

On the Transferability of Folding and Threading Potentials and Sequence-Independent Filters for Protein Folding Simulations

Rafal Adamczak¹ and Jaroslaw Meller^{1,2 *}

¹Division of Biomedical Informatics
Children's Hospital Research Foundation
3333 Burnet Avenue, Cincinnati, OH 45229, USA

²Department of Informatics,
Nicholas Copernicus University,
87-100 Toruń, Poland

Dedication:

This paper is dedicated to the memory of Professor Brian Wybourne, my scientific advisor and a truly inspiring mentor. JM

* Corresponding author

Phone (513) 636-0270

Fax (513) 636-2056

e-mail: jmeller@chmcc.org

Running title: "Sequence-Independent Filters for Protein Folding"

Keywords: Pair Correlation Function, Contact Order, Decoys, Contact Potentials,
Linear Programming, Maximum Feasibility

Abstract

Significant progress has recently been made in *de novo* protein structure prediction. The Rosetta method by Baker and colleagues, which is based on the idea of assembling putative models from a library of k-mer fragments derived from known three-dimensional protein structures, proved to be particularly successful. A critical component of the Rosetta approach are various sequence dependent as well as sequence independent measures that are used to rank alternative models and to enhance sampling of native-like conformations. In the present work we revisit several sequence independent filters that have been used before to enhance the discrimination of native and native-like structures from misfolded structures, such as the overall compactness of the structure and its contact order. We also propose a novel sequence independent filter, based on the shape of the mean inter-residue radial distribution function. Using the Rosetta, Park-Levitt and CASP4 sets of decoys we show that sequence independent filters are in fact more successful in distinguishing native structures in Rosetta and CASP4 tests than commonly used knowledge-based pairwise potentials. The latter are typically designed to distinguish native structures in a population of well-folded alternatives, and they fail to discriminate between native-like and non-physically packed misfolded structures from Rosetta simulations. Moreover, a rigorous attempt to optimize pairwise potentials for recognition of homologous structures in threading by using Linear Programming approach leads to further deterioration of performance in terms of recognition of native structures from the Rosetta set. Our findings shed light onto the success of tailored scoring functions used in

the Rosetta protocol and provide support for explicit inclusion of both sequence dependent and sequence independent measures in the design of scoring functions. A Web server that enables ranking of decoy structures according to sequence independent filters considered here is available at <http://sift.chmcc.org>.

I. Introduction

While predicting three-dimensional structure of a protein from its amino acid sequence remains one of the central challenges in computational biology, significant progress has been made in protein structure prediction in the last several years. Reliable predictions for distant homologs have become possible due to the progress in fold recognition methods (with the parallel growth of the sequence and structural databases), whereas *de novo* folding simulations proved to be increasingly reliable for independent domains and relatively small proteins [1-7].

The fold recognition approach relies on the fact that numerous native protein folds have already been determined. Given an appropriate scoring function (also referred to as *folding potentials* throughout the paper) these methods “simply” find the best (i.e. the most compatible) template from the library of known folds. The scoring functions for fold recognition typically incorporate some measures of sequence-to-structure fitness, helping to find distant homologs that share the same fold without detectable sequence similarity.

On the other hand, in *de novo* (or *ab initio*) folding simulations one attempts (at least in principle) to reproduce the actual physical folding process by sampling the conformational space without restriction to known protein structures. The unique three-dimensional structure of a protein is postulated to correspond to a global minimum of the free energy function, which may be approximated by a simplified folding potential. In

practice, mixed protocols that utilize similarity to known proteins in order to constrain the expensive search in the space of all possible conformations are often applied. The Rosetta protocol by Baker and colleagues [8] is a particularly successful example of the latter approach.

The success of the Rosetta and other protocols for protein structure prediction is critically dependent on the quality of scoring functions, which are used for conformational search and to rank the resulting putative structures. For example, the Rosetta protocol combines tailored sequence independent and sequence dependent measures in order to identify native-like structures (meaning structures close to the native structure in terms of RMSD or any other conveniently chosen measure of structural similarity) among large samples of putative models [8]. In this paper, we revisit the role of sequence independent filters in protein folding simulations and in the design of improved scoring functions for protein structure prediction. We also consider the problem of selecting appropriate decoy structures (i.e. misfolded models of a protein) for the design of folding potentials.

Park and Levitt have raised the importance of a proper choice of decoy structures for design and evaluation of folding potential before [9]. They argued that only structures with overall protein characteristics need to be included in studies on scoring functions since “non-physical” decoys could be, in principle, recognized by simpler (e.g. sequence independent) tests. The question remains, however, which features are critical and consequently which measures are more effective in filtering out “non-physical” structures depending on the particular technique used to generate the decoy set. It is also not clear to

what extent different decoy sets should be utilized to develop scoring functions for discrimination of native-like structures in folding simulations and fold recognition.

While all-atom or intermediate united atom models [10-13] may be more appropriate for conformational search in structure refinement and protein folding simulations, we focus here on simple contact models for structure recognition. The rationale for this choice lies in the conceptual and practical importance of simple models in protein folding studies [14-15]. For example, inter-residue pairwise potentials have been widely used in computational protein structure prediction to distinguish native-like from misfolded conformations [16-21]. In principle, the reduced representation of protein structures, with one interaction center per residue and pair specific interaction strength, proved to be insufficient for perfect recognition of all native structures in demanding sets of decoys [22-26]. However, an approximate ranking of native and misfolded conformations is often sufficient for successful applications in fold recognition [21] or threading (by which we mean a fold recognition technique that relies on sequence-to-structure matching with contact potentials) [27].

In threading the task is to distinguish between optimal and non-optimal effective inter-residue interactions imposed by an alignment of the sequence of interest with a known structure. In this case, relevant sequence dependent features are relatively well captured by the simple inter-residue pairwise model. However, a significant fraction of decoy structures generated in *de novo* folding simulations may be characterized by “non-physical” packing. Packing of 3D structures is defined here specifically in terms of inter-residue radial distribution function, as discussed in detail in the subsequent sections of the paper. Knowledge based pairwise potentials that are derived from known protein

structures fail to discriminate between structures with physical and non-physical packing. Therefore, one may be able to design improved folding potentials for recognition of native-like conformations by first applying sequence independent filters in order to separate physical models from non-physical ones.

Here, we use Linear Programming (LP) and the Maximum Feasibility (MaxF) heuristic for infeasible LP problems [25] in order to demonstrate that pairwise potentials optimized to perform well on the Rosetta set of decoys perform poorly on decoys generated by using threading and vice versa. This apparent lack of transferability of parameters between threading and folding potentials, together with the observation of relatively good performance of a simple filter based on the number of contacts for Rosetta decoys, indicates very different characteristics of the two types of decoy structures.

Starting from the above observation, we revisit several sequence-independent filters, including the contact order filter [29] and the contact number filter, which may be regarded as a simplified version of compactness filters used by Park and Levitt [9] and by Simons et. al. [8], for instance. We also propose a novel sequence independent filter that uses the shape of the inter-residue radial distribution function in order to discriminate the protein-like from non-physical packing. We next demonstrate, using the Rosetta, CASP4 and Park and Levitt sets of decoys that combinations of different filters may be advantageous in terms of discrimination of native and native-like from misfolded structures. We also suggest how this new filter may be used to define convergence criteria for folding simulations.

The paper is organized as follows. In the Methods and Materials section we briefly revisit the LP and MaxF approaches to optimization of folding and threading potentials. Next, we define the new sequence independent filter based on the shape of pair distribution functions as well as describe the various data sets of native and decoy structures used in the paper. In the Results section we discuss to what extent parameters for folding and threading potentials are interchangeable and the implications for the design of improved potentials for both folding and threading. We also present the results of the contact number, contact order and the new pair distribution function based filters using different set of decoys, followed by conclusions.

II. Methods

II.1 Maximum Feasibility protocol for optimization of folding potentials

An ideal folding potential, which will be represented throughout the paper as an effective energy function E , would be expected to distinguish all native-like from non-native conformations. Such discrimination may be achieved by imposing that for each pair of native and misfolded structures the following constraints are satisfied:

$$\Delta E_{\text{mis,nat}} = E_{\text{misfolded}} - E_{\text{native}} \geq \varepsilon . \quad (1)$$

Here, $E_{\text{native}} \equiv E(\mathbf{X}_{\text{nat}}; \mathbf{z})$ and $E_{\text{misfolded}} \equiv E(\mathbf{X}_{\text{mis}}; \mathbf{z})$ are the energies of the native, \mathbf{X}_{nat} , and misfolded, \mathbf{X}_{mis} , structures, respectively, whereas \mathbf{z} is the vector of parameters and ε is a positive constant. Assuming that dependence on the parameters \mathbf{z} is linear, the requirement that energies of native structures be lower than the energies of misfolded

structures allows one to apply linear programming techniques to design and optimize folding and threading potentials [30, 22-27].

