PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

An energy-based conformer library for side chain optimization: Improved prediction and adjustable sampling

Sabareesh Subramaniam and Alessandro Senes*

Department of Biochemistry, University of Wisconsin-Madison, Madison Wisconsin 53706

ABSTRACT

Side chain optimization is a fundamental component of protein modeling applications such as docking, structural prediction, and design. In these applications side chain flexibility is often provided by rotamer or conformer libraries, which are collections of representative side chain conformations. Here we demonstrate that the sampling provided by the library can be substantially improved by adding an energetic criterion to its creation. The result of the new procedure is the Energy-Based library, a conformer library selected according to the propensity of its elements to fit energetically into natural protein environments. The new library performs outstandingly well in side chain optimization, producing structures with significantly lower energies and resulting in improved side chain conformation prediction. In addition, because the library was created as an ordered list, its size can be adjusted to any desired level. This feature provides unprecedented versatility in tuning sampling. It allows to precisely balance the number of conformers required by each amino acid type, equalizing their chances to fit into structural environments. It also allows to scale the amount of sampling to the specific requirement of any given side optimization problem. A rotameric version of the library was also produced with the same method to support applications that require a dihedral-only description of side chain conformation. The libraries are available at http://seneslab.org/EBL.

Proteins 2012; 00:000–000.

Key words: side chain; rotamer library; conformer library; side chain optimization; protein design; modeling; structure prediction.

INTRODUCTION

Generally stated, the goal of side chain optimization is to identify the most favorable configuration of the side chains for a given backbone. It is a fundamental component of most protein structure prediction and design applications. While the specific details may vary, side chain optimization generally involves four key elements (Fig. 1): (1) a backbone that provides a structural template; (2) a side chain library that provides conformational freedom to the variable positions; (3) a set of physical and/or empirical energy functions of statistical derivation for scoring; and (4) a search strategy to identify the lowest energy state among all possible configurations.

Side chain optimization poses a difficult challenge, as the search space grows combinatorially with the number of positions involved and their conformational freedom. The side chain library is essential to transform what is a continuum search space into a discretized problem for which a number of powerful deterministic or stochastic algorithms are available (such as Dead End Elimination,[1] Branch and Bound,[2] and Graph Theory,[3] Monte Carlo,[4] Self Consistent Mean Field[5,6]). It is important to remark that the library is key to the quality of the outcome. This is demonstrated in Figure 1. The theoretical target of the optimization procedure is the global minimum of the side chain conformational energy landscape, but the landscape is sampled only in a finite number of locations, while the rest remains unknown. The "winner" can approach the global minimum only if the correct side chain conformations were provided by the library. Therefore, the choice of a library predetermines—even before the search is started—the best possible accuracy of the procedure.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Alessandro Senes, Department of Biochemistry, University of Wisconsin-Madison, 433 Backcock Dr., Madison WI 53706. E-mail: senes@wisc.edu.

Received 12 December 2011; Revised 22 April 2012; Accepted 2 May 2012
Published online 10 May 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24111

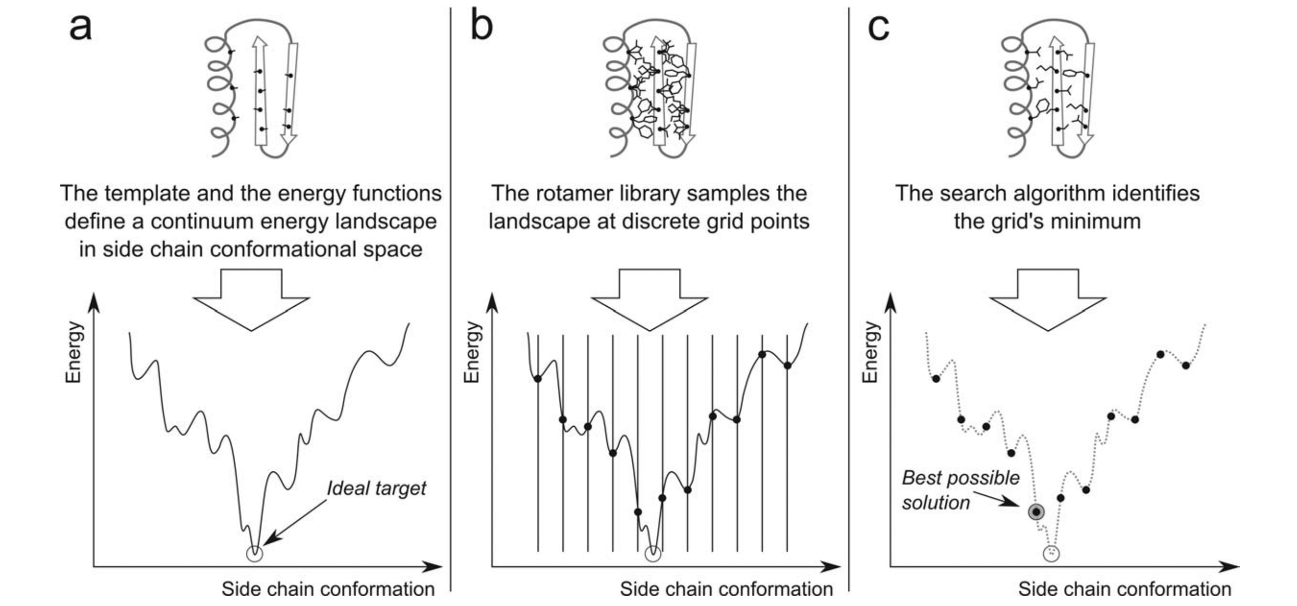
a
The template and the energy functions define a continuum energy landscape in side chain conformational space
Energy
Ideal target
Side chain conformation
b
The rotamer library samples the landscape at discrete grid points
Energy
Side chain conformation
c
The search algorithm identifies the grid's minimum
Energy
Best possible solution
Side chain conformation

Figure 1

The side chain library predetermines the best possible accuracy of a side chain optimization procedure. (**a**) The template and the energy functions define a multidimensional landscape (here schematized in 1-D) whose dimensions are degrees of freedom of the side chains. The global minimum of the landscape is the ideal target of the optimization. (**b**) The introduction of a side chain conformation library produces a grid that discretizes the space. (**c**) The search algorithm can identify the grid point with lowest energy. Depending on the choice of library this point may lie near or far from the global minimum of the entire landscape.

The most trivial way to increase accuracy would be increasing the size of the library, and indeed, it has been shown that high-sampling libraries can improve the outcome of side chain optimization.[7–11] Higher sampling, however, comes at a significant computational cost as the total number of states can easily reach an intractable number of combinations. Another possible solution to this problem is to adopt a continuous sampling of side chain space. The trade off is reduced efficiency but methods such as minDEE (minimized Dead End Elimination)[12–15] can be particularly suited for protein design applications in which correct prediction of the landscape global minimum is most critical. Nevertheless, library-based sampling is still very popular because it is simple to implement, it can be integrated with many algorithms and offers a good balance between speed and accuracy, which is essential, for example, when side chain optimization is repeated multiple times, such as in protein prediction methods. The questions that we ask here are: how do we improve the library accuracy without affecting its efficiency? Or, for applications for which speed is paramount, how do we improve the library efficiency without affecting its accuracy? The answer to both questions is to identify, for any given size of the library, the set of side chain conformations that will maximize its performance, which is the goal of this work.

Currently, the majority of the libraries used for side chain optimization are derivatives of statistical *rotamer* libraries,[16,17] such as the “Penultimate” library[18] and, most commonly, the backbone-dependent (BBD) library of Dunbrack,[19,20] which is still actively curated.[21,22] These statistical libraries are based on the analysis of the distribution of the amino acids’ χ angles (the torsional rotations around bonds), which are the main determinants of side chain conformation. The rotamer libraries define the clusters in torsional space, providing their average, dispersion and relative population. Figure 2(a) plots the rotamers for an amino acid that contains χ angles exclusively between sp^3 carbons (Leu) and one with an sp^2 carbon (Asn). The nine rotamers of Leu cluster at combinations of the classical staggered conformations, (near $-60°$, $180°$, $+60°$). The nine theoretical minima, however, are far from being evenly populated because some of the rotamers are disfavored by local conformational strains.[20] The tight clustering displayed by Leu side chains is not observed when the side chain torsions involve an sp^2 carbon, such as in the case of the χ_2 dimension of Asn, and the density is more dispersed.

The adoption of a rotamer library allows to focus the search only on the favorable regions of conformational space. However, the rotameric wells are generally too wide to be covered by a single conformation. Providing sufficient sampling is indeed a critical issue for side chain optimization because even small atomic clashes can prevent a favorable solution from being identified.[23,24] A commonly implemented scheme to increase sampling is

