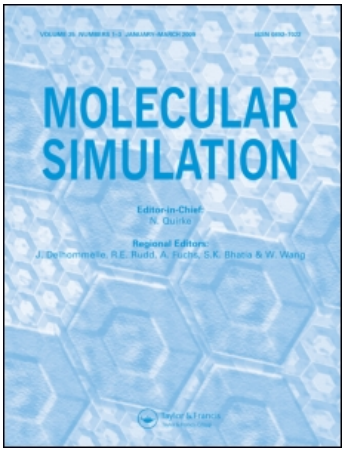

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Application of Systematic Conformational Search to Protein Modeling

Robert E. Bruccoleri[a]
[a] Department of Macromoleculer Modeling, Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, NJ, USA

To cite this Article Bruccoleri, Robert E.(1993) 'Application of Systematic Conformational Search to Protein Modeling', Molecular Simulation, 10: 2, 151 — 174
To link to this Article: DOI: 10.1080/08927029308022163
URL: http://dx.doi.org/10.1080/08927029308022163

Molecular Simulation, 1993, Vol. 10(2-6), pp. 151-174
Reprints available directly from the publisher
Photocopying permitted by license only

APPLICATION OF SYSTEMATIC CONFORMATIONAL SEARCH TO PROTEIN MODELING

ROBERT E. BRUCCOLERI

Department of Macromoleculer Modeling, Bristol-Myers Squibb Pharmaceutical Research Institute, P.O. Box 4000, Princeton, NJ 08543 USA

(Received September 1992, accepted December 1992)

Systematic conformational search is a powerful tool in the modeling of proteins and peptides. As a deterministic method for sampling comformational space, it provides an efficient mechanism for finding global energy minima. The program CONGEN has been developed to use conformational search in conjunction with other modeling methods. The search operators in CONGEN can be combined in arbitrary ways, and therefore, they can applied to a wide variety of problems. Typical applications include homology modeling, construction of protein coordinates from C_α positions, sidechain placement, peptide structure modeling, derivation of three-dimensional structure from NMR constraints, etc. In this paper, a detailed description will be provided of its conformational search capabilities and its previous applications to protein modeling.

KEY WORDS: Conformational search, protein modeling, homology modeling, CONGEN, prediction of protein folding

1 INTRODUCTION

The prediction of protein structure from the amino acid sequence remains one of the most significant unsolved problems in biochemistry [1]. However, it is a very difficult problem for primarily two reasons. First, the conformational space of proteins is very large [2-4], and second, it is not known how to accurately calculate the free energy of proteins in solution.

The approach that has been taken with CONGEN attempts to deal with both of these problems. Rather than try to work with an entire protein, efforts have been focused on the prediction of the structure of short segments (loops) within proteins. Such segments have the advantage that their conformational space is very restricted, and therefore, one can sample their entire conformational space in a reasonable amount of computer time. The availability of a complete sampling provides a powerful test of any free energy function that might be devised, since the true minimum energy of the conformational space can be evaluated. Comparison of the minimum energy conformer against an experimentally determined structure represents the ultimate test of a structure prediction method.

Systematic search has been chosen over common methods of traversing conformational space for a number of reasons. First, it is very efficient. Methods such as dynamics or Monte Carlo operate by sequentially perturbing one conformation into another. The distance between successive conformers is generally very small.

Thus, many steps are required to move between local minima in the energy surface. On the other hand, conformational search can be programmed to generate local minima every time. Thus, the number of energy evaluations, which are the most time consuming operations, can be reduced. In addition, conformational search does not require the calculation of the derivative of the energy. This represents a significant savings over molecular dynamics, and also permits the exploration of energy surfaces which are not continuous.

Second, systematic search is deterministic. When the algorithm completes, the space has been sampled. On the other hand, searches involving random processes must be run until no new conformations are found. Inevitably, this requires that many conformations will be repeatedly sampled before closure is achieved.

Third, systematic conformational search provides control over the search process. There are numerous options controlling how finely the degrees of freedom are sampled. The order of sampling may be specified as well as which conformations are examined first. In cases where the search is too long to completed in a reasonable amount of computer time, it is possible to order the search process to find the better conformations earlier.

Fourth, compared to knowledge based modeling [5–8], systematic search is not dependent on existing databases of structure. Although the use of fragments taken from known protein structure can yield accurate predictions [9], such methods will fail if there are no homologous structures available in the database. In addition, there are examples where identical sequences have very different structures in different proteins [10]. Finally, we will gain more insights into the energetics of protein folding if we can predict structures *ab initio*.

With regard to the calculation of the free energy, empirical free energy methods [11, 12] are being most heavily explored. However, in the work described in this paper, the available computer resources have been limited, and therefore, the CHARMM potential energy function [13] has been used as the energy function. For the antibody and T-cell receptor modeling, a surface area test has been added which improves the prediction by incorporating packing considerations.

The division of the search problem into the generation of conformers and their energetic evaluation implies that progress can be made on each aspect of the problem separately. Most of the previous effort on CONGEN has been on the generation of structures, but improvements in the energy calculations are underway.

In order to use conformational search effectively to model proteins, it is essential to have the capability to construct, manipulate, and analyze the molecule [13], all as part of the same system which perfoms the conformational search. Although many conformational search algorithms have been described [14–22], CONGEN integrates a number of methods together so that they can be applied easily in any way that is suitable for the problem at hand.

In this paper, the conformational search capabilities in CONGEN will first be described in detail, and then a number of modeling problems win be described for which conformational search has provided some insights, and for which there is other data to compare with the modeling.

2 CONFORMATIONAL SEARCH IN CONGEN

In its most general form, a conformational search is just a set of nested iterations of the degrees of freedom in the system. In the initial implementation of CONGEN

[23], the degrees of freedom were encoded directly into the program, and were capable of generating conformations for a single loop. It was clear from this prototype that the necessary operations inherent in such a search could be generalized to quasi-independent operators. This operators could then be combined in any reasonable way, and as a result, a great variety of searches could be performed.

2.1 *General Principles of the Conformational Search of Loops*

The fundamental problem of generating loop conformations is finding a set of atomic coordinates for the backbone and sidechains that satisfy all its stereochemical and steric constraints. For the sake of efficiency, it is presumed that bond lengths and bond angles are fixed [24], and in addition, it is assumed that the peptide ω torsion angle is also planar. Under these assumptions the only degrees of freedom in the loop are the torsion angles.

Given the chemical structure of proteins, the search process is divided into backbone and sidechain constructions. The backbone conformational space is normally sampled before the sidechains because the chain closure condition is very restrictive. As a result, fewer samples are generated early in the process, which helps to reduce the necessary computer time.

2.1.1 *Backbone construction*

The generation of backbone coordinates depends heavily on the modified Gō and Scheraga chain closure algorithm [19, 25]. The algorithm is designed to calculate local deformations of a polymer chain, i.e., finding all possible arrangements of a polymer anchored at two fixed endpoints. Given stereochemical parameters for the construction of the polymer, and six adjustable torsion angles between the two fixed points; this algorithm will calculate values for the six torsion angles in order to perfectly connect the polymer from one endpoint to the other. In the sampling of the backbone, the use of a planar ω torsion angle reduces the number of free backbone torsion angles per residue to two, and therefore, three residues are required for the application of the Gō and Scheraga algorithm. For generating conformations of loops with more than three residues, the backbone torsion angles of all but three residues are sampled, and the Gō and Scheraga procedure is used to close the backbone.

Bruccoleri and Karplus (1987) modified the Gō and Scheraga algorithm to allow small changes in the peptide bond angles [25]. After the first implementation of the algorithm, it was tested by deleting three residue segments in several different proteins, and calling the algorithm to reconstruct the peptide backbone. In the helices in flavodoxin [26], the algorithm failed to reconstruct a large number of these segments. An attempt was then made to reconstruct ideal α-helices where individual bond angles were perturbed by a few degrees. Many of these trials failed. Thus, it was clear that the normal variations in bond angles due to the limits of crystallographic resolution were interfering with the algorithm. By allowing the algorithm to adjust the bond angles by small amounts, typically no more than 5°, it was possible to find torsion angles that would close all three residue segments in the helices [25].

The free sampling of backbone torsion angles is done with the aid of a backbone energy map. Bruccoleri and Karplus (1987) calculated the energetics of constructing the backbone for three different classes of amino acids; glycine, proline, and all the rest. This information is stored as a map [27] which gives the energy as a

function of discrete values ϕ, ψ, and ω, where ω can only be 0° (*cis*) or 180° (*trans*). A set of maps corresponding to grids of 60°, 30°, 15°, 10°, and 5° have been calculated; typically, a 30° sampling is sufficiently fine for good agreement.

With regard to the peptide ω angle, only the proline ω angle is normally allowed to sample *cis* values. However, CONGEN can be directed to sample *cis* ω angles for all amino acids.

