PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Backbone dependency further improves side chain prediction efficiency in the Energy-based Conformer Library (bEBL)

Sabareesh Subramaniam and Alessandro Senes*

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

ABSTRACT

Side chain optimization is an integral component of many protein modeling applications. In these applications, the conformational freedom of the side chains is often explored using libraries of discrete, frequently occurring conformations. Because side chain optimization can pose a computationally intensive combinatorial problem, the nature of these conformer libraries is important for ensuring efficiency and accuracy in side chain prediction. We have previously developed an innovative method to create a conformer library with enhanced performance. The Energy-based Library (EBL) was obtained by analyzing the energetic interactions between conformers and a large number of natural protein environments from crystal structures. This process guided the selection of conformers with the highest propensity to fit into spaces that should accommodate a side chain. Because the method requires a large crystallographic data-set, the EBL was created in a backbone-independent fashion. However, it is well established that side chain conformation is strongly dependent on the local backbone geometry, and that backbone-dependent libraries are more efficient in side chain optimization. Here we present the backbone-dependent EBL (bEBL), whose conformers are independently sorted for each populated region of Ramachandran space. The resulting library closely mirrors the local backbone-dependent distribution of side chain conformation. Compared to the EBL, we demonstrate that the bEBL uses fewer conformers to produce similar side chain prediction outcomes, thus further improving performance with respect to the already efficient backbone-independent version of the library.

Proteins 2014; 82:3177–3187.

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

INTRODUCTION

Side chain optimization is the process of predicting the 3-dimensional conformation of the side chains of a protein given the structure of its backbone. Side chain optimization is an important component of several protein modeling applications such as homology modeling,[1–5] structure prediction,[5–7] protein design,[8–11] point mutation analysis,[12,13] protein and ligand docking,[12,14] and structure refinement.[15,16] Mechanistically, side chain optimization is a search for the lowest energy state among all the combinations of side chain conformations for a given protein backbone. The energy of each conformation is scored with a variety of physics-based[17,18] or knowledge-based potential functions.[19] Since the conformational space of side chains is extensive but sparsely populated, this space is generally sampled by adopting a library of discrete conformations.

The library-based approach provides two advantages. First, it allows the search algorithms to focus on those regions of side chain conformational space that occur frequently in proteins, while discarding or deprioritizing those that are rarely encountered. In addition, the library converts a continuous and multi-dimensional search problem into a discretized combinatorial problem, which reduces complexity and facilitates the application of a number of fast, deterministic[20–22] and probabilistic algorithms[23–25] for identifying the global minimum energy conformation of the side chains.

The discrete libraries used for side chain optimization are usually derived in one of two ways: *rotamer* or *conformer* libraries. The rotamer libraries[26–28] are compiled using a statistical analysis of the distribution of the side chain dihedral angles (χ angles), which are the major

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

Grant sponsor: National Institutes of Health; Grant number: R01GM099752.

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

Received 2 July 2014; Revised 21 August 2014; Accepted 3 September 2014
Published online 11 September 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24685

determinants of side chain conformation. This approach is based on clustering side chain conformations observed in high-resolution structures present in the Protein Data Bank (PDB). The clusters (rotamers) are reported as combinations of χ angles, by indicating the center of each cluster, a measure of their deviation, and their overall frequency. The rotamer libraries generally do not report bond angles and bond lengths, which are only minor determinants of conformation. Therefore, when rotamer libraries are applied to side chain prediction, these variables are generally held fixed at some optimal value.

A second approach is employing conformer libraries,[25,29,30] which are representative conformations from native structures found in the PDB. These libraries are created by extracting a large number of conformations from the structural database, which are then reduced to a manageable subset by applying a similarity filter, such as angular similarity[25] or root mean square deviation (RMSD).[29,30] Because the selected conformers are actual side chains found in proteins, the conformer libraries retain the natural variation of bond lengths and angles observed in proteins, which can be beneficial.[25]

An important issue with both rotamer and conformer libraries is controlling the specific granularity of the sampling. Increasing the size of the library can improve the outcome of side chain optimization.[25,29,31–33] However, side chain optimization is a combinatorial search and higher sampling often comes at significant computational cost. Particularly, if the number of side chains involved is large, or if multiple side chain optimizations are required by an application, side chain optimization can become a bottleneck. For example, we use side chain optimization to develop methods for the structural prediction of complexes of transmembrane helices.[7,13] These methods involve extensive exploration of backbone conformational space, and the cycle of side chain optimization required after each backbone move represents a major cost for the procedures. Therefore, it is important to find the level of sampling that provides the best compromise between two conflicting requirements, (i) reduction of the size of the library, for computational efficiency, and (ii) increase of its size, to achieve the best possible accuracy.[25,32]

To address this challenge, we previously introduced the Energy-based Library (EBL),[30] an efficient conformer library created with an energetic criterion, whose granularity of sampling is easily customizable. While the previously available rotamer and conformer libraries may be produced in different sizes,[25,29,32] the EBL is provided as a sorted list that can be truncated precisely at any desired length, enabling much finer control of the size of the library, down to the level of the single conformer.

In addition, the most important advantage of the EBL is its excellent performance in side chain optimization.