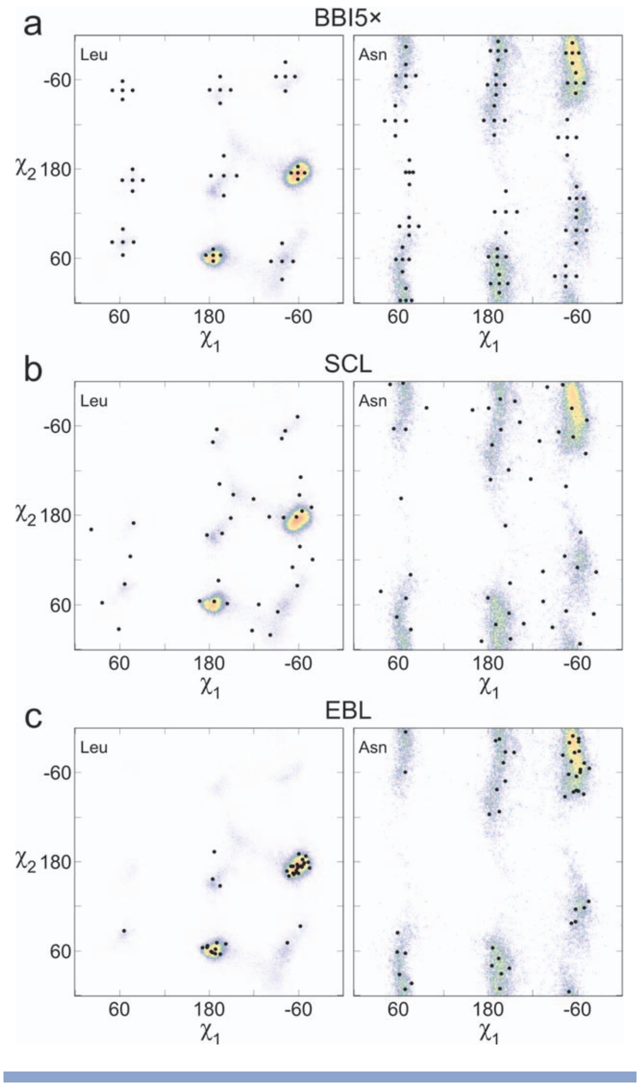

Figure 2

$\chi1/\chi2$ plot of an expanded rotamer library, a conformer library and the Energy-Based conformer library. (**a**) a 5× expansion of the Backbone Independent library in which each rotameric region has been evenly enriched with subrotamers that vary by ±1 S.D. in either $\chi1$ or $\chi2$ dimension. The figure shows the χ_1/χ_2 plot (black dots) for a side chain with torsions between two sp^3 carbons (Leu, $9 \times 5 = 45$ rotamers) and one characterized by a sp^2 carbon in the χ_2 dimension (Asn, $18 \times 5 = 90$ rotamers). The dots are overlaid on a color coded density map of the side chain distribution in the structural database. (**b**) χ_1/χ_2 distribution for the same two amino acids in the mid-sized (0.5 Å RMSD) conformer library of Shetty *et al.* Leu: 36 conformers. Asn: 48 conformers. (**c**) $\chi1/\chi2$ plot of the first 36 conformers of Leu and 48 conformers of Asn of the Energy-Based Library. The numbers are chosen to allow a direct visual comparison with the SCL (b). In the EBL the conformers are not evenly spaced but tend to cluster with a bias that is similar to the conformational distribution observed in the structural database. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to expand the main rotamers with a combinatorial addition of ±1 standard deviations in the χ_1 and χ_2 dimensions (for example, see Refs. [21], [23], [25], [26]), resulting into a nine-fold expansion of each rotameric center. Alternatively, expansions can be produced such that the addition of ±1 standard deviations is operated in the χ_1 or in the χ_2 dimensions, producing the five-fold expansion illustrated in Figure 2(a).[21] While rational, such expansion are in part arbitrary and do not consider the fact that the relative populations of these regions can range significantly, raising the question of what would be the most effective strategy. For example, the relative density of the nine clusters of Leu ranges from as high as 63% of the total in one rotameric region (−60°/180° cluster) down to a mere 0.02% of density in the least populated region (60°/−60° cluster). A distribution of sampling that somehow reflects this bias would likely be beneficial.

An alternative approach to side chain conformational sampling is the adoption of a conformer library.[7,9] These are collections of side chain conformations extracted from high-resolution structures. They are created from an exhaustive set of side chains that is reduced to a desired number by removing conformers that are too similar to each other using a filter based either on χ angle similarity[9] or on root mean square deviation (R.M.S.D.).[7] The conformer libraries do not involve clustering and expansion and are directly suitable for fine-grained sampling, as they can be created in different sizes by tuning the similarity filter. An advantage of conformer libraries is that they retain variation of all degrees of freedom, including, bond distances and angles, in addition to the dihedral angles. In particular, they capture any systematic bond angle variation occurring in sterically strained rotameric regions, which can be large enough to affect the energies.[9] It should be noted, however, that the application of a filter based on geometric similarity flattens the differences between the most populated and the rare regions. For example, while the Leu conformer library illustrated in Figure 2 does not sample the very rare 60°/−60° rotameric region, its relative coverage of the −60°/180° (63%), 180°/60° (30%), 180°/180° (2.6%) and −60°/60° regions (0.7%) is not proportional to their densities.

The rotamer and conformer libraries have been fundamental tools in protein modeling and design. Particularly the seminal backbone-dependent library is at the core of a number of modeling methods which have enabled important achievements in prediction and design (for example, see Refs. [1], [3], [26–32]). Their continued development is important to improve accuracy and reduce run time when applications require high throughput, high sampling, or when side chain optimization is repeated multiple times in concert with backbone motions. The expanded rotamer libraries and the conformer libraries are both based on the natural distribution of side chain conformations in proteins. Both approaches greatly reduce the size of the search space by providing good guidance on where sampling should be allocated, excluding any regions of conformational space that are energetically unfavorable. The question is up to what point the natural distribution of side chain conformation can inform how to best prioritize sampling within the rotameric regions, and what additional information could be used to improve the sampling strategies. An important consideration in this regard is that protein side chains are co-evolved with their environment to complement each other. Consistently, it has been observed that the conformational preferences of the amino acids in the structural database reflect primarily the internal steric constraints and local backbone interactions, and only marginally the effects imposed by the surrounding environment.[33] On the other hand, in side chain optimization the backbone is generally fixed, the side chains only have discrete mobility, and the structure is unable to undergo those small movements that would relax any minor clashes. Therefore, the primary factor that determines the probability that a given set of representative conformers will contain a fitting solution are the interactions of the side chain with the environment in which it is reconstructed. It follows that in order to maximize their chances of fitting, instead of using a pure geometric criterion either in cartesian or torsional space, it would be preferable to space the conformers evenly according to the energetic impact of their motions i.e. the likelihood that a motion would produce significant energy variation.

We hypothesized that the introduction of an energetic criterion into the selection of the conformer library would lead to more effective prioritization of sampling in side chain optimization. The relationship between side chain geometry and the energetic impact of their motions is complex and difficult to derive analytically, as they depend very specifically on their structures and the degrees of freedom altered. As illustrated in Supporting Information Figure S1, the energetic impact is related to the number of atoms that are displaced and also to the distance traveled by these atoms, which depends on their distance from the axis of rotation. For example, $\chi1$ rotations are likely to impact the energies more than $\chi2$ rotations because they translate more atoms and for a further distance. For the same reason, $\chi1$ rotations of the bulky Trp are more likely to impact the energies than $\chi1$ rotations of the smaller Leu. Therefore, it is clear that $\chi1$ should be allocated more sampling than $\chi2$, and that the bulky Trp should be allocated more sampling than Leu. The question is how much more? Here we address the problem with a practical approach based on the analysis of how an extensive library of conformers interacts with a wide variety of natural protein environments. The data is used to sort the conformers by their propensity to fit (energetically) into protein environments. We have compared the resulting library with three libraries from the literature and observed important performance improvements with the new approach, both in energetic terms as well as side chain conformation recovery. The approach also introduces a new beneficial feature: because the library is sorted, the number of conformers can be resized to any desired level of sampling. This feature provides unprecedented flexibility in adjusting, even dynamically, the combinatorial size of the optimization to match the precise needs and limits of a procedure.

MATERIAL AND METHODS

Structure database preparation

A collection of 2159 high resolution x-ray structures was obtained from the Protein Data Bank (PDB) using the following conditions: resolution <2.0 Å; deposition date: later than 01/01/1998; method: X-ray diffraction; molecule type: protein (no DNA, no RNA); no ligands. The proteins were filtered to allow no more than 30% sequence identity between individual chain. Hydrogen atoms were added with the program Reduce,[34] which also performed any necessary rotation of the hydroxyl groups, flipping the side chain of Asn, Gln, and His and determine the protonation state of His to optimize hydrogen bonding (*-BUILD −ROTEX* options). The three protonation states of His are referred here as His-δ (neutral, protonated on Nδ1), His-ϵ (neutral, protonated on Nϵ2), and His+ (doubly protonated and positively charged). The proteins were curated with an automated procedure that rebuilt missing side chain atoms, removed multiple side chain conformations, and converted any main chain missing amino acids into chain termini. All protein structures were then minimized with CHARMM[35] (using the CHARMM 22 potential), with 3 cycles and 50 steps of adopted basis Newton Raphson method using a harmonic potential with a force constant of 100 kcal mol^{-1} Å^{-2}. Minimization was required for two reasons. The bond lengths needed to be homogenized because differences in refinement methods create variability which is within experimental uncertainty but sufficient to produce significant energy penalties. Minimization can also resolve the occasional small clashes that may occur in poorly refined regions of the crystallographic models. The minimization procedure was selected to reduce these unwanted effects while preserving the natural conformation observed in the crystallographic models. The final RMSD of the crystallographic and minimized models was on average 0.05 Å. The differences in the side chain torsion angles are at most few degrees, and it has been previously demonstrated that premininization of the structures has no significant influence on side chain prediction.[9] A typical example of crystallographic and minimized models is shown in Supporting Information Figure S2.

Preparation of the fine grained (unsorted) conformer library

A set of 1000 proteins was randomly selected from the structural database for the creation of the conformer library and the selection of the environments. All side chains with a B-factor >= 40 and those with missing atoms in the original structure were not considered. Any side chain with a Cα to Cα distance below 8 Å from side chains with missing density was excluded in the selection of the environments. For each amino acid type, up to 5000 side chains were randomly selected as environments (except Cys: 1614; Met: 4296; and Trp: 2734). One single set of environments was selected for all protonation states of His. Up to 25,000 side chains were set aside for the creation of the initial (unsorted) conformer library. The conformers were selected at random and added to the conformer list if they had an RMSD >0.05 Å from all other previously collected conformers (RMSD filtering). A conformer library was created independently for the three protonation states of His-δ, His-ϵ, His+. Each conformer library was topped to 5000 conformers (except Cys: 1780; His-δ: 2906; His-ϵ: 4221; His+: 542; and Val: 3916). Cys residues in disulfide bonds were excluded from the analysis. For the creation of the rotameric version (dihedral only) of the Energy-Based library, the bond lengths and angles of the conformers were standardized to the standard values from CHARMM 22 topology prior to RMSD filtering. The remainder of the procedure followed was the same for both conformer and rotamer versions of the library.

