Full protein sequence redesign with an MMGBSA energy function

Thomas Gaillard* Thomas Simonson*

January 30, 2017

Laboratoire de Biochimie (CNRS UMR7654), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France.

*Corresponding authors: thomas.gaillard@polytechnique.edu, thomas.simonson@polytechnique.edu

Keywords: computational protein design, inverse folding problem, molecular mechanics, Generalized Born, surface area.

Short title: Full sequence redesign with MMGBSA

Abstract

Computational protein design aims to create proteins with novel properties. A key element is the energy or scoring function used to select the sequences and conformations. We study the performance of an "MMGBSA" energy function, which combines molecular mechanics terms, a Generalized Born and Surface Area (GBSA) solvent model, with approximations that make the model pairwise additive. Our approach is implemented in the Proteus software. The use of a physics-based energy function ensures a certain model transferability and explanatory power. As a first test, we redesign the sequence of nine proteins, one position at a time, with the rest of the protein having its native sequence and crystallographic conformation. As a second test, all positions are designed together. The contributions of individual energy terms are evaluated and various parameterizations are compared. We find that the GB term significantly improves the results, compared to simple Coulomb electrostatics, but is affected by pairwise decomposition errors when all positions are designed together. The SA term, with distinct energy coefficients for non-polar and polar atoms, makes a decisive contribution to obtain realistic protein sequences, and can partially compensate for the absence of a GB term. With the best GBSA protocol, we obtain native-like protein cores and Superfamily recognition of almost all our sequences.

1 Introduction

Protein design aims to create new proteins or modify existing ones to obtain new properties. Computational approaches are a valuable help for protein design, to rationalize the predictions and guide experimental tests.[1–9] Computational protein design (CPD) has been successfully applied to various engineering goals, such as sequence compatibility with a given fold [5, 10–19], ligand binding [20–23], or a new enzyme activity [24–26].

CPD is a difficult problem. It involves the exploration of an enormous sequence and conformational space. Common simplifications are to hold the backbone fixed and replace the continuous sidechain conformational space by a discrete set of frequently observed rotamers [27]. Assuming ten rotamers per amino acid, the size of the space is still $(20 \times 10)^{100}$ for a protein with 100 amino acids: googols of googols. Efficient exploration algorithms are thus required.[28–30] Another critical aspect is the accuracy of the energy function.[29, 31–34]

CPD has stimulated important methodological efforts and many programs are available. Among the most popular ones are Rosetta [5, 14, 35, 36], OSPREY [37], EGAD [16, 38, 39], and ORBIT [3, 40]. Rosetta is the most cited. Rosetta is a suite of tools originally developed for *de novo* folding prediction. It was then expanded to various problems, including protein design, with the RosettaDesign module. The scoring function used in RosettaDesign contains a knowledge-based sidechain rotamer energy term based on the backbone-dependent Shapovalov and Dunbrack library [41], a Lennard-Jones potential with a finite repulsion term, Coulomb electrostatics, an explicit hydrogen bonding term, a Lazaridis-Karplus implicit solvation energy, and unfolded state reference energies. The exploration is performed with a Monte Carlo simulated annealing scheme.

Here, we perform sequence design with the Proteus CPD program [18, 42, 43]. Proteus can address a wide variety of problems, including sidechain prediction [44, 45], mutation stability prediction [44, 46], redesign of full protein sequences [18], fold recognition [19, 47], active site engineering [22, 23], and pK_a predictions [48]. We use an "MMGBSA" energy function [49], composed of molecular mechanics terms and an implicit solvation model, with Generalized Born [50] and solvent accessible Surface Area terms [51]. We follow a precalculation strategy, where the interaction energies of all possible rotamer pairs of the protein are evaluated and stored in an energy matrix. From the matrix, the energy of any sequence and conformation (among the $(20 \times 10)^{100}$ possibilities) can be quickly calculated, greatly speeding up the exploration phase. The precalculation strategy requires that the energy function is pairwise additive. This is not normally true for GB and SA terms and so additional approximations have to be made, studied elsewhere [44, 52, 53]. The sequence-conformation exploration is done with a heuristic method

originally developed by Wernisch et al. [11] and available in Proteus.

This work has two main goals. First, we want to evaluate the performance of Proteus. An earlier, large-scale evaluation for full sequence design dates back to 2010 [19, 47] and used a rather simple energy function. A recent evaluation used an improved energy but considered a single protein family [54]. The improved energy function uses an all-atom force field (AMBER ff99SB [55, 56]), a GB term [23, 44, 52] and an improved decomposition of the SA term [42, 53]. Second, we want to investigate the performance of MMGBSA energy functions for sequence design. Most CPD programs use highly optimized, empirical energy functions.[5, 15, 57–61] Energy functions based on physical principles have the advantage of better transferability and explanatory power. Evaluating the strengths and weaknesses of physics-based energy functions applied to CPD can lead to a better fundamental understanding of the protein design problem. Molecular mechanics energy functions with GB or PB (Poisson-Boltzmann) [62] and SA solvent models are widely-used in biomolecular simulations [49, 63–68], but to our knowledge, they have not been systematically evaluated for protein design. We recently performed an assessment of MMGBSA energy functions for sidechain prediction. [45] The current work follows a parallel approach applied to sequence design.

Full sequence design for a given fold, also called the inverse folding problem, is an important CPD problem. It has been studied many times,[3, 5, 10–14, 16, 18, 69, 70] in particular as a test of CPD methods. We perform full sequence design on 9 proteins, belonging to 3 different fold families (SH3, PDZ, and SH2), with sizes ranging from 56 to 109 amino acids. Identity percentages and similarity scores of designed sequences with respect to the native sequences and Pfam family profiles are computed. Single-position designs are first presented, where a residue is designed with the rest of the protein in its crystallographic conformation and native sequence. All-position designs are then presented, where all residues are designed at the same time. The contributions of individual energy terms and protocol details are discussed. We find that, with the best GBSA protocol, we obtain native-like protein cores and Superfamily recognition of almost all our sequences.

2 Methods

2.1 Choice and preparation of structures

A set of protein structures was downloaded from the Protein Data Bank (PDB) [71]. The set comprises nine proteins belonging to three well-studied families, SH3, PDZ, and SH2. The set is described in Table 1. Cartoon representations of the structures are given in Figure 1.

2.2 Rotamer library

The rotamer library was prepared from the 1995 library of Tuffery *et al.* [72]. We built a "hydrogen-augmented" variant of this library, by adding rotamers controlling the conformation of selected polar hydrogens (Cys: Cα-Cβ-Sγ-Hγ = -60, 60, 180°; Ser: Cα-Cβ-Oγ-Hγ = -60, 60, 180°; Thr: Cα-Cβ-Oγ1-Hγ1 = -60, 60, 180°; Tyr: Cε1-Cζ-Oη-Hη = 0, 180°). The number of rotamers per amino acid type in the original and augmented libraries is shown in Supplementary Table S1.

2.3 Energy matrix calculation

An energy matrix was calculated for each protein. The matrices are symmetrical and their dimension is the total number of rotamers of all amino acids over all positions. Diagonal terms correspond to the interaction energy of a sidechain with itself and with the fixed part of the protein. Off-diagonal terms correspond to the interaction energy between sidechain pairs. The energy matrix calculation was performed with Proteus [18, 42], which uses scripts implemented for the XPLOR program [73].

We performed two different sequence design studies. In single-position designs, the amino acid type and conformation of one sidechain is predicted while the rest of the protein is kept in its native sequence and crystallographic conformation. In this case, the fixed part is the backbone and other sidechains, the matrix is diagonal, and there are no errors due to the pairwise additivity approximations. In all-position designs, the amino acid type and conformation of all sidechains are predicted at the same time. The fixed part is the backbone and fixed sidechains (Gly, Pro, and disulfide Cys). While the MM terms are pairwise additive, this is not the case of GB and SA terms. The pairwise decomposition of these terms necessitates approximations which are detailed elsewhere [53]. A Native Environment Approximation (NEA) is used for the GB term [23, 48, 52] and a truncated series expansion is employed for the SA term [53].

2.4 Energy function

The energy function was of the MMGBSA form. The molecular mechanics force field was AMBER ff99SB [55, 56]. Charges were slightly modified to make them uniform on the backbone of all amino acids except Gly and Pro. The Generalized Born variant was GB/HCT [74–76]. Atomic volumes and scaling factors optimized for protein design were used [44]. A solvent dielectric constant of 80 was chosen. No cutoff was employed for the van der Waals, Coulomb, and GB terms. Solvent accessible surface areas were calculated with the Lee and Richards algorithm [77], using a 1.5 Å probe radius, van der Waals half σ values as atomic radii, and

an accuracy parameter of 0.005, which was found to be a good compromise between speed and accuracy. Hydrogen atoms were not considered in surface area calculations. Surface energies were obtained by multiplying the atomic components of surface areas by atom-type specific solvation energy coefficients. Four classes of atoms were distinguished: non-polar (N), polar (P), aromatic (A), and ionic (I). Non-polar atoms are C and S; polar atoms are O, N, and carbonyl, carboxylate, and imine C; aromatic atoms are aromatic C; ionic atoms are those most affected by formal charges. Details of atom classes are given elsewhere [45].