The ring in proline creates special problems. The proline ring constrains the ϕ torsion to be close to −65 degrees; any deviation from −65 degrees distorts the ring. The minimum energy configuration of the proline ring (specifically, 1,2 dimethyl pyrrolidine) has been determined for a range of ϕ angles (±90 degrees) about −65 degrees using energy minimization with a constraint on ϕ, and a file has been constructed which contains these energies and the construction parameters necessary to calculate the position of C_β, C_γ and C_δ of the proline. All of these energies are adjusted relative to a minimum ring energy equal to zero. After a chain closure is performed, any conformations which have a proline ϕ angle whose energy exceeds the minimum energy by more than the parameter, ERINGPRO, are discarded. Generally, a large value for ERINGPRO is used (209.2 kJ/mole or 50 kcal/mole) so that the chain closure algorithm does not overly restrict proline closures. The *cis-trans* peptide isomerization is handled by trying all possible combinations of *cis* and *trans* configurations. The user has complete control over which residues can be built in the *cis* isomer. Since there are only three residues involved in the chain closure, this results in no more than eight (2^3) attempts at chain closure.

There are two optimizations performed during the sampling of backbone torsions. First, whenever any atom is constructed, a check is made to see if the atom overlaps with the van der Waals radius of any other atom in the system. If so, that conformation is discarded. Second, as backbone residues are generated, CONGEN calculates the distance from the growing end back to the other fixed point. If that distance is greater than can be reached by fully extended backbone, then those conformations are discarded.

The backbone can be constructed either forward from the N-terminus or backward from the C-terminus until only three residues remain. The N-terminus of the internal segment is anchored on the peptide nitrogen; the C-terminus is anchored on C_α. When the construction direction is from the N terminus to the C terminus, the first torsion to be sampled in a residue is the ω angle (which normally is sampled just at 180 degrees, and can be sampled at 0 degrees for prolines or, as an option, all the amino acids). It determines the C_α and the peptide hydrogen positions. The ϕ angle determines the position of the carbonyl carbon and the beta carbon of the sidechain; and finally, the ψ angle determines the carbonyl oxygen and peptide nitrogen of the next residue. When the construction is in the reverse direction; the ψ angle determines the peptide nitrogen; the ϕ angle determines the carbonyl carbon of the preceding residue, the peptide hydrogen, and the beta carbon; and the ω angle determines the position of the preceding residue's C_α and carbonyl oxygen.

2.1.2 *Sidechain Construction*

Given a set of backbone conformations, it remains to generate a set of side chain atom positions for each of the backbone conformations. This problem is divided into two parts, construction of individual sidechains and combining results from individual sidechains for all the residues.

As with the backbone atom placement, the atoms of a sidechain are positioned based on free torsion angles. The side chain torsions are processed from the backbone out as each succeeding atom requires the position of the previous atom for its placement. The sampling interval of each torsion can be either some fixed number of degrees or the period of the torsion energy. If the latter is used, and the parameters for the torsion energy specify only a single term in the Fourier series for the torsion energy, then the sidechain torsion energy is always zero.

It is common for one free torsion to generate the position of more than one atom because of side chain branching, non-rotatable bonds, and rings. For example, although an explicit hydrogen [13] tryptophan has 11 side chain atoms to be placed, it has only two free torsion angles. Also, some side chain branching is symmetric, e.g. phenylalanine, and CONGEN can use such symmetry to reduce the sampling necessary.

As with the backbone construction, a search of the surrounding space is made for any constructed atom to see if there are any close contacts. However, with the sidechains, there are two ways of checking for such overlaps. The first method is very simple: given the sampling of the torsion angles, each atom is constructed and checked for contacts.

The second method, van der Waals avoidance, is more time consuming, but it yields better quality structures. It is a straightforward geometrical problem to determine the range of torsion angles which will avoid constructing an atom within a given distance of other atoms in the system. As a side chain torsion angle, χ_i, varies, it specifies a circular locus of points on which atoms can be constructed (Figure 1) [23]. If atoms in the vicinity of this circle are examined, the sectors of the circle which will result in the repulsive overlap of the constructed atom with its spatial neighbors can be calculated. The complement of these sectors can be used to determine values for the χ_i angles which avoid bad contacts.

The information needed for side chain construction is stored in a side chain topology file. It is a straightforward matter to add new amino acids to this file so that the structure of unnatural amino acids can be predicted.

Given this method for constructing individual sidechains, it remains to combine sidechain conformations for all the sidechains attached to a particular backbone conformer.

Since the backbone construction process provides the position of C_β, there is a strong bias to the side chain orientation. Thus, an acceptable course of action is the generation of only one sidechain conformation for each backbone conformation. A substantial effort must be made to ensure that this one conformation the lowest energy possible for the given backbone. Second, because the side chains close together in sequence frequently are not close together in space, and therefore, do not interact strongly, it is a reasonable approximation to treat the side chains quasi-independently. Instead of finding all combinations of side chain atomic positions, the side chains can be processed sequentially so the time required for side chain placement increases linearly, rather than exponentially, with the number of residues.

In order not to limit the options for using the program, five possible methods for generating side chain positions have been implemented. Some of the methods can generate only one side chain conformer; others can generate many. The first two methods described, ALL and FIRST, assume no quasi-independence of the sidechains whereas the others do.

The first method, named ALL, generates all possible conformations by a series

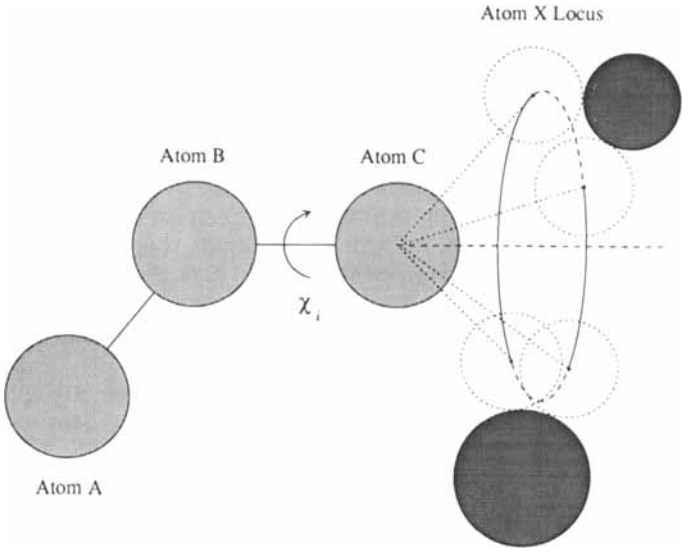

Figure 1 An illustration of van der Waals avoidance. The construction of atom X is based on the positions of atoms A, B, and C; the C-X bond distance, and B-C-X bond angle. Depending on the value of χ_i, the center of atom X can be located anywhere on the circle illustrated as the 'Atom X Locus'. Neighboring atoms, shown in dark gray, will block out parts of the circle, shown in dashed lines. The remaining part of the locus identifies the values of the torsion angle for which there are no van der Waals overlap. For illustrative purposes, the radii of the atoms in this figure are much smaller than actual values.

of nested iterations over every sidechain as described above. The second method, named FIRST, uses the same algorithm as ALL except that all the iterations terminate when the first conformation for all the sidechains has been found. This method is useful for determining if a backbone conformation will accommodate the sidechains when details about the sidechain energetics are not required.

The three other methods all depend on a function which evaluates the side chain positions as they are generated so that the best ones can be selected. "Best" is defined as the conformation whose evaluation function is numerically smallest. Two evaluation functions are currently provided, one based on positional deviations, and one based on the CHARMM energy function [13]. The evaluation function based on positional deviations is present for testing CONGEN as it provides a means for determining the limit of CONGEN's ability to generate a known structure. If coordinates are available for the sidechain atoms, this evaluation function will determine the RMS shift between a generated side chain conformation and the initial coordinates. The second evaluation function computes the CHARMM energy of the sidechain atoms.

In the first of these other methods, named INDEPENDENT, each side chain is placed independently, with atoms of the other side chains in the peptide being ignored; interactions with all other atoms in the system are included. The conformation which has the lowest value for the evaluation function is selected for each side chain. When the RMS evaluation function is used, this method gives the optimum conformation, though it may be sterically inappropriate. However, it cannot be used

when the energy is the evaluation function unless the possibility of large repulsive van der Waals is not important.

The COMBINATION method begins by generating a small number of the best side chain conformations for each side chain independently, as above. Then, these side chain conformations are assembled in all possible combinations, and those combinations which do not have bad van der Waals contacts are accepted. The number of conformations saved for each side chain must be small to avoid a combinatorial explosion.

The ITERATIVE method starts with an energetically acceptable side chain conformation for all the side chains. This conformation is generated, if possible, using the FIRST method. Starting with this conformation, all the possible positions for the side chain atoms of the first residue are recalculated, and the conformation with the lowest energy is selected. The value of the evaluation function is also saved. This regeneration is done with all the other side chain atoms present so that their effect can be accounted for. The process is repeated sequentially for the rest of sidechains in the gap. The process then returns to the first residue and it is repeated over each side chain until the energies of the side chain atoms do not change or until the number of passes reaches an iteration limit. This method has the virtue that only one conformation is generated per backbone conformation, and it is an energetically reasonable one. However, if there are significant interactions between the sidechain atoms, the initial state of the sidechains will bias the iterative process, and the lowest energy side chain conformation may be missed.

With any of the five methods described above, the CONGEN command can apply any of the minimization algorithms to the generated conformations. Minimization provides an ability to reduce the small van der Waals repulsions that are inevitable with coarse torsion grids.

