Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER) as a New Method for Protein Structure Optimization

Johan Desmet,[1*] **Jan Spriet,**[2] **and Ignace Lasters**[1]
[1]*AlgoNomics NV, Gent-Zwijnaarde, Belgium*
[2]*Interdisciplinary Research Centre, K.U. Leuven Campus Kortrijk, Universitaire Campus, Kortrijk, Belgium*

ABSTRACT **We have developed an original method for global optimization of protein side-chain conformations, called the Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER) method. The method operates by systematically overcoming local minima of increasing order. Comparison of the FASTER results with those of the dead-end elimination (DEE) algorithm showed that both methods produce nearly identical results, but the FASTER algorithm is 100–1000 times faster than the DEE method and scales in a stable and favorable way as a function of protein size. We also show that low-order local minima may be almost as accurate as the global minimum when evaluated against experimentally determined structures. In addition, the new algorithm provides significant information about the conformational flexibility of individual side-chains. We observed that strictly rigid side-chains are concentrated mainly in the core of the protein, whereas highly flexible side-chains are found almost exclusively among solvent-oriented residues. Proteins 2002;48:31–43.**

Key words: structure prediction; side-chain conformation; side-chain flexibility; dead-end elimination; combinatorial problem; local minima

INTRODUCTION

Side-chain structure prediction is an essential component of protein modeling. The determination of optimal side-chain conformations given a particular protein backbone structure (or template) emerges in many subdisciplines of protein structure prediction: homology modeling, protein design, loop structure prediction, peptide docking, and so forth. During the last 15 years, considerable progress has been made in this area, resulting in a variety of different methods.[1] This may lead to the impression that the side-chain packing problem is now greatly solved. On the other hand, it is also generally recognized that efforts are needed to further improve the quality of the predictions, especially by taking into account some small-scale flexibility of the backbone[1] or the side-chains.[2]

Over the last few years, some consensus has grown that the methods with the highest intrinsic accuracy are empowered by the dead-end elimination (DEE) principle.[3] DEE-based methods eliminate side-chain conformations that cannot be part of the global minimum energy conformation (GMEC), so that, upon convergence, they are bound to yield the GMEC. Compared with the original DEE method,[4] numerous improvements have been made extending the limits of DEE,[5–10] but a major problem remains the unfavorable scaling of the computational time with the size of the proteins studied.[3] The main reason for this is a fundamental one: the simultaneous placement of multiple, clustered side-chains essentially forms a combinatorial problem that is nontrivial to solve[11] and has been found to be NP-complex.[12]

The problem of computational performance becomes even more important in view of some recent trends in protein modeling. First, the field of protein design is rapidly gaining attention. Typically, in a design application, many different sequences have to be modeled onto a given structure, either by systematically threading each of them or by including sequence variation in the search process.[10,13,14] Second, side-chain placement algorithms are often plugged into high-level packages for complex applications such as loop structure prediction,[15] ab initio folding,[16] peptide docking,[17] or automated fitting of electron density maps.[18] Typically, in such environments, a side-chain placement module is repeatedly addressed to model residues onto different known or computer-generated main-chain structures. Such applications are extremely demanding because of the huge search space. Therefore, they often necessitate drastic simplifications, either in the number of residues modeled simultaneously, the number of residue types considered, the conforma-

Abbreviations: PDB, Protein Data Bank; RMSD, root-mean-square deviation; ASA, accessible surface area; DEE, dead-end elimination; GMEC, global minimum energy conformation; BMEC, backbone-determined minimum energy conformation; IMEC, intermediate minimum energy conformation; iBR, iterative batch relaxation; ciBR, conditional iterative batch relaxation; sPR, single-residue perturbation/relaxation; dPR, double-residue perturbation/relaxation

Grant sponsor: Vlaams Instituut voor de bevordering van het Wetenschappelijk-Technologisch onderzoek in de industrie; Grant number: 990255.

*Correspondence to: Dr. Johan Desmet, AlgoNomics NV, Technologiepark 4, B-9052 Gent, Belgium. E-mail: Johan.Desmet@algonomics.com

Received 31 October 2001; Accepted 31 October 2001

tional freedom allowed (e.g., the number of rotamers considered), the level of detail of the energy function, algorithm-specific execution parameters, or the search process itself.

A surprising observation has been made by Eisenmenger et al.[19] who found that most side-chains can be correctly positioned by taking into account only the interactions with the main-chain, thereby completely ignoring pair interactions. When applying a simple template-based method to a test set of six proteins, they found that, on average, 53% of all side-chains had dihedral angles that were in agreement with the known structure. For the buried side-groups (<25% exposed surface), the score increased to 74%. When each side-chain was modeled in the presence of the complete, known structure, the average prediction score increased by only about 10% (65% for all and 84% for the buried residues). From these observations the authors concluded that the combinatorial barrier in side-chain placement hardly exists. This conclusion has been supported later by the success of relatively simple modeling methods that sample the conformational space, rather than pruning and enumerating it. Examples of this type are Monte Carlo methods,[20–23] genetic algorithms,[24,25] mean-field optimisation,[26,27] and local optimization.[28–30] On the other hand, a recent study by Voigt et al.[3] has shown that such methods invariably yield semirelaxed structures that are higher in energy than the GMEC.

The discussion about the combinatorial nature of side-chain placement is a delicate matter because it not only depends on important execution variables (the force field, rotamer library, test set, parameters) but especially on the way it is approached: either from a theoretical or practical point of view. In the former case, predicted structures are evaluated by an internal criterion (usually an energy-based scoring function), whereas in the latter case the predictions are evaluated by an external criterion (e.g., by comparison with experimentally determined structures or experimental properties such as stability or function).

Several studies have focused on the optimization of the correlation between internal and external scoring functions (for a review, see Gordon et al.[31]), but the focus of the present work is different. The main question addressed here is: given a state-of-the-art internal, energy-based scoring function, a documented rotamer library and a high-quality protein test set, to what extent should the combinatorial aspect of side-chain placement be taken into account to obtain reasonable, good, or optimal predictions? The latter is examined both from a theoretical and from a practical point of view. An important, related question is: can one identify a method that deals with the combinatorial problem in such a way that the predictions are (near-)optimal in a theoretical and practical sense, while it remains fast enough to be applied in high-throughput mode?

The FASTER method presented here has been based on the hypothesis that the DEE method is overly cautious in the elimination of rotamers, whereas other methods may be too crude for the combinatorial nature of the problem. Therefore, we developed a completely new method that directly focuses on the problem of local minima. The FASTER method in its current state is an iterative method that "explores the energy landscape" and systematically attempts to surmount local minima of increasing order (as defined below). The landscape is formed by an energetic scoring function defined on a set of position-specific side-chain rotamers, interacting with the template (the fixed part of a protein structure) as well as with other side-chains in a particular rotameric state. In the current version, the energy landscape is treaded in a downhill direction until it is found that the system resides in a third or higher order local minimum.

The FASTER method has been developed with a focus on solving the combinatorial side-chain packing problem to such an extent that its results are statistically equivalent and biologically indistinguishable from those obtained with the DEE reference method.[4,6,7,32] Therefore, it guarantees an identical or equivalent prediction accuracy as can be reached with the DEE method. The main advantage of the new method is its unprecedented computational speed so that it is ideally suited either for handling very large systems or for incorporation into high-level modeling applications. Helpful in this respect is that the basic algorithm is much simpler to implement than many other methods, including DEE. Another advantage is the fact that the FASTER algorithm always converges (which is not guaranteed for DEE) and that multiple near-optimal solutions can be retrieved at a negligible extra computational cost. In view of the latter property, the FASTER method can also be applied as a tool to study the intrinsic conformational flexibility of protein side-chains.

We first describe the method from a theoretic and algorithmic point of view. Next, we compare the results obtained for 33 highly resolved protein structures where for each structure all side-chains have been placed back onto their native backbone, both by the DEE and the FASTER method. The comparison is done in four different ways: (i) by computing the percentage of FASTER-modeled residues adopting exactly their GMEC rotamer (as identified by DEE), (ii) by comparing the total energies of FASTER- and DEE-modeled structures, (iii) by comparing the computational performance of the two methods, and (iv) by comparing both the FASTER- and DEE-modeled structures with the known X-ray determined structures. Finally, we illustrate the applicability of FASTER as a tool to study the intrinsic conformational flexibility of side-chains.

