PROTEINS: Structure, Function, and Genetics 33:227–239 (1998)

Exploring the Conformational Space of Protein Side Chains Using Dead-End Elimination and the A* Algorithm

Andrew R. Leach[1]* and Andrew P. Lemon[2]
[1]*Glaxo Wellcome Medicines Research Centre, Stevenage Hertfordshire, United Kingdom*
[2]*Department of Chemistry, University of Southampton, Southampton, United Kingdom*

ABSTRACT **We describe an algorithm which enables us to search the conformational space of the side chains of a protein to identify the global minimum energy combination of side chain conformations as well as all other conformations within a specified energy cutoff of the global energy minimum. The program is used to explore the side chain conformational energy surface of a number of proteins, to investigate how this surface varies with the energy model used to describe the interactions within the system and the rotamer library. Enumeration of the rotamer combinations enables us to directly evaluate the partition function, and thus calculate the side chain contribution to the conformational entropy of the folded protein. An investigation of these conformations and the relationships between them shows that most of the conformations near to the global energy minimum arise from changes in side chain conformations that are essentially independent; very few result from a concerted change in conformation of two or more residues. Some of the limitations of the approach are discussed. Proteins 33:227–239, 1998.** © 1998 Wiley-Liss, Inc.

Key words: conformational search; dead-end elimination; A* algorithm; protein; side chain; rotamer library; protein folding; entropy

INTRODUCTION

The conformations of a molecule can significantly influence its properties and behavior. A crucial component of any conformational analysis is the identification of the thermally accessible conformations of a molecule. This can be a difficult problem as even "small" molecules frequently have very many minimum energy structures. A protein's side chains are vital to its stability and function. Side chains are also frequently implicated in the binding of substrates and inhibitors and are thus crucial to molecular recognition processes in biology. Unfortunately, reliable modeling has been limited by the vast number of possible combinations of side chain conformations. In this paper we are concerned with the nature of the conformational energy surface of the side chains of the amino acids in a protein and in estimating the conformational entropy of these side chains in the folded protein structure.

The problem of estimating the contribution to the free energy of folding of the conformational entropy of the side chains has previously been considered by a number of research groups. A review has been published by Doig and Sternberg.[1] These methods typically use the following equation for the entropy:

$$S = -\Sigma p_i \ln p_i. \quad (1)$$

p_i is the probability of state i. The contribution of the conformational degrees of freedom of the side chains to the overall entropy of folding is given by the difference between the entropies of the folded and unfolded states as calculated using equation (1). †

The side chain conformational entropy of the unfolded state has been estimated by both empirical and theoretical approaches. The empirical approaches such as those of Pickett and Sternberg[2] or Abagyan and Totrov [3] are based upon analyses of the protein databank, where it is assumed that the conformations adopted by side chains in protein crystal structures are representative of the unfolded state. In estimating the change in side chain conformational entropy they also assumed that each side chain is restricted to a single conformation in the folded state. This means that p_i in Equation (1) is zero for all side chains with the exception of sym-

Abbreviations: GMEC, global minimum energy conformation; DEE, dead-end elimination.

Grant sponsor: Biotechnology and Biological Sciences Research Council; Grant sponsor: Engineering and Physical Sciences Research Council.

*Correspondence to: Andrew R. Leach, Glaxo Wellcome Medicines Research Centre, Gunnels Wood Road, Stevenage Hertfordshire SG1 2NY, United Kingdom.

Received 3 February 1998; Accepted 5 June 1998

†As also pointed out by Doig and Sternberg, it is common to assume that each rotamer has the same potential energy well for both the folded and unfolded states, thus implying that the vibrational contribution to the entropy is zero. This will also be the case in the work described here.

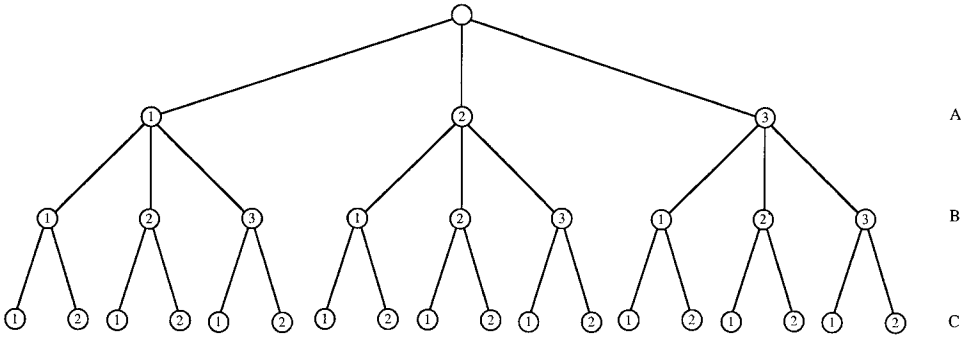


Fig. 1. Illustration of the relationship between conformational space and tree representation, for a tripeptide in which residues A and C has three rotamers each and residue B has two rotamers.

metrical residues such as Phe, Asp, and Glu for which the twofold symmetry means that the entropy in the folded state is Rln2. Various theoretical approaches have been used to estimate the conformational entropy change, including the mean field method of Koehl and Delarue,[4] the systematic conformational search method of Wang and Purisima,[5]and the Monte Carlo simulations of Creamer and Rose.[6,7] These approaches all provide an estimate of the entropy in the folded state though the systematic search method and the Monte Carlo simulations have thus far been restricted to rather small model systems (typically of α-helices); their application to larger systems would be severely limited due to the computational resources required. By contrast, the mean-field approach of Koehl and Delarue[4] has been applied to "real" proteins but again, is a probabilistic approach. The entropy calculation requires that the side chain conformations are not correlated. It is one of the objectives of our study to investigate the correlations between side chains, and thence the validity of this assumption, as well as the validity of the $p_i = 0$ approximation for the folded state.

METHODS

A Tree Representation of Conformational Space

If the side chains of a protein are restricted to distinct conformational states (which we will refer to as *rotamers*), then the problem of exploring the conformational space can be conveniently represented as a search tree. Each of the goal nodes in the tree represents a conformation in which every side chain has been assigned a particular rotameric state; all the intermediate nodes represent conformations in which only some of the side chains have been assigned. We can illustrate this concept using as an example the conformational space of a simple tripeptide (ABC) in which amino acids A and B both have three rotamers, and amino acid C has two. The search tree for this molecule is shown in Figure 1 and as can be seen there are a total of 18 ($= 3 \times 3 \times 2$) possible conformations. In fact, the number of possible conformations (i.e. the number of *goal nodes* in the search tree) is frequently a very large number for real proteins; 10^{100} or more is not uncommon. From this immense search space we wish to identify all combinations of side chain rotamers that give rise to low-energy conformations, that is all conformations that would have an appreciable probability, p_i.

Our strategy is to use a combination of two search algorithms,[8] the dead-end elimination (DEE) algorithm of Desmet et al.,[9] together with the A* algorithm[10] to explore the side chain conformational space. Here we provide some brief background information to these two methods.

