Bioinformatics Advance Access published September 27, 2013

Category

Optimized atomic statistical potentials: Assessment of protein interfaces and loops

Guang Qiang Dong[1], Hao Fan[1], Dina Schneidman-Duhovny[1], Ben Webb[1], and Andrej Sali[1,*]

[1]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences (QB3), University of California, San Francisco, UCSF MC 2552, Byers Hall at Mission Bay, Suite 503B, University of California, San Francisco, 1700 4th Street, San Francisco, CA 94158, USA

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Motivation: Statistical potentials have been widely used for modeling whole proteins and their parts (*e.g.*, sidechains and loops) as well as interactions between proteins, nucleic acids and small molecules. Here, we formulate the statistical potentials entirely within a statistical framework, avoiding questionable statistical mechanical assumptions and approximations, including a definition of the reference state.

Results: We derive a general Bayesian framework for inferring *Statistically Optimized Atomic Potentials* (SOAP), in which the reference state is replaced with data-driven "recovery" functions. Moreover, we restrain the relative orientation between two covalent bonds instead of a simple distance between two atoms, in an effort to capture orientation-dependent interactions such as hydrogen bonds. To demonstrate this general approach, we computed statistical potentials for protein-protein docking (SOAP-PP) and loop modeling (SOAP-Loop). For docking, a near-native model is within the top 10 scoring models in 52% of the PatchDock benchmark cases, compared to 23% and 27% for the state-of-the-art ZDOCK and FireDock scoring functions, respectively. Similarly, for modeling 12-residue loops in the PLOP benchmark, the average mainchain RMSD of the best scored conformations by SOAP-Loop is 1.5Å, close to the average RMSD of the best sampled conformations (1.2Å) and significantly better than that selected by Rosetta (2.1Å), DFIRE (2.3Å), DOPE (2.5Å), and PLOP scoring functions (3.0Å). Our Bayesian framework may also result in more accurate statistical potentials for additional modeling applications, thus affording better leverage of the experimentally determined protein structures.

Availability: SOAP-PP and SOAP-Loop are available as part of MODELLER (http://salilab.org/modeller).

*To whom correspondence should be addressed. Andrej Sali, UCSF MC 2552, Byers Hall at Mission Bay, Suite 503B, University of California at San Francisco, 1700 4th Street, San Francisco, CA 94158, USA. tel: +1 415 514 4227; fax: +1 415 514 4231; e-mail: sali@salilab.org.

1 INTRODUCTION

Computational modeling can be used to predict the structures of whole proteins or their parts (*e.g*., loops and sidechains) as well as complexes involving proteins, peptides, nucleic acids and small molecules (Skolnick *et al.*, 2013; Audie and Swanson, 2012; Dill and MacCallum, 2012; Wass *et al.*, 2011; Ding *et al.*, 2010; Baker and A Sali, 2001). A modeling method requires a conformational sampling scheme for proposing alternative structures and a scoring function for ranking them. Significant progress has been made on both fronts (Moult *et al.*, 2011; Fernández-Recio and Sternberg, 2010). In particular, many physics-based energy functions as well as statistical potentials computed from known protein structures have been described (Tanaka and Scheraga, 1975; Hendlich *et al.*, 1990; Sippl, 1993; Colovos and Yeates, 1993; Kocher *et al.*, 1994; Park and Levitt, 1996; Miyazawa and Jernigan, 1996; Melo and Feytmans, 1997; Reva *et al.*, 1997; Simons *et al.*, 1997; Samudrala and Moult, 1998; Rojnuckarin and Subramaniam, 1999; Jones, 1999; Betancourt and Thirumalai, 1999; Gatchell *et al.*, 2000; Lu and Skolnick, 2001; Melo *et al.*, 2002; Zhou and Zhou, 2002; Keasar and Michael Levitt, 2003; McConkey *et al.*, 2003; Betancourt and Skolnick, 2004; Wang *et al.*, 2004; Summa *et al.*, 2005; Qiu and Elber, 2005; Dehouck *et al.*, 2006; Shen and Sali, 2006; Ferrada *et al.*, 2007; Pierce and Weng, 2007; Andrusier *et al.*, 2007; Zhu *et al.*, 2008; Lu *et al.*, 2008; Benkert *et al.*, 2008; Chuang *et al.*, 2008; Rajgaria *et al.*, 2008; Gao and Jeffrey Skolnick, 2008; Xu *et al.*, 2009; Zhang and Zhang, 2010; Rata *et al.*, 2010; Rykunov and Fiser, 2010; Huang and Zou, 2010; Shapovalov and Dunbrack, 2011; Liu and Vakser, 2011; Fan *et al.*, 2011; Zhao and Xu, 2012; Brenke *et al.*, 2012; Liu and Gong, 2012; Zhou and Skolnick; Cossio *et al.*, 2012; Li *et al.*, 2013).