MATERIALS AND METHODS

Theory

A potential practical weakness of the DEE method is that it is an elimination method. This algorithm starts by assigning a set of amino acid type-specific rotamers for each modeled residue position. All rotamers are then screened for potential elimination on the basis of various DEE criteria. If this procedure leads to a unique structure (one rotamer for each residue), the latter corresponds to the GMEC. However, convergence is not guaranteed, and therefore, the basic DEE method must be combined with end-stage routines.[7,8,33] This has urged us to look for deterministic rather than eliminative criteria.

The total potential energy, E_{tot}, is a frequently used scoring function to assess the global fitness of a particular protein structure. For a protein comprising N rotatable side-chains, the total energy can be expressed as

$$E_{tot} = E_{template} + \sum_{i=1}^{N} E(r_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E(r_i,r_j) \qquad (1)$$

In this equation $E_{template}$ is the template self-energy, $E(r_i)$ is the self-energy of residue i in rotameric state r_i plus its interaction with the template, and $E(r_i,r_j)$ denotes the nonbonded interaction between two side-chain rotamers r_i and r_j. This equation shows that the total energy can be calculated as the sum of a constant term, a set of individual side-chain contributions, and a large number of pair-wise interactions.

When the template is kept fixed and pair-wise contributions are ignored, the problem of finding the optimal set of side-chain conformations $\{r_i^{min}\}$ reduces to the simple selection of the rotamer with the lowest value for $E(r_i)$, for all residues i.

$$r_i^{min} = \arg[\min_{r_i} E(r_i)] \qquad (2)$$

Structures determined on the basis of this selection criterion will hereafter be called "backbone-determined minimum energy conformation" (BMEC) structures. The elegance of this method is that it produces reasonable results[19,34] and, from a theoretical and practical viewpoint, that it is based on a simple, deterministic selection criterion. On the other hand, it completely ignores the complexity of the multiple side-chain packing problem.

When pair-wise side-chain contributions are not ignored, the problem becomes much more difficult. Even when side-chain conformations are represented by a limited number of discrete rotamers, the globally best conformation can, in principle, only be found by means of a combinatorial search. Yet, it is possible to calculate for each side-chain rotamer r_i an upper and lower bound to the pair-interactions $E(r_i,r_j)$ considered in Eq. 1. These values can then be used to detect and eliminate conformations that are incompatible with the GMEC. Concretely, the original DEE criterion stipulates that side-chain rotamers r_i for which

$$E(r_i) + \sum_{\substack{j=1\\(j\neq i)}}^{N} \min_{r_j} E(r_i,r_j) > E(r_i') + \sum_{\substack{j=1\\(j\neq i)}}^{N} \min_{r_j} E(r_i',r_j) \qquad (3)$$

can never be an element of the GMEC.[4] The minimal and maximal values for the pair-interactions in Eq. 3 form calculable and mathematically safe substitutes for the interactions with the unknown GMEC rotamers, here referred to as G_j. If the G_j rotamers were known, the following expression

$$E(r_i) + \sum_{\substack{j=1\\(j\neq i)}}^{N} E(r_i,G_j) > E(r_i') + \sum_{\substack{j=1\\(j\neq i)}}^{N} E(r_i',G_j) \qquad (4)$$

would also be a valid elimination criterion (in view of Eq. 6 in Desmet et al.[4]). Both sides of this expression can be read as "the total energy of a particular rotamer r_i placed in a GMEC environment," noted as $E(r_i|GMEC^{-i})$ and mathematically defined as:

$$E(r_i|GMEC^{-i}) \equiv E(r_i) + \sum_{\substack{j=1\\(j\neq i)}}^{N} E(r_i,G_j);$$

$$GMEC^{-i} \equiv \{G_j|G_j \in \ GMEC; \quad j \neq i\} \qquad (5)$$

When substituting Eq. 5 into Eq. 4, we find that r_i is not a GMEC rotamer if another rotamer $r_i^{'}$ exists for which

$$E(r_i|GMEC^{-i}) > E(r_i'|GMEC^{-i}) \qquad (6)$$

In contrast to the original DEE-criterion (Eq. 3), this expression has the unique property that it allows the elimination of all rotameric states except $r_i = G_i$. Consequently, G_i can be identified, if all $G_{j\neq i}$ rotamers are known, by performing the operation:

$$G_i = \arg[\min_{r_i} E(r_i|GMEC^{-i})] \qquad (7)$$

In practice, this equation is of little use because the GMEC is generally unknown. On the other hand, it may be interesting to investigate which rotamer would result from the min-operation in Eq. 7 when the non-i residues adopt a non-GMEC state, here referred to as "intermediate minimum energy conformation" or $IMEC^{-i}$. A rotamer obtained in such way is noted as r_i^{min}:

$$r_i^{min} = \arg[\min_{r_i} E(r_i|IMEC^{-i})] \qquad (8)$$

An IMEC can be any global structure, including the BMEC as defined above or any possible randomly or methodologically generated structure.

A comparison between Eqs. 2 and 8 may be interesting. The min-operation in Eq. 2 is performed in an environment where the $j\neq i$ residues are "invisible," which is equivalent to setting the pair-interactions $E(r_i,r_j)$ to zero. In Eq. 8, these interactions have assigned values that fall into two categories: (i) those with residues in the GMEC state [i.e., $E(r_i^{min},G_j)$] and (ii) those with residues adopting a non-GMEC state [i.e., $E(r_i^{min},r_j\neq G_j)$]. Equation 7 suggests that the former will have a beneficial influence when using Eq. 8 to predict side-chain conformations, whereas the latter will have an unpredictable effect: either advantageous or disadvantageous compared to the situation wherein pair-interactions are set to zero.

Equation 8 provides us with a simple deterministic selection criterion, the performance of which can be tested in practice by systematically applying it to all modeled residues and where the IMEC structure consists of either a random structure, the BMEC structure or, importantly, the structure resulting from a previous round of r_i^{min}

calculations. The latter mode of operation corresponds to an iterative application in which IMECs are updated after each round (this is called a "batch" mode, as opposed to a "gradual" mode in which an IMEC is updated immediately for each new $r_i^{\min}$ rotamer). The central question is then whether the prediction quality statistically improves in consecutive rounds, which is discussed in Results and Discussion.

From a theoretical point of view, this iterative method can show two types of behavior: convergence to a uniquely defined system or not. Ignoring the latter possibility, convergence is bound to lead to a local minimum of the first or a higher order (where the order of a local minimum is defined in accordance with Goldstein[5]: in an *n*th order local minimum, the change of any *n* or fewer residues must result in a higher energy). Indeed, if in a given round not any new $r_i^{\min}$ rotamer is found, this is because all non-$r_i^{\min}$ rotamers led to a global structure with higher energy than that of the current IMEC.

Here we want to introduce the "perturbation-relaxation-evaluation" (PRE) method, which forms the most innovative and effective part of the FASTER method. The PRE approach aims to systematically surmount the barriers of local minima in which the protein system may have become trapped. One PRE step corresponds to the (temporary) fixation of a residue in a selected rotameric state (the perturbation step), followed by a relaxation step in which essentially the same approach is followed as described above (i.e., iterative optimization), and concluded by an evaluation step in which the total energy of the resulting structure is computed and used to decide about its retention or rejection. When only energetically downhill changes are retained and when the PRE cycle is systematically and exhaustively applied to all modeled residues in all available rotameric states until the system converges to a stable global state, the latter must reside in a local minimum of the second (or a higher) order. Indeed, because in such a case all possible rotamer pairs have been considered (one loop over the rotamers of the perturbed and another over the rotamers of the relaxed residue, and this for all pairs of residues) and not any of these situations has led to a lower global energy, the condition for a second-order local minimum is fulfilled.

It will be clear that the perturbation step is not necessarily restricted to single residues. It is conceptually trivial to systematically perturb any possible pair of residues in any possible combined rotameric state. If the same strategy is then followed as for single residues, a local minimum of the third (or higher) order must be reached when the system converges. Analogously, higher-order local minima may be reached. It is important to stress that a local minimum of the order *n* may also be of any higher order, possibly the N-th order where N is the total number of residues, in which case the global minimum is found. However, the latter cannot be guaranteed unless the PRE method is applied to all possible N-1 residue sets in all possible rotameric combinations. On the other hand, a comparison of the PRE results with the GMEC structures obtained by DEE may provide practical rules in this respect.

