doi:10.1016/j.jmb.2004.07.038 J. Mol. Biol. (2004) 342, 651–664

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

# Analysis of Anisotropic Side-chain Packing in Proteins and Application to High-resolution Structure Prediction

**Kira M.S. Misura, Alexandre V. Morozov and David Baker***

*Department of Biochemistry University of Washington Box 357350, J-567 Health Sciences, Seattle, WA 98195-7350 USA*

**Corresponding author*

π–π, cation–π, and hydrophobic packing interactions contribute specificity to protein folding and stability to the native state. As a step towards developing improved models of these interactions in proteins, we compare the side-chain packing arrangements in native proteins to those found in compact decoys produced by the Rosetta *de novo* structure prediction method. We find enrichments in the native distributions for T-shaped and parallel offset arrangements of aromatic residue pairs, in parallel stacked arrangements of cation–aromatic pairs, in parallel stacked pairs involving proline residues, and in parallel offset arrangements for aliphatic residue pairs. We then investigate the extent to which the distinctive features of native packing can be explained using Lennard-Jones and electrostatics models. Finally, we derive orientation-dependent π–π, cation–π and hydrophobic interaction potentials based on the differences between the native and compact decoy distributions and investigate their efficacy for high-resolution protein structure prediction. Surprisingly, the orientation-dependent potential derived from the packing arrangements of aliphatic side-chain pairs distinguishes the native structure from compact decoys better than the orientation-dependent potentials describing π–π and cation–π interactions.



*Keywords:* side-chain packing; Rosetta; potential energy function; hydrophobic core; protein structure prediction

## Introduction

Close complementary packing between side-chains is a characteristic feature of native proteins.[1,2] To improve high-resolution structure prediction methods, it is necessary to develop accurate models of the forces that give rise to the side-chain packing arrangements in native structures, including π–π, cation–π and van der Waals interactions. An approach to evaluating existing force fields is to identify the features of side-chain packing in native protein structures which are distinct from those observed in random compact polypeptide chains, and to determine the extent to which these arrangements constitute local energy minima of the force field. To improve force fields and high-resolution structure prediction methods, the features of side-chain packing not captured by existing force fields can be incorporated into orientation-dependent side-chain interaction potentials.

π–π, cation–π and hydrophobic interactions play important roles in stabilizing protein structures and have been investigated using quantum chemistry calculations,[3–14] molecular mechanics calculations,[5,9,15,16] and protein structural analysis.[17–23] Quantum chemistry and molecular mechanics calculations have identified the energy minima for pairs of interacting residues, and protein structural analysis has described the distributions of interaction geometries for different residue pairs. Previously, we used density functional theory to compute dimer energy landscapes for pairs of ring-containing amino acids, and compared these landscapes to molecular mechanics energy landscapes.[6] We found that there was a reasonable correlation between the two landscapes, and that molecular mechanics landscapes in turn were correlated to some extent with the distributions of side-chain packing geometries in protein structures. This and the other studies[5,9] suggested that molecular mechanics

Present address: A. V. Morozov, Center for Studies in Physics and Biology, The Rockefeller University, Box 25, 1230 York Ave., New York, NY 10021 USA.

Abbreviations used: PDB, Protein Data Bank.

E-mail address of the corresponding author: dabaker@u.washington.edu

potentials capture π–π and cation–π interactions to an extent greater than might be expected for simple point charge-based models, but that there is still considerable room for improvement.

Due to the geometric constraints associated with compact structures, deviations from random side-chain orientation distributions might be expected in compact non-native polypeptide chain conformations as well as in native structures. Thus, the characteristic features of side-chain packing in native structures might be best highlighted by comparison to those in compact non-native conformations produced by *de novo* structure prediction methods rather than to purely random distributions. These "decoy" conformations for each protein sequence have approximately the same radius of gyration as native proteins with similar sequence lengths, and are composed of secondary structure elements that pack together in a manner similar to that seen in native structures. The side-chains in the decoy conformations are subjected to many of the constraints present in protein interiors: hydrophobic burial, van der Waals attractive forces and excluded volume constraints due to main-chain backbone atoms and other side-chain atoms.

Here, we compare the side-chain packing arrangements in native structures to those in a set of compact non-native conformations produced by the Rosetta *de novo* structure prediction method.[24–26] We then investigate the extent to which the Lennard-Jones and point charge based Coulomb potentials used in molecular mechanics force fields account for the native side-chain packing distributions. Finally, we develop an orientation-dependent potential based on the differences in side-chain packing between native structures and non-native decoy conformations, and investigate the efficacy of the resulting potential function for high-resolution structure prediction.

## Results

### Collection of statistics from native protein structures and Rosetta decoys

We determined the distribution of side-chain packing geometries for pairwise combinations of F,Y,W,H,P,R and F,L,V,I,W,Y in a set of 3500 crystal structures deposited in the Protein Data Bank (PDB) with resolutions of better that 2.5 Å and with less than 40% sequence identity with other proteins in the set. Each side-chain was represented by a plane through a set of at least three atoms (Figure 1B). For the non-aromatic hydrophobic amino acids, different planes can be defined depending on the choice of atoms used. We chose the atom set and planes they define to maximize the differences between the native and decoy distributions. We measured the angle between the planes and the horizontal and vertical offset of the neighboring side-chain with respect to the reference side-chain (θ, *r*1 and *r*2, respectively, Figure 1A). The centers of the side-chains were defined as the average of the coordinates used to define the plane. The *r*1 and *r*2 parameters are the horizontal and vertical components, respectively, of the distance between the averaged coordinates of the atoms used to define the plane.

The parameters θ, *r*1 and *r*2 were calculated for 104,620 hydrophobic-containing and 63,476 aromatic-containing side-chain pairs in the PDB subset, and 214,244 hydrophobic-containing and 128,321 aromatic-containing side-chain pairs in Rosetta. The native and compact decoy

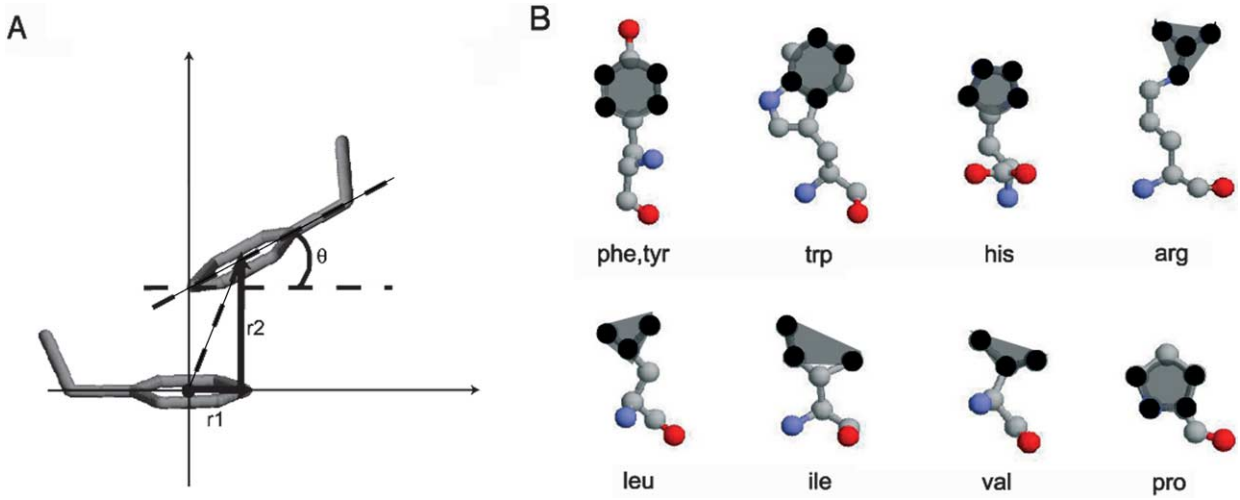


**Figure 1.** A, Schematic representation of the parameters used in derivation of the side-chain orientation potential. θ is the angle between planes of the side-chains as defined by three or more atoms, *r*1 and *r*2 are the horizontal and vertical components of the distance between the centers of the side-chain planar groups, respectively. B, Schematic of the atoms used to define planes in side-chains involved in this study. Atoms used to define the planes are shown in black, and the planes they define are shaded in gray. This Figure was made using Rasmol 2.7.1 (http://www.umass.edu/microbio/rasmol).

