Atomically detailed potentials to recognize native and approximate protein structures

Jian Qiu and Ron Elber*

Department of Computer Science

Cornell University

4130 Upson Hall

Ithaca NY 14853

*Corresponding author, e-mail ron@cs.cornell.edu, fax 607-255-4428

keywords: threading potentials, linear programming, decoy structures, protein structure prediction

Abstract

Atomically detailed potentials for recognition of protein folds are presented. The potentials consist of pair interactions between atoms. One or three distance steps are used to describe the range of interactions between a pair. Training is carried out with the mathematical programming approach on the decoy sets of Baker, of Levitt, and of our own design. Recognition is required not only for decoy-native structural pairs but also for pairs of decoy and homologous structures. Performance is tested on the targets of CASP5 using templates from the protein databank, on two test ab-initio decoy sets from the Skolnick's laboratory, and on decoy sets from the Moult's laboratory. It is concluded that the newly derived potentials have significant recognition capacity that is comparable to the best models derived by other techniques using a significantly smaller number of parameters. The enhanced recognition capacity extends primarily to the identification of structures generated by ab-initio simulation and less to the recognition of approximate shapes created by homology.

I. Introduction

One of the outstanding problems in the design of folding and threading potentials is the recognition of *approximate* structures. Fold recognition and ab-initio folding do not give exact models but only approximate protein structures. Nevertheless most efforts of potential design are based on the recognition of (exact) native shapes [1-13] and are not necessarily addressing the need for approximate recognition. In studies of exact structures the goal is to make the energy of the correct structure lower than the energy of other (wrong) structures. This condition can be written as a set of inequalities

$$E\left(S_n, X_i; P\right) - E\left(S_n, X_n; P\right) > 0 \qquad \forall i, n \tag{1}$$

The energy is $E\left(S, X; P\right)$, S_n and X_n are the sequence and the structure of the native protein, and X_i is the structure of a wrong structure (decoy). The elements of the vector P are the parameters of the potential that we wish to determine given the inequality (1). The inequality in equation (1) is repeated for all available decoys $\left(i\right)$ and native structures $\left(n\right)$. The number of inequalities that we attempt to solve is typically in the hundreds of thousands to millions while the number of parameters is less than ten thousands. We [11, 12, 14-17] and others [2, 7, 18, 19] have used equation (1) to derive competitive folding potentials for functions that depend linearly on their parameters. These energy functions were no worse than potentials derived from optimization of T_f / T_g [20], Z score [7], the σ parameter [21], or from direct counting of contacts (the so called statistical potentials) [3, 6, 9, 10, 22-24].

Nevertheless, a significant conceptual flaw in the straightforward application of (1) is the neglect of approximate structures. The designed function is guaranteed to have a low energy value only at one point in configuration space– the native structure. This may be insufficient, as can be exemplified by a "golf course" potential model. Hits that are very close to the native structure (or to the holes in a golf course) still count as a miss. Without the ability of recognizing approximate structures sound models remain undetected. This recognition is of particular importance in modeling applications, since in practice the exact structures are not available.

It is therefore possible that a highly refined potential with excellent recognition capacity of native folds will perform poorly when only approximate folds are at hand. An ideal potential will have a gradual "sensitivity" so that the energy decreases smoothly when the structure is more similar to the native shape. This statement is another reflection of the funnel picture [25] in which the potential is globally attractive (on the average) towards the native structure. Coordinates of choice to measure the similarity of a given structure to the native fold are (for example) the RMS (Root Mean Square distance) or Q (the fraction of the native contacts). Each of the measures has its problems in detecting similarities, and it is far from obvious that it is monotonic as a function of the folding pathway. Monotonic behavior is required in addition to the learning of equation (1), making the direct application of equation (1) difficult for the design of "ideal" funnels.

The idea of learning also approximate structures in addition to the native shapes in order to have a more global view of the energy surface was put forward by Zhang, Kolinski and Skolnick in their Touchstone II program [26]. In their approach they optimize the correlation between the RMS and the energy and maximize the average energy gap,

between the native and the set of misfolded structures. The main difference between the earlier study and the current investigation is in the treatment of the energy gap. Previously the average energy gap (averaged over all or a subset of decoys) was optimized. Here an attempt is made to make every single energy difference between native and decoy structures (or approximate and decoy structures) positive. Introducing global attractiveness to a potential is difficult and a more modest goal is in place. Instead of a global view, the focus here is on the neighborhood of the native structure. The extension to equation (1) that we propose considers native-like shapes in addition to the native conformation. The native-like shapes are required to have energy values lower than the energies of decoy structures. The definition of "native-like" or "similar" structures used in this paper is given in the context of practical modeling applications, and is described in the next section. However, we emphasize that this definition is not unique, and we therefore anticipate that extensions and further exploration of the ideas presented in this paper will follow in future works. For example we do not impose constraints differentiating between approximate folds and native structures, putting all reasonable folds in a single basket of indistinguishable objects. A refinement of the present approach, ordering approximate structures according to their proximity to the native shape, should be possible by setting corresponding inequalities (e.g. $E\left(S_n, X_{\text{homolog}}; P\right) - E\left(S_n, X_{\text{native}}; P\right) > 0$).

Traditionally, potentials parameterized on the basis of known protein structures are coarse grained (in contrast to parameters fitted to ab-initio calculations or to data of small molecules). Most potential energy functions were developed on the residue level in which an amino acid is described by one or two points. The low resolution has the benefit of

potentially recognizing approximate, nearby structures that fall into the same class of coarse-grained shapes. Hence, limited structural information can be beneficial if approximate shapes are compared and scored.

Nevertheless, other (coarse grained) descriptors of protein structures are likely to help. For example, protein side chains have specific shapes that cannot be captured by the two-point representation, and are influencing sequence to structure fitness. One approach to describe the side chain contributions is the use of orientation dependent residue-residue (ODRR) potential [9]. Another approach for more detailed description of side chains and backbone is to use point representation of the heavy atoms in the side chains. Atomic Potentials (AP) can (in principle) capture the same features (and more) as an ODRR while using a number of parameters that is not substantially larger. Interpretation and analysis of AP-s are somewhat easier than ODRR due to the significant number of AP potentials available that are based on chemical physics principles. A potential disadvantage of AP is the smaller range of interactions (sizes) of the atoms compared to the amino acids. The averaging and coarse graining, desirable in approximate recognition, are therefore quite limited in AP. The use of atoms can produce an overly sensitive, golf course like potential, and the feasibility of a successful and simple AP for practical applications is one of the questions that we wish to explore here.

A number of AP potentials were designed in the past [3, 27-29]. They were designed based on statistical analysis of contacts, which is different from the mathematical programming approach that we use here. In addition to the alternative approach to learn approximate folds, this paper presents a comparison between information-based atomic-

potentials that are derived by either mathematical programming or by Bayesian statistical approaches. The quality of the potentials derived here seems comparable to another highly successful potential [3].

II. Methods

II.1 The training sets

The training set includes learning of (exact) native folds (equation (1)) as well as learning of approximate (homologous) structures to be defined below. We denote an approximate structure by X_{app} and in the spirit of equation (1) we require

$$E\left(S_n, X_i; P\right) - E\left(S_n, X_{app}; P\right) > 0 \qquad \forall i \tag{2}$$

We attempt to satisfy this requirement by searching for a vector of parameters P that satisfies all the inequalities formulated in (2). We consider three training sets.

The first set of structures aims at exact recognition and include decoys from the groups of Levitt [6, 30] and of Baker [31]. We call this set of structures, the ER set (Exact recognition). From the Levitt's group Decoys R Us databases we consider the subsets 4state_reduced, fisa, fisa_casp3, lattice_ssfit, lmds, and semfold. From the Baker's set that includes 91 structures we have selected proteins that are longer than 50 amino acids and the decoy structures contain at least 70% of the native residues to provide a set with 64 native structures (and their associate decoys). These sets include atomically detailed structures that are used "as are" in contrast to the learning sets that follow. The total number of inequalities generated from the ER set was 172,215. All of these inequalities are of the type formulated in equation (1). These sets are often used to test different models, however, since we have an alternative and very extensive test that was built in

the Skolnick group we prefer to use the above collections of decoy structures to train our potential.

The second set is called BA (Best Alignment) and is based on structural alignment of homologous (and decoy) proteins. Structural alignment is usually the best we can hope for once a template was identified (unfortunately this alignment is not available for structure prediction and can be used only for learning). The BA set consists of 642 native proteins. For each protein we assign a few homologous templates and up to two thousand decoy structures, all are selected from the protein databank. The complete BA set can be found on the web http://www.cs.cornell.edu/~jianq/research.htm. A template $\left(X_{app}\right)$ is chosen if the Z score of the structural alignment program CE [32], (comparing the native and the template shapes), is larger than 4.5 (high structural similarity). A decoy structure is so defined if the CE Z score is at most 3.5. Inequalities are then generated for both exact and approximate matches (following equation (1) and (2)). For the native structure we have (of course) the position of all the heavy atoms. However, we do not have all the heavy atom coordinates of the approximate models or the decoys. The MODELLER program [33, 34] is used to generate the coordinates based (in the present set) on the CE alignments and the coordinates of the templates (homolog or decoy structures). The use of a specific modeling program may affect our training of approximate structures. However, by examining atomic structures produced with an alternative approach (see Results section) we are able to check the sensitivity of our design to the way decoys and approximate structures were constructed.

Some of the structural alignments (especially against decoy structures) are very short and are unlikely to score highly and provide useful input to the learning process. We therefore attempted to build an atomic model only if the Z score of the structural alignment was at least 1.0, and the minimum length of an alignment was ninety residues. Another condition is that the alignment length must be greater than 70 percent of the length of at least one protein that participates in the alignment. Moreover, since the typical lengths of decoy alignments are shorter than alignments to approximate structures we compare the energy of only the residues that they share. Decoy and approximate structures that share less than 50 amino acids do not generate a corresponding inequality. The total number of inequalities that were generated, based on the structural alignment procedure described above, was 739,587.