2.5 Sidechain minimizations

Before the calculation of diagonal energy terms, sidechains were energy minimized for 15 steps using Powell's conjugate gradient algorithm, in the environment of the fixed part (the rest of the protein in single-position designs; the backbone plus Gly, Pro, and disulfide Cys in all-position designs). Harmonic restraints of 200 kcal/mol/rad^2 with a tolerance range of $\pm 5^\circ$ around the targeted angle were applied to sidechain dihedrals to enforce selected rotamer conformations. Before calculation of off-diagonal energy terms, a minimization of sidechain pairs was conducted with the same protocol. MM and GB terms were activated in sidechain minimizations and an interior dielectric constant of 4 was used.

2.6 Designed sequences

In single-position designs, the optimal amino acid type was obtained by selecting the rotamer with the lowest folding energy for each position. In all-position designs, the designed sequences were generated from the energy matrix using a heuristic optimization [11] implemented in the Proteus program [42]. The heuristic procedure starts by assigning a random rotamer to each position. The positions are then considered one at a time and the rotamer with the best folding energy is found, given the current rotamers at all other positions. The sequence is cycled 500 times. The whole procedure is repeated 200,000 times with different random initial rotamers. The 100 sequences with the lowest folding energy are then selected for the analysis.

2.7 Unfolded state

The folding energy is obtained as the energy of the folded state calculated from the energy matrix minus the energy of the unfolded state. A simple description of the unfolded state is adopted, where the positions are independent. The unfolded state energy is calculated as a sum of amino-acid specific values termed reference energies.

2.8 Reference energies

The reference energies were optimized for each energy protocol by performing cycles of sequence predictions and reference energies corrections, targeting a natural composition of amino acids [18]. The target criterion corresponds to a maximum likelihood criterion [54]. Natural occurrence frequencies of amino acids were calculated from Pfam sequence libraries of PDZ, SH2, and SH3 domains. The optimization was stopped when converged amino acid compositions were obtained. An initial guess was obtained by averaging the energy matrix diagonal terms by amino acid type. In the single-position study, the reference energies were not optimized beyond the initial guess.

2.9 Evaluation of sequences

Designed sequences were evaluated by calculating different descriptors and comparing them to reference sequences.

2.9.1 Reference sequences

The reference sequences considered were the native sequences of the designed proteins and the sequence profiles computed from the Pfam libraries of the protein families studied. When the designed sequences are compared to the native sequence, no alignment is necessary and the whole sequence is considered. When the Pfam profile was the reference, a preliminary step consists in aligning the native sequence of the designed protein with the Pfam profile. The sequence-profile alignment was performed with Clustal Omega [78] followed by manual inspection.

2.9.2 Identity and similarity scores

The identity percentage and the similarity score were calculated with respect to the reference sequences. The similarity matrix was BLOSUM62 [79].

2.9.3 Superfamily classification

SCOP superfamily classification of designed sequences was performed using the SUPERFAMILY database and tools [80].

2.10 Statistics

Statistics were calculated from the series of score values. The dataset composition is available in Supplementary Table S2. Averages were calculated over the whole dataset, and also on core and surface residues. Core residues were defined as those with a solvent accessible surface area

exposure ratio lower than 0.15. Surface residues were those with an exposure ratio higher than 0.30.

To test whether the score differences between two protocols is significant, paired t-tests were performed. The null hypothesis is that the score averages of the two protocols are the same. To test whether there is a protocol significantly different from the others within a group, repeated measures analysis of variance (ANOVA) was performed. The null hypothesis is that the score averages of all the protocols are the same.

2.11 Rosetta design

The Rosetta design program [5, 35] was also used to redesign the sequences of our protein set. Default parameters were employed. For each protein, we predicted 1,000 sequences with Rosetta design and the 100 having the lowest folding energy were analyzed with the same protocol as our designs.

3 Results

Single-position, and all-position sequence designs were performed on 9 proteins (Table 1 and Figure 1). Different protocols were compared. In the following, noC, C, and GB, refer to a molecular mechanics protocol without electrostatics, with Coulomb electrostatics, and with GB electrostatics, respectively. noCSA, CSA, and GBSA are the same protocols with an additional surface area term. The results obtained with the best protocols and for the native sequences and Pfam profiles are presented in Table 2. The native sequences have a similarity of 5.13 with themselves. They have an identity of 26.5% and a similarity of 1.05 with the Pfam profiles. The Pfam profiles have an identity of 27.6% and a similarity of 1.02 with themselves. The native and Pfam sequences have 100% of Superfamily recognition. The effects of different factors and their significance are summarized in Table 3.

3.1 Single-position designs

We first predicted the amino acid type of individual positions while the rest of the protein was maintained in the native sequence and crystallographic conformation. We investigated first the contribution of the electrostatics treatment (no electrostatics, Coulomb, or GB) for different protein dielectric constant values (1 to 32). Then the contribution of the surface area term was assessed. The results are presented in Figure 2 and in the upper half of Table 2. The best dielectric constant is defined as the one maximizing the identity vs native and similarity vs Pfam scores.

3.1.1 Electrostatic terms

The contribution of the electrostatic terms was first investigated. The electrostatics treatments compared were a simple Coulomb term and a Generalized Born term with different dielectric constants (1–8, 10, 12, 16, 24, or 32), or no electrostatics. 27 electrostatics protocols were thus tested. The average identity vs native sequences and similarity vs Pfam profiles as a function of the electrostatics treatment are shown in Figure 2 for noC, C, and GB protocols (solid lines). Additional scores at the optimal dielectric constants are given in Table 2 (first three lines). We observe that the MM energy function without electrostatics and SA predicts single amino acids with 22.6% of identity vs native and -0.37 of similarity vs Pfam. Adding Coulomb electrostatics significantly improves the results to 25.8% of identity vs native and -0.12 of similarity vs Pfam, at a dielectric constant of 24. The GB term yields a significantly better performance with a dielectric constant of 2 for the protein (and 80 for the solvent): 32.0% of identity vs native and 0.02 of similarity vs Pfam. The score profiles as a function of the dielectric constant are comparable for the identity vs native and similarity vs Pfam scores.

3.1.2 Surface area term

We next investigated the contribution of a surface area term. This term is obtained from the atomic components of the solvent accessible surface area of the protein, multiplied by atom-type specific surface energy coefficients. An optimization of these coefficients was conducted for each electrostatics protocol. Three coefficient models were investigated. First, a uniform σ coefficient was applied to all atoms. A second model used two coefficients: σNA for non-polar and aromatic atoms, and σPI for polar and ionic atoms. A third model used four classes of atoms: non-polar (σN), polar (σP), aromatic (σA), and ionic (σI) (see our sidechain prediction article [45] for details). The optimization was done by grid search with an increment of 10 cal/mol/Å^2. Optimal values of the coefficients according to the identity score vs native sequences are given in Table 4 for each electrostatics protocol.

The average identity vs native sequences and similarity vs Pfam profiles for noCSA, CSA, and GBSA with the optimal σNPAI coefficients are shown in Figure 2 (dashed lines). Additional descriptors at the optimal dielectric constants are given in Table 2 (lines 4 to 6). The SA term brings a significant and approximately uniform improvement to the different protocols. For noC, the gain is 8.5% for the identity vs native, and 0.68 for the similarity vs Pfam. For Coulomb, the improvement is 7.9% and 0.52, for an optimal dielectric constant of 16. For GB, it is 4.9% and 0.44, for an optimal dielectric constant of 2. The score profiles as a function of the dielectric constant are not significantly modified by the SA term. SA results obtained

with the two-coefficient model are not significantly poorer than with the four-coefficient model; the uniform-coefficient model, however, barely improves the results obtained without SA (see Supplementary Figure S1). With GB, the optimal surface coefficients ranged from −90 to 10 cal/mol/Å^2 for non-polar atoms, from −150 to −30 cal/mol/Å^2 for polar and ionic atoms, and from 0 to 70 cal/mol/Å^2 for aromatic atoms.

3.1.3 Comparison between protocols

The protocol hierarchy is GBSA > CSA, noCSA, GB > C > noC, in terms of identity vs native sequences, and GBSA, CSA, noCSA > GB, C > noC, in terms of similarity vs Pfam profiles. Equivalence between protocols was determined with ANOVA tests at the 0.05 significance level. The GBSA protocol, at a dielectric constant of 2 and σNPAI coefficients of −90, −130, 40, and −110 cal/mol/Å^2, has an identity vs native of 36.9% and a similarity vs Pfam of 0.45.

