Minimizing and Learning Energy Functions for Side-Chain Prediction

Chen Yanover[1], Ora Schueler-Furman[2], and Yair Weiss[1]

[1] School of Computer Science and Engineering,
The Hebrew University of Jerusalem, 91904 Jerusalem, Israel
{cheny,yweiss}@cs.huji.ac.il
[2] Department of Molecular Genetics and Biotechnology, Hadassah Medical School
The Hebrew University of Jerusalem, 91120 Jerusalem, Israel
oraf@ekmd.huji.ac.il

Abstract. Side-chain prediction is an important subproblem of the general protein folding problem. Despite much progress in side-chain prediction, performance is far from satisfactory. As an example, the ROSETTA program that uses simulated annealing to select the minimum energy conformations, correctly predicts the first two side-chain angles for approximately 72% of the buried residues in a standard data set. Is further improvement more likely to come from better search methods, or from better energy functions? Given that exact minimization of the energy is NP hard, it is difficult to get a systematic answer to this question.

In this paper, we present a novel search method and a novel method for learning energy functions from training data that are both based on Tree Reweighted Belief Propagation (TRBP). We find that TRBP can find the *global* optimum of the ROSETTA energy function in a few minutes of computation for approximately 85% of the proteins in a standard benchmark set. TRBP can also effectively bound the partition function which enables using the Conditional Random Fields (CRF) framework for learning.

Interestingly, finding the global minimum does not significantly improve side-chain prediction for an energy function based on ROSETTA's default energy terms (less than 0.1%), while learning new weights gives a significant boost from 72% to 78%. Using a recently modified ROSETTA energy function with a softer Lennard-Jones repulsive term, the global optimum does improve prediction accuracy from 77% to 78%. Here again, learning new weights improves side-chain modeling even further to 80%. Finally, the highest accuracy (82.6%) is obtained using an extended rotamer library and CRF learned weights. Our results suggest that combining machine learning with approximate inference can improve the state-of-the-art in side-chain prediction.

1 Introduction

Proteins are chains of *residues*, each containing one of 20 possible *amino acids*. All amino acids are connected together by a common backbone structure, onto which amino-specific *side-chains* are attached. The 3-dimensional structure of a protein can thus be fully defined by the dihedral angles that specify the backbone conformation on the one hand (ϕ, ψ and ω angles), and the side-chain conformations on the other hand (up to 4 dihedral angles, denoted χ_1 to χ_4).

T. Speed and H. Huang (Eds.): RECOMB 2007, LNBI 4453, pp. 381–395, 2007.

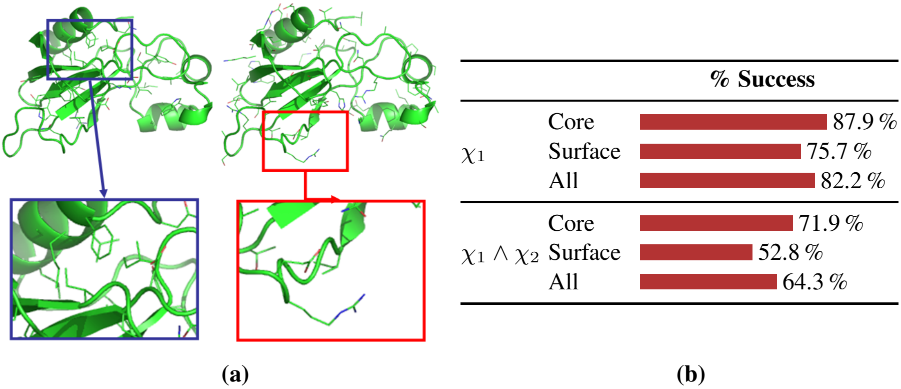


Fig. 1. **(a)** Buried and exposed residues of Barnase (PDB code 1brn). The challenge in side-chain prediction is to locate the native side-chain conformation (sticks), starting from the protein backbone (depicted as a cartoon), and its amino acid sequence. Blowups for specific regions are shown for buried residues (left) and exposed residues (right). Note that due to packing, core residues are significantly more constrained than their exposed counterparts. **(b)** Success rate of the state-of-the-art ROSETTA package using default parameters. It can be seen that even for the core, the fraction of residues for which either the χ_1 or χ_2 angles are incorrectly modeled is about 30%. Is improvement more likely to come from better search methods or from better energy functions?

The problem of predicting the residue side-chain conformations given a backbone structure is considered of central importance in protein-folding and molecular design and has been tackled extensively using a wide variety of methods (for a recent review, see [1]). The typical way to predict side-chain configurations is to define an energy function and a discrete set of possible side-chain conformations, and then search for the minimal energy configuration.

Despite much progress, the performance of side-chain prediction is far from satisfactory. To illustrate the state-of-the-art, Figure 1b shows the results of the ROSETTA package [2] on a standard benchmark set. The prediction success is typically reported separately for core residues and surface residues, since core residues are much more tightly constrained (see Figure 1a). ROSETTA uses an elaborate energy function for side-chain modeling that contains 8 energy terms. Simulated annealing is used to search for the minimal energy configuration. As can be seen, the success rate for the first two angles is around 72% for core residues and 53% for surface residues. Thus even for the better constrained residues, the prediction is wrong for almost one third of the residues.

One can think of two different approaches to improve this performance: (1) using a better optimization algorithm to find a lower energy conformation; and (2) changing the energy function. Deciding between these two approaches is currently difficult because simulated annealing and many of the other minimizers used in side-chain prediction are only guaranteed to find *local* minima of the energy function. We therefore do not know if a better optimizer would find a better solution.