Algorithm

The core of the FASTER algorithm consists of a suite of four modules, here referred to as passes 1–4, in which an input IMEC structure is transformed into an output IMEC structure that is used as input in a next pass or, eventually, as the final result (Fig. 1). The initialization steps are described in Desmet et al.[35] and are essentially identical to those of the DEE method.[32]

Pass 1: iterative batch relaxation (iBR)

The FASTER method requires a starting structure to be input to pass 1. Apart from a randomized structure, the most obvious starting conformation is the one in which each rotatable side-chain is placed in its optimal interaction with the template, ignoring other side-chains, that is, the BMEC structure. The latter is calculated by application of Eq. 2, and the resulting $r_i^{\min}$ values (actually indices on the rotamer library for the amino acid type of each residue *i*) are stored in a one-dimensional array. The value of the *i*th element (residue) in this array is noted as I_i. The set of values $\{I_i\}$ defines the starting IMEC.

Each iteration round in pass 1 comprises three steps. First, for all available rotamers of each residue *i,* the total energy in the current environment, corresponding with $E(r_i \mid IMEC^{-i})$ in Eq. 8, is calculated, and the values are stored in a two-dimensional array. These values are noted as $E_{i,r}$. In the second step, the minimal energy is calculated for each residue (Eq. 8), and the corresponding rotamer number is transferred to the IMEC array, where previous I_i values are overwritten. Finally, the total energy of the protein (E_{tot}, Eq. 1) is calculated and printed so that the user can follow the progress of the optimization process. These three steps are executed in an iterative way until E_{tot} stabilizes or starts oscillating (see below). Importantly, $E_{i,r}$ values are recalculated only after all I_i values have been assigned. Such approach is typically characterized as a "batch process." The main advantages of a batch compared to a gradual approach are the higher computational speed and, especially, the independence of the order in which the residues are processed.

Pass 2: conditional iterative batch relaxation (ciBR)

This pass is basically identical to the previous routine, although some major and minor modifications have been introduced. Pass 2 also alternates between step 1 (updating of $E_{i,r}$ values) and step 2 (calculation of new I_i rotamers), but step 3 (calculation of E_{tot}) is skipped for reasons of computational speed. A more important difference is the fact that new rotamers are accepted only in 80% of the cases on a random basis. This acceptance probability, *P*, was introduced as a simple means to "break" the commonly observed oscillation of E_{tot} in consecutive cycles. An acceptance rate of 80% was found to be the optimal compromise between computational speed and energetic improvement (data not shown). A test version of the program revealed that by using P = 0.8, 20 optimization cycles are sufficient to reach energetic stability. Another consequence of introducing a stochastic factor is that the

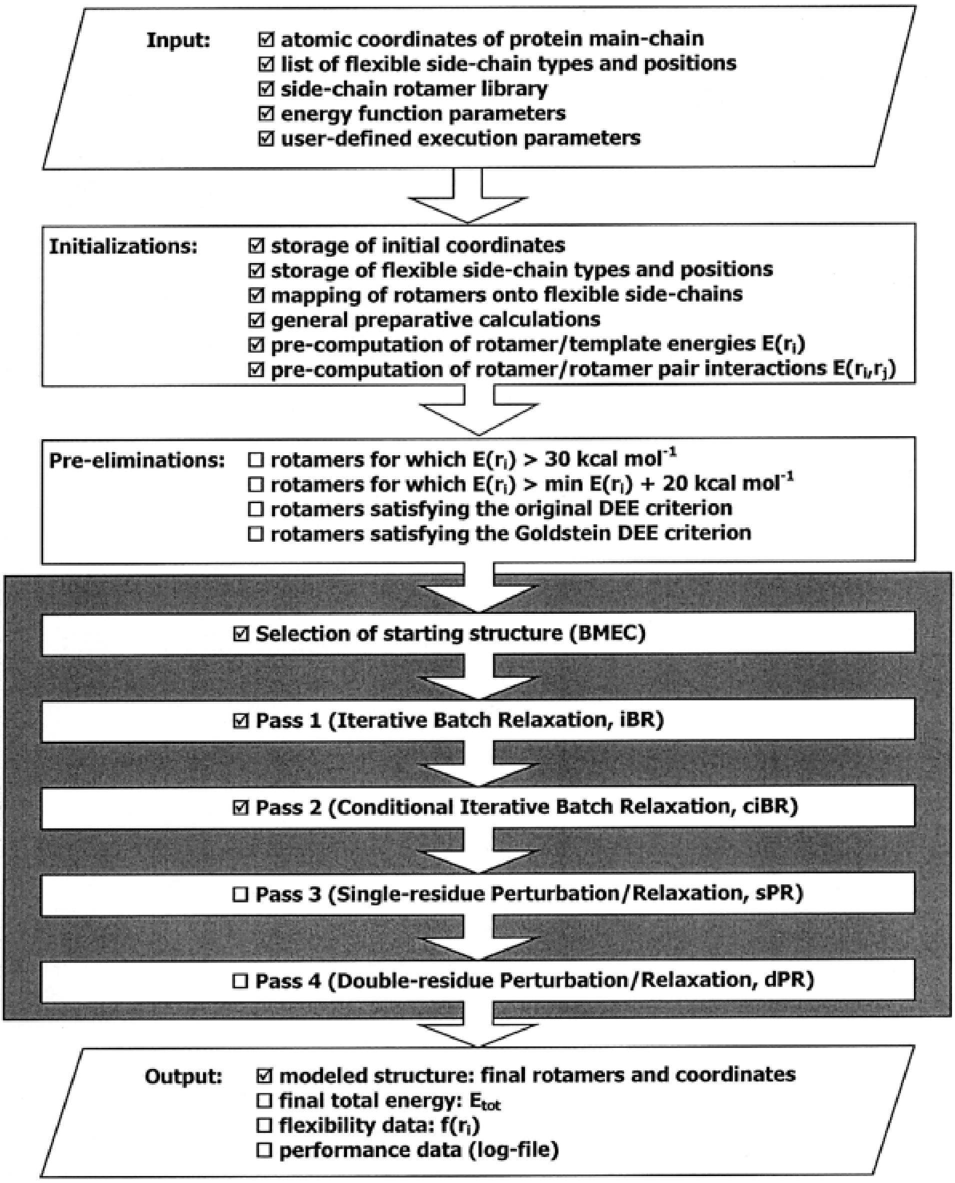
Input:
☑ atomic coordinates of protein main-chain
☑ list of flexible side-chain types and positions
☑ side-chain rotamer library
☑ energy function parameters
☑ user-defined execution parameters
Initializations:
☑ storage of initial coordinates
☑ storage of flexible side-chain types and positions
☑ mapping of rotamers onto flexible side-chains
☑ general preparative calculations
☑ pre-computation of rotamer/template energies $E(r_i)$
☑ pre-computation of rotamer/rotamer pair interactions $E(r_i,r_j)$
Pre-eliminations:
□ rotamers for which $E(r_i) > 30$ kcal mol^{-1}
□ rotamers for which $E(r_i) >$ min $E(r_i) + 20$ kcal mol^{-1}
□ rotamers satisfying the original DEE criterion
□ rotamers satisfying the Goldstein DEE criterion
☑ Selection of starting structure (BMEC)
☑ Pass 1 (Iterative Batch Relaxation, iBR)
☑ Pass 2 (Conditional Iterative Batch Relaxation, ciBR)
□ Pass 3 (Single-residue Perturbation/Relaxation, sPR)
□ Pass 4 (Double-residue Perturbation/Relaxation, dPR)
Output:
☑ modeled structure: final rotamers and coordinates
□ final total energy: E_{tot}
□ flexibility data: $f(r_i)$
□ performance data (log-file)

Fig. 1. Flow diagram of the complete FASTER method. The gray frame shows the pure FASTER routines. Required steps are marked with the symbol ☑, whereas optional steps are marked with □. In the present work, all options were selected, except for the flexibility analysis of 3LZT where the DEE pre-eliminations were skipped.

search path and, thus, also the final result of this pass is nondeterministic. We have exploited this feature by repeating the entire pass 2 procedure 10 times and retaining only the energetically best result as the input IMEC for pass 3.

Pass 3: single-residue perturbation/relaxation (sPR)

We have observed that the iBR and ciBR passes always get trapped in a local minimum different from the global minimum. The sPR pass systematically attempts to surmount such local minima to allow further relaxation. Central to this routine is the PRE concept, that is, the temporary fixation of a single side-chain in a particular conformation (perturbation) followed by energetic relaxation of the other side-chains and an energetic evaluation of the global result. This routine is basically the same as pass 1, enwrapped by a loop over all rotamers being temporarily fixed, and supplemented with an energy-based evaluation step.