3.2 All-position designs

In the second phase of this study, we performed full sequence designs, optimizing the amino acid type of all positions of a protein at the same time. We investigated first the contribution of the electrostatics treatment (noC, C, GB) for different dielectric constant values (1 to 32). Then the contribution of the SA term was assessed. The results are presented in Figures 3 and 4, and in the lower half of Table 2. The best dielectric constant is defined as the one maximizing the identity vs native, similarity vs Pfam, and Superfamily scores. The results obtained with the Rosetta program are also given in Table 2. Rosetta results were 31.5% identity vs native, a similarity vs Pfam of 0.32, and 100% Superfamily recognition. Results for core and surface positions, and for family-specific reference energies are also given. Finally, examples of designed core sequences are shown.

3.2.1 Electrostatic terms

The contribution of the electrostatic terms was first investigated. The average identity vs native sequences, similarity vs Pfam profiles, and Superfamily recognition as a function of the electrostatics treatment are shown in Figures 3 and 4 for noC, C, and GB protocols (solid lines). Additional scores at the optimal dielectric constants are given in Table 2 (lines 7 to 9). With noC, we obtain 12.7% identity vs native, a similarity vs Pfam of −0.82, and 1.6% Superfamily recognition. With Coulomb and a dielectric constant of 12, the predictions are not statistically improved (except for the identity vs Pfam) with 13.1% identity vs native, a similarity vs Pfam of −0.77, and 2.6% Superfamily recognition. The GB term, with a dielectric constant of 6, brings

a significant improvement to the Coulomb results with 16.7% identity vs native, similarity vs Pfam of −0.49, and 16.6% of Superfamily recognition. Compared to the single-position designs, the scores as a function of the dielectric constant (Figure 3) present a deterioration for GB at the lowest ε values (1–4) and, as a consequence, a shift of the optimal ε towards higher values (6–10) instead of 2.

3.2.2 Surface area term

The contribution of the SA term was then investigated for each electrostatics protocol. The σNPAI coefficient model was used with the coefficients optimized above, in the single-position study. The average identity vs native sequences, similarity vs Pfam profiles, and Superfamily recognition for noCSA, CSA, and GBSA are shown in Figures 3 and 4 (dashed lines). Additional descriptors are given in Table 2 (lines 10 to 12). The SA term brings a significant improvement to the design scores, for all electrostatics protocol. The noC results are increased by 7.2% for the identity vs native, 0.81 for the similarity vs Pfam, and 83.8% for the Superfamily recognition. The Coulomb results are increased by 8.1% for the identity vs native, 0.79 for the similarity vs Pfam, and 87.1% for the Superfamily recognition, at $\varepsilon = 16$. The GB results are also significantly improved, by 4.7% for identity vs native, 0.49 for similarity vs Pfam, and 80.6% for Superfamily recognition, at $\varepsilon = 8$.

3.2.3 Results for core and surface positions

The average identity and similarity vs native sequences and Pfam profiles for core and surface positions are shown in Supplementary Tables S4 and S5. For core positions, the scores are much better than average, in particular with the SA term. For example, with GBSA and $\varepsilon = 8$, we obtain a 42.2% identity vs native and a similarity vs Pfam of 0.81. The difference between our best protocol and Rosetta (48.7% for identity vs native and 0.96 for similarity vs Pfam) is less important than for all positions. For surface positions, the scores are poorer than average. The SA term makes no improvement to the identity percentages but improves the similarity scores. The noCSA protocol achieves 8.7% of identity vs native and a similarity vs Pfam of −0.54. Rosetta performs better, with 16.4% and −0.22.

3.2.4 Comparison between protocols

The protocol hierarchy is GBSA, CSA, noCSA > GB > C, noC. Equivalence between protocols was determined with ANOVA tests at the 0.05 significance level. The GBSA protocol, at a dielectric constant of 8 and σNPAI coefficients of −30, −90, 10, and −90 cal/mol/Å^2, has an

identity vs native of 21.4%, a similarity vs Pfam of 0.00, and 97.2% Superfamily recognition. The initial guess and optimized reference energies for this protocol are given in Supplementary Table S3.

3.2.5 Results with family-specific reference energies

We have examined the effect of optimizing the reference energies separately for each family (SH3, SH2, and PDZ) instead of globally. A small, not significant improvement of identity and similarity scores is obtained. The identity vs native is 21.6%, the similarity vs Pfam is 0.05, and the Superfamily recognition is 96%.

3.2.6 Examples of designed cores

Examples of designed sequences are shown in Figure 5, and Supplementary Figures S2 and S3, for hydrophobic core positions of SH3, PDZ, and SH2 proteins. The designed profiles are visualized as sequence logos. The protocol employed is GBSA with $\varepsilon = 8$, which gives the best overall results. For a lot of positions, we observe a remarkable recovery of the native amino acid type. When this is not the case, the correct chemical class of the amino acid is often respected. In particular, our method is able to correctly distinguish between aromatic and aliphatic amino acids in many cases.

4 Discussion

We have redesigned the full sequences of 9 proteins, to assess the performance of an MMGBSA energy function and of the Proteus software, which applies some specific approximations. We discuss the contributions of the different energy terms. Then, we consider the compatibility of the optimal parameters found here with those found for sidechain conformation predictions. Finally, we compare our results to other studies by our group and to those obtained with Rosetta.

4.1 Single-position designs

Single-position sequence design is a simplified problem that can be solved rapidly, by enumeration, so that extensive tests and parameter exploration are possible. Single-position predictions of sidechain conformations have been performed several times (reviewed in Gaillard et al. [45]) as a test and to optimize scoring functions. Single-position sequence designs were less often studied. They were used in the optimization of the Rosetta energy function [5, 15]. In this work, we used single-position designs to optimize the surface energy coefficients. Single-position designs are also valuable in the interpretation of the results. They provide an upper limit on

the accuracy we can expect for the full problem. Indeed there is no influence of badly-predicted amino acids on the others and no pairwise-additivity error for the solvation energy, in contrast to all-position designs, where the GB and SA terms require specific approximations for pairwise additivity. Here, going from single- to all-position designs, there was a performance loss of 15.6% for the identity vs native score, and 0.45 for the similarity vs Pfam score.

4.2 Coulomb and GB energy terms

Electrostatics is an important component of an energy function for CPD [29, 34, 81]. Different models have been used including knowledge-based terms [82], Coulomb with a fixed [11, 18] or distance-dependent [37, 83, 84] dielectric constant, Generalized Born [16, 23, 38, 42, 53, 85], or Poisson-Boltzmann [86, 87]. A separate term to model hydrogen bonds is sometimes added [60, 83, 84]. Here, we tested pure Coulomb and GB electrostatics, with different values of the interior dielectric constant.

In single-position designs, the scores of the Coulomb term increase with the value of the dielectric constant and reach a plateau for $\varepsilon \geq 16$, while the scores of the GB term are better at the lowest values of ε (1–4, optimum at 2) then decrease and reach a plateau. The scores of the protocol without electrostatics are lower than the best Coulomb scores. We have thus noC < C(ε=16–32) < GB(ε=1–4). We recall that with the Coulomb term, the same dielectric constant is applied to the different regions of the proteins. This is a severe approximation as a protein interior is known to have a dielectric constant of about 2–5 while the protein surface value is larger.[88] The best results with the Coulomb term are obtained at $\varepsilon = 24$, a compromise between interior and surface values. The degradation of Coulomb at low ε values corresponds to an excessive weight of electrostatic interactions, leading to sequences containing mostly ionic amino acids.

In all-position designs, the most striking difference with single-position designs is the behaviour of the GB term. A degradation is now observed at the lowest values of ε (1–4), the best scores are obtained at ε = 6–10, an intermediate plateau is then reached. The Coulomb behaviour and the relative position of the noC protocol are similar to single-position designs. The protocol hierarchy is thus noC < C(ε=10–32) < GB(ε=6–10). The degradation of GB at low ε values can be attributed to pairwise decomposition errors. Due to the many-body character of the Born radii calculation, it is indeed necessary in all-position designs to replace the exact GB term by a "native environment" approximation to be able to decompose it into a sum over pairs. This approximation introduces errors that were shown in a previous work to be inversely proportional to the dielectric constant value [53], so that they are largest with a low ε.

