BIOCHEMISTRY
including biophysical chemistry & molecular biology

New Concept

pubs.acs.org/biochemistry

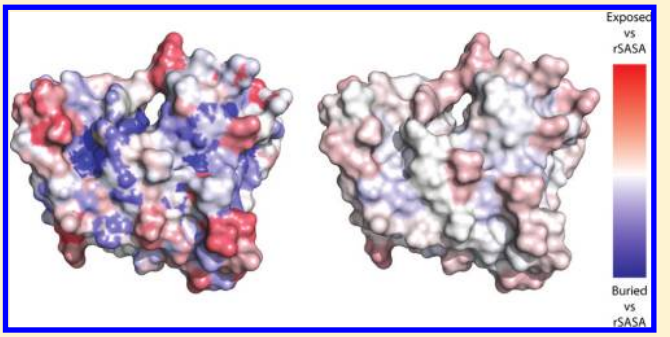
Design of Native-like Proteins through an Exposure-Dependent Environment Potential

Samuel DeLuca, Brent Dorr,[†] and Jens Meiler*

Departments of Chemistry, Pharmacology, and Biomedical Informatics, Center for Structural Biology, Institute for Chemical Biology, Vanderbilt University, Nashville, Tennessee 37212, United States

Supporting Information

ABSTRACT: We hypothesize that the degree of surface exposure of amino acid side chains within a globular, soluble protein has been optimized in evolution, not only to minimize the solvation free energy of the monomeric protein but also to prevent protein aggregation. This effect needs to be taken into account when engineering proteins de novo. We test this hypothesis through addition of a knowledge-based, exposure-dependent energy term to the RosettaDesign solvation potential [Lazaridis, T., and Karplus, M. (1999) *Proteins* 35, 133−152]. Correlation between amino acid type and surface exposure is determined from a representative set of experimental protein structures. The amino acid solvent accessible surface area (SASA) is estimated with a neighbor vector measure that increases in accuracy compared to the neighbor count measure while remaining pairwise decomposable [Durham, E., et al. (2009) *J. Mol. Model.* 15, 1093−1108]. Benchmarking of this potential in protein design displays a 3.2% improvement in the overall sequence recovery and an 8.5% improvement in recovery of amino acid types tolerated in evolution.



Computational design of proteins is an active area of research. The design of protein surfaces with proper amino acid composition is critical to preventing aggregation and allowing for correct protein folding.[3] Thermostabilization of enzymes and design of proteins with novel folds are two possible applications of this research.

As there are relatively few explicit interactions of amino acids on the protein surface, the total energy of a residue is dominated by Rosetta's implicit solvation model. The solvation model currently used by Rosetta is a function developed by Lazaridis and Karplus.[1] This potential estimates the solvation free energy of an atom from a reference free energy, where the atom is essentially fully solvent-exposed. For every proximal atom, a cost of "desolvation" is added in a pairwise decomposable and distance-dependent manner. This procedure aligns with the protein folding process, in which amino acids move from a completely exposed location (reference state) into varying degrees of burial. While the model is parametrized for all amino acid atom types, it is driven by high desolvation penalties of polar atoms. It is quite insensitive to the burial of apolar atoms because desolvation energies are small.

This paradigm of desolvation is useful for determining energy changes in the folding of a monomeric protein. However, hydrophobic patches on the surface of a de novo-designed protein are hardly penalized, as the environment of these amino acids did not change in the folding process. At present, RosettaDesign excels in the design of tightly packed protein cores, while the protein surface is often poorly composed and requires manual adjustment.[4] We hypothesize that native proteins have evolved to minimize unspecific aggregation, a fact that is ignored by the desolvation potential. Evolutionary pressures exerted on protein sequence composition by the requirement of protein solubility are difficult to model with a typical physics-based model but can be modeled effectively with a knowledge-based energy potential.

The RosettaDesign energy function is a weighted composite of the Lazaridis−Karplus solvation free energy potential, attractive and repulsive interactions, an action center pairwise potential to approximate electrostatic interactions, an orientation-dependent hydrogen bonding potential,[5] and reference energies for amino acid type and conformation.[4] Amino acid reference energies and scoring function weights are optimized to maximize sequence recovery in a protein design benchmark. Reference energies can be viewed as the ground state energy of an amino acid in an essentially fully exposed, unfolded peptide chain. Hence, these reference energies can disfavor apolar amino acids on the surface, thereby representing some of the evolutionary pressure to prevent aggregation. However, the same reference energies are also fitted to reflect amino acid propensities in nature in a manner independent of burial. In addition, the reference energies are fitted to maximize sequence recovery and thereby counterweight other inaccuracies in the

Received: May 4, 2011
Revised: August 19, 2011
Published: September 9, 2011

RosettaDesign energy function. As a result, the reference energies form a container term that combines multiple effects that can be difficult to disentangle, and it provides a corrective power against exposed hydrophobic amino acids on the surface.

To improve upon the shortcomings of RosettaDesign described above, we implemented the neighbor vector (NV)-based knowledge-based potential (KBP) previously described by Durham et al.[2] This neighbor vector environment KBP converts the likelihood to see an amino acid at a given level of exposure into an environment energy. The NV environment KBP encapsulates both desolvation energy and evolutionary biases against apolar amino acids at the protein surface with amino acid-level resolution.