The DEE algorithm identifies side chain conformations that are incompatible with the global minimum energy arrangement. The energy of a combination of side chain rotamers is given by:

$$E_{total} = E_{template} + \sum_{i=1}^{nres} E_{i_r,template} + \sum_{i=1}^{nres-1} \sum_{j>i}^{nres} \epsilon_{i_r,j_s} \quad (2)$$

where $E_{template}$ is the internal energy of the backbone and conformationally inflexible residues (the template), $E_{i_r,template}$ is the energy of interaction between residue i in rotamer r and the template, and ϵ_{i_r,j_s} is the interaction energy between rotamer r of residue i and rotamer s of rotamer j. As originally formulated,[9] the DEE algorithm stated that a rotamer i_r is incompatible with the global energy minimum structure if it satisfies the following inequality:

$$E_{i_r,template} + \sum_{j} \min_{s} \epsilon_{i_r,j_s} > E_{i_t,template} + \sum_{j} \max_{s} \epsilon_{i_t,j_s}. \quad (3)$$

$\min_{s} \epsilon_{i_r,j_s}$ is the minimum interaction energy between rotamer r of residue i with all permitted rotamers s of residue j and $\max_{s} \epsilon_{i_t,j_s}$ is the corresponding maximum value for rotamer i_t. An analogous expression was derived for rotamer pairs, and it was shown that successive application of the single and pairwise

DEE algorithms could reduce the total number of possible combinations of side chain conformations from its often very large initial value to a much smaller number of combinations.

More recently, Goldstein[11] has shown that the original DEE theorem is a special case of a more generally applicable series of inequalities that are even more effective at eliminating rotamers incompatible with the global minimum energy conformation and this work has been further extended by Desmet and colleagues.[12] We use these more general expressions in our current implementation.

The DEE algorithm is designed to identify the global minimum energy conformation (GMEC) of side chain rotamers. To calculate other properties, it is important to probe the nature of the energy surface and to determine other structures that may contribute to the partition function. To achieve this, we use a modification of the DEE algorithm together with an alternative searching method (the A* algorithm), our objective being to locate not only the global energy minimum, but also provide *all* structures within some specified energy E_{cut} of the lowest energy structure.

The first step is to modify the DEE inequalities to eliminate those side chain rotamers that are incompatible with a conformation within E_{cut} of the global energy minimum using the following trivial modification of the DEE expression:

$$\mathbf{E}_{i_r,\text{template}} + \sum_j \min_s \epsilon_{i_r,j_s} > \mathbf{E}_{i_t,\text{template}} + \sum_j \max_s \epsilon_{i_t,j_s} + \mathbf{E}_{cut}. \quad (4)$$

Having applied our modified DEE algorithm, Equation (4), we then search the remaining space (which often contains many more possible combinations of side chain rotamers than remain when using the original DEE expression) using the A* algorithm.[10,13] The A* algorithm is a method for finding the optimal ("least-cost") path from the root node to a goal node in a search tree or search graph. The algorithm uses an evaluation function termed **f***. There are two components of **f*** for any particular node **n**; the cost of reaching the node from the start node (termed **g***) and the estimated cost of reaching a goal from **n** (**h***). Thus:

$$\mathbf{f^*} = \mathbf{g^*} + \mathbf{h^*}. \quad (5)$$

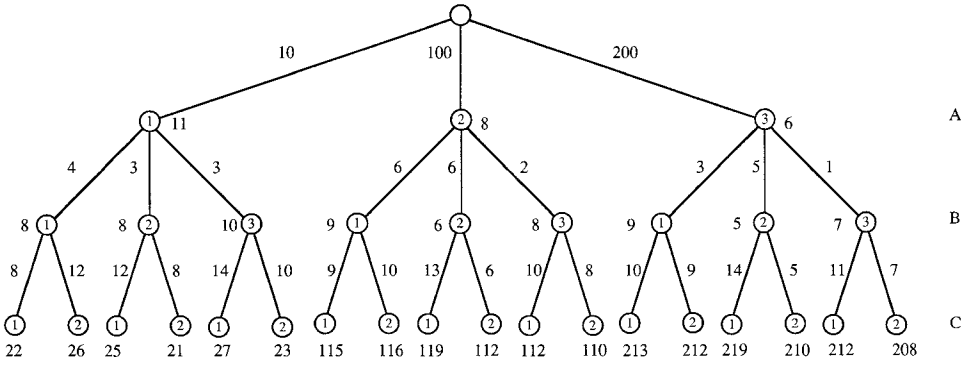
g* is the cheapest path found so far for the node. **h*** carries heuristic information, defined in a way appropriate to the problem domain. It must, however, never overestimate the cost of reaching a goal node. If **h*** were to overestimate the actual path cost then we might ignore a path that would be lower in cost than the one ultimately found. A list of nodes is stored in an array by the algorithm, ordered according to their values of **f***. At each stage, the node at the head of the list (the one with the smallest value of **f***) is expanded and new values of **f*** calculated for its successor nodes. These successor nodes are added to the list in the appropriate position. The first goal node to reach the head of the list has the minimum cost path from the root node. In the current application, each node **n** represents a partially constructed model in which some of the amino acids have been assigned a rotameric state. **g*** corresponds to the energy of this partial conformation, calculated using Equation (2). **h*** is then an estimate of the minimum energy required to complete the model, which can be obtained from the pre-calculated values of $E_{i_r,\text{template}}$ and ϵ_{i_r,j_s}. The secret to success lies in obtaining the best possible estimates of the energies and in processing the residues in the correct order, otherwise the algorithm can become bogged down in the immense space of possible solutions. To calculate the values of **h*** we consider the (N-n-1) residues that have not yet been assigned a rotameric state. For each rotamer s of these residues j the minimum value of the following function is determined:[8]

$$E_{j_s,\text{template}} + \sum_{i=1}^{n+1} \epsilon_{i_r,j_s} + \sum_{k=(n+2)}^{j-1} \min_t \epsilon_{k_t,j_s}. \quad (6)$$

In this expression the individual terms have the same meaning as before. k runs over the unassigned residues from n + 2 to j − 1. The minimum value of this function (which, it should be noted, is guaranteed to be less than the energetic penalty of completing the conformation) is added to **h***, thereby permitting the calculation of **f*** for that node.

The first goal node to reach the head of the list is the global energy minimum; the list of nodes can then be processed further in exactly the same fashion to generate a succession of conformations of increasing energy. As a simple example, consider the search tree for the tripeptide described above. We have redrawn the tripeptide search tree in Figure 2 where we have also assigned **h*** and **g*** values. The A* algorithm would operate as indicated in the figure; as can be seen, not only does the global energy minimum correspond to the first goal node to reach the head of the list, but other solutions are found in order of increasing energy.