The last training set is called the SA (Sequence Alignment) set. It employs approximate and decoy structures similar in spirit to the BA set. However, two versions of LOOPP alignments [35] are used as input to MODELLER instead of structural alignments. The two alternative LOOPP alignments are based either on a local sequence alignment with the BLOSUM50 substitution matrix [36] (alignment I), or on a substitution matrix derived from structural alignments (alignment II) (Jian Qiu, submitted). The lack of direct structural alignment makes the SA set more difficult (and more related to real life applications) compared to the BA set. The SA set contains 803 proteins and their corresponding decoys (each protein has at most 50 decoys) that were modeled with the MODELLER program (The data set generated from the first version of alignment contains 633 proteins, and that from the second version has 660 proteins.). Complete files of the structures can be found in http://www.cs.cornell.edu/~jianq/research.htm. In this

set a structure is considered homologous if at least 60 percent of the structure has an RMS distance of 6Å (or less) to the native shape. A structure is considered a decoy if the RMS to the native structure is greater than 8Å, and at most 30 percent of its structure are within 6Å from the native shape. The generation of inequalities is done in a similar way to the BA set. The total number of inequalities generated here is 148,745. The number of inequalities generated with the SA protocol is significantly smaller than the number of inequalities of the BA procedure due to the difficulties in obtaining meaningful decoy structures in the SA protocol.

The total number of constraints (inequalities) generated in the ER, BA and SA sets is 1,059,547. This set of inequalities was used to train the four different potentials discussed below.

II.2 Functional forms of the potentials

We consider four atomic potentials. The first two are comparable in complexity to residue-based contact-potentials. The other two potentials are more complex and explore the impact of distance dependent potential. While we anticipated some interplay between the two approaches (contact potentials are more coarse grained and may have better recognition of approximate structures) the results suggest that the more detailed approach is more effective. The last two potentials demonstrate higher recognition capacity for exact matches and comparable performance for approximate matches. Nevertheless, we shall demonstrate on an independent set that the most detailed potential is not the best and some subtle over learning is evident.

We use the heavy atom definitions provided by the extended atom model of the OPLS force field [37] as implemented in the moil program [38]. Different protonation states did not change the atom type, and bonded and non-bonded cysteine residues (through the sulfur bridge) were modeled by a single set of parameters. This analysis provides 46 types of atoms. Collapsing a few similar atoms into one type provided an alternative and smaller set of 32 types. A complete list of the atom types used in the present study is provided in tables I.

*** PLACE TABLES I HERE ***

In addition to different number of atom types we also consider two distance-dependent formulations. In the first implementation, a square well potential, different from zero between 3.5 and 6.5Å, is used. In the second formulation, three steps are defined in the range of 2.0-3.5Å, 3.5-5.0Å, and 5.0-6.5Å.

Combining the variations in atom types and in the number of steps we trained four potentials (using the inequalities described above). The first is called T32S1, the second T46S1, the third T32S3, and the fourth T46S3. The "T" stands for Type of atoms and the number that follows is the total number of atom types of this model. The "S" is for the number of steps, which is followed by the number of steps that are used. Note that the potentials as described above depend linearly on their parameters. It is convenient to think on the distance (and steps) as just an extension of the atom types. For example, the energy of the T32S3 model can be written as

$$\begin{aligned} E_{T32S3}(S,X;P) &= \sum_{i>j} u\left(atom\_type_i, atom\_type_j, d_{ij}\right) \qquad i,j=1,...,N \\ &= \sum_{\alpha} n_\alpha p_\alpha \qquad \alpha=1,...,1584=3\times 32\times\left(32+1\right)/2 \end{aligned} \tag{3}$$

The sum in the first line is over the indices of the atoms where N is the total number of atoms, and u is a table that assigns an energy value to a given contact type. The "contact type" is determined by the types of the two atoms and by their distance. An equivalent expression is written more compactly in the second line where the index α is running over contact types. The vector element, n_α, provides the number of times that a contact type α occurs in the structure, and the vector element, p_α, is the energy value associated with a contact type α. From the second line it is obvious that the energy is linear with the parameters -- p_α, and equations (1) and (2) yield linear inequalities. For example equation (1) is now written

$$\sum_\alpha p_\alpha \left(n_\alpha^i - n_\alpha^n \right) > 0 \qquad \forall i \tag{4}$$

The number of contacts of type α in decoy structure i is denoted by n_α^i; while the number of contacts of the same type α in the native structure is denoted by n_α^n.

The numbers of parameters that we need to determine for the T32S1, T46S1, T32S3, and T46S3 potentials are 528, 1081, 1584, and 3243 respectively. These numbers are about three orders of magnitude smaller than the number of inequalities that we generated, suggesting that if our training protocol succeeds, significant over learning of the data is unlikely. Nevertheless, we will show later, on independent training set that the potential T32S3 is the most accurate, adding more parameters to create the model T46S3 does not help recognition.

The increase in the number of parameters used in the atomic potentials compared to the number of parameters in residue-based contact potentials (210 parameters) is nevertheless

significant. To justify the new formulation we expect the atomic potential to significantly enhance recognition. Note also that the number of parameters that we use is smaller than the number that was used in another atomic potential (167 atom types and 3 distance steps, providing a total of 42,084 parameters [3]) to which we compare out results.

II.3 Calculation of the parameter sets

We used the interior point algorithm, implemented in the two programs BPMPD [39] and PF3 [12], to determine the potential parameters for the four potentials considered above. The task is expected to be easier for a smaller number of parameters. This is however not the case since all the potential forms that we have examined led to infeasible sets of inequalities for the training set described in section II.1 (i.e. that there is not a vector of parameters such that all the inequalities are satisfied). The more pronounced is the infeasibility, the more difficult is the determination of an optimal parameter set. In that sense a smaller set of parameters does not imply easier calculations since it may be infeasible. Since our learning includes approximate structures, it is not obvious that a physical energy exists such that the approximate structures are better than all decoys. Therefore infeasibilities do not necessarily point into deficiencies in the model potential (though they do not exclude this possibility). By forcing the algorithm to learn approximate shapes of our design, which may be inconsistent with physical energies, we are not restricted to physically based approaches.

The optimal parameters were determined using an approach related to the Maximum Feasibility Principle (MFP) [16]. We start the calculations with the program PF3 [12].

PF3 is a mathematical programming code that solves the set of inequalities using the interior point algorithm [40]. A unique feature of PF3 is efficient parallelization on a massively parallel computer, making it possible to distribute the inequalities between many computing nodes and to study a very large set of inequalities. PF3 is especially tuned to the design of folding potentials taking into account the large asymmetry between the number of inequalities (millions) and the number of the parameters (a few thousands). For example, PF3 relies on the solution of the dual problem. In the dual problem the dimension of the squared matrix, which is used in the optimization, is the number of parameters. This is in contrast to the prime formulation with a matrix dimension equal to the number of constraints. Finally (and in contrast to BPMPD, which is another program we use extensively) PF3 has a built in heuristic to estimate a solution for infeasible sets (by adding slack variables [40] and minimizing the number of slack variables that are not zero). While the user can add slack variables manually, the PF3 formulation is especially economic since the PF3 slack variables do not change the matrix dimensionality (in the dual formulation) and therefore do not add much to the required computer memory. Reduction in memory requirement is crucial for the large-scale optimization problem at hand. With all the advantages of PF3 mentioned above it is useful to note an advantage of BPMPD compared to the current version of PF3. BPMPD allows quadratic programming and analytical centering while PF3 does not. The analytical centering is important in the second phase of parameter estimation (see below).

Once an approximate solution is obtained from PF3 (for an infeasible set) a refinement step is made. Only inequalities that are satisfied with the recently computed parameter set

are considered for the second round. These inequalities are chosen from those that are as close as possible to the boundaries of infeasibility (the constraint values are close to zero), and help to determine the feasibility volume. To select a concrete set of parameters for the potential we choose a point in the feasible volume (of the subset of the inequalities) by centering. A point at the center is as far as possible from being infeasible and is likely to provide a good potential for future proteins. Using the interior point algorithm this point is estimated as the analytical center of the feasible volume [16].

The BPMPD program determines the center of the feasible volume starting from any feasible point. This option is not available in PF3, making the BPMPD program the code of choice for the refinement part. Note that in the PF3 calculations all the constraints (inequalities) were considered simultaneously, while in the second phase of parameter refinement, only a fraction of the total number of constraints (typically 60,000 to 160,000) was used due to the limited memory available for the BPMPD serial program.

III. Results

III.1 The potentials

To demonstrate the enhancement in potential capacity after the refinement procedure, we consider the T32S3 potential. Before the refinement of the parameter set with the BPMPD program 7,231 inequalities were not satisfied. The unsatisfied inequalities were distributed 159, 2749, and 4323 in the ER, BA, and SA sets. After the refinement procedure which centers the parameter vector 6,544 inequalities were not satisfied, distributed this time 113, 2461, and 3970, in the ER, BA and SA sets respectively. While

the improvement is not earthshaking it is consistent in all training sets. Note also the larger number of unsatisfied inequalities in the SA set (especially compared to the number of inequalities that each set contributed to the total).

Not surprisingly (due to the use of a larger number of parameters) the performance of the potentials T32S1, T46S1 and T32S3 increases monotonically on the training set (the number of inequalities solved are 1046352, 1049950, and 1053003 with 528, 1081, and 1584 parameters respectively). The increase in the number of solved inequalities is however small and raise questions regarding the generality of the computed scores. Therefore tests on independent sets, which are more meaningful, must be performed and will be described next.

