C. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. U.S.A. 81:1075–1078, 1984.
52. van Gunsteren, W.F., Berendsen, H.J.C., Hermans, J., Hol, W.G.J., Postma, J.P.M. Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. Proc. Natl. Acad. Sci. U.S.A. 80:4315–4319, 1983.
53. Schulz, G.E., Schirmer, R.H. Thermodynamics and kinetics of polypeptide chain folding. In: Cantor, C.R. (ed), "Principles of Protein Structure." Heidelberg, New York: Springer-Verlag. 1979:150.
54. Warshel, A. Calculations of chemical processes in solution. J. Phys. Chem. 83:1640–1652, 1978.
55. Gō, N., Scheraga, H.A. Ring closure and local conformational deformations of chain molecules. Macromolecules 3:178–187, 1970.
56. Barr, A., Feigenbaum, E.A. (eds.): "The Handbook of Artificial Intelligence, Vol. II." Stanford, California: Heuristech Press, 1982.