Let us consider widely used inter-residue folding potentials. In contact pairwise models [16-20] the energy of the protein with sequence S and a structure \mathbf{X} is a sum of pair energies from all pairs of interacting amino acids:

$$E(S, \mathbf{X}; \mathbf{z}) = \sum_{\gamma} z_{\gamma} n_{\gamma}(S, \mathbf{X}). \quad (2)$$

The summation index, $\gamma \equiv \alpha\beta$, runs over 210 different contact types, where α and β denote the types of amino acids at certain sites i and j , and $n_{\gamma}(S, \mathbf{X})$ denotes the number of contacts of a specific type found in \mathbf{X} . Sites i and j are said to be in contact, if their distance, r_{ij} , is sufficiently small. In this work we consider a model that was used before [27], with geometric side chain centers as interaction sites that are assumed to be in contact if their distance satisfies: $1.0 < r_{ij} < 6.4 \text{ \AA}$. We also consider an alternative model, in which short-range contacts are excluded, $4.0 < r_{ij} < 6.4 \text{ \AA}$. Pairs of residues that are separated by fewer than four virtual bonds are excluded, i.e. $|i - j| \geq 4$.

The parameters $z_{\gamma} \equiv z_{\alpha\beta}$ are the target for LP optimization. Given a set of native and misfolded structures and the resulting frequency of different types of contacts in native and non-native structures, one obtains the corresponding set of linear inequalities:

$$E(S_n, \mathbf{X}_{j_n}; \mathbf{z}) - E(S_n, \mathbf{X}_n; \mathbf{z}) = \sum_{\gamma} z_{\gamma} (n_{\gamma}(S_n, \mathbf{X}_{j_n}) - n_{\gamma}(S_n, \mathbf{X}_n)) \geq \varepsilon \quad \forall (j_n, n). \quad (3)$$

Here, the index j_n runs over the misfolded structures for protein sequence S_n and n runs over the native structures in the training set. The goal is to find a set of effective pair

energies, $z_{\alpha\beta}$, satisfying the inequality constraints (3). If the problem is feasible, then the set of inequalities (3) may be solved efficiently for \mathbf{z} by using LP solvers, otherwise an indication of infeasibility is obtained.

The LP approach has been applied before to the design of pairwise potentials for protein folding and protein threading [30, 22, 27]. In particular, pairwise models were found to be insufficient for perfect recognition of native protein structures [22, 25-27]. In other words, the set of inequalities (3), proves infeasible for sufficiently large sample of native and misfolded structures when using pairwise contact models. As discussed in the next section, this is also the case for the Rosetta and threading sets of decoys considered in this work.

The recently introduced Maximum Feasibility (MaxF) [25] heuristic may be used in such a case to find an approximate solution, which satisfies a possibly large subset of an infeasible set of inequalities. Using MaxF allows one to go beyond the simple feasibility test when assessing the quality of a given model (note that including just few special cases in the training may result in infeasibility [25]). It also provides a simple way to improve potentials that are not explicitly optimized to satisfy inequality constraints in (1), as for example the commonly used statistical pairwise potentials [25-26].

The MaxF procedure is based on a special property of **interior point** algorithms for LP [31-33]. Without a function to optimize the interior point algorithm places the solution at the “maximally feasible” point, which is away from any individual constraint. The idea behind MaxF heuristic is that the “maximally feasible” partial solution is likely to satisfy more constraints than an off-centered guess. The MaxF heuristic starts from a certain initial guess and then a series of “maximally feasible” approximations is

computed. The (feasible) subset of all the inequalities satisfied by the previous approximation, is solved using an interior point method and the new solution becomes our next approximation that satisfies at least as many constraints as the previous partial solution. If no further constraints are satisfied the procedure stops [25]. The pPCx package by M. Wagner [26] was used to obtain results presented in this paper.

The choice of the initial guess of the solution is critical for the success of the MaxF heuristic. The problem of finding the largest feasible subset of an infeasible set of inequalities is NP-hard [34-35] and obtaining a satisfactory approximation cannot be guaranteed. However, in practice we observe significant improvement with respect to initial approximate solutions, provided that they are carefully chosen using a priori knowledge [25-26]. For example, the statistical potentials have been demonstrated before to provide reasonable initial approximations that may be improved using MaxF [26].

We would like to comment that another way to obtain an appropriate initial guess could be to solve an “elastic” (or soft) LP problem, with positive slack variables added to constraints in equation (1). Such a problem is always feasible and, by adding the sum of slack variables as the objective function, allows one to find approximate solutions of the original infeasible problem [36-37]. Although coupling MaxF with an elastic LP may provide solutions with a smaller total number of violated inequalities [38], we attempt to improve here well characterized threading and folding potentials from the literature. Since the final solution satisfies all the constraints that are not violated by the starting guess, the MaxF potentials tend to preserve the general characteristics of the initial potential. This may, in turn, be used to limit the overfitting, which is important for the analysis of transferability between potentials optimized for threading or Rosetta.

II.2 Sequence independent filters

In order to enhance the performance of Rosetta simulations, Baker and colleagues incorporated a number of sequence independent measures into their scoring functions [8, 29, 39]. Moreover, additional filters were used to eliminate non protein-like structures and to bias the Rosetta simulations towards structures with desired characteristics. For example, sequence independent terms monitor strand-strand pairing and beta sheet formation as well as helix-strand interactions [8]. Another filter that proved to be important for enhancing the performance of Rosetta simulations was based on the notion of the contact order. For a structure consisting of L amino acid residues the contact order is defined as follows:

$$CO = \frac{1}{L} \Delta \bar{S}_{ij} = \frac{1}{LN} \sum_{i < j} \Delta S_{ij} \quad , \quad (4)$$

where ΔS_{ij} denotes the sequence separation of residues i and j in contact, the summation runs over all contacts and N is the total number of contacts. The larger the average separation of residues in contact, $\Delta \bar{S}_{ij}$, the higher the contact order. As observed experimentally [29], structures with higher contact order tend to fold slower due to non-local contacts. This may be accounted for in the simulation by biasing towards structures with higher contact order.

We assume here that two residues are in contact if they are not immediate neighbors along the sequence (i.e. $|i - j| \geq 2$) and the distance between their side chain centers satisfies $1.0 < r_{ij} < 6.4 \text{ \AA}$. Thus, in accord with the paper by Plaxco et. al. [29]

(and contrary to what we used in section II.1), short range contacts due to local helical structures will be included, decreasing the overall contact order for helical proteins. Consequently, filtering out conformations with relatively low values of contact order for mostly helical proteins might favor incorrectly folded decoys with high beta strand content. This problem may be addressed using the overall consistency with the predicted secondary structures as additional filter.

INSERT HERE Figure 1.

The packing of amino acid residues may be characterized in terms of the pair correlation function, also called the radial distribution function (RDF). For the spherically symmetric inter-residue interactions considered here, the pair correlation function, $g_{\alpha\beta}(r)$, for a pair of residues of type α and β , located at a distance $r \pm \Delta r$ from each other is proportional to the number of pairs $[\alpha, \beta]$ found at this separation, $N_{\alpha\beta}(r)$. Following the definition adopted by Bahar and Jernigan [20], the normalized RDF may be expressed as follows:

$$\bar{g}_{\alpha\beta}(r) = \frac{g_{\alpha\beta}(r)}{\sum_r g_{\alpha\beta}(r)}, \quad (5)$$

$$g_{\alpha\beta}(r) = \frac{N_{\alpha\beta}(r)}{4\pi r^2}, \quad N_{\alpha\beta}(r) = \sum_{i < j} \delta(|\mathbf{r}_{\alpha i} - \mathbf{r}_{\beta j}| - r), \quad (6)$$

where $\mathbf{r}_{\alpha i}$ is the position vector of the i th residue of type α , $\delta(x)$ indicates the Kronecker delta and the summation runs over all pairs of type $[\alpha, \beta]$. The number of pairs at a given separation (and thus RDF) is computed for 40 discrete bins on the $1.0 \leq r \leq 11 \text{ \AA}$ interval. Since the interactions between the nearest neighbors are strongly influenced by

chain connectivity, only pairs of residues that are separated by six or more virtual bonds (i.e. $|i - j| \geq 6$) are included here.