The usefulness of an environment potential based on burial is contingent on an accurate measure of burial. Solvent accessible surface area (SASA) is the most accurate means of calculating amino acid burial but is generally time-consuming to compute, limiting its usefulness in protein design. RosettaDesign currently uses a neighbor count method (NCR) for estimating solvent accessibility in the pair potential. While the NCR method correlates with residue burial, high inaccuracies are common in surface and partially exposed positions (Figure 1).

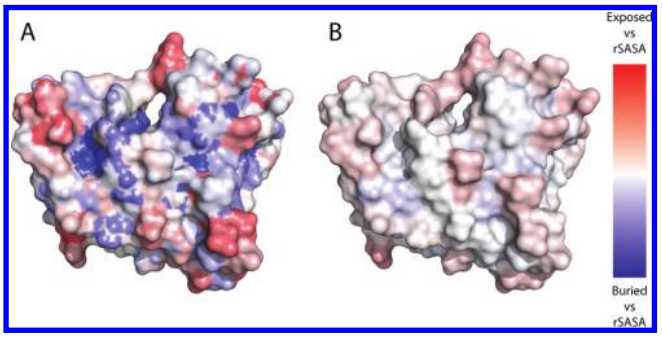


Figure 1. Comparison of NV and NCR measures to rSASA. In both panels, a color map plots the difference between a surface approximation method and the normalized rSASA value. A residue for which the SASA approximation matches rSASA exactly would have a score of 0 and be colored white. Regions of the surface colored red are categorized as more solvent exposed than by rSASA, while regions colored blue are categorized as less solvent exposed than by rSASA. (A) Protein 7DFR colored by the neighbor count approximation of surface accessibility as used in Rosetta (NCR). (B) Protein 7DFR colored using the neighbor vector (NV) approximation. The NV measure has significantly smaller deviations from the rSASA standard with a mean of 0.14 compared to the mean deviation of 0.20 seen with the NCR measure. Additionally, the NV measure is more consistent, with a standard deviation of 0.11 compared to the standard deviation of 0.46 seen with the NCR metric. Panels A and B illustrate the improvement in consistency, as areas of score deviation in panel B are smaller and generally less "patchy" in their appearance.

To overcome the limitations of the NCR burial approximation, an NV approximation of residue burial was implemented. For a schematic representation of the NV algorithm, see Figure 2 of the Supporting Information. The NV algorithm and KBP generated and described by Durham et al.[2] was used in our implementation. Proteins selected for deriving the KBP were monomeric, globular proteins, which do not engage in obligate, and therefore strong, protein−protein interactions. It is expected that some of these proteins will engage in transient interactions with other proteins; however, these interactions will be weaker. As a result, the noise added to the KBP by these interactions will be of low magnitude and uniform.

The half-sphere approximation method developed by Hamelryck in 2005[6] approximates surface accessibility by counting the number of residues in a half-sphere below the side chain of each amino acid. The half-sphere count is directly related to residue burial. Half-sphere exposure (HSE) is implemented in the freely available BioPython library, and this library was used to compare performance of HSE and NV to relative SASA (calculated using NACCESS). The per residue exposure was calculated for each of the proteins in the 42-protein benchmark set, and adjusted R^2 values were calculated for the correlation of each measure to relative SASA (rSASA). The adjusted correlation factor R^2 value for HSE versus rSASA was 0.68, while the adjusted correlation factor R^2 for the NV method was 0.86. This suggests that while HSE is conceptually simpler, it does not perform as accurately as NV for proteins in our benchmark set.

A linear regression modeling the correlation between rSASA values (in the range of 0−1 calculated using NACCESS) and NV score (range of 0−1) was generated on the basis of all proteins in the 42-protein benchmark set. The resulting linear regression model was rSASA = 1.29(NV) − 0.11 and had an R^2 of 0.86. On the basis of this model, residues with NV scores between 0.00 and 0.24 will have an approximate rSASA value of 0−0.19, residues with NV scores from 0.25 to 0.39 will have an approximate rSASA value of 0.21−0.39, and residues with NV scores between 0.40 and 1.00 will have an approximate rSASA value of 0.40−1.10.

Terms in the RosettaDesign energy function can take the form of either single-body or two-body terms. Two-body terms describe energies that pertain to the interaction between residues, such as the energy associated with hydrogen bonding, while single-body terms describe energies that pertain only to a single residue. The resulting NV environment KBP was implemented as a single-body term in the RosettaDesign energy function. RosettaDesign revision 39040 was used in all calculations.

Computationally assessing the performance of a protein design algorithm is inherently challenging. Historically, percent sequence recovery has been used as a metric for the quality of a protein design, as it has been observed that protein sequences are frequently close to optimal for a given fold.[7] However, many protein folds having large variations in sequence are frequently seen in nature.[8] Of the 74608 protein chains present in the Structural Classification of Proteins (SCOP) database as of 2009, only 1280 individual folds are observed.[9] In many positions, particularly on the surface of proteins, multiple residues can be tolerated with similar energies. This finding limits sequence recovery as a measure for successful protein design because the design of a different but tolerated amino acid is counted as a failure. To resolve this problem, we introduce a metric based on sequence homology. A Position Specific Scoring Matrix (PSSM) is derived from a Basic Local Alignment Search Tool (BLAST) query of the native sequence of a protein. The percent recovery of amino acids with positive values in the PSSM determines the recovery of evolutionarily tolerated amino acids.