The parameters of the T32S1 potential are given below. The complete set of parameters for the four potentials is available from the web

http://www.cs.cornell.edu/~jianq/research.htm

*** PLACE TABLE II HERE ***

It is of interest to examine the properties of the calculated potential and to check if the learning procedure found a set of parameters that agree with our biophysical intuition. Consider the potential T46S1. We plot the interaction energy of the atom types CFH, and OX1 with the rest of atoms in figure 1. CFH is found in hydrophobic residues -- phenylalanine, tyrosine, and tryptophan, and the polar atom OX1 is found in aspartic acid. The two of them therefore provide a useful contrast. It is interesting that no clear

pattern is observed indicating that the simple polar/apolar picture for individual atoms is inappropriate for the present model potentials. This observation is surprising since a similar optimization protocol [11,12,14,15] was able to recover hydrophobic relationship between residues. The failure of the same approach to recover hydrophobic relationships between atoms may suggest that the atoms with their much smaller size are inappropriate to describe hydrophobic interactions in a pairwise model.

III.1 Tests of the model potentials

We consider several tests of the potential models. In the first test we use the sequences from the CASP5 exercise. For each target, the program LOOPP [35] picks the fifty top scoring templates. After refinement (excluding models that are too short, or do not satisfy other criteria mentioned earlier) we are left with 22 to 50 candidate structures per protein, depending on the protein under consideration. The procedure to generate the structural models is similar to the model generation for the SA set, and it follows the same protocol we regularly use to predict protein structures [35]. The set is available from http://www.cs.cornell.edu/~jianq/research.htm. We (of course) made sure that our template database and the training set did not include the CASP5 targets. We also tested that targets have at least one nearby structural template in our database. For alignment type I we have constructed models for 30 targets (in some cases, even with correct template we were not able to obtain a model within 6Å of the native). For alignment II only 15 targets were used.

The MODELLER program [33, 34] was used to construct atomic models from templates. The program LOOPP developed in our laboratory identified the templates and aligned the probe sequence into their structure. When the overall RMS of the model is less than 6Å with respect to the native structure, we consider it a success. In table III we summarize the results of the prediction. For comparison we are also providing the ranking computed by alternative potentials including: (a) residue contact potentials (Miyazawa and Jernigan potential [1], and Hinds Levitt [41]), (b) distance dependent residue-residue potential (Tobi and Elber [11]), and atomically detailed potential (Lu and Skolnick [3]).

** PLACE TABLE III HERE **

In figure 2.a we show an overlap of the target protein 1M2E_A with a correctly predicted model that also has a low T32S3 energy, and in figure 2.b a worse model (but still acceptable with RMSD 6.9) that has a high T32S3 energy. In figure 2.c we overlap a wrong prediction that has a low T32S3 energy with the structure of the target. Note that the number of correct predictions using T32S3 is not significantly higher than the number of correct predictions using T32S1 in the CASP5 test.

The above test emphasizes approximate recognition. In the next test we consider the recognition of exact structures employing considerably more extensive sets of decoy structures. We consider two native/decoy sets generated in the group of Jeff Skolnick [42]. The first set we received includes only 17 proteins with 1,000 decoy structures for each of the proteins

PLACE TABLE IV HERE

The second test set of 125 proteins (each with 24,000 decoy structures) was received more recently. After eliminating proteins included in our training set we were left with 74 proteins with 24,000 decoy structures for each. The lattice structures recorded in both sets includes only C_α and side chain center coordinates. We therefore generated atomically detailed models based on this reduced representation with the program phoenix2 (Jaroslaw Pillardy, private communication). Phoenix2 uses the (available) C_α and side-chain center coordinates to create side chain positions. The models are then scored using the four potentials we have developed, and the alternative potentials mentioned earlier. A summary of these scores is given in table V

*** PLACE TABLE V HERE ***

On the two sets quoted above it is clear that the residue based potentials are doing considerably worse. Moreover, it is also clear that the multi-step atomic potentials are better than the single step potentials. This enhanced performance on independently constructed sets of decoy (by a different group and different technology) is reassuring.

Another measure of the potential quality and its global attraction is the dependence of the energy on the proximity to the native state. Our design protocol (by including approximate structures) attempted to make the energy more globally attractive. What is the proper coordinate to measure proximity to the native structure is not obvious, however in numerous cases the RMS is used. In figure 3 we show a scattered plot of the energy as a function of the decoy RMS value. We are showing two cases of exact recognition and one case in which the native structure was not recognized as one of the

top 10 hits. These examples were taken from the 17 set of Skolnick. Perhaps the most striking feature of this plot is the relatively poor correlation of the energy with the RMS. This suggests that either the RMS is a bad measure to the native shape, or that the potentials we have designed are not as broad in the RMS space as we have hoped. A similar plot was obtained also for the percent of native contacts – Q.

In order to compare our potential to another atomically detailed formulation, we also considered the potential RAPDF from the Moult's laboratory [29]. In the RAPDF reference, decoy test sets at the PROSTAR website of the Moult's laboratory [43] are mentioned and we considered two of them (1) MISFOLD [44], and (2) IFU [45]. The MISFOLD test is not difficult and all our potentials identified 100% of the correct fold (as did RAPDF). The IFU is more challenging and RAPDF predicted correctly 73% of the structures. Our potentials score as follows T32S1 64%, T46S1 77%, T32S3 80%, T46S3 73%, underlining our previous observation that T32S3 is the best of our designed potentials. The potential T32S3 performs better than RAPDF on a test of Moult's group design.

III.2 Potential stability

Design of potentials means the determination of a large number of parameters according to even a larger number of data points (number of decoys). Even using the same training set and optimization algorithm it is not obvious that identical potential (or even a potential with the same performance) will be obtained. To explore the convergence of our optimization algorithm we re-examine the performance of two T46S1 potentials learned

on the same training set and tested on Skolnick_17 set. One of these potentials was discussed earlier. The second potential was trained using T32S1 potential to generate an initial guess for the training. For atoms that were identical in the 46 and 32 parameter sets we simple copied the parameters from the 32 set as an initial guess for the 46 set. Atom types that were split into more than one type when moving from 32 to 46 were given an exact and identical value (of the 32 set) for all the types of the 46 set. For example, values associated with the atom type CH2 in the 32 set were copied to atoms CH2, CBH and CR1 in the 46 set. We emphasize that the parameter values of the initial guess are quite different from each other, as can be seen in http://www.cs.cornell.edu/~jianq/research.htm. With the T32S1 initial guess for the T46S1 potential we repeated the optimization of the parameters as described in the text. The new T46S1 potential (called T46S1_from_T32S1) was tested on the independent decoy set Skolnick_17. The results shown on figure 4 clearly indicate that the performances of the two potentials are very similar, suggesting that our optimization protocol indeed converges to potentials with similar recognition capacity.

IV. Summary statement

We have developed atomically detailed information-based potential that is showing significant improvement compared to residue-based potentials both in template recognition and in "fishing-out" native folds from alternative sets generated by ab-initio methods. The potential has comparable recognition capacity to the potential of Lu and Skolnick [3] on a variety of tests using significantly smaller number of parameters. In fact, when we tried to increase the number of atom types from 32 to 46 no significant

enhancement in recognition was obtained. This suggests that 32 atom types is an upper bound on the number of types that can be learned in the present functional form. The T32S3 performed better than RAPDF from the Moult's group [29] on a test designed by the same group [45]. We have shown that the mathematical programming approach is one of the most effective ways of learning threading and folding potentials, by combining learning of "individual" decoys (individual inequalities) with centering algorithms that make the parameter sets as significant and compact as possible. The complete potential parameters and supplementary material for the tests are available from http://www.cs.cornell.edu/~jianq/research.htm

From the perspective of performance we differentiate between homology modeling and ab-initio simulations. In our hands the residue-based potentials are having comparable performance to atomically detailed potentials on the homology tests. However, on structures generated from ab-initio studies the atomically detailed potentials perform considerably better. On the test sets that we examined the LS and the T32S3 are performing (on the average) the best.

VII. Acknowledgements

This research is supported by an NIH grant GM67823. We thank J. Pillardy and T. Galor-Naeh for their help in this project, and to H. Lu and J. Skolnick for providing us with their potential. We also thank M. Levitt, D. Baker, J. Moult, and J. Skolnick for making their decoy sets available.

Tables

Table I

A list is provided of 32 or 46 atom types that were used in the training of the potentials. The atom types are identical to the types used in the Molecular Dynamics program moil [38]. Note that hydrogen atoms (polar or non-polar) are not included in the training set. The table indicates the atom types that were merged when we contracted our set from 46 to 32 atom types.

Atom types:

A32	A46	Type definition
NX	NX	Lys-N^{ζ}
NH	NH	N(all amino acids)
CO	CO	C(all amino acids); Asn-C^{γ}; Gln-C^{δ}
OC	OC	O(all amino acids); Asn-$O^{\delta 1}$; Gln-$O^{\varepsilon 1}$
CAH	CAH	C^{α}(all amino acids, except Gly and Pro)
	CAG	Gly-C^{α}
CH3	CH3	Ala-C^{β}; Ile-$C^{\gamma 2}$; Leu-{$C^{\delta 1}$,$C^{\delta 2}$}; Thr-$C^{\gamma 2}$; Val-{$C^{\gamma 1}$, $C^{\gamma 2}$}
	CH3D	Ile-C^{δ}
CH2	CH2	Arg-C^{β}; Asn-C^{β}; Gln-{C^{β},C^{γ}}; Glu-C^{β}; His-C^{β}; Ile-$C^{\gamma 1}$; Leu-C^{β}; Lys-{ C^{β},C^{γ},C^{δ}}; Met-C^{β}; Phe-C^{β}; Pro-{C^{β},C^{γ}}; Trp-C^{β}; Tyr-C^{β}
	CBH	Ile-C^{β}; Leu-C^{γ}; Val-C^{β}
	CR1	Arg-C^{γ}
CFH	CG	Phe-C^{γ}; Tyr-C^{γ}
	CFH	Phe-{$C^{\delta 1}$,$C^{\delta 2}$,$C^{\varepsilon 1}$,$C^{\varepsilon 2}$,C^{ζ}}; Trp-{$C^{\varepsilon 3}$,$C^{\zeta 2}$,$C^{\zeta 3}$,$C^{\eta 2}$}; Tyr-{$C^{\delta 1}$,$C^{\delta 2}$,$C^{\varepsilon 1}$,$C^{\varepsilon 2}$}
CZ	CZ	Tyr-C^{ζ}
OH	OH	Ser-O^{γ}; Thr-$O^{\gamma 1}$; Tyr-O^{η}
CGTR	CGTR	Trp-C^{γ}
	CTR	Trp-$C^{\delta 2}$
CHTR	CHTR	Trp-$C^{\delta 1}$
	CGHT	Trp-$C^{\varepsilon 2}$
NDHS	NDHS	Trp-$N^{\varepsilon 1}$
CH2M	CH2M	Met-C^{γ}
SM	SM	Met-S^{δ}
CH3M	CH3M	Met-C^{ε}
CH2K	CH2K	Lys-C^{ε}
CH2S	CH2S	Ser-C^{β}
	CHT	Thr-C^{β}
CHPR	CH2D	Pro-C^{δ}
	CHPR	Pro-C^{α}
CH2C	CH2C	Cys-C^{β}
SH	SH	Cys-S^{γ}