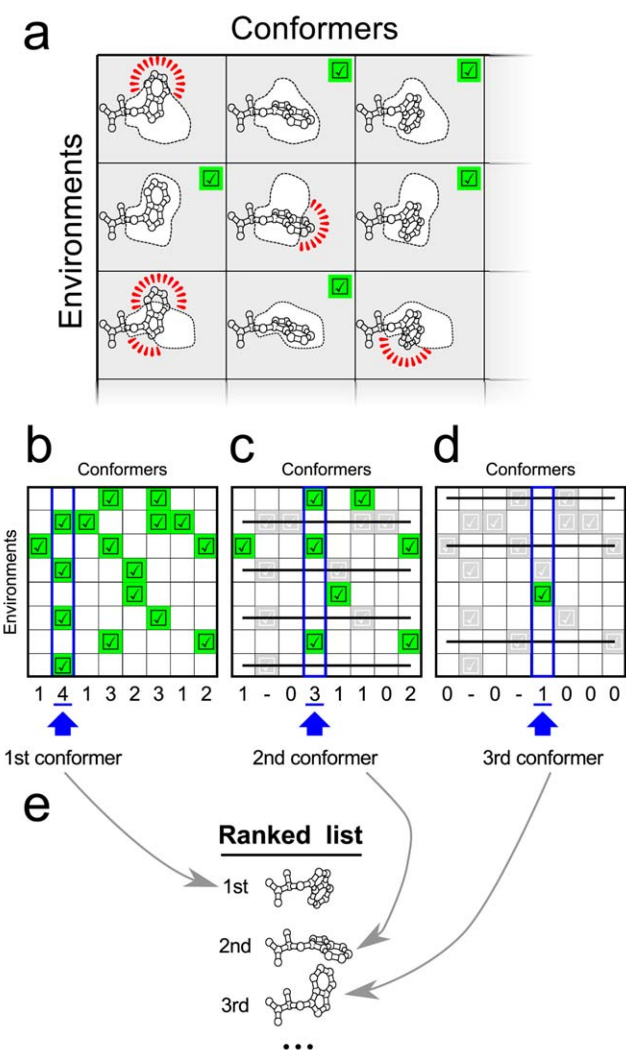

Figure 1

Procedure for the creation of an Energy-based Conformer Library. (**a**) Each of *N* conformers from a fine-grained library is built into each one of *M* environments that contain the same amino acid (Trp in the figure) in protein crystal structures. The interaction energies of each conformer-environment pair 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, where *true* means that the conformer satisfies the environment. The number of environments satisfied by each conformer is determined (number under the table). The conformer that satisfies 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 largest number of the remaining environments. (**d**) The procedure is repeated until all conformers are ranked. Not represented: to avoid that all environments are rapidly consumed by the procedure, at every cycle the energy threshold that determines if an environment is satisfied is increased and any excluded environments that would be no longer satisfied at the new more stringent threshold are brought back into consideration. (**e**) The resulting library is compiled as a ranked list, in which every additional element complements the previous. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The EBL was constructed using the same metric used in side chain optimization—that is, energy. This new strategy resulted in a conformer library that has improved computational efficiency. We previously showed that, compared to other commonly used state-of-the-art libraries, the EBL improves modeling accuracy (lower energies and better dihedral prediction) for a similar number of conformers, and that the EBL requires fewer conformers for achieving the same level of accuracy as the other libraries.[30]

Although the EBL is very efficient and has important advantages, a potential limitation is that it was developed as a backbone-independent library, which suggests it could be further improved. Indeed, it is well established that side chain conformation is strongly dependent on the local backbone geometry,[34] and has been shown that a backbone-dependent library performs significantly better than an equivalent backbone-independent library.[35] This is because small changes in φ/ψ angles can produce significant variations in the rotameric distribution even within regions with the same secondary structure classification (helix, sheet).[28,36,37] In addition, the local variations of backbone conformation also influence the average value of χ angles (to relax strained interactions between the side chain and the backbone),[35] and these variations should ideally be captured by a side chain conformer library.

An important hurdle for the creation of a backbone-dependent version of the EBL was obtaining sufficient structural data. The EBL was derived by remodeling individual side chains within the fixed context of protein crystal structures and evaluating the energetic interactions between large sets of natural protein environments and conformers (Fig. 1). The resulting data-set of conformer/environment interactions was used to sort the conformers by their propensity to fit (energetically) into natural proteins, thereby producing a very efficient library. However, this data-driven approach posed difficulties in obtaining sufficient environments for all backbone-dependent subdivisions once the environments were binned according to their φ/ψ coordinates.

Here, we test a new version of the EBL, which we call the backbone-dependent Energy-based Conformer Library (bEBL). The bEBL was created in a backbone-dependent fashion for the most populated regions of φ/ψ space, that is, those for which the sampling issue does not occur. Although these regions represent only a small minority of Ramachandran space, they contain a majority of side chain conformation density, given the highly uneven distribution of the population of the backbone space. In a side-by-side comparison, we demonstrate that the bEBL is more efficient than the already efficient EBL, achieving similar or better performance with a much smaller number of conformers. The comparison is performed using three important parameters: (i) energy of the predicted protein structures; (ii) correct prediction of crystallographic side chain conformation; and (iii) reduction of computational time. Support for the bEBL is implemented in the Molecular Software Libraries (MSL) v. 1.2, a C++ open source library for molecular modeling, analysis, and design.[38]

MATERIALS AND METHODS

Structure database preparation

A collection of 2159 high-resolution X-ray structures was obtained from the PDB and curated, as previously described,[30] using the following conditions: resolution <2.0Å; deposition date: later than 01/01/1998; method: X-ray diffraction; molecule type: protein (no DNA, no RNA); no ligands. Hydrogen atoms were added with the program Reduce,[39] which also performed any necessary rotation of the hydroxyl groups, flipping the side chain of Asn, Gln, and His and determined the protonation state of His to optimize hydrogen bonding (*-BUILD -ROTEX* options). The three protonation states of His were analyzed separately. 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 structures were then minimized 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} using CHARMM.[17] A set of 480 proteins was reserved for testing.