Derivation of a statistical potential has often been guided by an analogy between a sample of known native structures and the canonical ensemble in statistical mechanics, suggesting that the distributions of spatial features in the sample of native structures follow the Boltzmann distribution (Sippl *et al.*, 1990). Thus, statistical potentials are generally calculated in two steps: (1) extracting a probability distribution of a spatial feature (*e.g.,* a distance spanned by a specific pair of atom types) from a sample of known protein structures and (2) normalizing this distribution by a reference distribution (*e.g.*, the distribution of all distances, regardless of the atom types). Statistical potentials can differ in a number of aspects, including the sample of known protein structures, the protein representation (*e.g.*, centroids of amino acid residues, $\mathbf{C}_\alpha$ atoms, and all atoms), the restrained spatial feature (*e.g.*, solvent accessibility, distance, angles and orientation between two sets of atoms), the sequence features (*e.g.*, amino acid residue types, atom types, residue separation in sequence and chain separation), the treatment of sparse samples and the definition of the reference state. Here, we optimize the accuracy of a statistical potential over most of these aspects. This optimization challenge is addressed by formulating a statistical potential independently from any assumptions grounded in statistical mechanics; instead, we rely on a Bayesian approach based on data alone. While the proposed theory applies to any kind of a statistical potential, we illustrate it by deriving specific statistical potentials for protein-protein docking and loop modeling.

2 METHOD

We begin by defining statistical potentials in terms of distributions extracted from known protein structures (Section 2.1), followed by a description of a protocol to actually compute a statistical potential (Sections 2.2-2.7, Fig. 1).

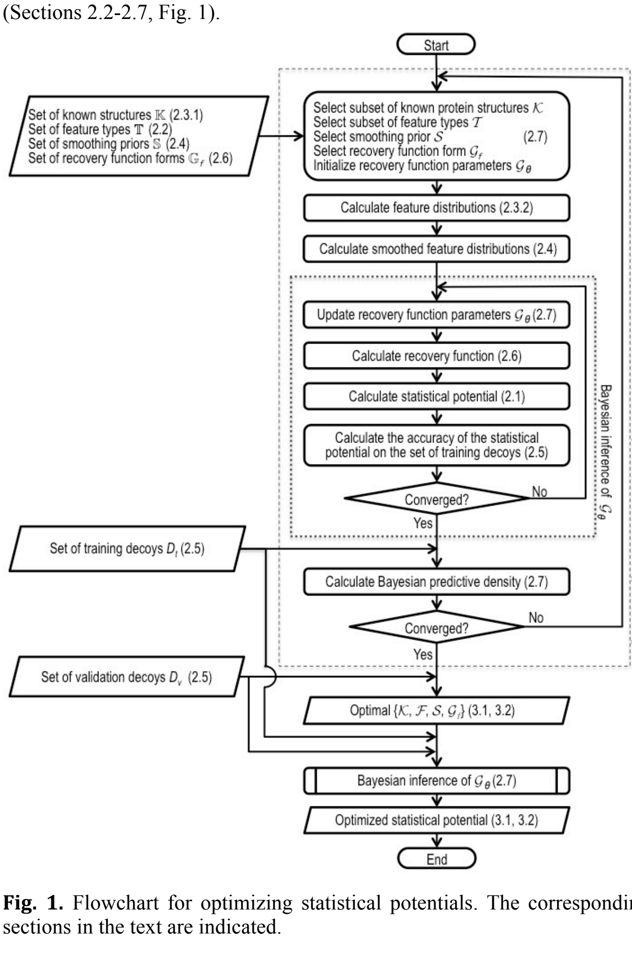

Fig. 1. Flowchart for optimizing statistical potentials. The corresponding sections in the text are indicated.