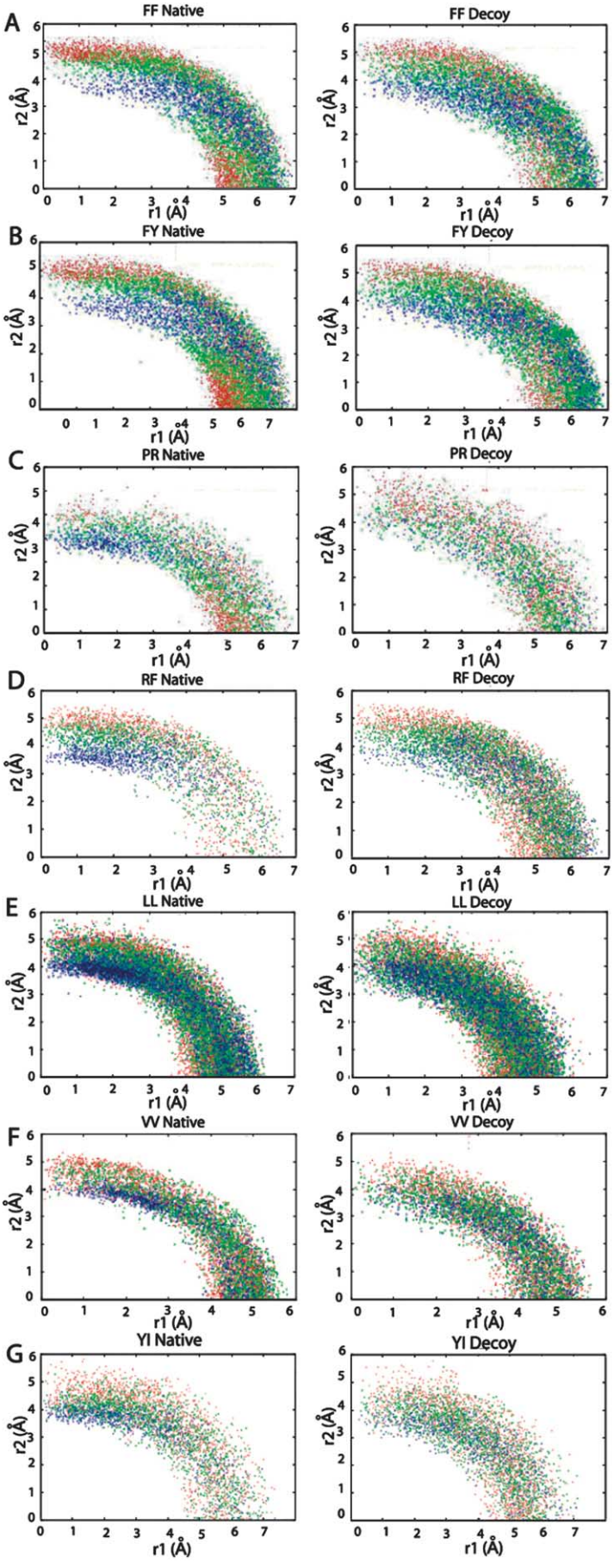


**Figure 2.** Distributions of the parameters θ, $r1$ and $r2$ of native proteins and Rosetta for selected aromatic and hydrophobic containing side-chain pairs. Blue, green and red represent the angle (θ) bins 0–30°, 30–60° and 60–90°, respectively. The horizontal axis represents the horizontal offset $r1$ of the plane centers (in Å), and the vertical axis represents the vertical offset $r2$ of the plane centers (in Å). This Figure was made using Gnuplot 4.0 (http://www.gnuplot.info).

**Table 1.** Percentage of counts in angle bins for selected side-chain pairs

| Side-chain pair | | No. of counts | 0–30° (%) | 30–60° (%) | 60–90° (%) |
|---|---|---|---|---|---|
| Phe-Phe | Native | 7023 | 17.5 | 34.4 | 48.1 |
| | Decoy | 7157 | 19.9 | 37.9 | 42.2 |
| Phe-Tyr | Native | 7627 | 17.4 | 33.1 | 49.5 |
| | Decoy | 7696 | 18.0 | 42.6 | 39.4 |
| Phe-Arg | Native | 3600 | 20.4 | 38.0 | 41.6 |
| | Decoy | 3711 | 11.3 | 40.8 | 47.9 |
| Pro-Arg | Native | 3011 | 28.7 | 33.9 | 37.4 |
| | Decoy | 3423 | 14.2 | 37.7 | 48.1 |
| Leu-Leu | Native | 15,099 | 22.0 | 37.5 | 40.5 |
| | Decoy | 40,524 | 15.8 | 36.5 | 47.7 |
| Val-Val | Native | 5943 | 16.4 | 32.5 | 51.1 |
| | Decoy | 16,610 | 14.8 | 36.8 | 48.3 |
| Leu-Phe | Native | 11,869 | 17.8 | 36.9 | 45.3 |
| | Decoy | 16,392 | 15.1 | 39.2 | 45.6 |
| Val-Phe | Native | 8045 | 20.1 | 34.1 | 45.8 |
| | Decoy | 13,287 | 18.1 | 36.8 | 45.1 |
| Random | – | – | 13.4 | 36.6 | 50.0 |

distributions are compared in Figure 2. For clarity, the three-dimensional native and decoy $r1$, $r2$ and θ distributions are separated into two-dimensional $r1$ *versus* $r2$ distributions for three different θ intervals (0–30° (blue), 30–60° (green) and 60–90° (red)). The populated region of the plots is restricted by atomic overlaps at small $r1$ and $r2$, and the cutoff criteria for contacting atoms at large $r1$ and $r2$. The native distributions generally cover a smaller area and form sharper peaks than the decoy distributions and are shifted towards the 0–30° θ bin more than the decoy distributions (Table 1). Both the native and decoy θ distributions are shifted towards small angles more than the sin distribution expected for randomly oriented planes. The preferences for parallel arrangements thus appear to be, at least in part, due to geometrical constraints present in compact polypeptide chains.

It is useful to distinguish several important regions in the θ, $r1$ and $r2$ distributions, which are highlighted in Figure 3A. The 0–30° bin consists of amino acid pairs with largely parallel planes. For $r1 = 0.0$ Å, the "parallel stacked" arrangement, the planes are parallel and their centers displaced only along the vertical axis $r2$. For $2\ \text{Å} < r1 < 4\ \text{Å}$, the "parallel offset" arrangement, the planes are stacked but the centers of the rings are displaced from one another along the $r1$ as well as $r2$ axes. In the 60–90° bin, the planes of the two amino acids are largely perpendicular to one another and the centers of the planes are displaced along the $r1$ axis. For the aromatic pairs FF and FY, the T-shaped arrangements are enriched in native structures compared to decoys (Figure 2A and B, clusters of red points near $r1 = 0.0$ Å, $r2 = 5.0$ Å and $r1 = 5.0$ Å, $r2 = 0.0$ Å). In contrast, the PR and RF pairs favor the parallel stacked and parallel offset conformations (Figure 2C and D, clusters of blue points at $r1 = 1.5$ Å, $r2 = 3.5$ Å). These conformations are enriched in the native structures relative to the compact decoys. These observations are consistent with quantum mechanics calculations, which have shown that the T-shaped and parallel offset arrangements of aromatic rings are considerably lower in energy than the parallel stacked conformations,[3,4,7] and that the T-shaped conformation is unfavorable for PP pairs.[6]

Among the aliphatic side-chain pairs, there are also notable differences between the native and decoy distributions (Figure 2). The LL and VV native pair distributions show a preference for the parallel offset over the parallel stacked conformation, as shown by the clusters of blue points in the LL and VV distributions for small $r2$ values. The side-chain pairs that contribute counts to the VV peak are typically located on β-strands and adopt parallel offset conformations (Figure 3B). The side-chain pairs that contribute to the LL peak can be found either on β-strands or α-helices and also adopt parallel offset conformations (Figure 3B). The YI pairs are enriched in the parallel stacked and parallel offset conformations relative to the compact decoys (Figure 2), and a similar trend can be observed for many other aromatic-hydrophobic pairs (data not shown). One notable difference between the native and decoy distributions for all hydrophobic pairs is the range of $r1$, $r2$ values. The decoy $r1$, $r2$ distributions frequently adopt smaller values than the native structures, and thus cover a larger area on the plots in Figure 2. This region represents close atomic contacts and seems to be more prominent for the pairs consisting of hydrophobic side-chain pairs than for those consisting of aromatic, arginine or proline residues.