Obviously a method that can find the *global* optimum of the energy function could shed light on this question. Unfortunately, it has been shown that for energy functions typically used in side-chain prediction, finding the global optimum is NP complete [3]. While this makes it extremely unlikely that we will be able to find the global optimum for *all* proteins in polynomial time, it leaves open the option for finding the global optimum for *some* proteins. Indeed, methods such as dead-end-elimination (DEE) [4,5,6] and linear programming relaxations [7] have been shown to find the global optimum for simple energy functions in side-chain prediction. However, as reported in [7], these techniques do not work well for more complicated energy functions and to the best of our knowledge, no one has successfully found the global optimum for the elaborate ROSETTA energy function.

In this paper, we present a novel search method and a novel method for learning energy functions from training data that are both based on Tree Reweighted Belief Propagation (TRBP). We find that TRBP can find the *global* optimum of the ROSETTA energy function in a few minutes of computation for approximately 85% of the proteins in a standard benchmark set. TRBP can also effectively bound the partition function which enables using the Conditional Random Fields (CRF) framework for learning of better energy functions.

Interestingly, finding the global minimum does not significantly improve side chain prediction for an energy function based on ROSETTA's default energy terms (less than 0.1%), while learning new weights gives a significant boost from 72% to 78%. A recent modification of the ROSETTA energy function is aimed at optimal side-chain modeling and uses a softer van der Waals term [8]. This energy function yields significantly better results than ROSETTA's default parameters (77% with simulated annealing). In this case, the global optimum improves prediction accuracy by 1.2%. Learning new weights again improves side-chain modeling, to 80%. Finally, not unexpectedly, the use of extended rotamer libraries improves modeling: combined with CRF learned weights it yields the highest accuracy (82.6%). Our results suggest that combining machine learning with approximate inference can improve the state-of-the-art in side-chain prediction.

2 Side-Chain Prediction

The input to the side-chain prediction task, which we will denote by y, is a list of amino-acids that make up the protein as well as the three-dimensional shape of the backbone. The output, which we will denote by x, is up to 4 dihedral angles, denoted χ_1 to χ_4, for each amino acid. In principle, the output is a continuous valued vector whose length is 4 times the number of amino acids in the protein. However, the common practice is to discretize the output space into a small number of possible angles. These discrete angles (usually up to 3 possibilities per angle) define a discrete set of possible side-chain configurations called *rotamers* [9]. Side-chain prediction thus becomes a discrete optimization problem:

$$x^* = \arg\min_{x \in \mathcal{R}} E(x, y) \tag{1}$$

where $\mathcal{R}$ is the discrete set of rotamer configurations, and the energy function $E(x, y)$ is, typically, defined in terms of pairwise interactions among nearby residues and interactions between a residue and the backbone. Approaches to side-chain prediction differ in their choices of energy functions and search methods.

Search Methods. Although the minimization problem for side-chain prediction has been shown to be NP hard [3], recent years have shown significant progress in search methods. Simulated annealing with Monte Carlo sampling used in Rosetta is a fast and efficient method to locate energy minima, but is not guaranteed to find the *global* minimum energy conformation.

The dead end elimination (DEE) algorithm is an exhaustive search algorithm that tries to reduce the search space as much as possible. It is based on a simple condition that identifies rotamers that cannot be members of the global minimum energy conformation [4,5,6]. In cases where enough rotamers can be eliminated, the *global minimum energy conformation* can be found by an exhaustive search of the remaining rotamers.

Kingsford *et al.* [7] used the method of *Linear Programming (LP) Relaxation* to locate the global optimum. They rewrote equation (1) as an integer program and then relaxed the integer constraints to obtain a linear program. They found that for an energy function similar to SCWRL [1], the LP solution was almost always integral, meaning that the LP relaxation found the *global minimum*. However, once they added a second energy term, the percentage of problems for which LP found an integer solution dropped dramatically. They also discussed using a commercial Integer-Programming package (CPLEX) and found it could work on the two-term energy functions that LP could not solve.

Energy Functions. Many of the early energy functions were primarily based on the repulsive part of the van der Waals energy term. The successful SCWRL program [1,9] approximates the repulsive portion of the 12-6 Lennard-Jones potential with a piecewise linear function. SCWRL also takes into account the prior probabilities of rotamers in a training set.

ROSETTA's energy function that is used for side-chain prediction also includes a repulsive term and prior probabilities of rotamers, but combines these with six other terms to obtain an atomic level, physically realistic energy function. Specifically it contains the following energy terms [10]:

1. The attractive portion of a 12-6 Lennard-Jones potential (herein denoted by *atr*).
2. The repulsive portion of a 12-6 Lennard-Jones potential (*rep*). This term is dampened in order to compensate for the use of a fixed backbone and rotamer set.
3. A solvation term, calculated using the model of Lazaridis and Karplus [11] (*sol*).
4. Rotamer energy: Backbone dependent internal free energies of rotamers, estimated from PDB statistics performed by Dunbrack and Karplus [9] (*dun*).
5. A hydrogen-bonding potential, dependent both on distance and angles [12]. For historical reasons, this term was divided into: **(a)** Side-chain to side-chain interactions (*hbond* 1); **(b)** Side-chain to backbone interactions (*hbond* 2); and **(c)** Backbone to backbone interactions (constant for the task of side-chain prediction).
6. A pair term that primarily reflects the electrostatic attraction and repulsion (*pair*). It describes the tendency of polar amino acid residues to contact each other, based on a statistical analysis of PDB structures of seeing two amino acids close together in space (after accounting for the intrinsic probabilities of these amino acids to be in that environment).
7. An internal term that reflects clashes within a side-chain conformation (*intra*).