Calculation of the conformer/environment interactions and creation of the energy tables

The energy data used to derive the library was collected in this phase. For each amino acid type, the native side chain of each of the *M* environments was remodeled as each one of the *N* conformers and their interaction energy was calculated, producing a matrix of $N \times M$ energies. The side chain reconstruction was performed from internal coordinates using the distance, the angle and the dihedral angle relationships relative to three preceding atoms (see Supporting Information Figs. S3 and S4). The interaction energies included the internal interactions of the side chain (including the bonded terms) and the interactions of the side chain with all other atoms. The energies were calculated according to the CHARMM 22 force field[36] (bond, angle, urey-bradley, dihedral, improper, van der Waals, and Coulomb electrostatics with an R-dependent dielectric), plus an additional hydrogen bond term as described in the program SCWRL4.[21] The nonbonded interactions were calculated with a distance dependent cutoff of 10 Å, using a switching function (cut-on 9 Å, cut-off 10 Å). The calculations were repeated with the van der Waals radii rescaled to 95 and 90% of their parameter 22 size. The table of energies were computed in three different conditions: (1) with full electrostatics and no hydrogen bonding term, (2) with a full hydrogen bonding term and no electrostatics and, (3) with electrostatics plus a hydrogen bonding term, both rescaled to 50%.

The $N \times M$ energy matrix was converted into an $N \times M$ boolean matrix in which a true value indicated that an element's energy was below a given threshold, and thus the environment was satisfied. Because the best energy achievable in each environment varied substantially, an environment dependent threshold was adopted. The threshold was calculated in the following way: first, the best interaction energy in the row (all conformers in the environment) was identified. All the elements of the row were adjusted by subtracting the best energy. The distribution of all the adjusted energy in the entire table was plotted. As shown in Supporting Information Figure S5, these distributions have a typical peak near the best energy. This peak represents conformers that are near the very best energy. The distance of the modal peak from the minimum is thus indicative of the typical energy spread of conformers that fit the environments favorably. For this reason, we chose the mode of this peak as the threshold to be added to the best environment energy. For example, Arg displays a peak at 7.0 kcal mol^{-1} from the best energy, thus the threshold for each Arg environment was set 7.0 kcal mol^{-1} above the best energy for that environment.

Creation of the sorted energy-based library

The fine grained conformer library was sorted by the propensity of its elements to fit in the largest number of natural environments, creating the Energy-Based library. The sorting procedure is schematically explained in Figure 3. For each amino acid type, the conformer that satisfied the largest number of environments was selected as the top conformer. All the environments satisfied by the first conformer were marked and no longer considered. The conformer that satisfied the largest number of remaining environments was then selected and the process was repeated. After each selection, however, the threshold was lowered and made more stringent: if an environment that was previously excluded was no longer satisfied at the lower threshold, it was put back into consideration. The process was repeated until all conformers were sorted. The threshold was scaled down linearly from its initial value to reach zero at the end of the sorting process.

Preparation of the benchmark libraries

Three previously published rotamer and conformer libraries of different sizes were selected for comparison. The 5× expansion of the 2010 version of the Backbone Dependent library[21,22] (here referred as BBD5×) was built with standard bond lengths and bond angles from CHARMM 22 topology. The rotamer library mean rotamers were expanded by ± 1 standard deviation in χ1

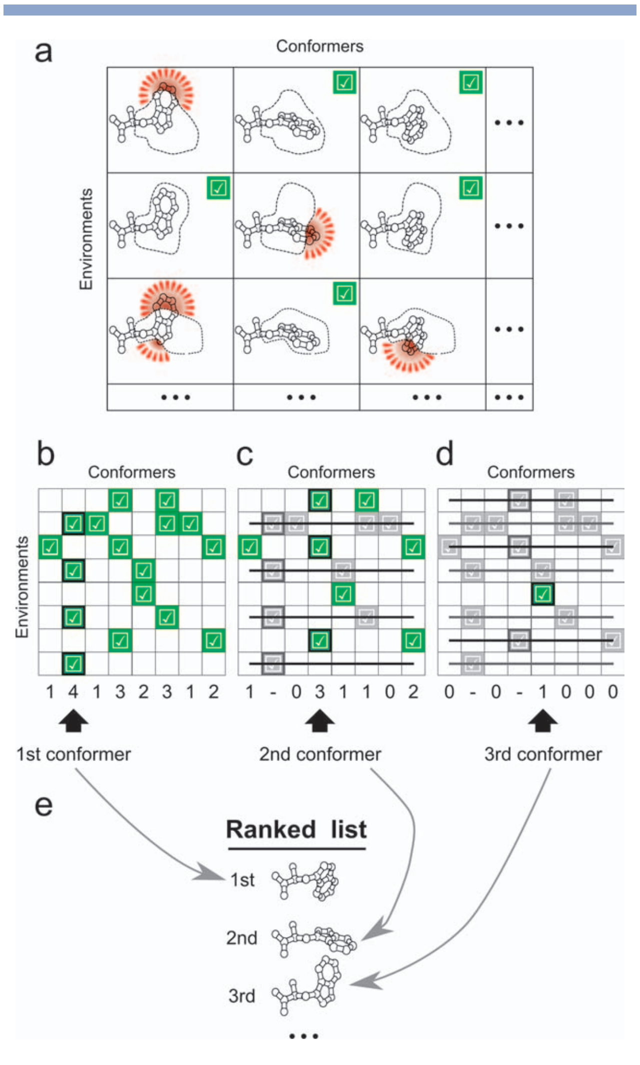

Figure 3

Procedure for the creation of the Energy-Based conformer library. (**a**) Each conformer of a fine-grained library of size N is built in each one of M of environments that contain the same side chain type (Trp in the figure) in protein crystal structures. The interaction energies of each conformer in each environment are calculated and if the energy is below a certain threshold, the conformer is considered a fit for the environment (illustrated as a green check mark in the figure). (**b**) The results are stored in a $N \times M$ boolean table. The number of environments satisfied by each conformer is determined (number under the table). The conformer that fits the largest number of environments is the first to be selected (black arrow). (**c**) The environments that were satisfied by the first conformer are no longer considered, and the procedure is repeated to find the conformer that would satisfy the most environments that are still uncovered. (**d**) The procedure is repeated until completion. (**e**) The resulting library is compiled as a ranked list, in which every additional element complements the previous. The major advantage of ranking the conformers is that it allows the user to truncate the library at any desired size, which is not possible with a traditional conformer library. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

or $\chi 2$ but not in both dimensions simultaneously. This led to a 5× expansion of amino acids with at least two χ angles, and 3× expansion of amino acids with a single χ angle. The dihedral relative to the hydrogen atom of hydroxyl groups was sampled at the canonical $-60°$, $180°$, and $+60°$ minima, each one expanded by $\pm 30°$ (9 total steps) for Ser and Thr, and every 45° (8 steps) for Tyr. The expansion generated a total of 3,755 conformers (Supporting Information Table S3). The two benchmark conformer libraries selected were the "medium" size library (0.5 Å RMSD) from Shetty *et al.*[7] (here referred as SCL) which contains 1547 conformers, and a small conformer library from Xiang and Honig[9] created from a database of 297 proteins, 100% coverage and 40° tolerance, which contains 1134 conformers (here referred as XCL). For consistency, and in particular to avoid any significant differences in the bonded energies of the conformers, both conformer libraries were subjected to constrained minimization (conformers were built and minimized in a Gly-X-Gly tripeptide).

Single side chain repack tests

Single side chain repack tests were performed on a set of 2000 environments obtained from 700 proteins that were set aside from the initial structural database for testing purposes. The test is similar to the conformer sorting procedure, in which the native side chain found in an environment is remodeled into a conformer and the interaction energies are calculated. Conformers were defined to satisfy the environment by the condition previously explained. An environment was defined to be satisfied by a set of n conformers if at least one of the elements satisfied the environment.

Complete protein repacks

Complete side chain repacks were performed on a subset of 560 of the 700 proteins set aside for testing purposes. All side chains were removed and predicted except Gly, Ala, and Pro. His residues were predicted using in the protonation state assigned by Reduce.[34] The optimization was performed with the program *repackSideChains* using a sequence of algorithms: first a run of Dead End Elimination (DEE) using Goldstein single criterion[37] was used to reduce the combinatorial space. A round of Self Consistent Mean Field (100 cycles, temperature 300 K) was performed on the conformers that were not eliminated during the DEE phase and the protein was set in the resulting most probable state. Finally a Monte Carlo simulated annealing procedure was run (50,000 cycles, with exponential cooling from 1000 to 0.5 K). The structure with the lowest energy identified by the Monte Carlo run was the final product of the optimization.

Analysis

Conformation prediction of the crystallographic side chain conformation (side chain conformation recovery) was performed by matching the $\chi 1$ and $\chi 2$ of the predicted and crystallographic structure with a tolerance of

40°. The analysis was performed on all side chains and on a subset of buried side chains. A side chain was defined buried if it had a Solvent Accessible Surface Area (SASA) below 25% of the maximum possible SASA for the side chain reconstructed into a Gly-X-Gly backbone (with X being the amino acid type under exam). The hydrogen bonding recovery was calculated as follows. First, all side chain-to-side chain and side chain-to-backbone hydrogen bonds were identified in the native structure if they had nonzero energies using the explicit hydrogen bonding function. A hydrogen bond was considered recovered if the interaction between the same donor and acceptor had nonzero hydrogen bonding energy in the predicted structure.

Programs

All calculations (modeling, energy evaluations, conformational analysis, SASA measurements, etc.) were performed with *ad hoc* programs written using MSL,[38] a C++ object oriented software library for molecular modeling and analysis, which is freely distributed under an open source license at http://msl-libraries.org. The total protein repacks were performed with the program *repackSideChains*, which is distributed with MSL.