## Comparison of native and compact decoy distributions to Lennard-Jones and Coulomb energy landscapes

The differences in side-chain packing observed between the native protein structures and Rosetta decoys suggested that high-resolution structure prediction in Rosetta could be improved by incorporating a potential function describing favorable side-chain orientations into the energy function. Significant differences are observed among amino acid pairs likely forming π–π and cation–π interactions, and it has been suggested that a

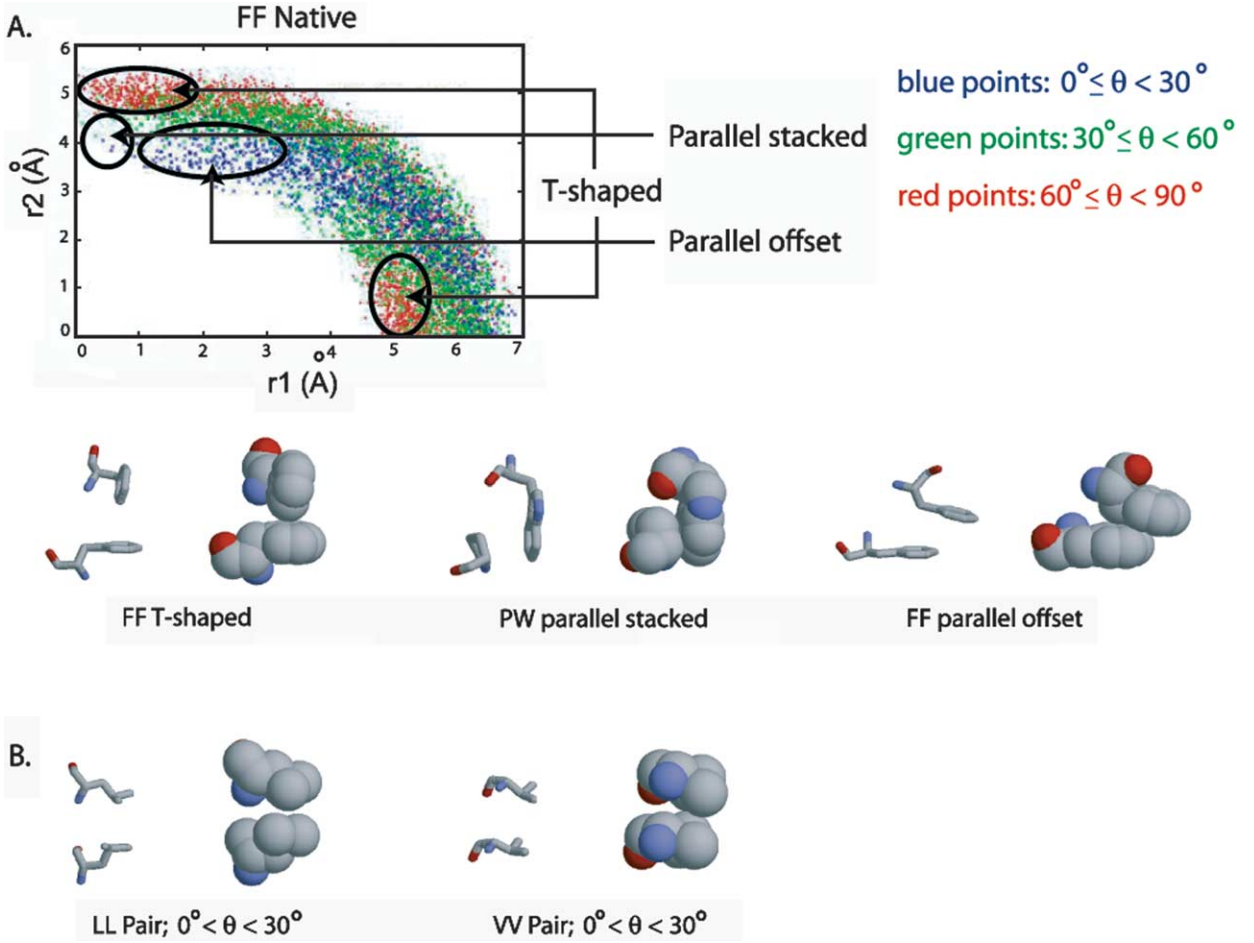


**Figure 3.** A, Illustration of regions of the distributions corresponding to T-shaped, parallel stacked and parallel offset comformations. Ball-and-stick and space-fill representations of selected side-chain pair orientations. T-shaped, parallel stacked and parallel offset conformations are shown for selected aromatic pairs and correspond to the labeled regions in the plot. This panel was made using Gnuplot 4.0 (http://www.gnuplot.info). B, Ball-and-stick and space-fill representations of favorable side-chain pair orientations for VV and LL pairs. The orientations shown correspond to the clusters of blue points (representing $0° < \theta < 30°$, 1.0 Å $< r1 <$ 2.5 Å, 3.5 Å $< r2 <$ 4.2 Å) in Figure 2E "native" and F "native". This Figure was made using Rasmol 2.7.1 (http://www.umass.edu/microbio/rasmol).

Coulomb model with partial charges on the aromatic carbon atoms can capture such interactions to some degree.[6] As the Rosetta potential includes a Lennard-Jones term but lacks an explicit electrostatic term, modeling interactions among aromatic residues may be improved by inclusion of a Coulomb term in the Rosetta potential. Therefore we first investigated the extent to which the native distributions and the differences between the native and decoy distributions could be explained using only a Lennard-Jones potential and Coulomb electrostatics.

We calculated the Lennard-Jones, Coulomb, and total CHARMM27[27] force field (Lennard-Jones plus Coulomb) dimerization energy landscapes *in vacuo* as a function of $r1$, $r2$ and $\theta$ using TINKER[28] and the CHARMM27[27] parameter set for each amino acid pair containing aromatic, cation or proline groups. We compared the landscapes to the native and decoy $r1$, $r2$ and $\theta$ distributions as well as to the differences between the native and decoy distributions (defined as the negative logarithm of $P_{native}/P_{decoy}$, see Materials and Methods) for the dimers. Assuming that side-chains interact primarily in a pairwise manner and are not influenced by higher-order effects, one might expect a correlation between the side-chain orientation distributions observed in native structures and the total dimerization energy landscapes. There is a reasonable correlation between the peaks in the native distributions (Figure 4A) and the minima in the total dimerization energy landscapes (Figure 4B) for the FF pairs but less so for the YY pairs.

Because side-chains interact in Rosetta predominately *via* a Lennard-Jones potential, we expect that the decoy distributions (Figure 4C) will resemble the Lennard-Jones potential energy landscape (Figure 4D). We also anticipate that the log ratio of the decoy and the native distributions (Figure 4E) will resemble the Coulomb energy landscape (Figure 4F), as the log ratio of the decoy and the native distributions should reflect the missing components in the Rosetta potential. Again, these