CGHP	CGHP	His-C^{γ}
	CHDP	His-$C^{\delta 2}$
NDHP	NDHP	His-$N^{\delta 1}$
	NEHP	His-$N^{\varepsilon 2}$
CHEP	CHEP	His-$C^{\varepsilon 1}$
CR2	CR2	Arg-C^{δ}
NR1	NR1	Arg-N^{ε}
CR3	CR3	Arg-C^{ζ}
NR2	NR2	Arg-$\{N^{\eta 1},N^{\eta 2}\}$
NAS	NAS	Asn-$N^{\delta 2}$; Gln-$N^{\varepsilon 2}$
CH2B	CH2B	Asp-C^{β}
	CH2A	Glu-C^{γ}
CX1	CX1	Asp-C^{γ}
	CSX1	Glu-C^{δ}
OX1	OX1	Asp-$\{O^{\delta 1},O^{\delta 2}\}$
	OSX1	Glu-$\{O^{\varepsilon 1},O^{\varepsilon 2}\}$

Table II

The energy parameters for the atomically detailed potential T32S1. The potential includes 32 atom types described in table I, and only one distance step (between 3.5 and 6.5Å) is considered. For clarity the interaction values in this table are truncated two digits after the decimal point. The complete set of parameters for the four potentials designed in this work – T32S1, T46S1, T32S3, and T46S3 (with no truncation) are available from the web http://www.cs.cornell.edu/~jianq/research.htm)

	NX	NH	CO	OC	CAH	CH3	CH2	CFH	CZ	OH	CGTR	CHTR	NDHS	CH2M	SM	CH3M
NX	0.267	0.076	-0.24	0.027	-0.04	3E-04	-0.04	0.259	-0.34	-0.15	1.891	-0.34	-2.39	-0.91	-0.47	-1.04
NH	0.076	0.025	-0.24	0.168	0.079	0.011	-0.08	0.02	-0.09	0.054	0.142	-0.81	0.59	0.038	-0.26	-0.38
CO	-0.24	-0.24	-0.15	0.14	-0.18	0.151	-0	0.028	-0.07	0.102	-0.27	-0.02	-0.18	-0.03	0.491	0.472
OC	0.027	0.168	0.14	-0	0.12	0.012	0.047	-0.03	-0.33	0.031	-0.02	0.245	-0.36	0.004	-0.46	-0.3
CAH	-0.04	0.079	-0.18	0.12	-0.16	-0.03	0.02	-0.05	0.136	0.07	0.502	0.189	-0.08	0.358	0.291	-0.06
CH3	3E-04	0.011	0.151	0.012	-0.03	-0.5	-0.11	-0.13	-0	0.193	0.097	-0.4	0.657	-0.22	-0.37	-0.12
CH2	-0.04	-0.08	-0	0.047	0.02	-0.11	-0.04	0.001	0.086	0.009	-0.19	0.119	-0.18	-0.13	0.012	-0.08
CFH	0.259	0.02	0.028	-0.03	-0.05	-0.13	0.001	-0.06	0.06	0.101	-0.26	0.126	0.116	0.211	-0.37	0.107
CZ	-0.34	-0.09	-0.07	-0.33	0.136	-0	0.086	0.06	0.047	-0.63	1.315	-0.72	-0.17	0.968	-1.52	0.63
OH	-0.15	0.054	0.102	0.031	0.07	0.193	0.009	0.101	-0.63	-0.18	-0.04	0.145	-0.12	-0.39	0.425	-1.2
CGTR	1.891	0.142	-0.27	-0.02	0.502	0.097	-0.19	-0.26	1.315	-0.04	2.458	-1.05	0.062	0.374	-0.28	-1.08
CHTR	-0.34	-0.81	-0.02	0.245	0.189	-0.4	0.119	0.126	-0.72	0.145	-1.05	-1.2	1.214	-0.24	0.824	1.252
NDHS	-2.39	0.59	-0.18	-0.36	-0.08	0.657	-0.18	0.116	-0.17	-0.12	0.062	1.214	2.365	-2.47	-1.64	0.653
CH2M	-0.91	0.038	-0.03	0.004	0.358	-0.22	-0.13	0.211	0.968	-0.39	0.374	-0.24	-2.47	-1.08	1.439	0.358
SM	-0.47	-0.26	0.491	-0.46	0.291	-0.37	0.012	-0.37	-1.52	0.425	-0.28	0.824	-1.64	1.439	-3.18	1.053
CH3M	-1.04	-0.38	0.472	-0.3	-0.06	-0.12	-0.08	0.107	0.63	-1.2	-1.08	1.252	0.653	0.358	1.053	-0.96
CH2K	0.307	0.315	0.094	0.034	-0.08	0.272	0.124	-0.13	0.266	0.025	-0.1	-0.68	1.084	1.474	1.039	-0.46
CH2S	0.144	-0.31	-0.23	0.208	0.334	-0.03	-0.04	-0.01	0.377	0.143	-0.39	0.093	-0.39	-0.4	0.449	0.633
CHPR	0.321	0.058	0.205	-0.04	-0.12	0.209	0.142	0.011	-0.52	0.151	-0.76	1.23	-0.78	-0.74	-0.45	0.903
CH2C	-1.16	-0.59	-0.36	0.626	0.672	-0.29	-0.08	-0.24	1.537	-0.48	-1.03	-0.75	-1.94	-0.15	-1.25	-1.05
SH	-0.85	0.1	0.389	-0.68	-0.23	-0.23	-0.23	-0.14	0.828	0.059	-0.23	1.222	0.778	0.277	0.195	0.365
CGHP	-0.36	0.051	-0.12	-0.04	0.151	0.138	-0.06	-0.18	-0	-0.18	0.541	0.622	-0.41	1.682	-1.62	-0.67
NDHP	-0.32	-0.33	-0.09	0.246	0.142	-0.26	-0.16	-0.11	-0.72	-0.34	-0.73	0.238	-2.33	-2.41	-0.32	0.51
CHEP	1.321	0.536	-0.22	0.222	0.284	0.184	0.199	0.269	2.002	-0.19	-0.25	1.943	-1.12	2.493	0.104	1.253
CR2	0.862	0.384	0.017	-0.07	0.222	0.104	-0.02	-0.2	-0.42	0.199	1.116	-2	1.706	-2.09	0.703	-0.6
NR1	-1.62	-0.05	-0.51	0.302	-0.13	-0.23	0.033	-0.04	1.104	0.01	-0.85	0.087	0.358	1.526	-0.1	-0.58
CR3	0.27	0.055	-0.27	-0.1	0.387	-0.43	-0.26	0.238	-1.99	-0.91	-1.34	1.186	-1.44	1.634	-1.66	-0.89
NR2	0.402	0.086	-0.05	0.047	-0.24	0.266	0.341	-0.05	-0.05	0.35	0.654	0.412	-0.32	-0.91	1.111	1.158
NAS	0.172	0.548	-0.31	0.083	0.076	0.124	-0.02	-0.06	0.814	-0.32	1.128	-0.73	-0.17	0.125	1.104	-0.97
CH2B	0.477	-0.25	0.164	0.205	0.146	-0.09	-0.04	-0.06	0.131	0.03	0.068	-0.2	0.508	-0.16	0.155	1.202
CX1	0.062	-0.05	-0.04	0.136	0.095	-0.18	-0.24	-0.15	-0.92	-0.13	0.396	-0.32	0.388	0.758	-1.46	0.083
OX1	-0.37	-0.18	0.088	-0.09	0.213	0.217	0.077	0.157	0.277	-0.2	0.158	-0.48	0.141	-0.11	0.065	-0.2

CH2K	CH2S	CHPR	CH2C	SH	CGHP	NDHP	CHEP	CR2	NR1	CR3	NR2	NAS	CH2B	CX1	OX1	
0.307	0.144	0.321	-1.16	-0.85	-0.36	-0.32	1.321	0.862	-1.62	0.27	0.402	0.172	0.477	0.062	-0.37	NX
0.315	-0.31	0.058	-0.59	0.1	0.051	-0.33	0.536	0.384	-0.05	0.055	0.086	0.548	-0.25	-0.05	-0.18	NH