Concretely, for each available rotamer of each rotatable side-chain (hereafter referred to as r_fix and i_fix, respectively) the following actions take place. First, the values for $E_{i,r}$ (where $i \neq$ i_fix) are corrected for the fixed side-chain. These values, noted as $E_{i,r}'$, are stored in a two-dimensional working array. Thus, the $E_{i,r}'$ values reflect the energetic consequences for all rotamers r_i of perturbing the IMEC by a single rotamer r_fix$_{\text{i\_fix}}$. Next, for all residues except the fixed one the minimum energy rotamer according to Eq. 8 is calculated on the $E_{i,r}'$ values and stored in the IMEC array (I_i values are updated). This step, referred to as the relaxation step, is thus identical to step 2 in pass 1. Also the next step, the calculation of E_{tot}, is the same as in pass 1, but the value for E_{tot} is here used as a decision criterion to accept or reject the new global conformation. Only in case E_{tot} is lower than the lowest value found thus far, the latter conformation is accepted as the new IMEC, which is consolidated by updating the $E_{i,r}$ with the $E_{i,r}'$ values. On completion of a PRE cycle, the algorithm continues by selecting the next rotamer, until all rotamers have been fixed once. Residues are selected for fixation in random order, and their rotamers are selected in the order they appear in the rotamer library. After one full round of fixations, a next round is executed. The process terminates when E_{tot} has not altered in the last full round.

Pass 4: double-residue perturbation/relaxation (dPR)

Exactly the same strategy is followed as in pass 3, except that the IMEC is here perturbed by systematic fixation of, in principle, all possible rotamer pairs. However, we have observed that the efficiency of finding an energetically better conformation in a given PRE cycle is strongly dependent on the distance between the perturbed residues. Therefore, only couples of residues (i,j) are selected for which the distance between their C_β-atoms is below $(6.0 + n_i + n_j)$ Å, where n_i and n_j is a measure of the size of side-chains i and j, respectively (more precisely, n is the number of heavy atoms counted from the C_β-atom along the longest branch in the side-chain). Pass 4 is the slowest step and is optional. It may be skipped if computational speed is preferred over maximal accuracy.

Algorithmic optimizations

The computational speed of the FASTER method strongly depends on the total number of side-chain rotamers. Therefore, we have introduced two logical optimization procedures that focus on a mathematically safe reduction in rotamers.

The most important optimization is the preelimination of rotamers before the start of the FASTER routines. As in the DEE-algorithm,[32] rotamers for which the interaction energy with the template, $E(r_i)$, is higher than 30 kcal mol^{-1} or higher than 20 kcal mol^{-1} above the lowest value for the same residue, are removed. Optionally, some DEE steps may be executed as well to further reduce the number of rotamers at minimal computational cost. Preferably, rotamers satisfying the original DEE criterion (Eq. 3 and Ref. 4) and the Goldstein DEE criterion (Eq. 4 from Ref. 5) are eliminated beforehand. In the current work, these routines were switched on and were executed in an iterative way until in a given cycle $< 5\%$ of the total number of remaining rotamers were eliminated.

The second program optimization only concerns pass 4 and exploits the experimental finding that the structures obtained after pass 3 already closely approximate the GMEC (see Results and Discussion). In such condition, it follows from a comparison between Eqs. 3 and 4 that rotamers r_i satisfying the "hybrid" criterion

$$E(r_i) + \sum_{\substack{j=1\\(j\neq i)}}^{N} \min_{r_i} E(r_i,r_j) > E(I_i) + \sum_{\substack{j=1\\(j\neq i)}}^{N} E(I_i,I_j) \qquad (9)$$

are "most likely" not GMEC rotamers. The theoretical validity of the latter criterion is of little importance here, because rotamers matching Eq. 9 are only skipped for perturbation but remain present during the relaxation steps. Moreover, we have experimentally observed that this algorithmic optimization has absolutely no influence on the resulting structures but drastically reduces the computational time.

Test Set and Technical Aspects

The test set of protein structures was identical to the combined first and second test set of Xiang and Honig.[30] In short, it consisted of 33 structures with a resolution ranging between 0.83 and 1.4 Å, a pair-wise sequence identity below 20%, and a size of 46–328 residues. Their PDB codes are: 1AAC, 1AHO, 1B9O, 1C5E, 1C9O, 1CBN, 1CC7, 1CEX, 1CKU, 1CTJ, 1CZ9, 1CZP, 1D4T, 1ECA, 1IGD, 1IXH, 1MFM, 1PLC, 1QJ4, 1QL0, 1QLW, 1QNJ, 1QQ4, 1QTN, 1QTW, 1QU9, 1RCF, 1VFY, 2PTH, 3LZT, 5P21, 5PTI, and 7RSA.

Hydrogen atoms were automatically added by the Brugel modeling package.[36] All structures were refined by 100 steps steepest descent minimization. The CHARMM force field of Brooks et al.[37] was used throughout this work to calculate potential energies. No modifications were made to the energy function nor to the parameters. No specific solvent terms were taken into account. Coulombic interactions were calculated by using a distance-dependent dielectric constant ($\epsilon = r$). The pair-wise atomic interaction cutoff was set at 8 Å. All hydrogen atoms were explicitly included in the computations. Side-chains were modeled in standard geometry (i.e., they were generated with ideal bond lengths and angles). The protein backbones were kept fixed during the modeling. For each structure, the template was composed of the main-chain, C_β-atoms and full Gly, Ala, Pro, and disulfide bond forming Cys residues. The calculations were performed on a cluster of four SGI Origin 200 computers, each equipped with four 270 MHz R12000 processors and 4 GB of memory.

The rotamer library was the same as used in previous work,[32] with the exception that for β-branched side-chains (Val, Thr, Ile) one step of 10° was taken above and below their basic χ_1 dihedral angle values, leading to three subrotamers per basic rotamer. The total number of rotamers in the library was 899.

The DEE computations were performed by the latest version of the algorithm, as published before.[7] The DEE-modeled structures were stored on disk for later comparison with the FASTER-derived structures. Both in the DEE and FASTER experiments, side-chain rotamers having an absolute "template interaction energy" [$E(r_i)$ in Eq. 1] higher than 30 kcal mol^{-1} and a relative energy higher

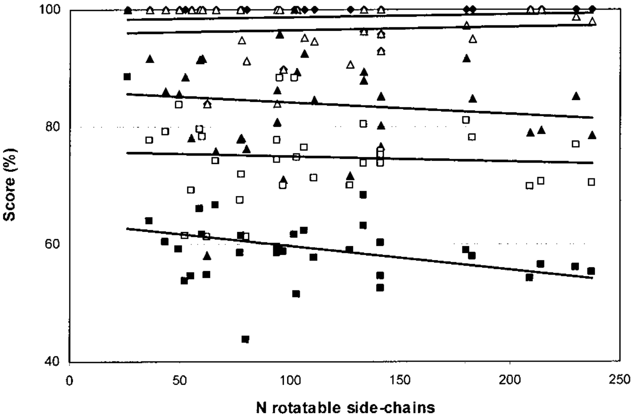

Fig. 2. Percentage of identical rotamers obtained by FASTER and DEE. Prediction scores are given for each protein of the test set and for the structures obtained after each pass of the FASTER algorithm, including the initial BMEC. All rotatable residues are taken into account. Symbols: (■), BMEC; (□), pass 1; (▲), pass 2; (△), pass 3; (◆), pass 4. The scores are plotted as a function of the protein size, expressed as the number of rotatable side-chains. Size-weighed average values and standard deviations are given in Table I.

than 20 kcal mol^{-1} above the lowest $E(r_i)$ value for each residue i were eliminated during the initialization stage (Fig. 1). Pair-wise rotamer/rotamer interaction energies were also precomputed during initialization.

Comparison between FASTER, DEE, and X-ray structures was performed by an automated procedure providing information about side-chain RMSD, volume overlap,[32] χ_1 and χ_{1+2} dihedral angle correlation, both for individual residues and complete molecules. All average values mentioned in this work have been weighed by the number of rotatable side-chains in the molecules.

