An energy trained conformer library for side chain optimization

Sabareesh Subramaniam[1,2] and Alessandro Senes[1]

[1]Dept of Biochemistry and [2]Dept of Computer Sciences, University of Wisconsin-Madison, Madison WI 53706

Abstract

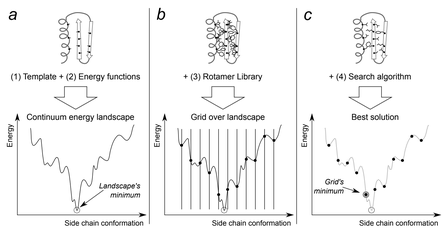
Side chain optimization is an essential component of applications such as protein structure prediction, docking and design. Currently, rotamer or conformer libraries are used to define the search space for the side-chain optimization problem. The *rotamer* libraries are based on the statistical distribution of side chain dihedral (χ) angles in the structural database. The *conformer* libraries are a collection of naturally occuring side-chain conformations obtained from high-resolution crystal structures. The conformer libraries have the advantage in that they retain the natural ranges of bond distances and angles.

Both rotamer- and conformer-based libraries for side chain optimization are built with the goal of a geometric representation of conformational space (that is, the members of the library are selected to represent those that are similar in shape). Here we introduce the concept of *energetic representation*, taking into account also the nature of the environment in which the conformations are built.

This criterion allocates sampling strategically proportional to energy variation, which is the same metric that select the "winner" of a side chain optimization procedure. This approach significantly improves both efficiency and accuracy in side chain optimization.

A search for side-chain packing is, at best, only as good as the library used.

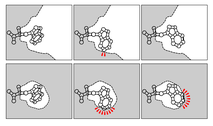
A side chain optimization problem is defined by a template (backbone) and a set of energy functions, which define a continuum energy lanscape (*a*) covering all possible theoretical side chain conformations.



The introduction of a rotamer library forms a "grid" that samples the space at intervals (*b*). Only these grid points are searched and the rest of the space remains unexplored.

How far or how near the grid's minimum will lie from the best point on the landscape – the actual target of the optmization procedure – depends on the library.

The amount of sampling required by a side chain is strongly dependent on its surrounding environment.



A surface exposed side chain obviously requires much less sampling to find a conformation that will fit favorably compared to a position buried in the core. Other factors of the local environment (φ/ψ, seconday structure, crowdedness, nature of the neighboring side chians) also matter.

In this project we consider the nature of the environment in allocating sampling strategically for side chain optimization.

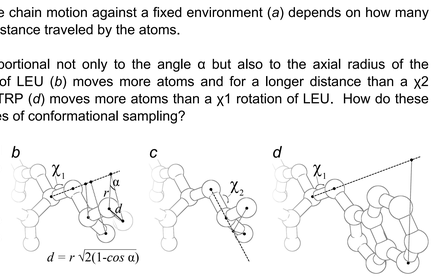
Sampling requirements also vary for each degree of freedom and for the residue type.

The energetic impact of a side chain motion against a fixed environment (*a*) depends on how many atoms are translated and to distance traveled by the atoms.

The distance traveled is proportional not only to the angle α but also to the axial radius of the rotation *r* (*b*). A χ1 rotation of LEU (*b*) moves more atoms and for a longer distance than a χ2 rotation (*c*). A χ1 rotation of TRP (*d*) moves more atoms than a χ1 rotation of LEU. How do these differences impact the energies of conformational sampling?

To estimate that, we have calculated the 2 kcal/mol "energy wells" of various χ angle rotations for a large number of side chains in their crystal enviroment (*e*).

The 2 kcal well for a χ1 rotation of LEU is on average much narrower than the one for a χ2 rotation (19° vs 28°) (*f*).



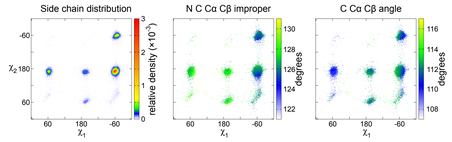
The average wells of the χ rotations for the bulky TRP are, as expected, much narrower (χ1=9° vs χ2=11°) than those for LEU.