4.3 Solvent accessible surface area term

The hydrophobic effect corresponds to the unfavorable disruption of the network of water molecules by a non-polar solute [89]. A linear relationship has been observed between the solvation energy of alkane molecules and their surface area [90]. This prompted the use of solvent accessible surface area terms with a uniform surface energy coefficient to model the hydrophobic effect [51, 91, 92]. It seems reasonable that the perturbation of the water hydrogen-bond network is modulated by the polarity of the solute atoms. Non-polar atoms will have a more disruptive effect than polar ones. This led to the use of atom-dependent surface energy coefficients [93–96]. Solvent accessible surface area terms have been used in CPD [11, 16, 18, 38, 83]. An alternative has been the Lazaridis-Karplus solvation model [35, 37, 58, 97]. In this work, we tested SA models with a uniform coefficient, two coefficients (non-polar/aromatic and polar/ionic atoms), and four coefficients (non-polar, polar, aromatic, ionic atoms).

The SA term with the four coefficient model brings an important improvement to the noC, C, and GB protocols, in both single- and all-position designs. The score increase is approximately independent of the dielectric constant value. In single-position designs, we have shown that the benefit of the SA term is also important with the two coefficient model while it is almost null with the uniform coefficient model. This indicates that the polar/non-polar atom distinction is critical in the choice of the surface energy coefficients. The improvement is less important for GB than for noC and C. This can be explained by some redundancy between the effects modeled by the SA term and the GB self term. The GB self term, or Born equation, models the favorable interaction between a charge and the solvent, decreasing with the charge distance to the solvent, and thus favors polar residues at the surface. When the GB self term is absent, this effect can be approximately captured by the SA term with surface energy coefficients favoring polar atoms. The benefit of the SA term is particularly spectacular in terms of Superfamily recognition. The GBSA protocol indeed has 97% sequence recognition. Without the SA term, the GB protocol only reaches 17% sequence recognition while noC and C protocols are below 3%.

For core positions, the effect of the SA term on the scores is similar but more important in magnitude than for all positions. The core composition is modified by the SA term, with a decrease in polarity. For example, comparing GB and GBSA protocols, the core polarity measured by the percentage of polar and ionic amino acids drops from 65 to 26%. For surface positions, the SA term has no effect on the identity scores, but significantly improves the similarity scores. This indicates that, while the native sequence is not recovered better with the SA term, the assignment of amino acids into correct physical-chemical classes is improved. The surface com-

position is altered by the SA term, with an increase of polarity. For example, comparing GB and GBSA protocols, the surface polarity goes from 36 to 69%.

The main benefit of the SA term in sequence design is to favor polar over non-polar residues at the surface. The exposure of polar residues to the solvent and the burial of non-polar residues are fundamental traits of protein composition. Such an amino acid polarity gradient with respect to the solvent exposure cannot be obtained with van der Waals and Coulomb terms alone, and only partially by the addition of the GB self term, at least when ε is large (≥ 8). We observe that the SA term not only favors polar residues at the surface, as expected, but also favors non-polar residues in the core. As the solvent accessible surface area of core residues is almost zero, it would be expected that the SA term has almost no effect in the core. The unfolded state has to be taken into account. In the absence of an SA term, polar amino acids are distributed all over the protein, including in the core. We now add an SA term with energy coefficients favoring polar over non-polar atoms ($\sigma\mathrm{P} < \sigma\mathrm{N}$). At the surface of the folded state, we consider that SA = 2, while in the core, SA = 0. In the unfolded state, we consider that all residues are intermediately exposed, SA = 1. The folding energy difference between a polar and a non-polar amino acid at the surface of the folded state is thus $(\sigma\mathrm{P} - \sigma\mathrm{N})(2 - 1) < 0$, while in the core it is $(\sigma\mathrm{P} - \sigma\mathrm{N})(0 - 1) > 0$. The effect of the SA term is thus to favor polar amino acids at the surface and non-polar amino acids in the core.

It has to be noted that, in theory, the reference energies are not explicitly related to a structural and energetical model of the unfolded state, as they are optimized targeting an amino acid composition criterion. In practice, we observe that the optimized reference energies are close to the initial guess (correlation coefficient > 0.99, RMSD = 1.14 kcal/mol, see Supplementary Table S3), obtained by averaging the matrix diagonal terms (see Methods). Thus a qualitative interpretation of the reference energies as the energy of the unfolded state remains valid.

4.4 Compatibility with sidechain conformation predictions

We recently evaluated an MMGBSA energy function for sidechain predictions [45], following a parallel approach to the one presented here. The contributions of the different energy terms to the accuracy of sidechain predictions were quite different from those found in the current work. This is not surprising as distinguishing between conformations of an amino acid is a problem of a different nature than distinguishing between amino acid types. In particular, the challenge of respecting the polarity gradient between the hydrophobic core and the hydrophilic surface is not present at fixed sequence. Sidechain prediction is a component of protein design. In structure-based CPD approaches, the conformations are predicted at the same time as the

amino acid types. When designing an active site, the orientation of the sidechains represents valuable information. It is of interest to assess the compatibility of the optimal parameters found for sidechain conformation and for sequence design.

We previously found that, for sidechain conformation predictions, the Coulomb term can bring a significant improvement over van der Waals alone with a specific choice of the dielectric constant. The GB term does not improve or degrade the Coulomb performance level, but the optimal protein dielectric constant with GB is physically plausible, contrary to the value used with the Coulomb term alone. The SA term also does not improve or degrade performance, except for a small improvement in single-position predictions. The GBSA, GB, CSA, and Coulomb protocols were found to be statistically equivalent. Here, for sequence design, we find that the best protocol is GBSA with $\varepsilon = 6$–10. We have tested the optimal energy function obtained here for sidechain conformation predictions. We find a significant degradation with respect to the optimal sidechain conformation prediction protocol (0.84 vs 0.67 Å for the RMSD, 71 vs 79% for χ_{all}). The degradation is less important for core residues (0.49 vs 0.43 Å for the RMSD, 85 vs 87% for χ_{all}). A two-stage approach is possible in CPD applications where the sidechain conformations are of interest. Sequence design is performed first with the optimal energy function found in this work. Sidechain conformation optimization at fixed sequence is then performed with the optimal energy function found in Gaillard et al. [45]. The additional cost is marginal as the same energy matrix can be used in the two stages (if enough information has been saved).

4.5 Comparison to other Proteus studies

In an earlier study, full sequence design of SH3, SH2, SKI, chemokine, PDZ, and caspase proteins was performed [19, 47] with Proteus. The energy function was of the CSA form, with a dielectric constant of 10, and surface energy coefficients of 12 cal/mol/Å^2 for carbon and sulfur, -60 for oxygen and nitrogen, -150 for ionic groups, and zero for hydrogens. The reference energies were optimized to obtain a natural amino acid composition, as here. The SH3, SH2, and PDZ lowest energy sequences of Schmidt am Busch et al. were reanalyzed with the protocol used in this work. We obtained an identity vs native of 19.5%, compared to 21.2/21.4% obtained here with CSA/GBSA. The similarity vs native was 0.52, compared to 0.44/0.53 obtained here with CSA/GBSA. The Superfamily recognition was 84.9%, compared to 89.7/97.2% obtained here with CSA/GBSA. The main protocol differences here are the all-atom force field (AMBER ff99SB, vs. CHARMM19 previously) and the GB term.[42, 53] While the GB term was found in this work to significantly improve the Coulomb results, we found that the GBSA protocol is statistically equivalent to CSA. Indeed the SA term can serve as a substitute for the improve-

ment brought by GB. The comparison also shows that the ff99SB all-atom force field does not importantly improve the CHARMM19 sequence design capabilities. The major contributor to design scores was found in this work to be a SA term with differentiated polar and non-polar atom coefficients, already present in the earlier study.

In a recent study, an improved energy was used but only PDZ proteins were designed [54]. The energy function was of the same GBSA form as here, with a dielectric constant of 8, and surface energy coefficients of -5, -80, -40, and -100 cal/mol/Å^2 for non-polar, polar, aromatic, and ionic atoms. The sequences produced for the Tiam1 and CASK proteins were reanalyzed with the protocol used here. We found an equivalent quality (15.7 and 22.8% identity vs native, -0.25 and 0.04 similarity vs Pfam, 100% Superfamily recognition). Design of our protein set with the energy function of Mignon et al. was also performed. We obtained 21.1% identity vs native, -0.13 similarity vs Pfam, and 81% Superfamily recognition. These scores are improved to 20.7%, 0.01, and 96% if a family-specific instead of global reference energy optimization is conducted. The proximity of these results to those obtained here indicates that the energy function is robust to small differences in the surface energy coefficients, providing that the polar/non-polar distinction is maintained.

4.6 Comparison to Rosetta

We also compared our design results to Rosetta [5, 14, 35, 36], the most popular specialized protein design program. Our predictions with the GBSA protocol ($\varepsilon = 8$ and σNPAI coefficients) are behind Rosetta in terms of identity vs native (21 vs 32%) and similarity vs Pfam (0.00 or 0.05 with family-specific reference energies vs 0.32), and close in terms of Superfamily recognition (97 vs 100%). The difference is reduced for core residues (42 vs 49% of identity vs native, 0.81 vs 0.96 of similarity vs Pfam).