RESULTS AND DISCUSSION

The energy-based conformer library

The Energy-Based library (EBL) is an extremely fine-grained conformer library sorted by the propensity of its elements to fit in a wide variety of natural protein environments. The procedure used to derive the library—explained in detail in the Methods section and illustrated in Figure 3—is the following:

1. A very finely grained library of N conformers is created for each amino acid type.
2. A large number M of environments that contain the same amino acid type is selected at random from high-resolution crystal structures.
3. The native side chain of each environment is remodeled into each of the conformers, and the interaction energy between the conformer and the environment is measured [Fig. 3(a)].
4. The data is collected in an $N \times M$ table of energies.
5. Each energy is converted to a boolean value, indicating if the environment is "satisfied" (True, if energy < threshold) by the conformer, or not satisfied (False).
6. The conformer that satisfies the largest number of environments is added to the library [Fig. 3(b)].
7. All environments satisfied by the conformer are marked and no longer considered.
8. The threshold is lowered by a small amount. Any previously satisfied environment that would no longer be satisfied at the new more stringent threshold is brought back for consideration.
9. The procedure is repeated from #6 until all conformers are sorted.

This procedure selects the conformers with the highest propensity to energetically fit in environments that contain the amino acid type in natural proteins. The first conformer selected is invariably a conformer near the center of the most populated region of side chain conformational space. The second conformer complements the first by covering another dense region, most often the center of the second most populated cluster. Step 7 is the key step that ensures this complementarity. Without step 7, the second and the other top conformers would most likely be very close structural neighbors of the first pick. By removing from consideration the environments satisfied by all previous conformers, the procedure ensures that each element extends the coverage to new areas of conformational space (the problem would be classified as a classical Set Cover Problem in complexity theory). The use of a variable threshold that becomes more stringent at every cycle (Step 8) allows to sample further between conformers as the library becomes larger.

The method is based on an energetic criterion for the selection of conformers, but it also incorporates two sources of natural conformational bias. The fine-grained conformer library—albeit being "flattened" by the application of a similarity filter—excludes any energetically unfavorable regions of the conformational space. The most important factor, however, are the environments. They were randomly selected and not filtered, and thus the environments reflect the natural conformational preferences of the amino acids side chains they contained. During the sorting process, the environments essentially "vote" for conformers, and are more likely to chose those that belong to the same rotameric region of the side chain they originally contained. A second important aspect that is likely to affect the selection process is how tightly packed the environments are around their side chain. The environments of surface exposed positions are more likely to accommodate a variety of conformers, voting more indiscriminately than those that belong to core positions. This aspect not only affects the selection of the conformers, but as discussed later (in the "sampling level" section) it also has important ramifications for balancing sampling between the various amino acid types.

Choice of energy functions

An important initial step in defining the procedure for the creation of the library was to identify a good choice of energy functions. All bonded terms (bond, angle, dihedral, improper terms, from the CHARMM22 parameter set[36]) were included to penalize conformers that are internally strained. The first function analyzed was the van der Waals function. A common practice in side chain optimization is to soften the repulsive component of this function, reducing the negative impact of any small clashes that may occur, under the rationale that they would be readily relaxed by small side chain and backbone motions in a flexible protein structure. This is often accomplished by the adoption of *ad hoc* functions and/or by rescaling the van der Waals radii.[24] Here we tested whether the use of reduced radii was beneficial for the creation of the library. The second issue tested was related to the electrostatics, salvation, and hydrogen bonding functions. These three inter-related forces are notoriously difficult to model in side chain optimization and their treatment varies widely between applications.[21,39–44] The simple inclusion of partial charges may not improve side chain prediction when the effect of solvent are unaccounted for.[9,39] Moreover, the hydrogen bond—a key factor in predicting the structural organization of protein folds and protein–protein interfaces—has a complex geometry dependency and is not well modeled by an integration of Coulomb and Lennard-Jones interactions.[44] To try to maximize the hydrogen bonding prediction capabilities of the library we chose to test three simple conditions: (1) pure coulombic interactions, (2) an explicit hydrogen bonding function without the electrostatic term, and (3) an equal weight of both terms. We selected the hydrogen bonding function implemented in the SCWRL4 program[21] because it is based on elements of the CHARMM force field and has multiple angle dependencies.

To test van der Waals radii rescaling, we created three separate conformer libraries using 100% (full), 95% and 90% radii. We tested the libraries in a series of procedures in which a single side chain was placed in fixed protein environment (referred here as "single side chains repacks"). In the procedure we determine what percentage of the environments was satisfied by each truncation of N conformers of the library (for $N = 1$ to the size of the library). A direct comparison of the performance of the resulting libraries, performed under all three conditions, that is, 100, 95, and 90% radii, revealed minor differences and did not identify a significant advantage in using rescaled radii. In the second test we found that the repacking efficiencies were similar but the hydrogen bonding recovery was higher when an explicit hydrogen bond function was used without electrostatics. These conditions were chosen for the remainder of the work.

The energy-based library

The Energy-Based library is a sorted conformer library of up to 5000 conformers for each amino acid type, except Gly, Ala, and Pro (Supporting Information Table S1). The three protonation states of His were treated separately because they are chemically distinct (referred here as His-δ: neutral, protonated on Nδ1; His-ϵ: neutral, protonated on Nϵ2; and His+: charged). A library of 5000 conformers per amino acid is exceedingly large but the sorted list can be truncated to any desired number of elements. Figure 2(c) shows a plot in χ1/χ2 coordinates of the top ranking elements of the EBL library of Leu and Asn. To allow a direct visual comparison with the conformer library of panel *b*, the number of conformers shown is identical. The EBL conformers are less evenly spaced and have a higher propensity to sample the most common regions, which is particularly evident in the case of Leu. To illustrate the precise order in which the conformers are ranked and how the algorithm initially prioritizes sampling of the most populated area and gradually extends coverage, Figure 4 shows a "walk" through the first 12 conformers of Trp. The first conformer lands in the center of the −60°/90° region (35% of the total density), which is sampled six times within the first 12 conformers. The library then visits the second most populated region (−180°/90°, 14% of the density) and remaining conformers gradually extend sampling, roughly in the order of the relative density of the clusters. The structural superimposition of the first twelve Trp conformers in Figure 4(b) shows how the conformers complement their coverage of tridimensional space. Even the closely spaced conformers that belong to the −60°/90 region are sufficiently shifted by χ1 and χ2 variations (and in parts also by bond angle variations) to cover different portions of tridimensional space.

Selection of benchmark libraries for testing

To test the performance of the Energy-Based library we selected three previously published libraries: (1) a 5× expansion of the Backbone Dependent rotamer library of Dunbrack[20,22] (BBD5×); (2) a medium-size conformer library from Shetty *et al.*[7] (SCL, for Shetty Conformer Library); and (3) a small conformer library from Xiang and Honig[9] (XCL). The benchmarks were chosen based on their sizes, to compare the performance of the EBL over a range of sampling levels. The Backbone Dependent library is the most popular rotamer library in the literature. The expansion scheme adopted is the one implemented in SCWRL4,[21] in which "sub-rotamers" are added by expanding either χ1 or χ2 by ±1 S.D. but not both at the same time (illustrated in Fig. 2). The version of the BBD library adopted is the most recent,[22] as it demonstrated significant performance enhancement compared to the previous version[20] in our preliminary tests. The BBD5× contains a total of 3755 conformers, which were built with the bond lengths and bond angles defined in the CHARMM 22 topology file. The SCL selected was the intermediate possibility (0.5 Å R.M.S.D. similarity filter), which contains a total of 1549 conformers. The conformer libraries of Xiang and Honig are at the core of SCAP and Jackal, a suite for modeling and

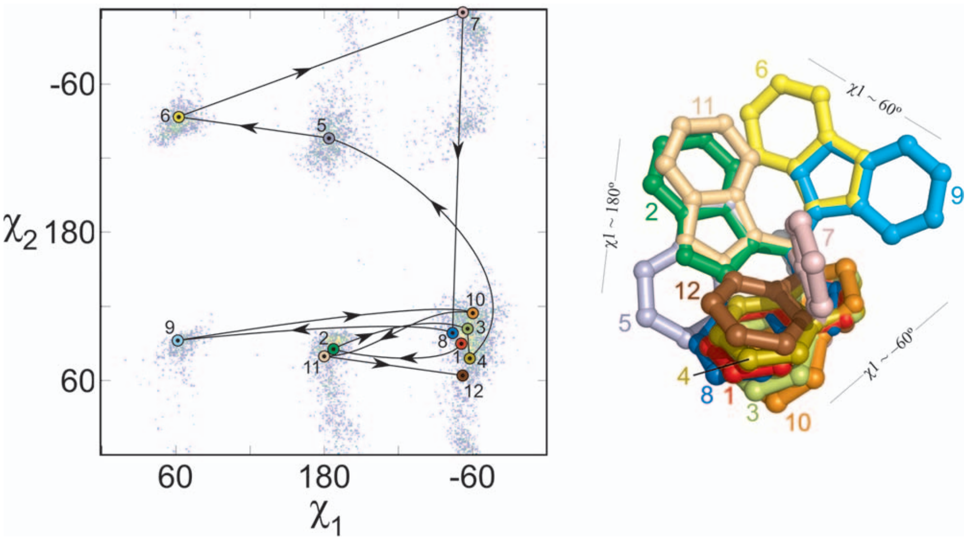

Figure 4

A "walk" in Trp space. (**a**) χ1/χ2 plot of the first 12 conformers of Trp overlaid on a color coded density map of the side chain distribution in the structural database. The most populated region (−60°/90°) is sampled multiple times while the coverage gradually extends to the other regions of density. (**b**) Structural representation using the same color coding of panel a. The figure demonstrates how the conformers are arranged to cover complementary regions of tridimensional space.

analysis.[9,41] Among the many possible sizes, we selected the library derived from 297 proteins with 100% coverage and 40° bins totaling 1136 conformers, which provided the opportunity to test the EBL at a relatively low level of sampling.