A consequence of this ability to identify the global minimum energy combination of side chain rotamers and then, in strict order of increasing energy, all of the low-energy combinations of side chain rotamers up to the threshold energy value E_{cut} (or until the number of combinations found exceeds a user-specified maximum) is that the algorithm can provide a detailed picture of the energy surface of the side chains and the way in which that energy surface varies with the amino acid sequence and the back-



1. $A_1(21)\ A_2(108)\ A_3(206)$
2. $A_1B_2(21)\ A_1B_1(22)\ A_1B_3(22)\ A_2(108)\ A_3(206)$
3. $\mathbf{A_1B_2C_2}(21)\ A_1B_1(22)\ A_1B_3(23)\ A_1B_2C_1(25)\ A_2(108)\ A_3(206)$
4. $\mathbf{A_1B_1C_1}(22)\ A_1B_3(23)\ A_1B_2C_1(25)\ A_1B_1C_2(26)\ A_2(108)\ A_3(206)$
5. $A_1B_3(23)\ A_1B_2C_1(25)\ A_1B_1C_2(26)\ A_2(108)\ A_3(206)$
6. $\mathbf{A_1B_3C_2}(23)\ A_1B_2C_1(25)\ A_1B_1C_2(26)\ A_1B_3C_1(27)\ A_2(108)\ A_3(206)$
7. $\mathbf{A_1B_2C_1}(25)\ A_1B_1C_2(26)\ A_1B_3C_1(27)\ A_2(108)\ A_3(206)$

Fig. 2. Application of A* algorithm to the exploration of conformational space for the tripeptide. The number next to each edge is the energy associated with assigning that rotamer. The numbers next to each node are the h* estimates of the cost to reach a goal node. The order in which the nodes are expanded is shown; the conformations obtained are $A_1B_2C_2$ (energy 21), $A_1B_1C_1$ (22), $A_1B_3C_2$ (23), $A_1B_2C_1$ (25) etc. (shown in bold).

bone conformation. The algorithm provides a means to determine the conformational entropy of the side chains because it generates all states in a deterministic fashion from which a side chain conformational partition function **Z** can be directly calculated via the following expression:

$$\mathbf{Z} = \Sigma\ \mathbf{e}^{-(E_i/k_BT)}. \tag{7}$$

E_i is the energy of conformation i; k_B is Boltzmann's constant and T is the temperature. Having calculated the partition function, it is straightforward to determine p_i and thence the entropy.

Implementation and Application to Exploring the Side Chain Conformational Space of Proteins

It is clear from previous studies that the results of any side chain positioning algorithm are dependent upon the energy function used to calculate the intramolecular interactions and upon the conformational possibilities for the side chains (i.e. the rotamer library). Previous approaches to the problem have typically used either a "standard" force field to determine the energies or, more commonly, a simplified energy function that may, for example, only use a van der Waals function. A number of rotamer libraries have been published; all are based upon analyses of the conformational preferences of side chains in high-resolution protein X-ray structures but others also use energy minimization to refine these structures.

The current implementation of our conformational search protocol is based around the "Anal" module[14] of AMBER with the 1986 force field,[15,16] and one of the objectives was to assess the utility of this force field in tackling the side chain assignment problem. The energy of a given protein conformation is calculated using this force field using the following expression:

$$\mathbf{V} = \sum_{\mathbf{bonds}} \mathbf{K_r}(\mathbf{r} - \mathbf{r_{eq}})^2 + \sum_{\mathbf{angles}} \mathbf{K}_\theta(\theta - \theta_{\mathbf{eq}})^2 + \sum_{\mathbf{dihedrals}} \frac{\mathbf{V_n}}{2}[1 + \cos(\mathbf{n}\phi - \gamma)] + \sum_{\mathbf{i<j}} \left[\frac{\mathbf{A_{ij}}}{\mathbf{R_{ij}^{12}}} - \frac{\mathbf{B_{ij}}}{\mathbf{R_{ij}^{6}}} + \frac{\mathbf{q_i q_j}}{\epsilon \mathbf{R_{ij}}}\right]. \tag{8}$$

In Equation (8), r is the bond length, r_{eq} the reference bond length, K_r the bond force constant, θ the bond angle, θ_{eq} the reference bond angle, K_θ the angle bending constant, V_n the torsional potential barrier height for a torsion angle ϕ, n being the multiplicity and γ the phase factor. A and B are the Lennard-Jones parameters for two atoms i and j separated by a distance R_{ij} with q_i and q_j being their partial atomic charges and ϵ is the dielectric constant. As we keep the bonds and angles fixed at their original values, the first two terms in Equation (8) are constant for all the conformations constructed for a particular protein. It is possible within this force field to modify some of the parameters used to determine the various energy terms. These include the dielectric model (both distance-dependent and constant dielectric models are available), the relative permittivity, the non-bonded cutoff and the

TABLE I. The Proteins Considered in This Study; The Total Number of Residues for Those Proteins; the Number of Variable Residues;* the Number of Rotamer Combinations Using the Desmet and Lavery Rotamer Libraries

Name	PDB code	Number of residues	Number of variable residues	Desmet combinations	Lavery combinations
Crambin	1crn	46	26	10^{22}	10^{21}
Ribosomal protein	1ctf	68	46	10^{42}	10^{44}
Complement control protein	1hcc	59	39	10^{34}	10^{35}
Snake venom neurotoxin	1nxb	62	45	10^{43}	10^{43}
Ovomucoid third domain	2ovo	56	39	10^{35}	10^{34}
Erabutoxin B	3ebx	62	45	10^{43}	10^{43}
Bovine pancreatic trypsin inhibitor	5pti	58	36	10^{37}	10^{35}
Rubredoxin	5rxn	54	43	10^{35}	10^{35}

*Those with more than one rotamer: Ala, Gly, Pro and Cys in disulphide bonds.

representation (all atom or united atom). In this work we have used both united and all atom models, with an 8Å non-bonded cutoff. In order to investigate the influence of the electrostatic interactions we used a standard distance-dependent dielectric and also a model in which the electrostatic interactions were effectively damped to zero (achieved by setting the relative permittivity to 10^5) thus retaining just the van der Waals contributions via the Lennard-Jones equation. In both cases the dihedral angle terms were retained as in the original AMBER potential.

To investigate the effects of the rotamer library we have used two rotamer libraries; one the library of Desmet et al,[9] which is based upon the work of Ponder and Richards[17] and the other an extended version of a library published by Lavery and coworkers.[18,19] This extended library permits rotations of the hydroxyl hydrogen in serine, threonine, tyrosine, and the sulphydryl hydrogen in cysteine. We felt that this would be a more appropriate library to use with the AMBER force field which can assign quite significant partial atomic charges to polar hydrogens. Three staggered torsion angle values are permitted to the hydrogen in these cases. This multiplies the number of rotamers permitted to these residues by three over the original set. Our extended Lavery library also increases the number of rotamers available to lysine and arginine, which in the original publication have χ_4 in the extended conformation only. The number of rotamers in the Desmet and extended Lavery libraries is 216 and 182 respectively.

A set of 8 proteins was used, Table I. These were chosen to cover a range of protein fold families.[20] The conformational space of each protein was explored using our DEE/A* algorithm with the two rotamer libraries, using both united atom and all atom models, and with the "standard" and "reduced" electrostatic representations. This corresponds to a total of 8 searches per protein. For each combination of force field model and rotamer library we systematically generated conformations and accumulated the side chain conformational entropy. Conformations were generated until one of the following four criteria was satisfied: either the number of conformations exceeded a preset value, or the energy of the node at the head of the A* list exceeded by more than E_{cut} the energy of the GMEC, or the incremental change in TS (T = 300 K) between successive conformations fell below a specified threshold, or the ordered array used by the A* algorithm filled up. In this work we used a limit of 2.5 million conformations, an entropy increment of 10^{-12} kcal/mol and the A* array was dimensioned to hold approximately 250,000 nodes. The objective was to find a value of E_{cut} for each of the 64 conformational searches which was not so high that the A* array filled up before any conformations were generated but at the same time not too low that the search terminated before the entropy difference criterion was met. For each search we initially employed an E_{cut} value of 5 kcal/mol. In some cases this was an appropriate value to enable the search to be terminated by the entropy increment criterion. In other cases the entropy did not converge and so the energy limit was increased in an attempt to meet this criterion or until the value of E_{cut} gave such a large search space that no conformations could be found before the array used to store them in the A* algorithm became full. In other cases the A* array filled up at the 5 kcal/mol level without any conformations being generated and so E_{cut} was systematically reduced until the conformational search did generate conformations. Thus each of the conformational searches was terminated either because the entropy criterion was met or because no conformations remained with an energy lower than E_{cut}. In all cases the search terminated due to one of these two criteria and before the 2.5 million conformation limit was reached.