Despite its reduced accuracy, the advantages of our method are that it is more general and transferable, and has more explanatory power. The Rosetta energy function is highly optimized and of empirical character (see Introduction), while ours is based as much as possible on physical principles. We use a standard, widely-used molecular mechanics force field (AMBER ff99SB). This allows us to easily model the effect of non-protein environments, such as ligands, nucleic acids, or lipids. We can benefit from the latest developments and additions in force field libraries and solvation models. We use an empirically-derived rotamer library, but the rotamer probabilities are not included in our energy function, contrary to Rosetta. In addition to the reference energies, there are five adjustable parameters in our energy function (the dielectric constant and σNPAI coefficients) and only three if the σNP model is used. The Rosetta scoring

function contains 8 overall weight factors, plus the reference energies, the rotamer probabilities, the distance-dependent dielectric constant, the parameters of the hydrogen-bond term and of the Lazaridis-Karplus solvation model, not counting the terms involving the backbone only. Thus, the physical nature of the different contributions to protein stability is not obvious in knowledge-based scoring functions.

5 Conclusion

We have evaluated a pairwise additive MMGBSA energy function, implemented in our CPD program Proteus, for the design of full protein sequences. The dielectric constant, the surface energy coefficients, and the reference energies were optimized. The respective contribution of the MMGBSA energy components was assessed. In addition to all-position designs, we also performed single-position designs. This simpler problem is useful to provide an upper bound for the accuracy of all-position designs, removing in particular the effect of pairwise decomposition errors. Overall, we obtained native-like protein cores and almost 100% recognition of our sequences by Superfamily, an HMM-based sequence classification tool. Our sequences are of slightly lower quality than Rosetta. However, the use of a physical-based energy function has advantages, such as compatibility with biomolecular simulation force fields and a better power of interpretation.

To our knowledge, this is the first systematic evaluation of an MMGBSA energy function and its components for full sequence design. We found that the GB term can bring a significant improvement to pure Coulomb electrostatics with a specific choice of the interior dielectric constant. In all-position designs, the GB improvement is reduced as it is affected by pairwise decomposition errors. This error could be removed in future work, since an efficient, exact (many-body) GB method [85] was recently implemented in Proteus (manuscript in preparation). The surface area term, with distinct energy coefficients for polar and non-polar atoms, brings a spectacular improvement, in particular in terms of Superfamily recognition. This term is needed to establish an amino acid polarity gradient between the protein core and surface. The improvement is roughly independent of the electrostatics protocol, but slightly more important with Coulomb than with GB. This is interpreted as a partial redundancy between the physical effects captured by the GB and SA terms. Further developments in physical-based energy functions for protein design could include the use of GB or PB models less sensitive to pairwise decomposition errors [85], and an improved model of solute-solvent dispersion interactions [98].

References

1. Pabo C. Designing proteins and peptides. Nature 1983;301:200.

2. Ponder JW, Richards FM. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775–791.

3. Dahiyat BI, Mayo SL. De novo protein design: Fully automated sequence selection. Science 1997;278:82–87.

4. Koehl P, Levitt M. De novo protein design. I. In search of stability and specificity. J Mol Biol 1999;293:1161–1181.

5. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97:10383–10388.

6. Suárez M, Jaramillo A. Challenges in the computational design of proteins. J R Soc Interface 2009;6:S477–S491.

7. Saven JG. Computational protein design: Engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. Curr Opin Chem Biol 2011;15:452–457.

8. Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG. Theoretical and computational protein design. Annu Rev Phys Chem 2011;62:129–149.

9. Pantazes RJ, Grisewood MJ, Maranas CD. Recent advances in computational protein design. Curr Opin Struct Biol 2011;21:467–472.

10. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282:1462–1467.

11. Wernisch L, Hery S, Wodak SJ. Automatic protein design with all atom force-fields by exact and heuristic optimization. J Mol Biol 2000;301:713–736.

12. Jaramillo A, Wernisch L, Héry S, Wodak SJ. Folding free energy function selects native-like protein sequences in the core but not on the surface. Proc Natl Acad Sci USA 2002; 99:13554–13559.

13. Larson SM, England JL, Desjarlais JR, Pande VS. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. Protein Sci 2002;11:2804–2813.

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

15. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.

16. Pokala N, Handel TM. Energy functions for protein design: Adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J Mol Biol 2005;347:203–227.

17. Ambroggio XI, Kuhlman B. Computational design of a single amino acid sequence that can switch between two distinct protein folds. J Am Chem Soc 2006;128:1154–1161.

18. Schmidt am Busch M, Lopes A, Mignon D, Simonson T. Computational protein design: Software implementation, parameter optimization, and performance of a simple model. J Comput Chem 2008;29:1092–1102.

19. Schmidt am Busch M, Sedano A, Simonson T. Computational protein design: Validation and possible relevance as a tool for homology searching and fold recognition. PLOS ONE 2010;5:e10410.

20. Clark LA, Boriack-Sjodin PA, Eldredge J, Fitch C, Friedman B, Hanf KJ, Jarpe M, Liparoto SF, Li Y, Lugovskoy A, Miller S, Rushe M, Sherman W, Simon K, van Vlijmen H. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. Protein Sci 2006;15:949–960.

21. Lazar GA, Dang W, Karki S, Vafa O, Peng JS, Hyun L, Chan C, Chung HS, Eivazi A, Yoder SC, Vielmetter J, Carmichael DF, Hayes RJ, Dahiyat BI. Engineered antibody Fc variants with enhanced effector function. Proc Natl Acad Sci USA 2006;103:4005–4010.

22. Lopes A, Schmidt am Busch M, Simonson T. Computational design of protein-ligand binding: Modifying the specificity of asparaginyl-tRNA synthetase. J Comput Chem 2010; 31:1273–1286.

23. Polydorides S, Amara N, Aubard C, Plateau P, Simonson T, Archontis G. Computational protein design with a Generalized Born solvent model: Application to asparaginyl-tRNA synthetase. Proteins 2011;79:3448–3468.

24. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ Jr, Stoddard BL, Baker D. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 2006;441:656–659.

25. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D. De novo computational design of retro-aldol enzymes. Science 2008;319:1387–1391.

26. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. Nature 2008;453:190–195.

27. Janin J, Wodak S, Levitt M, Maigret B. Conformation of amino acid side-chains in proteins. J Mol Biol 1978;125:357–386.

28. Voigt CA, Gordon DB, Mayo SL. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. J Mol Biol 2000;299:789–803.

29. Pokala N, Handel T. Review: Protein design–where we were, where we are, where we're going. J Struct Biol 2001;134:269–281.

30. Gainza P, Nisonoff HM, Donald BR. Algorithms for protein design. Curr Opin Struct Biol 2016;39:16–26.

31. Gordon DB, Marshall SA, Mayo SL. Energy functions for protein design. Curr Opin Struct Biol 1999;9:509–513.

32. Mendes J, Guerois R, Serrano L. Energy estimation in protein design. Curr Opin Struct Biol 2002;12:441–446.

33. Boas FE, Harbury PB. Potential energy functions for protein design. Curr Opin Struct Biol 2007;17:199–204.

34. Li Z, Yang Y, Zhan J, Dai L, Zhou Y. Energy functions in de novo protein design: Current challenges and future prospects. Annu Rev Biophys 2013;42:315–335.

35. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.

36. Liu Y, Kuhlman B. RosettaDesign server for protein design. Nucleic Acids Res 2006; 34:W235–W238.

37. Gainza P, Roberts KE, Georgiev I, Lilien RH, Keedy DA, Chen CY, Reza F, Anderson AC, Richardson DC, Richardson JS, Donald BR. Osprey: Protein design with ensembles, flexibility, and provable algorithms. Methods Enzymol 2013;523:87–107.

38. Pokala N, Handel TM. Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. Protein Sci 2004;13:925–936.

39. Chowdry AB, Reynolds KA, Hanes MS, Voorhies M, Pokala N, Handel TM. An object-oriented library for computational protein design. J Comput Chem 2007;28:2378–2388.

40. Dahiyat BI, Mayo SL. Protein design automation. Protein Sci 1996;5:895–903.

41. Shapovalov MV, Dunbrack RL Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 2011; 19:844–858.

42. Simonson T, Gaillard T, Mignon D, Schmidt am Busch M, Lopes A, Amara N, Polydorides S, Sedano A, Druart K, Archontis G. Computational protein design: The Proteus software and selected applications. J Comput Chem 2013;34:2472–2484.

43. Polydorides S, Michael E, Mignon D, Druart K, Simonson T, Archontis G. Proteus and the design of ligand binding sites. In: Stoddard B, Baker D, editors. Methods in Molecular Biology: Design and Creation of Protein Ligand Binding Proteins. New York: Springer Verlag; 2016.