Performance test using single side chain repacks

To test the performance of each individual amino acid of the new library against the benchmarks we first performed a series of single side chain repacks in fixed protein environment using a set of proteins that was set aside for testing purposes. The results are shown in Figure 5. The histograms show the total number of EBL conformers that are necessary to match the performance of the benchmark library. The data was obtained in the following way: for each individual amino acid we calculated the fraction of protein environments that was satisfied by at least one of the conformers of the benchmark library and then we determined the number of EBL conformers that were necessary to satisfy the same fraction of environments. In all cases the difference in performance is extremely significant. All across the three comparisons, the sampling requirements of the EBL are always lower, often by a factor or 10 or more and always by at least a factor of 2. These results demonstrate that the EBL conformers have a high propensity to fit into protein environments that should be able to accommodate the side chain, which was the original premise behind the selection procedure. The next question was whether the improved performance would also be observed in side chain optimization procedures in which multiple side chains are modeled at the same time.

Performance test using total protein side chain predictions

We tested the Energy-Based library in a series of protein side chain prediction runs in which all positions in a protein (excluding Gly, Ala, and Pro) were remodeled using side chain optimization (referred here as total protein repacks). The scatter plots in Figure 7(a) compare the final energies of a set of 560 proteins after optimization with the EBL against the three benchmarks. In these repacks the number of conformers for each individual amino acid was exactly matched to the benchmark library. In all comparisons the majority of the points lay above the diagonal (97.3% against the BBD5×, 88.8% against the SCL, and 73.1% against the XCL), demonstrating that the EBL is much more likely to reach lower energy solutions. For ease of comparison, the energies are plotted after subtracting the "crystal energy," that is the energy of the native structure after constrained minimization. The crystal energy is used here solely as a convenient reference under the assumption that in many cases—but certainly not in all cases—the minimized crystal structure is devoid of strains and represents a good target for an optimization. Panels *b* of Figure 6 represents the same data of panel *a* in histogram form. This view highlights the distribution of energies obtained with the different libraries with respect to the crystal energy (zero

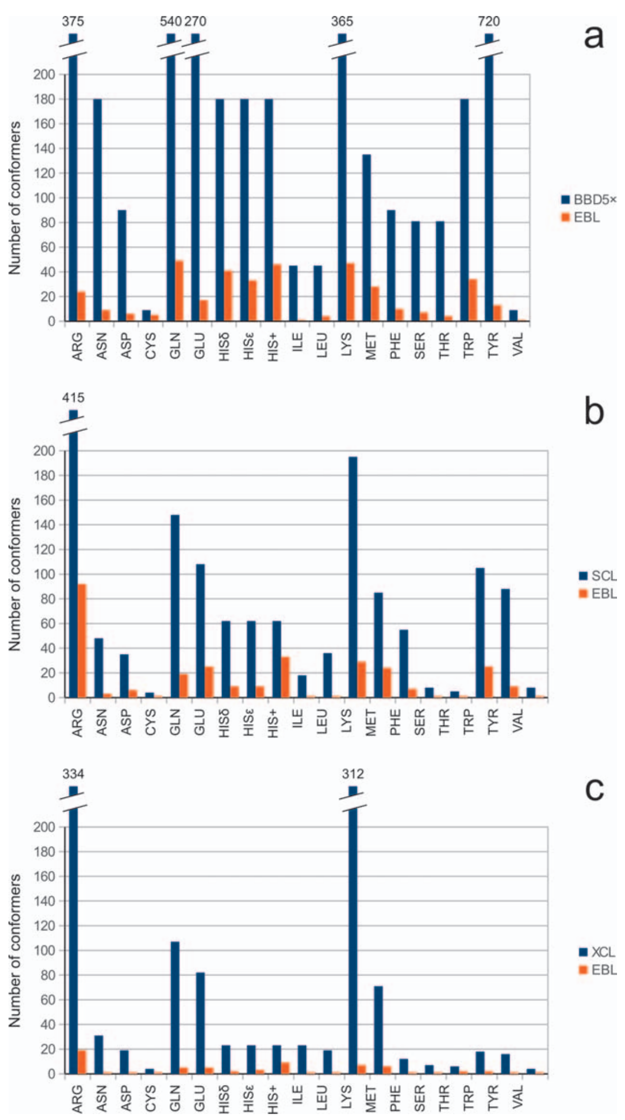

Figure 5

Performance test on single environment repacks. The figure compares the number of conformers that are required for equivalent performance between the Energy-Based library and three benchmark libraries (BBD5×: 5× expansion of the 2010 backbone-dependent library; SCL: a medium conformer library from Shetty *et al.*; XCL: a medium conformer library from Xiang and Honig). The number of conformers of the benchmarks is a fixed number (blue bar). We determined the fraction of environments that are satisfied by at least one of these conformers. The red bar represents the number of EBL conformers that are required to satisfy the same fraction of environments. For example, the XCL has 334 Arg conformers, which satisfy 55.0% of Arg environments. It takes only 19 conformers of the EBL to satisfy at least the same fraction. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

value). In the three sets of calculations the EBL fares well, with a number of solutions below the crystal energy (87.3%, 70.8%, 32.0%) that is significantly higher than the respective benchmarks (BBD5× 13.3%, SCL 10.3%, and XCL 1.0%, respectively). The modes of the energy distributions are shifted by hundreds of kcal/mol compared to those of the benchmarks. It should also be noted that even at the lowest level of sampling (1136, the same number of conformers of the XCL) the EBL produces a number of proteins below crystal energy that is greater than the larger SCL (1549 conformers) and the BBD5× (3755 conformers). The data demonstrates that the introduction of an energetic criterion in the creation of the conformer library greatly enhanced the energetic performance of the library in side chain optimization.

Sampling levels

In the first performance test, the number of conformers of the Energy-Based library was matched exactly with the respective benchmark library. However, since the number of conformers can be adjusted to any desired number, it is possible that the optimal distribution of sampling between the various amino acid types could be different. We addressed this question using data from the single side chain repacks against fixed protein environments. We determined the number of conformers that are required to satisfy a certain percentage of protein environments in the test set. This led to the creation of a series of "sampling levels" which, at least in principle, should provide each amino acid type with an equal chance to fit into protein environments. We created 14 levels, from very sparse sampling (282 total conformers in the 60% level) up to very high sampling (6985 conformers in the 99% level). The percentage refers to the number of environments that are energetically satisfied by the set of conformers in single side chain repacks. These levels will be referred as SL60 (Sampling Level 60%) to SL99. The number of conformers in each sampling level is reported in Table I. The balance within each level is consistent with the expectation that larger amino acids would require more sampling than smaller amino acids. A second factor that presumably contributes is the propensity of an amino acid type to occur in tightly packed positions (which likely require more sampling) versus solvent exposed positions (which can be satisfied by a larger variety of conformers). A substantially larger number of conformers is given to amino acids with hydroxyl groups compared to other amino acids with similar structure (for example, Thr vs Val, and Phe vs Tyr), an indication that a significant amount of sampling is required to satisfy hydrogen bonding.

Complete protein repacks using the EBL sampling levels

The total repack tests of the same set of 560 proteins were repeated using the sampling levels. The results are shown in Figure 7(c,d). Since the number of conformers for each individual amino acid is no longer matched to the benchmark libraries, to ensure a fair comparison we