The energy function, $E(x, y)$ is defined as a weighted sum of the eight terms. Denoting by λ_i the weight of the ith term, the energy is:

$$E(x, y; \lambda) = \sum_{i=1}^{8} \lambda_i E_i(x, y) \qquad (2)$$

2.1 Learning Energy Functions

Most current energy functions are based on a combination of parameters that describe different aspects of a protein structure, some from physical chemistry (such as *atr* and *rep*), others from analyses of given protein structures (such as *dun*).

What is the relative importance of the different terms in the energy function? The supervised learning problem of setting the relative contribution, i.e. the *weights*, of the energy terms can be formulated as follows: given a set of training proteins $\{x_t, y_t\}_{t=1}^{T}$, where x_t is the side-chain configuration in the crystal structure of protein t and y_t denotes its backbone structure, seek parameters λ that maximize the prediction success rate. Kuhlman and Baker [2] used a conjugate gradient-based optimization method to optimize the weights of these energy terms by decreasing the energy of the native state relative to a small number of decoy configurations.

Conditional Random Fields [13] provide a principled way of learning energy functions from labeled data [14, 15, 16]. Defining the probability of the native side-chain configuration (for a given backbone structure) as:

$$\Pr(x_t | y_t; \lambda) = \frac{1}{Z_t(\lambda)} e^{-E(x_t, y_t; \lambda)} \qquad (3)$$

with:

$$Z_t(\lambda) = \sum_{x \in \mathcal{R}} e^{-E(x, y_t; \lambda)} \qquad (4)$$

CRFs seek to maximize the product of the probabilities $\Pr(x_t | y_t; \lambda)$ over all training proteins $\{x_t, y_t\}_{t=1}^{T}$. The term "Conditional Random Fields" comes from the fact that we are maximizing the conditional likelihood – we are not maximizing the joint probability of side-chain and backbone, but rather the conditional probability of a side-chain configuration given the backbone. CRFs have several attractive properties for learning energy functions: the conditional log likelihood is a convex function of the parameters λ and the gradient of the log likelihood is simply:

$$\frac{\partial \ln \Pr(x_t | y_z; \lambda)}{\partial \lambda_i} = -E_i(x_t) + < E_i >_{\lambda} \qquad (5)$$

Hidden Conditional Random Fields (HCRFs) [17, 18] extend conditional random fields to settings where some of the variables are hidden. This is simply done by marginalizing out the hidden variables. In practice, a Viterbi approximation in which the marginalization is replaced with maximization, is often used [15]. This leads to maximizing:

$$\Pr(x_t | y_t; \lambda) \approx \max_{h} \frac{1}{Z_t(\lambda)} e^{-E(x_t, y_t, h; \lambda)} \qquad (6)$$

where h are the hidden variables.

Applying the CRF framework to side-chain prediction raises a tremendous computational challenge. Note that calculating Z_t (equation (4)) requires summing over all possible side-chain configurations for a given protein. For the vast majority of proteins this summation is intractable. Similarly, calculating the gradient in equation (5) is based on taking expectations which requires a weighted sum over all possible side-chain configuration for a given protein. Finally, equation (6) requires maximizing over all possible configurations for the hidden variables. Similar computational problems arise with other supervised learning methods for learning energy functions [19, 15].

3 Tree Reweighted Belief Propagation

To summarize the results of the previous section, side-chain prediction raises major computational difficulties, either in finding the global minimum of the energy or calculating the partition function with respect to an energy function. In this work, we use tree-reweighted belief propagation (TRBP) to address both problems.

Tree-reweighted belief propagation (TRBP) is a variant of belief propagation introduced by Wainwright and colleagues [20]. We start by briefly reviewing ordinary max-product belief propagation (see e.g. [21,22]). The algorithm receives as input a graph G and the potentials Ψ_{ij}, Ψ_i. In energy minimization settings, the potentials are inversely related to the energy: $\Psi_{ij}(x_i, x_j) = e^{-E(x_i,x_j)}, \Psi_i(x_i) = e^{-E(x_i)}$. In the side-chain prediction setting the nodes of the graphs correspond to residues, and there are edges between any two residues that interact [23].

At each iteration, a node i sends a message $m_{ij}(x_j)$ to its neighbor in the graph j. The messages are updated as follows:

$$m_{ij}(x_j) \leftarrow \alpha_{ij} \max_{x_i} \Psi_{ij}(x_i, x_j)\Psi_i(x_i) \prod_{k \in N_i \setminus j} m_{ki}(x_i) \tag{7}$$

where $N_i \setminus j$ refers to all neighbors of node i except j. The constant α_{ij} is a normalization constant typically chosen so that the messages sum to one (the normalization has no influence on the final beliefs). After the messages have converged, each node can form an estimate of its local “belief” defined as:

$$b_i(x_i) \propto \prod_{j \in N_i} m_{ji}(x_i)\Psi_i(x_i) \tag{8}$$

It is easy to show that when the graph is singly-connected, choosing an assignment that maximizes the local belief will give the minimal energy configuration [22]. In fact, when the graph is a chain, equation (7) is simply a distributed computation of dynamic programming. When the graph has cycles, ordinary belief propagation (BP) is no longer guaranteed to converge, nor is there a guarantee that it can be used to find the minimal energy configuration.