RESULTS AND DISCUSSION

Program Performance

The FASTER method has been applied to a test set of 33 high-resolution protein structures, corresponding exactly to the first and second test set of Xiang and Honig.[30] The test set was prepared as described in Materials and Methods. In addition to the FASTER experiments, all proteins have been modeled by our latest version of the DEE algorithm.[7,32] Because DEE is guaranteed to find the GMEC (i.e., the theoretically best possible global solution), we were able to check whether the FASTER method could reach the same accuracy.

Figure 2 shows for each protein the percentage of side-chains that were exactly in the GMEC after each consecutive pass of the FASTER algorithm. Table I shows the average values and standard deviations. The scores in Figure 2 are plotted as a function of the size of the proteins, expressed in terms of the number of rotatable side-groups. For each pass, including the initial BMEC structure, we found that the scores are essentially independent of the protein size. It is also clearly seen that each individual pass leads to a gradual and significant improvement of the global conformation. For the BMEC structure, where each residue was assigned the rotamer with the best possible interaction with the template, ignoring other side-chains, the average score was 58.2% with a standard deviation of 7.1%. Application of pass 1 (iBR) led to the most significant improvement by about 16% to an average score of 74.6 ± 7.0%. Pass 2 (ciBR) further increased the score to 83.3 ± 8.7%. The third pass (sPR) turned out to be the most effective, considering the fact that most of the side-chains were already in the GMEC. The average score increased by 13.4% to 96.7 ± 4.8%, and in 17 of the 33 cases, the entire structure was identical to the global minimum. Finally, pass 4 (dPR) yielded a score of 98.9 ± 3.5% on average, and the GMEC was reached for 28 proteins.

Another way to evaluate the FASTER results is by measuring the volume overlap of each side-chain with the GMEC. We assumed two side-chain conformations to be equivalent when they showed an overlap $> 70\%$.[32] By using this criterion we again observed a gradual increase in the average prediction scores as a function of the different passes (Table I). All scores based on the volume overlap criterion were significantly higher than those based on the identity with the GMEC. This obviously arises from the fact that the overlap criterion is less restrictive (but probably more significant) than the identity criterion. The average volume-based scores varied from 76.9% for the starting structures, to 86.5% in pass 1, 91.5% in pass 2, 98.3% in pass 3, to 99.5% for the final structures in pass 4. Thus, on completion of pass 4, only 0.5% of the modeled side-chains has a conformation that is spatially significantly distinct from that in the GMEC.

An important measure of the quality of computed structures is their total energy, E_{tot}. Figure 3 and Table I (fourth column) show the difference in energy between the structures obtained after passes 2–4 and the GMEC. All values have been divided by the number of rotatable side-chains in the protein because large proteins obviously contain more deviations, which makes the difference in total energy size-dependent. The starting structures and those obtained after pass 1 always contained severe atomic overlaps, yielding strongly positive energy values with a large spread. Pass 2 was extremely effective in removing the worst clashes. In all but one cases, the pass 2-derived structures were clash-free. Only the 1CKU pass 2 structure (hipip, 55 rotatable side-chains) had a high-energy difference with the GMEC of 2.88×10^5 kcal mol^{-1} (5.24×10^3 kcal mol^{-1} res^{-1}). Ignoring this value, the average difference with the GMEC was 0.21 ± 0.19 kcal mol^{-1} res^{-1}. This means that for a medium-sized protein of 200 residues, the expected total energy is about 42 kcal mol^{-1} higher than that of the GMEC. Passes 3 and 4 were able to further reduce the energy differences (also for 1CKU) to only 0.014 ± 0.022 and 0.0014 ± 0.0045 kcal mol^{-1} res^{-1}, respectively. For a 200-residue protein, the energy of the global minimum can thus be approximated to about 3 kcal mol^{-1} after the sPR pass and to 0.3 kcal mol^{-1} after the full FASTER method. However, these values only reflect the globally expected error per residue and not the local strain associated with badly placed side-chains. When the energy differences are averaged over only the incorrectly placed residues (Table I, last column), significantly higher

TABLE I. Comparison of FASTER Structures With the GMEC

	Score ident.[a] (%)	Score vol.[b] (%)	ΔE^c (kcal mol^{-1} res^{-1})	ΔE^d (kcal mol^{-1} res^{-1})
BMEC	58.2 ± 7.1	76.9 ± 5.7	$(1.8 \pm 4.9) \times 10^{16}$	$(4.1 \pm 11.1) \times 10^{16}$
Pass 1	74.6 ± 7.0	86.5 ± 5.6	$(0.6 \pm 1.3) \times 10^{6}$	$(2.2 \pm 4.5) \times 10^{6}$
Pass 2	83.3 ± 8.7	91.5 ± 5.2	0.21 ± 0.19[e]	1.19 ± 1.18[e]
Pass 3	96.7 ± 4.8	98.3 ± 3.4	0.014 ± 0.022	0.45 ± 0.86
Pass 4	98.9 ± 3.5	99.5 ± 2.6	0.0014 ± 0.0045	0.027 ± 0.087

[a] Average percentage of identical rotamers after FASTER and DEE modeling.
[b] Average percentage of side-chains with a volume overlap of at least 70%.
[c] Difference in E_{tot} per residue when averaged over all residues.
[d] Difference in E_{tot} per residue when averaged over the wrong (non-GMEC) residues.
[e] Value for ICKU ignored (see text).

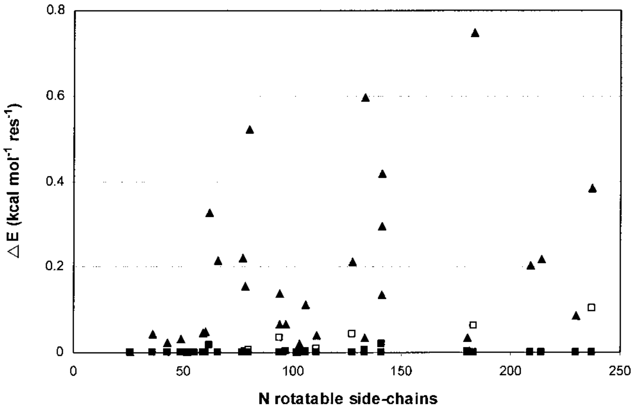

Fig. 3. Difference in energy between FASTER and GMEC structures. For each protein of the test set, the difference in total energy between the FASTER-derived structures obtained after passes 2–4 and the DEE-computed GMEC is plotted. The energy differences are expressed in kcal mol^{-1} per rotatable residue and plotted as a function of the protein size. Symbols: (▲), pass 2; (□), pass 3; (■), pass 4.

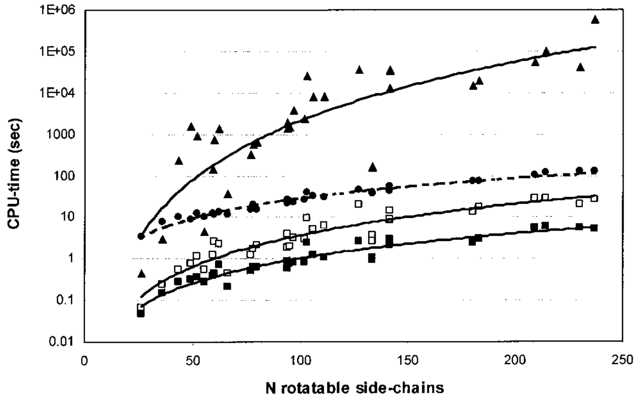

Fig. 4. Comparison of FASTER and DEE CPU times. The total CPU times required to model the test proteins by means of the DEE method (▲) and the 4-pass (□) and 3-pass (■) FASTER method, excluding initializations, are plotted as a function of the protein size. The CPU times for the initializations are given separately (●). All values are expressed in CPU seconds (sec) and are plotted on a logarithmic scale.

values are obtained. Now it is seen that the 3.3% wrong side-chains after pass 3 have an average energy difference with the GMEC of 0.45 ± 0.86 kcal mol^{-1} res^{-1}, which is more or less comparable to the value of 1.19 ± 1.18 kcal mol^{-1} res^{-1} obtained after pass 2. Thus, the few wrong side-chains after pass 3 still have a relatively high local strain. This is not the case for the wrong side-chains after pass 4: they show a local strain of only 0.027 ± 0.087 kcal mol^{-1} res^{-1} compared to the GMEC. Finally, we remark that the relatively high standard deviations on all values are due to the fact that the energies do not follow a normal distribution. The largest deviation from the GMEC, after pass 4, was found for 1CC7 (Metallochaperone Atx1, 62 rotatable residues), which had an energy difference of 0.0162 kcal mol^{-1} res^{-1} (3.24 kcal mol^{-1} per 200 residues), associated with 10 badly placed side-chains (i.e., 0.10 kcal mol^{-1} per wrong residue).