In a "experimental" test of side chain-building capabilities of CONGEN, the backbones of an immunoglobulin VL domain (McPc 603) and myohemerythrin were stripped of their sidechains and then were rebuilt. Two different orders for iteration over the sidechains were used; sequential from the N-terminus, or ordered by increasing distance from the center of gravity of the domain. In both cases the same result was obtained, namely, the buried core of the domain, where the side chain packing density was the highest, was rebuilt fairly accurately while side chains on the surface, particularly those carrying formal charges, were often placed into positions significantly different from the crystallographic positions [11].

2.2 *Organization of a CONGEN Conformational Search*

Within CONGEN, a degree of freedom signifies a computer operation applied to a group (zero or more) of atoms by either sampling a set of variables or performing an operation on existing atoms. When a conformational search is specified, the user indicates which degrees of freedom are to be sampled and also their order. The program automatically sets up a nested iteration over all of them. Only successful samples of a degree of freedom will invoke the succeeding degrees of freedom. The process can be visualized graphically in Figure 2.

There are two reasons for taking this abstract approach to the operation of the search. First, it allows searches of arbitrary complexity to be performed. Second, the operations inherent in sampling a degree of freedom can be separated from the process of managing the search. Such modularity greatly simplifies the

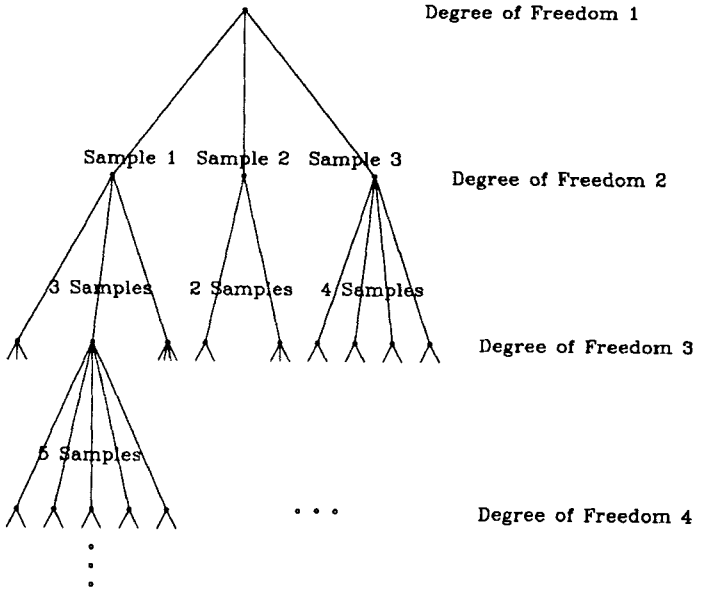

Figure 2 Presented is the top portion of a search tree showing how the sampling process can be represented. Each successively deeper (meaning lower) level in the tree corresponds to a sampling of further degrees of freedom. The root node of the tree, which is at the top of the figure, represents the segment with no intervening residues constructed. The leaf nodes, which have no nodes emanating from them and which are not shown, are either completely constructed conformations or partially constructed conformations which were blocked because of the various constraints on the sampling. The completely constructed conformations will be as deep in the tree as are there are degrees of freedom to be sampled. The blocked conformations will be higher up. Intermediate nodes represent partially constructed conformations with a particular sampling of the degrees of freedom above it. Constraints applied early serve to delete many more possible nodes than pruning later on in the generation process. Adapted from reference [23] and used by permission.

implementation of the program. In addition, once can apply the methods of state space search as developed in research into artificial intelligence [28].

Currently, six "degree of freedom" operators are provided in CONGEN. Three of them deal with atomic construction using stereochemistry. The backbone degree of freedom generates the position of the peptide backbone atoms, and the chain closure degree of freedom closes a loop. Because the Gō and Scheraga procedure finds multiple solutions to the chain closure, each solution is treated as a separate sample. The sidechain degree of freedom will construct sidechains onto any number of backbone residues, and depending on the method, it will generate either single samples or multiple ones.

Two of the degrees of freedom are involved with input and output. The "WRITE" degree of freedom writes a conformation to a file each time it is invoked. It can also do some limited filtering of what is written by comparing the energy of each conformer against the minimum energy seen thus far. Normally, this filter will greatly reduce the number of conformers written to a file. In all cases, this degree of freedom does not generate any atomic positions, and it always succeeds. The "READ" degree of freedom can be viewed as an inverse of "WRITE". It will read

Table 1 Testing CONGEN on known proteins[a].

Secondary Structure	*Protein*	*Segment*	*Min RMS*[b]	*Min E RMS*[c]
Helix	Flavodoxin[26]	127–131	0.610	0.896
Sheet	Plastocyanin[84]	80–84	0.860	1.150
Turn	McPC 603[35]	L 95–99	1.349	2.660
Turn	McPC 603	L 98–102	1.074	1.436
Turn	Kol[85]	H 41–45	0.650	1.114
Turn	Kol H	41–46	0.685	0.789
Turn	Kol H	41–47	0.743	0.882

[a] Tests of CONGEN using known proteins. Adapted from ref[23] and used by permission. © 1987 John Wiley & Sons, Inc.
[b] This column presents the minimum RMS deviation to the crystal structure. For this calculation, the INDEPENDENT sidechain construction method was used, and sidechains were evaluated based on their agreement to the crystal structure. Thus, this column shows the theoretical lower limit of agreement for conformations generated by CONGEN.
[c] Here, the ITERATIVE sidechain construction method was used, sidechains were evaluated based on their CHARMM energies.

a set of conformations from a file, and then invoke succeeding degrees of freedom on each one. Conformations can be selected based on their energies, so it is possible to set up a "build-up" procedure [29] where the best conformations from one search are used as the starting point for adding additional residues. In addition, this degree of freedom allows a user to input his own set of conformations, which can be generated by arbitrary means. This approach was used by Martin *et al.* to process loop conformations as found in a database [6].

The final degree of freedom is the "Evaluate" degree of freedom. This operation is responsible for calculating either energies or root mean square (RMS) deviations. When used for energy evaluations, this degree of freedom can either calculate the energy, or it can perform minimization or dynamics on each of the conformers. When used for RMS deviations, it will compare the coordinates of the conformations against a reference coordinate set. This is used for testing the search process; in particular, to see if a search is capable of generating the original experimental coordinates.

2.3 *Testing on Single Loops*

CONGEN was tested on a number of short segments in proteins with different secondary structures [23]. The testing was designed to answer two different questions. First, does the limitation in degree of freedom to torsion angles, restrict the conformational space so that native structures cannot be generated? Second, how well the CHARMM energy function predict correct structures? Table 1 summarizes the results of these tests. Provided that the backbone is sampled well, the answer to the first question is that there is no limitation. As seen in the middle column of the table, conformers are usually found within 1 Å of the crystal structure. The second question is answered by the last column in Table 1. In many cases, single loops are predicted within 1 Å. However, in other cases, they are not. Figure 3 shows a comparison between the CONGEN predicted structure and the X-ray structure for the last loop in the table.

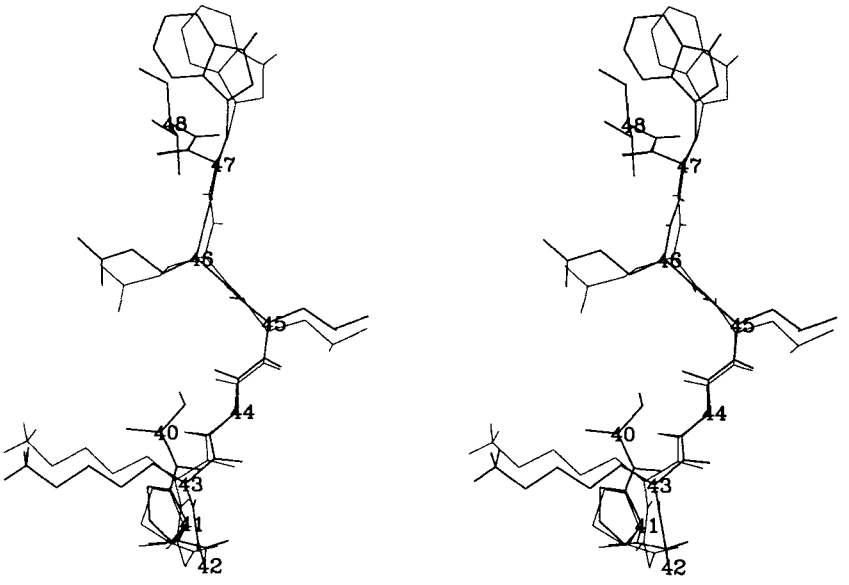

Figure 3 A stereo, stick figure showing the lowest energy conformation generated by CONGEN for heavy chain residues 41 to 47 in the antibody KOL[85] (thin lines) and the X-ray conformation (thick lines). The RMS deviation between the two is 0.882 Å. Residues 40 and 48 which surround the generated segment are also included. Adapted from reference [23] and used by permission. © 1987 John Wiley & Sons, Inc.