44. Lopes A, Alexandrov A, Bathelt C, Archontis G, Simonson T. Computational sidechain placement and protein mutagenesis with implicit solvent models. Proteins 2007;67:853–867.

45. Gaillard T, Panel N, Simonson T. Protein side chain conformation predictions with an MMGBSA energy function. Proteins 2016;84:803–819.

46. Schmidt am Busch M, Lopes A, Amara N, Bathelt C, Simonson T. Testing the Coulomb/accessible surface area solvent model for protein stability, ligand binding, and protein design. BMC Bioinform 2008;9:148.

47. Schmidt am Busch M, Mignon D, Simonson T. Computational protein design as a tool for fold recognition. Proteins 2009;77:139–158.

48. Polydorides S, Simonson T. Monte Carlo simulations of proteins at constant pH with Generalized Born solvent. J Comput Chem 2013;34:2742–2756.

49. Srinivasan J, Cheatham TE 3rd, Cieplak P, Kollman PA, Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. J Am Chem Soc 1998;120:9401–9409.

50. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.

51. Chothia C. Hydrophobic bonding and accessible surface area in proteins. Nature 1974; 248:338–339.

52. Aleksandrov A, Polydorides S, Archontis G, Simonson T. Predicting the acid/base behavior of proteins: A constant-pH Monte Carlo approach with Generalized Born solvent. J Phys Chem B 2010;114:10634–10648.

53. Gaillard T, Simonson T. Pairwise decomposition of an MMGBSA energy function for computational protein design. J Comput Chem 2014;35:1371–1387.

54. Mignon D, Panel N, Chen X, Fuentes EJ, Simonson T. Computational design of PDZ domains: Parameterization and performance of a simple model. 2017;in preparation.

55. Wang J, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comput Chem 2000;21:1049–1074.

56. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 2006;65:712–725.

57. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. Proc Natl Acad Sci USA 1997;94:10172–10177.

58. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999; 35:133–152.

59. Looger LL, Hellinga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. J Mol Biol 2001;307:429–445.

60. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326:1239–1259.

61. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature 2003;423:185–190.

62. Fogolari F, Brigo A, Molinari H. The Poisson–Boltzmann equation for biomolecular electrostatics: A tool for structural biology. J Mol Recognit 2002;15:377–392.

63. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE 3rd. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. Acc Chem Res 2000;33:889–897.

64. Simonson T, Archontis G, Karplus M. Free energy simulations come of age: Protein-ligand recognition. Acc Chem Res 2002;35:430–437.

65. Homeyer N, Gohlke H. Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. Mol Inform 2012;31:114–122.

66. Yang T, Wu JC, Yan C, Wang Y, Luo R, Gonzales MB, Dalby KN, Ren P. Virtual screening using molecular simulations. Proteins 2011;79:1940–1951.

67. Genheden S, Ryde U. Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. Proteins 2012;80:1326–1342.

68. Hou T, Wang J, Li Y, Wang W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J Chem Inf Model 2011;51:69–82.

69. Shah PS, Hom GK, Ross SA, Lassila JK, Crowhurst KA, Mayo SL. Full-sequence computational design and solution structure of a thermostable protein variant. J Mol Biol 2007; 372:1–6.

70. Liu H, Chen Q. Computational protein design for given backbone: Recent progresses in general method-related aspects. Curr Opin Struct Biol 2016;39:89–95.

71. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.

72. Tufféry P, Etchebest C, Hazout S. Prediction of protein side chain conformations: A study on the influence of backbone accuracy on conformation stability in the rotamer space. Protein Eng 1997;10:361–372.

73. Brünger AT. X-Plor Version 3.1. A System for X-Ray Crystallography and NMR. New Haven: Yale University Press; 1992.

74. Hawkins GD, Cramer CJ, Truhlar DG. Pairwise solute descreening of solute charges from a dielectric medium. Chem Phys Lett 1995;246:122–129.

75. Hawkins GD, Cramer CJ, Truhlar DG. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. J Phys Chem 1996;100:19824–19839.

76. Moulinier L, Case DA, Simonson T. Reintroducing electrostatics into protein X-ray structure refinement: Bulk solvent treated as a dielectric continuum. Acta Crystallogr Sect D 2003; 59:2094–2103.

77. Lee B, Richards FM. The interpretation of protein structures: Estimation of static accessibility. J Mol Biol 1971;55:379–400.

78. Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539.

79. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.

80. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 2001;313:903–919.

81. Vizcarra CL, Mayo SL. Electrostatics in computational protein design. Curr Opin Chem Biol 2005;9:622–626.

82. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.

83. Liang S, Grishin NV. Effective scoring function for protein sequence design. Proteins 2004; 54:271–281.

84. O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, DiMaio F, Bradley P, Kortemme T, Baker D, Snoeyink J, Kuhlman B. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. J Chem Theory Comput 2015;11:609–622.

85. Archontis G, Simonson T. A residue-pairwise Generalized Born scheme suitable for protein design calculations. J Phys Chem B 2005;109:22667–22673.

86. Marshall SA, Vizcarra CL, Mayo SL. One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. Protein Sci 2005;14:1293–1304.

87. Vizcarra CL, Zhang N, Marshall SA, Wingreen NS, Zeng C, Mayo SL. An improved pairwise decomposable finite-difference Poisson-Boltzmann method for computational protein design. J Comput Chem 2008;29:1153–1162.

88. Simonson T. Electrostatics and dynamics of proteins. Rep Prog Phys 2003;66:737–787.

89. Tanford C. The Hydrophobic Effect: Formation of Micelles and Biological Membranes. New York: Wiley; 1973.

90. Hermann RB. Theory of hydrophobic bonding. ii. the correlation of hydrocarbon solubility in water with solvent cavity surface area. J Phys Chem 1972;76:2754–2759.

91. Simonson T, Brünger AT. Solvation free energies estimated from macroscopic continuum theory: An accuracy assessment. J Phys Chem 1994;98:4683–4694.

92. Sitkoff D, Sharp KA, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.

93. Eisenberg D, McClachlan A. Solvation energy in protein folding and binding. Nature 1986; 319:199–203.

94. Ooi T, Oobatake M, Nemethy G, Scheraga H. Accessible surface areas as a measure of the thermodynamic hydration parameters of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.

95. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.

96. Fraternali F, van Gunsteren WF. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. J Mol Biol 1996;256:939–948.

97. Wagner F, Simonson T. Implicit solvent models: Combining an analytical formulation of continuum electrostatics with simple models of the hydrophobic effect. J Comput Chem 1999;20:322–335.

98. Aguilar B, Shadrach R, Onufriev AV. Reducing the secondary structure bias in the generalized born model via R6 effective radii. J Chem Theory Comput 2010;6:3613–3630.

Figure legends

Figure 1: Cartoon representations of the structures studied.

Figure 2: Single-position design scores as a function of the protocol. The identity score to the native sequence is above, the similarity score to the Pfam profile is below. The horizontal axis is the dielectric constant. noC (blue), C (green, cross points), and GB (red, plus points) protocols are shown as plain lines. noCSA, CSA, and GBSA protocols with optimal σNPAI surface energy coefficients are shown as dashed lines.

Figure 3: All-position design scores as a function of the protocol. See Figure 2 legend.

Figure 4: All-position design Superfamily scores as a function of the protocol. See Figure 2 legend.

Figure 5: Sequence logos for hydrophobic core positions of SH3 proteins. Designed sequences are compared to the native sequences and to the SH3 Pfam profile.

Table 1: Studied structures and properties: PDB code, chain, selection, length, and family.