Figure 4 shows the computational time required for the modeling of each protein by using the DEE algorithm and the 3- and 4-pass FASTER algorithm. Because of the extreme differences in CPU times, the values are plotted on a logarithmic scale. Analysis of the data showed that the CPU times scale with the number of rotatable residues to the power of 4.71, 2.50, and 1.97 for the DEE method and the 4- and 3-pass FASTER method, with correlation coefficients (R^2) of 0.70, 0.88, and 0.91, respectively. The fact that the CPU times for DEE are very dependent on the protein size (and are poorly predictable) is primarily due to the combinatorial end-stage routine. In contrast, passes 4 and 3 theoretically scale in a cubic and quadratic fashion, respectively, that is, the number of perturbations [in the big-O notation: $O(n^2p^2$ for pass 4 and $O(np)$ for pass 3] times the number of relaxations [$O(np)$ for both passes]. In practice, pass 4 is somewhat more efficient than theoretically predicted, which is due to the fact that only proximate couples of side-chains are perturbed simultaneously. More important, however, is a comparison between absolute CPU times. For the largest proteins studied (>200 rotatable residues), DEE takes about 10^5 CPU-sec or more than one CPU-day. In marked contrast is the performance of the FASTER methods. The 4-pass FASTER method consumes <30 CPU-sec for the largest proteins, which is more than three orders of magnitude faster than DEE. The 3-pass method accomplishes these jobs in only about 5 CPU-sec, thereby reducing the computational times by another factor of 6. For small systems (≈30 residues), DEE is known to be relatively fast, with CPU times lying in the range of seconds. Yet, even in these cases, FASTER remains about 100 times faster, with CPU times in the range of hundredths of seconds.

Importantly, the FASTER routines in the 3- and 4-pass versions remain in all cases significantly below the initial-

TABLE II. Comparison of FASTER and DEE Structures With X-ray structures

	Buried side-chains (<10% ASA)					All side-chains				
	BMEC	Pass 2	Pass 3	Pass 4	GMEC	BMEC	Pass 2	Pass 3	Pass 4	GMEC
RMSD	1.36 ± 0.47	1.06 ± 0.37	1.06 ± 0.41	1.06 ± 0.42	1.06 ± 0.41	1.85 ± 0.31	1.67 ± 0.30	1.65 ± 0.28	1.66 ± 0.30	1.65 ± 0.30
Vol[a]	80.4 ± 7.7	87.3 ± 7.4	86.9 ± 7.6	87.2 ± 7.4	87.2 ± 7.2	64.2 ± 6.5	69.2 ± 6.8	69.9 ± 6.5	69.9 ± 6.8	70.0 ± 6.7
χ_{1+2}[b]	79.0 ± 8.5	86.3 ± 7.7	86.3 ± 7.8	86.6 ± 7.5	86.6 ± 7.6	64.4 ± 6.1	69.7 ± 6.7	70.5 ± 6.1	70.4 ± 6.4	70.4 ± 6.3

[a] Average percentage of side-chains with a volume overlap of at least 70%.

[b] Average percentage of side-chains with a difference in χ_1 and χ_2 angles of <40°.

ization times (Fig. 4, dotted line). The initializations, dominated by the computation of single and pair-wise energy terms now become rate limiting, whereas they were almost negligible compared to the DEE routines.

Comparison With Experimental Structures

The difference in side-chain positioning accuracy between the FASTER and the DEE methods completely vanishes when the modeled proteins are evaluated against the experimentally determined structures. Table II lists the results obtained by the template-based method producing BMEC structures, the FASTER pass 2–4 structures and the DEE-generated GMEC structures. The predicted structures are evaluated by using three applicable criteria, and separate values are given for the buried and for all side-chains.

On the basis of the 70% volume overlap criterion, we found that the DEE algorithm places 87.2 ± 7.2% of the buried side-chains and 70.0 ± 6.7% of all rotatable side-chains in a correct conformation. Using the classical χ_{1+2} ± 40° criterion we obtained 86.6 ± 7.6% and 70.4 ± 6.3%, respectively. It is of interest that both scoring criteria yield very similar results, although discrepancies often occur at the level of individual side-chains. The average side-chain RMSD of modeled versus experimental structures was 1.06 ± 0.41 Å for buried and 1.65 ± 0.30 Å for all side-chains. In general, we conclude that the score/RMSD with the DEE method and current settings (see Materials and Methods) is about 87%/1 Å and 70%/1.6 Å for core and all side-chains, respectively. These results are in very good agreement with our earlier findings[32] and are at the top of what can be reached today in this field. Most recently, two studies have been published in which the apparent barriers of ≈90%/1 Å and ≈70%/1.5 Å were crossed: Xiang and Honig[30] reported an RMSD of 0.74 Å for core residues (1.66 Å for all residues). Mendes et al.[2] obtained χ_{1+2} ± 40°/RMSD results of 88%/0.70 Å and 76%/1.34 Å for core and all residues, respectively. Both groups applied an extremely large library comprising thousands of rotamers to represent the conformational space of the natural amino acid side-chains. In addition, Mendes et al.[2] introduced entropical contributions by using the concept of "flexible rotamers." They attributed half of the increase in accuracy to the high number of rotamers and an equal improvement to the entropy terms.

The standard deviations on the prediction scores are very large, consistent with earlier work.[10,29,30,32,38] A first reason is the calculation of percentages for small numbers of residues: the standard deviations for core residues are about 1% higher than for all residues (see also Fig. 1, where the spread in scores for small proteins is significantly greater than for large molecules). However, a variety of other factors may play a role, but this is beyond the scope of the present work. Yet, it indicates that side-chain placement studies must be performed on large test sets.

The main objective of the present work is to compare the results obtained by the FASTER and DEE methods. Table II shows that both methods yield structures of identical quality compared with the experimental structures. This is logical because Table I showed an exact (or equivalent) correspondence with the GMEC for 98.9% (or 99.5%) of all side-chains after pass 4. Even after pass 3, there was already a 96.7% identity or 98.3% equivalence.

Before discussing the scores after pass 2, it is worth going into the BMEC scores. When the side-chains are placed in the optimal conformation with respect to the template (≈backbone), about 77% of them assume a conformation equivalent to the GMEC (Table I) and about 64% are equivalent to the experimental structure (Table II). These are remarkably high scores in view of the fact that side-chain/side-chain interactions are completely ignored in BMEC structures. In contrast, when the pair interactions are "switched on" and the GMEC is identified by a suitable method, this involves the repositioning of 23% of the side-chains (GMEC-based score increasing from 77 to 100%) but only about 6% increase in score when evaluated against experimental structures (from 64 to 70%). These results illustrate the very strong driving force of local interactions compared with global side-chain packing interactions.

Figure 5 illustrates the evolution of the average FASTER prediction scores, based on a comparison with the GMEC or X-ray structure. As discussed before, passes 1, 2, and 3 progressively improve the quality of the structures into the direction of the GMEC. When the same structures are compared with the experimental structures, it is seen that the scores become stagnant after pass 2. It followed from Table I that 8.5% of pass 2 determined side-chains are significantly different from the GMEC, but their further modeling by passes 3 and 4 is of little relevance in the light of experimental structures. Similar results are obtained for the buried side-chains, irrespective of the evaluation criterion (Table II). For the latter residues, starting from the BMEC, an increase of about 7% is obtained by the FASTER modeling (which is substantial given that already about 80% are correctly positioned in the BMEC), but again, this increase is already reached at the end of

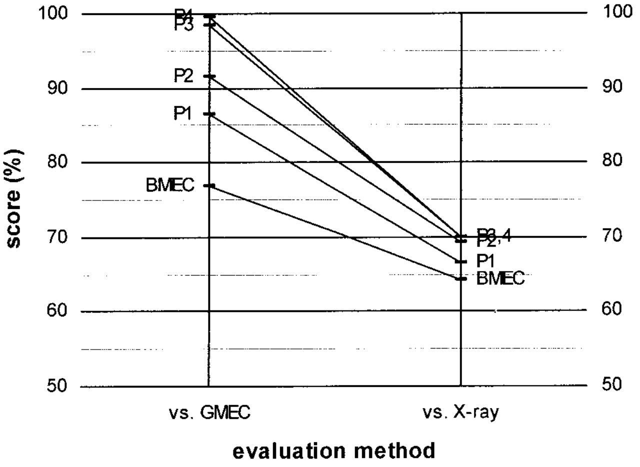

Fig. 5. Comparison of FASTER prediction scores relative to the GMEC and X-ray structures. The average prediction scores are based on the 70% volume overlap criterion. All rotatable residues are taken into account.

pass 2. From these results we can draw the following general conclusions: (i) the protein backbone is a very strong conformational determinant; (ii) side-chain/side-chain packing appears to be a relatively weak determinant; (iii) the maximal accuracy can be achieved by, in principle, a simple local search; and (iv) the energetic improvement obtained after passes 3 and 4 ($\approx$0.2 kcal mol^{-1} res^{-1}, see Table I) is mainly of theoretical importance.