In tree-reweighted BP (TRBP), the algorithm receives an additional set of input *edge appearance probabilities*, ρ_{ij}. These edge appearance probabilities are essentially free parameters of the algorithm and are derived from a distribution over spanning trees of the graph G. They represent the probability of an edge (i, j) to appear in a spanning tree under the chosen distribution. As in standard belief propagation, at each iteration a node i sends a message $m_{ij}(x_j)$ to its neighbor in the graph j. The messages are updated as follows:

$$m_{ij}(x_j) \leftarrow \alpha_{ij} \max_{x_i} \Psi_{ij}^{1/\rho_{ij}}(x_i, x_j) \Psi_i(x_i) \frac{\prod_{k \in N_i \setminus j} m_{ki}^{\rho_{ki}}(x_i)}{m_{ji}^{1-\rho_{ji}}(x_i)} \tag{9}$$

Note that for $\rho_{ij} = 1$ the algorithm reduces to standard belief propagation.

After one has found a fixed-point of these message update equations, the singleton and pairwise beliefs are defined as:

$$b_i(x_i) \propto \Psi_i(x_i) \prod_{j \in N_i} m_{ji}^{\rho_{ji}}(x_i)$$

$$b_{ij}(x_i, x_j) \propto \Psi_i(x_i) \Psi_j(x_j) \Psi_{ij}^{1/\rho_{ij}}(x_i, x_j) \cdot \frac{\prod_{k \in N_i \setminus j} m_{ki}^{\rho_{ki}}(x_i)}{m_{ji}^{1-\rho_{ji}}(x_i)} \frac{\prod_{k \in N_j \setminus i} m_{kj}^{\rho_{kj}}(x_j)}{m_{ij}^{1-\rho_{ij}}(x_j)}$$

The theoretical properties of TRBP are a subject of ongoing research [20,24,25,26]. We briefly summarize some relevant properties:

- If the TRBP beliefs contain no ties, that is for every i the maximum of $b_i(x_i)$ is attained at a unique value, then the assignment that locally maximizes the beliefs is the global minimum of the energy function.
- If the TRBP beliefs contain ties, running an additional algorithm on a problem defined only on nodes that have ties, gives an easily verified condition for the solution to be a global optimum (see [26] for details).
- Using the sum-product version of TRBP (in which the maximization in equation (9) is replaced with summation) it is possible to calculate a rigorous upper bound Z_{TRBP} on the partition function.

$$-\log Z_{TRBP} =< E >_b - \left(\sum_{ij} \rho_{ij} H(b_{ij}) + \sum_i c_i H(b_i) \right) \tag{10}$$

  where $c_i = 1 - \sum_j \rho_{ij}$ and $< E >_b$ is the average energy with respect to the TRBP beliefs, and $H(b_{ij}), H(b_i)$ are the entropies of the beliefs.

We used these properties of TRBP for minimizing and learning energy functions for side-chain prediction. For minimizing energy functions, we used the max-product version of TRBP followed by post-processing as described in [26]. For learning energy functions, we replaced the partition function $Z(\lambda)$ in equations (3),(6) with the TRBP bound $Z_{TRBP}(\lambda)$. This enables us to maximize a lower bound on the probability:

$$\Pr(x_t | y_t; \lambda) = \frac{1}{Z(\lambda)} e^{-E(x_t, y_t; \lambda)} \geq \frac{1}{Z_{TRBP(\lambda)}} e^{-E(x_t, y_t; \lambda)} \tag{11}$$

We used the implementation of TRBP publically available at `www.cs.huji.ac.il/~talyam/inference.html`. The same package was also used to solve the LP relaxation, as discussed in [27].

4 Results

In the first part of this study, we evaluate whether location of the global minimum energy conformation improves side-chain modeling accuracy. We then proceed to improving the energy function by optimizing the weights of the different parameters in the energy function to maximize the probability of native side-chain conformations. We show that this improves side-chain prediction accuracy more than finding the minimum energy conformation. The next step evaluates those approaches on an additional energy function with a softer repulsive term, and finally we investigate the use of extended rotamer libraries.

Data set and Evaluation. A data set of 276 single chain proteins, up to 700 amino acids long (all in all 64,397 positions) was used for this study (taken from the *RosettaDesign* webserver [28]). We randomly selected 20% of these proteins (55 proteins, 11,067 positions) as a training set and used the remaining 80% (221 proteins, 53,330 positions) as a test set.

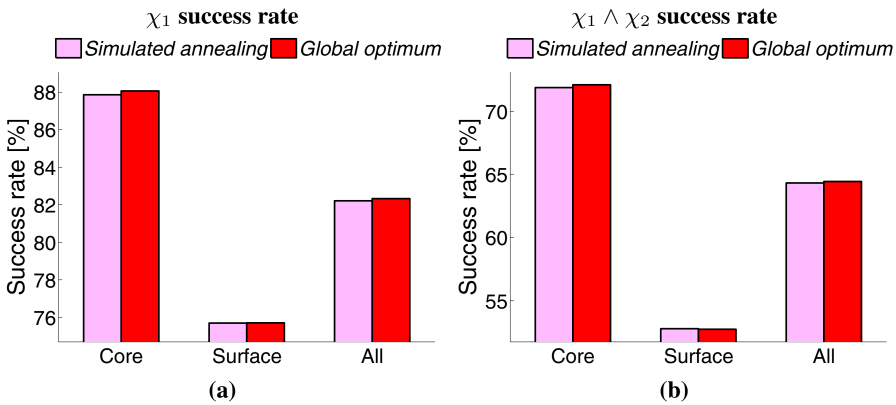
We define the success rate of an energy function as the percentage of side-chain angles that are predicted correctly, i.e. when the predicted angles are in the same bin as those of native side-chain conformation in the crystal (e.g. gauche+, gauche−, or trans). As widely accepted, we report the success rates for the first angle (χ_1) and the first two angles (χ_1 and χ_2) on all test set proteins. We also calculated the success rate separately for core residues, defined as residues with more than 19 interacting neighbors, and surface residues (up to 19 interacting neighbors), where residues are termed neighbors if the distance between their C_β atoms is less than $10\AA$.