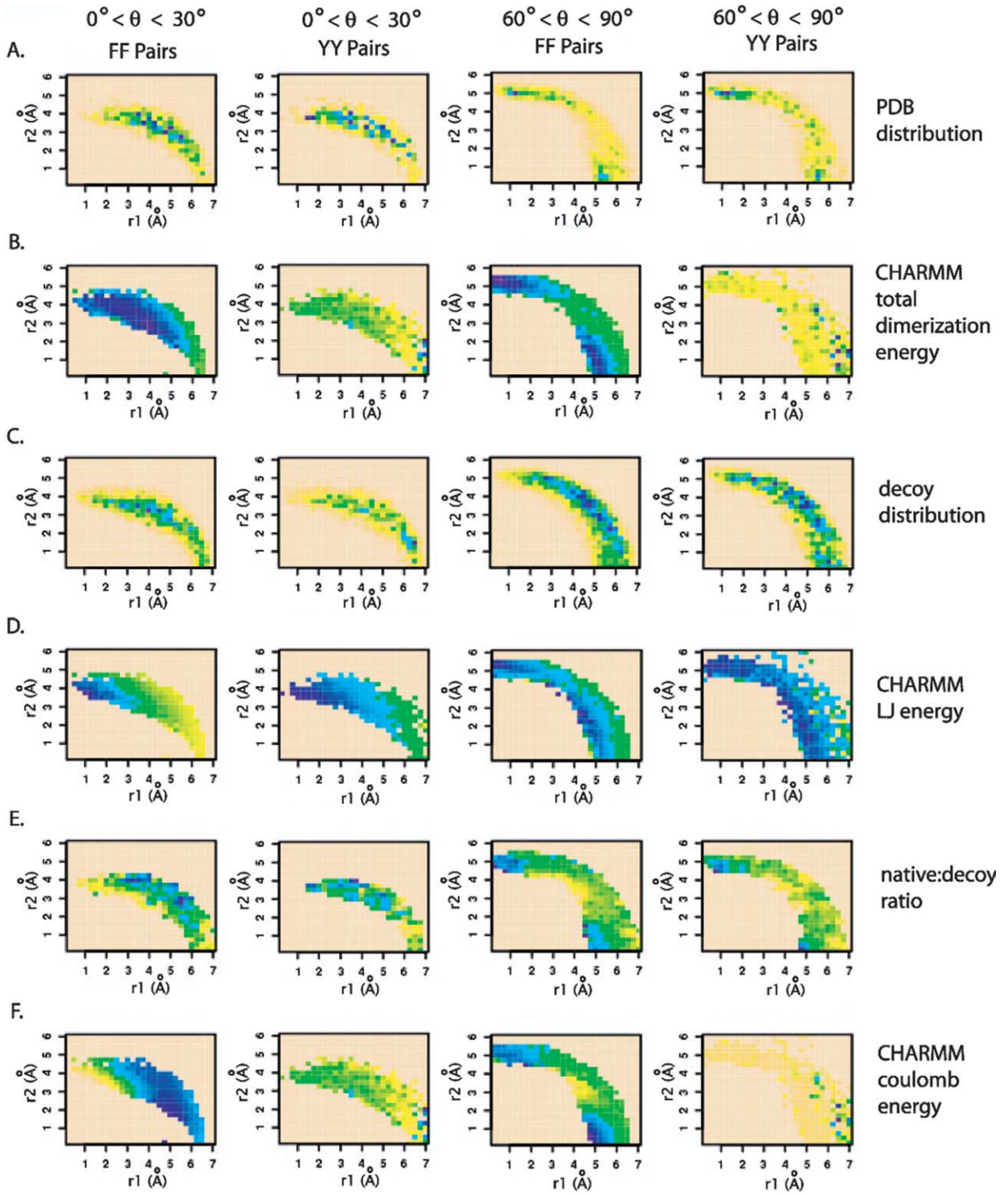


**Figure 4.** Comparison of molecular mechanics potential energy surfaces with PDB and decoy distributions for FF and FY pairs. Blue represents favorable energies, green represents neutral energies and yellow represents unfavorable energies (B, D, E, F). Blue, green and yellow represent highly populated, moderately populated, and less populated, respectively for A and C. The plots were calibrated so that the maximum value in the data set corresponds to yellow, the minimum value corresponds to dark blue, and the background value (colored peach) = 0. The color scale was determined using the TOPO color scheme in the R statistics package (http://www.r-project.org), with 24 colors representing equal intervals between the minimum and maximum values. The horizontal offset $r1$ and the vertical offset $r2$ in Å are plotted on the $x$ and $y$ axes, respectively. The left-hand panels show the 0–30° angle bins, the right-hand panels show the 60°-90° bins. For clarity, the 30–60° angle bins are not shown. This Figure and Figures 5 and 6 were made using the R statistics package (http://www.r-project.org).

expectations hold up reasonably well for the FF pair (columns FF $0° < \theta < 30°$ and FF $60° < \theta < 90°$), especially in the 90° angle bin where the T-shaped conformation is favored. The Coulomb energies and ratio between the native and decoy distributions are less similar for the YY pair (Figure 4). Although the FF and YY native distributions are remarkably similar (Figure 4A), the total dimerization energy landscape for the YY pair is different from the FF pair (Figure 4B). The discrepancies between the FF and YY dimerization landscapes arise primarily from the electrostatic model, as can be seen through the differences in the isolated charge–charge components for FF and YY (Figure 4F). Evidently the CHARMM27 partial charges assigned to the tyrosine hydroxyl group distort the reasonably accurate electrostatic description of the aromatic ring.

To extend this analysis to the non-polar side-chain pairs, we calculated the Lennard-Jones energies for all native LX, IX and VX side-chain pairs, where X is L, V, I, W, Y, F or P. We then compared the energy landscapes to the native distributions in order to determine whether the packing arrangements of non-polar side-chains that are enriched in native structures represent minima of the Lennard-Jones interactions between them. We used the same criteria to select contacting side-chain pairs as for collecting statistics from hydrophobic amino acid side-chains, and again we considered only buried residues for the analysis. The main-chain atoms C, O, N and $C^{\alpha}$ were excluded from the energy calculation. Pairs that contained very close atomic contacts in the crystal structures produced sharp peaks in the energy landscapes and therefore were discarded.

The results, which are shown in Figure 5, are quite surprising. We find that while the Lennard-Jones interaction energies are favorable over a broad range of packing orientations, only a small subset of these arrangements are frequently observed in native structures. For example, for LL pairs with roughly parallel planes, favorable Lennard-Jones interactions are made in both the parallel stacked and "side-by-side" packing arrangements, but the parallel stacked arrangement