Programs

All calculations (modeling, energy evaluations, conformational analysis, SASA measurements, etc.) were performed with programs implemented in MSL, a C++ object oriented software library for molecular modeling and analysis.[38]

Input (unsorted) conformer library

The input fine-grained, unsorted conformer library was the same used for the creation of the original EBL, prepared as previously described.[30] Side chains with a B-factor ≥ 40 and those with missing atoms in the original structure were not considered. Conformers were RMSD filtered by selecting them at random from a large pool and adding them to the conformer list if they had an RMSD >0.05Å from all other previously collected conformers.

Selection of the backbone-dependent protein environments

The positions in high-resolution crystal structures to be used as environments for the creation of the conformer library were selected as previously described,[30] except that the environments were subdivided according to their backbone conformation. To ensure adequate sampling, for each amino acid, only the 10° × 10° subdivisions of φ/ψ space that contained a minimum of 100 side chains were considered individually for the creation of a sorted library. All the remaining sparsely populated 10° × 10° subdivision were then combined and treated as a single partition.

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

For each amino acid type, the crystallographic side chain of each of the M environments was remodeled as each one of the N conformers and the conformer/environment interaction energy was calculated, producing a matrix of $N \times M$ energies. Energies were calculated as described previously[30] using the CHARMM 22 force field[17] (bond, angle, Urey-Bradley, dihedral, improper, and van der Waals terms) supplemented by an explicit hydrogen bond term from SCWRL 4.[40] The non-bonded interactions were calculated with a distance-dependent cutoff using a switching function (cut-on 9Å, cut-off 10Å).

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* by the conformer. Because the best energy achievable in each environment varied substantially, an environment dependent threshold was adopted, as previously reported.[30]

Creation of the sorted bEBL

The fine grained conformer library was sorted by the propensity of its elements to fit in the largest number of natural environments, as described previously,[30] but the procedure was applied to each backbone partition independently. The sorting procedure is schematically explained in Figure 1. 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. To avoid that the environments are rapidly consumed by the procedure, after each cycle of selection the threshold that determined if a conformer satisfies an environment was lowered. This brought back into consideration any excluded environment that was no longer satisfied by any of the previously selected conformers at the new, more stringent threshold. 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.

Sampling levels

A series of sampling levels was created as a means to balance conformational sampling across amino acid types. These were constructed from conformer/environment interactions, as reported previously.[30] For each amino acid, the number of conformers that are necessary to satisfy a certain fraction of environments constitutes a sampling level. For example, eight conformers may be required to satisfy 85% of the environments of a certain amino acid: therefore the 85% sampling level (SL85) for the amino acid would be eight conformers. Fourteen sampling levels were created, from very sparse sampling (60% level) to very high sampling (99% level). The sampling levels were calculated independently for each φ/ψ partition.

Side chain prediction tests

Side chain optimization was performed on a test set of 480 proteins set aside for this purpose. All side chains except Ala, Pro, and Gly were removed and simultaneously predicted using both EBL and bEBL. His residues were modeled according to the predicted protonation state. Side chain optimization was performed with a variant of the *repackSideChains* program updated to support the bEBL library.[30]

Testing of the bEBL was performed against the backbone-independent EBL. A comparison with other libraries was reported previously,[30] where we showed that the EBL performs favorably compared to three state-of-the-art rotamer and conformer libraries: a 5× expansion of the 2010 version of the Backbone Dependent rotamer library;[28,40] the "medium" size library (0.5Å RMSD) from Shetty *et al.*;[29] and a small conformer library from Xiang and Honig[25] created from a database of 297 proteins, 100% coverage and 40° tolerance.

Dihedral recovery

Assessment of the conformation prediction of crystallographic side chains (side chain conformation recovery) was performed with the *getChiRecovery* program in MSL[38] by matching the χ^1 and χ^2 of the predicted and crystallographic structure with a tolerance of 40°. The analysis was performed on the subset of buried side chains that had Solvent Accessible Surface Area (SASA) below 25% of the maximum possible SASA for the side chain reconstructed into Gly-X-Gly backbone (with X being the amino acid type under consideration).

Library format

The format of the library is illustrated in Supporting Information Figure S1. The format is identical to that of the original EBL, with the addition of the "BBDEP" tag, which marks the beginning of the conformer list relative to each backbone bin. The library is available for download at http://seneslab.org/EBL.

RESULTS AND DISCUSSION

Library creation procedure

The procedure followed for the creation and testing of the bEBL is nearly identical to the procedure described previously.[30] The base was the same fine conformer library that was sorted for the creation of the EBL. In the present case, however, the protein environments were subdivided in bins based on their backbone dihedral angle (φ/ψ) coordinates, and the library was sorted independently for each bin.

The procedure is schematically summarized in Figure 1. The conformers were reconstructed within fixed protein environments (classified by their φ/ψ bin) and the interaction energies between the conformers and the protein environments were measured [Fig. 1(a)]. If a conformer/environment interaction energy was below a certain threshold, the conformer was considered to satisfy the environment [green tick marks in Fig. 1(a)]. This resulted in a 2-dimensional boolean table [Fig. 1(b)], which was used to sort the conformers so that those that were most efficient to satisfy protein environments would be ranked higher. First, the conformer that satisfied the most environments was chosen as the top element of the sorted library [Fig. 1(b)]. All the environments that were satisfied by this conformer were then removed and no longer considered in the selection of the next conformers [Fig. 1(c)]. This exclusion ensures that each round selects complementary conformers that cover different regions of side chain space [Fig. 1(c,d)]. However, to avoid having this procedure rapidly consume all environments, the energy threshold that determines if an environment is satisfied was increased after every cycle. This allows for the excluded environments to be considered again if they are no longer satisfied by the previously selected conformers at the new, more stringent threshold. The procedure produced a series of conformer libraries sorted independently for each bin of backbone conformational space.