Therefore, we hypothesize that if we allocate sampling proportionally to the relative energetic impact of conformational variation we will obtain more effective libraries for side chain optimization.

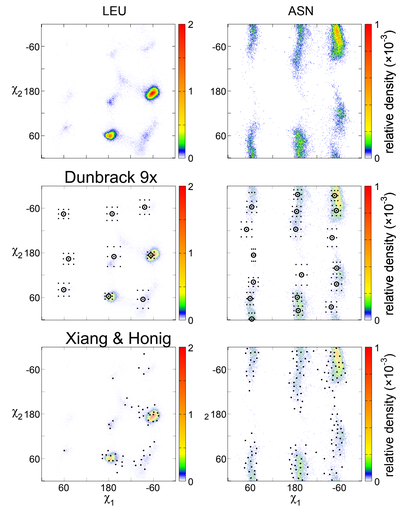
Systematic angle variations occur in side chains.

An example of systematic variations of bond angles can be observed in the different rotameric regions of ILE.

The "N C Cα Cβ" improper and the "C Cα Cβ" are the two degrees of freedom that position the Cβ relative to the backbone.



The current libraries for side chain optimization are constructed with a criterion of geometric representation.



The figure shows two examples of χ angle distributions for a residue that includes only *sp3* carbons (LEU, clustered around the canonical -60, 180, 60 minima) and one with a χ angle that involves a *sp2* carbon (the χ2 of ASN, showing a much more diffused distribution).

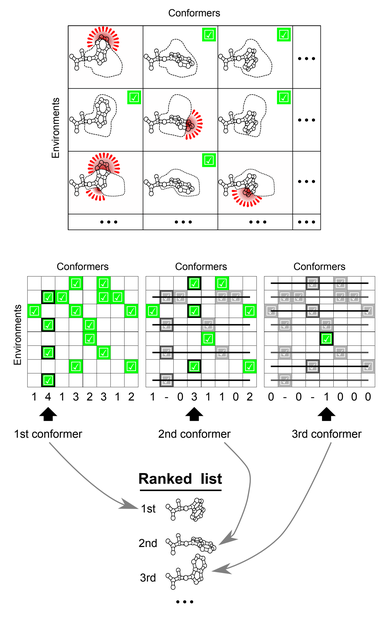
To cover the entire range of conformation, the libraries that are based on statistical data, such as the Dunbrack backbone dependent library[1], need to be supplemented with other conformations in addition to the average rotamer.

The example shows an expansion by ±1 standard deviation in the χ1 and χ2 dimension, a commonly adopted scheme.

The conformer libraries are more directly suitable for fine-grained representation of conformational space. However, they are still built with a criterion of geometrical representation of the space that flattens out the most populated regions and over-represents the sparse regions.

In the example, a library from Xiang and Honig[2] (LEU 53; ASN 152 conformers).

An energy based, ranked fine-grained conformer library.



We have produced a conformer library that disributes sampling based on the energetic impact of side chain conformation instead of pure geometry, using the following methods:

1) An exceedingly fine-grained library of *N* conformers is created
2) A large number *M* of natural environments that contain the same amino acid type are selected from the PDB
3) The native side chains in the enviroments are remodeled into the shape of each one of the conformers and the energy is measured (with the CHARMM 22 all-atoms force field)
4) The operation produces an *N*×*M* table of energies
5) Using a threshold, each energy is converted to a boolean value to indicate if the conformer fits (*T*) or does not fit (*F*) the enviroment
6) The conformer that fits the largest number of enviroments is selected as the top-ranking conformer of the sorted library
7) All environments that have been already satisfied by the previous conformer are no longer considered
8) The conformer that fits the largest number of remaining environments is selected as the best complement to the previous and added to the list
9) "GOTO" #7 and repeat till completion

The method produces a ranked list that is **efficient** because it is built with the same metric (energy) that selects the winner in a side chain optimization procedure.