There is an close relationship between potentials of mean force, and statistical pairwise potentials in particular, and RDFs. The effective distance dependent interaction energy between residues of type α and β relative to the average interactions $z_{XX}(r)$ may be expressed as:

$$\Delta z_{\alpha\beta}(r) = z_{\alpha\beta}(r) - z_{XX}(r) = -RT \ln[\bar{g}_{\alpha\beta}(r) / \bar{g}_{XX}(r)], \quad (7)$$

where $\bar{g}_{XX}(r)$ is the mean inter-residue radial distribution function [16]. Thus, effective pairwise potentials (including also considered here approximate stepwise potentials) discriminate between contact type specific variations in the shape of the pair correlation function. Since the reference, $\bar{g}_{XX}(r)$, is derived from a set of well packed, native protein structures (see Figure 1), such potentials may fail to discriminate between physical (“protein-like”) and unphysical structures.

The average (reference) RDF is defined here using the representative native structures from the Pfam database as shown in Figure 1. Note that there is a pronounced maximum corresponding to the first contact shell, as well as a minimum between the first and second contact shells. Note also that CASP4 structures contain on average more disulphide bridges, manifesting by an increased probability of observing two residues at a separation of about 3 Å. On the other hand, individual structures j from protein folding simulations may result in RDFs $\bar{g}_{XX}^j(r)$ that deviate significantly from the ideal shape (see Figure 2). In order to capture those deviations and to filter out putative structures characterized by non-native packing we introduce the following distance measure:

$$d_s(\bar{g}_{XX}^j(r), f(r)) = \int_a^b |f(r) - \bar{g}_{XX}^j(r)| dr; \quad (8)$$

$$f(r) = \frac{1}{2} [\max_{r \in [a,c]} \bar{g}_{XX}^j(r) + \min_{r \in [c,b]} \bar{g}_{XX}^j(r)] \quad (9)$$

Thus, d_s , estimates how strongly the first contact shell peak and the minimum between the first and the second contact shells of $\bar{g}_{XX}^j(r)$ are pronounced, with the [a,b] interval chosen to include these two features. After some experimentation the interval [a,b] was set as [3.5,8.5] Å (the choice of c is discussed below). However, instead of measuring the surface area between the structure specific and the ideal (average) RDF, for example derived from the Pfam set of structures, the surface above and below the central horizontal line defined in equation (9) is computed (see also Figure 3). The latter measure discriminates better between protein-like and unphysical structures because of shifts in the position of the maximum of $\bar{g}_{XX}^j(r)$ for individual structures and because of the need to apply smoothing, as described below.

INSERT HERE Figure 2.

Individual structures may contain only few contacts, resulting in noisy RDFs. Therefore, we applied Gaussian smoothing that introduces an uncertainty as to the exact position of the residues in contact: each contact at a separation r is replaced by a Gaussian density centered at r , with the standard deviation that may be varied in order to tune the level of smoothing. Here we use $\sigma = 0.25$ Å. Despite the smoothing, some structures with very low number of contacts may still result in rapidly changing (noisy) RDFs. Therefore, the definition of d_s was adjusted: only the surface above the central line $f(r)$ on the interval [3.5,6.0] Å and only the surface below the central line on the interval [6.5,8.5] Å, respectively, is taken into account.

Since the alternative conformations are compared in terms of the shape of the contact type (and thus sequence) independent mean inter-residue RDF, $\bar{g}_{xx}^j(r)$, the new filter could be better suited to distinguish structures with non-physical packing than sequence dependent mean field potentials of equation (7). The performance of the contact order, contact number and RDF based filters is discussed in the Results section. We would like to remark that these results are not sensitive to small changes in cutoff distances, interaction centers (e.g. side chain centers vs. C_α carbons provided that cutoff distances are adjusted accordingly) or extent of nearest neighbor exclusions.

II.3 Decoy sets

In our analysis of the performance of pairwise potentials and sequence independent filters we use Rosetta [5], CASP4 [1] and Park-Levitt [15] sets of decoys. The **Rosetta** set of decoys was developed by Simons et. al. [5] using their protocol for assembling protein tertiary structures from a library of short fragments [8]. A set of 92 families of structures, consisting of the native state and approximately one thousand decoy structures that were generated by the Rosetta simulation protocol, was first converted to contact representation. The linear inequalities in (3) were then generated for each family, resulting in a set of approximately 93 thousand constraints. This set of decoys and native structures as well as the resulting set of inequalities will be simply referred to as the Rosetta set throughout the paper.

INSERT HERE Figure 3.

The second set of decoys that we used consists of a subset of 25 non-homology **CASP4** targets [1], for which the native structures were available, and the corresponding models submitted by the predictors. On average, about 120 models (decoy structures) per target were included (alignment based and backbone only models were excluded). Structures with a wide range of RMSD with respect to native conformations are present in the set of CASP4 models. We will refer to this set of decoys and native structures as CASP4 set.

In order to further sample different types of decoys we also considered several sets of decoys developed by Levitt and colleagues [9], and used before to design and evaluate various folding potentials [40]. The following decoy sets were merged to obtain a more representative sample of misfolded structures: the actual Park-Levitt set of decoys for 7 small proteins, the local minima decoy set for 10 proteins derived by Kesar and Levitt and decoy sets for 12 proteins designed by Simons et. al. using fragment assembly and optimization with the Charmm force field [41]. The resulting decoy set consists of 29 families and on the average about 380 decoy structures for each family, representing a wide range of deviations from the experimental structures. This set of decoys will be referred to as the **PKLS** (Park-Kesar-Levitt-Simons et. al.) set.

In addition to the above sets of decoys from the literature, we also developed a new set of threading decoys for design and optimization of improved threading potentials. A subset of the Protein Families (Pfam) database (version 6.6) [31] covering known protein domains was used. Starting from 3071 protein families, of which about 45% had known three-dimensional structures in the PDB as of Jan. 2002, and removing all the problematic PDB files, families with just one structure known and membrane

proteins, a subset of 773 families was obtained. This subset was then used to construct a set of homology based pseudo-native structures for each family, as described below.

The Pfam alignments of homologous structures belonging to the same family were used to define a number of native-like structures by overlaying the homologous sequences (instead of the native one) with the actual native structures. Up to ten such pseudo-native structures were constructed for each family and compared with populations of non-native structures. The latter ones were generated using the so-called gapless threading protocol [27], i.e., by threading (without gaps) the sequences of pseudo-native structures through structures of non-homologous proteins from the database. The resulting set of about 9.945 million decoys (on the average about 13 thousand per family) and the corresponding inequalities will be referred to as the **Pfam** set. The Pfam test is geared towards threading: perfect recognition of a family means here that not only the native but also homologous structures are recognized as native-like. Improving pairwise potentials by imposing that as many as possible native-like structures should be recognized with respect to populations of structurally unrelated conformations (decoy structures) is expected to result in threading potentials that perform better in threading and homologous structure recognition [43].

The contact representations of the structures from the respective databases as well as the linear constraints of equation (3) in terms of sequence dependent pairwise models were derived using the Loopp program [44]. The lists of proteins and decoy structures in each database and the resulting potentials optimized using them are available on-line at <http://sift.chmcc.org>.

III. Results and Discussion

III.1 Transferability between folding and threading potentials.

The summary of results for several potentials optimized here for recognition of native structures in Rosetta or Pfam sets as well as some potentials from the literature, including the HP model [14], Miyazawa-Jerningan (MJ) [19] and Tobi-Elber (TE) [24] potentials, are presented in Table 1. The number of native structures ranked best, N_{nat} (perfect recognition), and as one of the best five, N_{casp} (CASP criterion) as well as the average number of decoys that are ranked higher than the native structures, N_{mis} (misclassified decoys), are reported for the set of 92 families of Rosetta decoys in the second column. The results for the Pfam set, with up to ten pseudo-native structures per family (see section II), are reported in the third column. The number of families for which all of the homology derived pseudo-native structures were ranked higher than any misfolded structure, N_{nat} , and the number of misclassified decoys (in thousands, out of about 9,945 thousand), N_{mis} , are given.

INSERT HERE Table 1.

