An Experimental Analysis of the Performance of Side Chain Packing Algorithms

[Extended Abstract]

Carlos A. Brizuela
CICESE Research Center
Carr. Ensenada-Tijuana 3918
Ensenada, B.C., Mexico
cbrizuel@cicese.mx

Rosario I. Corona
Department of Bioinformatics and Genomics
University of North Carolina at Charlotte
Charlotte, NC 282233, USA
rcorona1@uncc.edu

Christian Lezcano
Facultad Politécnica - UNA
San Lorenzo, Paraguay
clezcanopy@gmail.com

David Rodriguez
CICESE Research Center
Carr. Ensenada-Tijuana 3918
Ensenada, B.C., Mexico
drodrigu@cicese.edu.mx

Jose D. Colbes
CICESE Research Center
Carr. Ensenada-Tijuana 3918
Ensenada, B.C., Mexico
jcolbes@cicese.edu.mx

ABSTRACT

This paper presents a brief description of the protein side chain packing problem (PSCPP) and a performance assessment, on this problem, of three state-of-the-art algorithms: SCWRL4, OPUS-Rota, and CIS-RR. In order to perform a fair comparison, the algorithms are evaluated on three data sets, two of them were previously proposed in the literature and a set of 723 protein structures proposed here. Experimental results show that the achieved accuracy when evaluating the side chain's first torsion angle (χ_1) is of approximately 86% and around 69% for the first and the second torsion angles (χ_{1+2}), for all methods. Although all the algorithms achieve similar accuracies, SCWRL4 requires on average, less computation effort than the others. We highlight relevant aspects that need to be considered in order to verify whether or not this 86% is a theoretical upper bound for the algorithms' performance as well as what might become a promising direction to follow in case an improvement is possible.

Categories and Subject Descriptors

J.1.9 [**Applied Computing**]: Life and Medical Sciences—*Computational Biology, Molecular structural biology*

Keywords

Structure prediction; Protein side-chain packing; Combinatorial Optimization; Rotamer library; Energy function

GECCO '15, July 11 - 15, 2015, Madrid, Spain

DOI: http://dx.doi.org/10.1145/2739482.2768440

1. INTRODUCTION

The protein side-chain packing problem (PSCPP) consist of: given the backbone co-ordinates of each amino acid, select a set of rotamers (one for every amino acid), from a rotamer library, such that a given energy function is minimized. This problem has been proven to be NP-hard [1]. The PSCPP is an essential part of a protein structure prediction method known as homology modeling [22] as well as of the protein design problem [9].

Although many methods such as RASP [17], OSCAR-star [14], SIDEPRO [18], SCMF-PDRL [8], OPUS-Rota [16], CIS-RR [3], and SCWRL4 [11] have been proposed to deal with the PSCPP, comparison works have been rather scarce [19, 17]. Most of the methods present only a brief comparison analysis of their new proposed method against previously proposed ones on particular test instances. In order to assess the performance of each method, over a set that is different from the ones used to tune their parameters, we propose to use a new set consisting of 723 protein structures along with two other sets of 65 and 373 structures, previously used in the literature.

2. PROBLEM STATEMENT

The PSCPP associated to a backbone independent rotamer library is defined as:

Given a sequence of amino acids $\vec{a} = (a_1, a_2, \ldots, a_n)$, the co-ordinates of the backbone atoms $\vec{c} = (\vec{c_1}, \vec{c_2}, \ldots, \vec{c_n})$ and the backbone independent rotamer library r, the PSCPP consist on finding the side chains' torsion angles $\vec{t} = (\vec{t_1}, \vec{t_2}, \ldots, \vec{t_n})$ such that the energy function $f(\vec{a}, \vec{c}, r, \vec{t})$ is minimum, where

$a_i \in \vec{A}$, with $\vec{A}$ the set of 20 amino acids;
$\vec{c_i} = \{\vec{N_i}, \vec{C_i^\alpha}, \vec{C_i}, \vec{O_i}\}$, with $N_i, C_i^\alpha, C_i, O_i \in \Re^3$; and
$\vec{t_i} \in r(a_i)$.

If the rotamer library is backbone dependent then $r(a_i)$ changes to $r(a_i, \phi_i, \psi_i)$, with ϕ_i and ψ_i the backbone torsion angles corresponding to the $i-th$ amnino acid.