The ranked library also adds an unprecedented level of **flexibility** because it allows to control sampling – even dynamically – precisely to the level that is needed by a side chain optimization procedure.

The performance of the conformers in test environments provides information about the sampling requirements of the different residues.

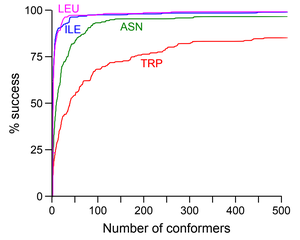
A number of environments were set apart for testing.

The figure shows the percent success in fitting the environments as a function of number of conformers.

Smaller residues with relatively low mobility as LEU and ILE satisfy the majority of the environment with a small number of conformers.

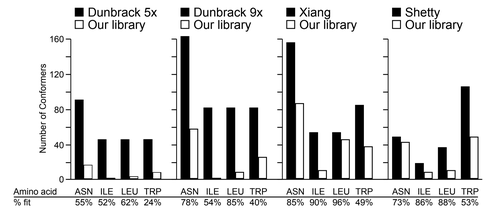
ASN, which is more diffuse, requires more sampling.

The bulky TRP has not fully converged even after 500 conformers.



The energy ranked library is extremely efficient compared to current benchmarks.

To compare our library to a number of existing libraries, we have rebuilt the benchmarks into the test environments and measured the level of success in fitting the environments displayed by each library.

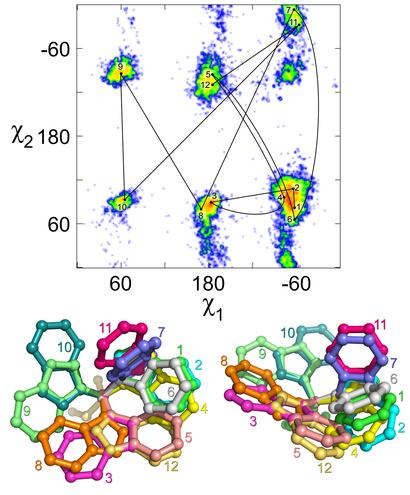


Dunbrack 5x : Dunbrack backbone dependent, incremented by ±1 S.D. in χ_1 OR χ_2 (5x expansion)
Dunbrack 9x : Dunbrack backbone dependent, incremented by ±1 S.D. in χ_1 AND χ_2 (9x expansion)
Xiang : Xiang and Honig, 297 proteins, 94%, 10° conformer library
Shetty : Shetty *et al*. 0.5Å RMSD conformer library

For example, the 5x expansion of Dunbrack's library has 45 rotamers (black bar). When these rotamers are built in the test set, they satisfy (fit) 52% of environments. To satisfy the same percent (or more) of environments we need just one conformer with our library (white bar).

Our library outperforms all benchmarks, and in most cases by a very significant number.

A walk in Trp space: the interactions with the environment direct the sampling to privilege the most populated regions.



The figure shows the top-12 conformers of TRP in the ranked library, displayed over the frequency of the regions in the structural database (as a density map).

The sampling of the energy trained library shows proportionality to the population of the various rotameric regions. For example, four of the top six conformer are focused in the most frequent region (-60°, 90°) of space.

The structural representation of the conformers is shown in the bottom panel from two different perspectives.

Next steps

1) Introduction of an explicit hydrogen bonding function for hydrogen bond optimization.
2) Multi-body testing: we will compare the performance of the library in multi side chain optimization against the benchmarks
3) Creation of a backbone dependent energy optimized library

References

1. Dunbrack RL, Jr. and Cohen FE. (1997) *Protein Science* **6**, 1661-1681
2. Xiang Z, Honig B. (2001) *J Mol Biol* **311**, 421-30
3. Shetty RP, De Bakker PIW, DePristo MA, Blundell TL. (2003) *Protein Eng* **16** 963-9.