As can be seen from Table 1, all the pairwise potentials included in our analysis reach very limited accuracy in terms of ranking of native vs. non-native structures in the Rosetta test. Let us consider first the simple HP model that rewards any contact between hydrophobic residues while neglecting contacts involving polar residues (we adopted here the convention from [27] to classify each of the 20 amino acids as either hydrophobic or polar). Thus, the HP model reduces to a simple counting of contacts between hydrophobic residues. For reasons that will become apparent when we analyze

the contact number filter, the HP potential performs relatively well compared to pairwise potentials with 210 parameters, recognizing correctly 16 (or 28 according to CASP criteria) out of 92 native structures. On the other hand, the HP model performs significantly worse than the Miyazawa-Jernigan (MJ) [19] and Tobi-Elber (TE) [24] potentials in the Pfam test, misclassifying 251 thousand decoys and recognizing perfectly 564 families. The statistical MJ potential or the LP optimized TE potential for large-scale threading self-recognition misclassify 128 and 118 thousand decoys, respectively, and recognize all the homologs from 675 (656) families.

We next attempted to further improve pairwise potentials for Rosetta simulations and for threading by using the MaxF heuristic described in Section II.1. First, the threading optimized TE potential is used as the initial guess for further refinement with the MaxF approach. Starting from the subset of Pfam constraints that are satisfied by the TE potential, or in other words including initially only those pairs of native-like and decoy structures that are correctly ranked, three MaxF iterations suffice to obtain a potential that violates only 35 thousand (as opposed to the initial 118 thousand) constraints and recognizes perfectly 672 (as opposed to initial 656) families on the training set. The new potential is referred to as MaxF TE-T, where TE denotes the starting guess and T denotes threading optimized. While yielding an improved recognition of homologs with respect to self-recognition trained TE potential, the threading optimized MaxF TE-T potential performs much worse than the original TE potential in the Rosetta test, however. Namely, the MaxF TE-T potential misclassifies on the average 313 decoys per family, compared with 188 for the TE potential.

The TE potential was next refined in order to recognize more decoys in the Rosetta set. Starting from the subset of constraints satisfied by the original TE, one obtains after three MaxF iterations potential referred to as MaxF TE-R (with R denoting Rosetta optimized), which misclassifies on the average only 70 decoys per family in the Rosetta training set. However, MaxF TE-R potential violates as many as 616 thousand inequalities from the Pfam set (which plays a role of the control set in this case), recognizing perfectly only 254 families in this test.

The last potential we consider here was optimized to further improve recognition of native structures in the Rosetta set. The initial feasible subset of constraints was obtained this time by considering only those decoys that are recognized as non-native by the HP potential. However, short range contacts from the interval $1.0 < r_{ij} < 4.5 \text{ \AA}$ were excluded. Such modified HP contact potential ranks 27 native structures as best (and 34 as one of the best five) in the Rosetta test and it is further improved by three MaxF iterations in the space of 210 parameters corresponding to the full pairwise model. In other words, the initial energy term is equal to -1 for each pair of hydrophobic residues and 0 for any pair involving a polar residue, respectively. However, all the 210 pair energies are varied independently during the optimization procedure. The resulting potential is referred to as MaxF HP210-R to indicate that the initial solution is that of HP model projected into the space of 210 types of contacts and that it was optimized for recognition of Rosetta decoys as non-native structures. As can be seen from Table 1, the MaxF HP210-R potential performs best in the Rosetta test (recognizing using the CASP criteria as many as 60 native structures) and, on the other hand, much worse than any other potential on the Pfam test set.

We would like to comment that poor performance observed on the Pfam test indicates certain degree of overfitting for the potentials optimized using the Rosetta set as training. It is not our goal here to find the best folding potential for Rosetta simulation, but rather to illustrate the increased deterioration of the performance on the Pfam control set for potentials achieving an improved recognition for Rosetta decoys. On the other hand, however, the number of training vectors is still much larger than the number of parameters to be optimized and, additionally, the effect of overfitting is further reduced by the use of MaxF. All the Rosetta decoys recognized initially by the original TE potentials must also be recognized in the subsequent iterations, constraining the optimization to a specific region in the parametric space. Therefore, the clear trend observed here is likely to indicate the very different nature of the two sets of decoys.

The results of MaxF optimization of potentials for threading and Rosetta decoys are also consistent with previous observation suggesting very specific characteristics of the decoys generated by Rosetta simulations [5,8,28,40]. In particular, the relatively good performance of the simple HP potential in the Rosetta test may be explained by the fact that many of the Rosetta decoys are not compact enough. Indeed, 34 out 92 native structures have larger number of contacts than any decoy, suggesting that a simple sequence independent filter based on the number of contacts (discussed in detail in the next section) may play a useful role.

On the other hand, one can significantly improve upon the HP (or, in fact, simple contact counting) model by using MaxF in the space of the full contact model with 210 parameters, indicating a sequence dependent structure in the packing of Rosetta decoys, which is exploited by the optimization protocol. This structure appears to be, however,

very different compared to the packing of threading decoys. As a result, Rosetta optimized potentials (such as the MaxF HP210-R potential) fail in the Pfam test, whereas the threading optimized potentials perform poorly in the Rosetta test.

In light of the above, further analysis of packing and other characteristics of decoys generated using different protocols may help identify the nature of those differences. This in turn may allow one to select appropriate set of decoys to train enhanced potentials for folding simulations and for recognition by threading on one hand, and to facilitate folding simulations by additional measures that filter out non physical structures on the other hand.

III.2 Sequence independent filters.

The first sequence independent filter that we consider here is the simple contact number filter. The presumption is that correctly folded structures are more likely to achieve denser packing, while structures with relatively low number of contacts may be discarded as unphysical. In fact, one may recognize about one third of native structures in the Rosetta set by simple counting of contacts (in the first contact shell) because none of the alternative structures is packed densely enough.

INSERT Figure 4.

Several examples of distributions of the number of contacts for structures from the Rosetta simulations are included in Figure 4. The solid and dashed vertical lines indicate the number of contacts in the native structure and the average number of contacts in the decoy structures, respectively. In addition, subsets of native-like and grossly

misfolded structures are defined for each family of decoys. We define these subsets in terms of the fraction of native contacts measure, Q . Namely, structures that have the value of Q larger than the average plus one standard deviation, $Q > \bar{Q} + \sigma$, measured for the distribution of Q values in a given family of decoys, are regarded as native-like. On the other hand, decoys that have very few native contacts, $Q < \bar{Q} - \sigma$, are regarded as grossly misfolded. The distributions of native-like structures are shown in Figure 5 using shaded bars.

There are only 4 structures in the Rosetta set (1hsn, 1res, 1tih and 2bds) that have fewer contacts in the native conformation than the average number of contacts in the decoy conformations. On average, 86% of the native-like structures per family have more contacts than the average number of contacts observed in this family, as opposed to only 15% of the grossly misfolded structures. This suggests a simple filter that enriches the population of putative conformations in native-like structures for globular proteins by removing those structures that have less than average number of contacts in the first contact shell.

The contact order and the radial distribution function shape scores are less successful in the Rosetta test than the contact number score. The performance of RDF based score is similar to that of the MJ pairwise potential, which performed somewhat better than other potentials from the literature in our tests (see Table 2). Examples of histograms for values of d_s measure of the RDF shape are included in Figure 5. Using the RDF based filter, i.e. filtering out structures with d_s value smaller than the mean, retains on average only about 53% of the native-like structures per family.

Nevertheless, as can be seen from Figures 4 and 5, for some families the RDF filter works better than contact filter, suggesting that combination of different filters might be advantageous. Note, for example, that 1hsn native structure, which is a relatively open DNA binding protein, is not filtered out by the RDF filter, although it was removed by the contact number filter. The opposite happens for the native structure of the actin binding protein 1ksr. The latter results in an usual RDF with the second contact shell maximum shifted into the region of the expected minimum between the first and second contact shell (see also discussion of the results for CASP4 decoys below).

INSERT Figure 5.

While clearly successful in Rosetta test, contact number and other compactness filters must be applied with care. For example, the contact number filter should not be applied to extended (“open”) structures with relatively only few contacts. On the other hand, however, when the structure is known (or is predicted) to be relatively compact it may be useful to exploit the premise that native-like structures from folding simulations can be simply recognized by their relatively dense packing.

INSERT Table 2.