3. ALGORITHMS FOR THE SIDE-CHAIN PACKING PROBLEM

Several algorithms have been developed to approach the side-chain packing problem, using a heuristic (either deterministic or randomized) or an exact search method. The methods to solve the PSCPP consist of three main components: a rotamer library, an energy function, and a search algorithm to find the set of rotamers minimizing the energy function.

Recently, there have been some advances in all of the three components mentioned above. The growth and improvement in determining experimentally protein structures have lead to better rotamer libraries which are generated using larger data sets of better quality structures. The energy functions have also benefited from these new structures, either by refining the parameters of the functions or by adding new terms that represent knowledge obtained from them.

Regarding the search algorithms, the improvements concentrated mainly on the efficient computation of the energy function, the bottleneck of the search algorithm. Efficiency is important due to the huge search space that needs to be explored.

Next, we briefly describe each of the methods considered for the performance assessment.

OPUS − Rota [16] uses a backbone-dependent rotamer library [7], and its energy function is described by:

$$E_{total} = w_1 E_{rot} + w_2 E_{vdw} + w_3 E_{orient} + w_4 E_{solvation} \quad (1)$$

The term E_{rot} is related to the probabilities associated with the selected rotamers in a given predicted structure, and the term E_{vdw} represents the Van der Waals interactions in the structure. These terms are used in almost all the methods that tackle the PSCPP; however, the terms E_{orient} and $E_{solvation}$ are unique to OPUS-Rota, these terms incorporate information about the energy associated with an angle conformation and the solvent-accessible surface area of each atom. Each component of the energy function is weighted by a constant factor w_i.

The search method in OPUS-Rota is based on simulated annealing. The process is initialized by positioning all side chains at the rotamers with minimum main-chain/side-chain energy.

SCWRL4 [11] is one of the most accurate, fastest and most frequently used method to solve the side-chain packing problem. It uses one of the latest backbone-dependent rotamer library [21]. The total energy of the protein is expressed by Equation 2.

$$E(S) = \sum_{i=1}^{N} E_{self}(r_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E_{pair}(r_i, r_j), \quad (2)$$

where vector $\mathbf{r}$ specifies a single rotamer for each of the N residues in S.

In this case, the self-energy term (E_{self}) expresses the rotamer energy relative to the most populated rotamer, given the backbone dihedrals, in addition to the energy from interactions (E_{pair}) of the side chain with the backbone and any ligand or other fix atoms in the structure. The pairwise rotamer energies consist of repulsive and attractive van der Waals terms as well as a hydrogen bonding term.

SCWRL4 uses a deterministic search method based on the Dead End Elimination (DEE) [5] algorithm and a tree decomposition approach to solve the combinatorial problem.

CIS − RR [3] uses a backbone-dependent rotamer library [6]. The scoring function was adapted from that of SCWRL3 [2], which consists of two terms:

$$E = E_{vdW} + k_{rot} E_{rot} \quad (3)$$

The term E_{vdW} is an empirical van der Waals potential modified from SCWRL3, and E_{rot} is the rotamer term, which measures the preferences of the side-chain conformers, k_{rot} is a weighting factor to balance the relative importance of E_{vdW} and E_{rot}.

The search method of CIS-RR focuses on minimizing the atomic clashes in the predicted structure. This is done by a phase called *Rotamer Relaxation* (*RR*), which has shown to lower significantly the number of clashes.

The starting side-chain conformation of each residue is constructed by the rotamers with the highest probability at each position. Then, for each residue i, every one of its rotamers will be optimized (by *RR*) and tested for clashes with the other residues that are kept fix [3].

There are other recent approaches for the PSCPP like RASP [17], OSCAR-star [14], SIDEPRO [18], and SCMF-PDRL [8]; however, these are not considered since they essentially achieve similar results as those of the methods studied here.