Partitioning the protein backbone space

The major challenge for the creation of a backbone-dependent energy-based library was the identification of an effective strategy for subdividing the Ramachandran (φ/ψ) space. The subdivision should result in sufficient data for each partition while at the same time capturing the natural variation of rotamer propensity across the backbone space. We tested a scheme based on dividing the φ/ψ space into 10° × 10° bins, which was previously used by Dunbrack and colleagues.[28,36]

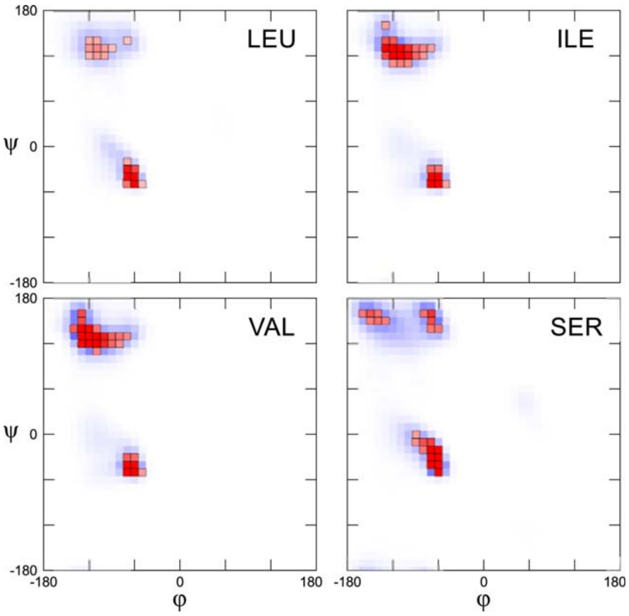

Figure 2

Partitioning scheme for four amino acids. The bEBL was computed for regions of Ramachandran space that contained a sufficient number of training environments. The 10° × 10° individual partitions selected for bEBL creation are in red, the remaining large partition is highlighted in blue. The intensity of the colors is proportional to the probability of backbone conformation. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

To avoid over-fitting the training data during the sorting procedure, we estimated that each 10° × 10° partition should contain a minimum of 100 environments. Therefore all 10° × 10° partitions that contained at least 100 examples for each amino acid were collected for the application of the EBL algorithm. Because the Ramachandran distribution is highly uneven, with small areas of high density and large sparsely populated regions, only a minority of all the possible 1296 (36 × 36) bins contained sufficient environments, with a number ranging from 27 for Val to only 4 for Cys (Supporting Information Table SI). However, because the selected partitions are highly populated, the scheme resulted in a backbone-dependent library that, on average, covers 50–60% of the effective side chain conformational frequency for most amino acids, and as high as 70%, as in the case of Ile. The remaining space that could not be assigned to a 10° × 10° bin was treated as a single large partition. The partitioning schemes of Leu, Ile, Var, and Ser are illustrated in Figure 2.

The bEBL library matches the natural side chain bias of each partition

It is well established that rotamer frequency is strongly biased by the conformation of the backbone.[37] Figure 3 graphically illustrates how side chain preference varies