2.4 *Surface Area Rule*

Because the energy function currently used in CONGEN is the *in vacuo* CHARMM potential energy, solvent effects are largely ignored. In modeling experiments with McPC 603, the conformational space of each of the loops was explored. In some cases, the lowest energy conformation was close to the X-ray structure. In other cases, the lowest energy conformation deviated significantly from the X-ray structure, but the next lowest energy conformation agreed well. In all of these cases, the structures with the best agreement had the lowest accessible surface area. This made intuitive sense, as the solvent-accessible area of molecules is an approximate measure of the magnitude of solvent effects acting on the molecule; and proteins have been known to minimize their (nonpolar) solvent-exposed areas [11, 30]. Therefore, until solvent effects are incorporated into the energy calculation, loop conformations can be selected by examining all the conformations within 8.4 kJ/mole (2 kcal/mole) of the minimum and using the one with lowest accessible surface energy. The "XCONF" command in CONGEN automates this task.

3 APPLICATIONS OF CONGEN TO PROTEIN MODELING

There have been a large number of applications of systematic conformational search to protein modeling. These range over homology modeling, small peptide structure

determination, theoretical studies of protein folding, and reconstruction of protein coordinates from C_α coordinates.

3.1 *Homology Modeling*

Since many proteins evolve through the accumulation of small mutations, there are large families of proteins which have both similar sequence and structure. The problem of predicting the structure of one protein from a homologous one is generally reduced to the prediction of the effect of the substitutions. In the case of many single substitutions, the sidechain construction operators in CONGEN can be used to model the replacement sidechains. Where there are more extensive substitutions or short deletions and insertions, then the full loop searching capabilities can be brought to bear.

Nearly all of the homology modeling problems which have been attacked using CONGEN involve antibodies. In this section antibody modeling examples will be described for which there is experimental evidence to compare. Also, the modeling of a closely related protein, the T-cell receptor, will be presented.

3.1.1 *Antibody Modeling*

The antigen-combining sites of antibodies present ideal opportunities for experimenting with protein modeling procedures. First, the antibody structure, with its conserved β-sheeted framework and hypervariable combining site loops (antigen complementarity-determining regions or, CDR's) is perfectly suited for loop-splicing experiments, both in the computer and in the laboratory [31]. Second, some 15 X-ray crystallographic structures of antibody Fab fragments are currently available through the Brookhaven Protein Data Bank [32], and the number is steadily growing into an impressive structural base supporting both homology modeling and structural analysis of various antibody antigeriic specificities. Finally, combining site model-building is often the only means of obtaining three-dimensional structural information to guide protein engineering of practically important antibodies (e.g., anti-tumor therapeutic inimunoglobulins).

The antigen combining site resides in the antibody Fv fragment, a non-covalent dimer of heavy and light chain variable domains (VH and VL). The domains themselves consist of conserved framework regions (essentially, an antiparallel, 8-stranded β-sheet sandwich) and the six CDR loops (the LI-L3 loops in the VL domain and the H1-H3 loops in the VH domain). The loops vary in length and sequence among different antibodies thus creating combining sites complementary to diverse antigens.

The protocol for modeling antibodies consists of three steps: matching sequence by homology, construction of the framework, and construction of the hypervariable loops.

The goal of the sequence homology step is to find the antibody of known structure whose sequence is closest to the antibody being modeled. This can be done using the standard homology matching algorithms [33, 34]. A second factor that must be included in the choice of antibody is the length of loops that must be constructed using CONGEN. If there is a choice between two nearly homologous structures, then one should choose the one which gives the shortest difference in the hypervariable loops. Although the time will vary significantly with the number of glycine and proline residues in the loop, the local environment, and the orientation and

location of the endpoints; CONGEN can perform complete searches over loops up to about 10 residues using current computer technology.

The next step in the antibody construction protocol is building the framework. First, one must use the "SPLICE" command in CONGEN to change the sequence of the antibody of known structure into the antibody being modeled. Second, the coordinates of the hypervariable loops of the reference structure must be deleted.

Finally, one must build the coordinates for the residues which changed in the framework. The "SPLICE" command will preserve the backbone and C_β coordinates if the sequence substitutions are of equal length. In such cases, only a search over the sidechains is necessary to build them. If a substitution requires a change in the length of the sequence, or if the substitution requires the removal or addition of a proline or glycine, then a loop search should be performed. The loop should be centered on the sequence changes, it should incorporate conserved positions from both sides of the changes, and it should be at least 4 residues long.

Typically, all the individual sidechain substitutions are modeled using a single sidechain degree of freedom using the iterative search method. Then each of framework loop searches is done sequentially.

Once all the framework substitutions have been made, the hypervariable loops can be constructed. Because of computer time limitations, the searches must be performed in sequential order. Since the framework structure is known in advance, the hypervariable loops are constructed from the framework out, so that the known structure can provide a partial template for constructing the lower loops. The order used is L2, H1, L3, H2, H3, and L1.

Although the above protocol will work under ideal circumstances, Nature rarely provides ideal circumstances. However, procedures have been developed to deal with a number of problems than have arisen during antibody modeling.

One of the most common problems is the failure of CONGEN is find any conformations for a loop. A related problem is finding only high energy conformers. In both cases, this is usually due to a small number of atoms that block the space near the endpoints of the loops because of imprecise positioning of side chains in the preceding framework construction. Usually, a visual inspection of the endpoints will reveal the problem. (The PEER program provided with CONGEN can be used for this purpose.) Alternatively, the conformational search debugging variables can be activated, and examination of the output will usually show a common residue involved in bad contacts. This second method must be used with caution since the debugging output of the program can be voluminous.

Once the offending residue is found, the simplest solution is to incorporate it into the search of the loop itself. Because the degrees of freedom are general purpose operators, this is straightforward. Another option is to use a finer grid for the backbone residues in this region. If the space through which a loop must pass is narrow, the finer grid should help. A third option, which is less likely to solve the problem, is to change the order of searching the backbone. Depending on the residues where the chain closure degree of freedom is applied, the conformations of the backbone will vary by a small amount.

Another common problem that arises in the search is loops that are too long to be searched in reasonable time. The only option which has been tested is to break the search into pieces. Typically, two residues (both backbone and sidechain) are sampled at a time from each end of the loop, and the resulting conformers are written to a file. Succeeding searches use a small subset consisting of the lowest

Table 2 Summary of agreements for McPC 603[a].

Loop	*Length*	*RMS (Å)*		*CPU (μ Vax II)*
		Total	*Backbone*	
H1	5	1.7	0.7	4 hours
H2	9	2.1	1.6	5 days
H3	8	2.9	1.1	7 days
L1	12	3.0	2.6	7 days
L2	6	1.9	1.6	8 hours
L3	6	1.4	0.8	5 hours
Totals	46	2.4	1.7	20 days

[a] The RMS deviations for the loops (complete and backbone only) as well as the loop lengths and CPU times for the searches are given for each loop individually and for the six loops jointly. (Adapted from ref.[37], © 1988 Macmillan Magazines Limited.)

Table 3 Reconstruction of Hyhel15[a].

Loop	*Length*	*RMS (Å)*		*CPU (μ Vax II)*
		Total	*Backbone*	
H1	5	1.8	1.1	3 hours
H2	7	3.1	2.1	20 min.
H3	4	2.7	1.0	2 hours
L1	5	1.8	0.6	40 min.
L2	6	1.7	0.8	37 hours
L3	5	4.1	1.1	12 hours
Totals	32	2.6	1.4	2.25 days

[a] The RMS deviations for the loops (complete and backbone only) as well as the loop lengths and CPU times for the searches are given for each loop individually and for the six loops jointly. (Adapted from ref.[37], © 1988 Macmillan Magazines Limited.)

energy conformers of the previous searches to extend the chain. The order of construction usually works from each end toward the center, although it is desirable to start at endpoints which are less exposed. Since less exposed endpoints have less space for sampling conformers, fewer conformers will result, and the likelihood that the correct conformer will be selected is increased.

3.1.1.1 *McPC 603*

The first published antibody modeling with CONGEN was performed on McPC 603, a phosphorylcholine binding antibody [35]. In this study, CONGEN was used to generate conformations of each of the loops, and the results were compared against the X-ray structure as progress was made. This analysis of these conformations provided the basis for the accessible surface rule. The final results are shown in Table 2.

3.1.1.2 *HyHEL5*

Using the protocol developed for McPC 603, a model of HyHEL5 [36] was constructed [37]. The results are shown in Table 3. Because the loops in HyHEL5 were shorter than those in McPC 603, the construction was straightforward. However, two problems were noted upon comparison with the crystal structure. First, the

Table 4 Loop definitions and deviations for modeling 26-10.

Loop	*Residues*	*Number*	*RMS (Å)*	
			Total	*Backbone*
H1	Not done		5.1[a]	3.5[a]
H2	50-57	8	3.6	3.5
H3	98-106	9	5.5	5.2
L1	31-37	7	2.9	2.1
L2	55-66	6	5.0	3.6
L3	94-101	8	3.3	2.8
Totals			4.4	3.6

[a] These deviations were calculated for residues 26-35.

conformer selected for the H1 loop protruded into the space occupied by the H2 loop. As a result, the H2 loop also deviated from the correct structure. This highlighted the difficulties with the sequential construction of the loops. Second, a number of the sidechains deviated from the X-ray structure. This was due to the fact that the modeling was done in the absence of the lysozyme antigen, whereas the crystal structure was of the antibody, lysozyme complex. All of the deviations occurred with hydrophobic sidechain which protruded to make contact with lysozyme, but were modeled to lay against the antibody.