EXPERIMENTAL PROCEDURES

The RosettaDesign energy function is a linear combination of individual energy terms. As a result, the addition of a new energy term will impact the energy function as a whole. To

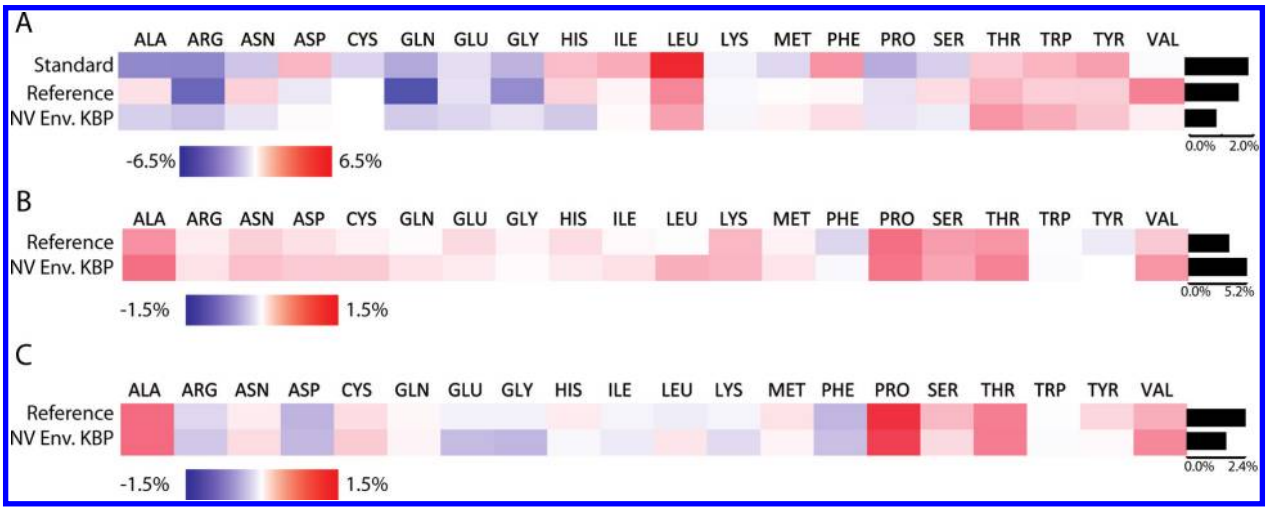


Figure 2. (A) Percent change in overall sequence composition between native and designed proteins for all 100 structures in the five-way cross-validation set. The black bars show the rms percent composition change. (B) Percent PSSM recovery for all 100 structures in the five-way cross-validation set. The black bars show rms percent PSSM recovery. (C) Percent sequence recovery for all 100 structures in the five-way cross-validation set. The black bars show rms percent sequence recovery.

address this, each energy term is multiplied by a weight, and these weights must be carefully optimized following the introduction of a new term. In most cases, it is not necessary to optimize the entire scoring function when a new term is added. Instead, only the terms that describe similar information as the new term are optimized. In the case of the NV environment KBP, the solvation free energy potential and the reference energies must also be optimized.

To ensure that the optimized weights would apply to a wide range of proteins, a set of 100 soluble protein crystal structures from the Protein Data Bank (PDB) were used in optimization. Structures were selected to have a sequence homology of less than 25%, a length of 67−179 amino acids, and a resolution better than 2.0 Å. The optimization was conducted using a five-way cross-validation protocol. In this protocol, the 100 crystal structures described above were split into five groups of 20 structures each. In each component of the five-way validation, 80 proteins were used during optimization, and the remaining 20 were used to benchmark the resulting weights. In statistics generated from the benchmarking phase of the optimization process, results from all five sets of 20 proteins are combined, resulting in a total benchmark set of 100 proteins.

An iterative particle swarm approach[10] was used to optimize the weights. The RosettaDesign standard energy function was used as an initial point for optimization, and the weight of the NV environment KBP was arbitrarily given an initial value of 1.0. Twenty rounds of particle swarm optimization were performed for each component of the five-way cross validation described above. The weights were optimized to maximize the PSSM score of proteins designed using the energy function (Table 1 of the Supporting Information). The PSSM for each protein was generated from a PSI-PRED BLAST query of the protein structure sequence using an e threshold of 0.001 and three iterations. The nonredundant (NR) sequence database was used. The average level of sequence identity between the query sequence and all other sequences in the generated PSSMs was 30% for both benchmark sets

Because the standard deviation of the averaged reference energies was relatively high, the reference energies of the averaged energy function are optimized to reduce the overall sequence composition biases introduced during design (Table 2 of the Supporting Information).

Two separate optimization experiments were performed. In the first experiment, the reference energies, solvation free energy potential, and the NV environment KBP were optimized. In the second experiment, the NV environment KBP was excluded from the energy function, and only the reference energies were optimized. This second experiment acts as a control and makes it possible to distinguish between design improvements caused by reference energy optimization and design improvements caused by the addition of the NV environment KBP itself. While both the NV environment KBP and the solvation free energy potential describe overlapping but different phenomena at different levels of resolution, the NV environment KBP is an indirect measure of solvation free energy and evolutionary biases against aggregation. It is a measure at amino acid resolution and will be independent of side chain conformation. In contrast, the solvation free energy potential is at atomic resolution incorporating a specific model of solvation. While the solvation free energy potential does an inadequate job of accounting for biases against aggregation on the protein surface, it is highly accurate in avoiding burial of polar atoms and is important in identifying side chain conformations.