The Rosetta decoys considered here were obtained using a scoring function that favors compact structures [8, 39]. The overall compactness is measured in terms of a “density” term, P_{density} , which is a function of the observed number of inter-residue contacts in the first and second contact shells, and in terms of the radius of gyration, which is defined as the square root of the averaged squares of inter-residue distances. Thus, structures with higher number of residue pairs at close separation (contacts in the first contact shell) will generally have a favorable “density” and a lower radius of

gyration. Relatively open structures, on the other hand, should be eliminated in the course of simulations. Nevertheless, a significant fraction of structures with relatively sparse packing is included in the Rosetta set and the number of contacts filter turns out to be useful in filtering out such decoys. It is plausible that the simple contact number score could enhance convergence of Rosetta simulations in combination with other recent improvements in Rosetta, which are based on global measures of hydrophobic core formation and implicitly account for compactness [45].

While the contact number score is clearly more successful than other sequence independent or sequence dependent scores for recognition of native structures from the Rosetta set, this is not the case for the PKLS and CASP4 decoys (see Table 2). The native structures from the PKLS set are best ranked by the MJ pairwise potentials. However, contact filter is still successful in enriching decoy populations in native-like structures. For example, on average about 70% of native-like and only about 30% of grossly misfolded structures per family contain more contacts than the mean number of contacts in the family. On the other hand, the performance of contact order filter is poor, which is consistent with the fact that relatively large fraction of short, helical proteins are included in the PKLS set.

For CASP4 decoys using a combination of the contact score and the RDF shape score performs better than individual scores (including the pairwise MJ potential). Some CASP models with very few contacts lead to an artificially pronounced first contact shell peak of individual RDFs. As a result, spuriously large values of d_s may be obtained. Therefore, for CASP4 we report results of the RDF filter in combination with contact filter, which is first applied in order to remove structure with few contacts. In fact, the

results of the RDF shape filter may also be improved by removing structures with a low number of contacts in case of Rosetta and PKLS decoys. For example, using such a combination for ranking of decoys in PKLS set, 18 native structures are retained in the top 100 structures, as opposed to 16 with just the RDF shape filter or 15 with the contact number score. Note that the star symbol (*) is used in Table 2 to indicate when the RDF shape filter was combined with the contact number filter.

The improvement of the results due to combination of RDF and contact number filters is, however, more significant in case of CASP4 decoys. When the RDF based filter is applied independently, it ranks (similarly to contact number filter) only 11 native structures as one of the best ten structures. On the other hand, as can be seen from Table 2, the RDF filter combined with initial filtering out of structures that have fewer than mean number of contacts ranks 10 native structures as best and 18 (out of 25) as one of best five models, significantly outperforming other scores, including sequence dependent MJ potential. Thus, more than 70% of native structures in the CASP4 set can be recognized using the CASP criterion by a simple sequence independent measure.

INSERT Figure 6.

The analysis of ranking of native-like structures is included below and in Figure 6. We regard here as native-like all the models that obtained non-zero scores in Murzin's evaluation of CASP4 predictions [3], including partially correct models. The distributions for "native-like" structures are shown using shaded bars. For a subset of 15 difficult targets (T0097-98, T0102, T0105-106, T0114, T0089, T0094, T0100-101, T0107-109, T0120-121), at least one native-like structure was scored as one of the top five models for eight targets. As can be seen from the examples included in Figure 6, however, many

native-like models are in fact characterized by values of d_s which are much lower than those for the respective native structure (often in accord with low Murzin scores for such models). While similar values of d_s do not necessarily imply similar packing, structures close to the native conformation should result in similar values of d_s measure (as is the case for several models for target T0120, for instance).

It is worth noting that the new filter based on the shape of RDF may be used to indicate if a given set of decoys contains native-like structure even when the actual native structure is unknown. Decoys characterized by non-physical packing may lead by chance (e.g. because of very few contacts at a specific separation) to the values of d_s similar to that of native structure. However, it is unlikely that the actual RDF would be similar. Therefore, clusters of similar structures (either in terms of RMSD or overall RDF shape) with large values of d_s are likely to reveal populations of native-like structures. Testing this hypothesis remains the target for the future work.

IV. Conclusions

Protein structures obtained in *de novo* folding simulations, such as Rosetta decoys or CASP models, are often packed in non-physical way. This non-physical packing may manifest itself by non-compactness of the decoy, its low contact order or unphysical shape of the inter-residue radial distribution functions. Therefore, pairwise threading potentials that are trained to recognize native structures against other folded conformations fail when presented with populations of decoys generated in folding simulations.

Using Linear Programming, coupled with our Maximum Feasibility approach, we showed that pairwise potentials optimized for threading perform significantly worse on the Rosetta set of decoys. The reverse is also true, i.e., the potentials optimized for Rosetta perform much worse in threading, indicating the **lack of transferability** between potentials for folding simulations and fold recognition. Even though our analysis is limited to pairwise potentials, it suggests that separate strategies for design of folding and threading potentials are required. Namely, rather than attempting to improve the accuracy of folding and threading potentials by including both types of decoys in the training, separate potentials and filters should be developed for different tasks. Moreover, if decoys generated by folding simulations, such as Rosetta, are to be used in the design and optimization of improved potentials for protein folding, then structures that do not achieve protein-like packing should be first filtered out.

The **sequence independent filters** that we consider here, including a novel filter based on the shape of inter-residue pair correlation function, achieve surprisingly good performance not only in filtering out non-physical structures but also in terms of ranking of native and, to a lesser degree, native-like structures. In light of the above, one may observe that relative success of the Rosetta approach by Baker and colleagues lies to a large extent in the development of tailored scoring functions. Moreover, successful recognition of native structures in CASP4 test by simple sequence independent filters exposes obvious weaknesses of many models submitted to the CASP4 competition. Applying some of these simple filters might have prevented some groups from submitting incorrect models.

Acknowledgments

This work was supported by the National Institutes of Health Grant PA-02-046 (to JM).

We are grateful to Dr. Michael Wagner for making his LP solver available to us and to Dr. Aleksey Porollo for technical assistance.

References

1. Venclovas C., Zemla A., Fidelis K., and Moult J., 2001, *Proteins: Structure, Function, and Genetics*, Suppl. 5: 163-170
2. Fischer D., Elofsson A., Rychlewski L., Pazos F., Valencia A., Rost B., Ortiz A. R., and Dunbrack R. L., 2001, *Proteins: Structure, Function, and Genetics*, Suppl. 5: 171-183
3. Critical Assessment of Techniques for Protein Structure Prediction (CASP), <http://predictioncenter.llnl.gov>
4. Schonbrun J., Wedemeyer W. J. and Baker D., 2002, *Curr. Opin. Struct. Biol.*, Vol. 12, pp. 348-354
5. Bonneau R., Chivian D., Strauss C.E.M, Rohl C., Baker D., 2002, *J. Mol. Biol.* 322 (1): 65
6. Bonneau R., Tsai J., Ruczinski I. and Baker D., 2001, *J. Struct. Biol.* 134 (2-3): 186-90
7. Fischer D., 2003, *Proteins: Structure, Function, and Genetics* 51: 434-441
8. Simons K. T., Kooperberg C., Huang E. and Baker D., 1997, *J. Mol. Biol.* 268: 209-25
9. Park B. and Levitt M., 1996, *J. Mol. Biol.* 258: 367-92
10. Melo F. and Feytmans E. J., 1997, *J. Mol. Biol.* 267: 207-222
11. Samudrala R. and Moult J., 1998, *J. Mol. Biol.* 275: 895-916