3.1.1.3 *The Anti-digoxin Antibody, 26-10*

Numerous monoclonal antibodies have been raised against digoxin a cardiac glycoside [38]. Several properties of digoxin make it a valuable antigen in the study of antibody-antigen interactions. The antigen is quite large, and the steroid moiety is nearly rigid, with only a single rotatable torsion angle, the bond to the lactone ring. Even this torsion angle is sterically hindered so that there are only two states for the rotation. There are dozens of analogs for digoxin so that one can easily probe the specificity of the interaction. Many of the antibodies against digoxin have dissociation constants in the nanomolar range or better [38-40].

The antibody, 26-10, was the first of the anti-digoxin antibodies modeled. McPC 603 was used as the reference antibody. Table 4 gives the definition of the loops used for the first model. Loop H1 was not modeled because the only sequence difference in the definition of the loop was H 35 Glu → Asn (this turned out to be an error.)

In the first attempts at construction of the model, residues H Tyr 50 and H Trp 104 in the center of combining site were modeled in a solvent exposed position. It appeared from a visual inspection of the model that H Phe 32 and H Tyr 33 were interacting with H 50. Since the definition for H2 started at H 50, we believed that extending the search to H 49 would allow for additional sampling for H 50. Therefore, the model was reconstructed with the extension of H2, and the sidechains of residues H 32 and 33 were included in the searches over H2 and H3. The resulting model had a cleft about the size of digoxin. About 10 side chains which were exposed to solvent and prominently located at the putative digoxin binding site were subjected to side-directed mutagenesis, and in almost all the cases, mutations to the selected positions resulted in alterations of antibody affinity for digoxin [39, 40, 41].

Nevertheless, the X-ray structure of 26-10 [42] turned out to differ from the

model that was constructed. The correct structure for 26–10 has a binding cavity lined with residues H 33 Tyr, H 50 Tyr, and H 104 Trp as well as the framework residue, H 47 Tyr.

Upon comparing the two structures in detail, it was found that the H1 loop was substantially different for residues 28–30. These residues were outside the range for the H1 loop as we had determined [43]. As a result of this error and the sequential construction of the loops, H2 and H3 were constructed incorrectly.

The results with 26–10 clearly demonstrate the importance of determining the endpoints to be used for modeling the loops, and the need to improve upon the sequential modeling protocol. However, the fact that polypeptide segments with essentially identical sequences may differ radically in their conformations presents a major challenge to all the protein modeling protocols [10].

3.1.1.4 *Antibody 40–150 Modeling*

A curious mutation in the antibody 40–150 [44] has also been analyzed [45]. The mutation of H94 serine to arginine results in a 1000 fold reduction in binding constant. A subsequent deletion of the first two residues of the amino terminus of the heavy chain restores much of the binding energy.

At first, reconstruction of the entire molecule of 40–150 was attempted. Unfortunately, the third hypervariable loop was too long to perform a successful search. CONGEN was then used to model just the environment around position H 94, and it was discovered that the H arginine 94 can participate in a hydrogen bond network with residues H Asp 101, L Arg 46, and L Asp 55. Thus, in this case, dealing with a group of polar atoms buried inside and required to satisfy their hydrogen bonding potentials, the CONGEN-generated conformations were ranked not according to their calculated in vacuo energy or their solvent exposed surface but according to strength and quality of hydrogen-bonding they participated in (i.e., the CONGEN-calculated H-bond potential). A serine in position H 94 cannot participate in this network. When the two amino terminal residues are deleted, H Arg 94 becomes much more solvent exposed, and its charge would be solvated. As a result, it would not participate in the hydrogen bond network, and it would restore the network back to the state when serine was in position, H 94.

3.1.1.5 *AN02*

AN02 is an anti-dinitrophenyl antibody [46] with very high homology to HyHel-5 and Hyhel-10. It was modeled with CONGEN using two different starting structures [47]. The modeling was completed prior to the solution of the X-ray structure [48], and the two models were each compared against the X-ray structure, see Table 5. In both cases the RMS deviation to the heavy chain was about 2.5 Å and the deviations to the light chain were either 2.0 or 2.1 Å.

It is illustrative to review the protocols used for AN02. The heavy chain of AN02 is 73% homologous with the heavy chain of HyHel-10, and the light chain of AN02 is 83% homologous with the light chain of HyHel-5. Thus, two models were built, one where the heavy chains are superimposed using a least squares fit and the light chains are carried along, and a second model where the light chains are superimposed and the heavy chains are carried along. Both models are similar to each other except for modest variations in L3 (1.3 Å RMS), H1 (2.8 Å RMS), and H3 (2.0 Å RMS).

Some of the loop constructions were different than encountered before. Loops

Table 5 Summary of agreements for AN02, Model 1.[a]

Loop	*Length*	*RMS (Å)*	
		Total	*Back*
H1	9	3.4	2.2
H2	7	2.1	1.7
H3	7	5.5	3.6
L1	6	2.7	1.4
L2	7	0.8	0.8
L3	7	3.8	2.4
Totals	43	3.9	2.6

[a] The RMS deviations for the loops (complete and backbone only) as well as the loop lengths are given for each loop individually and for the six loops jointly for Model 1. The results from model 2 are very similar. (Adapted from ref.[47]).

L2 and H2 were nearly identical to the parent anti-lysozyme antibodies, and therefore, no loop construction was necessary. Loop L1 was constructed with no problems.

Loop H1 had a change from Asp to Tyr at position 27 and an alanine insertion into position 34. The loop from position 26 to 34 was too long for a single CONGEN run, so the loop was split into two overlapping searches from 26 to 30 and from 29 to 34.

Loop L3 was a seven residue loop contain two sequential proliness. The first attempt to construct this loop failed to find any conformations. The search was then repeated with the backbone torsion grid for the prolines was reduced to 15°, and low energy conformations were obtained.

The H3 loop was also seven residues containing a proline. Using a 30° grid, no low energy conformations were found. Since there were no atoms blocking the endpoints, the search was repeated using a 15° search, and many good conformations were found.

3.1.2 *The T-cell Receptor*

The T-cell receptor, coupled with proteins of the major histocompatibility complex (MHC), is the primary sensor of the cellular immune response [49, 50]. The antigen/MHC binding element is a dimer of two subunits, α and β, linked by a disulphide. There are sufficient sequence and structural similarities to strongly suggest that the T-cell receptor has a variable domain with the same structure as the variable domain of antibodies [51–53].

Novotny *et al.* [54] devised a single chain flourescein binding T-cell receptor (named RFL3.8) from the variable domains of the receptor and constructed a model of it using the antibody construction protocol described above, except that the $\beta 3$ hypervariable loop was too long to completed in reasonable time. The conformational search on this loop was truncated after 7 days of computation [55], and the lowest energy conformer found was used in the model. In addition, the 23 residue linker between the α and β chains was modeled by using a very narrow range of ϕ, ψ torsion angles for the backbone conformations. Since the structure of the linker is expected to be disordered in solution because of its large percentage of glycines, the goal of this linker model was to generate a plausible structure.

Although experimental structural studies on the single chain T-cell receptor remain to be performed, two site directed mutation experiments suggest that the structure is largely correct. First, after the receptor was first cloned, it was noted that it was poorly soluble. Examination of the model revealed that five hydrophobic sidechains were present on the surface opposite to the combining site. In immunoglobulins, these residues were all hydrophilic. All five of these hydrophobic residues were mutated to hydrophilic residues. The resulting receptor was more soluble than the original and had nearly identical fractionation and binding properties.

Second, the model of the single chain T-cell receptor was superimposed on the anti-flourescein antibody 4-4-20 [56]. The model shows a cavity whose size and relative location is similar to that of the antibody. Six residues within the cavity of the RFL3.8 T-cell receptor as well as a lysine residue outside the combining site were mutated to alanine. Five of the six mutants made from residues within the cavity lost binding to flourescein, and the mutant of the external lysine had no change in binding relative to the wild type receptor [57]. Although the experience with 26-10 illustrates that the T-cell receptor model could be incorrect, these experimental results are very encouraging.

3.2 *Small Peptides*

Conformational search has been extensively applied to the structure of small peptides [58-63]. CONGEN has the necessary code and topology files for D amino acids, as well as the logic for handling cyclic peptides where the residues are joined in the backbone.

This capability was used in the determination of the structure of cyclo-(D-Trp-D-Asp-L-Pro-D-Val-L-Leu) [64], an antagonist for the endothelin ET_A receptor. The structure was determined by both NMR spectroscopy and by global energy minimization using conformational search. Since the peptide was cyclic, the modified Gō and Scheraga chain closure algorithm could be used on the backbone, and therefore, the free backbone torsions could be sampled using a 10° or 15° grid. All backbone conformations were minimized by 50 steps of Adopted Basis Newton-Raphson minimization [13].

The RMS deviation for heavy atoms between the average experimental structure and the lowest two energy CONGEN structures were 0.25 Å and 2.12 Å. The backbone RMS deviations for these two structures were 0.22 Å and 0.37 Å. The difference between the two lowest energy structures was the orientation of tryptophan sidechain. Otherwise, the CONGEN determined structures was nearly identical to the experimental structure.

3.3 *Sidechain Reconstructions*

Because the sidechain degree of freedom can be invoked independently of any other degree of freedom, CONGEN has been used in a number of applications where sidechain positions needed to be modeled [11, 65-68].