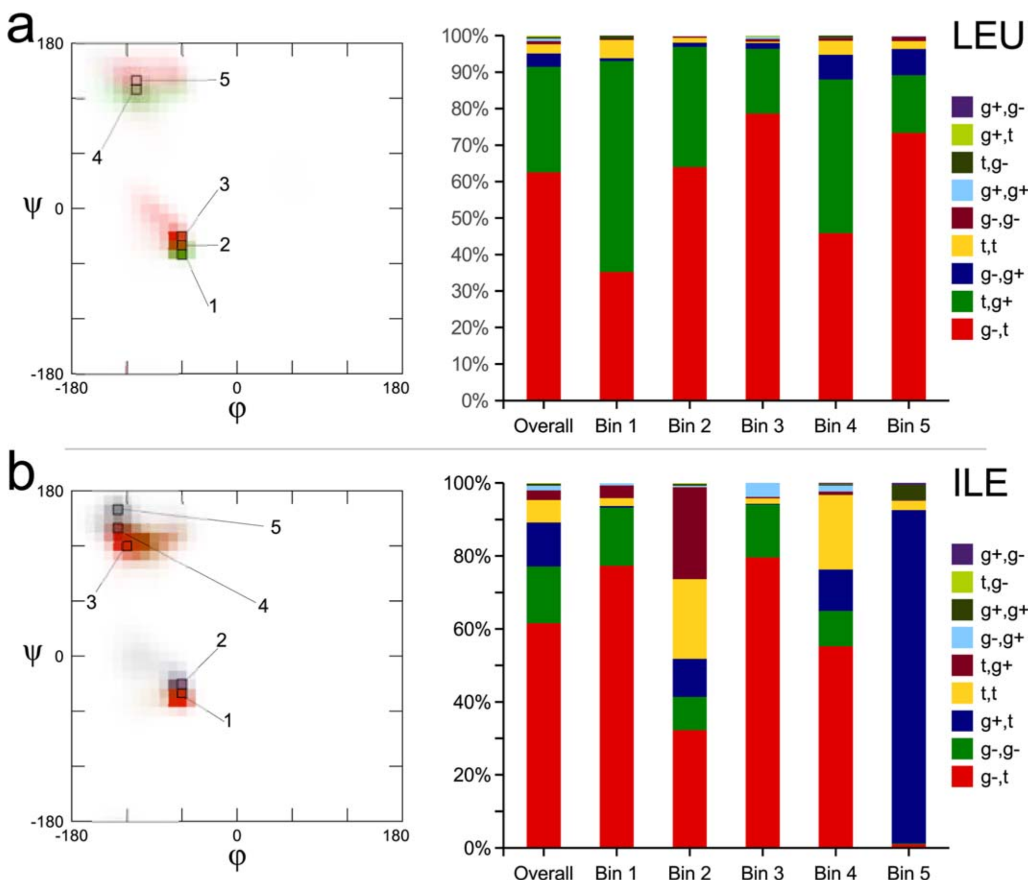

Figure 3

Side chain conformation is strongly dependent on backbone conformation. The stacked histograms show the distribution of side chain conformation in the entire PDB (all) and in five 10° × 10° φ/ψ bins, for Leu (**a**) and Ile (**b**). The colors of the bins in the Ramachandran graphs are a sum of colors in the histogram (weighted by rotamer probability in the bin), and their opacity is proportional to the probability of backbone conformation. The probability of each rotamer is profoundly affected by the backbone, with important variation observed even between adjacent 10° × 10° bins.

dramatically even between adjacent 10° × 10° bins for Leu and Ile. For example, the overall preferred rotamer of Leu is [*g*-,*t*] (62%, represented in red in the figure), but its local probability changes from 35% to nearly 80% across three adjacent bins in the α-region (marked as 1, 2, and 3 in the figure). The change is roughly compensated by the [*t*,*g*+] rotamer (green), which becomes predominant in bins 2 and 3. Similar changes are also noticeable for the β-region (bins 4 and 5).

A similar phenomenon can also be observed for Ile in Figure 3(b). The fraction of side chains in [*g*-,*t*] conformation drops from 80% to 30% between the two adjacent bins 1 and 2 in the α-region. Even more dramatic changes are observable in the β-region (bins 3, 4, and 5): in particular, bin 5 is composed by over 95% of the rare [*g*+,*t*] conformation (blue), which represents only 10% of the total frequency of conformers in Ile ("Overall" bin).

We found previously that the composition of the top conformers of the original backbone-independent EBL closely matched the overall frequency observed in proteins, a feature that is likely important for its efficiency.[30] This raised the question of whether the bEBL would also mirror the dramatic changes in composition of each individual 10° × 10° bin. As shown in Figure 4, this expectation was confirmed: the side-by-side comparison between natural side chain conformational preferences and the composition of the top conformers of each bEBL bin shows a remarkable correlation.

The bEBL requires fewer conformers for comparable "sampling levels"

To evaluate the efficiency of the bEBL relative to the original backbone-independent version, we first compared the number of conformers required by their equivalent "sampling levels." The sampling level is a concept that was introduced with the original EBL to balance sampling for the various amino acids at all different

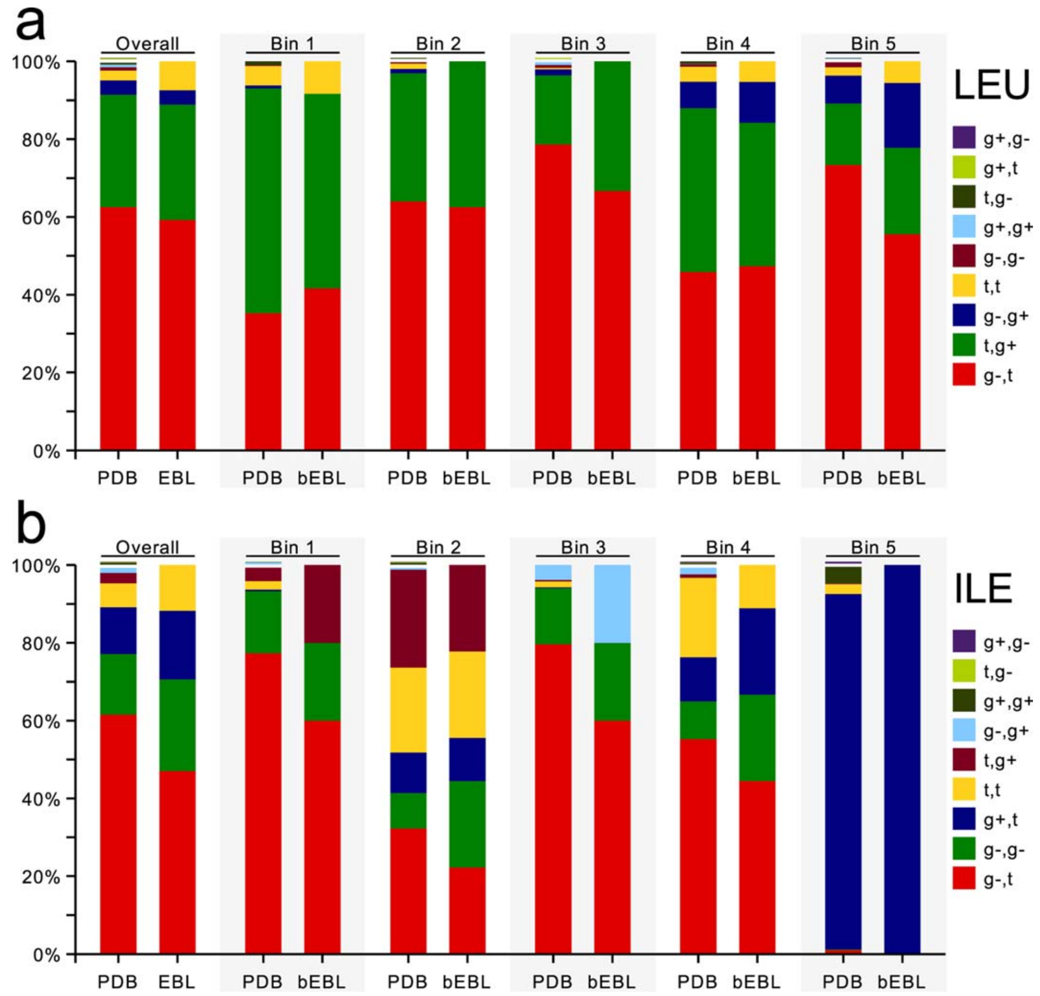

Figure 4

The bEBL captures the local rotamer distribution. Comparison of the probability of side chain conformation in crystal structures ("PDB") and in top conformers of the Energy-based Libraries for (**a**) Leu and (**b**) Ile. "Overall": backbone-independent probability, compared to the original EBL. Bin 1–5: side chain frequency in the 10° × 10° φ/ψ bins identified in Figure 3, compared to the corresponding bin for the bEBL. The composition of the EBL and bEBL bins corresponds to the top conformers of the SL85 sampling level. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