0.094	-0.23	0.205	-0.36	0.389	-0.12	-0.09	-0.22	0.017	-0.51	-0.27	-0.05	-0.31	0.164	-0.04	0.088	CO
0.034	0.208	-0.04	0.626	-0.68	-0.04	0.246	0.222	-0.07	0.302	-0.1	0.047	0.083	0.205	0.136	-0.09	OC
-0.08	0.334	-0.12	0.672	-0.23	0.151	0.142	0.284	0.222	-0.13	0.387	-0.24	0.076	0.146	0.095	0.213	CAH
0.272	-0.03	0.209	-0.29	-0.23	0.138	-0.26	0.184	0.104	-0.23	-0.43	0.266	0.124	-0.09	-0.18	0.217	CH3
0.124	-0.04	0.142	-0.08	-0.23	-0.06	-0.16	0.199	-0.02	0.033	-0.26	0.341	-0.02	-0.04	-0.24	0.077	CH2
-0.13	-0.01	0.011	-0.24	-0.14	-0.18	-0.11	0.269	-0.2	-0.04	0.238	-0.05	-0.06	-0.06	-0.15	0.157	CFH
0.266	0.377	-0.52	1.537	0.828	-0	-0.72	2.002	-0.42	1.104	-1.99	-0.05	0.814	0.131	-0.92	0.277	CZ
0.025	0.143	0.151	-0.48	0.059	-0.18	-0.34	-0.19	0.199	0.01	-0.91	0.35	-0.32	0.03	-0.13	-0.2	OH
-0.1	-0.39	-0.76	-1.03	-0.23	0.541	-0.73	-0.25	1.116	-0.85	-1.34	0.654	1.128	0.068	0.396	0.158	CGTR
-0.68	0.093	1.23	-0.75	1.222	0.622	0.238	1.943	-2	0.087	1.186	0.412	-0.73	-0.2	-0.32	-0.48	CHTR
1.084	-0.39	-0.78	-1.94	0.778	-0.41	-2.33	-1.12	1.706	0.358	-1.44	-0.32	-0.17	0.508	0.388	0.141	NDHS
1.474	-0.4	-0.74	-0.15	0.277	1.682	-2.41	2.493	-2.09	1.526	1.634	-0.91	0.125	-0.16	0.758	-0.11	CH2M
1.039	0.449	-0.45	-1.25	0.195	-1.62	-0.32	0.104	0.703	-0.1	-1.66	1.111	1.104	0.155	-1.46	0.065	SM
-0.46	0.633	0.903	-1.05	0.365	-0.67	0.51	1.253	-0.6	-0.58	-0.89	1.158	-0.97	1.202	0.083	-0.2	CH3M
0.997	-0.58	-1.11	1.804	-0.76	-2.19	2.598	-1.15	0.392	0.374	-1.51	0.56	-0.33	-0.03	0.576	-0.72	CH2K
-0.58	-0.48	0.183	0.022	0.433	0.112	0.306	-0.76	-0.43	0.493	-0.64	0.139	-0.25	0.039	0.033	-0.28	CH2S
-1.11	0.183	-0.22	0.999	-0.63	-0.01	-0.2	0.23	1.008	-0.17	-0.6	-0.05	-0.37	0.578	0.152	-0.04	CHPR
1.804	0.022	0.999	-9.98	4.788	0.361	0.76	-1.42	-0.75	0.84	-0.32	-1.02	-0.4	0.515	-0.9	-0.43	CH2C
-0.76	0.433	-0.63	4.788	-1.81	-2.58	0.486	1.883	1.101	-1.82	-1.5	1.711	0.198	1.009	1.524	-0.57	SH
-2.19	0.112	-0.01	0.361	-2.58	0.299	-0.25	0.796	-0.58	-1.31	0.35	0.503	0.193	0.719	-0.81	-0.38	CGHP
2.598	0.306	-0.2	0.76	0.486	-0.25	0.582	-2.14	-0.92	-0.29	-0.17	0.326	0.186	-0.09	0.611	0.64	NDHP
-1.15	-0.76	0.23	-1.42	1.883	0.796	-2.14	3.733	2.608	0.739	-0.68	0.399	-0.89	-1.15	-0.76	-0.58	CHEP
0.392	-0.43	1.008	-0.75	1.101	-0.58	-0.92	2.608	1.48	-1.38	1.775	-0.78	-0.52	0.402	-0.57	-0.44	CR2
0.374	0.493	-0.17	0.84	-1.82	-1.31	-0.29	0.739	-1.38	-0.39	-0.04	-0.33	1.248	0.245	0.12	-0.01	NR1
-1.51	-0.64	-0.6	-0.32	-1.5	0.35	-0.17	-0.68	1.775	-0.04	-1.43	0.936	0.449	-0.59	-0.22	-0.58	CR3
0.56	0.139	-0.05	-1.02	1.711	0.503	0.326	0.399	-0.78	-0.33	0.936	-0.2	-0.68	-0.34	0.113	0.285	NR2
-0.33	-0.25	-0.37	-0.4	0.198	0.193	0.186	-0.89	-0.52	1.248	0.449	-0.68	-0.5	0.226	-0.42	-0.41	NAS
-0.03	0.039	0.578	0.515	1.009	0.719	-0.09	-1.15	0.402	0.245	-0.59	-0.34	0.226	-0.11	-0.47	0.518	CH2B
0.576	0.033	0.152	-0.9	1.524	-0.81	0.611	-0.76	-0.57	0.12	-0.22	0.113	-0.42	-0.47	-0.95	0.158	CX1
-0.72	-0.28	-0.04	-0.43	-0.57	-0.38	0.64	-0.58	-0.44	-0.01	-0.58	0.285	-0.41	0.518	0.158	-0.18	OX1

Table III

Summary of approximate recognition test using the potentials designed in this paper and other widely used scoring functions. On the left of the table we list the potentials under consideration. T32S1, T46S1, T32S3, and T46S3 are the potential computed in the present manuscript (see text for more details). LS (Lu and Skolnick [3]) denotes an atomically detailed potential with 3 steps which is similar in functional form to T46S3 and T32S3, except that 167 atom types are used instead of 46 and 32. HL (Hind Levitt [41]) and MJ (Miyazawa Jernigan [1]) are residue-based contact-potentials and TE13 (Tobi and Elber [11]) is a distance dependent residue-residue potential. The horizontal line provides the rank. The first column is the number of proteins with correct models ranked first, the second the number of proteins with correct models that were ranked either first or second and so on. A model is considered correct if the RMS distance of the C_α to the native is at most 6Å. Note the fluctuations in performance of different potentials when the alignment changes, suggesting that on this test the potential designed in the present manuscript are comparable and similar in performance to other scoring functions. Note that the larger number of parameters in T46S3 does not yield a significantly better performance.

(a) The recognition capacity of different potentials on a subset of CASP5 exercise performed on 14 proteins using alignment I (This alignment is largely based on BLOSUM 50 with structure dependent gap penalties). The total number of decoy structures for each template was between 44 and 50. The maximal number of correct predictions that could have been made is 14.

T32S1	11	12	12	12	12	12	12	12	12	12
T46S1	11	11	12	12	12	12	12	12	12	12
T32S3	11	11	11	11	11	11	11	11	11	11
T46S3	11	11	12	12	12	12	12	12	12	12
LS	9	11	11	12	13	13	13	13	13	13
HL	10	11	12	13	13	13	13	13	13	13
MJ	11	11	13	13	13	13	13	13	13	13
TE13	7	9	10	11	12	12	12	12	13	13

(b) The recognition capacity of different potentials on a subset of CASP5 exercise performed on only 15 proteins using alignment II (the substitution matrix learned from structural alignments, Jian Qiu unpublished). The total number of decoy structures for each template varies between 22 and 49. The maximal number of correct predictions that could have been made is 15.

T32S1	9	10	10	10	10	10	10	11	11	11
T46S1	8	9	10	10	11	11	11	11	11	11
T32S3	11	11	11	11	12	12	12	12	12	12
T46S3	10	10	12	12	12	12	12	12	12	12
LS	10	11	12	12	12	13	13	13	14	14
HL	7	9	11	11	11	11	12	12	12	12
MJ	8	9	10	11	11	11	11	11	11	11
TE13	8	9	10	10	10	10	10	10	11	11

(c) Similarly to (a) but using an alternative (more extensive) database of templates that enables the (maximal) recognition of 30 proteins. The number of decoys for each protein varies between 24 and 50.

T32S1	26	27	27	28	28	28	28	30	30	30
T46S1	25	27	28	28	30	30	30	30	30	30
T32S3	25	26	28	29	29	30	30	30	30	30
T46S3	25	26	28	28	28	29	29	29	29	29
LS	24	27	27	27	27	27	27	27	28	28
HL	24	26	27	28	29	29	29	29	29	29
MJ	20	24	26	27	27	28	29	29	29	29
TE13	19	22	22	23	24	24	24	25	25	25

Table IV

Identification tests of a set of 17 proteins (protein databank codes 1acp 1ah9 1cei 1cewI 1coo 1difA 1emn 1fow 1fvl 1hdj 1hoe 1iba 1kjs 1kptA 1krn 1ngr 1npoA) against decoy structures generated ab-initio. One thousand decoy structures with the approximate side chains built up with the program Phoenix2 are considered. T46S3, which was trained on the same data set as T32S3, is performing worse on an independent dataset, suggesting over learning. For a description of the potentials see table III.