3.3.1 *Incorrectly Folded Structures*

An important test of any protein modeling procedure is the ability to discriminate correctly from incorrectly folded proteins. Novotny *et al.* [11, 66] constructed models of sea-worm hemerythrin and the variable domain of the mouse κ light chain

where the structures were clearly incorrect. Both of these proteins have the same length. The incorrect models were generated by swapping the sidechains of one protein onto the backbone of the other. In the first of these studies [66], the sidechains were initially constructed using *trans* sidechain torsion angles, and then, energy minimization was used to refine the sidechain positions. In the second study [11], the sidechain modeling was improved by using the sidechain construction operator in CONGEN to rebuild the sidechains. The energies of the new models were improved, but the incorrectly folded models still had excessive non-polar solvent exposed sidechain surfaces.

The sidechain construction operator was tested on the native protein structures. For both proteins, sidechain atoms beyond C_β were removed, and the iterative sidechain algorithm was used to rebuild all the sidechains. Although there was great variation in the individual sidechain deviations (from 0.0 Å to 6.5 Å), the largest differences were found with exposed sidechains. As the energy functions used to select sidechain conformations are improved, the agreement in the interior should also improve.

3.3.2 *Coiled coils*

Supercoiled dimers of α-helices are a common structural motif in transcriptional factors [69, 70] and fibrous proteins [71]. The structural parameters for the supercoiling [72] can be used to construct the the peptide backbone in these dimers [67]. The sidechains for a particular sequence can then be added using the sidechain construction operator from CONGEN. It is possible to optimize the dimerization by calculating the dimerization energy while varying the supercoiling parameters. Two examples of such calculations are given below.

Synthetic oligoheptapeptides having an alternating pattern of hydrophobic and hydrophilic residues $(Lys\text{-}Leu\text{-}Glu\text{-}Ala\text{-}Leu\text{-}Glu\text{-}Gly)_n$ were shown to dimerize when the repeat length was at least four. Shorter peptides showed random structure [73]. The experiments were performed on peptides with repeats of one to five.

Models of the five peptides were constructed by using the supercoiling parameters of Crick and optimizing the distance between the helices and rotation angle of the helices. The peptide sidechains for each of these trials was constructed with the iterative sidechain construction method of CONGEN. Using an empirical free energy potential [66], the free energy of helix and dimer formation was calculated, and it was found that the energy of helix and dimer formation was positive for repeat lengths of one through three, and negative for four and five.

Using the construction protocols developed for the coiled coils, models of the leucine zipper DNA-binding proteins; *fos, jun, GCN4*; were used to calculate the stability of various homo- and heterodimers [68]. Recently, the X-ray structure for GCN4 was published [74], and the model structure for GCN4 was compared against the X-ray structure. The overall RMS deviation for a least squares superposition of all atoms was 2.70 Å, but most of this deviation is due to the hydrophilic sidechains which would be expected to be disordered in solution. The RMS deviation for the backbone atoms is 1.08 Å, and the deviation for all the sidechains was 3.58 Å. The agreement for the leucine and valine sidechains found in the interface was only 1.51 Å. Visually, the packing of the leucine and valine sidechains in the interface is also preserved.

Table 6 RMS deviations for reconstructions from C_α coordinates[a].

Protein	*RMS C_α (Å)*	*RMS backbone (Å)*
Myohemerythrin[86]	0.87	0.89
Flavodoxin[26]	0.36	0.60
Concavalin A[87]	0.73	0.99
Triose phosphate isomerase[88]	0.46	0.71
Carboxypeptidase A[89]	0.62	0.84
McPC 603 Heavy Chain[35]	0.30	0.50
Thioredoxin[90]	1.28	NA[b]
Triacylglycerol acylhydrolase[91]	1.02	NA[b]

[a] Adapted from ref.[75] and used by permission. © 1992 Wiley-Liss Inc.
[b] Not available because only C_α coordinates were published.

3.4 *Reconstruction of the Backbone Coordinates from C_α Coordinates*

A substantial fraction of the protein coordinate sets in the Brookhaven Protein Data Bank [32] contain only coordinates for the α-carbons. It is highly desirable to reconstruct the remaining coordinates from them.

We have begun exploring the use of a directed search of conformational space in order to find good conformations for segments too long to be sampled exhaustively. Since a systematic search can be described as a traversal of a search tree, information about partial conformations can be used to direct the traversal (see Figure 2). For example, the best first search method begins by sampling the first degree of freedom. The partial conformer generated in this first step which has the best RMS deviation for the C_α position will be used as first partial conformation for sampling the second degree of freedom. Thereafter, the search tree is examined for the partial conformer that has the best RMS deviations for its C_α's, and the program selects for sampling the degree of freedom which extends the best partial conformation. Although this best first search method does not lead by itself to good quality conformations for large searches, variants of this procedure do work very well [75, 76]. The α-carbon reconstruction problem is well suited to a directed search because any backbone conformation which deviates significantly from the α-carbons can be ignored. Once the backbone is constructed, the sidechain construction operator can then be used to generate sidechain positions. Because the construction of sidechains depend much more on the accuracy of the non-bonded interaction potential, the accuracy of the sidechains will not be as good as the backbone.

The results of a number of test fits are given in Table 6. In all of these cases, the complete structures were found in a single search using a 5° grid and generating approximately 5×10^7 partial conformers. The branching factor on each backbone degree of freedom was around 600 conformers. Yet, these results are comparable to other methods which use geometric information [77], energy minimization [78], or protein fragments from a database [79–81].

3.5 *Comparison to Dynamics*

Molecular dynamics is a widely used technique for sampling conformational space. It has been compared to conformational search both in terms of efficiency and completeness [82].

In this study, the motion of the six hypervariable loops of the variable domains of McPC 603 was simulated at temperatures of 500°K, 800°K, and 1500°K. The framework residues were rigidly fixed. The simulations were run for periods of 111.5, 101.7, and 76.3 picoseconds, respectively, and they were analyzed in a variety of ways to measure the conformational space that was explored. In all cases, the trajectories never converged to any region of conformational space.

Two of the hypervariable loops, H1 and L2, were defined with lengths short enough that a conformational search was completed within a day of VAX 780 CPU time, and the conformational spaces sampled by dynamics and CONGEN were compared.

At 500°K, the space generated by CONGEN was larger by all measures that were used. In addition, for the H1 loop, many conformations were found by CONGEN which were far from any conformations in the trajectory, whereas all dynamics conformations were found close to some CONGEN conformation. For the L2 loop, the conformational samplings generated by each method had unique conformers.

At 800°K, most of the measures of conformational space were larger for the CONGEN sampling, but the volume measure for H1 was larger for the dynamical sampling. For both loops, the two samplings each had unique conformers.

At 1500°K, the dynamics sampling was larger by all measures. Given the extensive *cis-trans* peptide bond isomerization, this was predictable.

The result of this comparison suggest that for problems where the conformational space is short enough for systematic search, then it should be employed because of its greater coverage and efficiency. However, it must be realized that conformational search does not sample over bond angles, and this could result in missing some regions of conformational space that molecular dynamics at high temperature might encounter [83].

4 CONCLUSION

Systematic conformational search is a useful tool for a variety of problems in protein and peptide modeling. A number of problems have been described for which systematic search provided useful insights and reasonable agreement with experiment. Nonetheless, the full potential of this method has not been realized. Improvements to the energy functions and to the search methods will extend the range and type of problems to which this technique can be applied. In addition, we will gain further insight into the energetics of protein folding as a whole.

For information about obtaining the CONGEN program, please contact the author.

Acknowledgements

I thank Drs Edgar Haber and Jiri Novotny for their unwavering support, guidance, and insights over the years. This work was partially funded by NIH Grant PO1-HL 19259 and other grants from NIH, NSF, and ONR. I thank Jiri Novotny, Donna Bassolino, Joel Schildbach, and Phil Jeffrey for the use of data prior to publication, and Howard Alper for a careful reading of the manuscript. I am grateful to John Wiley and Sons, Inc., Wiley-Liss Inc., the American Association for the Advancement of Science, Cambridge University Press, and Macmillan Magazines Ltd. for permission to use copyrighted material. Some of the text describing CONGEN was adapted from our contribution to reference [1].

References and Notes

[1] L.M. Gierasch and J. King, *Protein Folding: Deciphering the Second Half of the Genetic Code*, American Association for the Advancement of Science, 1990, Washington, D.C.

[2] C. Levinthal in *Mossbauer Spectroscopy in Biological Systems*, P. Debrunner, J.C.M. Tsibris, and E. Münck, eds., University of Illinois Press, Urbana 1969 pp 22-24.

[3] K.A. Dill, "Dominant Forces in Protein Folding", *Biochemistry*, **29**, 7133-7155 (1990).

[4] R. Zwanzig, A. Szabo, and B. Bagchi, "Levinthal's paradox", *Proc. Natl. Acad. Sci. USA*, **89**, 20-22 (1992).

[5] T.A. Jones and S. Thirup, "Using known substructures in protein model building and crystallography", *EMBO J.*, **5**, 819-822 (1986).

[6] A.C.R. Martin, J.C. Cheetham, and A.R. Rees, "Modeling antibody hypervariable loops: a combined algorithm", *Proc. Natl. Acad. Sci. USA*, **86**, 9268-9272 (1989).