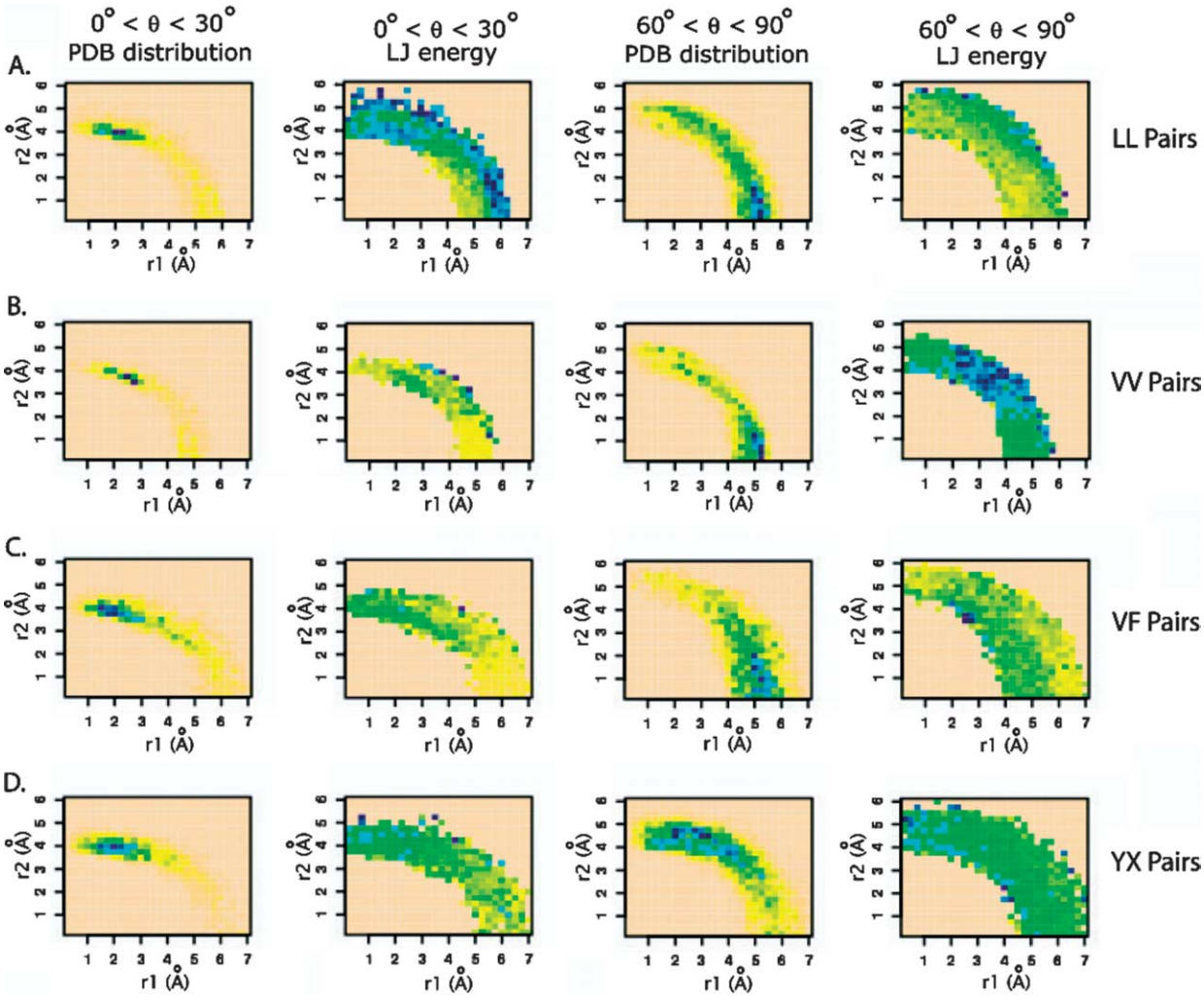


**Figure 5.** Comparison of native distributions and Lennard-Jones energy landscapes for selected hydrophobic pairs. The horizontal offset $r1$ and the vertical offset $r2$ in Å are plotted on the $x$ and $y$ axes, respectively. For clarity, only the 0–30° and 60–90° bins are shown. Color schemes and calibration are the same as in Figure 4.

is strongly preferred in native structures (Figure 2). This may indicate that the compact parallel stacked arrangement is more compatible with the formation of a tightly packed hydrophobic core composed of many side-chains than the more extended side-by-side packing arrangement. The native distributions for many side-chain pairs often have peaks corresponding to the T-shaped arrangement, but this arrangement is not particularly favorable in the Lennard-Jones energy landscape. This can be seen by comparing the distributions with the Lennard-Jones energy when $60° < \theta < 90°$ (Figure 5A–D). For the VV pairs, the peaks are somewhat anti-correlated. Again, the T-shaped arrangement may be favored in native structures because of its compatibility with overall core formation; both the stacked and T-shaped arrangements are relatively compact compared to other side-chain packing arrangements. These results suggest that the packing arrangements of aliphatic residues in native structures are not determined primarily by their pairwise interaction energies, but by the overall geometric constraints associated with formation of a tightly packed core.

### Derivation of an orientation-dependent side-chain–side-chain interaction potential

Given the imperfect correlation between the Lennard-Jones and Coulomb energy landscapes and distributions observed in the PDB, we derived a side-chain orientation potential in order to better describe the orientation dependence of π–π, cation–π and hydrophobic interactions in proteins. We used the differences between the native and decoy distributions of θ, $r1$ and $r2$ to construct a potential that rewards native-like side-chain pair geometries. For many amino acid pairs, the data were sparse and several regions of the distributions were not well-populated. π–π, cation–π and proline interactions occur far less frequently in proteins than interactions between aliphatic residues, especially for pairs that include a rare amino acid such as tryptophan. To compensate for this problem, counts for these pairs were first combined into four classes (see Materials and Methods) in order to ensure reliable statistics for each pair. In most cases, there were sufficient counts for the hydrophobic pairs that grouping into classes was not necessary. We did not correct for the dependence of the bin volume on θ and $r2$ (proportional to sin θ and $r2$, respectively), as we are primarily interested in comparing the native and decoy distributions, in which case they cancel. The combined counts were then binned and for each bin we computed:

$$P_{\text{native}}(\theta, r1, r2) = N(\theta, r1, r2)/N_{\text{total,native}}$$

and

$$P_{\text{decoy}}(\theta, r1, r2) = N(\theta, r1, r2)/N_{\text{total,decoy}}$$

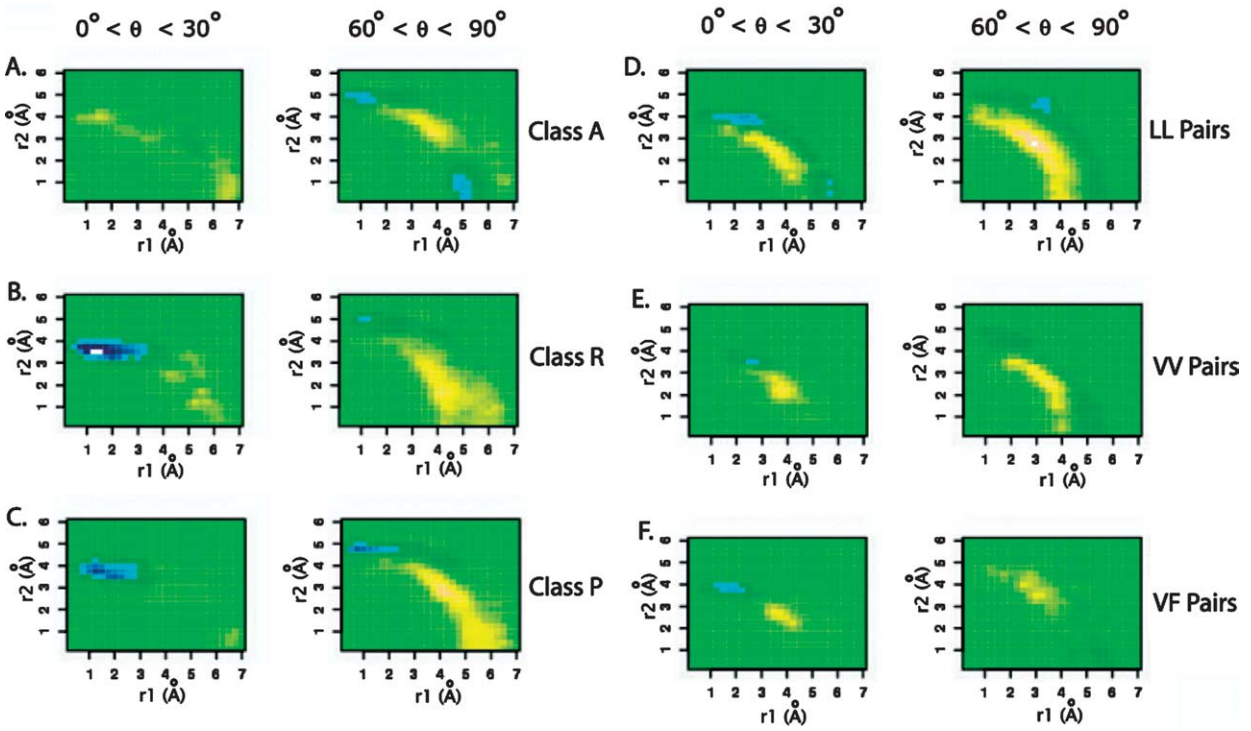


**Figure 6.** Energy landscapes of the derived side-chain orientation energy function for selected amino acid pairs. Class A includes FF, FY, FW, YY and YW pairs; class R includes RF, RW, RY, RR and RH pairs; class P includes FP, RP, HP, WP and YP pairs. The $x$ and $y$ axes represent the horizontal and vertical displacement, respectively. Green represents neutral energy (areas where there are either no counts from natives or decoys or where the distributions of the decoys match those of the native structure). Blue represents favorable energies (native counts are enriched relative to decoys), green represents neutral energies and yellow represents unfavorable energies (decoy counts are enriched relative to natives).

Finally, the potential was computed as:

$$E = -\ln(P_{\text{native}}/P_{\text{decoy}})$$

Figure 6 shows examples of the knowledge-based energy landscapes derived using this procedure.

### The side-chain orientation potential discriminates native structures from Rosetta decoys

We tested the new potential for its ability to discriminate native structures from Rosetta decoys. Side-chain interaction energies were calculated for a set of 54 native protein structures and 500 compact decoys built from the native sequence for each protein. Table 2 shows the percentage of the test set proteins for which the side-chain orientation potential ranked the native structure in the top 15% of the range of energies for each protein sequence. The total score is a sum of the energies contributed by all side-chain pairs; energies for each side-chain pair or class of pairs were evaluated separately in order to assess the relative energetic contribution of each type of side-chain interaction. The aromatic–aromatic and cation–aromatic energy components are somewhat successful at discriminating native structures from the corresponding set of decoys. Energies derived from pairs containing histidine are relatively unsuccessful, perhaps because frequent involvement of histidine in metal binding and choice of protonation states gives the ring a wider distribution of angles relative to other ring-containing side-chain pairs.