An explanation for the latter findings readily follows from the calculation of the average difference in energy between the modeled and experimental structures: after pass 4, this energy difference is 2.95 ± 3.52 kcal mol^{-1} res^{-1}, which is $>$10 times higher than the difference between pass 2-derived structures and the GMEC. The value of $\approx$3 kcal mol^{-1} res^{-1} includes the suboptimal relaxation after pass 2 ($\approx$0.2 kcal mol^{-1} res^{-1}) but, more importantly, it primarily results from the usage of discrete rotamers having standard bond length and angle geometry and predefined torsional angles. Within a given protein environment, a rotameric geometry is by definition not optimally adapted. When applying a sensitive force field model (like the all-hydrogen CHARMM function used in this study), the rotameric approximations cause significant, noiselike energetic approximations. This can only be resolved by either using a less sensitive energy function (with the risk of losing specificity, as pointed out by Lazaridis and Karplus,[39] or by including a multiple of the current number of rotamers in the library. Consistent with our earlier findings[32] and those of Xiang and Honig[30] and Mendes et al.[2] we believe that the latter is a preferred approach.

Analysis of Rotameric Side-Chain Flexibility

The sPR pass provides the opportunity to estimate the rotameric flexibility of individual side-chains within their specific context. Such flexibility analysis is postulated to measure the true rotameric flexibility of side-chains because it incorporates potential rotameric adaptations in the environment of tested rotamers. Moreover, a flexibility analysis can be performed at almost no extra computational cost because it is derived from data that are computed during normal execution of the sPR pass.

The basic idea was to capture in a data array the values for E_{tot}, computed in the third step of the PRE cycle, for each temporarily fixed (perturbed) rotamer during the last iteration round of the sPR pass. Each such value can be written as $E_{tot}(r_i)$ and read as the total energy of the pass 3 IMEC structure, in which side-chain i has been forced into rotameric state r_i (step 1 of the PRE cycle) and its environment has been energetically adapted to this state (step 2 of the PRE cycle). The final IMEC structure of the sPR pass can be seen as the "ground state" from where the perturbations are executed (including adaptation of the environment). Thus, the difference between the total energy $E_{tot}(r_i)$ and that of the ground state, $\Delta E_{tot}(r_i)$, reflects the energetic cost to force a side-chain from the ground state rotamer to r_i and is a true perturbation energy. All perturbation energies below a given cutoff value, in the present work set to 5 kcal mol^{-1}, are considered as energetically feasible. Formally, a flexibility coefficient f(i) for each side-chain i can be defined as

$$f(i) \equiv (N_{allowed}(i) - 1)/(N_{library}(i) - 1) \qquad (10)$$

where $N_{library}(i)$ is the number of library rotamers for residue i and $N_{allowed}(i)$ is the number of rotamers having an energetically acceptable perturbation energy. Finally, all side-chains can be assigned as either rigid, semiflexible, or flexible, where rigid means $f(i) < 0.05$, semiflexible is $0.05 \leq f(i) < 0.34$, and flexible is $f(i) \geq 0.34$.

Such flexibility analysis is exemplified for 3LZT (hen lysozyme, 95 rotatable side-chains) in Figure 6, where strictly rigid and flexible residues are indicated by white and gray spheres, respectively. The 22 rigid side-chains had the following distribution: 6 Leu, 1 Ile, 5 Trp, 1 Phe, 1 Tyr, 2 Met, 2 Asn, 1 Glu, 2 Arg, and 1 Lys. The distribution of the 18 flexible side-chains was 1 Val, 9 Ser, 3 Thr, 1 Asn, 3 Asp, 1 Arg. Thus, 16 of the 22 rigid side-chains were apolar (73%), and 17 of the 18 flexible side-groups (94%) had a polar character, the latter group being dominated by Ser. It is not surprising that the distribution in the structure is such that rigid side-chains are mainly located in the core of the molecule, whereas flexible side-chains are found almost exclusively at the surface (Fig. 6). These results are consistent with the general idea that packing at the interior of a protein is very dense and specific, whereas it is much less dense at the exterior, resulting in fewer energetic constraints and, consequently, a greater conformational freedom. The FASTER method can provide a valuable tool to further explore and/or quantify this important property.

CONCLUSIONS

In the present work we have focused on the development and application of an original method for side-chain placement. The FASTER method enables the modeling of large sets of side-chains onto a predefined backbone structure

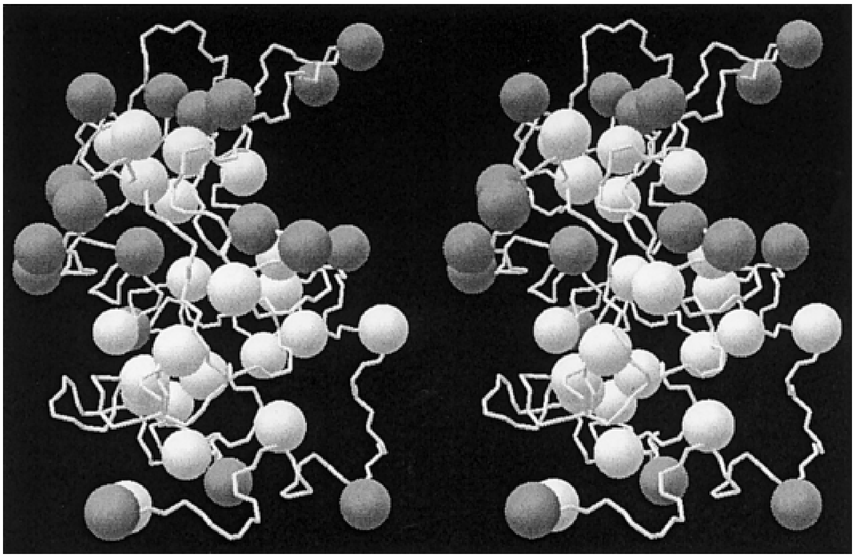

Fig. 6. Stereo plot showing the results of a flexibility analysis of 3LZT (hen lysozyme). The residues classified by the FASTER method as rigid ($f < 0.05$) or flexible ($f \geq 0.34$) are depicted by white or gray spheres, respectively. All spheres are mapped onto the C_β-atoms of the side-chains concerned. Semiflexible residues are not shown.

with calculation times in the range of seconds or below (when the rate-limiting precalculations are ignored). The computations scale in a stable and favorable way with the size of the proteins studied. This makes the method ideally suited for high-throughput modeling and the prediction of very large systems like those encountered in protein design, as we have shown earlier.[40]

In the margin of this work, we have obtained valuable theoretical and practical insights into the complexity of the energy landscape involved with side-chain placement. The FASTER algorithm is particularly well suited to study this subject because it searches for local minima of progressively higher order. The latter can be compared with the global minimum obtained by DEE. If the global minimum is found to coincide with a local minimum of lower order, this proves that there is at least one simple route, different from a brute-force combinatorial approach, which leads to this state. The structures obtained after pass 2 probably correspond to local minima of the first or higher order. Those after passes 3 and 4 are of the order ≥2 and ≥3, respectively. We found that for 2 of the 33 proteins studied (1CBN and 1B9O), the pass 2-derived structures coincided exactly with the global minimum. After passes 3 and 4, this was the case for 17 and 28 proteins, respectively. Thus, for most proteins, the global minimum coincided with a detectable local minimum of low order. Five proteins (1CC7, 7RSA, 1CEX, 1RCF, and 5P21) got trapped in a local minimum of at least order 3 but different from the GMEC. On the energy landscape, these local minima were very close to the global minimum, the highest difference being observed for 5P21 having a difference in total energy with the GMEC of 2.8 kcal mol^{-1} (0.020 kcal mol^{-1} res^{-1} or 0.47 kcal mol^{-1} res^{-1} when averaged over all side-chains or the 6 non-GMEC side-chains, respectively). In general, it can be concluded that the energy landscape in side-chain placement shows two phases. Away from the GMEC, it consists of a rugged hypersurface, but steep downhill movements can be realized by simple local optimizations (passes 1 and 2). Once this rugged phase has been passed, the landscape abruptly changes to a smooth surface with plenty of shallow wells that are of second, third, or occasionally, higher order.