[7] C. Chothia, A.M. Lesk, A. Tramontano, M. Levitt, S.J. Smith-Gill, G. Air, S. Sheriff, E.A. Padlan, D. Davies, W.R. Tulip, P.M. Colman, S. Spinelli, P.M. Alzari, and R.J. Polyak, "Conformations of inimunoglobulin hypervariable regions", *Nature*, **342**, 877-883 (1989).

[8] M.J. Sutcliffe, I. Haneef, D. Carney, and T.L. Blundell, "Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures", *Protein Engineering*, **1**, 377-384 (1987).

[9] A.C.R. Martin, J.C. Cheetham, and A.R. Rees, "Molecular Modeling of Antibody Combining Sites", *Methods in Enzymology*, **203**, 121-153 (1991)

[10] W. Kabsch and C. Sander, "Identical pentapeptides with different backbones", *Nature*, **317**, 207 (1985).

[11] J. Novotny, A.A. Rashin, R.E. Bruccoleri, "Criteria that Discriminate Between Native Proteins and Incorrectly Folded Models", *Proteins*, **4**, 19-30 (1988).

[12] J. Novotny, R.E. Bruccoleri, F. Saul, "On the Attribution of Binding Energy in Antigen-Antibody Complexes McPC 603, D1.3, and HyHEL-5", *Biochemistry*, **28**, 4735-4749 (1989).

[13] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, "CHARMM - A Program for Macromolecular Energy, Minimization, and Dynamics Calculations", *J. Comput. Chem.*, **4**, 187-217 (1983).

[14] J. Moult and M.N.G. James, "An algorithm for determining the conformation of polypeptide segments in proteins by systematic search", *Proteins: Structure, Function, and Genetics*, **1**, 146-163 (1986).

[15] M. Lipton and W.C. Still, "The Multiple Minimum Problem in Molecular Modeling. Tree Searching Internal Coordinate Conformational Space", *J. Comput. Chem.*, **9**, 343-355 (1988).

[16] D.P. Dolata, A.R. Leach, and K. Prout, "Wizard: AI in conformational analysis", *J. Comput.-Aided Mol. Des.*, **1**, 73-85 (1987).

[17] P.S. Shenkin, D.L. Yarmush, R.M. Fine, H. Wang, and C. Levinthal, "Predicting the antibody hypervariable loop conformation I. Ensembles of random conformations for ringlike structures", *Biopolymers*, **26**, 2053-2085 (1987).

[18] R.A. Dammkoehler, S.F. Karasek, E.F. Berkley Shands and G.R. Marshall, "Constrained search of conformational hyperspace", *J. Comput-Aided Mol. Des.*, **3**, 3-21 (1989).

[19] N. Gō and H.A. Scheraga, "Ring Closure and Local Conformational Deformations of Chain Molecules", *Macromolecules*, **3**, 178-187 (1970).

[20] T.F. Havel, I.D. Kuntz, and G.M. Crippen, "The combinational distance geometry method for the calculation of molecular conformations. I. A new approach to an old problem", *J. Theor. Biol.*, **104**, 359-381 (1983).

[21] G.M. Smith and D.F. Weber, "Computer aided systematic search of peptide conformations constrained by NMR data", *Biochem. Biophys. Res. Commun.*, **134**, 907-914 (1986).

[22] Z. Li and H.A. Scheraga, "Monte-Carlo minimization approach to the multiple minima problem in protein folding", *Proc. Natl. Acad. Sci. USA*, **84**, 6611-6615 (1987).

[23] R.E. Bruccoleri, M. Karplus, "Prediction of the Folding of Short Polypeptide Segments by Uniform Conformational Sampling", *Biopolymers*, **26**, 137-168 (1987).

[24] The chain closure algorithm can perturb the bond angles in the peptide backbone a small amount.

[25] R.E. Bruccoleri, and M. Karplus, "Chain Closure with Bond Angle Variations", *Macromolecules*, **18**, 2767-2773 (1985).

[26] W.W. Smith, R.M. Burnet, G.D. Darling, M.L. Ludwig, "Structure of the semiquinone form of flavodoxin from clostridium M.P.", *J. Mol. Biol.*, **117**, 195-226 (1977).

[27] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations", *J. Mol. Biol.*, **7**, 95-99 (1963).

[28] J. Pearl and R.E. Korf, "Search Techniques", *Ann. Rev. Comput. Sci.*, **2**, 451-467 (1987).
[29] M.R. Pincus, R.D. Klausner, and H.A. Scheraga, "Calculation of the three dimensional structure of the membrane-bound portion of melittin from its amino acid sequence", *Proc. Nat. Acad. Sci. USA*, **79**, 5107-5110 (1982).
[30] J. Novotny, R.E. Bruccoleri, and M. Karplus, "An Analysis of Incorrectly Folded Protein Models, Implications for Structure Prediction", *J. Mol. Biol.*, **177**, 787-818 (1984).
[31] P.T. Jones, P.H. Dear, J. Foote, M.S. Newberger, and G. Winter, "Replacing the complementarity-detemining regions in a human antibody with those from a mouse", *Nature*, **321**, 522-525 (1986).
[32] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shirmanouchi, and M. Tasumi, "The Protein Data Bank: a computer-based archival file for macromoleculer structures", *J. Mol. Biol.*, **112**, 535-542 (1977).
[33] S.B. Needleman and C.D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *J. Mol. Biol.*, **48**, 443-453 (1970).
[34] R.A. Wagner and M.J. Fischer, "The string to string correction problem", *J. Association of Computing Machinery*, **21**, 168-173 (1974).
[35] Y. Satow, G.H. Cohen, E.A. Padlan, and D.R. Davies, "Phosphorylcholine binding immunoglobulin Fab McPC603 - an X-ray diffraction study at 2.7 Å", *J. Mol. Biol.*, **190**, 593-604 (1986).
[36] S. Sheriff, E.W. Silverton, E.A. Padlan, G.H. Cohen, S.J. Smith-Gill, B.C. Finzel, and D.R. Davies, "The three dimensional structure of an antibody-antigen complex", *Proc. Nat. Acad. Sci. USA* **84**, 8075-8079 (1987).
[37] R.E. Bruccoleri, E. Haber, J. Novotny, "Structure of Antibody Hypervariable Loops Reproduced by a Conformational Search Algorithm", *Nature*, **335**, 564-568 (1988); Errata: *Nature*, **336**, 266 (1988).
[38] M. Mudgett-Hunter, W. Anderson, E. Haber, and M.N. Margolies, "Binding and structural diversity among high-affinity monoclonal anti-digoxin antibodies", *Molecular Immunology*, **22**, 477-488 (1985).
[39] J.F. Schildbach, D.J. Panka, D.R. Parks, G.C. Jager, J. Novotny, L.A. Herzenberg, M. Mudgett-Hunter, R.E. Bruccoleri, E. Haber, and M.N. Margolies, "Altered Hapten Recognition by Two Anti-Digoxin Hybridoma Variants due to Variable Region Point Mutations", *J. Biol. Chem.*, **266**, 4640-4647 (1991).
[40] J.F. Schildbach, personal communication.
[41] R.I. Near, R.E. Bruccoleri, J. Novotny, N.W. Hudson, A. White, M. Mudgett-Hunter, "The Specificity Properties that Distinguish Members of a Set of Homologous Anti-Digoxin Antibodies are Controlled by H Chain Mutations", *J. Immunology*, **146**, 627-633 (1991).
[42] P. Jeffrey, personal communication.
[43] J. Novotny, R.E. Bruccoleri, J. Newell, D. Murphy, E. Haber, and M. Karplus, "Molecular Anatomy of the Antibody Binding Site", *J. Biol. Chem.*, **258**, 14433-14437 (1983)
[44] D.J. Panka, M. Mudgett-Hunter, D.R. Parks, L.L. Peterson, L.A. Herzenberg, E. Haber, and M.N. Margolies, "Variable region framework differences result in decreased or increased affinity of variant anti-digoxin antibodies", *Proc. Nati. Acad. Sci. USA*, **85**, 3080-3084 (1988).
[45] J. Novotny, R.E. Bruccoleri, E. Haber, "Computer Analysis of Mutations that Affect Antibody Specificity", *Proteins*, **7**, 93-98 (1990).
[46] J. Anglister, T. Frey, and H.M. McConnell, "Magnetic resonance of a monoclonal anti-spin label antibody", *Biochemistry*, **23**, 1138-1142 (1984).
[47] D. Bassolino, R.E. Bruccoleri, and S. Subramaniam "Modeling the Antigen Combining Site of an Anti-dinitrophenyl Antibody, AN02", *Protein Science*, **1**, 1465-1476 (1992).
[48] A.T. Brünger, D.J. Leahy, T.R. Hynes, and R.O. Fox, "2.9 Å resolution structure of an anti-dinitrophenyl-spin-label monoclonal antibody Fab fragment with bound hapten", *J. Mol. Biol.*, **221**, 239-256 (1991).
[49] S.C. Meuer, O. Acuto, T. Hercend, S.F. Schlossman, and E.L. Reinherz, "The human T-cell receptor", *Annu. Rev. Immunol.*, **2**, 23-50 (1984).
[50] J.D. Ashwell, and R.D. Klausner, "Genetic and mutational analysis of the T-cell antigen receptor", *Annu. Rev. Immunol.*, **8**, 139-167 (1990).
[51] J. Novotny, S. Tonegawa, H. Saito, D.M. Kranz, and H.N. Eisen, "Secondary, tertiary and quaternary structure of T-cell-specific immunoglobulin-like polypeptide chains", *Proc. Natl. Acad. Sci USA*, **83**, 742-746 (1986).
[52] C. Chothia, D.R. Boswell, and A.M. Lesk, "The outline structure of T-cell alpha-beta receptor", *EMBO J.*, **7**, 3745-3755 (1988).