The terms representing hydrophobic and proline groups are significantly more successful at discriminating native structures from decoys than the aromatic and cation components. We considered for this analysis only hydrophobic pairs that were buried in the protein core; in this environment two side-chains are likely to have high surface complimentarity.[17] Well-packed hydrophobic cores are distinguishing features of native proteins;[29,30] the aliphatic component of the side-chain orientation potential may capture packing arrangements conducive to formation of tightly packed cores. The somewhat unexpected poorer performance of the π–π and cation–π interactions may reflect their relatively low abundance: the proteins in our test set contain significant numbers of hydrophobic interactions but only a few π–π or cation–π interactions. Hence, the fluctuations in the total energy for any individual structure are greater for the π–π or cation–π interactions. We tested this hypothesis by removing a random subset of the hydrophobic counts such that the numbers of hydrophobic and aromatic–aromatic interactions were equal, and found that the hydrophobic component of the side-chain orientation potential did not discriminate native structures from decoys nearly as well (data not shown).

**Table 2.** Native ranks (% of test set for which native receives score in the top 15%)

| | Side-chain-orientation energy | Coulomb energy |
|---|---|---|
| Aromatic | 43 | 17 |
| Histidine | 13 | 7 |
| Cation | 32 | 17 |
| Proline | 39 | 26 |
| Hydrophobic | 83 | n/a |
| Total | 92 | 30 |

As discussed above, the ratio of the native and compact decoy distributions resembles the Coulomb interaction energy between the pairs for some π–π and cation–π interactions. As the differences between the native and compact decoy distributions reflect deficiencies in the current Rosetta potential, we were interested in determining the effectiveness of a simple Coulomb model in distinguishing the native structure from decoys. If the Coulomb potential accurately accounted for the deficiencies in the Rosetta potential as well as the side-chain orientation potentials, both potentials in isolation should discriminate native structures equally well. This would also indicate that the side-chain orientation potentials are redundant with the Coulomb potential. To test this we computed the Coulomb energy for each side-chain pair using the distance-dependent dielectric of Warshel *et al.*,[31] and using the same criteria to select pairs of π–π and cation–π side-chains in native proteins and Rosetta decoys as to construct the side-chain orientation potential. We compared the results to those obtained from the side-chain orientation potential. Table 2 shows that the side-chain orientation potential can discriminate native structures from decoys more effectively than the energy calculated from the simple electrostatic model. This result is perhaps not surprising, as certain aromatic interactions appeared to be well modeled by the Coulomb plus Lennard-Jones model (Figure 4, FF pair) while others showed poor correlation (Figure 4, YY pair). The poorer performance of the electrostatic model may involve other charged groups, such as the tyrosine OH, which can perturb the interaction energy landscapes considerably.[6]

### Correlation between the side-chain orientation potential and van der Waals interactions

The side-chain orientation potential may capture interactions that could be modeled by a Lennard-Jones potential. This is evident in Figure 6D, where the unfavorable energies at *r*1–3.5 Å, *r*2–2.5 Å likely correspond to atom pairs that make close contacts. As the Rosetta potential energy function already includes a Lennard-Jones potential, we wanted to show that the side-chain orientation potential is not redundant with the Lennard-Jones potential. To determine the extent of the correlation between the two potentials, we computed the energies of the decoys and native structures for the 54 proteins in the test set using only the Lennard-Jones potential or a combination of the Lennard-Jones potential and

the side-chain orientation potential. For each of the 45 proteins where the energy of the native structure was lower than the decoys, we computed the energy gap between the native structure and the lowest scoring decoy normalized by the standard deviation of the decoy population. A large energy gap between the native structure and the lowest energy near-native decoy is a desirable feature because it indicates that the energy function effectively discriminates native structures from decoys.

Addition of the side-chain orientation energy to the Lennard-Jones energy gives rise to a larger energy gap between native structures and decoys on average, showing that there is additional information in the side-chain orientation model not captured by the Lennard-Jones model (Table 3). The energies of the hydrophobic, cation–π, π–π and proline-containing amino acid pairs are evaluated individually and combined. We did not observe significant differences in the performance of the individual components, and the combination of all components shows the largest enhancement of the energy gap. We performed a similar test using the Lennard-Jones energy plus the Coulomb energy computed as described above, and found that this combination did not significantly alter the energy gaps (Table 3).

## Discussion

The distributions of side-chain orientations found in native protein structures differ substantially from those in compact decoy conformations. The T-shaped arrangement is significantly enriched and the parallel stacked arrangement slightly depleted in the native structures compared with the decoys for aromatic side-chain pairs (Figure 2A and B). For side-chain pairs involving cation–π interactions, the parallel stacked arrangement is strongly enriched in native structures relative to decoys (Figure 2C). The differences are likely to reflect interactions between the π electrons as well as pairwise and higher-order packing effects. The arrangements that are disfavored in native structures appear to correspond with poor shape complementarity between side-chains. These differences suggested that structure prediction using Rosetta could be improved by inclusion of a term in the potential representing aromatic–aromatic and cation–aromatic interactions explicitly.

Although not investigated previously in such detail, many of the overall features of the aromatic–aromatic and cation–aromatic native distributions noted in the above paragraphs have been, to some extent, anticipated by previous studies. The differences between the native structures and compact decoys for pairs involving proline are less anticipated. For most pairs containing a proline side-chain, the parallel stacked arrangement is favored over other packing arrangements in native structures relative to decoys, as illustrated by the negative logarithm of the native-to-decoy ratio (Figure 6C). The origins of this strong preference are not entirely clear, but the preference for parallel stacked orientations among pairs containing proline residues may reflect favorable higher-order packing interactions.

There are also informative differences in the packing arrangements of aliphatic amino acid pairs in native structures and compact decoys. The conformations adopted most frequently in native structures correspond to only a subset of the arrangements with favorable pairwise Lennard-Jones interaction energies (Figure 5). For example, the parallel stacked and T-shaped arrangements of LL pairs occur much more frequently in native structures than other packing arrangements with equivalent Lennard-Jones interaction energies, perhaps because these relatively compact packing arrangements are most compatible with formation of a tightly packed hydrophobic core involving many residues. Overall, the observation of differences in packing arrangements in native structures for aliphatic amino acids is of interest as it suggests that there may be commonalities to the way aliphatic amino acids assemble to form the tightly packed cores of native proteins. Again, the comparison to the compact decoy conformations is useful in highlighting these features, as it controls for general packing constraints in compact polypeptide chains.

Are the features of native side-chain packing captured by the interaction energies of two isolated residues, or are higher-order interactions important as well? We have previously carried out studies in which high-level quantum mechanics calculations were used to determine the orientation dependence of the interaction energies of selected aromatic,

**Table 3.** Energy gap change between native structures and low energy compact decoys

| | Lennard-Jones (LJ) + side-chain orientation energy | | | Lennard-Jones (LJ) + Coulomb energy | | |
|---|---|---|---|---|---|---|
| | Increase (%) | Decrease (%) | Unchanged (%) | Increase (%) | Decrease (%) | Unchanged (%) |
| LJ + aromatic-X | 52 | 20 | 28 | 24 | 35 | 41 |
| LJ + histidine-X | 25 | 16 | 59 | 14 | 14 | 72 |
| LJ + cation-X | 41 | 13 | 46 | 22 | 18 | 60 |
| LJ + proline-X | 56 | 16 | 28 | 24 | 12 | 64 |
| LJ + hydrophobic | 65 | 35 | 0 | n/a | n/a | n/a |
| LJ + total | 72 | 26 | 2 | 40 | 40 | 20 |

polar and proline-containing amino acid pairs using the same three degrees of freedom (θ, *r*1, and *r*2) used here.[6] We also investigated the extent to which the CHARMM27 and OPLS-AA[32] total dimerization energy correlated with the more accurate quantum mechanics calculations and found reasonable agreement. If side-chain interactions can be accurately modeled using the pairwise molecular mechanics potentials, we anticipated a correlation between the CHARMM27 total dimerization energy landscapes and the geometrical distributions found in native structures. We tested this hypothesis here, and found correlations as well as differences. Solvent effects and the environment provided by neighboring side-chains are also likely to influence the parameters *r*1, *r*2 and θ. Modeling these complex multi-body interactions using pairwise potentials may be difficult, and in this context it may be particularly appropriate to model side-chain packing using knowledge-based potentials.