All conformations were compared to the conformation assigned in the deposited X-ray structure. RMS deviations (in Å) were calculated (non-hydrogen atoms not including C_β and taking account of symmetry in residues such as Phe and Asp). The proportion of chi torsions within 40° of the value in the X-ray structure was also determined. In addition we identi-

TABLE II. Summary of Results Comparing the Conformations Generated by the Conformational Search to the X-ray Structure[†]

	RMS[a] of GMEC	Lowest RMS for all confs	Best possible RMS for this library	Dihedral percentage[b]	Lowest percentage for all structures
All atom normal electrostatics Lavery library	1.92 (1.74–2.33)	1.73 (1.39–2.03)	0.83	65 (49–79)	70 (43–85)
United atom normal electrostatics Lavery library	1.84 (1.48–2.11)	1.67 (1.40–2.17)	0.83	68 (45–81)	72 (46–83)
All atom reduced electrostatics Lavery library	1.97 (1.66–2.04)	1.79 (1.60–2.00)	0.83	67 (44–83)	71 (46–89)
United atom reduced electrostatics Lavery library	1.83 (1.48–2.11)	1.66 (1.31–1.92)	0.83	71 (46–87)	74 (49–89)
All atom normal electrostatics Desmet library	1.84 (1.24–2.16)	1.62 (0.96–2.07)	0.75	66 (47–83)	73 (49–94)
United atom normal electrostatics Desmet library	1.76 (1.18–2.16)	1.55 (0.69–2.01)	0.75	69 (45–85)	75 (49–98)
All atom reduced electrostatics Desmet library	1.72 (1.11–2.13)	1.59 (1.04–2.08)	0.75	71 (51–91)	75 (53–94)
United atom reduced electrostatics Desmet library	1.77 (1.26–2.26)	1.51 (1.09–2.12)	0.75	73 (48–87)	76 (52–91)

[†]Where figures are given in brackets these represent the minimum and maximum values obtained.
[a]RMS values are calculated for all non-hydrogen side chain atoms (excluding C_β), taking account of symmetry where relevant.
[b]Dihedral percentage refers to the number of side chain torsion angles that are within 40° of the value in the X-ray structure.

fied the "best" value for each criterion from the conformations generated in each case and, in the case of the atomic RMS values, the lowest value that could be obtained for that particular library (i.e. where each residue is permitted to adopt the closest rotamer irrespective of the energy of the resulting structure). Summary statistics for these data are presented in Table II.

The number of conformations generated for each conformational search is reported in Table III together with the E_{cut} value eventually employed. As indicated above, we accumulated the side chain conformational entropy during the search and thus a value of TS (at 300 K). These TS values correspond to the difference in side chain conformational entropy between our conformational searches and the result that would arise from assuming that there is only a single conformational state for the folded protein and are also reported in Table III. We also indicate whether the conformational search was terminated because the difference in TS between successive conformations fell below 10^{-12} kcal/mol or whether the search terminated because no conformations remained at that energy threshold.

We performed a detailed analysis of each conformation generated to determine its relationship to the lower energy conformations previously obtained. This analysis was designed to identify those higher energy structures which involved two or more residues simultaneously changing their conformation, rather than being due to the consequence of a combination of changes in two (or more) side chains acting independently. Each conformation was compared with all previously generated conformations to determine how many rotamers were different. This series of calculations was restricted to the first 25,000 conformations as the time taken for the analysis scales with the square of the number of conformations. Particular attention was paid to the number of occasions on which two or more new rotamers simultaneously appeared, indicating cooperativity between the rotamers.

To illustrate this aspect of the analysis, consider our earlier tripeptide example, Figure 2.

Here, the second conformation obtained, $A_1B_1C_1$ differs from the global minimum ($A_1B_2C_2$) by the simultaneous change of residue B from rotamer B_2 to rotamer B_1 and residue C from rotamer C_2 to rotamer C_1 and so this conformation would be flagged as having arisen from the concerted change in conformation of two residues. Conformation 3 ($A_1B_3C_2$; energy 23 units) differs from the global minimum by just one residue (B changes from B_1 to B_3) which would be flagged as merely a single residue change. The fourth conformation, $A_1B_2C_1$ (energy 25 units) differs by two rotamers from the GMEC but by only a single

TABLE III. Number of Conformations Generated, TS Values† and E_{cut} Value Employed

	1crn	1ctf	1hcc	1nxb	2ovo	3ebx	5pti	5rxn
All atom	33705*	35*	63507*	166758	84257	65454*	43504*	31581*
normal electrostatics	3.5	1.5	6.3	6.0	4.4	6.4	3.6	4.0
Desmet library	8.0	2.065	3.0	8.0	8.0	3.0	8.0	7.0
All atom	378836	380328	399358	305468	353638	353601	514142	68047*
reduced electrostatics	7.5	7.6	7.7	7.5	7.6	7.6	7.8	6.6
Desmet library	5.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
United atom	23128*	6129*	141417*	130109	148176	31380*	9364*	57004*
normal electrostatics	3.7	3.9	6.0	5.4	5.3	5.8	2.7	4.4
Desmet library	7.0	5.0	5.0	8.0	8.0	3.0	8.0	7.0
United atom	247322	75621*	548610	310789	460102	459244	317550	386867
reduced electrostatics	7.2	6.6	7.8	7.5	7.7	7.8	6.9	7.3
Desmet library	5.0	2.0	5.0	5.0	2.0	3.0	5.0	5.0
All atom	30390	15111*	105628*	186715	185869	290251	56470	24195
normal electrostatics	3.6	4.6	5.0	6.4	6.0	6.5	4.4	3.9
Lavery library	8.0	5.0	6.5	8.0	6.5	8.0	8.0	8.0
All atom	392027	10298*	403447	482921	374828	155795*	259869*	108792*
reduced electrostatics	7.46	5.5	7.7	7.8	7.7	7.1	7.3	6.8
Lavery library	3.0	0.925	1.5	5.0	1.5	0.4	2.0	2.0
United atom	30156*	1141*	212282	131977	134765	32050*	17663	47735
normal electrostatics	3.5	3.3	5.8	5.2	5.6	5.4	3.5	4.0
Lavery library	8.06	4.00	8.0	8.0	6.5	4.75	8.0	8.0
United atom	405091	381983	509119	472172	318810	333409	333781	297739
reduced electrostatics	7.36	7.4	7.7	7.8	7.5	7.6	7.2	7.3
Lavery library	5.06	5.0	3.0	3.0	5.0	3.0	5.0	5.0

*Indicates that the search was terminated because no more conformations with an energy less than E_{cut} were present. Lack of an asterisk indicates that the search terminated when the difference in TS between two successive conformations fell below 10^{-12} kcal/mol (see text).
†Which correspond to the difference in entropy between value here and value assuming only one conformation is accessible.

rotamer from conformation number 2 and so would be classified as a single change. In Table IV we indicate how many conformations differ by a change in two, three etc. rotamers over the conformational space examined.

