Determination of optimal effective interactions between amino acids in globular proteins.

*Optimal effective interactions of amino acids*

Giovanni Settanni\(^1\), Cristian Micheletti\(^1\), Jayanth Banavar\(^2\), Amos Maritan\(^1\)

\(^1\)International School for Advanced Studies (SISSA) and INFM, Via Beirut 2, 34014 Trieste, Italy; The Abdus Salam Centre for Theoretical Physics - Trieste, Italy; 
\(^2\)Department of Physics and Center for Materials Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802

February 1, 2008

Abstract

An optimization technique is used to determine the pairwise interactions between amino acids in globular proteins. A numerical strategy is applied to a set of proteins for maximizing the native fold stability with respect to alternative structures obtained by gapless threading. The extracted parameters are shown to be very reliable for identifying the native states of proteins (unrelated to those in the training set) among thousands of conformations. The only poor performers are proteins with heme groups and/or poor compactness whose complexity cannot be captured by standard pairwise energy functionals.

**Keywords:** potential extraction, fold recognition, optimal stability, energy gap maximization.

1 Introduction

A knowledge of the interaction potentials between amino acids is of crucial importance both for predicting the three-dimensional structure of a protein’s native state and for designing novel proteins, folding on a desired target conformation (Bauer & Lawrence 1993, Casari & Sipp 1992, Godzik 1994, Bowie, Lüthy & Eisenberg 1991, Bryant & Wolynes 1992, Socci & Onuchic 1994, Huang, Subbiah & Levitt 1995, Jones, Taylor & Thornton 1992, Ouzounis, Sander, Sharf & Schneider 1993, Bowie & Skolnick 1994, Levitt 1976, Levitt 1983, Skolnick 1994, Kolinsky & Skolnick 1994, Levitt 1976, Levitt 1983, Skolnick 1994, Kolinsky, Brooks, Godzik & Rey 1993, Sun 1993, Walqvist & Ulner 1994, Deutsch & Kuroskey 1996, Srinivasan & Rose 1993, Micheletti, Seno, Maritan & Banavar 1998, Sano, Micheletti, Maritan & Banavar 1998, Creighton 1993, Branden & Tooze 1991, Anfinsen 1973). This builds on the assumption that the interactions between amino acids (and the solvent) are principally responsible for driving the folding of a protein to its native state. This is supported by considerable experimental evidence that the native states of many globular proteins correspond to free energy minima (Anfinsen 1973, Wolynes, Onuchic & Thirumalai 1995, Creighton 1993, Branden & Tooze 1991, Anfin- sen 1973). On a microscopic scale, all-atom potentials are used to carry out “first principle” molecular dynamics for folding (van Gunsteren 1989). Due to the high level of details included in such calculations,
the folding processes of short peptides can be followed only for rather short time-scales (of the order of 1 µs as in Duan & Kollman (1998)). While the impact of such “ab initio” calculations is destined to grow rapidly, at present, highly satisfactory results can be obtained by adopting mesoscopic phenomenological approaches. Within this framework, a commonly used approach is to avoid a detailed description of an amino acid but represent it as a sphere or an ellipsoid centered on the $C_{\alpha}$ or $C_{\beta}$ position (Maiorov & Crippen 1992, Srinivasan & Rose 1995, Kolinsky, Godzik & Skolnick 1993, Sun, Brem, Chan & Dill 1995, Micheletti et al. 1998). This coarse-grained procedure amounts to integrating out the fine degrees of freedom of a peptide chain and introduces effective interactions between the surviving degrees of freedom.

One commonly used strategy to extract coarse grained potentials between pairs of amino acids has been proposed by Miyazawa & Jernigan (1985). The method is based on the quasichemical approximation and it entails the calculation of pairing frequencies of amino acids observed in native structures of naturally occurring proteins. Similar approaches have been reviewed by Sippl (1995) and Wodak & Rooman (1993). Thomas & Dill (1996) have recently tested the validity of this procedure on exactly solvable lattice models for proteins. In all the cases they considered the extracted potentials did not correlate too well with the true potentials, although the two sets shared a common trend.

A different strategy for extracting potentials was suggested by Maiorov & Crippen (1992) and recently an optimized version has been introduced (van Mourik, Clementi, Maritan, Seno & Banavar 1998, Seno, Maritan & Banavar 1998). Rigorous tests, similar to the ones in Thomas & Dill (1996), carried out both for lattice and off-lattice models have shown that the optimized strategy converges to the exact potentials for increasing chain length and/or number of proteins in the training set. The method, explained in detail in section 4, uses the following basic ingredients: the potentials parametrizing a suitably chosen Hamiltonian must be such that the energy of a protein sequence in its own native state is lower than in any other alternative conformations that the protein can attain. For each sequence this yields a set of linear inequalities involving the unknown interaction potentials. Two key points need to be addressed carefully when applying this procedure: the parametrization of the Hamiltonian and the generation of alternative conformations. If the parametrization of the energy is too poor and/or there are unphysical conformations among the decoys (i.e., violating steric contraints), then no consistent solution can be found (unlearnable problem, for an example see van Mourik et al. (1998) and Vendruscolo, Najmanovich & Domany (1999)). On the other hand, if the parametrization is reliable and there are no unphysical decoy conformations, the energy parameter satisfying the inequalities lie in a convex region of parameter space. While all points inside the cell satisfy the whole set of inequalities, there is an optimal point, typically equidistant from the hyperplanes bounding the cell. The potential parameters corresponding to the optimal point, ensure that the native states of proteins are maximally stable with respect to alternative structures. Our strategy aims at pinpointing the optimal solution, while the original Maiorov and Crippen strategy stopped when reaching an unspecified sub-optimal point inside the cell. Our approach differs from the one employed in (Maiorov & Crippen 1992) also because of the different interaction matrix: in our scheme (as in Miyazawa & Jernigan (1985) or Kolinsky et al. (1993)) the interaction energy of amino acids pairs does not depend on their sequence separation, while a complementary strategy was followed by Maiorov and Crippen. In the next Section we introduce the coarse grained model for proteins and give an overview of the optimal potential extraction technique. The latter is discussed in detail in Section 4. An assessment of the performance of extracted potentials, and a comparison with previously known interactions, are given in Section 3.