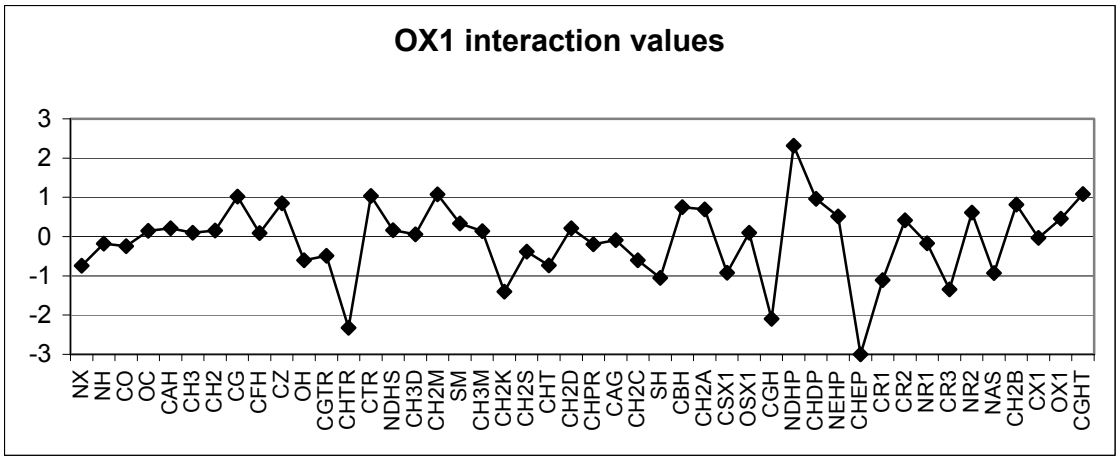
T32S1	7	7	7	8	10	10	11	12	12	12
T46S1	7	8	8	8	8	9	9	9	9	9
T32S3	11	11	11	12	12	12	13	13	13	13
T46S3	9	10	10	10	10	10	10	10	10	10
HL	0	0	0	0	0	0	0	0	0	0
MJ	0	0	0	0	0	0	0	0	0	0
LS	4	5	6	8	8	8	8	8	8	8
TE13	2	2	3	4	4	4	4	4	4	4

Table V
An identification test of 74 proteins. On the left hand side of each row we report the potential examined, the numbers that follow are the natives that were identified as the first (the first column), natives that were identified as the first or the second (the second column) and so on. Each of the 74 proteins (not included in our training set) has a set of 24,000 decoy structures and one correct hit (native). We have generated side chains for this decoy set (that include only alpha carbon and geometric center of side chains) using the Phoenix2 program. For a description of the potentials considered see table III. We provide ranking up to the tenth place. The maximal number of correct predictions is 74. Note that single step potentials (residue or atomic based) perform poorly on this set and that the LS potential performs the best for getting the natives to the top 3. T32S3 brings the largest number of the native sequences to the top 4. The list of the proteins is available from http://www.cs.cornell.edu/~jianq/research.htm)

T32S1	3	6	7	8	8	9	10	10	10	10
T46S1	5	10	10	12	12	12	13	13	13	13
T32S3	32	39	40	44	44	47	48	50	50	50
T46S3	24	27	31	36	38	38	38	38	39	41
LS	38	42	43	43	43	43	43	45	46	47
HL	2	4	4	4	4	4	4	4	4	4
MJ	4	4	4	4	4	4	4	4	4	4
TE13	7	13	13	15	15	16	16	16	16	16

Figure 1

The interaction energies of two "extreme" atoms (charged OX1 and hydrophobic CFH) with the rest of the atoms according to the potential A46S1. Note that the interaction of OX1 with the backbone atoms (OC CAH), is positive (repulsive) while the interactions of CFH with the backbone atoms (CAH CO OC) are negative (but CO, OC are quite close to zero), only NH is positive, suggesting that the CFH feels overall attraction from the backbone. This interaction is significant since it is abundant and mimics hydrophobicity.



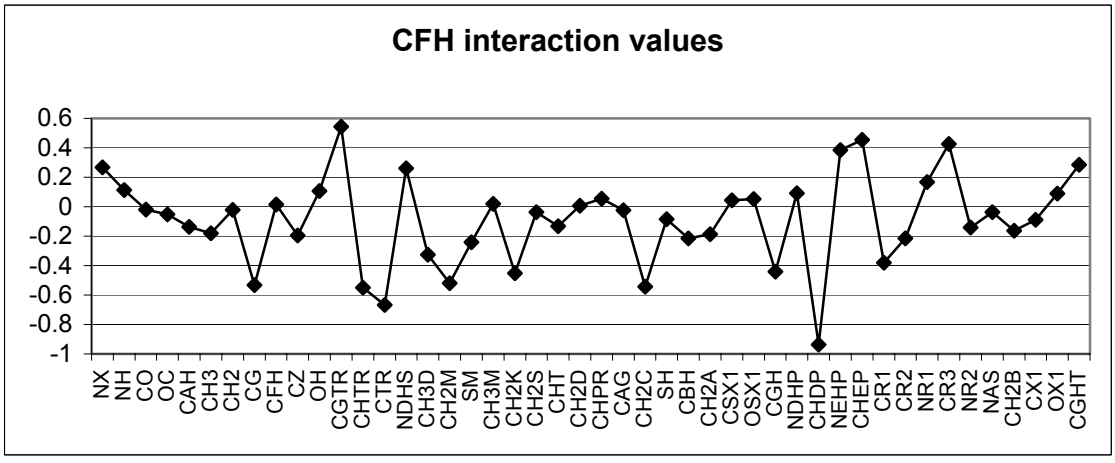


Figure 2

Overlap of the backbone of the structural models with the backbone of the native structure. In 2.a we show the overlap with a correct model (RMSD= 5.03) that also has a low T32S3 energy (-591.26) and in 2.b we show a worse model (but still acceptable, RMSD =6.90) that has a high T32S3 energy (-438.13). In 2.c we show an overlap with a wrong structure (RMSD =9.49) that has a low T32S3 energy (-533.97).