The optimization experiments described above produce five individual energy functions, each generated from one section of the five-way cross validation. To produce a single optimized scoring function for general use, the weights from the five optimized energy functions are averaged together, and the reference energies of the averaged energy function are optimized using the set of 100 proteins used in the initial cross validation. The averaged energy function is benchmarked on an independent set of 42 protein crystal structures, in which the proteins have a sequence homology of less than 15%, a size range of 150−225, and a resolution of less than 1.5 Å. Note that these proteins are larger and more complex than the proteins used in the more time-consuming weight optimization procedure. As a result, this benchmark poses a formidable challenge for the RosettaDesign fixed backbone design algorithm. Several different metrics were used during benchmarking to assess the quality of designed proteins.

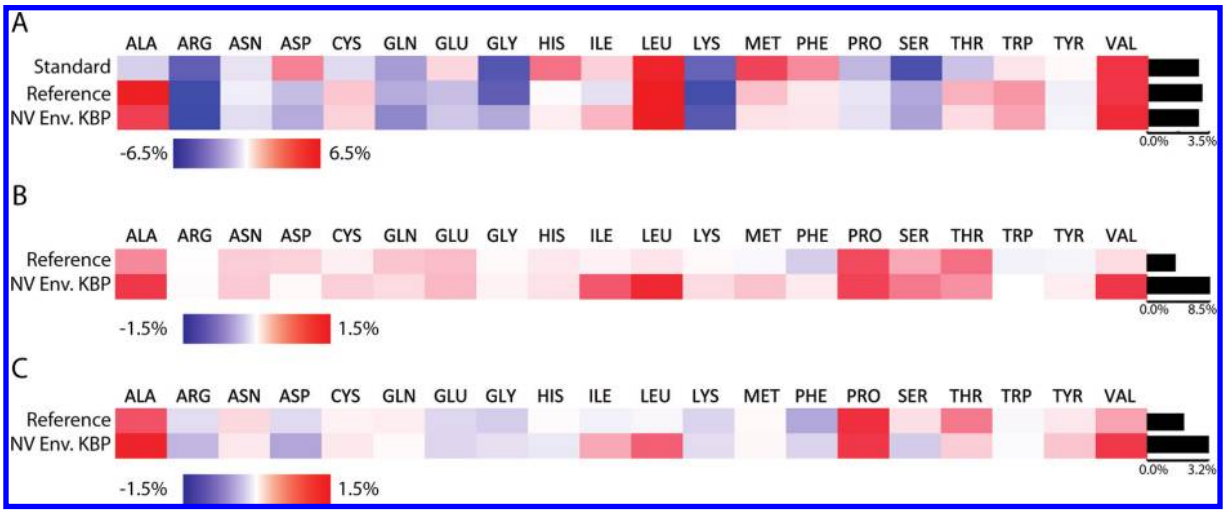


Figure 3. (A) Percent change in overall sequence composition between native and designed proteins for all 42 structures in the independent benchmark set. The black bars show the rms percent composition change. (B) Percent PSSM recovery for all 42 structures in the independent benchmark set. The black bars show rms percent PSSM recovery. (C) Percent sequence recovery for all 42 structures in the independent benchmark set. The black bars show rms percent sequence recovery.

Table 1. Percent PSSM Recovery and Percent Sequence Recovery by Degree of Burial for 100 Proteins Used in Optimization[a]

	percent PSSM recovery (%)			percent sequence recovery (%)		
	standard	reference	NV environment KBP	standard	reference	NV environment KBP
buried	73.4	77.1	78.9	64.9	66.5	65.5
boundary	72.1	75.3	77.3	44.3	46.6	45.5
surface	70.4	74.4	75.9	32.8	35.9	35.5
overall	72.0	75.6	77.2	45.7	48.1	47.3

[a]Standard refers to the standard energy function. Reference refers to the modified standard energy function in which the reference energies were reoptimized. NV environment KBP refers to the optimized energy function incorporating the NV environment energy term.

Percent PSSM recovery was the primary benchmarking metric used in the study. Percent PSSM recovery was calculated as the percentage of residues that were designed as residues with a positive score in the PSSM of the native protein. In addition, the percent sequence recovery was measured as the percentage of residues that remained as the native residue after design.

The percent PSSM recovery per residue, percent sequence recovery per residue, and the change in overall sequence composition were also calculated for each designed protein. Percent PSSM recovery per residue is calculated as (num_pssm_recovered)/(num_designed), where num_pssm_recovered is the number of residues with a given identity that were designed to a residue with a positive PSSM score and num_designed is the total number of residues designed. The percent sequence recovery per residue was calculated as (num_recovered)/(num_designed), where num_recovered is the number of residues with a given identity that were designed to an identical residue. In addition to calculation of the overall percent sequence recovery, sequence recovery by chemical group was also calculated. In this metric, residues were grouped into the categories polar (Ser, Thr, Asn, and Gln), nonpolar (Ala, Val, Leu, Met, and Ile), aromatic (Phe, Tyr, and Trp), charged (Lys, Arg, His, Asp, and Glu), and other (Cys, Pro, and Gly). A residue was counted as recovered if it was mutated to another residue within the same group.