4.1 Location of Global Minimum Energy Configuration

Our first set of experiments was designed to measure the importance of locating the global minimum energy conformation of the energy functions currently used in side-chain prediction. We first asked which methods can find the global optimum in reasonable time? Consistent with Kingsford *et al.*'s report, the LP relaxation works well for the simple SCWRL energy function (over 90% in a database of 370 proteins) but rarely does so for the ROSETTA function (less than 5%). In other words, the LP solution is almost never integer for the ROSETTA energy functions. In contrast, the TRBP method finds the global optimum in over 80% of the proteins in our database for ROSETTA, while the commercial Integer Programming package (CPLEX) can find the minimum for all the examples in our database (although its run time is generally much larger than that of TRBP). Also consistent with the report in [7], DEE [5] never found the global optimum for these problems, indicating that not enough rotamers could be eliminated.

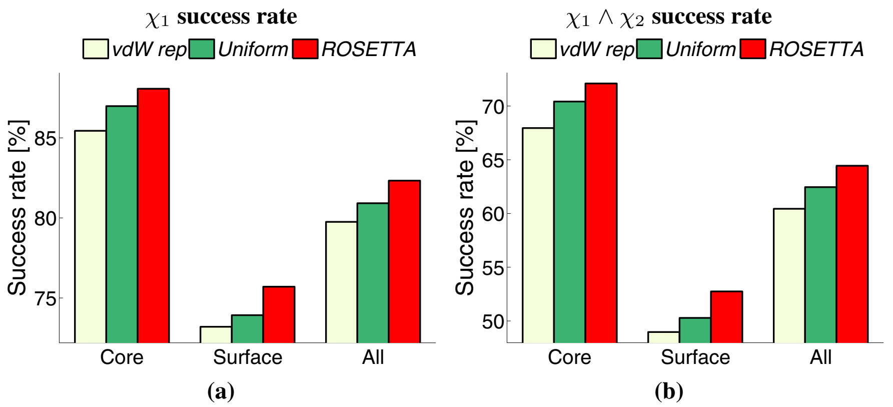
How much then does location of the global minimum energy conformation improve performance? Our results indicate that the improvement obtained from locating the global minimum energy (compared to simulated annealing) is negligible: side-chain modeling accuracy for the first two χ angles of core, surface and all residues are essentially unchanged (Figure 2).

Given a method that can find the global minimum energy, a better comparison of the usefulness of different energy functions for side-chain modeling can be performed:



(a) **(b)**

Fig. 2. The location of the global minimum energy conformation does not improve side-chain modeling for the ROSETTA original (default) energy function. The percentages of correctly predicted χ_1 side-chain angles **(a)**, and both χ_1 and χ_2 angles **(b)** are indicated for the whole set of side-chains, as well as for the buried and exposed subsets separately.



(a) **(b)**

Fig. 3. Comparison of different energy functions with global minimization. Side-chain prediction success rate for energy functions that use either ROSETTA's repulsive van der Waals and rotamer energy terms (*vdw rep*) only, the full ROSETTA energy function with uniform weights, or ROSETTA's default weights.

For which energy function do the global energy minima coincide best with near-native models? Figure 3 compares different energy functions defined by different weightings of ROSETTA's eight energy terms – using only the repulsive van der Waals (rep) and rotamer energy (dun) terms (which simulates the setup of SCWRL [1]), using a uniform weighting on all eight terms, and using ROSETTA's default weights. It can be seen that the van der Waals and rotamer terms on its own give the worst performance, followed by a uniform weighting of ROSETTA's eight terms and the best performance is given by ROSETTA's weights. These results are consistent with previously reported conclusions. Note however that in the present study effects due to correlation between the energy function and the search algorithm are excluded since only proven global energy minima are considered. For all three cases, we can therefore conclusively attribute the improvement in performance to a more accurate energy function.

4.2 Learning

Our second set of experiments deals with the effect of reweighting the energy terms in ROSETTA. We compared the default ROSETTA weights to those obtained using supervised learning by two learning methods: (1) the standard CRF framework – when all angles are considered observed; and (2) the Hidden CRF framework – when angles χ_3, χ_4 are considered hidden. Note that our database includes ground truth for all angles based on crystallography, but we hypothesized that due to the large variability in the angles far from the backbone, ignoring the crystallographic "ground truth" might enable better performance on the first two angles. As mentioned earlier, we used a small subset of the proteins as a training set, and report here results for the *test set*—proteins that were not seen by the learning algorithm.