A satisfying result is that the Coulomb landscapes resemble the logarithm of the ratio of native and decoy distributions for FF pairs (Figure 4E and F). This validates one of the main premises underlying this work: that comparison of native and decoy distributions would highlight energetic contributions not captured by the current Rosetta force field. The example of YY pairs, however, shows the potential problems in incorporating a Coulomb treatment of interactions into the Rosetta force fields. The OH group and associated large partial charge dominates the Coulomb energy landscape, and is likely the reason why the total dimerization energy landscape is distorted away from the native distribution. This hypothesis is supported by the observation that the Coulomb and total dimerization energy landscapes are similar for FF but not for YY side-chain pairs. Optimizing the hydroxyl proton positions in the isolated dimers does not improve the correlation between the molecular mechanics landscape and the observed distributions of tyrosine-containing dimers.[6]

The lack of a term describing aromatic interactions in the current Rosetta force field in part gives rise to the differences in side-chain packing between native structures and decoys. We were therefore disappointed to find that the explicit descriptions of π–π and cation–π interactions for separate classes of side-chain pairs did not perform better in the native structure discrimination test (Table 2). This may reflect the relatively rare occurrence of such pairs in the native structures. Low counts for many side-chain pairs forced us to group the data into four classes instead of treating each pair separately, and contributed noise to the resulting potential function. Also, because of the small numbers, the total aromatic interaction energy can fluctuate widely for different conformations, and decoy conformations can by chance have more favorable energies than native structures. Another possible explanation for the marginal performance of the π–π and cation–π potentials is that there are likely to be energetic differences between the native structures and decoys more significant than those involving the π electrons of aromatic residues. While not alone sufficient to distinguish the native structures from decoys, incorporation of a description of aromatic interactions is likely to improve the accuracy of the Rosetta force field.

We were also surprised by the results with the aliphatic pairs, but this time more positively. We had assumed that the Lennard-Jones pairwise model was sufficient to capture aliphatic side-chain packing and therefore did not anticipate differences between the native and decoy distributions. Instead, we observed significant differences between the geometrical distributions of aliphatic pairs in native and decoy structures and, as indicated in Table 2, the aliphatic component of the side-chain orientation potential was considerably more effective in recognizing native structures than any of the other components. The aliphatic–aliphatic interaction potential is more effective in distinguishing native structures from decoys despite the inclusion of a Lennard-Jones interaction term in the Rosetta force field; decoy conformations may make favorable pairwise Lennard-Jones interactions but have overall less well-packed cores (K.M.S.M. & D.B., unpublished results). We also showed that this component was independent, to some extent, from the Lennard-Jones potential. Isoleucine, leucine, methionine and valine are relatively abundant, thus the distributions for hydrophobic side-chain pairs are better sampled than those produced by the aromatic, proline and arginine-containing pairs. This allowed us to treat each side-chain pair explicitly rather than grouping counts from different side-chain pairs with visibly similar distributions, as was done for the aromatic, arginine and proline-containing pairs. These factors contributed to higher quality data and thus more reliable energies derived for the buried hydrophobic pairs. Additionally, unlike the aromatic–aromatic and cation–aromatic side-chain interactions, all compact structures in our test set contain large numbers of aliphatic interactions. Therefore the decoy energy distribution will be sharper for the aliphatic pairs than for pairs involved in cation–π or π–π interactions.

## Conclusions

The detailed comparison of side-chain packing arrangements in native structures with compact non-native structures has highlighted the distinctive features of side-chain packing in proteins. Some of these features, such as the enrichment of parallel offset arrangements of aromatic rings, can be partly explained by the interaction energy landscape of the isolated residue pair, and are consistent with Coulomb and Lennard-Jones models. Other features, notably those involving aliphatic pairs, are not well

accounted for by the interaction energies of the isolated pairs and are likely to reflect higher-order packing constraints. Future research will involve combining the knowledge-based potentials derived here with the full Rosetta potential function, and we are hopeful that this will improve high-resolution protein structure prediction. We anticipate that detailed comparison of the properties of native structures with those of compact non-native decoys will continue to reveal routes to improving the current generation of force fields.

## Materials and Methods

### Datasets

Statistics for angle and distance distributions were taken from atomic coordinates deposited in the Protein Data Bank. The set consisted of 3500 protein structures experimentally determined using X-ray crystallography at 2.5 Å or better resolution, from which the sequences used for the decoy set have been removed. Sequences in this set were no more than 40% identical with any other sequence in the set. Decoy structures were initially generated using fragment insertion as implemented in Rosetta with side-chains represented as centroids.[24] Side-chains were then added and the energies of the full atom representation of the models were minimized by repacking the side-chains[33] and optimizing the backbone torsion angles using a Monte Carlo procedure coupled to the full-atom energy function (Rohl *et al.*[26] and K.M.S.M. & D.B., unpublished results). The side-chain orientation potential was not included in the energy function used to generate the decoy sets. The proteins in the decoy set consisted of approximately 500 decoys produced for each of the 54 proteins ranging from 36 to 127 residues in length.

### Constructing the side-chain orientation potential function

Side-chains are included in the analysis if at least one non-hydrogen atom–atom contact is made between the residues, excluding backbone atoms and $C^{\beta}$. Pairs that contain histidine must have at least three contacts, and pairs with both residues hydrophobic must have four contacts. A contact is defined as two heavy atoms separated by a distance of 4.2 Å or less, which was determined empirically to be the value that maximizes the differences between the native and decoy distributions. We considered for our analysis aromatic, proline and arginine residues in any environment but only buried hydrophobic side-chains. We defined a side-chain pair as buried if both members have more than 18 neighboring centroids within 6 Å, where the centroid is defined as the average of the atomic coordinates for all non-main-chain atoms in a given residue. The distance and number of contacts parameters were chosen in order to maximize the differences between native and compact decoy distributions.

For amino acid pairs that meet these criteria, the plane of each side-chain was defined using the atoms illustrated in Figure 1B. The angle θ between the planes of the side-chains is given by the dot product of the unit vectors normal to the planes of the side-chains. $r1$ and $r2$ are the horizontal and vertical components, respectively, of the distance between the averaged coordinates of the atoms used to define the plane. We divided the area defined by these parameters into bins and computed the probabilities for each bin. We experimented with bin sizes for each of the three parameters, attempting to extract as much information as possible from the data while retaining enough counts in each bin to create a smooth energy landscape. The angles 90–120°, 120–150° and 150–180° were folded into 60–90°, 30–60° and 0–30° θ bins, respectively; not doing so decreased the counts in each bin, thus increasing noise in the data, and did not appear to add additional information. Counts from the side-chain pairs were binned into three angle bins (0–30°, 30–60°, and 60–90°) and 0.25 Å horizontal and vertical distance bins. The energy for each bin was defined as:

$$-\ln(P_{\text{native}}/P_{\text{decoy}})$$

where $P_{\text{native}}$ and $P_{\text{decoy}}$ are the normalized probabilities of each bin for native and decoy side-chains. Due to low counts for several side-chain pairs in the native and decoy sets, some pairs were classified into groups and their counts combined based on visual inspection of the native distributions. The classes consisted of FF, FY, FW, YY and YW (class A), RF, RH, RW, RY and RW (class R), FP, HP, RP, WP and YP (class P), HF, HW and HY (class H) for aromatic pairs. Interactions between PP pairs were excluded. For pairs contributing to the hydrophobic potential, the classes consisted of YI, YL and YV (class Y) and WI, WL and WV (class W). LL, LI, LV, LF, II, IV, IF, VV and VF pairs were kept separate. Only hydrophobic pairs for which the native distributions looked significantly different from the decoy distributions were included in the potential. Bins containing fewer than 25 counts for natives and decoys were set to 0.0 to reduce noise.