Figure 2.a

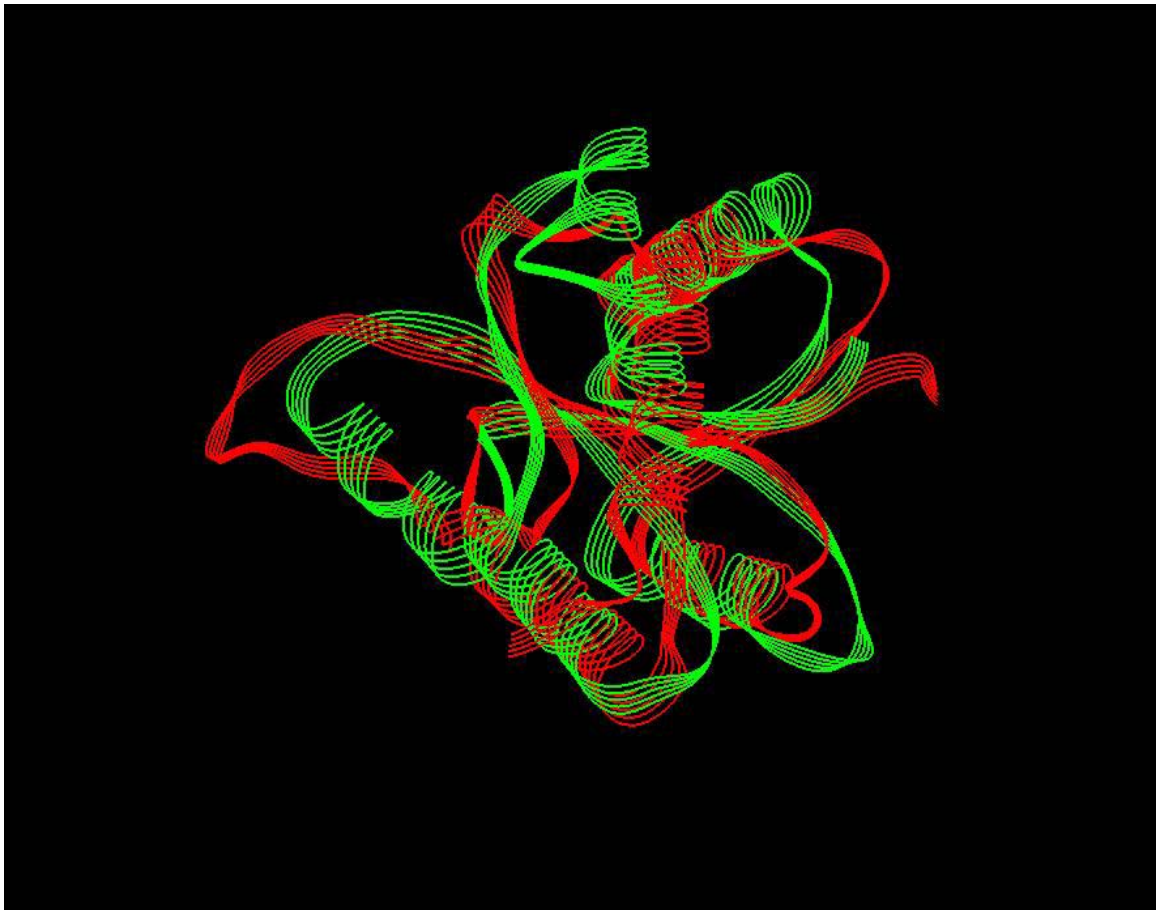


Figure 2b

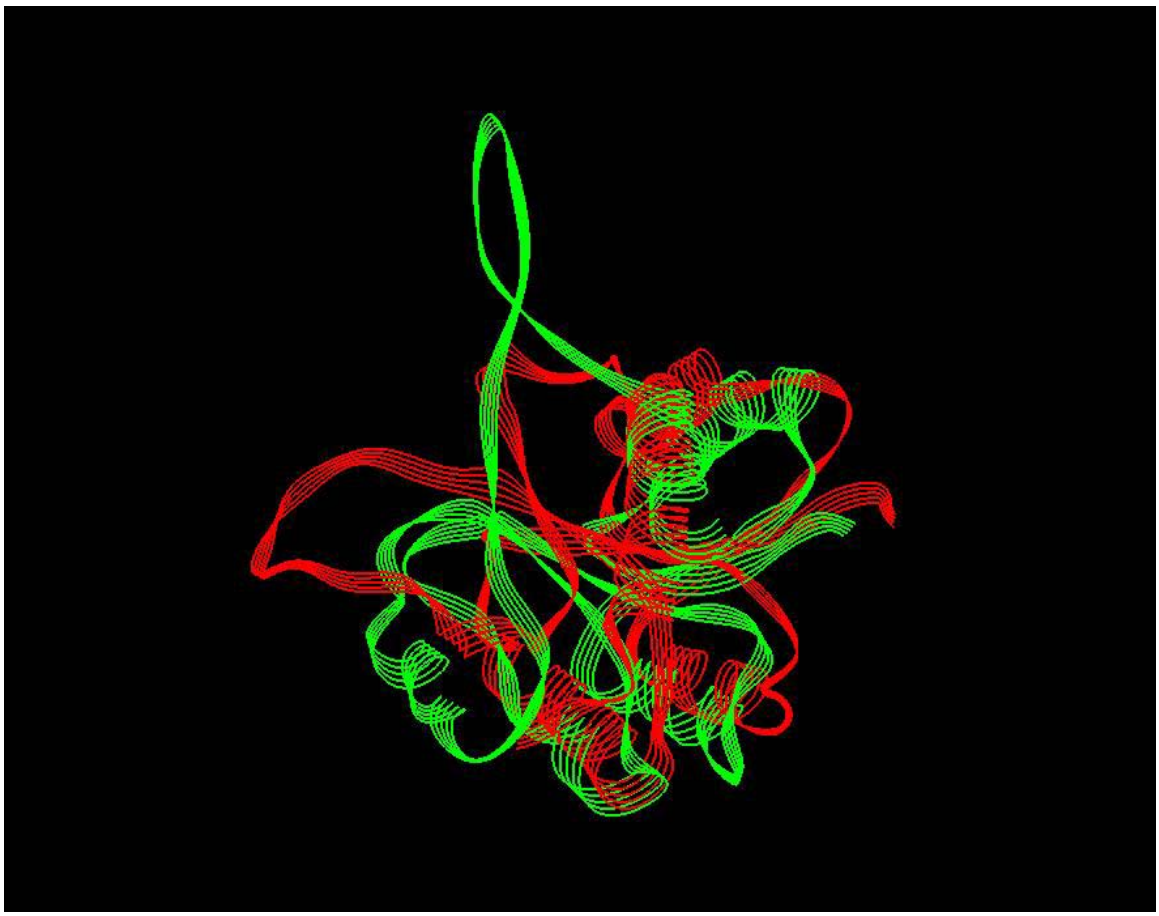


Figure 2c

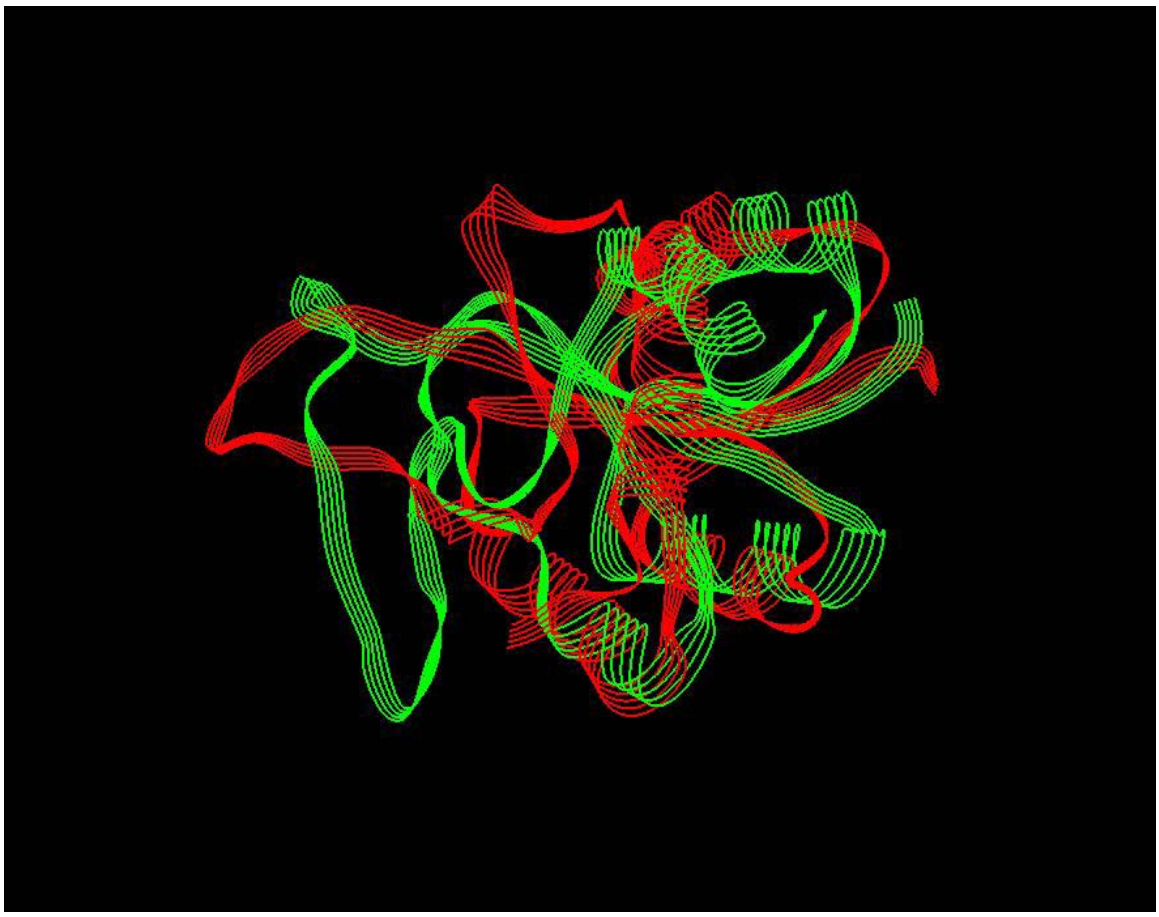
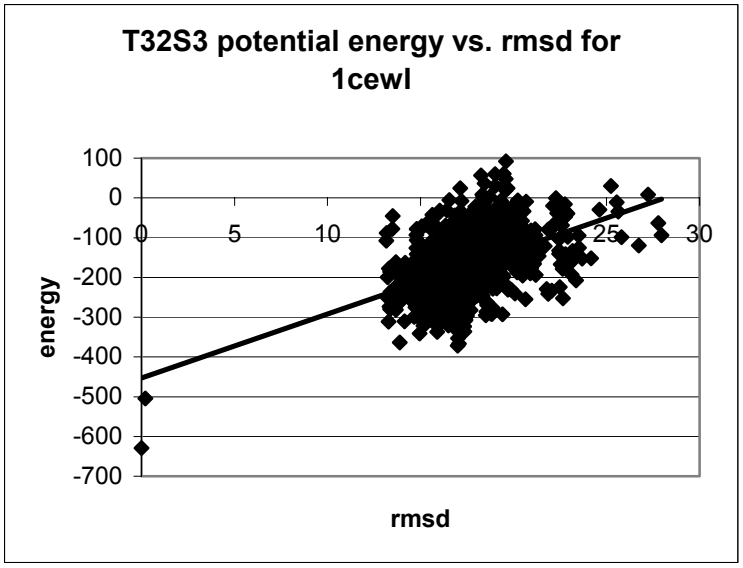


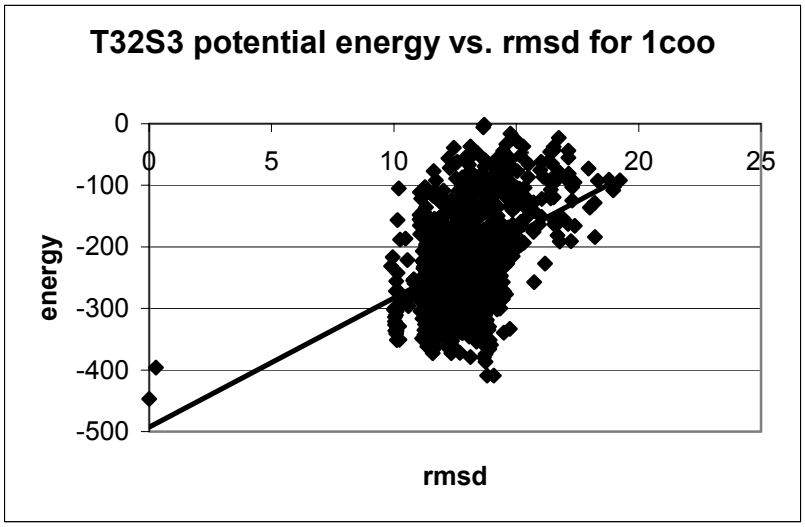
Figure 3.

A scatter plot of energy versus RMS

3.a The protein 1cewI from the 17 set of Skolnick. We observe a reasonable correlation between energy and RMS though the decoys are quite far from the native. The total number of structures included in the plot was 1002.



3.b A successful recognition of the native shape but poor correlation of the energy with the RMS value for the protein 1coo. Note however, that most decoy structures are above 10Å in distance from the native.



3.c A failure to recognize the native fold and a poor correlation of the energy with the RMS for the protein 1fvl