[53] J.M. Claverie, A. Prochnicka-Chalufour, and L. Bougeleret, "Implications of a Fab-like structure for the T-cell receptor", *Immunol. Today.*, **10**, 10–14 (1989).
[54] J. Novotny, R.K. Ganju, S.T. Smiley, R.E. Hussey, M.A. Luther, M.A. Recny, R.F. Siliciano, and E.L. Reinherz, "A soluble, single-chain T-cell receptor fragment endowed with antigen combining properties", *Proc. Natl. Acad. Sci. USA*, **88**, 8646–8650 (1991).
[55] Since this work was completed, CONGEN has been modified to run in parallel on the Silicon Graphics multiprocessor workstations, and therefore, it runs about 10 times faster. It is likely that the same search could be completed today.
[56] J.N. Herron, X.M. Hei, M.L. Mason, E.W. Voss, and A.B. Edmundson, "Three-dimensional structure of a fluorescein-Fab complex crystallized in 2-methyl-2,4-pentanediol", *Proteins*, **5**, 271–280 (1989).
[57] R.K. Ganju. S.T. Smiley, J. Bajorath, and J. Novotny, "Similarity between fluorescein-specific T-cell receptor and antibody in chemical details of antigen recognition", *Proc. Natl. Acad. Sci. USA*, **89**, 11552–11556 (1992).
[58] D. Hall and N. Pavitt, "Conformation of cyclic analogs of enkephalin. III. Effect of varying ring size.", *Biopolymers*, **24**, 935–945 (1985).
[59] D. Hall, N. Pavitt, and M.K. Wood, "The conformation of pithomycolide", *J. Comput. Chem.*, **3**, 381–384 (1982).
[60] V. Madison, "Cyclic peptides revisited", *Biopolymers*, **24**, 97–103 (1985).
[61] C.M. Deber, V. Madison, and E.R. Blout, "Why cyclic peptides? Complementary approaches to conformation", *Acc. Chem. Res.*, **9**, 106–113 (1976).
[62] C.M. Venkatachalam, M.A. Khaled, H. Sugano, and D.W. Urry, "Nuclear magnetic resonance and conformational energy calculations of repeat peptides of elastin. Conformational characterization of cyclopentadecapeptide cyclo-(L-Val-L-Pro-Gly-L-Val-Gly)$_3$", *J. Am. Chem. Soc.*, **103**, 2372-2379 (1981).
[63] M. Dygert, N. Gō, and H.A. Scheraga, "Use of a symmetry condition to compute the conformation of Gramicidin S", *Macromolecules*, **8**, 750–761 (1975).
[64] S.R. Krystek, Jr., D.A. Bassolino, R.E. Bruccoleri, J.T. Hunt, M.A. Porubcan, C.F. Wandler, N.H. Andersen, "Solution Conformation of a Cyclic Pentapeptide Endothelin Antagonist: Comparison of Structures Obtained From Constrained Dynamics and Conformational Search", *FEBS Letters*, **299**, 255–261 (1992).
[65] J. Allen, J. Novotny, J. Martin, and G. Heinrich, "Molecular structure of mammalian neuropeptide Y: Analysis by molecular cloning and computer-aided comparison with crystal structure of avain homologue", *Proc. Natl. Acad. Sci USA*, **84**, 2532–2536 (1987).
[66] J. Novotny, R.E. Bruccoleri, and M. Karplus, "An analysis of incorrectly folded protein models. Implications for structure predictions", *J. Mol. Biol.*, **177**, 787–818 (1984).
[67] R.E. Bruccoleri, J. Novotny, P. Keck, and C. Cohen, "Two-stranded α-helical coiled-coils of fibrous proteins. Theoretical Analysis of Supercoil Formation", *Biophys. J.*, **49**, 79–81 (1986).
[68] S.R. Krystek, R.E. Bruccoleri, and J. Novotny, "Stabilities of leucine zipper dimers estimated by an empirical free energy method". *Int J. Peptide Protein Res.*, **38**, 229–236 (1991).
[69] W.H. Landschutz, P.F. Johnson, and S.L. McKnight, "The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins", *Science*, **240**, 1759–1764 (1988).
[70] T. Kouzarides and E. Ziff, "The role of the leucine zipper in the fos-jun interaction", *Nature*, **336**, 646–651 (1988).
[71] N. Geisler and K. Weber, "Amino acid sequence data on glial fibrillary acidic protein (GFA); implications for the subdivision of intermediate filaments into epithelial and non-epithelial members", *EMBO J.*, **2**, 2059–2063 (1983).
[72] F.H.C. Crick, "The Fourier transform of a coiled-coil." *Acta Cryst.*, **6**, 685–689 (1953).
[73] S.Y.M. Lau, A.K. Taneja, and R.S. Hodges. 1984. "Synthesis of a model protein of defined secondary and quaternary structure. Effect of chain length on the stabilization and formation of two-stranded α-helical coiled-coils." *J. Biol. Chem.*, **259**, 13253-13261.
[74] E.J. O'Shea, J.D. Klemm, P.S. Kim, and T. Alber. "X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil", *Science*, **254**, 539–544 (1991).
[75] D. Bassolino-Klimas and R.E. Bruccoleri. "The application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates", *Proteins: Structure, Function and Genetics*, **14**, 465–474 (1992).
[76] J. Novotny, R.E. Bruccoleri, and P. Kourilsky. "On the Molecular Nature of 'Restrictive' Antigenic Elements Present on Major Histocompatibility Complex (MHC) Proteins", *Annals of Institute Pasteur/Immunology*, **140**, 145–158 (1989).

[77] E.O. Purisma and H.A. Scheraga, "Conversion of virtual bond chain to a complete polypeptide chain", *Biopolymers*, **23**, 1207–1224 (1984).
[78] P. Correa, "The building of protein structures from α-carbon coordinates", *Proteins: Structure, Function and Genetics*, **7**, 366–377 (1990).
[79] L.S. Reid and J.M. Thornton, "Rebuilding Flavodoxin from C_α coordinates: A test study", *Proteins: Structure, Function, and Genetics*, **5**, 170–182 (1989).
[80] M. Classens, E. Van Cutsem, I. Lasters, and S. Wodak, "Modelling the polypeptide backbone with spare parts from known protein structures", *Protein Engineering*, **2**, 335–345 (1989).
[81] L. Holm and C. Sander, "Database algorithm for generating protein backbone and sidechain coordinates from a C_α trace", *J. Mol. Biol.*, **218**, 183–194 (1991).
[82] R.E. Bruccoleri, M. Karplus, "Conformational Sampling using High Temperature 'Molecular Dynamics' ", *Biopolymers*, **29**, 1847–1862 (1990).
[83] W.F. van Gunsteren and M. Karplus, "Effects of constraints, solvent and crystal environment on protein dynamics", *Nature*, **293**, 677–678 (1981).
[84] P.M. Colman, H.C. Freeman, J.M. Guss, M. Murata, V.A. Norris, J.A.M. Ramshaw, and M.P. Venkatappa, "X-ray crystal structure analysis of plastocyanin at 2.7 Å resolution", *Nature*, **272**, 319–324 (1978).
[85] M. Marquart, J. Deisenhofer, R. Huber, and W. Palm, "Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3.0 Å and 1.0 Å resolution", *J. Mol. Biol.*, **141**, 369–391 (1980).
[86] S. Sheriff and W. Hendrickson, "Structure of Myohemerythrin in the azidomet state at 1.7/1.3 Å resolution", *J. Mol. Biol.*, **197**, 273–296 (1987).
[87] K.D. Hardman and C.F. Ainsworth, "Structure of Con A at 2.4 Å Resolution", *Biochemistry*, **11**, 4910–4919 (1972).
[88] D.W. Banner, A.C. Bloomer, G.A. Petsko, D.C. Phillips, and I.A. Wilson, "Atomic Coordinates for Triose Phosphate Isomerase from Chicken Muscle", *Biochem. Biophys. Res. Comm.*, **72**, 146–155 (1976).
[89] D.C. Rees and W.N. Lipscomb, "Crystallographic studies on Apocarboxypeptidase A at 1.54 Å resolution", *J. Mol. Biol.*, **168**, 367–387 (1983).
[90] A. Holmgren, B. Söderberg, H. Eklund, C. Bränden, "Three dimensional structure of E. Coli thioredoxin-S2 to 2.8 Å resolution", *Proc. Natl. Acad. Sci. USA*, **72**, 2305–2309 (1975).
[91] L. Brady, A.M. Brzozowski, Z.S. Derewenda, E.J. Dodson, G.G. Dodson, S.P. Tolley, J.P. Turkenburg, "A serine protease triad forms the catalytic center of a triacyclglycerol lipase", *Nature*, **343**, 767–770 (1990).