We show in Table V the number of residues in each search for which more than one rotamer was used during the conformational search, the total number of rotamers that were observed, and the number of rotamers that correspond merely to a change in position of a hydrogen atom.

RESULTS AND DISCUSSION

The results displayed in Table 2 are typical of published side chain placement algorithms using current rotamer libraries. These results are for all residues. In these relatively small proteins the proportion of surface residues is rather high. If the RMS deviation for "buried" residues alone in the predicted GMEC is determined (calculated for the protein X-ray structure using the Shrake and Rupley algorithm[21] and taking 25% accessibility as the cutoff value) then the average values are 1.0 Å for the Lavery library and 0.9 Å for the Desmet library. These values are also consistent with other side chain prediction methods and with the observation that the buried side chains in protein X-ray structures tend to be more consistent with the distributions present in rotamer libraries. There is little to choose between the force field models and the rotamer libraries in terms of their ability to correctly predict the X-ray structure, though the more comprehensive Desmet library does appear to be marginally better.

The average accessibility of these variable residues is rather high; for all combinations of rotamer library and force field the average varied between 52% and 57%. As all conformationally flexible residues are permitted to vary in these searches this figure is not surprising; one would expect the surface residues to be more flexible than the buried residues. Moreover, the surfaces of proteins often have a relatively higher concentration of conformationally flexible residues such as lysine and arginine whereas the bulkier, conformationally less flexible aromatic residues are often found in protein cores. A study of the conformational space of protein cores alone will be reported elsewhere.

The results presented in Table III suggest that the number of conformations obtained with a given energy cutoff is typically greater for the searches using a model with much reduced electrostatics than

TABLE IV. Number of Occasions (Figure in Brackets) on Which 2, 3, or 4 Residues Simultaneously Changed Conformation

	1crn	1ctf	1hcc	1nxb	2ovo	3ebx	5pti	5rxn
All atom	2 (1)	2 (1)	2 (4)	3 (1)	2 (11)	2 (3)	2 (3)	2 (3)
normal electrostatics			3 (1)					
Desmet library								
All atom		2 (4)	2 (1)				2 (1)	
reduced electrostatics								
Desmet library								
United atom	2 (1)	2 (3)	2 (13)	2 (1)	2 (4)	3 (1)	2 (1)	
normal electrostatics				3 (1)				
Desmet library								
United atom		2 (1)	3 (1)					
reduced electrostatics								
Desmet library								
All atom	2 (1)	2 (1)	2 (1)		2 (14)	2 (1)	2 (2)	2 (1)
normal electrostatics		3 (2)						
Lavery library		4 (1)						
All atom		2 (2)	2 (1)		2 (1)		2 (1)	
reduced electrostatics		4 (1)						
Lavery library								
United atom	2 (1)		2 (8)		2 (3)	2 (3)		2 (2)
normal electrostatics			3 (1)					3 (1)
Lavery library								
United atom		2 (1)	2 (1)	3 (1)			2 (1)	
reduced electrostatics								
Lavery library								

TABLE V. The Number of Residues for Which More Than One Rotamer was Observed During the Conformational Search, the Total Number of Rotamers Used and the Number of These Rotamers That Correspond to Rotation of a Hydrogen Atom

	1crn	1ctf	1hcc	1nxb	2ovo	3ebx	5pti	5rxn
All atom	20	7	20	21	26	24	22	24
normal electrostatics	58	14	39	48	74	54	50	49
Desmet library	26	2	10	28	24	24	11	16
All atom	12	22	15	15	20	17	20	22
reduced electrostatics	27	59	38	32	44	41	37	44
Desmet library	18	9	18	22	22	28	10	9
United atom	21	20	24	27	25	18	23	24
normal electrostatics	56	44	55	60	59	42	52	47
Desmet library	25	6	19	29	25	22	11	15
United atom	13	20	16	14	15	15	16	23
reduced electrostatics	28	40	46	32	34	38	36	32
Desmet library	19	9	23	23	23	30	13	9
All atom	22	14	17	17	24	25	23	24
normal electrostatics	57	48	53	36	67	61	53	50
Lavery library	26	4	15	22	21	28	8	17
All atom	16	24	15	19	23	18	21	20
reduced electrostatics	31	57	34	38	42	45	43	47
Lavery library	19	5	18	26	21	30	8	14
United atom	22	14	23	20	23	25	22	24
normal electrostatics	51	26	57	43	56	71	55	50
Lavery library	24	3	15	24	21	36	10	11
United atom	14	20	17	14	16	18	20	16
reduced electrostatics	31	44	44	37	29	44	45	30
Lavery library	18	9	23	27	19	32	11	13

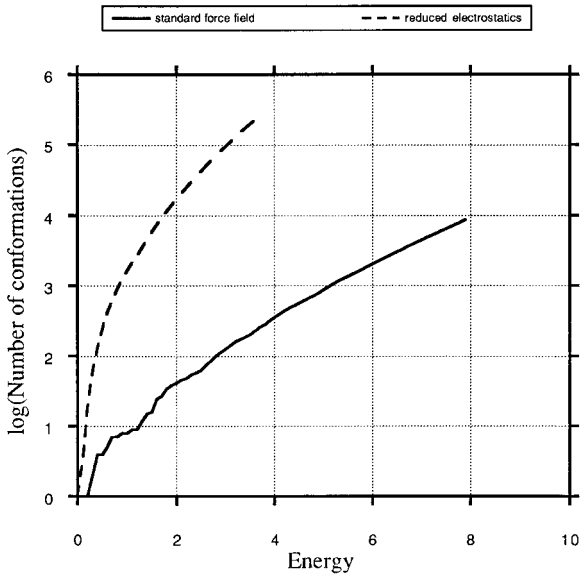


Fig. 3. Graph showing the number of conformations generated as a function of their energy (in kcal/mol) above the global minumum energy combination of rotamers for 5pti using the Desmet library with a united-atom force field.

for a "standard" force field. This is not unexpected; electrostatic interactions play a significant role in a force field such as AMBER and accentuate the energy differences between different conformations. This is illustrated in Figure 3 where we plot the logarithm of the number of conformations as a function of the energy above the GMEC for 5pti using the Desmet library with a united-atom force field. In this particular case 9364 conformations were generated using the standard force field and an energy threshold of 8 kcal/mol. More than thirty times more conformations (317550) were generated for the reduced electrostatics model with a smaller energy threshold of 5 kcal/mol. Indeed, in the latter case the search terminated due to the entropy criterion being satisfied when the conformational energy was 3.8 kcal/mol.