12. DeBolt E. E. and Skolnick J., 1996, *Prot. Eng.* 8: 175-186
13. Liwo A., Oldziej S., Pincus M. R., Wawak R. J., Rackowsky S. and Scheraga H. A., 1997, *J. Comp. Chem.* 18: 849-873
14. Thomas P. D. and Dill K. A., 1996, *J. Mol. Biol.* 257: 457-469
15. Hinds D. A. and Levitt M., 1994, *J. Mol. Biol.* 243: 668-682
16. Sippl M., 1990, *J. Mol. Biol.* 213: 859-883
17. Bryant H. S. and Lawrence C. E., 1993, *Proteins: Structure, Function and Genetics*, 16:92-112
18. Rooman M. J. and Wodak S. J., 1995, *Protein Engineering* 8 (9): 849-858
19. Miyazawa S. and Jernigan R. L., 1996, *J. Mol. Biol.* 256: 623-644
20. Bahar I. and Jernigan R. L., 1997, *J. Mol. Biol.* 266: 195-214
21. Jones D. T., 1999, *J. Mol. Biol.* 287(4): 797-815
22. Vendruscolo M. and Domany E., 1998, *J. Chem. Phys.* 109: 11101-11108
23. Tobi D and Elber R., 2000, *Proteins: Struct. Func. Gen.* 41: 40-46
24. Tobi D., Shafran G., Linial N. and Elber R., 2000, *Proteins: Struct. Func. Gen.* 39: 71-85
25. Meller J., Wagner M., Elber R., 2002, *Journal of Computational Chemistry*, **23**: 111-118
26. Wagner M., Meller J. and Elber R., 2004, *Mathematical Programming*, to appear
27. Meller J., Elber R., 2001, *Proteins: Struct. Funct. Gen.* 45: 241-261
28. Galor T., Meller J. and Elber R., unpublished result
29. Plaxco K. W., Simons K. T., and Baker D., 1998, *J. Mol. Biol.* 277: 985-994
30. Maiorov V. N. and Crippen G. M., 1992, *J. Mol. Biol.* 227: 876-888
31. Karmakar N. K., 1984, *Combinatorica* 4: 373-395
32. Ye Y, 1997, "Interior Point Algorithms: Theory and Analysis", Wiley

33. Adler I. and Monteiro R. D. C., 1991, *Math. Program.* 50:29-51
34. Chakravarti N., 1994, *Eur. J. Oper. Res.* 73: 139
35. Garey MR and Johnson DS, 1979, W.H. Freeman and Company, New York
36. Brown G and Graves G, 1975, “Elastic Programming: A New Approach to Large-Scale Mixed Integer Optimization”, presented at ORSA/TIMS conference, Las Vegas
37. Parker M and Ryan J, 1996, *Annals of Mathematics and Artificial Intelligence* 17: 107-126
38. Porollo A., Adamczak R., Wagner M. and Meller J., 2003, “Maximum Feasibility Approach for Consensus Classifiers”, *Proceedings of The Second International Conference on Computational Intelligence, Robotics and Autonomous Systems*, Singapore
39. Simons K. T., Ruczinski I., Kooperberg C., Fox B. A., Bystroff C. and Baker D., 1999, *Proteins: Struct. Fun. Genet.* 34: 82-95
40. Gatchell D. W., Dennis S. and Vajda S., 2000, *Proteins: Struct. Fun. Genet.* 41: 518-534
41. Park-Levitt, Kesar-Levitt and Simons-Kooperberg-Huang-Baker sets of decoys are available from <http://dd.stanford.edu>
42. Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., and Sonnhammer E.L.L., 2002, *Nucleic Acids Research* 30(1): 276-280
43. Adamczak R. and Meller J., unpublished result
44. Meller J. and Elber R., 2000, “LOOPP: Learning, Observing and Outputting Protein Patterns (LOOPP) – a program for protein recognition and design of folding potentials”, <http://www.tc.cornell.edu/CBIO/loopp>

45. Bonneau R., Strauss C. E. M. and Baker D., 2001, Proteins: Struct. Fun. Genet. 43: 1-11

Table 1. Performance of pairwise folding potentials on Rosetta [5] and threading (Pfam) sets of decoys (see text for details).

Potential	Rosetta	Pfam
	$N_{\text{nat}} (N_{\text{casp}}) / N_{\text{mis}}$	$N_{\text{nat}} / N_{\text{mis}}$
HP	16 (28) / 211	564 / 251
MJ	19 (26) / 216	675 / 128
TE	14 (22) / 188	656 / 118
MaxF TE – T	4 (13) / 313	672 / 35
MaxF TE – R	28 (43) / 70	254 / 616
MaxF HP210 – R	45 (60) / 32	160 / 975

Table 2. Recognition of native structures in Rosetta, PKLS (Park-Levitt set of decoys merged with some other sets as described in the text) and CASP4 sets of decoys by sequence independent filters: contact number, contact order and the radial distribution function (RDF) shape. For each filter the number of native structures ranked as the top, top five and top hundred (ten for CASP4) structures is reported.

Decoy set:	Rosetta			PKLS			CASP4		
Results for top N structures:	1	5	100	1	5	100	1	5	10
MJ pairwise potential	19	26	52	10	14	22	8	14	18
Contact number filter	33	48	70	7	8	15	2	10	12
Contact order filter	7	21	57	0	1	9	2	6	11
RDF shape filter	21	28	52	1	4	16	10*	18*	19*

Figure 1. Normalized mean inter-residue radial distribution function (also known as the pair correlation function), $g_{XX}(r)$, computed using equation (5) and averaged over the native structures from Pfam, Rosetta, PKLS (referred to as Park-Levitt) and CASP4 databases, respectively.

Figure 2. Mean radial distribution functions (RDF) for individual structures from the Rosetta set of decoys: mean RDFs for the native structure (denoted by circles) and for a randomly selected subset of non-native structures are shown.

Figure 3. Pictorial representation of the d_s measure defined in equation (8) used in this work to capture deviations of individual radial distribution functions from the ideal RDF shape (see text for details).

Figure 4. Performance of the contact number filter for Rosetta decoys, as illustrated by histograms for the number of decoy structures with a given number of contacts in the first contact shell for four families of Rosetta decoys. Solid vertical lines indicate the number of contacts in the native conformation and the dashed vertical line shows the position of the mean for each distribution. Shaded bars are used to illustrate the distribution of native-like structures as defined in the text.

Figure 5. Examples of distributions for the values of the RDF shape measure, d_s , for families of Rosetta decoys included in Figure 4 (Panel A). Further examples of the mean RDFs for the native structures and samples of decoy structures are included in Panel B.

Figure 6. Distributions of the RDF shape measure, d_s , for several families of difficult CASP4 models (the histograms are smoothed by averaging over neighboring bars). The RDF shape filter is combined here with a weak contact number filter (see text for details).

Figure 1.

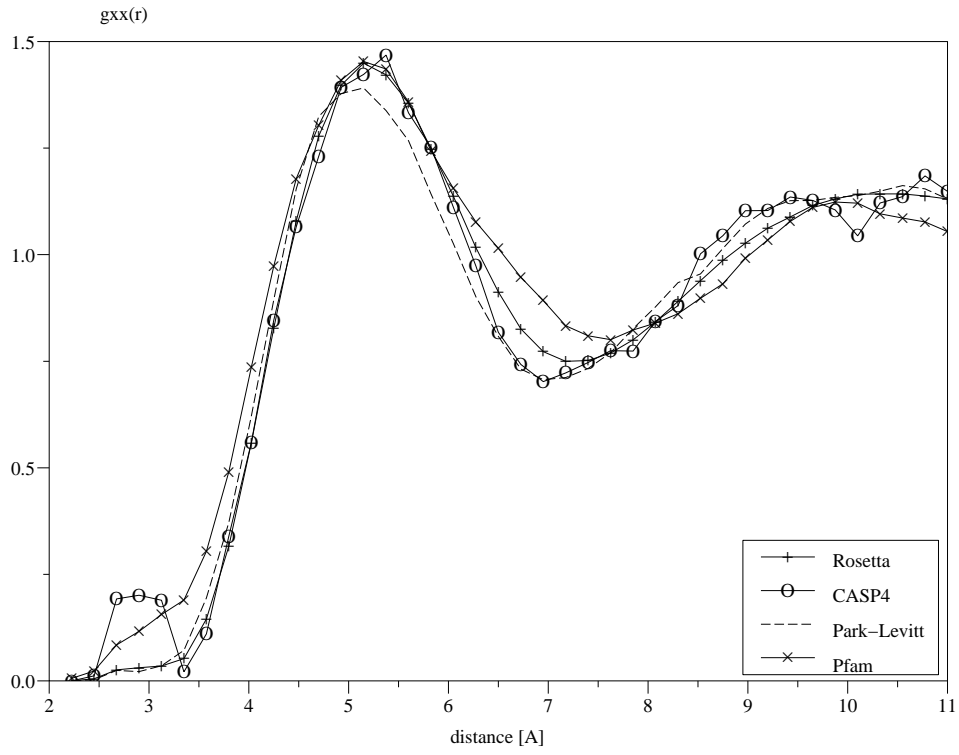
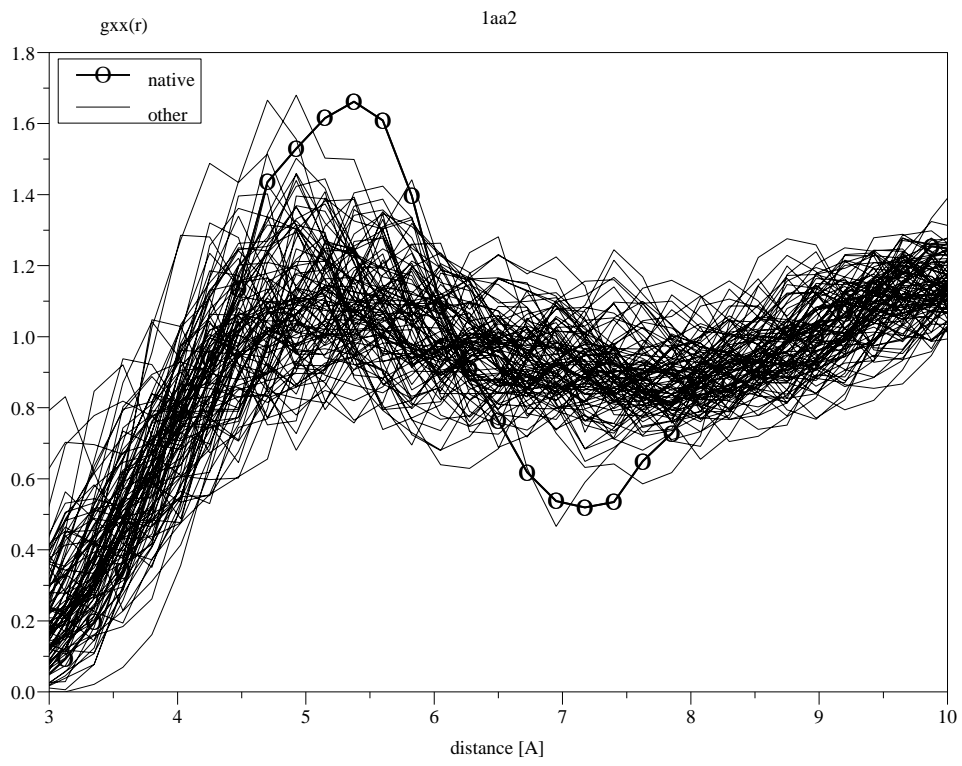