## 2 Theory

### 2.1 The Model

We choose not to introduce any subdivision of amino acids in classes and retain the full repertoire of 20
For simplicity of notation, in the following, we will use the 20 × 20 matrix of contact energies. Since ε is symmetric, there are only 210 distinct entries in the matrix. We also considered a second form with 20 additional terms related to the degree of solvation of amino acid types:

$$E(S, \Gamma) = \sum_{i>j+1}^{N} \epsilon(A_i, A_j) \cdot \Delta(r_i, r_j),$$  \(2\)

where the sum is over all pairs of non-consecutive residues, N is the protein length and A_i is the amino acid type (there are altogether 20 types) at r = r_i. ε is the 20 × 20 matrix of contact energies. Since ε is symmetric, there are only 210 distinct entries in the matrix. We also considered a second form with 20 additional terms related to the degree of solvation of amino acid types:
2.2 Optimal strategy

The key prescription at the heart of the potential extraction scheme is that a protein sequence attains the lowest possible energy when mounted on its correct native state. Hence, assuming that the energy parametrizations (4) and (3) are reliable, the correct potentials will be such that the native state has the lowest energy when compared to alternative conformations.

The first step of the analysis was to compile a list of non-homologous proteins representing a variety of folds (see section 4 for details). For each protein sequence in this training set, \( S_i \) (with known native state \( \Gamma_i \)), the alternative structures are obtained by threading on conformations in the training set of equal or longer length (Jones et al. 1992). Thus, for the correct set of potentials:

\[
E(S_i, \Gamma_i) < E(S_i, \Gamma_D) ,
\]

for all the decoy structures, \( \Gamma_D \), obtained by threading. Therefore, for each sequence in the training set, one obtains an array of inequalities. Due to the finite number of proteins in the training set, the whole ensemble of inequalities will be satisfied by more than a single set of potentials. Indeed, there will typically be a whole region of points in parameter space each corresponding to a set of potentials consistent with" inequalities (6). The optimal solution is attained by simultaneously maximizing the stability gap for all proteins in the set. The stability gap is defined as the smallest energy difference between a protein’s native state and one of the decoy conformations. The optimal stability requirement implies that the following inequalities should hold simultaneously for each training protein

\[
E(S_i, \Gamma_D) - E(S_i, \Gamma_i) > c \quad \forall \quad \Gamma_D ,
\]

where \( c \) is a positive quantity to be made as large as possible, the \( \Gamma_D \)'s belong to the set of decoy conformations and the energy interactions satisfy to (3).

The function \( f \) in the denominator of (7) is a function of the structural distance between \( \Gamma_D \) and \( \Gamma_i \). This serves the purpose of making inequalities (7) more stringent when mounting \( S_i \) on structurally dissimilar conformations. We used three different trial functions for \( f \):

\[
\begin{align*}
    f_1(x) &= 1 , \\
    f_2(x) &= x , \\
    f_3(x) &= x^2 .
\end{align*}
\]

For the distance function \( D(\Gamma, \Gamma') \), appearing in eq. (11), we used the Euclidean distance in contact-map space:

\[
D(\Gamma, \Gamma') \equiv \left[ \sum_{i=1}^{N} \sum_{j=1}^{N} (\Delta(r_{ij}) - \Delta(r'_{ij}))^2 \right]^{1/2} / (N-1)(N-2)/2 .
\]

\( D(\Gamma, \Gamma') \) can be viewed as a close relative in terms of contact maps of the standard distance root mean square deviation (DRMSD) but related to our definition of the energy functional.

By threading the training sequences on longer structures, we generated the whole set of inequalities (6). Each of these identifies a hyperplane in parameter space dividing space into two semi-infinite regions; one of which is compatible with the inequality and contains the physical set of parameters (van Mourik et al. 1998). When more inequalities are used, the physical region containing the correct parameters reduces to the intersection of all physical hyperspaces. Eventually, the region reduces to a small, convex cell (not necessarily closed) whose walls are determined by a number of inequalities of the order of the dimension of parameter space.

The optimal point in the cell is found by using perceptron strategy, as described in Section 4. This procedure has been shown to converge to the true potential when used in exact models where rigorous test are available (van Mourik et al. 1998) (Clementi, Maritan & Banavar 1998). It is also possible that parametrizations (4) or (3) may not be sufficient to guarantee that a solution to inequalities (6) exists. Indeed, if the decoys structures are very competitive with the native structures, three or further body interactions might be necessary to solve inequality (6) consistently (Vendruscolo et al. 1999).
This procedure differs significantly from the one of Maiorov & Crippen (1992) where the parameters were determined in a sub-optimal manner.

3 Results and discussion

We succeeded in finding an optimal solution to the different systems of inequalities (7): the optimal parameters obtained for Hamiltonians (8) and \( f = 1 \) are given in Table 1.

We found that only a tiny fraction of all inequalities (7) determine the optimal stability solution (more or less 100 out of 1551196 according to the \( f \) used or whether solvation term is present). It is important to ensure that the optimal solution does not fluctuate wildly when stringent inequalities are added or removed. To check this, we eliminated the 100 most stringent inequalities. Even though this completely replaces the walls of the physical cell, the new optimal solution slightly differed from the first one: representing the parameters in a 230-dimensional vector space the two vectors were only 15° apart. Such a degree of correlation is significant because the expected angle between two uncorrelated vectors in a space of \( \approx 200 \) dimensions is about 90° ± 4°. This gives confidence in the robustness of the procedure and the statistics of the training set.