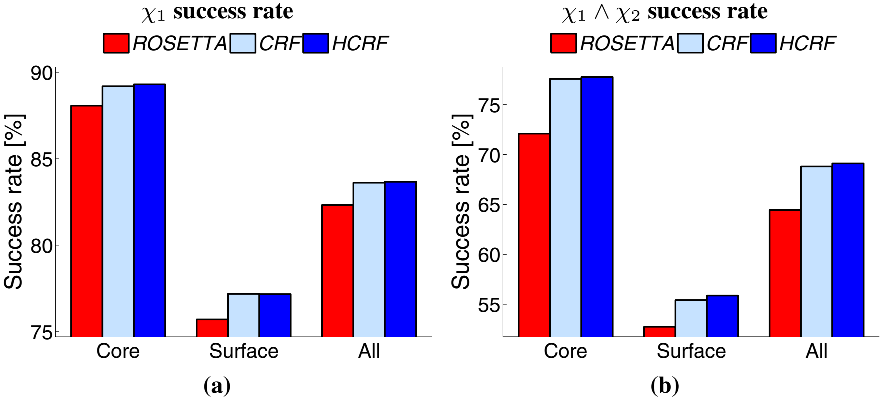


Fig. 4. The success rates, obtained using CRF and HCRF learned weights compared to ROSETTA's weights. Learning gives a significant improvement in performance.

Figure 4 shows χ_1 and χ_2 success rates on the test set using ROSETTA's original weights and the weights learned by the CRF and the HCRF algorithms. Both learning algorithms improve over ROSETTA's weightings.

Note that the improvement obtained by reweighting the terms (either using CRFs or using Hidden CRFs) is far larger than that obtained by using a better minimizer. Whereas going from simulated annealing to global minimization yields less than 0.1% improvement for the first two angles in core residues, reweighting the energy terms increases performance by almost 6%.

Figure 5a shows the weights learned by CRF and HCRF compared to ROSETTA's weights. While the change in most weights is mild, the repulsive van der Waals weight almost vanishes. Note however that complete exclusion of the repulsive term from ROSETTA's default energy function significantly decreases the success rates. The reason for the significant reduction of the van der Waals repulsive term is its sensitivity to discretization. Native structures are well-packed, therefore modeling with near-to-native, discrete conformations (that is, using rotamers) can easily lead to clashes. Consequently, when optimizing an energy function that distinguishes near-native conformations from wrong conformations, the repulsive term will be down-weighted. While an energy function with low repulsive weight might be useful for selecting correct side-chain conformations from a discrete set of possible combinations, procedures that involve continuous minimization will be impeded by the missing term that contributes significantly in guiding the structure towards the correct conformation.

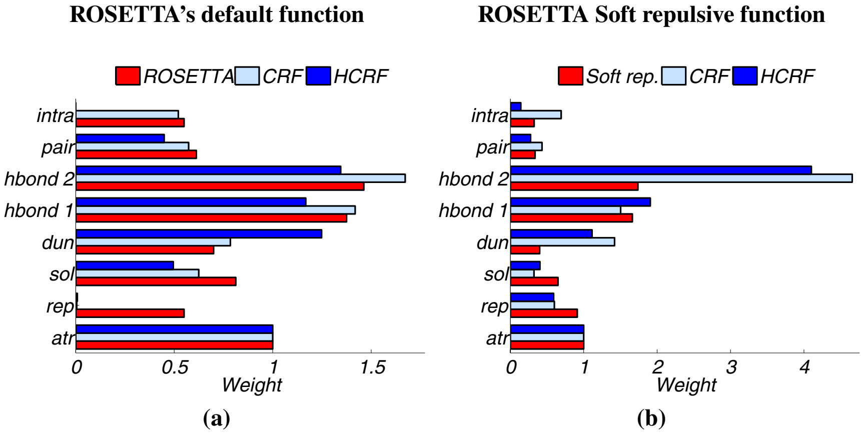
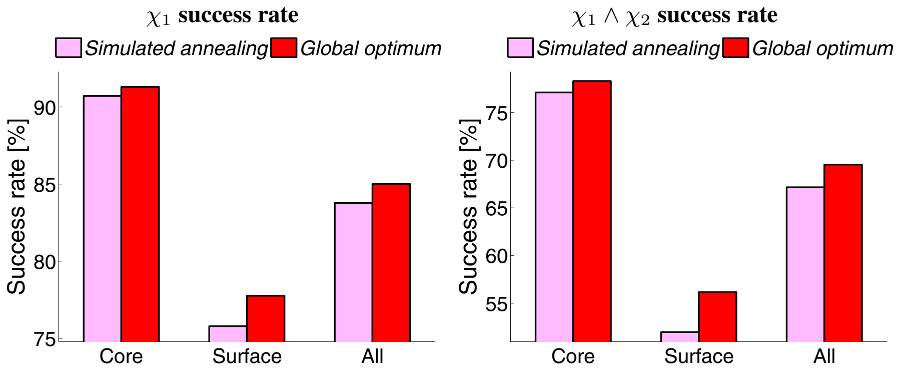


Fig. 5. ROSETTA's weights (a) and ROSETTA soft repulsive weights (b) compared to CRF and HCRF learned weights. Weights are normalized so that attractive weights equal 1.

When we analyzed the performance for different amino acids, we found that the greatest improvements were obtained on aromatic amino acids – Phenylalanine (F), Tyrosine (Y), Tryptophan (W) and Histidine (H). These bulky aromatic rings tend to clash if no extra rotamers are included in the rotamer library [29]. Since the repulsive contribution to the energy function is significantly reduced as a consequence of the low repulsive weight (Figure 5a), the selection of near-native conformations that clash with the surrounding environment – but still create favorable contacts that contribute to other terms in the energy function – is improved.