Figure 2.

A.



B.

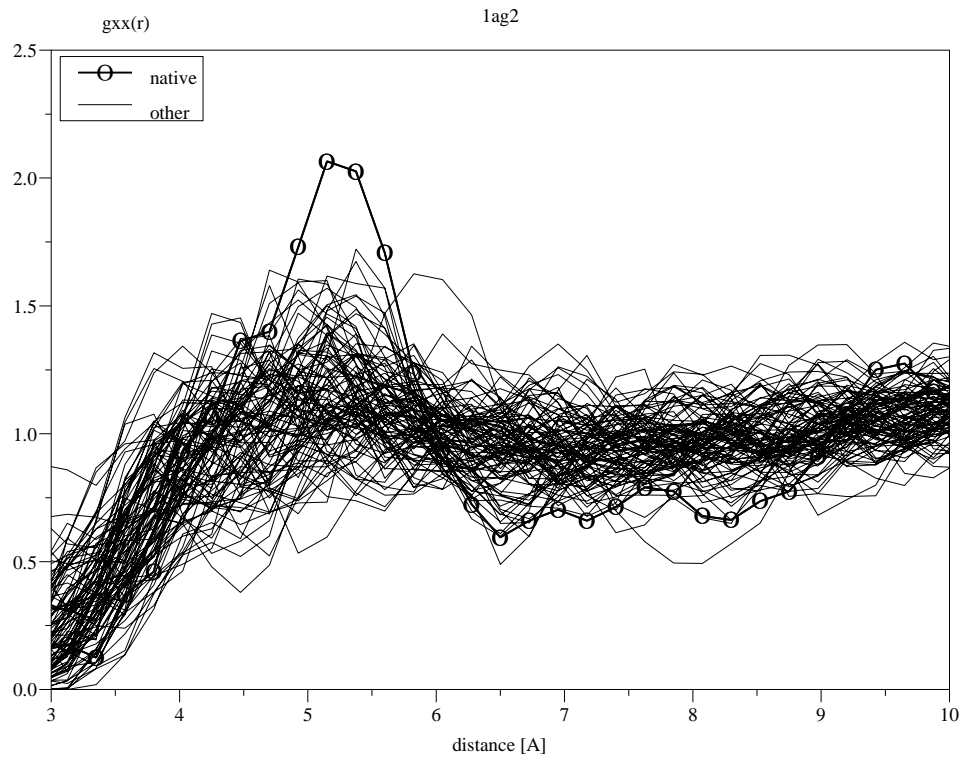


Figure 3.

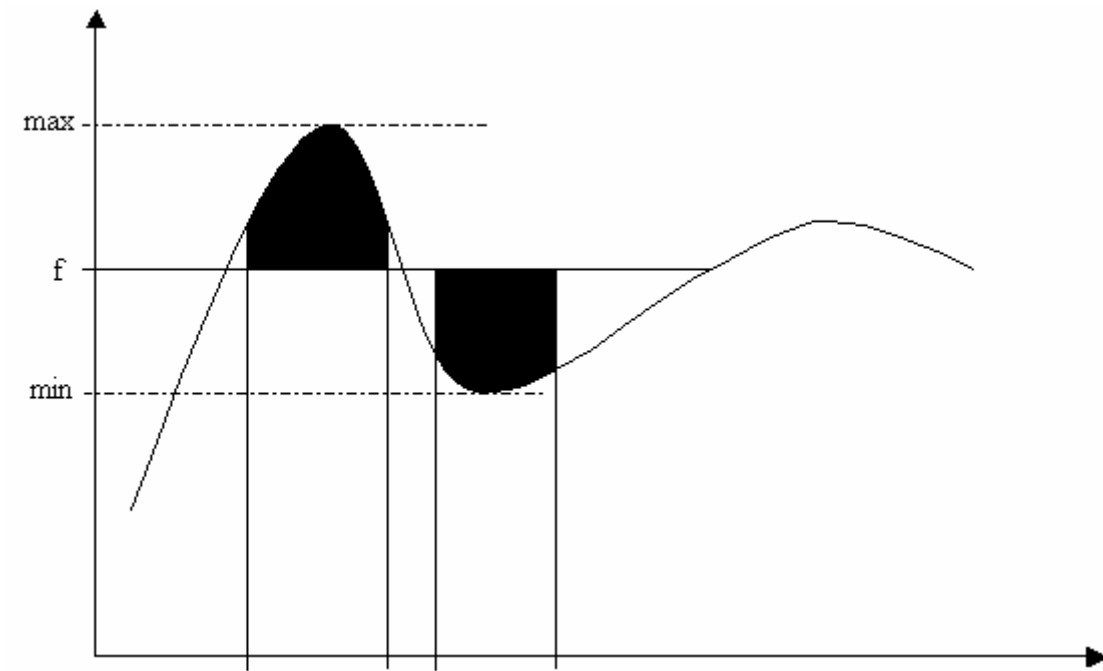


Figure 4.