The optimal parameters extracted with different trial forms of \( f \) in (7) were also closely correlated. As summarized in Table 2, their relative angle was always less than 15°. On the contrary, sub-optimal vectors, for which inequalities (7) are satisfied for \( c \approx 0 \) (in which case the detailed form of \( f \) is not relevant) form, on average, an angle of 50° with the optimal solution. This fact underscores the importance of introducing an extremal criterion when maximizing (7).

The extracted solvation parameters showed a very good correlation (0.67 correlation coefficient) with the hydrophobicity scales as given by Creighton (1993). As shown in Fig. 4.2 the agreement is quite good except, perhaps, for proline. The discrepancy with proline finds a natural explanation within the scheme that we used. In fact, while the hydrophobicity scales in Fig. 1.2 relate to the propensities of individual, isolated amino acids, the solvation parameter reflects also their structural functionality in a peptide context. In fact, because the prolines are typically located in loop regions, they appear to have an effective hydrophilic propensity larger than their bare value.

Finally, we carried out a stringent validation of the extracted potentials by performing a blind ground-state recognition on a test set. The test set (see Table 3) was comprised of proteins taken from those used in Miyazawa & Jernigan (1996) and chosen so that they would meet some of the criteria used to select the training set (see section 4). We deliberately introduced proteins with hetero groups, low degree of compactness and also pairs with high structural homology. In all cases we ensured that no protein in the test set had a significant degree of structural homology with those in the training one.

We took, in turn, the sequences of the test set and threaded them on structures in the set with equal or longer length. Hence, we checked whether using the optimal potential parameters of Table 1, the true native state was recognized as the lowest energy one. Indeed, this turned out to be the case for all but 6 proteins. No higher success rate was found on using some other known sets of potentials consistent with the form of our Hamiltonian.

Another relevant quantity related to the performance of the algorithm is given by the number of wrongly satisfied inequalities of type (7) for the test set. This quantity shows a much higher degree of variability than the number of correctly identified ground states and is given in column 3 of Table 4.2. It can be seen that the optimal parameters extracted with the solvent and \( f = 1 \) perform far better than those without the solvent and previously extracted ones. It also appears that, enforcing optimality provides a dramatic reduction of wrong inequalities compared to the sub-optimal cases. This provides a sound a posteriori justification for the optimal extraction procedure as well as giving confidence in the parameters.

The few cases where the extracted potentials fail
are due to one of the following situations: a) the native protein is not too compact or b) it contains stabilizing hetero groups. Situations in which a highly homologous structure has a lower energy score than the native one are not deemed as errors. A typical energy/structural distance plot is shown in Fig. 5. It is apparent that homologous structures have energies similar to that of native conformations, while distant structures lie higher in energy. Some differences in the performance were observed for the sets of 210 and 230 potentials. While the latter only fail to recognize native states containing heme groups etc., the former occasionally fail to recognize the native states with no atypical feature (e.g. interleukin-4, 1rcb).

For proteins with heme groups, several structures score better: they usually present a smaller number of contacts than the native structure being less compact than the native state. This is possibly related to the presence of proline in an unusually buried position, namely the heme pocket. In fact, due to the high effective solvation term assigned to proline, the native structure is penalized with respect to decoy ones where it is confined in more solvent-exposed positions.

An interesting case where the failure relates to a non-compact protein, is given by 3wrp, for which several better scoring decoys exist. The explanation lies in the fact that 3wrp is always found as a dimer: the side of the protein binding its counterpart has non-polar surface residues usually in contact with non-polar residues on the other dimer, which is not accounted for by our procedure.

Nevertheless, the algorithm appears to work in other instances of non-compact conformations such as troponin c (4tnl) and calmodulin (1c1l) and on some cytochrome-c as 1ccr or 1yeb, showing that the optimization procedure succeeds in extracting a potential with a wider applicability range than that given by the folds used in the training set.

Over 15 different pairs of homologous structures (contact map distance less than 0.1), the energy functional is able to rank the true native state as the lowest in just 8 cases. In the other cases the native state attains an energy value slightly higher than the homologous one. As expected, the simple contact potential cannot distinguish the native state among very similar structures but it consistently assigns similar value of energy to similar conformations according to the degree of similarity (see Fig. 5).

It is important to note that there is a well-defined trend for the protein ground-state energies as a function of protein length. Deviations from this typical trend could be used to assess the reliability of the predicted fold of a sequence with unknown structure (Fig. 6).

4 Methods

4.1 Protein data sets

We selected 142 protein structures from PDB [Bernstein, Koetzle, Williams, Meyer, Brice Rodgers, Kennard, Shimanouchi & Tasaki 1977], listed in Tab. 1, with lengths varying from 36 to 823, following criteria very similar to Maiorov & Crippen [1992]. For each reference protein, we built a set of alternative conformations by threading its sequence on all the other structures in the set with a greater or equal number of amino acids. As explained in Jones et al. [1992], threading a sequence of $L$ amino acids on a structure, $I'$, of length $L' > L$, involves mounting the sequence on all the $(L' - L + 1)$ segments (of contiguous amino acids) taken from $I'$. This procedure assigns the contact map of the threaded segment to the threading sequence. The inherited contact map is used to calculate the energy of the sequence in the alternative, threaded, conformation, and compared with the energy in its native state, which is required to be the global energy minimum.