levels of granularity.[30] Each amino acid type has a unique geometry, number of atoms and number of rotatable bonds, and therefore, a different sampling requirement. For example, Val has a small side chain and requires fewer conformers compared to the large and flexible Arg. Therefore, to optimize sampling, it is necessary to employ a different number of conformations for each amino acid. The "sampling levels" help to rationalize and optimize this difficult decision by creating a series of balanced libraries of increasing size to choose from, depending on whether a calculation needs to prioritize speed or maximize accuracy, and any compromise in between. For example, in the original EBL, Arg, Leu, and Val side chains are assigned 52, 9, and 3 conformers respectively at the 70% sampling level (SL70). The number of conformers for the same amino acids increased to 102, 17, and 5 at the 80% sampling level (SL80), and to 222, 39, and 10 at SL90.

The sampling levels of the bEBL were created with the same method, except that the number of conformers for every sampling level is calculated independently for each individual bin. The number of conformers in the sampling levels of the bEBL can be taken as an initial indication of its performance: if the bEBL is more efficient than the EBL, it is expected that it will require a smaller number of conformers for the equivalent sampling levels. As shown in Figure 5, this expectation was met. For example, SL70 for Leu consists of nine conformers in the EBL, but only four conformers are required by the bEBL (weighted average across all bins) [Fig. 5(a)]. Similarly,

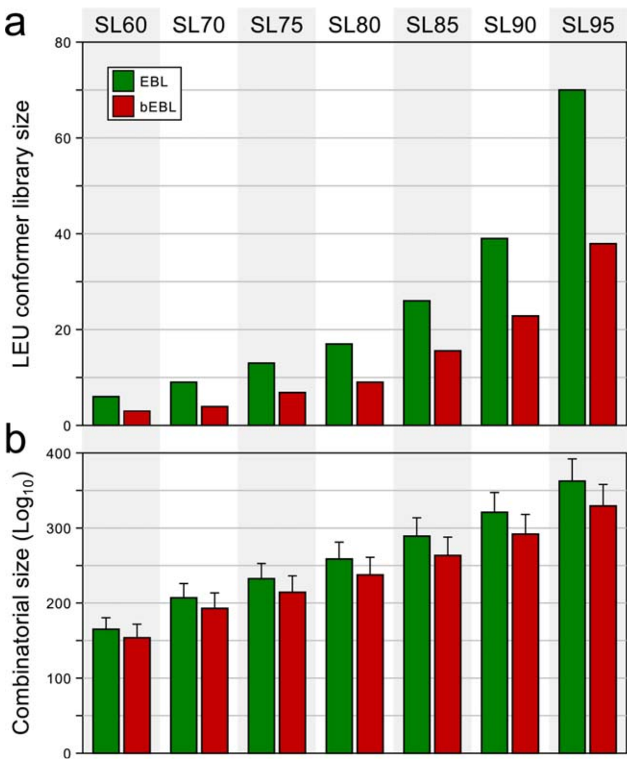

Figure 5

The bEBL has smaller size for comparable sampling levels. (**a**) Comparison of number of conformers in each sampling level for EBL and bEBL for Leu. The bEBL requires approximately half the conformers as the EBL. (**b**) Logarithm of the combinatorial size of side chain conformational space for 480 full-protein predictions, that is, the product of the number of conformers n_i at each positions i of P total positions, normalized for a 100 amino acid protein: $\log(\Pi n_i^{100/P})$. On average, the bEBL reduces the combinatorial size by 11 (SL60) to 33 (SL95) orders of magnitude at the various sampling levels. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the sampling requirements of the SL80 and SL90 levels are also reduced in the bEBL, from 17 to 9 and from 39 to 23, respectively. A full comparison of the sampling levels in the EBL and bEBL for all amino acids is reported in Supporting Information Table SII.

The reduction in the number of conformers necessary at each sampling level should translate into a substantial decrease of the search space in side chain optimization while preserving the quality of the outcome. Figure 5(b) shows the combinatorial size of the overall search space—the product of the number of conformers at all positions—at different sampling levels. The data show that the bEBL reduces the search space by 10 to over 30 orders of magnitude at the different sampling levels. Although the speed of side chain optimization is generally not linearly proportional to the full combinatorial size of the search problem, a reduction of the search space is likely to result in significant performance enhancements, as shown later.

bEBL leads to models with similar or lower energy using fewer conformers

Side chain prediction is a search for the lowest energy state of a structure; therefore energy is an important parameter in estimating efficiency. Side chain prediction was performed on the 480 proteins from the curated test dataset, after removing the crystallographic side chains. Figure 6(a) shows the comparison of the final energy of each protein for side chain predictions performed at the SL85 level, using either the bEBL (x-axis) or the EBL (y-axis). Despite the fact that the bEBL is a smaller library, the energies are similar, with most points lying near the diagonal of the graph. In fact, the points in the upper-left side of the graph (better bEBL energy)