The percent sequence composition change per amino acid type was calculated as $(d - n)/(\text{num\_designed})$, where d is the number of designed residues of a given type and n is the number of native residues. To compute the change in overall sequence composition, a root-mean-square (rms) deviation method was used. The rms percent sequence composition change was calculated as follows, where statistical_metric is one of the metrics described above (shown as black bars in Figures 2 and 3):

$$\text{rms} = \sqrt{\frac{1}{20}\sum_{i=1-20}^{20}(\text{statistical\_metric})^2}$$

All of the metrics described above were calculated for the entire protein, as well as the deeply buried region, surface region, and a boundary layer between the two. For this study, the buried region is defined as all residues with an NV score between 0.00 and 0.24, the boundary is defined as residues with an NV score between 0.25 and 0.39, and the surface region is defined as residues with an NV score between 0.40 and 1.00. The performance of the optimized energy functions via these benchmarks was compared to the performance of the standard RosettaDesign energy function.

The benchmarks described above are intended as a measure of how well RosettaDesign is accomplishing its goal of generating low-energy, native-like protein sequences. In a well-optimized energy function, we expect that the percent PSSM recovery will increase compared to the standard RosettaDesign energy function. We also expect that the percent sequence recovery will remain similar to that obtained with the standard energy function. Finally, we expect that a well-optimized energy function will exhibit smaller biases in sequence composition

Table 2. Percent PSSM Recovery and Percent Sequence Recovery by Degree of Burial for 42 Proteins Used in Benchmarking[a]

	percent PSSM recovery (%)			percent sequence recovery (%)		
	standard	reference	NV environment KBP	standard	reference	NV environment KBP
buried	68.0	70.7	76.0	49.5	49.8	51.5
boundary	66.1	70.6	75.7	32.1	34.1	35.3
surface	67.2	72.2	75.8	22.7	26.3	27.3
overall	67.4	71.2	75.9	35.7	37.6	38.9

[a]Standard refers to the standard energy function. Reference refers to the modified standard energy function in which the reference energies were reoptimized. NV environment KBP refers to the optimized energy function incorporating the NV environment energy term.

compared to proteins designed with the standard energy function.

RESULTS

The percent PSSM recovery and percent sequence recovery calculated for the 100 proteins used in the five-way cross validation are listed in Table 1. The results of PSSM recovery and sequence recovery analysis show that the optimized NV environment KBP energy function exhibits a 5.2% improvement in percent PSSM recovery compared to the standard energy function and that 3.6% of this improvement was a result of reference energy optimization. The NV environment KBP energy function showed a 1.6% improvement in percent sequence recovery compared to the standard RosettaDesign energy function and a 2.4% improvement if only reference energies are optimized.

The percent change in composition between native and designed sequences for the 100 proteins used in the 5-fold cross validation is shown in Figure 2A. Proteins designed with the NV environment KBP energy function show a decrease in the average magnitude of sequence composition biases introduced during design compared to proteins designed with the standard energy function. Proteins designed with the standard energy function exhibit an rms percent change in sequence composition of 2.0%, while proteins designed with the NV environment KBP show an rms percent change in sequence composition of 1.0%. Figure 2B shows that rms per residue PSSM recovery increased from 3.8% with the standard energy function to 4.2% with the NV environment KBP, and Figure 3C shows that rms per residue sequence recovery remained relatively constant between the standard energy function and NV environment KBP.

The energy functions produced with the five-way validation were averaged to produce a single energy function; the reference energies of this averaged function were optimized, and the benchmarking analysis used above was repeated using the averaged energy function. In this case, the independent benchmark set of 42 proteins was used. Table 2 shows the percent PSSM recovery and percent sequence recovery calculated for the 42 proteins designed using the averaged energy function. The NV environment KBP showed an 8.8% improvement in PSSM recovery compared to the standard energy function and that 3.8% of this improvement was a result of the reference energy optimization. The NV environment KBP showed a 3.2% overall improvement in sequence recovery, of which 1.9% was due to the reference energy optimization.

When sequence recovery is broken down by group (Figure 4), a large improvement (decrease) in the percentage of unrecovered buried charged residues is observed; from 8.78 to 1.96% unrecovered residues function in the 100-protein benchmark set. Additionally, a decrease from 8.77 to 3.46% unrecovered nonpolar residues on the surface is observed.

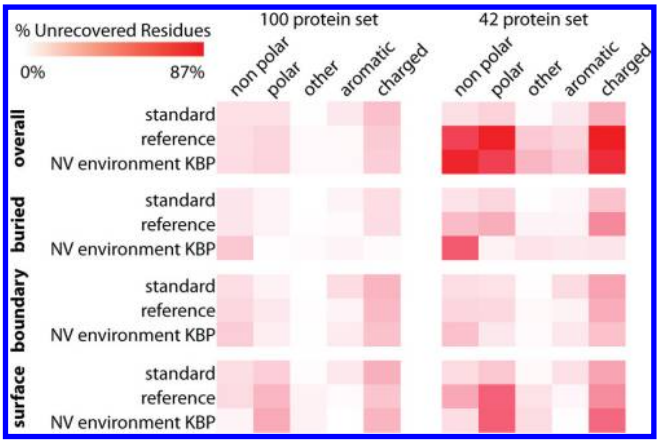


Figure 4. Percentage of unrecovered residues (number of recovered residues divided by total number of residues in the benchmark set) by amino acid category in the 100- and 42-protein benchmark sets. The color scale ranges from white (small number of mistakes) to red (large number of mistakes). In this metric, residues were grouped into the categories polar (Ser, Thr, Asn, and Gln), nonpolar (Ala, Val, Leu, Met, and Ile), aromatic (Phe, Tyr, and Trp), charged (Lys, Arg, His, Asp, and Glu), and other (Cys, Pro, and Gly). A residue was counted as recovered if it was mutated to another residue within the same group.