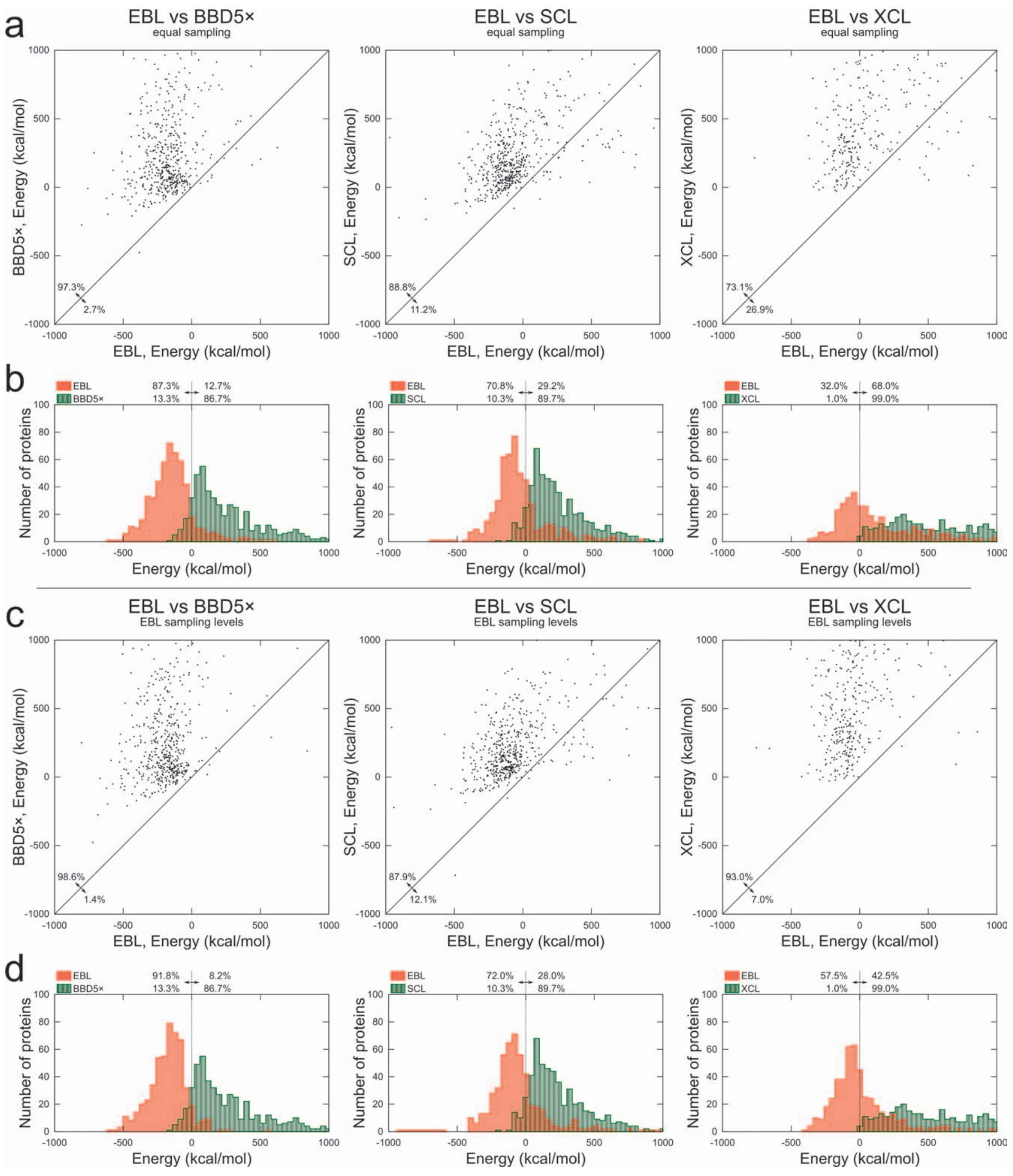

Figure 6

Performance of the Energy-Based library in total protein repacks. (**a**) The scatter plots graph the final energy after optimization of all side chains in 560 proteins, for the Energy-Based library (EBL, x axis) and three representative rotamer and conformer libraries (see Methods). The majority of the points lie above the diagonal indicating that the EBL on average achieves better performance that the benchmarks. For easier comparison energies are plotted after subtracting the energy of the minimized crystal structure. (**b**) Representation of the same data as histograms. The dashed line separates the proteins that score better than the crystal energy (percentages indicated), a convenient reference under the assumption that in most cases it represents a good target for an optimization. In a and b the calculations were made with an equal number of conformers compared to the benchmark for each amino acid type (equal sampling). The BBD5× has 3,755 conformers, the SCL 1,549 and the XCL 1,136. Panels c and d report the results of the same calculation performed using EBL sampling levels of similar total complexity of that of the benchmarks. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

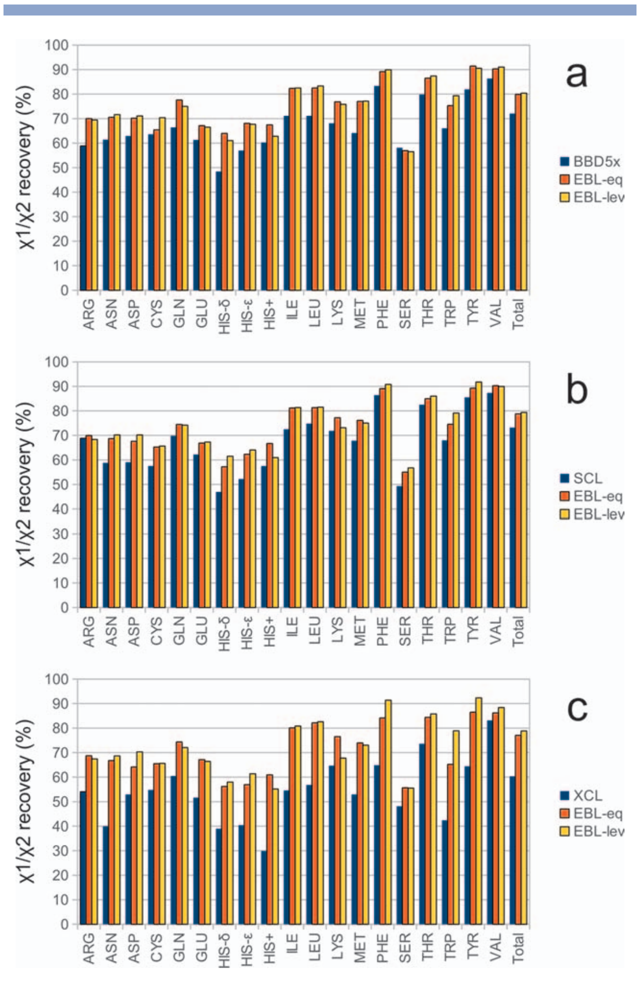

Figure 7

Recovery of the crystallographic side chain conformation in total protein repacks. Recoveries obtained with the EBL are compared to (**a**) the BBD5×, (**b**) the SCL, and (**c**) the XCL. The data is expressed as χ1+χ2 recovery with a tolerance of ±40° for buried side chains (<25% SASA). The orange bar represents the recovery in repacks made with an equal (eq) number of conformers compared with the benchmark for each amino acid type. The red bar represents the recovery in side chain optimizations made using an EBL sampling level (lev) of similar complexity with respect to the benchmark. With very few exceptions, the EBL performs better than the benchmark, often significantly. In Supporting Information Figure S6 the data is dissected by χ1, χ1+χ2, χ1+χ2+χ3, χ1+χ2+χ3+χ4 recoveries. The data relative to all positions, independently of burial, is shown in Supporting Information Figure S7. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

matched the total combinatorial complexity of the search space, that is the product of the number of conformers given to each position. The total combinatorial complexity of each optimization was calculated for the benchmark library, and the same protein was repacked with the largest EBL sampling level that did not exceed the benchmark's total complexity. It should be noted that this criterion always puts the EBL at a disadvantage—at times minimally, at times significantly—ensuring a stringent test. The most frequently selected levels were (in order of frequency) the SL95, SL96, and SL92.5 against the BBD5× library, the SL85, SL87.5, and SL82.5 against the SCL, and the SL80, SL75, and SL82.5 against the XCL. This discussion will refer to this strategy as "sampling levels" and to the previous strategy, in which benchmark and EBL were equally matched, as "equal sampling."

Optimization with the sampling levels produced a significant improvement of the energies in the comparisons against two of the three libraries, the BBD5× and XCL. In the comparison against the BBD5× library, the fraction of proteins below the crystal energy increased from 87.3% to 91.8%. This improvement can be appreciated visually by comparing the frequency of proteins just above zero energy in Figure 6(b,d). The most significant improvement was found in the comparison against the smallest of the three libraries, the XCL. In this case the number of proteins below crystal energy almost doubled, going from 32.0% to 57.5%. The introduction of sampling levels was not as beneficial in the test against the SCL (the number of proteins below crystal energy increased from 70.8% to 72.0%). It should be noted, however, that the adopted strategy to compare with the "equal sampling" puts the "sampling levels" strategy at disadvantage and likely contributes for this small improvement. This can be appreciated by comparing the total size of the most commonly used sampling levels. The main sampling levels used against the SCL (SL85, SL87.5, and SL82.5 levels, with 1231, 1464, and 1039 conformers, respectively) are all smaller in size than the size of the "equal sampling" strategy (1549 conformers).

In fairness, it should also be noted that the overall size of the BBD5× can be reduced by excluding the most rare rotamers from the library.[3,21] If this filtering is applied to maintain at least 99% of the cumulative density, the total complexity of the BBD5× in total repacks decreases approximately to the same level of the SCL. The application of a 90% filter reduces the BBD5× to approximately the XCL complexity. If we compare the energetic performance of the EBL at the reduced sampling levels of the SCL and XCL [Fig. 7(d), center and right panels, red areas] against the full-size BBD5× [Fig. 7(d), left panel, green area] we observe, however, that the smaller size EBLs still maintain a significant edge against the full size BBD5×. This advantage is likely in part due to the fact that with the EBL some representation of these rare areas can still be maintained while sampling is gradually reduced, while with a transitional rotamer library entire rotameric regions need to be completely removed.

Recovery of crystal structure conformation

After establishing that the Energy-Based library performs well in total protein repacks from an energetic stand point, we investigated if the performance translated

Table I

Suggested Sampling Level

Level	ARG	ASN	ASP	CYS	GLN	GLU	HIS-δ[a]	HIS-ϵ[a]	HIS+[a]	ILE	LEU	LYS	MET	PHE	SER	THR	TRP	TYR	VAL	Total
60%	27	16	12	3	7	8	14	11	3	4	6	4	12	30	3	4	43	72	3	282
70%	52	28	22	4	13	18	28	20	5	7	9	7	19	48	6	7	66	126	3	488
75%	73	38	31	6	19	28	39	26	6	9	13	10	25	60	8	10	83	167	4	655
80%	102	51	43	7	27	41	52	37	9	11	17	16	33	76	13	13	111	222	5	886
82.5%	123	60	49	8	33	50	63	44	10	13	21	21	39	87	16	16	126	254	6	1039
85%	149	70	59	9	40	61	76	53	13	16	26	26	47	100	21	20	144	294	7	1231
87.5%	177	83	70	11	49	73	95	63	17	20	32	34	57	116	27	27	164	341	8	1464
90%	222	100	86	16	61	90	121	76	22	26	39	44	72	138	34	34	196	408	10	1795
92.5%	273	122	106	21	77	111	152	94	30	35	52	58	93	169	43	44	237	487	13	2217
95%	354	154	138	32	106	144	226	125	44	50	70	81	126	214	57	62	298	613	16	2910
96%	397	173	152	39	121	163	293	150	51	60	82	94	144	242	65	74	340	687	19	3346
97%	449	201	173	48	143	182	424	172	71	78	98	111	176	278	74	89	382	767	22	3938
98%	498	227	201	71	177	215	649	255	89	108	124	132	206	342	88	108	430	838	28	4786
99%	589	292	231	100	233	261	1107	569	140	165	183	184	281	428	111	152	663	1252	44	6985

The table reports the suggested number of conformers for each amino acid type at different sampling levels. The levels have been obtained by matching the efficiency of repacking single side chain environments. For example, the top 27 Arg conformers satisfy on average 60% of Arg protein environments, and 589 are required to satisfy 99% environments. The top 16 and 292 conformers of Asn provide roughly the same chances to satisfy Asn protein environments.