Our results in Table IV show that higher energy conformations are overwhelmingly due to independent changes in rotamers, for this force field and these rotamer libraries at least. Thus the number of occasions when two or more residues simultaneously change their conformation, in a concerted manner, is extremely small. Most of the conformations are related to previously generated structures by a change in just a single rotamer. Nevertheless, there are a number of occasions when concerted changes do occur. Of the proteins we examined, the largest number of simultaneous rotamer changes was 4. This occurred for 1ctf using the Lavery rotamer library with an all-atom force field and is illustrated in Figure 4a. Two different conformational arrangements of these residues have nearly the same interaction energy. However, one should be careful of over-interpreting this result as all of these residues are solvent-exposed for which both the force-field model and the rotamer library would be expected to be most seriously inadequate. Moreover, in this particular case two of the rotamers for Glu64 adopt very similar torsion angles and should probably be considered the "same" rotamer (with the extensions noted above we used the libraries unchanged, as published). Visual inspection of the pairs of conformations that differed by three rotamers revealed that most were similarly due to concerted movements between spatially close residues. A second example is that of 1hcc (Figure 4b) where the interplay between Leu48, Lys51 and Ser53 involves both steric and electrostatic components. In some cases the concerted changes in residue conformation involved residues between which there is no such close interaction. Two examples are 1nxb and 5rxn (Figures 4c and 4d) where there are occurrences of simultaneous changes in conformation involving three residues. In these cases there is clearly a fine balance involving the energetics of interaction, at least with the current force field model. It is worth noting that the number of such concerted changes in residue conformations is larger for the "normal" force field than for the energy model with reduced electrostatics. This is not surprising given the greater modulation of the energy surface with the "normal" force field. Table V suggests that a large proportion of the residues that are potentially variable do vary their conformation. Moreover, these variations in conformation are by no means restricted to changes in the location of hydrogen atoms, though these are a major contribution in some cases.

The entropy calculations summarized in Table III suggest that the conformational freedom of the folded protein certainly plays a role in the overall free energy for the change from the unfolded to the folded state. Nevertheless, despite the fact that large number of distinct conformations were obtained in many cases, the free energy difference due to this source is relatively small compared to the other contributions to the folding free energy. For example, Pickett and Sternberg suggest that the entropy loss associated with a single arginine residue is 2.03 kcal/mol from the unfolded to the folded state and the total entropy loss for the proteins examined in this paper would be of the order of 50 kcal/mol.[2] Koehl and Delarue have also determined the entropy of the folded protein using their mean field method.[4] They employ the Lavery et. al. rotamer library with a potential energy function that includes just Lennard-Jones terms (no electrostatics). For 13 proteins with fewer than 100 amino acid residues their most reliable calculations for the entropy give values from 6.3 kcal/mol to 17.2 kcal/mol. These entropy values were derived using Equations (1) and (7) from a systematic generation of conformations based on the final probabilities in

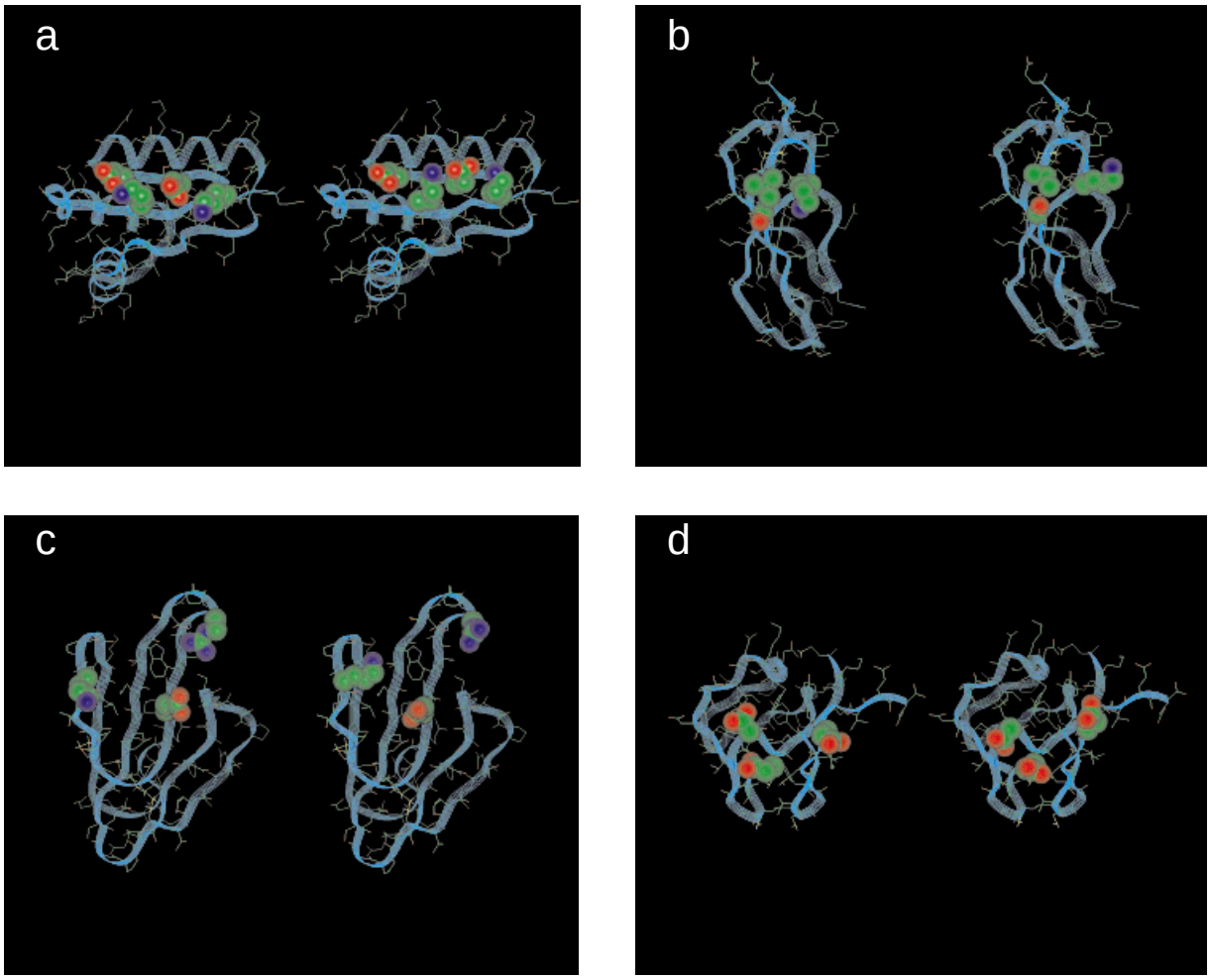


Fig. 4. Graphical representations of the simultaneous changes in conformation of side chains for **(a)** 1ctf (Lys7, Glu64, Glu66 and Lys68), **(b)** 1hcc (Leu48, Lys51, Ser53), **(c)** 1nxb (Arg33, Glu38, Lys47), and **(d)** 5rxn (Asp47, Glu48, Glu50). In each case the lower energy structure is on the left.

their conformational probability matrix. These values are generally comparable to the results we have obtained (Table III). The major differences between their approach and ours is in the search philosophy; ours is deterministic (i.e. the global energy minimum is guaranteed to be the first structure generated, with other structures being generated in strict order of energy). As indicated above, the approach of Koehl and Delarue is probabilistic, with the final result depending upon the initial probabilities assigned to the rotamers (though it was shown that reproducible convergence was obtained using the typical operation of the program).