PDB	chain	selection	length	family
1ABO	A	64-119	56	SH3
1CSK	B	11-66	56	SH3
1CKA	A	134-190	57	SH3
1R6J	A	192-273	82	PDZ
1G9O	A	9-99	91	PDZ
2BYG	A	186-282	97	PDZ
1BM2	A	55-152	98	SH2
1O4C	A	1-105	105	SH2
1M61	A	4-112	109	SH2

Table 2: Single- and all-position design scores for the best protocols. The protocols presented are: noC, C, and GB with the best dielectric constant; noCSA, CSA, and GBSA with the best dielectric constant and optimal σNPAI surface energy coefficients. Results with the Rosetta program, of the native sequences, and of the Pfam profiles are also given. The scores presented are the identity (%) and similarity to the native sequences and to the Pfam profiles, and the percentage of Superfamily hits.

pos.	model	ε	id.	sim.	id.	sim.	Super-
			vs native		vs Pfam		family
Single	noC	-	22.6	0.40	13.1	-0.37	-
Single	C	24	25.8	0.84	13.7	-0.12	-
Single	GB	2	32.0	1.11	16.1	0.02	-
Single	noCSA	-	31.1	1.23	16.4	0.31	-
Single	CSA	16	33.7	1.51	17.3	0.40	-
Single	GBSA	2	36.9	1.66	18.6	0.45	-
All	noC	-	12.7	-0.41	8.6	-0.82	1.6
All	C	12	13.1	-0.38	9.3	-0.77	2.6
All	GB	6	16.7	0.02	11.4	-0.49	16.6
All	noCSA	-	19.9	0.45	14.1	-0.01	85.4
All	CSA	16	21.2	0.44	14.6	0.02	89.7
All	GBSA	8	21.4	0.53	14.5	0.00	97.2
All	Rosetta		31.5	1.14	18.1	0.32	100.0
	Natives		100.0	5.13	26.5	1.05	100.0
	Pfam		26.5	1.05	27.6	1.02	100.0

Table 3: Effect of different factors on sequence design scores. The factors studied are: Coulomb (C vs. noC), GB (GB vs. C), SA(noC) (noCSA vs. noC), SA(C) (CSA vs. C), SA(GB) (GBSA vs. GB), and All vs. Single (all- vs. single-position designs; GBSA). Significant differences, according to a paired t-test at a 0.05 significance level, are shown in boldface.

pos.	factor	id.	sim.	id.	sim.
		vs native		vs Pfam	
Single	C	**+3.2**	**+0.45**	+0.6	**+0.25**
Single	GB	**+6.2**	**+0.26**	**+2.3**	+0.13
Single	SA(noC)	**+8.5**	**+0.83**	**+3.3**	**+0.68**
Single	SA(C)	**+7.9**	**+0.67**	**+3.6**	**+0.52**
Single	SA(GB)	**+4.9**	**+0.56**	**+2.5**	**+0.44**
All	C	+0.4	+0.03	**+0.7**	+0.05
All	GB	**+3.7**	**+0.40**	**+2.1**	**+0.28**
All	SA(noC)	**+7.2**	**+0.86**	**+5.5**	**+0.81**
All	SA(C)	**+8.1**	**+0.82**	**+5.3**	**+0.79**
All	SA(GB)	**+4.7**	**+0.51**	**+3.1**	**+0.49**
All vs Single		**−15.6**	**−1.13**	**−4.0**	**−0.45**

Table 4: Optimal surface energy coefficients (cal/mol/Å^2) for each electrostatics protocol. Models with four, two or one surface coefficient are given in each case.

		4 coeff.				2 coeff.		1 coeff.
model	ε	N	P	A	I	NA	PI	all
noCSA	-	−40	−140	10	−170	10	−80	−20
CSA	1	−180	−680	30	−3150	280	−380	20
CSA	2	−20	−960	120	−850	160	−230	40
CSA	3	−10	−720	140	−640	70	−290	120
CSA	4	40	−710	180	−600	120	−220	90
CSA	5	50	−180	160	−140	90	−200	−40
CSA	6	20	−140	160	−130	60	−170	10
CSA	7	10	−140	130	−120	30	−160	−10
CSA	8	20	−130	120	−120	40	−150	−10
CSA	10	−10	−150	60	−130	40	−130	0
CSA	12	40	−110	80	−50	30	−120	−30
CSA	16	10	−120	50	−110	0	−130	−30
CSA	24	−10	−110	20	−110	40	−40	−40
CSA	32	−20	−80	40	−100	0	−40	−50
GBSA	1	−40	−130	70	−100	30	−70	70
GBSA	2	−90	−130	40	−110	40	−80	20
GBSA	3	−70	−120	30	−120	20	−100	0
GBSA	4	−60	−120	30	−110	−10	−80	−80
GBSA	5	−70	−120	20	−120	−20	−90	−90
GBSA	6	−30	−80	20	−90	−10	−80	−90
GBSA	7	−20	−80	20	−80	−10	−80	−80
GBSA	8	−30	−90	10	−90	−30	−90	−80
GBSA	10	−20	−80	40	−90	10	−90	−80
GBSA	12	−70	−130	10	−140	0	−110	−50
GBSA	16	−70	−140	0	−150	10	−90	−40
GBSA	24	10	−110	50	−30	−20	−70	−40
GBSA	32	10	−110	50	−30	−30	−60	−50

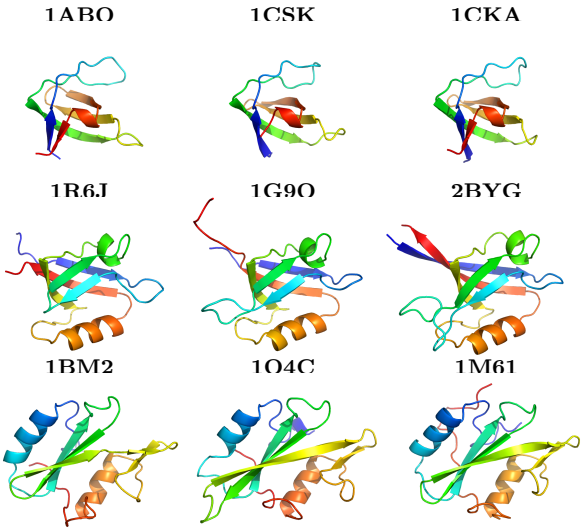

Figure 1

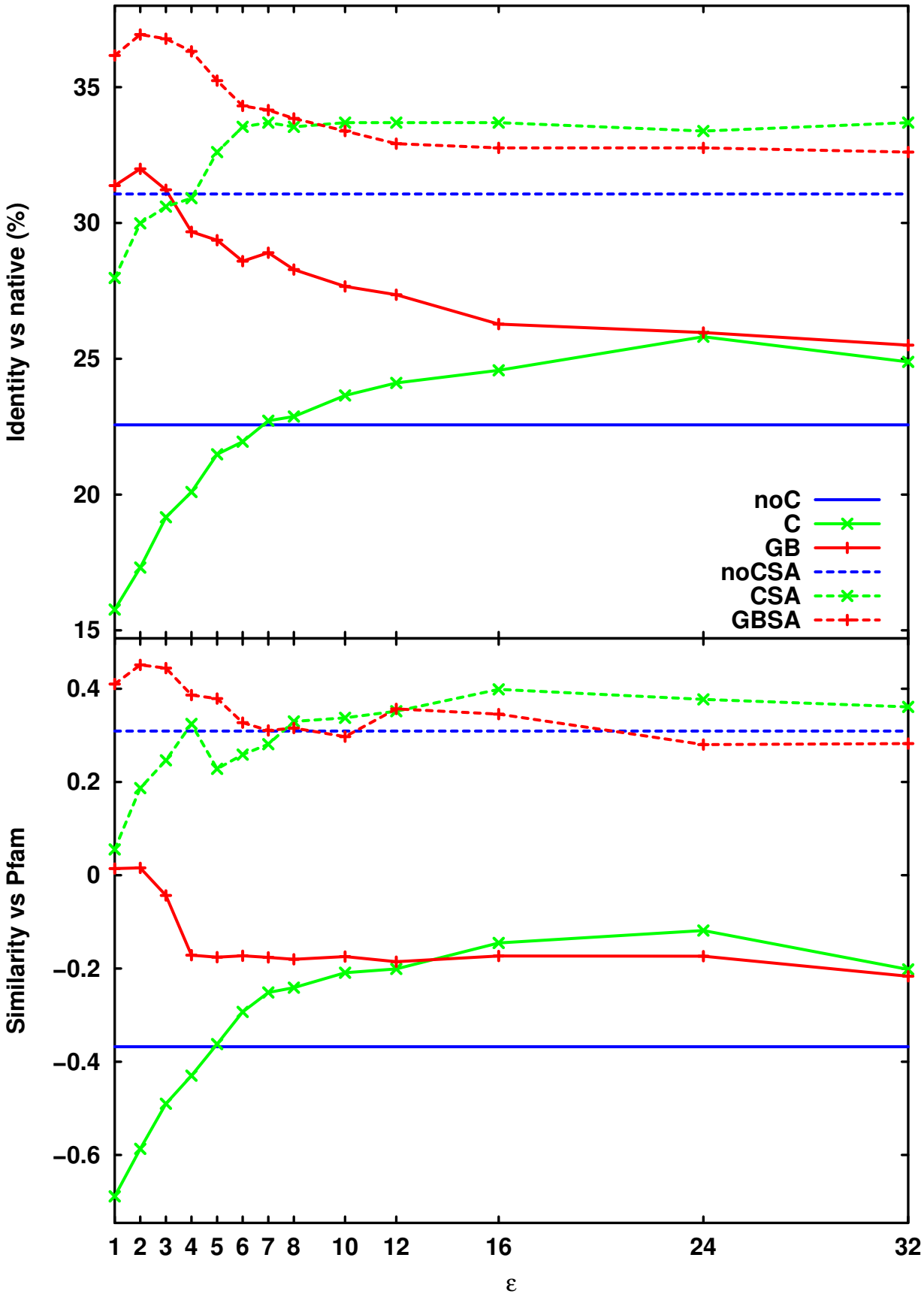
Identity vs native (%)
35
30
25
20
15
noC
C
GB
noCSA
CSA
GBSA
Similarity vs Pfam
0.4
0.2
0
−0.2
−0.4
−0.6
1 2 3 4 5 6 7 8 10 12 16 24 32
ε