Only single chain structures have been selected in order to avoid the occurrence of interchain contacts between amino acids, that are not detected by our procedure and that could cause the stabilization of hydrophobic residues on a protein’s surface. Considering multiple chain structures would have introduced spurious effects in the extracted potentials, since inter-chain contacts would not be present in threaded conformations. For simplicity, however, we decided to retain proteins which may be found in polymeric forms. Because the presence of large hetero groups can distort the usual geometry of dihedral
angles between amino acids and cannot be treated in a simple way by a pairwise potential, we discarded protein structures with high percentages of non-water HETATM records in their PDB files (like HEM or CPS groups).

We used the classification of 3-D protein structures SCOP \cite{Murzin:1995aa} to select proteins spanning a wide range of different three dimensional folds: no pair of proteins in our training set belong to the same SCOP family. Furthermore, we have included only proteins in the first 4 SCOP classes: all-\(\alpha\), all-\(\beta\), \(\alpha/\beta\) and \(\alpha + \beta\) proteins. Cell membrane or surface proteins and very short peptide chains are excluded because usually they are not stabilized by just amino acid interactions but by some external factors, such as the hydrophobic environment, metal ligands, heme groups etc. No unresolved backbone atoms inside a chain are allowed; disordered or unresolved terminal backbone atoms are eliminated.

We also disregarded proteins that were not typically compact: because there is a clear dependence from the interaction parameters between amino acids on chain length, we rejected from our training set proteins with too large a radius of gyration or with significantly fewer contacts than expected for their length. The rejection was based using the quantitative procedures discussed in \cite{Maiorov:1992aa}.

\subsection{Optimal Stability Perceptron}

It is convenient to recast expression (\ref{eqn:interaction}) so that the dependence from the interaction parameters between amino acid types \(A\) and \(A'\), \(\epsilon(A, A')\) appears explicitly,

\begin{equation}
\sum_{\alpha \leq A'} \frac{\epsilon(A, A')}{|\epsilon|} \cdot \frac{n_{S, \Gamma_D}(A, A') - n_{S, \Gamma}(A, A')}{f(D(\Gamma_D, \Gamma_i))} > \epsilon \equiv \{\epsilon(1, 1), \epsilon(1, 2), \ldots, \epsilon(20, 20)\} .
\end{equation}

and likewise for the vector

\begin{equation}
\vec{\epsilon} \cdot \vec{N}_{S, \Gamma, \Gamma_D} > c .
\end{equation}

A formally equivalent expression is obtained when using 230 parameters as in eq. (3).

Expression (12) leads to a geometrically appealing interpretation of the stability requirement. The optimal stability is reached when the interaction vector has the largest possible inner product with all the \(\vec{N}_{S, \Gamma, \Gamma_D}\) vectors, also termed “patterns”, originated from the training set. A rigorous solution for this geometrical problem was given by \cite{Krauth:1987aa} who suggested an iterative procedure called optimal stability perceptron.

The procedure is the following. Starting from a random (or an otherwise assigned) set of interactions satisfying the norm constraint (\ref{eqn:norm}), the stability score of all inequalities is computed. Then, the potentials are updated so to increase the stability of the lowest scoring inequality. This is done by adding to the original potentials vector, \(\vec{\epsilon}\), a small term proportional to the worst scoring pattern (see Fig. \ref{fig:perceptron}):\footnote{\cite{Maiorov:1992aa}}

\begin{equation}
\forall A, A' \quad \epsilon(A, A') \rightarrow \epsilon(A, A') + \frac{1}{d} N_{S, \Gamma, \Gamma'}(A, A') ,
\end{equation}

where \(d\) is the dimension of the parameter space (210 or 230). Then, the inequalities are re-computed with the updated interaction parameters. The lowest scoring one is identified again and a new update of \(\vec{\epsilon}\) is carried out. Note that the update (16) will typically change the norm of \(\vec{\epsilon}\). The unit norm (see eqn. 3) can be conveniently enforced after convergence has been achieved.

While convergence is guaranteed to be reached in a finite number of steps, the time required for each
iteration grows linearly with the number of inequalities. In our case, we typically dealt with $\approx 10^6$ inequalities, and convergence sometimes required several thousand iterations (each taking few seconds of CPU time). Hence, we devised a scheme to speed up convergence based on the observation that the stability variation due to the change of parameters is proportional to the distance between the inequality point and the parameter direction,

$$\Delta s = \vec{N} \cdot \Delta \vec{\epsilon} = \vec{N}_{\perp} \cdot \Delta \vec{\epsilon}_{\perp} + \vec{N}_{\parallel} \cdot \Delta \vec{\epsilon}_{\parallel} \propto \begin{cases} \geq 0 & \text{if } a \neq 0, \\ \geq 0 & \text{if } b \geq 0. \end{cases} \quad (17)$$

This implies that inequality vectors far from the parameter direction will get the largest score variations (positive or negative) and so they are more likely to become the lowest scoring ones. Accordingly, the standard perceptron procedure was run until reaching a sub-optimal value for the stability threshold, $c > 0$; this typically needed 300 iterations. Then we temporarily restricted the updating procedure to those inequalities lying outside a cone with axis along the parameter direction and vertex at distance larger than $c$ from origin (see Fig. 3). The cone width was determined to limit the number of inequalities to less than 20000 (Fig. 3). In this way we had to deal with $10^4$ inequalities that are 2 orders of magnitude smaller than the original ones, thus decreasing enormously the CPU time needed for optimization. Furthermore, after convergence, the neglected inequalities are found to be satisfied well above the optimal stability threshold $c_{\text{max}}$, thus justifying the numerical shortcut.

We conclude by remarking that, if the relative correction to $\epsilon(A, A')$ in eq. (14) is too large, this may result in a slowing down of the convergence. This difficulty can be readily circumvented by increasing the size of $\vec{\epsilon}$ by an order of magnitude. It was typically necessary to repeat this “inflation” procedure 3-4 times during each run towards convergence (see Fig. 3). This was sufficient to reach optimal convergence to the solution: $\Delta c/c < 10^{-3}$.