One possible source of concern with the entropy calculations is that the values in Table III are not converged, even when a very low entropy difference between successive conformations was required for termination of the search (10^{-12} kcal/mol). To investigate this further we plotted the variation in entropy as a function of the energy above the global minimum for each of the searches. These results are shown in Figure 5 from which it can be seen that in some cases at least the entropy has converged. This is generally the case for those searches using the "standard" force field where we were able to use a relatively high E_{cut} threshold of at least 7 kcal/mol. By contrast, entropy convergence was never achieved with the "reduced electrostatics" model due to the density of states effect; a large number of conformations were found at relatively low energies (less than 2 kcal/mol). To investigate this further, we generated all conformations for two proteins (1crn and 5rxn) using a united atom, reduced electrostatics model and both the Desmet and Lavery libraries with an E_{cut} threshold of 6 kcal/mol for 1crn and 5 kcal/mol for 5rxn. As shown in Figure 6 entropy convergence does now appear to have been achieved in all four searches, though this has only been achieved at the

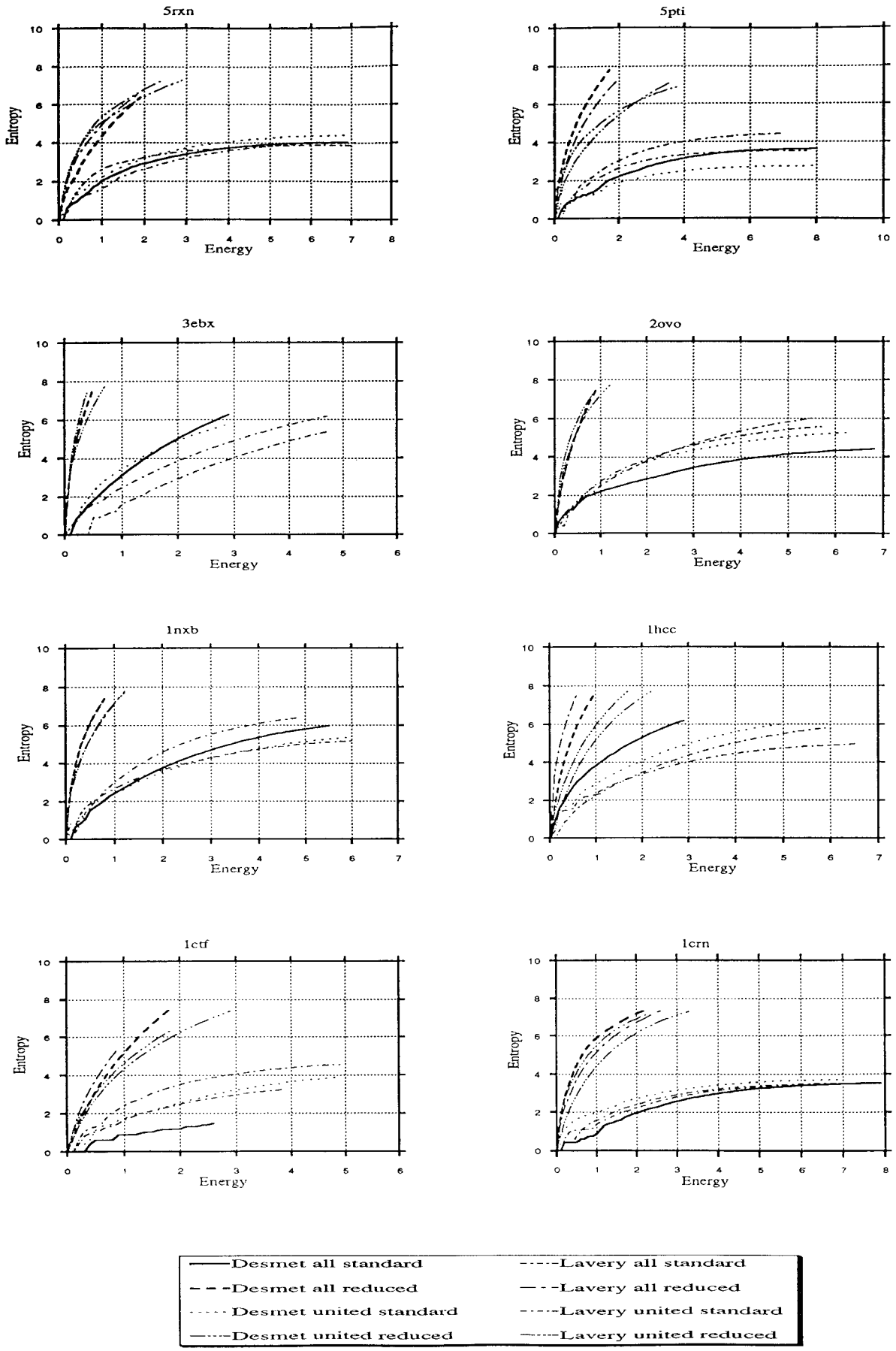


Fig. 5. Variation in entropy as a function of the energy above the global minimum energy conformation for each of the 8 different search protocols for the eight different proteins. ''Standard'' and ''reduced'' refer to the electronics model.

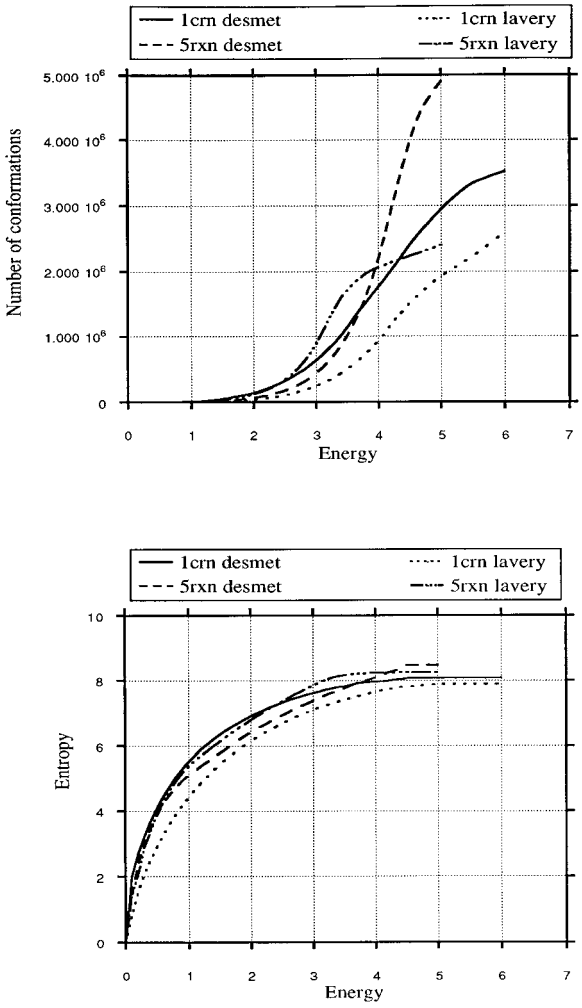


Fig. 6. Variation in the number of conformations and the entropy as a function of energy above the GMEC (in kcal/mol) for 1crn and 5rxn using a united atom, reduced electrostatics model for both the Desment and Lavery libraries.

cost of generating up to 5 million conformations. While it is difficult to draw any general conclusions, the side chain conformational entropy does appear to converge to values between approximately 3.5 and 10 kcal/mol for this size of protein. This range is at the lower end of the values calculated by Koehl and Delarue. Their use of a van der Waals energy model (no electrostatics) together with the assumptions inherent in the mean-field approach may in part explain these differences, though of course there are many other possible reasons.