Additionally, Figure 4 reveals a fundamental difference in the two data sets. The lower recovery values seen in all categories in the 42-protein benchmark set suggest that it is a much more challenging target for design than the 100-protein benchmark set used in optimization. The proteins of the 42-protein benchmark set are substantially larger (average length of 207 residues) than those in the 100-protein benchmark set (average length of 120 residues). Each additional residue drastically increases the number of possible sequences to consider, decreasing the probability of a high-quality design. Despite this more challenging independent benchmark, improvement was still observed.

DISCUSSION

The results of both the 100-protein five-way cross validation and the 42-protein independent benchmark set are consistent. In both cases, introduction of the NV environment KBP into the energy function and optimization of the energy function weights lead to an overall improvement in the quality of designed sequences. As the independent benchmark set tests an averaged scoring function that would be generally useful, the remaining analysis will focus on this benchmark set.

The results of the benchmarking show that, in general, structures designed using the NV environment KBP exhibit smaller conformation biases and more evolutionarily favorable mutations. A detailed analysis of these results also provides

dx.doi.org/10.1021/bi200664b | Biochemistry 2011, 50, 8521–8528

some insight into the behavior of the RosettaDesign scoring function.

Because of the lack of an explicit water model in RosettaDesign, the standard RosettaDesign energy function is dominated by the solvation free energy potential. As a result, there are few constraints on amino acid mutations on the protein surface. Because of this lack of constraints, proteins designed with the standard energy function exhibit large biases in sequence composition on the protein surface. Proteins designed with the standard energy function show large numbers of aromatic residues on the protein surface. Specifically, there is a 3.1% increase in the number of phenylalanines, a 1.9% increase in the number of tryptophans, and a 2.5% increase in the number of tyrosines on the protein surface in the benchmark set designed with the standard RosettaDesign energy function compared to the native structure. Proteins designed with the NV environment KBP show a large reduction in these biases. Proteins designed with the NV environment KBP showed a 1.4% increase in the number of phenylalanines, a 0.8% increase in the number of tryptophans, and a 1.5% increase in the number of tyrosines compared to native proteins. While still large, these biases are much smaller than the biases observed with the standard energy function.

It was expected that improvements in the quality of surface sequence design would be the primary benefit of the NV environment KBP. However, an analysis of the overall PSSM recovery, sequence recovery, and sequence composition biases suggests that the improvement given by the NV environment KBP implementation occurred across the board rather than merely at the protein surface. Table 2 shows the overall impact of the NV environment KBP at various levels of burial. The percent PSSM recovery improved using the NV environment KBP by 8.0% in the buried region, 9.6% in the boundary region, and 8.5% on the surface region compared to the standard energy function. The percent sequence recovery improved by 2.0% in the buried region, 3.2% in the boundary region, and 4.6% in the surface region compared to the standard energy function.

When percent sequence recovery is broken down by group (Table 2), a large increase in the recovery of buried charged residues is observed, with a 6.2% increase compared to the standard energy function. Additionally, a 3.8% increase in recovery of nonpolar residues is observed on the surface. Not all groups show improvement, and this is expected, as the scoring function was not directly optimized for percent sequence recovery.

While the overall percent changes are relatively small, these changes are both statistically and scientifically significant. To assess the statistical significance of the data, standard deviations were calculated and are listed in Tables 3 and 4 of the Supporting Information. The standard deviations were calculated for both percent PSSM recovery and percent sequence recovery. Each of the five scoring functions generated during the five-way cross-validation weight optimization using 100 proteins was used to design the independent set of 42 proteins. The standard deviations of PSSM and sequence recovery are listed in Tables 3 and 4 of the Supporting Information. The standard deviations listed in Tables 3 and 4 of the Supporting Information range from 0.1 to 1.2%. The average error is 0.4% and therefore smaller than the observed improvements in recovery rates.

It is important to consider not only the absolute change in percent recovery but also the change relative to the maximum possible recovery value. In the case of sequence recovery, the maximum possible sequence recovery can be estimated by analyzing the amino acids tolerated in each position in BLAST-derived PSSMs. In this case, the average percentage of time that the native residue is seen in the PSSM is used as an estimate for expected sequence recovery. For the 100-protein benchmark set, the average was 34% with a standard deviation of 12%, while in the 42-protein benchmark set, the average was 34% with a standard deviation of 7%. While the achievable sequence recovery is somewhat higher because of correlation between individual positions, these values suggest that sequence recovery rates of 40−50% would approach the maximum. Tables 1 and 2 show that for the 100-protein benchmark set, the total overall sequence recovery is 45.7% with the standard energy function and 47.0% with the NV environment KBP. For the 42-protein benchmark set, the total overall sequence recovery is 35.7% with the standard energy function and 38.9% with the NV environment KBP. This explains the relatively small increases in sequence recovery, as current recovery values are approaching the practical maximum. For that reason, we introduce the PSSM recovery value. In this context, it is important to note that the scoring functions were not directly optimized for sequence recovery but rather PSSM recovery. As a result, it is not surprising that the sequence recovery is not necessarily maximized during optimization.