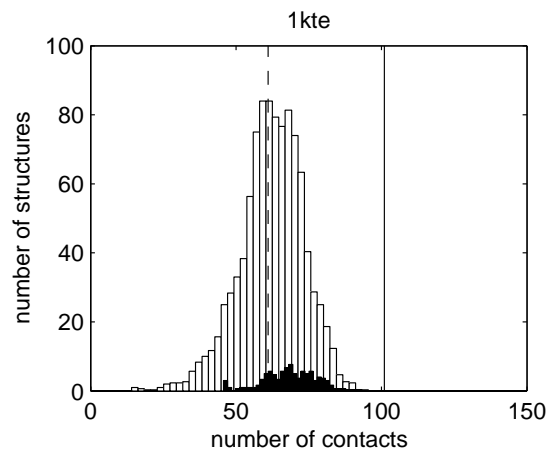
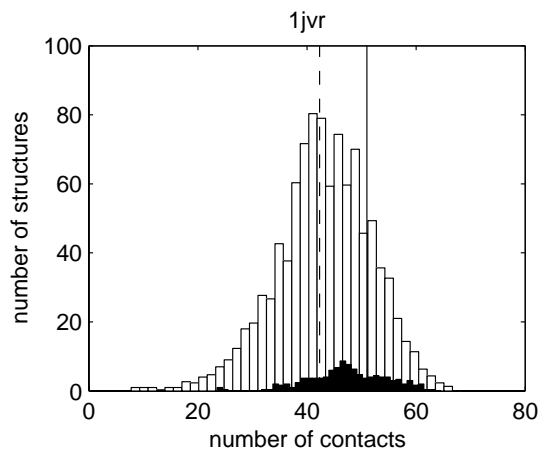
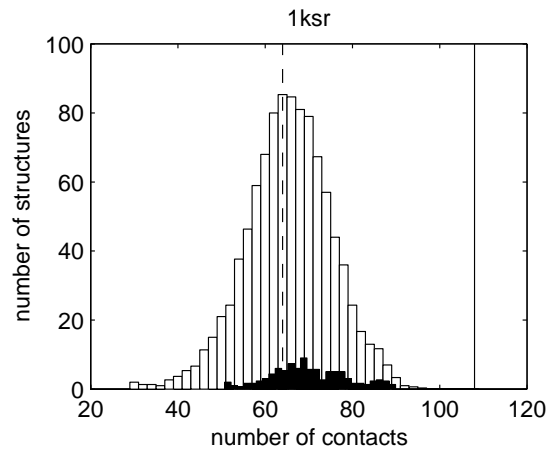
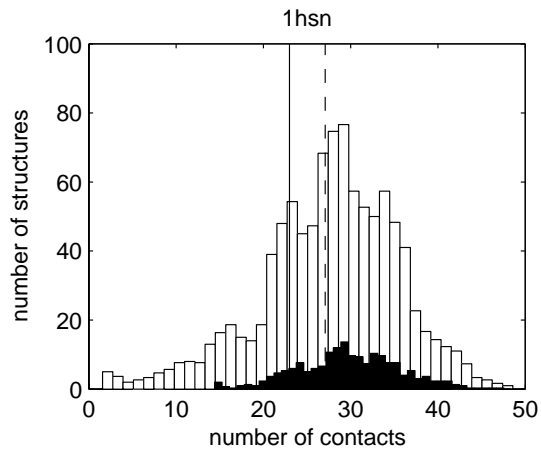
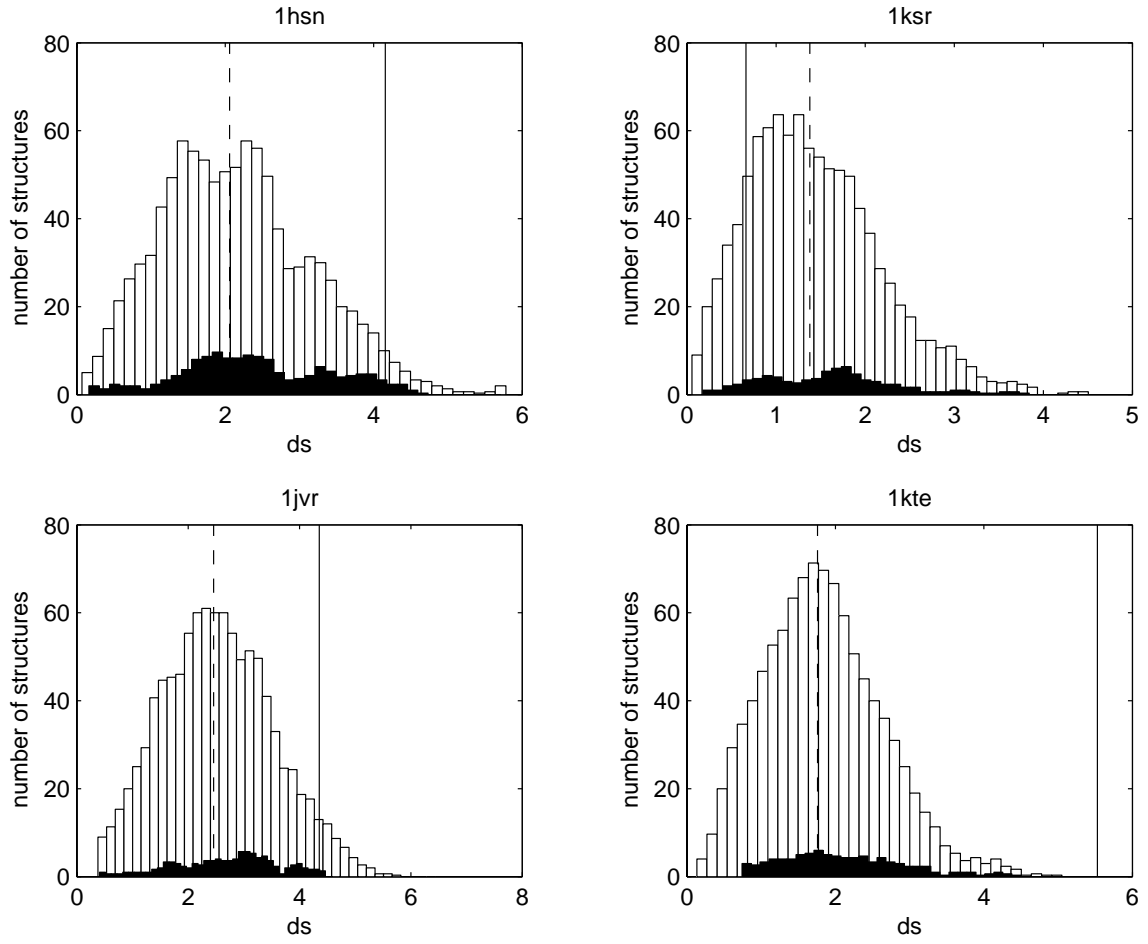


Figure 5.

A.



B.

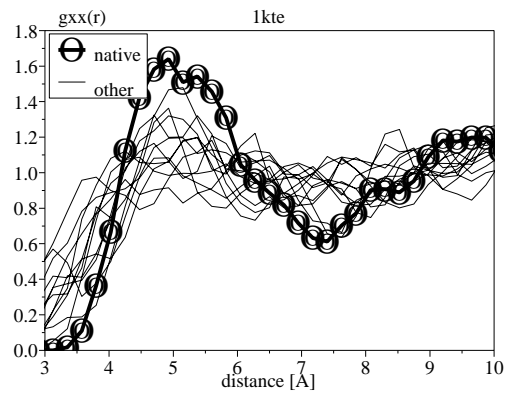
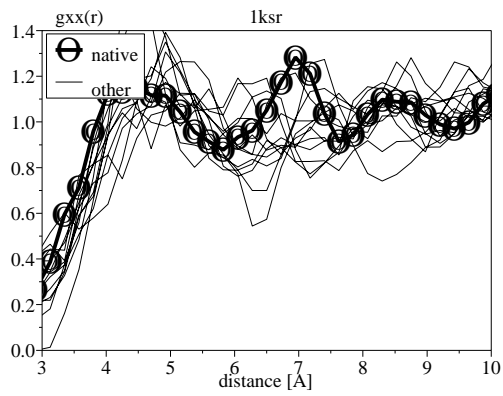
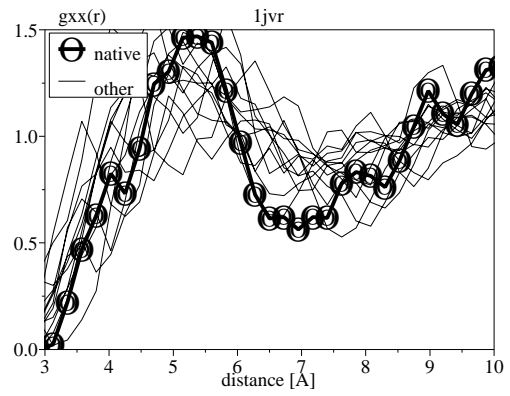
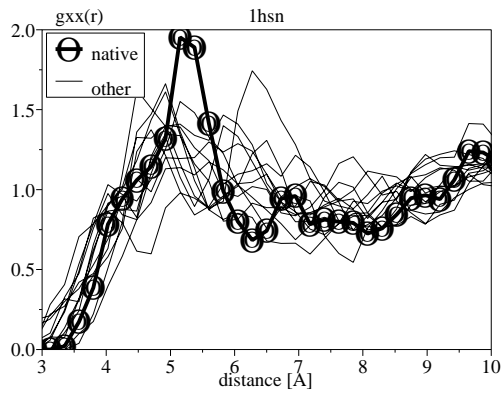


Figure 6.