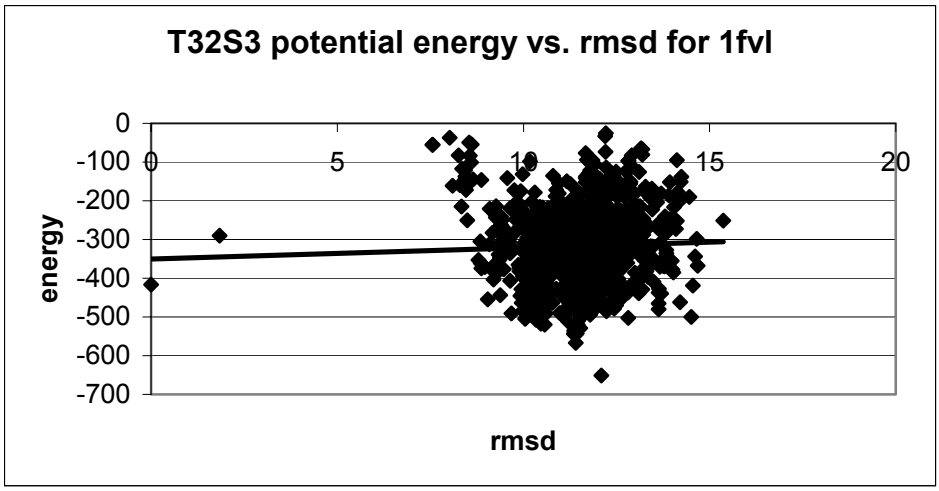
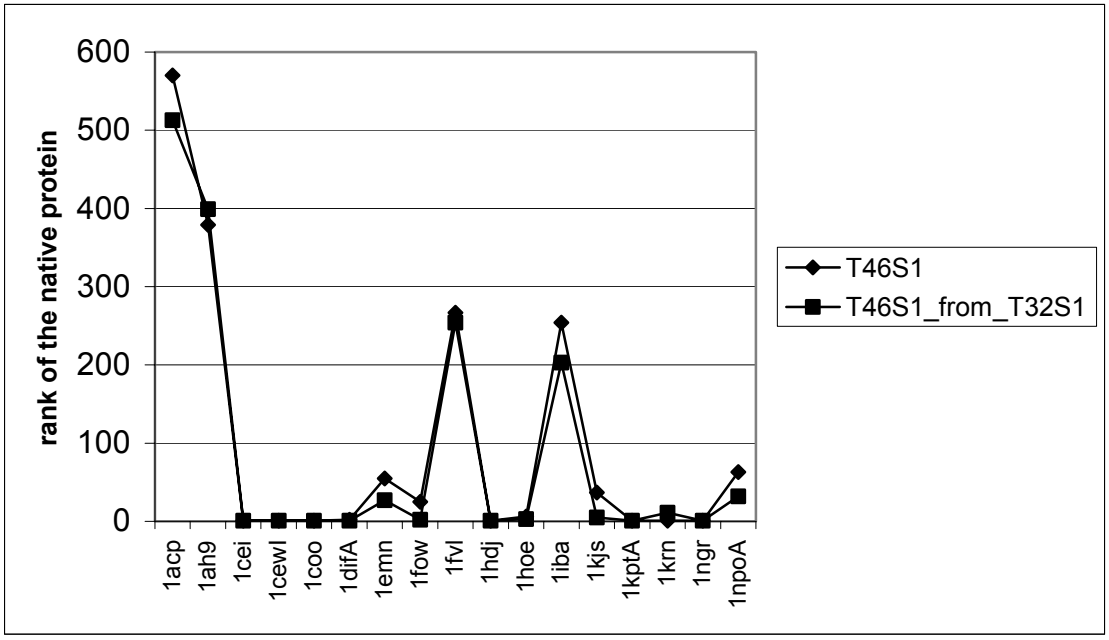


Figure 4

Testing the convergence of the optimization algorithm for the potential T46S1. The performances of two T46S1 potentials are compared on the Skolnick_17 set. One of the potential is the same as reported in the text; the second is a re-optimization of the T32S1 potential (see text for more details).



References

1. Miyazawa, S. and R.L. Jernigan, *Estimation of effective interresidue contact energies from protein crystal structures: Quasi chemical approximation.* Macromolecules, 1984. **18**(3): p. 534-552.
2. Maiorov, V.N. and G.M. Crippen, *Contac potential that recognizes the correct folding of globular proteins.* Journal of Molecular Biology, 1992. **227**: p. 876-888.
3. Lu, H. and J. Skolnick, *A distance dependent atomic knowledge-based potential for improved protein structure selection.* Proteins, Structure, Function and Genetics, 2001. **44**(3): p. 223-232.
4. Godzik, A., A. Kolinksi, and J. Skolnick, *Knowledge-based potential for protein folding: What can we learn from protein structures?* Proteins, Structure, Function and Genetics, 1996. **4**: p. 363-366.
5. Zhou H.Y. and Zhou Y.Q., *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.* Protein science, 2002. **11**(11): p. 2714-2726.
6. Park, B.H., E.S. Huang, and M. Levitt, *Factors affecting the ability of energy functions to discriminate correct from incorrect folds.* Journal of Molecular Biology, 1997. **266**(4): p. 831-846.
7. Vendruscolo, M., et al., *Comparison of two optimization methods to derive energy parameters for protein folding: Percepton and Z score.* Proteins, Structure, Function and Genetics, 2000. **41**(2): p. 192-201.
8. Dominy, B.N. and E.I. Shakhnovich, *Native atom types for knowledge-based potentials: Application to binding energy prediction.* Journal of Medicinal Chemistry, 2004. **47**(18): p. 4538-4558.
9. Buchete, N.-V., J.E. Straub, and D. Thiramulai, *Development of novel statistical potentials for protein fold recognition.* Current Opinion in Structural Biology, 2004. **14**(2).
10. Betancourt, M.R. and D. Thirumalai, *Pair potentials for protein folding: choice of reference states and sensitivity of predicted states to variations in the interaction schemes.* Protein science, 1999. **2**: p. 361-369.
11. Tobi, D. and R. Elber, *Distance dependent, pair potential for protein folding: Results from linear optimization.* Proteins, Structure, Function and Genetics, 2000. **41**: p. 40-46.
12. Wagner, M., J. Meller, and R. Elber, *Large-scale linear programming techniques for the design of protein folding potentials.* Mathematical Programming Ser. B, 2004. **101**(2): p. 301-318.
13. Loose, C., J.L. Kelpeis, and C.A. Floudas, *A new pairwise folding potential based on improved decoy generation and side chain packing.* Proteins Structure Function and Genetics, 2004. **54**(2): p. 303-314.
14. Tobi, D., et al., *On the design and analysis of protein folding potentials.* Proteins, Structure, Function and Genetics, 2000. **40**: p. 71-85.
15. Meller, J. and R. Elber, *Protein recognition by sequence-to-structure fitness: Bridging efficiency and capacity of threading models*, in *Advances in chemical physics*, F. Richard, Editor. 2002, John Wiley & Sons. p. 77-130.

16. Meller, J., M. Wagner, and E. R., *Maximum Feasibility Guidline in the Design and Analysis of Protein Folding Potentials.* J Comput Chem, 2002. **23**: p. 111-118.
17. Meller, J. and R. Elber, *Linear Optimization and a double statistical filter for protein threading protocols.* Proteins, Structure, Function and Genetics, 2001. **45**: p. 241.
18. Clementi, C., et al., *Folding Lennard-Jones proteins by a contact potential.* Proteins Structure Function and Genetics, 1999. **37**(4): p. 544-553.
19. Vendruscolo, V. and E. Domany, *Pairwise contact potentials are unsuitable for protein folding.* J. Chem. Phys., 1998. **109**: p. 11101-11108.
20. Goldstein, R., A.,, Z. Luthey-Schulten, and P.G. Wolynes, *Protein tertiary structure recognition using optimized hamiltnians with local interactions.* Proc. Natl. Acad. Sci USA, 1992. **89**(19): p. 9029-9033.
21. Klimov, D.K. and D. Thirumalai, *Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics.* Journal of Chemical Physics, 1998. **109**(10): p. 4119-4125.
22. Miyazawa, S. and R.L. Jernigan, *An empirical energy potential with a reference state for protein fold and sequence recognition.* Proteins, Structure, Function and Genetics, 1999. **36**(3): p. 357-369.
23. Miyazawa, S. and R.L. Jernigan, *Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition.* Proteins Structure Function and Genetics, 1999. **36**(3): p. 347-356.
24. Skolnick, J., et al., *Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?* Protein science, 1997. **6**(3): p. 676-688.
25. Wolynes, P.G., J.N. Onuchic, and D. Thirumalai, *Navigating the folding routes.* Science, 2004. **267**(5204): p. 1619-1620.
26. Zhang, Y., A. Kolinksi, and J. Skolnick, *Touchstone II: A new approach to ab initio protein structure prediction.* Biophysical Journal, 2003. **85**: p. 1145-1164.
27. Delarue, M. and P. Koehl, *Atomic environment energies in proteins defined from statistics of accessible and contact surface-areas.* Journal of Molecular Biology, 1995. **249**(3): p. 675-690.
28. Liu, S., et al., *A physical reference state unifies the structure-derived potential of mean force for protein folding and binding.* Proteins Structure Function and Bioinformatics, 2004. **56**(1): p. 93-101.
29. Samudrala, R. and J. Moult, *An all-atom distance dependent conditional probability discriminatory function for protein structure prediction.* Journal of Molecular Biology, 1998. **275**: p. 895-916.
30. Levitt, M., *Decoy structures: http://dd.stanford.edu/.* 2004.
31. Baker, D., *decoy structures: http://depts.washington.edu/bakerpg/decoys/.* 2004.
32. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Engineering, 1998. **11**: p. 739-747.
33. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J. Mol. Biol., 1993. **234**: p. 779-815.

34. Marti-Renom, M.A., et al., *Comparative protein structure modeling of genes and genomes.* Annual Review Biophys. Biomol. Structure, 200. **29**: p. 291-325.
35. Elber, R., et al., *LOOPP: Learning Observing and Outputing Protein Patterns: http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm.* 2004.
36. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proceeding of the National Acadamy of Science USA., 1989. **89**: p. 10915-10919.
37. Jorgensen, W.L. and J. Tirado-Rives, *The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin.* Journal of the American Chemical Society, 1988. **110**(6): p. 1666-1671.
38. Elber, R., et al., *Moil - a Program for Simulations of Macromolecules.* Computer Physics Communications, 1995. **91**(1-3): p. 159-189.
39. Meszaros, C.S., *Fast Cholesky factorization for interior point methods of linear programming.* Comput. Math. Appl., 1996. **31**: p. 59-51.
40. Wright, S.J., *Primal-dual interior-point methods*. 1997, Philadelphia: SIAM. 289.
41. Hinds, D.A. and M. Levitt, *Exploring conformational space with a simple lattice model for protein-structure.* Journal of Molecular Biology, 1994. **243**(4): p. 668-682.
42. Zhang, Y. and J. Skolnick, *Automated structure prediction of weakly homologous proteins on a genomic scale.* Proc. Natl. Acad. Sci USA, 2004. **101**: p. 7594-7599.
43. Braxenthaler, M., et al., *PROSTAR: The protein potential test site.* http://prostar.carb.nist.gov, 1997.
44. Holm, L. and C. Sander, *Evaluation of protein models by atomic solvation preference.* Journal of Molecular Biology, 1992. **225**: p. 93-105.
45. Pedersen, J.T. and J. Moult, *Folding simulation with genetic algorithms and a detailed molecular description.* Journal of Molecular Biology, 1997. **269**: p. 240-259.