5 Results with "Soft Repulsion"

The fact that better performance can be obtained by decreasing the weight of the repulsive term has been observed previously in ROSETTA (e.g. [30]). In order to overcome this unnaturally small contribution of the repulsive part of the Lennard-Jones potential, a "dampened" version has been developed (the "-soft_rep" option, referred to as *DampRep* in [8]). In this function, the repulsive energy increases less dramatically when two atoms are brought together, and therefore, clashes are penalized less in the course of discrete optimization. This energy function was shown to allow improved side-chain modeling in ROSETTA [8]. In order to evaluate the importance of the search strategy and the energy function optimization, we conducted additional experiments based on this energy function.



(a) **(b)**

Fig. 6. The location of the global minimum energy conformation improves side-chain modeling for the ROSETTA soft repulsive energy function. Legend as in Figure 2.

We again found that TRBP can obtain the global optimum in a few minutes of computation for the majority of the proteins in our database (approximately 80%) while the LP relaxation and DEE could not. Figure 6 again shows that using the global optimizer leads to only a small improvement in prediction accuracy (approximately 1% improvement for the first two χ angles of core residues). Consistent with our earlier experiments, the gain from using a different energy function is larger than that using better minimizers – note that using simulated annealing with the "soft repulsion" energy gives better results than global optimization of the default ROSETTA function.

Figure 7 shows the results of applying reweighting to ROSETTA with soft repulsion. Even though this energy function had been optimized for side-chain modeling, supervised learning is able to find better reweighting of the energy terms. In particular, the new weights allow an improvement of correct modeling of $\chi 1$ and $\chi 2$ angles from 78% to 80%. Note that our test set contains approximately 32,000 residues for which both χ_1 and χ_2 are defined, so that a 2% improvement corresponds to approximately 640 residues and is highly significant. For this data set the HCRF learning criterion performed slightly better than the CRF criterion.

Figure 5b confirms that indeed, in the soft repulsive function the Lennard-Jones repulsive term is of comparable size to the Lennard-Jones attractive term. Interestingly, the contribution of hydrogen bonds is significantly increased.

Using a large rotamer library. A bottleneck in further improvement of side-chain modeling is the rotamer library from which side-chain conformations are selected. Side-chains that are not adequately represented in the library, cannot be correctly modeled. Therefore, in addition to energy functions and search methods, another direction of possible improvement is to modify the discrete set of rotamers that define the search space

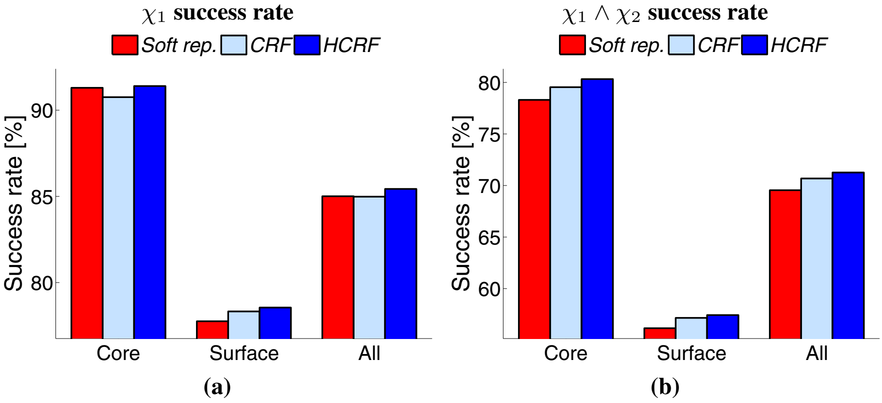


Fig. 7. The success rates, obtained using CRF and HCRF learned weights compared to ROSETTA's soft repulsive weights

[31, 32]. We therefore repeated our experiments using extra rotamers to core residues, for which accurate modeling is especially important to guarantee tight packing.

The much larger number of rotamers makes minimization much slower; TRBP can still obtain the global optimum for 70%-90% of the proteins in our data set, whereas the commercial Integer Programming package (CPLEX) fails to find the optimum for many proteins due to memory limitations (even after pruning the search space using DEE).

Indeed, extended sampling improves the performance of ROSETTA's soft repulsive energy function by more than 1% even when simulated annealing is used (81.4% for $\chi_1 \wedge \chi_2$ in core positions). Using the global minimum energy configurations when available (and the configurations obtained by simulated annealing otherwise) only slightly improved accuracy (less than 0.25%). In this case, using CRF learned weights leads to only a small improvement (0.35%, to 81.9%). The highest accuracy (82.6%) is obtained with weights learned using a local HCRF variant, in which we maximize the sum of the marginal log likelihoods of the native rotamers (and treat all other positions as hidden). For speed reasons we used ordinary BP in this variant.

6 Discussion

Side-chain prediction is an important subtask of the protein folding problem and has multiple applications in linking protein structure and function. Traditionally, it has been approached by formulating energy functions over a discrete set of angles and using discrete optimization algorithms to find the minimal energy configuration. Despite much progress in search methods and energy functions, performance is far from satisfactory and it has been difficult to systematically determine whether the energy functions or the search methods are to blame. In this paper, we have shown that using tree-reweighted belief propagation (TRBP) it is possible to find the global minimum for many side-chain prediction problems in a few minutes. TRBP can also be used to bound the partition function and this is useful for learning new weights in the CRF framework. Using these computational tools we have shown that (1) global optimization tends to yield a smaller improvement in performance than adapting the energy function, and that (2) supervised learning can be used to automatically reweight the energy terms to obtain relatively large improvement in side-chain modeling. By combining our learned weights with global optimization we obtain significantly better performance on test data compared to the ROSETTA package, widely considered the state-of-the-art.