### Ranking native structures

The decoy set used to rank the energies of native structures consisted of 500 decoys for each of 54 small proteins with sequences ranging from 36 to 127 amino acid residues. The set of sequences was independent from the set used to construct the side-chain orientation potential. The test set was chosen to include a range of near-native (<4 Å rmsd) and non-native (>4 Å rmsd) decoys. For each protein, the energies of the native structure and the decoys were evaluated and each decoy was assigned a rank from lowest to highest energy. The test was considered successful if the energy of the native protein ranked in the lowest 15% of energies. The test was performed using individual components and the sum of the components of the side-chain orientation potential, as well as the electrostatic energy. Pairwise electrostatic energies used for the native discrimination test were calculated using the Warshel model[31] with a distance-dependent dielectric constant parameterized by Morozov *et al.*:[34]

$$E = 322.0637(q1q2)/(r\varepsilon(r)), \quad \textit{if } r < 3, \quad \varepsilon(r)$$

$$= 16.55, \quad \text{else} \quad \varepsilon(r) = 1 + 60(1 - \exp(-0.1r))$$

where $r$ is the distance between two atoms. Energies between two atoms residing on the same side-chain were excluded. The energies were summed over all pairs in the protein sequence.

To calculate the change in energy gap between native structures and near-native decoys, each native structure was first subjected to bond length and angle idealization followed by energy minimization according to the

protocol used for the decoy models. Then the difference between the Lennard-Jones energies of the minimized native structure and the lowest scoring decoy was calculated. We normalized this difference by the standard deviation of the decoy energies. The normalized magnitude of this gap was compared to that which was obtained when the side-chain orientation energy was added to the Lennard-Jones energy of the minimized native and near-native decoys. The test was considered successful for proteins that had a larger energy gap with the side-chain orientation term included.

All molecular mechanics calculations were done with the TINKER 4.0 molecular modeling package.[28] In the calculations of the Lennard-Jones energy using Rosetta, pairs that contained close atomic contacts in the native structures produced sharp peaks in the energy landscape, which obscured other features. The maximum allowable energy value for a given amino acid pair type was defined as the magnitude of the lowest energy value observed for that pair. Pairs that had energies in excess of this value were discarded from the analysis.

## Acknowledgements

The authors thank Dylan Chivian, Bill Schief, and members of the Baker laboratory for helpful advice, discussion and comments on the manuscript. We also acknowledge Keith Laidig for expert management of computational resources. K.M.S.M. gratefully acknowledges funding from the Helen Hay Whitney Foundation. A.M. and D.B. were supported by the Howard Hughes Medical Institute.

## References

1. Crick, F. H. C. (1953). The packing of α-helices: simple coiled-coils. *Acta Crystallog.* **6**, 689–697.
2. Richards, F. M. (1974). The interpretation of protein structures total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1–14.
3. Chipot, C., Jaffe, R., Maigret, B., Pearman, D. A. & Kollman, P. A. (1996). Benzene dimer: a good model for pi–pi interactions in proteins? A comparison between the benzene and the toluene dimers in the gas phase and in an aqueous solution. *J. Am. Chem. Soc.* **118**, 11217–11224.
4. Gervasio, F. L., Chelli, R., Procacci, P. & Schettino, V. (2002). Is the T-shaped toluene dimer a stable intermolecular complex? *J. Phys. Chem. A*, **106**, 2945–2948.
5. Gervasio, F. L., Procacci, P., Cardini, G., Guarna, A., Giolitti, A. & Schettino, V. (2000). Interaction between aromatic residues. Molecular dynamics and *ab initio* exploration of the potential energy surface of the tryptophan-histadine pair. *J. Phys. Chem. B*, **104**, 1108–1114.
6. Morozov, A. V., Misura, K. M. S., Tsemekhman, K. & Baker, D. (2004). Comparison of quantum mechanics and molecular mechanics based energy landscapes for interactions between pairs of ring containing amino acids in proteins. *J. Phys. Chem. B*, **108**, 8489–8496.
7. Sinnokrot, M. O., Valeev, E. F. & Sherill, C. D. (2002). Estimates of the *ab initio* limit for π–π interactions: the benzene dimer. *J. Am. Chem. Soc.* **124**, 10887–10893.
8. Chelli, R., Gervasio, F. L., Procacci, P. & Schettino, V. (2001). Stacking and T-shape competition in aromatic–aromatic amino acid interactions. *J. Am. Chem. Soc.* **124**, 6133–6143.
9. Gervasio, F. L., Chelli, R., Procacci, P. & Schettino, V. (2002). The nature of intermolecular interactions between aromatic amino acid residues. *Proteins: Struct. Funct. Genet.* **48**, 117–125.
10. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.
11. Hobza, P., Selzle, H. L. & Schlag, E. W. (1996). Potential energy surface for the benzene dimer. Results of *ab initio* CCSD(T) calculations show two nearly isoenergetic structures: T-shaped and parallel displaced. *J. Phys. Chem.* **100**, 18790–18794.
12. Hobza, P., Selzle, H. L. & Schlag, E. W. (1994). Potential energy surface of the benzene dimer: *ab initio* theoretical study. *J. Am. Chem. Soc.* **116**, 3500–3506.
13. Hobza, P., Selzle, H. L. & Schlag, E. W. (1993). New structure for the most stable isomer of the benzene dimer: a quantum chemical study. *J. Phys. Chem.* **97**, 3937–3938.
14. Tsuzuki, S., Honda, K., Uchimaru, T., Mikami, M. & Tanabe, K. (2002). Origin of attraction and directionality of the π/π interaction: model chemistry calculations of benzene dimer interaction. *J. Am. Chem. Soc.* **124**, 104–112.
15. Alagona, G., Ghio, C. & Monti, S. (1999). *Ab initio* study of preferential interactions between aromatic side chains. *Int. J. Quantum Chem.* **73**, 175–186.
16. Alagona, G., Ghio, C. & Monti, S. (1998). *Ab initio* investigation of the methylimidazole-indole complexes as models of the histidine–tryptophan pair. *J. Phys. Chem A*, **102**, 6152–6160.
17. Banerjee, R., Sen, M., Bhattacharya, D. & Saha, P. (2003). The jigsaw puzzle model: search for conformational specificity in protein interiors. *J. Mol. Biol.* **333**, 211–226.
18. Dougherty, D. A. (1996). Cation–π interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* 1996;, 271.
19. Hunter, C. A., Singh, J. & Thornton, J. M. (1991). Pi–pi interactions: the geometry and energetics of phenylalanine–phenylalanine interactions in proteins. *J. Mol. Biol.* **218**, 837–846.
20. McGaughey, G. B., Gagne, M. & Rappe, A. K. (1998). Pi-stacking interactions. *J. Biol. Chem.* **273**, 15458–15463.
21. Mitchell, J. B. O., Laskowski, R. A. & Thornton, J. A. (1997). Non-randomness in side-chain packing: the distribution of interplanar angles. *Proteins: Struct. Funct. Genet.* **29**, 370–380.
22. Gallivan, J. P. & Dougherty, D. A. (1999). Cation–π interactions in structural biology. *Proc. Natl Acad. Sci.* **96**, 9459–9464.
23. Brocchieri, L. & Karlin, S. (1994). Geometry of interplanar residue contacts in protein structures. *Proc. Natl Acad. Sci. USA*, **91**, 9297–9301.
24. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.
25. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). Improved recognition

of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct. Funct. Genet.* **34**, 82–95.
26. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93.
27. MacKerell, A. D. J., Bashford, D., Bellot, M., Dunbrack, R. L., Jr, Evanseck, J. D., Field, M. J. *et al.* (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.
28. Ponder, J. W. & Richards, F. M. (1987). An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.* **8**, 1016–1026.
29. Chothia, C. (1984). Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **53**, 537–572.
30. Liang, J. & Dill, K. A. (2001). Are proteins well packed? *Biophys. J.* **81**, 751–766.
31. Warshel, A. & Russel, S. T. (1984). Calculations of electrostatic interactions in biological systems and in solutions. *Quart. Rev. Biophys.* **17**, 283–422.
32. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225–11236.
33. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci.* 2000;, 97.
34. Morozov, A. V., Kortemme, T. & Baker, D. (2003). Evaluation of models of electrostatic interactions in proteins. *J. Phys. Chem. B*, **107**, 2075–2090.

***Edited by J. Thornton***

*(Received 9 April 2004; received in revised form 6 July 2004; accepted 8 July 2004)*