4. EXPERIMENTAL SETUP AND RESULTS

4.1 Datasets

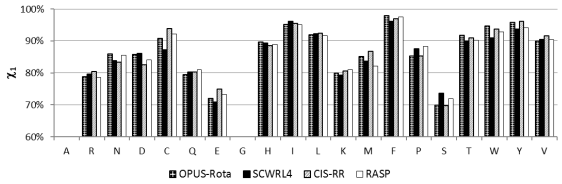
We use three datasets to make the performance comparison. They consist of 65, 373, and 723 protein structures from the Protein Data Bank (PDB). The first two data sets were used previously [20, 10, 16, 11, 3], and the last one is proposed in this work. A brief description of each dataset is given below:

- Dataset-65 [20]: 30 proteins were taken from [13]. For this subset, sequence identity cutoff was set to 50%, the resolution cutoff was set to 1.8 Å, and the R-factor cutoff was set to 20%. Only single-chain proteins with 100-500 residues and containing no incomplete side chains or ligands were selected. 28 proteins were taken from [24]; some of them have a resolution between 0.83-1.4 Å and a pair-wise sequence identity of less than 20%, while the others have a resolution better than 1.2 Å and more than 40 residues. The remaining 7 proteins were selected from the PDB using the criteria of having crystallographic resolutions better than 1.2 Å, and sequence length between 150-300 residues.

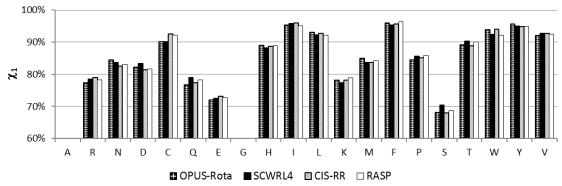
- Dataset-373 [11]: this is a subset from the dataset of 379 proteins proposed for testing SCWRL4, removing the ones already present in the Dataset-65. The proteins have a sequence length between 40-1000 residues, with resolutions better than 1.8 Å, maximum mutual sequence identity lower than 30%, and maximum R-factor of 20%.

- Dataset-723: proteins in this heterogeneous set have a single chain with 40-400 residues, a single domain under SCOP classification (within class a, b, c, and d),

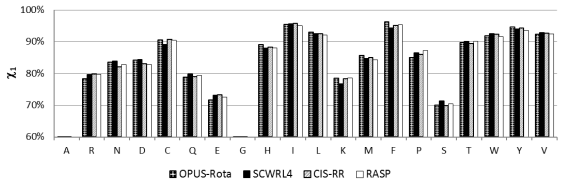
a maximum R-factor of 20%, a maximum resolution of 2 Å, a maximum sequence identity of 25%; and their structures were determined by X-ray crystallography.



(a) Dataset-65



(b) Dataset-373



(c) Dataset-723

Figure 1: χ_1 accuracy by amino acid for each data set

4.2 Quality measures

A comparison of the three considered methods is available [17]; however, it uses only the dataset proposed for SCWRL4 [11], with general and amino acid accuracy comparisons. One of the objectives of this study is to extend this work by using three datasets, providing a fair comparison of the most accurate methods available to date. We will consider aspects like accuracy over χ_1 and χ_{1+2}, by amino acid and global, as well as running time of each method.

The measure used in the literature to show the accuracy of a method is the percentage of correctly predicted side-chain torsion angles. This is called *absolute accuracy* [11]. A predicted side-chain torsion angle (χ') is *correct* if its error is less than or equal to a specified threshold (usually 40° [7]), with respect to the angle (χ) calculated from the original pdb entry. To calculate the error of a predicted side-chain torsion angle, we use the following expression:

$$e(\chi, \chi') = \min(|\chi - \chi'|, 360 - |\chi - \chi'|) \,, \qquad (4)$$

Table 1 shows the accuracies of OPUS-Rota [16], SCWRL4 [11], and CIS-RR [3] as they are reported in the literature. Fortunately, all these works used 40° as the threshold for the accuracy calculations; however, they reported slightly different results due to the differences in the data sets employed. For instance, in [16], [11], [3], and [15] the authors used datasets of 65, 379, 180, and 218 protein structures, respectively; besides, in [11] they used only side chains with electron density in a given range.

Table 1: Accuracy of OPUS-Rota, SCWRL4, and CIS-RR reported by previous work under different conditions.