The size of the conformational space displayed by these proteins can be measured by the number of possible combinations of rotamers that are present. Over both the Desmet and Lavery libraries the average number of possible combinations is 10^{36} (see Table I). Also of interest is the effectiveness of the dead-end elimination algorithm in reducing this initial space and the consequent size of the search space examined by the A* algorithm. The average here is $10^{14.5}$ with a minimum of 10^{10} and a maximum of 10^{21}. It should also be noted that a large number of the rotamers eliminated by the dead-end elimination step are of very high energy (due to clashes with the backbone etc.) and so the space that is explored by the A* algorithm consists very largely of rotamers which are all quite "reasonable" in terms of their potential for inclusion in a final structure. Here again, the effectiveness of the A* algorithm is dependent upon having as good estimates of the cost to reach a goal node as possible. This can be difficult to achieve because interactions between residues not yet assigned cannot be easily computed: the rotamer with the lowest template energy may interact unfavorably with all rotamers of another residue. The expression we currently employ is shown in Equation (6). This is guaranteed never to exceed the true cost and so does meet this requirement for successful operation of the A* algorithm. However, one consequence of using Equation (6) is that the residues must always be considered in the same order during the A* search, else the estimated cost can become greater than the true cost. The order is important because, if poorly defined, a large number of nodes with very similar values of **f*** are generated early in the search. If this occurs more nodes must be expanded before the optimal solution appears at the head of the list of nodes. We currently choose the residue order by computing for each rotamer of every residue the following function:

$$v_{i_r} = E_{i_r,\text{template}} + \sum_{j \neq i} \min \epsilon_{i_r,j_s}. \qquad (9)$$

The two lowest values for each residue are identified and their difference calculated. The residue with the greatest difference is the one expanded first, the residue with the second greatest difference is expanded second, and so on. Thus we try to expand early on those residues for which it is most likely that there will be a single rotamer that is preferred to all others, and thereby probe deeply into the search tree before encountering nodes which have similar **f*** values. For the systems we have examined thus far, we find this method of choosing the nodes combined with Equation (9) gives a satisfactory performance although it is possible that other ways of determining the **h*** values might be more efficient for specific cases. Moreover, in some cases, the algorithm is unable to distinguish between the large number of possibilities and the ordered array used by the A* algorithm fills up. This problem also often arises for larger proteins (using this force field and these rotamer libraries, at least) making it necessary to remove some of the residues from consideration. Nevertheless, to the best of our knowledge no other method is able to explore such a large conforma-

tional space in what amounts to a systematic fashion in order to identify a protein's low energy conformations.

CONCLUSIONS

In this paper we have described the extension and application of an approach originally introduced for performing molecular docking with flexibility of the protein to the problem of exploring the side chain conformational space of whole proteins. The two key features of our algorithm are that it is deterministic—provided it finds a solution, it is guaranteed to be the "correct" answer—and that it generates a number of solutions in strict order. Other approaches to this problem are either probabilistic, with no guarantee of finding the "ideal" solution or just locate the single global minimum energy structure. Notwithstanding these key attributes of our approach, it is important to recognize that in our current implementation it is generally restricted to relatively small proteins, with the current rotamer libraries. We have investigated various combinations of rotamer library and implementations of the AMBER potential energy function for predicting the conformations of side chains in proteins, obtaining results that are comparable to other methods. Our investigations of the relationships between the various conformations suggest that the side chains largely act independently, with relatively few concerted changes in conformation, at least for the conformations that would contribute significantly to the partition function. Our approach enables us to quantify the entropy of a folded protein and to suggest the magnitude of the error associated with assuming that this quantity is zero. There are significant differences in the density of conformational states for the side chains of proteins between a "standard" force field model and one in which the electrostatic component has been largely removed; this is reflected in lower converged entropy values when using the former energy model.

REFERENCES

1. Doig, A.J, Sternberg, M.J.E. Side chain conformational entropy in protein folding. Protein Sci. 4:2247–2251, 1995.
2. Pickett, S.D., Sternberg, M.J.E. Empirical scale of side chain conformational entropy in protein folding. J. Mol. Biol. 231:825–839, 1993.
3. Abagyan, R., Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235:983–1002, 1994.
4. Koehl, P., Delarue, M. Application of a self-consistent mean field theory to predict protein side chains conformation and estimate their conformational entropy. J. Mol. Biol. 239: 249–275, 1994.
5. Wang, J., Purisima, E.O. Analysis of thermodynamic determinants in helix propensities of non-polar amino acids through a novel free-energy calculation. J. Am. Chem. Soc. 118:995–1001, 1996.
6. Creamer, T.P., Rose, G.D. Side chain entropy opposes α-helical formation but rationalizes experimentally determined helix-forming propensities. Proc. Natl. Acad. Sci. USA 89:5937–5941, 1992.
7. Creamer, T.P., Rose, G.D. α-Helix-forming propensities in peptides and proteins. Proteins 19: 85–97, 1994.
8. Leach, A.R. Ligand docking to proteins with discrete side chain flexibility. J. Mol. Biol. 235:345–356, 1994.
9. Desmet, J., DeMaeyer, M., Hazes, B., Lasters, I. The dead-end elimination theorem and its use in protein side chain positioning. Nature 356:539–542, 1992.
10. Hart, P.E., Nilsson N.J., Raphael, B. A formal basis for the heuristic determination of minimum cost paths. IEEE Trans on SSC 4:100–114, 1968.
11. Goldstein R.F. Efficient rotamer elimination applied to protein side chains and related spin glasses. Biophys. J. 66:1335–1340, 1994.
12. Lasters, I., De Maeyer, M., Desmet, J. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. Protein Eng. 8:815–822, 1995.
13. Nilsson, N.J. "Principles of Artificial Intelligence," New York: Springer-Verlag 1982:74–88.
14. Pearlman, D.A., Case, D.A., Caldwell, J.C. et al. Amber 4.0, San Francisco:University of California, 1991.
15. Weiner, S.J., Kollman, P.A., Case, D.A. et al. A new force-field for molecular mechanical simulation of nucleic-acids and proteins. J. Am. Chem. Soc. 106:765–784, 1984.
16. Weiner, S.J., Kollman, P.A., Nguyen, D.T., Case, D.A. An all atom force-field for simulations of proteins and nucleic acids. J. Comp. Chem. 7:230–252, 1986.
17. Ponder, J.W., Richards, F.M. Tertiary templates for proteins—use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
18. Tuffery, P., Etchebest, C., Hazout, S., Lavery, R. A new approach to the rapid determination of protein side chain conformations. J. Biomol. Struct. Dyn. 8:1267–1289, 1991.
19. Tuffery, P., Lavery, R. Packing and recognition of protein structural elements: A new approach applied to the 4-helix bundle of myohemerythin. Proteins 15:413–425, 1993.
20. Orengo, C.A., Flores, T.P., Taylor W.R., Thornton J.M. Identification and classification of protein fold families. Protein Eng. 6:485–500, 1993.
21. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms. lysozyme and insulin. J. Mol. Biol. 79:351–371, 1973.