[a] The conformers of His are created separately for the three protonation states, indicated here using the following naming convention: HIS-δ, protonated in Nδ1; HIS-ϵ, protonated in Nϵ2; HIS+, doubly protonated, charged.

to improved prediction of side chain conformation. Figure 7 shows the recovery of the side chain crystallographic conformations in the 560 total repacks (buried positions only, χ1+χ2 recovery, with a tolerance threshold of 40°). In the conditions tested the EBL recovers on average nearly 80% of all side chain conformations, ranging from about 55% (Ser) to 90% (Phe, Tyr, and Val). In all three comparisons, the EBL performs better than the relative benchmark, by +8% against the BBD5× library, by +6% against the SCL, and by a substantial +18% margin against the smaller XCL. The use of sampling levels (EBL-lev) resulted in a slight improvement of the recoveries compared with "equal sampling" (EBL-eq).Comparing the performance of the EBL in the three trials, it is remarkable that the total χ1/χ2 recovery is already high at the lowest sampling levels (78.8% recovery in the test against the XCL) and does not further grow substantially (79.4% against the SCL and 80.3% against the BBD5×). In comparison, the energies significantly improved at every increase of sampling size (Fig. 6). This is an interesting finding that suggests that the lowest sampling levels could be effective in side chain optimization, particularly if used with a softened van der Waals function. The R.M.S.D. analysis of the repacked structures compared to the native structures provides further confirmation that the EBL achieves superior performance in side chain prediction. The R.M.S.D. obtained with the EBL is significantly lower than that observed with the benchmarks (1.55 Å to 1.85 Å, Table II) even at the lowest sampling level (1.38 Å). The data relative to χ1/χ2 recovery of all side chains (independently of burial) is shown in Supporting Information Figure S7. As expected the average recovery drops significantly compared with the buried positions (66% for all three tests) but the overall trend remains similar.

Table II

Average Root Mean Square Deviation of Total Repacks Compared with the Native Crystal Structure

	Benchmark	EBL, equal sampling	EBL, sampling levels
BBD5×	1.63 ± 0.36 Å	1.38 ± 0.35 Å	1.35 ± 0.33 Å
SCL	1.55 ± 0.32 Å	1.41 ± 0.36 Å	1.37 ± 0.33 Å
XCL	1.85 ± 0.34 Å	1.48 ± 0.34 Å	1.38 ± 0.33 Å

The table reports the average RMSDs (± standard deviation) between the predicted side chains of each protein and the native crystal structure. The set includes only buried side chains (<25% SASA) and all atoms (heavy and hydrogen atoms).

Figure 8 examines the hydrogen bond recovery in the protein repacks. The figure reports the fraction of the hydrogen bonds present in the original structure that were correctly predicted. In the three tests, the Energy-Based library recovers between 47% and 60% of the crystallographic hydrogen bonds. Unlike the χ1/χ2 recoveries, here we observed an improvement as sampling increases (47.5%, 52.5%, and 60.0% against the XCL, SCL, and BBD5×, respectively). It should be noted that while not all hydrogen bonds are correctly predicted, the total number of hydrogen bonds in the repacks exceeds that of the crystal structures (Supporting Information Fig. S8), which is likely a consequence of the absence of solvent (implicit or explicit) in the calculations. Once again, the new library's performance demonstrated to be outstanding. Compared with the benchmarks, the total recovery was higher by +12% (BBD5×), +15% (SCL), and +25% (XCL), an improvement that is even more marked than what is observed for the χ1/χ2 recoveries. Although in all three tests the total recovery is very similar for the "equal sampling" and "sampling levels" strategies, at the level of the individual amino acids there were noticeable differences.

Overall the three measures—energy, side chain conformation recovery, and hydrogen bond recovery—depict an

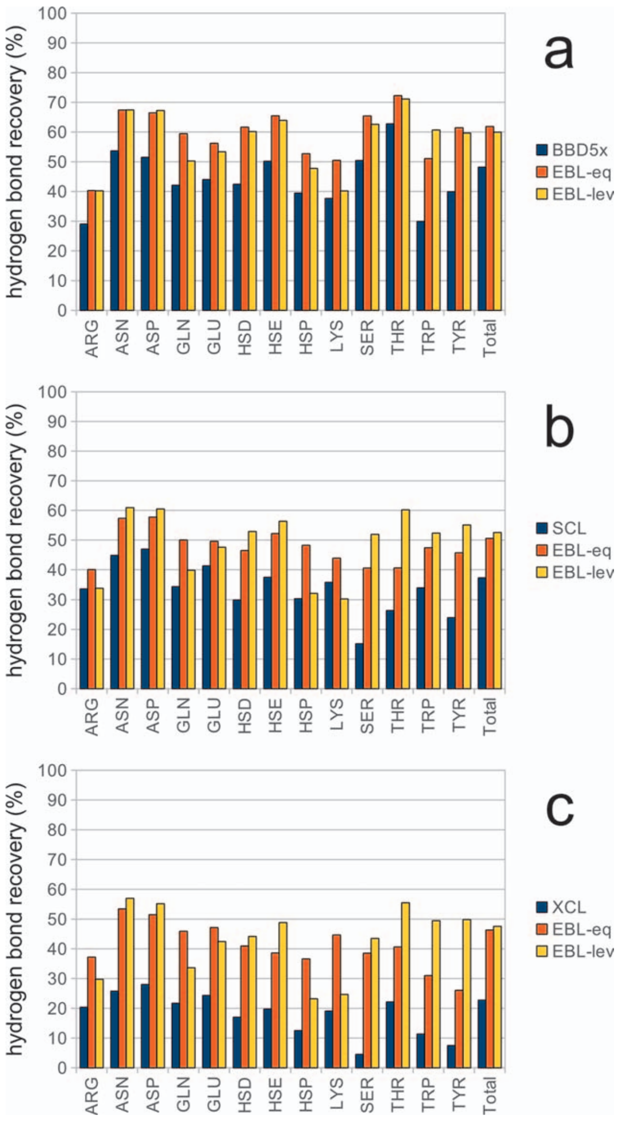

Figure 8

Recovery of the crystallographic hydrogen bonds in total protein repacks. Recoveries obtained with the EBL are compared to (**a**) the BBD5×, (**b**) the SCL, and (**c**) the XCL. The data indicates if a hydrogen bond present in the original crystal structure is recovered after side chain optimization. Any hydrogen bonds that are observed in the repacked structure but are not present in the original structure were not considered. The EBL demonstrated significantly better recoveries in all three cases. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

extremely favorable portrait of the Energy-Based library. The library displays superior performance in side chain optimization across a range of sampling levels. This efficiency, combined with its unique flexibility in rescaling sampling, means that the library can be a powerful and versatile tool that can be tailored precisely to improve quality and/or decrease run time in side chain optimization procedure.

Analysis of a sampling level

The three sets of total protein repacks demonstrated that increase in sampling produces substantial decrease of the energy. The 14 sampling levels proposed here, from SL60 to SL99, vary in the total number of conformers by a factor of 25 (Table I). To gain more precise information on their relative efficiency, we performed a series of total repacks systematically at each individual level on a random subset of 40 proteins. Figure 9 summarizes the results of this trial. In the figure the highest level (SL99) is chosen as the reference. The histograms shows the number of protein that reached an energy within a threshold of 2, 10, or 20 kcal mol^{-1} from the energy obtained with the SL99. The data indicates that to obtain approximately a 50% chance that a protein energy is within 50 kcal mol^{-1} of its SL99 energy, one should adopt at least the SL90 level (1795 conformers). At the SL95 (2910 conformers) about half of the proteins are within 10 kcal mol^{-1} from the best level. To obtain the same proportion below the 2 kcal mol^{-1} threshold the levels required are the SL97 (3938) or the SL98 (4786 conformers). The most important observation, however, is that the levels display a continuum and relatively smooth increase in performance. Although it is likely that an intensive analysis of performance based on total protein repack data would lead to the creation of even more effective levels, the data demonstrates that the proposed levels based on single side chain repacks are a suitable option.

An energy-based rotamer library

The method used for the creation of the Energy-Based conformer library can also be applied for the generation of a dihedral-only rotamer library. To create this library we standardized the bond lengths and angles of the

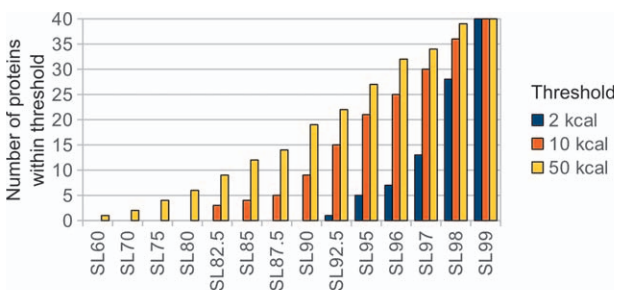

Figure 9

Comparison of the energetic performance of the EBL sampling levels. Forty proteins were repacked at each of the 14 proposed levels (from SL60 to SL99). The figure shows the number of proteins that had an energy below the energy of its SL99 optimization plus a threshold of 2, 10, or 50 kcal/mol. The results demonstrate a gradual increase in performance between the levels. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