**References**

Anfinsen, C. (1973). Principles that govern the folding of protein chains, *Science* **181**: 223–239.

Bauer, A. & Beyer, A. (1994). An improved pair potential to recognize native protein folds, *Proteins: Structure Function and Genetics* **18**: 254–261.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E. J., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures, *J. Mol. Biol.* **112**: 535–542.

Bowie, J. & Eisenberg, D. (1994). An evolutionary approach to folding small $\alpha$-helical proteins that uses sequence information and an empirical guiding fitness function, *Proc. Natl. Acad. Sci. USA* **91**: 4436–4440.

Bowie, J., Lütthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure, *Science* **253**: 164–170.

Branden, C. & Tooze, J. (1991). *Introduction to protein structure*, Garland Publishing, New York.

Bryant, S. & Lawrence, C. (1993). An empirical energy function for threading protein sequence through the folding motif, *Proteins: Structure Function and Genetics* **16**: 92–112.

Casari, G. & Sippl, M. (1992). Structure-derived hydrophobic potential: hydrophobic potential derived from x-ray structures of globular proteins is able to identify native folds, *J. Mol. Biol.* **224**: 725–732.

Acknowledgments. We acknowledge support from INFM, NASA, NATO and the Petroleum Research Fund administered by the American Chemical Society. We thank C. Clementi, R. Dima, L. Guidoni, S. Piana, A. Rossi, F. Seno and J. Van Mourik for useful discussions.
Clementi, C., Maritan, A. & Banavar, J. (1998). Folding, design and determination of interaction potentials using off-lattice dynamics of model heteropolymers, Phys. Rev. Lett. 81: 3287–3290.

Creighton, T. (1993). Proteins, structure and molecular properties, second edn, W.H.Freeman and Company, New York.

Dahiyat, B. & Mayo, S. (1997). De novo protein design: fully automated sequence selection, Science 278(5335): 82–87.

Dandekar, T. & Argos, P. (1994). Folding the main chain of small proteins with genetic algorithm, J. Mol. Biol. 236: 844–861.

Deutsch, J. & Kurosky, T. (1996). New algorithm for protein design, Phys. Rev. Lett. 76: 323–326.

Dima, R., Banavar, J., Maritan, A., Micheletti, C. & Settanni, G. (1999). Extraction of interaction potentials between amino acids from native protein structures. In preparation.

Duan, Y. & Kollman, P. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in water solution, Science 282: 740–744.

Giugliarelli, Maritan, A., Micheletti, C. & Banavar, J. (1999). In preparation.

Godzik, A., Kolinsky, A. & Skolnick, J. (1992). A general method for prediction of the three dimensional structure and folding pathway of globular proteins: application to designed globular proteins, J. Chem. Phys. 98: 7420–7433.

Kolinsky, A., Godzik, A. & Skolnick, J. (1994). Monte carlo simulations of protein folding. 1. lattice model and interaction scheme, Proteins: Structure Function and Genetics 18: 338–352.

Krauth, W. & Mezard, M. (1987). Learning algorithms with optimal stability in neural networks, J. Phys. A 20: L745–L752.

Levitt, M. (1976). A simplified representation of protein conformation for rapid simulation of protein folding, J. Mol. Biol. 104: 59–107.

Levitt, M. (1983). Protein folding by restrained energy minimization and molecular dynamics, J. Mol. Biol. 170: 723–764.

Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins, J. Mol. Biol. 227: 876–888.

Maiorov, V. N. & Crippen, G. M. (1994). Learning about protein folding via potential functions, Proteins: Structure Function and Genetics 20: 173–176.

Micheletti, C., Seno, F., Maritan, A. & Banavar, J. (1998). Design of proteins with hydrophobic and polar amino acids, Proteins: Structure Function and Genetics 32: 80.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal-structures - quasi-chemical approximation, Macromolecules 18: 534–552.

Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density
term, for simulation and threading, J. Mol. Biol. 256: 623–644.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). Scop: A structural classification of proteins database for investigation of sequences and structures, J. Mol. Biol. 247: 536–540.

Ouzounis, C., Sander, C., Sharf, M. & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from three-dimensional structure, J. Mol. Biol. 232: 805–825.

Park, B. & Levitt, M. (1996). Energy functions that discriminate x-ray and near-native folds from well-constructed decoys, J. Mol. Biol. 258: 367–392.

Seno, F., Maritan, A. & Banavar, J. (1998). Interaction potentials for protein folding, Proteins: Structure Function and Genetics 30: 224–248.

Seno, F., Micheletti, C., Maritan, A. & Banavar, J. (1998). Variational approach to protein design and extraction of interaction potentials, Phys. Rev. Lett. 81: 2172.

Sippl, M. (1995). Knowledge based potentials for proteins, Curr. Opin. Struct. Biol. 5: 229–235.

Skolnick, J., Kolinsky, A., Brooks, C., Godzik, A. & Rey, A. (1993). A method for predicting protein structure from sequence, Curr. Biol. 3: 414–423.

Socci, N. D. & Onuchic, J. N. (1994). Folding kinetics of proteinlike heteropolymers, J. Chem. Phys. 101: 1519.

Srinivasan, R. & Rose, G. D. (1995). Linus: a hierarchic procedure to predict the fold of a protein, Proteins: Structure Function and Genetics 22: 81–99.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms, Protein Sci. 2: 762–785.

Sun, S., Brem, R., Chan, R. & Dill, K. (1995). Designing amino acid sequences to fold with good hydrophobic cores, Protein Eng. 8: 1205–1213.