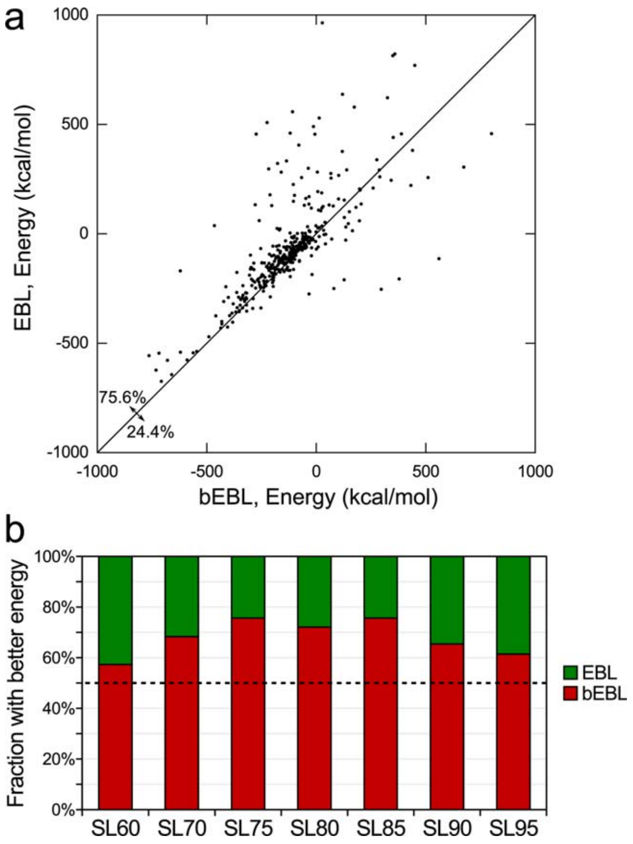

Figure 6

bEBL achieves similar energies with fewer conformers. Side chain optimization was performed on a set of 480 test proteins using different sampling levels. (**a**) Plot of the energies, after subtracting the energy of the crystal structure, obtained from side chain optimization at the SL85 level using bEBL and EBL (along the x- and y-axis, respectively). The fraction of proteins above the diagonal (better bEBL energy) and below the diagonal (better EBL energy) is indicated in the lower left-hand corner. (**b**) Fractions of predicted protein models that had a lower energy after side chain optimization with bEBL (red) and EBL (green) at six sampling levels. The bEBL performs consistently better than EBL at all sampling levels, in some case by over a 70%–30% margin. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

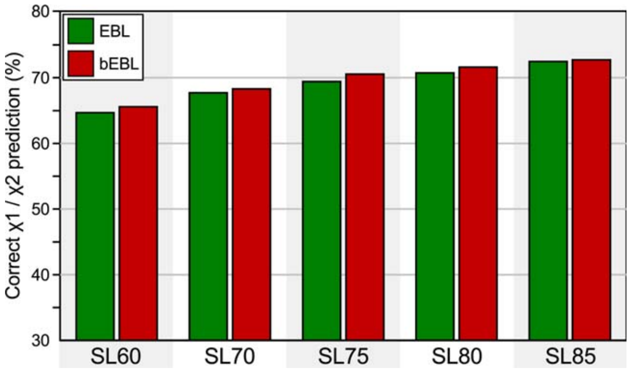

Figure 7

bEBL achieves comparable dihedral prediction with fewer conformers. Percentage of buried side chains correctly predicted in the test set of 480 full-protein repacks at the respective sampling levels of the EBL (green) and bEBL (red). A prediction is considered correct if χ^1 and χ^2 are close to the values observed in the original crystal structure, with a tolerance of ±40°. In spite of the smaller size, the bEBL achieves similar or slightly better prediction than the original EBL. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

outnumber those in the lower-right side (better EBL energy) by a 4:1 ratio. Figure 6(b) shows this same ratio for other sampling levels: in all cases the bEBL outperformed the EBL in a majority of the predictions, indicating that the smaller library is as efficient, if not more efficient, than the larger, backbone-independent version.

bEBL achieves similar side chain conformation prediction with fewer conformers

The accuracy of side chain modeling can be measured in terms of the percentage of predicted side chains that are in agreement with the natural conformation observed in a crystal structure. Figure 7 shows a comparison of average side chain prediction accuracy in the 480 proteins, using the EBL and bEBL at different sampling levels (prediction of χ^1 and χ^2, with a tolerance of 40° from the native structure).[30] As expected, the percentage of correct predictions increases with the sampling level in both libraries, which is consistent with the expectation that the use of more conformers would improve prediction accuracy.[25,29,32,33] At all sampling levels, the accuracy is similar and slightly better with the bEBL in spite of its smaller size, demonstrating that the backbone-dependent version is at least as effective, if not more effective, in predicting side chain conformation.

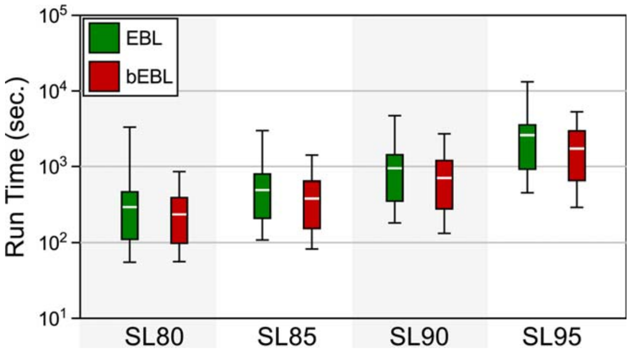

Figure 8

The bEBL reduces run time in side chain optimization. Distribution of total run time of each of the 480 full-protein repacks at various sampling levels, shown in logarithmic scale. White bar: average time. Box: 68% interval. Black bars: 95% interval. The bEBL reduces the average run time by 20–30% at the various levels. The backbone-dependent library is most effective at reducing run time for the slowest calculation (upper black bar), which can run faster by a factor of 2–4, depending on the sampling level. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