The discussion about whether protein side-chain placement faces a combinatorial problem has been going on for years. From a theoretical point of view, a combinatorial approach (or rather, an approach that respects the combinatorial nature of global side-chain packing) is certainly justified. However, this does not mean that it is impossible to identify criteria allowing a safe narrowing of the combinatorial space (like the DEE criteria) or to find paths leading straight to the global optimum, or very nearby (like the FASTER method). The present work shows that the side-chain placement problem is of a relatively simple type in that both the a priori space can be narrowed and steep downhill paths can be identified. Thus, the theoretical combinatorial problem, having a dimension proportional to the protein size, can, in practice, be reduced to a combinatorial problem of dimension 1 or 2 (i.e., the single or double perturbations in passes 3 and 4, respectively), with sufficient success in all cases studied. For an absolute success rate, we expect that one or two more dimensions (triple or quadruple perturbations) may be sufficient to reach the global minimum in all cases.

A most interesting observation is the fact that even the “locally modeled” pass 2 structures are of the same quality as the GMEC when evaluated against experimentally determined structures. This can be explained by the fact that the difference in total energy per residue between the modeled and experimental structures is relatively high (i.e., about 3 kcal mol^{-1} res^{-1}). This value can be interpreted, in essence, as the "rotamerization energy" resulting from the discretization of conformations into rotamers. It can be expected that discretization effects influence both the "correct" and "incorrect" rotamers in a stochastical way, so that they act as a nuisance factor. The rotamerization energy can be reduced by applying a smoother energy function, but chances are high that both the correct and incorrect rotamers benefit from it and that specificity is lost. Probably the only alternative is to work with large numbers of rotamers or flexible rotamers.[2] However, this not only extends the conformational space in an extreme fashion, it also reshapes the energetic landscape. Evidence for this comes from protein design experiments using DEE, in which it was found that the repertoire of allowed amino acid types and rotamers at the modeled residue positions dramatically influences the computational times.[10] Compared with DEE, the FASTER method has a considerable margin left for further extension of the rotamer library, for working with flexible rotamers, and for including amino acid variation into the rotameric concept (i.e., to perform protein design).

An important feature of the FASTER algorithm is that it provides significant information about the conformational flexibility of individual side-chains. We showed that strictly rigid side-chains are concentrated mainly in the core, whereas very flexible side-chains are found almost exclusively among solvent-oriented residues. Evidently, the FASTER method could be applied to monitor (changes in) rotameric flexibility for functional regions, interfaces, secondary structural units, and so forth. Analogously, if flexible rotamers are used, it is in principle also possible to examine small-scale, nonrotameric flexibility and to correlate this with crystallographic temperature factors. The same idea that led to the computation of rotameric perturbation energies has already been applied to amino acid substitutions (Ref. 40 and Vlieghe et al., to be published). It is indeed relatively straightforward to calculate, in a systematic way, the substitution energies for all residue types at each position in a protein structure. Such data can be correlated with stability data or natural sequence variation and used in protein design applications. These features illustrate the versatility of the FASTER approach for protein structure prediction, analysis, and design.

REFERENCES

1. Vásquez M. Modeling side-chain conformation. Curr Opin Struct Biol 1996;6:217–221.
2. Mendes J, Baptista AM, Carrondo MA, Soares CM. Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. Proteins 1999;37:530–543.
3. Voigt CA, Gordon DB, Mayo SL. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. J Mol Biol 2000;299:789–803.
4. Desmet J, De Maeyer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein sidechain positioning. Nature 1992;356:539–542.
5. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J 1994;66:1335–1340.
6. 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 1995;8:815–822.
7. Desmet J, De Maeyer M, Lasters I. Theoretical and algorithmical optimization of the dead-end elimination theorem. In: Altman RB, Dunker AK, Hunter L, Klein TE, editors. Pacific Symposium on Biocomputing '97, Hawaii, USA. New Jersey: World Scientific 1997. p 122–133.
8. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins 1998;33:227–239.
9. Pierce NA, Spriet JA, Desmet J, Mayo SL. Conformational splitting: a more powerful criterion for dead-end elimination. J Comp Chem 2000;21:999–1009.
10. Looger LL, Hellinga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol 2001;307:429–445.
11. Desjarlais JR, Clarke ND. Computer search algorithms in protein modification and design. Curr Opin Struct Biol 1998;8:471–475.
12. Fraenkel AS. Protein folding, spin glass and computational complexity. Third annual DIMACS workshop on DNA based computers, Philadelphia, June 23–25, 1997.
13. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82–87.
14. Wernisch L, Hery S, Wodak SJ. Automatic protein design with all atom force-fields by exact and heuristic optimization. J Mol Biol 2000;301:713–736.
15. Bruccoleri RE. Ab initio loop modeling and its application to homology modeling. Methods Mol Biol 2000;143:247–264.
16. Huang ES, Koehl P, Levitt M, Pappu RV, Ponder JW. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. Proteins 1998;33:204–217.
17. Desmet J, Wilson IA, Joniau M, De Maeyer M, Lasters I. Computation of the binding of fully flexible peptides to proteins with flexible side chains. FASEB J 1997;11:164–172.
18. Adams PD, Pannu NS, Read RJ, Brunger AT. Extending the limits of molecular replacement through combined simulated annealing and maximum-likelihood refinement. Acta Crystallogr Sect D 1999;55:181–190.
19. Eisenmenger F, Argos P, Abagyan R. A method to configure protein side-chains from the main-chain trace in homology modeling. J Mol Biol 1993;231:849–860.
20. Lee C, Levitt M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature 1991;352:448–451.
21. Holm L, Sander C. Fast and simple Monte Carlo algorithm for side-chain optimization in proteins: application to model building by homology. Proteins 1992;14:213–223.
22. Hellinga HW, Richards FM. Optimal sequence selection in proteins of known structure by simulated evolution. Proc Natl Acad Sci USA 1994;91:5803–5807.
23. Shenkin PS, Farid H, Fetrow JS. Prediction and evaluation of side-chain conformations for protein backbone structures. Proteins 1996;26:323–352.
24. Tufféry P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dynam 1991;8:1267–1289.
25. Pedersen JT, Moult J. Genetic algorithms for protein structure prediction. Curr Opin Struct Biol 1996;6:227–231.
26. 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 1994;239:249–275.
27. Lee C. Predicting protein mutant energetics by self-consistent ensemble optimization. J Mol Biol 1994;236:918–939.
28. Dunbrack RL, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein side-chains. Nat Struct Biol 1994;1:335–340.
29. Bower MJ, Cohen FE, Dunbrack RL, Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
30. Xiang Z, Honig B. Extending the limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.
31. Gordon DB, Marshall SA, Mayo SL. Energy functions for protein design. Curr Opin Struct Biol 1999;9:509–513.
32. De Maeyer M, Desmet J, Lasters I. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. Fold Des 1997;2:53–66.
33. Gordon DB, Mayo SL. Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure 1999;7:1089–1098.

34. Samudrala R, Moult J. Determinants of side chain conformational preferences in protein structures. Protein Eng 1998;11: 991–997.
35. Desmet J, Lasters I, Vlieghe D, Boutton C, Stas Ph, Spriet J. Method for generating information related to the molecular structure of a biomolecule. 2001;WO 01/33438 A2.
36. Delhaise P, Bardiaux M, Wodak S. Interactive computer animation of macromolecules. J Mol Graph 1984;2:103–106.
37. Brooks BR, Bruccoleri R, Olafson D, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. J Comput Chem 1983;4: 187–217.
38. Petrella RJ, Lazaridis T, Karplus M. Protein sidechain conformer prediction: a test of the energy function. Fold Des 1998;3:353–377.
39. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
40. Desmet J, Lasters I, Vlieghe D, Boutton C, Stas Ph. Apparatus and method for structure-based prediction of amino acid sequences. 2001; WO 01/37147 A2.