In the case of PSSM recovery, it is reasonable to expect that 100% PSSM recovery is unreachable as evolution might not have sampled all amino acids tolerated in a sequence position. A more realistic value for maximum possible PSSM recovery is between 80 and 90%, though the exact value of this upper bound is difficult to estimate. PSSM recovery with the standard energy function was 72.0%. The observed increase to 77.2% with the NV environment KBP represents a substantial increase relative to the 80−90% maximum and the 72% starting point. Generally, improvements in sequence recovery rates have been moderate when altering the energy function,[5] as the major contributors to the overall energy are already fine-tuned and remain unaltered.

Comparison of PSSM and sequence recovery results between the 42-protein benchmark set and the 100-protein set illustrates that the performance of the RosettaDesign algorithm varies on the basis of the characteristics of the protein being designed. For example, Tables 1 and 2 show the overall sequence recovery for proteins designed with the standard energy function. The overall recovery for the 42-protein benchmark set was 35.7%, while the overall recovery for the 100-protein benchmark set was 45.7%. This substantial difference is likely a result of the different criteria used to select the proteins in each set. The proteins in the 42-protein benchmark set are larger than those in the 100-protein set and will therefore have a larger total surface area and thus be more challenging targets for design.

Despite the difference inherent to different design targets, these values are similar to those obtained in the literature. Schneider et al. designed a set of proteins with 89−223 amino acids on the basis of high-resolution crystal structures. They observed surface sequence recovery rates of 22 ± 11% and buried recovery rates of 56 ± 13.7% when designed with RosettaDesign.[11] These values are similar to those seen in Tables 1 and 2. Additionally, Sharabi et al. reported overall sequence recovery values between 40 and 70% depending on the weights of the scoring function used during their design.[12]

These numbers are within the range of sequence recovery values obtained during the experiments described here.

In addition to improvements in PSSM and sequence recovery, the degree of sequence bias seen in the buried and boundary regions of designs made using the NV environment KBP decreased. When all residues in the benchmark set are considered, proteins designed with the NV environment KBP have an rms percent composition change of 2.8% compared to the native protein, while proteins designed with the standard energy function have an rms percent composition change of 2.9% (Figure 3A). When this overall value is broken down by region, the buried region designed with the NV environment KBP shows an increase in rms percent composition change compared to the standard energy function from 4.2 to 4.5%, the boundary region shows a decrease from 3.4 to 2.7%, and the surface region shows a reduction from 2.9 to 2.4%. While the improvements in sequence composition bias are minimal, Figure 3B shows increases in rms per residue PSSM recovery from 3.8% with the standard energy function to 4.2% with the NV environment KBP. Additionally, Figure 3C shows increases in rms per residue sequence recovery from 2.4% with the standard energy function to 2.6% with the NV environment KBP energy function, which is expected given the optimization of the scoring function toward PSSM improvement.

An investigation of the optimized weights provides some insight into the cause of the improvements in sequence design. Tables 1 and 2 of the Supporting Information show the scoring function and reference energy weights of the standard energy function and the optimized NV environment KBP. When the NV environment KBP term is added to the energy function, the weight of the free energy solvation potential decreases from 0.65 in the standard energy function to 0.56 in the NV environment KBP. The NV environment KBP term has a value of 1.01. As discussed earlier, in the standard energy function, the reference energies and solvation free energy potential are the dominant forces on surface residues because of the lack of explicit inter-residue interactions. Because the penalty given by the solvation free energy potential for apolar residues on the surface is relatively weak, the weight of this potential will need to be increased for it to adequately affect surface residues. However, because the energy function is applied evenly, regardless of the degree of burial, the increase in weight necessary to maintain a reasonable protein surface may cause the solvation free energy potential to apply too strongly to the boundary region. As the burial level increases, the number of inter-residue interactions will also increase, which explains the decrease in improvement in sequence bias seen in more highly buried regions of the protein. This idea is supported by the decrease in free energy solvation potential weight observed in the NV environment KBP energy function. The NV environment KBP provides additional information about protein surface composition, reducing the dominance of the free energy solvation potential. Figure 5 shows the effect of the NV environment KBP on the overall scoring function. All proteins used in the 100-protein benchmark set were scored using both the standard RosettaDesign energy function and the optimized NV environment KBP energy function. The average magnitude of each scoring term for each buried and surface residue was calculated and converted to the percentage of the total energy for each residue to measure the influence of each scoring term. We observe that the addition of the NV environment KBP term decreases the influence of the reference energies, solvation free energy term, and the attractive and repulsive terms throughout