Figure 2

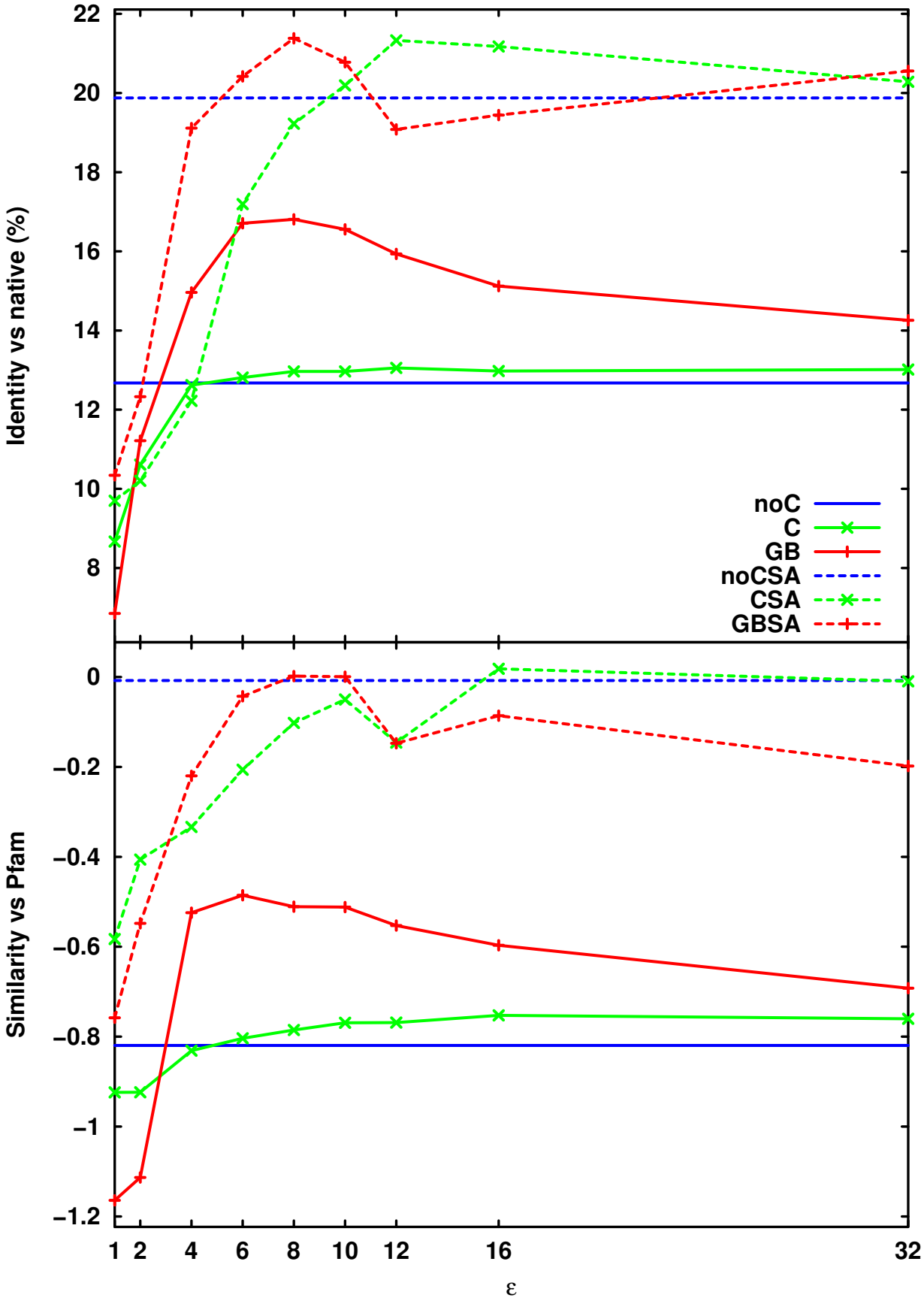
Identity vs native (%)
22
20
18
16
14
12
10
8
noC
C
GB
noCSA
CSA
GBSA
Similarity vs Pfam
0
−0.2
−0.4
−0.6
−0.8
−1
−1.2
1
2
4
6
8
10
12
16
32
ε

Figure 3

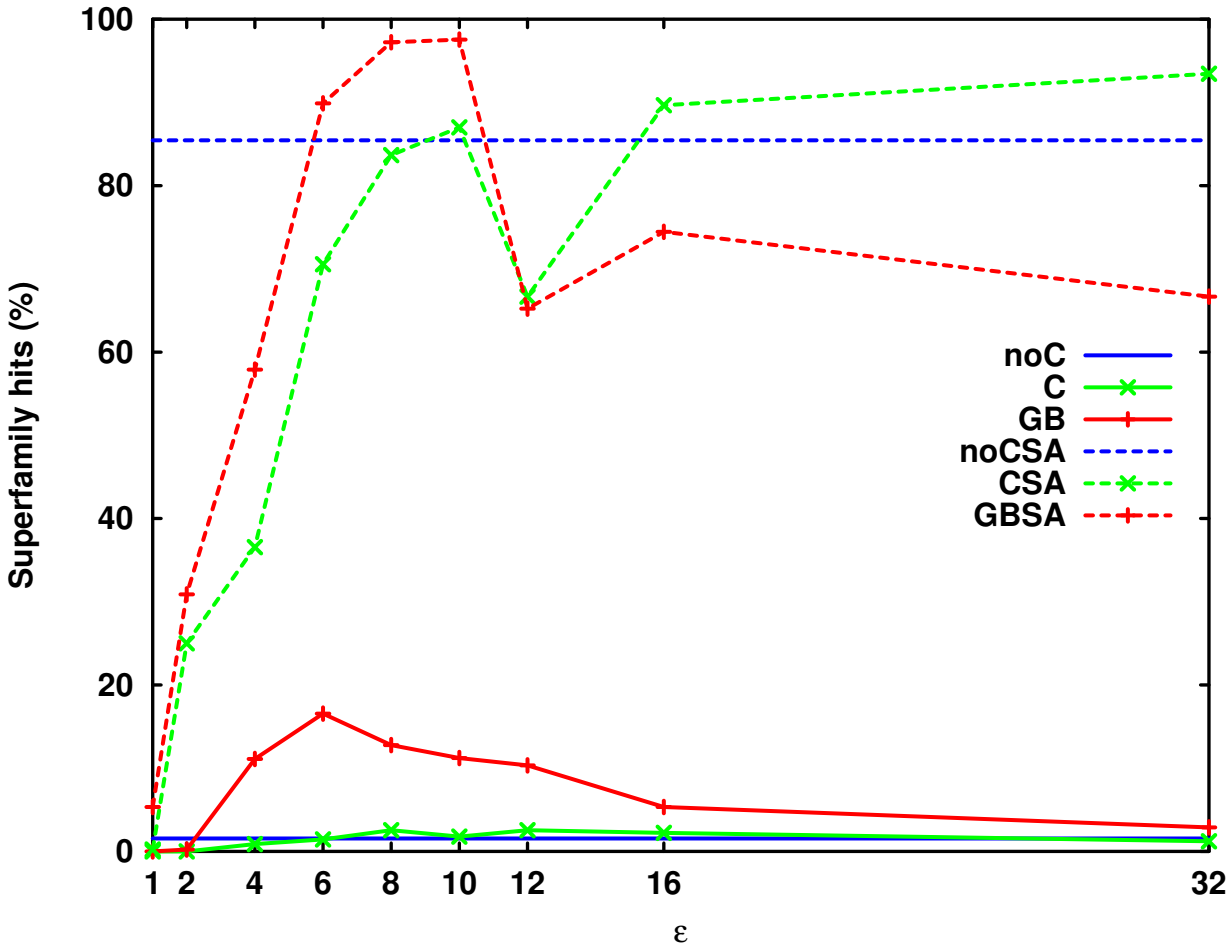
Superfamily hits (%)
100
80
60
40
20
0
1
2
4
6
8
10
12
16
32
ε
noC
C
GB
noCSA
CSA
GBSA

Figure 4


designed
1ABO
68 72 80 82 88 90 100 102 111
native
designed
1CSK
16 20 28 30 36 38 48 50 60
native
designed
1CKA
139 143 151 153 159 161 170 172 182
native
Pfam

Figure 5