Thomas, P. & Dill, K. (1996). An iterative method for extracting energy-like quantities from protein structures, Proc. Natl. Acad. Sci. USA 93: 11628–11633.

van Gunsteren, W. F. (1989). Computer Simulation of biomolecular systems, ESCOM Science Publishers, B.V. Leiden.

van Mourik, J., Clementi, C., Maritan, A., Seno, F. & Banavar, J. (1998). Determination of interaction potentials of amino acids from native protein structures: test on simple lattice models, E-print cond-mat/9801137.

Vendruscolo, M., Najmanovich, R. & Domany, E. (1999). Protein folding in contact map space, Phys. Rev. Lett. 82: 656–659.

Wallqvist, A. & Ullner, M. (1994). A simplified amino acid potential for use in structure prediction of proteins, Proteins: Structure Function and Genetics 18: 267–280.

Wodak, S. & Rooman, M. (1993). Generating and testing protein folds, Curr. Opin. Struct. Biol. 3: 247–259.

Wolynes, P., Onuchic, J. & Thirumalai, D. (1995). Navigating the folding routes, Science 267: 1619–1620.
Table 1: Interaction parameters extracted by using energy functional

|   | A     | B     | C     | D     | E     | F     | G     | H     | I     | J     | K     | L     | M     | N     | O     | P     | Q     | R     | S     | T     | U     | V     | W     | X     | Y     | Z     |
|---|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| A | 0.0491| -0.1560| -0.0189| 0.0027| -0.0130| 0.0137| 0.0072| 0.1696| -0.0196| -0.0241| 0.0046| 0.0127| -0.1752| 0.0491| 0.1343| -0.0769| 0.0491| 0.0376| 0.0095| 0.0036| 0.0140| 0.1203| 0.0514| 0.0765| 0.0654| 0.0765|
| B | 0.0222| 0.0007| 0.0007| 0.0007| 0.0024| 0.0220| 0.0765| 0.0007| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020|
| C | 0.0189| 0.0027| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007|
| D | 0.0027| -0.0130| 0.0137| 0.0072| 0.1696| -0.0196| -0.0241| 0.0046| 0.0127| -0.1752| 0.0491| 0.1343| -0.0769| 0.0491| 0.0376| 0.0095| 0.0036| 0.0140| 0.1203| 0.0514| 0.0765| 0.0654| 0.0765|
| E | 0.0024| 0.0220| 0.0765| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007| 0.0007|
| F | 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126| 0.0020| 0.0126|
| G | 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020|
| H | 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020|
| I | 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020| 0.0020|
| Solvent          | \( f = 1 \) | \( f = x \) | \( f = x^2 \) |
|------------------|-------------|--------------|---------------|
| \( f = 1 \)     |             | 10.1°        | 13.5°         |
| \( f = x \)     | 6.67°       |              |               |
| \( f = x^2 \)   |             |              |               |
| Non-optimal      | (average) 54°| (average) 55°|               |

| No-Solvent       | \( f = 1 \) | \( f = x \) | \( f = x^2 \) |
|------------------|-------------|--------------|---------------|
| \( f = 1 \)     |             | 8.5°         | 15.9°         |
| \( f = x \)     | 10.0°       |              |               |
| \( f = x^2 \)   |             |              |               |

Table 2: Angles formed by the optimal vectors for various forms of the Hamiltonian and \( f \) (see eqns. 2, 3, and 7).
Table 3. Proteins used in the test set. The symbol † denotes instances where the better scoring structure is homologous to the target protein, while * marks non compact native state. Hg has been used only as structural template.
| Solvent         | Score on Test | Score on Training |
|-----------------|---------------|-------------------|
|                 | Not Rec. Str. | Unsat. Ineq.      |
| $f = 1$         | 5             | 25                |
| $f = x$         | 5             | 33                |
| $f = x^2$       | 5             | 36                |
| Non-Optimal     | 5.25          | 75.7              |
|                 |               |                   |
| $f = 1$         | 6             | 118               |
| $f = x$         | 6             | 200               |
| $f = x^2$       | 6             | 254               |
| DBMS            | 7             | 51                |
| KGS             | 6             | 452               |
| MC              | 6             | 48                |
| DBMS            | 13            | 1091              |
| KGS             | 22            | 9789              |
| MC              | 20            | 1826              |