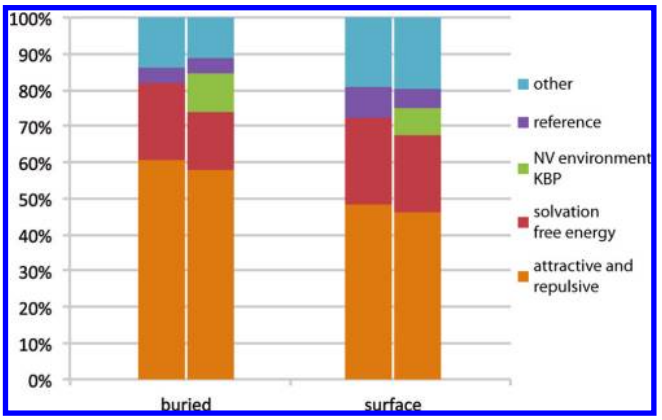


Figure 5. Contribution of individual scoring terms to the overall score of buried and surface residues. The introduction of the NV environment KBP reduces the reliance on solvation free energy and the attractive−repulsive forces at both levels of exposure.

all degrees of burial. Specifically, the influence of the solvation free energy decreases from 21 to 16% for buried residues and from 24 to 21% for surface residues. Additionally, the influence of the reference energy decreases from 8 to 5% on the surface, though it remains relatively unchanged for buried residues. The attractive and repulsive forces also change somewhat, with a decrease in influence from 60 to 57% in buried residues and 48 to 46% in surface residues. This change in influence is significantly smaller than the change in influence seen in the reference and solvation free energy functions. The NV environment KBP was designed to address shortcomings in the design of the protein surface. These shortcomings are the result of the energy function failing to model aspects of the protein surface that are not completely described through the solvation and reference energies. To achieve reasonably good performance despite these inaccuracies, both energy terms are overweighted in the standard energy function. As expected, addition of the NV environment KBP term reduces the impact of solvation and reference energies on the surface. As these adjustments apply throughout all degrees of burial, the artificially inflated weight of the solvation and reference energies can be decreased, improving performance also in the buried regions of the protein.

In addition to providing information about solvation effects, the NV environment potential also sheds light on the evolutionary and environmental forces on protein composition. Soluble proteins have evolved to be nonaggregative and generally stable in the environment of a cell. These properties are difficult to model via physics-based methods, as they arise from numerous interprotein interactions that are difficult to explicitly model. The implicit modeling of these environmental effects accounts in part for the improvements in native-like sequence design seen during design with the NV environment KBP. By optimizing the NV environment KBP energy function to maximize PSSM score rather than sequence recovery, the energy function is optimized to design proteins similar to those that are favored evolutionarily, rather than to merely reproduce the native sequence.

We thank David Baker and Kristian Kaufmann for valuable discussion and development assistance.

■ ASSOCIATED CONTENT

Supporting Information

Supporting figures and tables and a compressed archive containing the scripts, input files, and command lines used to perform the analysis described in this paper. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Vanderbilt University Medical Center, Chemistry Department, 5144B MRB III/BioSci, 465 21st Ave. S., Nashville, TN 37232. Phone: (615) 936-5662. Fax: (615) 936-2211. E-mail: jens.meiler@vanderbilt.edu.

Present Address

†Harvard University, Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, MA 02139.

Funding

This work was supported by the Protein Design Project of the Defense Advanced Research Projects Agency (DARPA Grant 04 12).

■ ABBREVIATIONS

SASA, solvent accessible surface area; NV, neighbor vector; KBP, knowledge-based potential; BLAST, Basic Local Alignment Search Tool; SCOP, Structural Classification of Proteins; HSE, half-sphere exposure; PSSM, position specific scoring matrix; NCR, neighbor count (ROSETTA); PDB, Protein Data Bank; NR, nonredundant; PSI-PRED, position specific iterated prediction.

■ REFERENCES

(1) Lazaridis, T., and Karplus, M. (1999) Effective energy function for proteins in solution. *Proteins 35*, 133−152.

(2) Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R., and Meiler, J. (2009) Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J. Mol. Model. 15*, 1093−1108.

(3) Chandler, D. (2005) Interfaces and the driving force of hydrophobic assembly. *Nature 437*, 640−647.

(4) Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003) A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol. 332*, 449−460.

(5) Kortemme, T., Morozov, A. V., and Baker, D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol. 326*, 1239−1259.

(6) Hamelryck, T. (2005) An amino acid has two sides: A new 2D measure provides a different view of solvent exposure. *Proteins: Struct., Funct., Bioinf. 59*, 38−48.

(7) Kuhlman, B., and Baker, D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U.S.A. 97*, 10383−10388.

(8) Chothia, C., and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J. 5*, 823−826.

(9) Schaeffer, R. D., Jonsson, A. L., Simms, A. M., and Daggett, V. (2011) Generation of a consensus protein domain dictionary. *Bioinformatics 27*, 46−54.

(10) Chen, H., Liu, B., Huang, H., Hwang, S., and Ho, S. (2007) SODOCK: Swarm optimization for highly flexible protein-ligand docking. *J. Comput. Chem. 28*, 612−623.

(11) Schneider, M., Fu, X., and Keating, A. E. (2009) X-ray vs. NMR structures as templates for computational protein design. *Proteins: Struct., Funct., Bioinf. 77*, 97−110.

(12) Sharabi, O., Dekel, A., and Shifman, J. M. (2011) Triathlon for energy functions: Who is the winner for design of protein-protein interactions? *Proteins: Struct., Funct., Bioinf. 79*, 1487−1498.