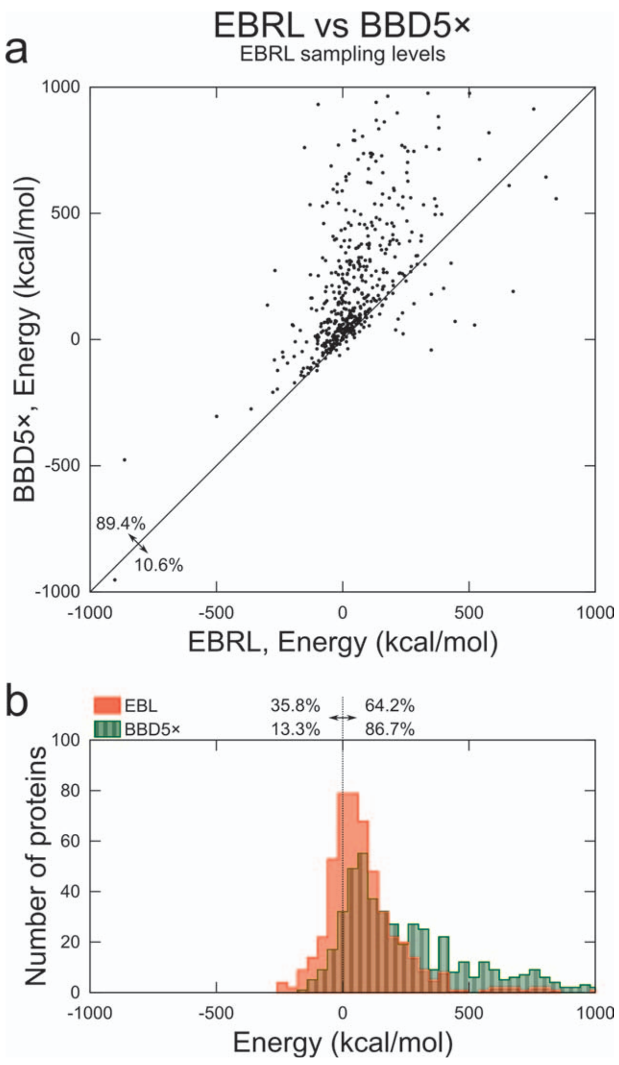

Figure 10

Performance of the Energy-Based rotamer library (EBRL) in total protein repacks. (**a**) The scatter plots and (**b**) histograms representation of the final energy after side chains optimization of 560 proteins, for the Energy-Based rotamer library compared to the BBD5× library. While the performance of the rotameric version of the Energy-Based library is decreased compared to the conformer library (Fig. 6), it represents an efficient alternative for applications that required a dihedral-only representation of the library. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

extremely fine grained library, and followed the same procedure (RMSD filtering, rebuilding in protein environments, energy-based sorting). Figure 10 shows a comparison of the performance of the Energy-Based Rotamer library (EBRL) against the BBD5× library. Although the performance of the EBRL is significantly decreased with respect to the conformer version, the rotamer version compares favorably against the benchmark. The EBRL scores better energies in 89.4% of the proteins compared to the BBD5× library. The percentage below crystal energy drops from 91.8% to 35.8%, but it remains significantly higher than the benchmark. The average side chain recoveries also drop (Supporting Information Fig. S9) but they still compare favorably against the benchmark. The data indicates that the conformers approach is substantially more effective than the use of rotamers. Nevertheless, the Energy-Based Rotamer library can be a useful alternative for applications that could benefit from the efficient and flexible sampling offered by an energy sorted library but require the use of a dihedral-only rotamer representation.

CONCLUSIONS

We have presented a new type of conformer library for protein modeling that introduces a number of innovations in side chain sampling. The library is in essence a sorted fine-grained conformer library. The library is so large that it needs to be trimmed down for most practical purposes, although an application that requires extremely precise positioning and is not of highly combinatorial nature could benefit from its exhaustive sampling. The method for sorting the Energy-Based library takes into account not only the conformational propensities of the side chains but also the nature of protein environments that host them. The selection of the conformers was made with an energetic criterion, under the hypothesis that using the same metric that selects the “winner” in a side chain optimization procedure would lead to a more efficient distribution of sampling. The results demonstrate that the strategy indeed provides important performance improvements.

The fact that the library is sorted and can be resized to any desired number represents *per se* a unique and important new feature. It introduces an unprecedented level of versatility in adjusting conformational sampling to match the specific needs of a modeling procedure. It allows to control the quality of the outcome and to meet any speed or memory requirements. The scalability of the library is also important for balancing the relative amount of sampling given to the different amino acid types and equalize their chances to fit in spaces that should accommodate them. Here we propose a series of sampling levels that gradually increase the library’s granularity while maintaining the mentioned balance. The versatility of the library can also be an important asset for developing more effective side chain sampling strategies. We have recently shown that that transferring sampling from positions that are likely to be satisfied by a variety of conformers (such as a relatively isolated solvent exposed position) to those that require a conformer from a very narrow and specific range (such as a tightly packed core position) can improve the economy of the calculation and the resulting energies.[45] The scalability of the EBL enables this and other similar strategies, opening new avenues for further improving performance in side chain optimization.

Compared with the current libraries, the EBL achieves significantly lower energies in side chain optimization. This is certainly a positive finding as it indicates that the EBL is very effective in exploring the energy landscape (Fig. 1). For this reason the library could aid the continued development of effective energy functions for protein prediction and reduce the need for artificial softening of the van der Waals function.[16,24,46] The fact that the EBL was tested with the same energy functions used for its creation may raise a concern that the performance could be in part due to over-training of the libraries to perform well with these specific functions. It is therefore important to note that the native structure recovery parameters tested—specifically the dihedral prediction, the hydrogen bonding recovery and the RMSD with the native structure—all improve alongside with the energies, indicating that the library captures well the physical aspects that determine side chain conformation in proteins. More tests will be necessary to understand how performance will be affected when the library is used with different energy functions. Our functions were selected specifically to favor efficient packing, hydrogen bonding and to prevent strains, which are factors that are present in a majority of modeling programs, and thus we are confident that the enhancement will translate well when the library is used with different functions. While nothing prevents the users from modifying their programs to adopt a set of functions similar to ours, should that be advantageous, we also encourage others to adopt the method to create specific Energy-Based libraries optimized *ad hoc* with the energy functions used in their own applications. A tutorial on how to create a library is made available on the EBL web site (http://seneslab.org/EBL) and all software and databases required for building a similar library are freely provided. The Energy-Based library and the dihedral-only version are distributed as supplementary information and in our website. The format of the library is described in Supporting Information Figure S4. All software used to create the EBL, the modules for reading the library, building conformers from internal coordinates, and for performing side chain optimization are implemented in C++ using the MSL package[38] (http://msl-libraries.org), a suite of molecular modeling tools freely available for download under an open source GPL v.3 license.

ACKNOWLEDGMENTS

The authors thank Ben Mueller, Dan Kulp and Gevorg Grigoryan for helpful discussion and critical reading of the manuscript, and Mohandas Pillai for assistance in program development. They also thank the Center for High-Throughput Computing of the University of Wisconsin-Madison for CPU time and Cathrin Weiss in particular for her kind assistance.

REFERENCES

1. Desmet J, Maeyer MD, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 1992;356:539–542.
2. Gordon DB, Mayo SL. Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure 1999;7:1089–1098.
3. Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12: 2001–2014.
4. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;Suppl 3:171–176.
5. 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.
6. Zou J, Saven JG. Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. J Mol Biol 2000;296:281–294.
7. Shetty RP, De Bakker PIW, DePristo MA, Blundell TL. Advantages of fine-grained side chain conformer libraries. Protein Eng 2003;16: 963–969.
8. Lassila JK, Privett HK, Allen BD, Mayo SL. Combinatorial methods for small-molecule placement in computational enzyme design. Proc Natl Acad Sci USA 2006;103:16710–16715.
9. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.
10. Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci 2004;13:735–751.
11. 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.
12. Georgiev I, Lilien RH, Donald BR. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. J Comput Chem 2008;29:1527–1542.
13. Chen C-Y, Georgiev I, Anderson AC, Donald BR. Computational structure-based redesign of enzyme activity. Proc Natl Acad Sci USA 2009;106:3764–3769.
14. Frey KM, Georgiev I, Donald BR, Anderson AC. Predicting resistance mutations using protein design algorithms. Proc Natl Acad Sci USA 2010;107:13707–13712.
15. Gainza P, Roberts KE, Donald BR. Protein design using continuous rotamers. PLoS Comput Biol 2012;8:e1002335.
16. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775–791.
17. Dunbrack RL. Rotamer libraries in the 21st century. Curr Opin Struct Biol 2002;12:431–440.
18. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40:389–408.
19. Dunbrack RL, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 1993;230: 543–574.
20. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
21. Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009;77: 778–795.
22. Shapovalov MV; Dunbrack RL, Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 2011;19:844–858.
23. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein–protein docking. Protein Sci 2005;14:1328–1339.

24. Grigoryan G, Ochoa A, Keating AE. Computing van der Waals energies in the context of the rotamer approximation. Proteins 2007;68:863–878.
25. Allen BD, Mayo SL. Dramatic performance enhancements for the FASTER optimization algorithm. J Comput Chem 2006;27:1071–1075.
26. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331:281–299.
27. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Meth Enzymol 2004;383:66–93.
28. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. Proc Natl Acad Sci USA 2002;99:14116–14121.
29. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.
30. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A. De novo design and structural characterization of proteins and metalloproteins. Annu Rev Biochem 1999;68:779–819.
31. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ Jr, Stoddard BL, Baker D. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 2006;441:656–659.
32. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. Nature 2008;453:190–195.
33. Petrella RJ, Karplus M. The energetics of off-rotamer protein side-chain conformations. J Mol Biol 2001;312:1161–1175.
34. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.
35. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 1983;4:187–217.
36. MacKerell DB, Dunbrack B, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
37. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J 1994;66:1335–1340.
38. Kulp DW, Subramaniam S, Donald JE, Hannigan BT, Mueller BK, Grigoryan G, Senes A. Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). J Comput Chem 2012; doi:10.1002/jcc.22968.
39. Pokala N, Handel TM. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. Protein Sci 2004;13:925–936.
40. Mendes J, Baptista AM, Carrondo MA, Soares CM. Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants. J Comput Aided Mol Des 2001;15:721–740.
41. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 2002;320:597–608.
42. Lu M, Dousis AD, Ma J. OPUS-Rota: a fast and accurate method for side-chain modeling. Protein Sci 2008;17:1576–1585.
43. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
44. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and proteinprotein complexes. J Mol Biol 2003;326:1239–1259.
45. Subramaniam S, Natarajan S, Senes A. A machine learning based approach to improve protein sidechain optimization. In: ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM BCB), 2012. pp 478–480.
46. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82–87.