bEBL reduces computational time

As discussed above, the bEBL matches or exceeds the performance of the original EBL, both in terms of energy and side chain conformation prediction, while using a smaller number of conformers. The smaller size of the library should in principle translate to a substantial reduction of the computational time required for side chain optimization. This expectation is confirmed by the results presented in Figure 8, which shows the time taken for side chain optimization using different sampling levels of the EBL and bEBL. The average calculation time is shorter with the bEBL compared to the EBL. The difference becomes more important for higher sampling levels: for example, the bEBL is on average 20% faster at SL80 and 34% at SL95.

When the black bars that correspond to the upper-end of the 95% interval are compared, the improvements become more noticeable, with speed increases of a factor of 2–4, depending on the sampling level. These slow runs correspond to the side chain optimization problems that have the highest complexity either because of the size of their proteins or because their solution has the highest intrinsic complexity. For this class of problems, clearly the utilization of a leaner, more efficient library has the greatest pay-off.

CONCLUSION

We have presented a backbone-dependent bEBL that further improves protein side chain optimization compared to its backbone-independent version. The original EBL already outperformed other state-of-the-art libraries, as reported,[30] therefore the bEBL is an extremely effective library. Our experiments further demonstrate that accounting for the conformation of the backbone is a very effective strategy for improving the outcome of side chain optimization.[35–37]

We have shown that application of the EBL algorithm to the most frequent 10° × 10° regions of backbone conformational space (which cumulatively account for 50–70% of all side chain conformations) is sufficient to significantly enhance the library. For these regions of Ramachandran space, the composition of the bEBL closely mirrors the conformational bias of rotamer distribution to a level that would be difficult to obtain with other traditional conformer or rotamer libraries. The bEBL performs similarly, or even slightly better, than the backbone-independent version both in terms of energy and modeling accuracy, while using a substantially smaller number of conformers. Utilization of a smaller library can lead to reduction of computational overhead and execution time while preserving quality. Alternatively, the more efficient bEBL can be used to improve accuracy in side prediction without increase in execution time. Therefore, the bEBL is an effective tool that can improve performance in a variety of protein modeling applications.

ACKNOWLEDGMENTS

The authors thank Ben Mueller, Sam Condon, Loren LaPointe and Sam Craven for discussion and critical reading of the manuscript.

REFERENCES

1. Fiser A, Feig M, Brooks CL 3rd, Sali A. Evolution and physics in comparative protein structure modeling. Acc Chem Res 2002;35:413–421.
2. Nayeem A, Sitkoff D, Krystek S. A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. Prot Sci 2006;15:808–824.
3. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. Prot Sci 2005;14:1315–1327.
4. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 2008;4:1–13.
5. Wang Q, Canutescu AA, Dunbrack RL. SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. Nat Protoc 2008;3:1832–1847.
6. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32(suppl 2):W526–W531.
7. Mueller BK, Subramaniam S, Senes A. A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds. Proc Natl Acad Sci USA 2014;111:E888–E895.
8. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. Proc Natl Acad Sci 1997;94:10172–10177.
9. Street AG, Mayo SL. Computational protein design. Structure 1999;7:R105–R109.
10. Senes A. Computational design of membrane proteins. Curr Opin Struct Biol 2011;21:460–466.
11. Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG. Theoretical and computational protein design. Annu Rev Phys Chem 2011;62:129–149.
12. Schaffer L, Verkhivker GM. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. Proteins Struct Funct Bioinforma 1998;33:295–310.
13. LaPointe LM, Taylor KC, Subramaniam S, Khadria A, Rayment I, Senes A. Structural organization of FtsB, a transmembrane protein of the bacterial divisome. Biochemistry (Mosc) 2013;52:2574–2585.
14. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein–protein docking. Protein Sci 2005;14:1328–1339.
15. Qian B. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259–264.
16. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. *MolProbity*: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2009;66:12–21.
17. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
18. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
19. Russ WP, Ranganathan R. Knowledge-based potential functions in protein design. Curr Opin Struct Biol 2002;12:447–452.
20. 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.
21. Gordon DB, Mayo SL. Branch-and-terminate: a combinatorial optimization algorithm for protein design. Struct Lond Engl 1993 1999;7:1089–1098.
22. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci Publ Protein Soc 2003;12:2001–2014.
23. 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.
24. 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.
25. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.
26. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci Publ Protein Soc 1997;6:1661–1681.
27. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40:389–408.
28. Shapovalov MV, Dunbrack RL Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Struct Lond Engl 1993 2011;19:844–858.
29. Shetty RP, De Bakker PIW, DePristo MA, Blundell TL. Advantages of fine-grained side chain conformer libraries. Protein Eng 2003;16:963–969.
30. Subramaniam S, Senes A. An energy-based conformer library for side chain optimization: improved prediction and adjustable sampling. Proteins 2012;80:2218–2234.
31. 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.
32. 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.
33. 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.
34. Chakrabarti P, Pal D. The interrelationships of side-chain and main-chain conformations in proteins. Prog Biophys Mol Biol 2001;76:1–102.
35. Dunbrack RL Jr. Rotamer libraries in the 21st century. Curr Opin Struct Biol 2002;12:431–440.
36. Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 1993;230:543–574.

37. Dunbrack RL, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Mol Biol 1994;1:334–340.
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;33:1645–1661.
39. 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.
40. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009;77:778–795.