Table 4: Performance of the potentials extracted in this work and other known sets. The second column gives the number of unrecognised native states among the 78 ones of Table 2. The associated number of violated inequalities (against a total of 444199) is given in column 3. The acronyms for the alternative potentials refer to: DBMS=(Dima et al. 1999), KGS=(Kolinsky et al. 1993), MC=(Maiorov & Crippen 1994). The last part of the table shows the scores of the alternative potentials applied to our training set of 142 proteins generating 1551196 inequalities (on which our potentials, by definition, scores 100% success).
| Prot. Code | Length | SCOP Classification | No. Decoys |
|------------|--------|---------------------|------------|
| 1vii       | 36     | 1001014001001001    | 001        | 25461      |
| 1erd       | 40     | 1001010001001001    | 002        | 24896      |
| 1pru       | 56     | 1001030001001001    | 001        | 22655      |
| 1fxd       | 58     | 1004033001001001    | 001        | 22376      |
| 1igl       | 61     | 1004012001001001    | 001        | 21961      |
| 1orc       | 64     | 1001030001001002    | 005        | 21549      |
| 1sap       | 66     | 1004009001001001    | 002        | 21276      |
| 1mit       | 67     | 1004022001001001    | 003        | 21140      |
| 1utg       | 69     | 1001072001001001    | 001        | 20871      |
| 1ail       | 70     | 1001015001001001    | 001        | 20737      |
| 1hoe       | 74     | 1002004001001001    | 001        | 20208      |
| 1kjs       | 74     | 1001040001001001    | 001        | 20208      |
| 1ubi       | 74     | 1004012002001001    | 001        | 20208      |
| 1hyp       | 75     | 1001042001001001    | 001        | 20076      |
| 5icb       | 75     | 1001034001001001    | 001        | 20076      |
| 1fow       | 76     | 1001004004001001    | 001        | 19947      |
| 1tif       | 76     | 1004012006001001    | 001        | 19947      |
| 1tnl       | 76     | 1001006001001001    | 001        | 19947      |
| 1acp       | 77     | 1001026001001001    | 001        | 19820      |
| 1hdj       | 77     | 1001002002001001    | 001        | 19820      |
| 1iba       | 77     | 1004053001001001    | 001        | 19820      |
| 1vcc       | 77     | 1004067001001001    | 001        | 19820      |
| 1coo       | 81     | 1001032001001001    | 001        | 19336      |
| 1cei       | 84     | 1001026002001001    | 001        | 18978      |
| 1ngr       | 84     | 1001062001001001    | 001        | 18978      |
| 1opd       | 85     | 1004052001001001    | 003        | 18859      |
| 1fna       | 90     | 1002001002001001    | 002        | 18278      |
| 1hqj       | 90     | 1004079001001001    | 001        | 18278      |
| 1who       | 94     | 1002006003001001    | 001        | 17820      |
| 1pdr       | 96     | 1002023001001001    | 001        | 17593      |
| 1beo       | 98     | 1001096001001001    | 001        | 17368      |
| 1tul       | 101    | 1002060004001001    | 001        | 17034      |
| 9rnt       | 104    | 1004001001001001    | 003        | 16703      |
| 1aaw       | 105    | 1002005001001001    | 001        | 16593      |
| 1erv       | 105    | 1003033001001001    | 004        | 16593      |
| 1jpc       | 108    | 1002054001001001    | 001        | 16270      |
| 1kum       | 108    | 1002003001001001    | 005        | 16270      |
| 1roo       | 108    | 1001034001001004    | 001        | 16270      |
| 3ssi       | 108    | 1004044001001001    | 002        | 16270      |
| 2mcm       | 112    | 1002001006001001    | 001        | 15854      |
| 1mai       | 118    | 1002037001001001    | 001        | 15241      |
| 1poa       | 118    | 1001025001002001    | 001        | 15241      |
| 1whi       | 122    | 1002025001001001    | 001        | 14839      |
| 1yua       | 122    | 1004067001001002    | 001        | 14839      |
| Protein | ID   | Accession           | Type | Description |
|---------|------|---------------------|------|-------------|
| 7rsa    | 124  | 1004004001001       | 001  | 14641       |
| 2phy    | 125  | 1004061002001       | 001  | 14543       |
| 1bfg    | 126  | 1002028001001       | 001  | 14446       |
| 3chy    | 128  | 1003013002001       | 001  | 14255       |
| 1pdo    | 129  | 1003040001001       | 001  | 14160       |
| 1tum    | 129  | 1004062001001       | 001  | 14160       |
| 1ife    | 131  | 1002041001002       | 002  | 13974       |
| 1kuh    | 131  | 1004050001001       | 001  | 13974       |
| 1lis    | 131  | 1001017001001       | 001  | 13974       |
| 1rsy    | 132  | 1002060001002       | 001  | 13882       |
| 1cof    | 135  | 1004060001002       | 001  | 13617       |
| 2end    | 137  | 1001016001001       | 001  | 13442       |
| 5nul    | 138  | 1003013004001       | 006  | 13355       |
| 2sns    | 140  | 1002026001001       | 001  | 13184       |
| 1icl    | 141  | 1002019001003       | 004  | 13099       |
| 1lba    | 145  | 1004064001001       | 001  | 12766       |
| 1pkp    | 145  | 1004011001001       | 002  | 12766       |
| 1vsl    | 145  | 1003041003002       | 001  | 12766       |
| 1npk    | 150  | 1004033006001       | 002  | 12363       |
| 1irp    | 153  | 1002028001002       | 003  | 12125       |
| 2rn2    | 155  | 1003041003001       | 001  | 11968       |
| 1vhh    | 157  | 1004034001002       | 001  | 11813       |
| 1gpr    | 158  | 1002059003001       | 001  | 11736       |
| 1ra9    | 159  | 1003053001001       | 001  | 11660       |
| 119f    | 162  | 1004020001003       | 001  | 11437       |
| 2cpl    | 164  | 1002043001001       | 001  | 11290       |
| 1sfe    | 165  | 1001004002001       | 001  | 11217       |
| 1wba    | 171  | 1002028003001       | 001  | 10790       |
| 2fha    | 171  | 1001024001001       | 003  | 10790       |
| 1amm    | 173  | 1002090001001       | 001  | 10650       |
| 2prd    | 173  | 1002026005001       | 003  | 10650       |
| 1ldo    | 184  | 1003045001001       | 002  | 9911        |
| 133f    | 185  | 1004002001004       | 001  | 9844        |
| 1xnb    | 185  | 1002019001008       | 001  | 9844        |
| 1knb    | 186  | 1002016001001       | 001  | 9778        |
| 1kid    | 192  | 100305003001       | 001  | 9399        |
| 1cex    | 197  | 1003013007001       | 001  | 9088        |
| 1chd    | 198  | 1003027001001       | 001  | 9026        |
| 1fua    | 206  | 1003055001001       | 001  | 8545        |
| 1thv    | 207  | 1002018001001       | 001  | 8485        |
| 2abk    | 211  | 1001066001001       | 001  | 8252        |
| 1ah6    | 213  | 1004068001001       | 001  | 8137        |
| 1lbu    | 213  | 1001019001001       | 001  | 8137        |
| 3cla    | 213  | 1003030001001       | 001  | 8137        |
| 2ayh    | 214  | 1002019001002       | 002  | 8080        |
| 1gpc    | 217  | 1002026004007       | 003  | 7920        |
| 1akz | 223 | 1003011001001 | 001 | 7607 |
| 1dad | 224 | 1003025001005 | 001 | 7555 |
| 1aol | 227 | 1002015001001 | 001 | 7404 |
| 1cby | 227 | 1004058001001 | 001 | 7404 |
| 1lbd | 238 | 1001087001001 | 001 | 6874 |
| 2baa | 243 | 100402001001 | 001 | 6638 |
| 1mjrj | 247 | 1004094001001 | 001 | 6453 |
| 3fib | 248 | 1004098001001 | 001 | 6407 |
| 1plq | 258 | 1004076001002 | 001 | 5966 |
| 2cba | 258 | 1002050001001 | 002 | 5966 |
| 1arb | 262 | 1002031001001 | 001 | 5796 |
| 1ako | 268 | 1004066001001 | 001 | 5549 |
| 2dri | 271 | 1003072001001 | 001 | 5428 |
| 1tml | 286 | 1003002001001 | 001 | 4842 |
| 1han | 287 | 1004020001003 | 002 | 4803 |
| 1mar | 289 | 1003001001005 | 002 | 4728 |
| 1amp | 290 | 1003052003004 | 001 | 4691 |
| 1ctt | 294 | 1003075001001 | 001 | 4550 |
| 2etc | 307 | 1003052003001 | 001 | 4107 |
| 1ede | 310 | 1003001001003 | 002 | 4007 |
| 1pgs | 311 | 1002011001001 | 001 | 3974 |
| 1lds | 315 | 1003001001001 | 002 | 3849 |
| 1hgyt | 316 | 1001053001001 | 002 | 3818 |
| 1tca | 317 | 1003050001007 | 001 | 3788 |
| 1pot | 321 | 1003073001001 | 011 | 3675 |
| 1axn | 323 | 1001051001001 | 001 | 3620 |
| 1dxy | 329 | 1003013009001 | 002 | 3463 |
| 1nif | 333 | 1002055001003 | 001 | 3362 |
| 1rpa | 341 | 1003043001002 | 001 | 3169 |
| 1uby | 348 | 1001091001001 | 001 | 3007 |
| 1ldk | 359 | 1002056001002 | 001 | 2764 |
| 1eur | 360 | 1002045001001 | 004 | 2742 |
| 1cem | 363 | 1001073001002 | 001 | 2681 |
| 1pud | 372 | 1003001001001 | 001 | 2509 |
| 1kaz | 377 | 1003041001001 | 001 | 2418 |
| 1edg | 380 | 1003001001001 | 001 | 2366 |
| 1php | 394 | 1003056001001 | 003 | 2141 |
| 1phc | 405 | 1001075001001 | 001 | 1975 |
| 1uae | 417 | 1004035001001 | 001 | 1806 |
| 1gnd | 430 | 1003004001003 | 001 | 1636 |
| 1csh | 433 | 1001074001001 | 001 | 1599 |
| 1pmi | 440 | 1002058002001 | 001 | 1521 |
| 1gcb | 452 | 1004003001001 | 008 | 1400 |
| 2bnh | 456 | 1003007001001 | 001 | 1363 |
| 3grs | 461 | 1003004001004 | 001 | 1322 |
| 1gai | 471 | 1001073001001 | 001 | 1251 |
|     | Code | Length | Structure1 | Structure2 | Structure3 | Structure4 |
|-----|------|--------|------------|------------|------------|------------|
| lam | 1lam | 484    | 1003036001001 | 001 | 1172 |
|    | 1vnc | 576    | 1001080001001 | 001 | 711 |
|    | 1ciy | 577    | 1002013001002 | 002 | 706 |
|    | 1amj | 753    | 1003050002001 | 002 | 177 |
|    | 1gpb | 823    | 1003068001002 | 001 | 36 |
|    | 1qba | 858    | 1002001001005 | 002 | 0 |