Method	Ref.	χ_1(%)	χ_{1+2}(%)
OPUS-Rota	[16]	89.00	79.10
OPUS-Rota	[15]	86.60	75.70
OPUS-Rota	[17]	85.03	75.05
SCWRL4	[11]	89.30	79.70
SCWRL4	[3]	85.80	76.30
SCWRL4	[15]	85.10	74.00
SCWRL4	[17]	85.03	75.44
CIS-RR	[3]	86.40	76.70
CIS-RR	[17]	84.88	74.88

As it can be noticed in Table 1, the reported accuracies are not consistent for each method, and cannot be compared easily due to the different experimental conditions. For instance in [16] Opus-Rota is run on a set of 65 proteins while in [15] and [17] it was evaluated on sets of 218 and 379 structures, respectively.

4.3 Accuracy comparison: Results

There are different ways to compute the accuracy depending on the characteristics we want to highlight. Figure 1 shows the accuracy regarding the χ_1 torsion angle with respect to each amino acid for the three different data sets. We can observe that the behavior of this accuracy is similar over the data sets. Alanine (A) and Glycine (G) do not have side chain. Serine (S) obtains the lowest accuracy, this is not an expected result since serine has a short side chain and shorter side chains should be easier to predict than longer ones since they have fewer rotamers.

In Table 2 we present the overall accuracy of the three methods for each dataset (65, 373, and 723 proteins, respectively). The overall accuracy was obtained calculating the percentage of the correct angles for all the residues in the datasets. From Table 2 we can observe that all methods perform similarly with respect to χ_1 and χ_{1+2}. There is a slight advantage for OPUS-Rota respect to the other methods.

Table 2: Results of accuracy for each method and each dataset.

	χ_1(%)			χ_{1+2}(%)		
Dataset	65	373	723	65	373	723
OPUS-Rota	86.70	85.66	85.90	69.68	69.01	68.60
SCWRL4	85.97	85.37	85.52	68.21	68.74	68.24
CIS-RR	86.18	85.10	85.35	68.70	68.46	68.10

Computation time is another characteristic, besides the accuracy, that needs to be taken into account when choosing an algorithm to predict the side-chain atom positions, specially when dealing with the protein design problem. Table 3 shows that the average running time is consistent among the three datasets. The algorithm that shows the lowest average computation time with the highest variability is SCWRL4 [11], while the other two methods have similar computation times.

We also observed (results not shown here) that SCWRL4 has many outliers (with respect to computation time), and

some of them are running times that double the maximum computation time of the other methods. This obeys to the following reason; when SCWRL4 spends more than a certain amount of time on a given structure it switches to a heuristic procedure to assure convergence (at the expense of losing optimality). It would be very interesting to characterize the structures that generate the outliers for the SCWRL4 in order to understand what makes a structure harder to predict than another.

After analyzing the results of this brief experimental comparison of side chain packing methods, the following questions arises:

1. Is 86% the maximum achievable value for the χ_1 average accuracy?

A simple idea to answer this question is to take rotamers from a simple library and compute the best match it can achieve for each amino acid on a given set of protein structures.

2. If the rotamer library is able to achieve higher accuracy values, then is the limitation in: i) the energy function and/or ii) the search algorithm?

A recent work on protein design algorithms [12] indicates that the energy functions still fail to correctly model the interactions within a protein, so they could be the main responsible for the marginal improvements in PSCPP. A similar conclusion was reached working with Rosetta for structure prediction of small proteins [4].

Table 3: Computation time (in seconds) for each method and each dataset.

	Dataset-65		Dataset-373		Dataset-723	
	AVG	SD	AVG	SD	AVG	SD
OPUS-Rota	10.66	5.94	15.77	9.88	10.46	4.93
SCWRL4	5.33	11.06	5.72	5.34	4.09	6.41
CIS-RR	11.48	9.01	18.95	16.90	10.22	6.84

5. CONCLUSION

In this work a performance assessment of three state-of-the-art methods for the PSCPP in terms of accuracy and running time is presented. To achieve this, we used two datasets proposed in the literature and added a larger heterogeneous one that considers SCOP classes, minimum values for resolution, R-factor, and a maximum sequence identity between every pair of proteins in the set.

Regarding the accuracy results, we have not found a significant difference between methods considering general and by amino acid accuracies. All the methods achieve approximately 86% for χ_1. However, a most important remaining question is to decide whether this 86% accuracy is a limit in some sense or there is still room for improvements. If the improvement is possible we need to determine if the current results are due to limitations in the search methods, the rotamer library, or the energy functions.