The present study suggests that supervised learning can also be used to devise novel energy terms, in addition to reweighting the existing ones. In addition, we plan to learn task-specific weights in a more general setting; for example, by focusing on interface modeling in protein-protein interactions (e.g. *docking*). We believe that the tools of approximate inference and machine learning will have great benefit in many applications of structural biology.

Acknowledgments. We are thankful for insightful discussions with Brian Kuhlman and Jim Havranek. C. Y. is supported by Yeshaya Horowitz Association through the Center for Complexity Science. Support from the Israeli Science Foundation is gratefully acknowledged.

References

1. Canutescu, A.A., Shelenkov, A.A., Dunbrack, Jr., R.L.: A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci **12**(9) (2003) 2001–2014
2. Kuhlman, B., Baker, D.: Native protein sequences are close to optimal for their structures. PNAS **97**(19) (2000) 10383–10388
3. Fraenkel, A.S.: Protein folding, spin glass and computational complexity. In: Proceedings of the 3rd DIMACS Workshop on DNA Based Computers, held at the University of Pennsylvania, June 23 – 25, 1997. (1997) 175–191
4. Desmet, J., Maeyer, M.D., Hazes, B., Lasters, I.: The dead-end elmination theorem and its use in protein side-chain positioning. Nature **356** (1992) 539–542
5. Goldstein, R.F.: Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys. J. **66**(5) (1994) 1335–1340
6. Pierce, N.A., Spriet, J.A., Desmet, J., Mayo, S.L.: Conformational splitting: A more powerful criterion for dead-end elimination. J. of Computational Chemistry **21**(11) (2000) 999–1009
7. Kingsford, C.L., Chazelle, B., Singh, M.: Solving and analyzing side-chain positioning problems using linear and integer programming. Bioinformatics **21**(7) (2005) 1028–1039
8. Dantas, G., Corrent, C., Reichow, S.L., Havranek, J.J., Eletr, Z.M., Isern, N.G., Kuhlman, B., Varani, G.Merritt, E.A., Baker, D.: High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design'. Journal of Molecular Biology **In Press** (2007)
9. Dunbrack, Jr., R.L., Karplus, M.: Back-bone dependent rotamer library for proteins: Application to side-chain predicrtion. J. Mol. Biol **230**(2) (1993) 543–574
10. Rohl, C.A., Strauss, C.E.M., Chivian, D., Baker, D.: Modeling structurally variable regions in homologous proteins with Rosetta. Proteins: Structure, Function, and Bioinformatics **55**(3) (2004) 656–677
11. Lazaridis, T., Karplus, M.: Effective energy function for proteins in solution. Proteins: Structure, Function, and Genetics **35**(2) (1999) 133–152
12. Kortemme, T., Morozov, A.V., Baker, D.: An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. Journal of Molecular Biology **326**(4) (2003) 1239–1259

13. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML, 2001. (2001) 282–289
14. Lafferty, J., Zhu, X., Liu, Y.: Kernel conditional random fields: Representation and clique selection. In: ICML. (2004)
15. LeCun, Y., Huang, F.: Loss functions for discriminative training of energy-based models. In: Proc. of the 10-th International Workshop on Artificial Intelligence and Statistics (AIStats'05). (2005)
16. Vishwanathan, S., Schraudolph, N., Schmidt, M., Murphy, K.: Accelerated training of conditional random fields with stochastic meta-descent. In: ICML. (2006)
17. Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C.: Hidden conditional random fields for phone classification. In: INTERSPEECH. (2005)
18. Quattoni, A., Collins, M., Darrell, T.: Conditional random fields for object recognition. In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA (2005)
19. Taskar, B., Guestrin, C., Koller, D.: Max-margin markov networks. In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA (2004)
20. Wainwright, M.J., Jaakkola, T., Willsky, A.S.: MAP estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. IEEE Transactions on Information Theory **51**(11) (2005) 3697–3717
21. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Understanding belief propagation and its generalizations. In: IJCAI (distinguished lecture track). (2001)
22. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
23. Yanover, C., Weiss, Y.: Approximate inference and protein folding. Advances in Neural Information Processing Systems (2002)
24. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. In: Proceedings AI Stats. (2005)
25. Kolmogorov, V., Wainwright, M.: On the optimality of tree-reweighted max-product message passing. In: Uncertainty in Artificial Intelligence (UAI). (2005)
26. Meltzer, T., Yanover, C., Weiss, Y.: Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In: Proceedings International Conference on Computer Vision (ICCV). (2005)
27. Yanover, C., Meltzer, T., Weiss, Y.: Linear programming relaxations and belief propagation – an empirical study. Journal of Machine Learning Research **7** (2006) 1887–1907
28. Liu, Y., Kuhlman, B.: Rosettadesign server for protein design. NAR **34** (2006) W235–238
29. Wang, C., Schueler-Furman, O., Baker, D.: Improved side-chain modeling for protein-protein docking. Protein Sci **14**(5) (2005) 1328–1339
30. Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., Baker, D.: Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of Molecular Biology **331**(1) (2003) 281–299
31. Leaver-Fay, A., Kuhlman, B., Snoeyink, J.: An adaptive dynamic programming algorithm for the side chain placement problem. Pacific Symposium on Biocomputing **10** (2005) 16–27
32. Peterson, R.W., Dutton, P.L., Wand, A.J.: Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci **13**(3) (2004) 735–751