Table 5: List of “training proteins” used to extract interaction potentials.

3The longest protein in the set, 1qba, was used only as a structural template
Figure 1: The extracted solvation parameters, $\epsilon(0, A)$ (see eqn. (3)) versus standard hydrophobicity values (Creighton 1993, p.154).

Figure 2: Schematic representation of a typical perceptron update in a two-dimensional space. Inequalities are represented by vectors connecting the origin to the points. At iteration $n$, The stability, $c$, of ineq. (7) is the smallest inner product between the parameter vector $\vec{\epsilon}_n$ and each of the inequalities. In case (a) $c$ is given by $\vec{\epsilon} \cdot \vec{N}$ and $\vec{\epsilon}_n$ accordingly acquires a small component parallel to $\vec{N}$. An equilibrium situation is shown in (b). Successive updates cause $\vec{\epsilon}$ to bounce on either side of the equilibrium direction. The latter is reached in a finite time, because the relative size of the added component decreases with the number of iterations.
Figure 3: After a sufficient number of iterations, successive perceptron updates will not change appreciably $\vec{\epsilon}$. To speed up convergence, it is convenient to temporarily retain only those inequalities (points in parameter space) lying outside a cone with axis along $\vec{\epsilon}$ and suitable vertex and width (see text). The edge of the cone is visible in this figure, where only the retained inequalities are shown.

Figure 4: Perceptron stability as a function of the number of iterations. The discontinuities are associated with the “inflation” of $\vec{\epsilon}$ used to speed up convergence (see text).

Figure 5: Energies of the protein sequence 1lz3 when threaded on decoy structures against structural dissimilarity. Very low energies are observed when threading on homologous conformations (in particular protein 1321).

Figure 6: When using the extracted parameters of Table 1, the native state energy of proteins shows an approximately linear behaviour as a function of their length. Points refer to proteins for both the training and test sets. Proteins with less than 200 amino acids and atypical compactness present significant deviations from the average trend.