Future work is aimed at analyzing the performance of state-of-the-art methods' energy functions. For this, the correlation between the energy and accuracy could be analyzed as it is done elsewhere [23], or a local search method could be applied to the native structure. In the latter case, an ideal energy function would at least consider the native structure as a local minimum.

6. ACKNOWLEDGMENTS

This work was partially supported by the National Council of Science and Technology of Mexico (www.conacyt.mx) under grant SEP-CONACYT-CB-2010-154737.

7. REFERENCES

[1] T. Akutsu. Np-hardness results for protein side-chain packing. *Genome Informatics*, 8:180–186, 1997.

[2] A. Canutescu, A. Shelenkov, and R. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12(9):2001–2014, 2003.

[3] Y. Cao, L. Song, Z. Miao, Y. Hu, L. Tian, and T. Jiang. Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. *Bioinformatics*, 27(6):785–790, 2011.

[4] R. Das. Four small puzzles that rosetta doesn't solve. *PLOS One*, 6(5):e20044, 2011.

[5] J. Desmet, M. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in side-chain positioning. *Nature*, 356:539–542, 1992.

[6] R. Dunbrack and F. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, 6:1661–1681, 1997.

[7] R. Dunbrack and M. Karplus. Backbone-dependent rotamer library for proteins. application to side-chain prediction. *Journal of Molecular Biology*, 230:543–574, 1993.

[8] P. Francis-Lyon and P. Koehl. Protein side-chain modeling with a protein-dependent optimized rotamer library. *Proteins: Structure, Function, and Bioinformatics*, 82(9):2000–2017, 2014.

[9] P. Gainza, K. Roberts, and B. Donald. "Protein Design Using Continuous Rotamers". *PLOS Computational Biology*, 8(1):1–15, 2012.

[10] T. Jain, D. Cerutti, and J. McCammon. Configurational-bias sampling technique for predicting side-chain conformations in proteins. *Protein Science*, 15(9):2029–2039, 2006.

[11] G. Krivov, M. Shapovalov, and R. Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 27(6):785–790, 2009.

[12] Z. Li, Y. Yang, J. Zhan, L. Dai, and Y. Zhou. Energy functions in de novo protein design: Current challenges and future prospects. *Annual Review of Biophysics*, 42:315–335, 2013.

[13] S. Liang and N. Grishin. Side-chain modeling with an optimized scoring function. *Protein Science*, 11(2):322–331, 2002.

[14] S. Liang, D. Zheng, C. Zhang, and D. Standley. Fast and accurate prediction of protein side-chain conformations. *Bioinformatics*, 27(20):2913–2914, 2011.

[15] S. Liang, Y. Zhou, N. Grishin, and D. Standley. Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions. *Journal of Computacional Chemistry*, 32(8):1680–1686, 2011.

[16] M. Lu, A. Dousis, and J. Ma. OPUS-Rota: a fast and accurate method for side-chain modeling. *Protein Science*, 17(9):1576–1585, 2008.

[17] Z. Miao, Y. Cao, and T. Jiang. RASP: rapid modeling of protein side chain conformations. *Bioinformatics*, 27(22):3117–3122, 2011.

[18] K. Nagata, A. Randall, and P. Baldi. Sidepro: A novel machine learning approach for the fast and accurate prediction of side-chain conformations. *Proteins: Structure, Function, and Bioinformatics*, 80(1):142–153, 2012.

[19] L. X. Peterson, X. Kang, and D. Kihara. Assessment of protein side-chain conformation prediction methods in different residue environments. *Proteins: Structure, Function, and Bioinformatics*, 82(9):1971–1984, 2014.

[20] R. Peterson, P. Dutton, and A. Wand. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Science*, 13(3):735–751, 2004.

[21] M. Shapovalov and R. Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Bioinformatics*, 19(22):844–858, 2011.

[22] H. Venselaar, E. Krieger, and G. Vriend. Homology modeling. *Structural Bioinformatics*, pages 715–732, 2009.

[23] P. Widera, J. M. Garibaldi, and N. Krasnogor. Gp challenge: Evolving energy function for protein structure prediction. *Genetic Programming and Evolvable Machines*, 11(1):61–88, 2010.

[24] Z. Xianga and B. Honig. Extending the accuracy limits of prediction for side-chain conformations. *Journal of Molecular Biology*, 311(2):421–430, 2001.