2.1 Theory

For structure characterization of a given protein sequence by either experiment or theory, we ideally need a joint probability density function (pdf) for the structure, given everything we know about it (Shen and Sali, 2006). In general, our knowledge can come from different kinds of experiments with the protein (*e.g.*, X-ray crystallography), physical theories (*e.g.,* a molecular mechanics force field), and/or statistical inference (*e.g.*, all known structures or only homologous known structures). Here, we focus on a joint pdf for a given sequence based on the knowledge of all known protein structures deposited in the Protein Data Bank (PDB) (Kouranov *et al*, 2006); thus, our joint pdf is a statistical potential.

To derive the joint pdf for a structure of a sequence, we need to approximate it by using terms that can actually be computed from the PDB. The structure X of an amino acid sequence is defined by the set of its features $\{f^{c(m)}\}, m = 1 \dots n$, such as a distance between two specific atoms. Thus, we can approximate the joint pdf by the product of pdfs (restraints) for individual features:

$$p(X) \approx \prod_{1 \le m \le n} p\left(f^{c(m)}\right) \tag{1}$$

Without any loss of accuracy, we define the restraint $p\left(f^{c(m)}\right)$ as the ratio between the feature distribution $p(f^c|\mathcal{Q}_\mathcal{K})$ from a sample of informative features in a set of proteins $\mathcal{Q}_\mathcal{K}$ with known structures (*e.g.*, for a distance, all distances spanned by the same atom types in $\mathcal{Q}_\mathcal{K}$) and an unknown recovery function $g(f^c|\mathcal{Q}_\mathcal{K})$:

$$p\left(f^{c(m)}\right) = p(f^c|\mathcal{Q}_\mathcal{K})/g(f^c|\mathcal{Q}_\mathcal{K}) \tag{2}$$

In other words, the recovery function is defined such that the product of restraints approximates the joint pdf as well as possible (*c.f.*, Eq. 1), while minimizing the number of parameters that need to be fit to the data. Construction of the sample of informative features involves a compromise between including only features of known structures that are most likely to resemble the predicted feature $f^{c(m)}$ (which minimizes sample size) and minimizing the statistical noise (which maximizes sample size). The features used in the sample are termed to be of the same type c as the inferred feature (Section 2.2). The restraints on all features of X of type c are calculated from the same set of informative features, and thus are the same. Here, the sample of informative features includes all features of the same type from representative known protein structures (Section 2.3).

2.2 Feature types

To illustrate the general theory above, we derive optimized statistical potentials for assessing protein-protein interfaces (SOAP-PP) and loop conformations (SOAP-Loop). We restrain the following feature types:

2.2.1 Atomic distance Distance $d|a_1, a_2, b_s$ is considered to depend on atom types a_1 and a_2 as well as the "covalent separation" between the two atoms (b_s). The atom type depends on the residue type, resulting in the total of 158 atoms types for the 20 standard residue types (Shen and Sali, 2006). Covalent separation is measured in three ways. First, by the minimum number of covalent bonds between the two atoms (bond separation). Second, by the number of residues separating the two atoms in the polypeptide chain (residue separation). Third, by chain separation, which is 0 if the atoms are in the same chain and 1 otherwise. The distance is mapped in the range from 0 to a parameterized distance cutoff, such as 15Å.

2.2.2 Orientation between a pair of covalent bonds Orientation $d, \alpha_1, \alpha_2, \psi|t_1, t_2, b_s$ is defined by a distance d, two angles α_1, α_2 and a dihedral angle ψ (Fig. 2). It is considered to depend on covalent bond types (t_1, t_2) defined in turn by their atom types and covalent separation (b_s); there are 316 covalent bond types for the 20 standard residue types.

2.2.3 Relative atomic surface accessibility Accessibility $s|a$ is considered to depend on the atom type (a) (Sali and Blundell, 1993).

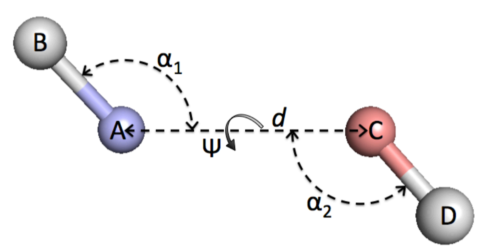

Fig. 2. Distance and angles between two covalent bonds, A-B and C-D. $\boldsymbol{d}$, distance between atoms A and C. $\boldsymbol{\alpha_1}$, angle between atoms B, A and C. $\boldsymbol{\alpha_2}$, angle between atoms A, C and D. $\boldsymbol{\psi}$, dihedral angle between atoms B, A, C and D. $\boldsymbol{b_s}$ is defined using atoms A and C.

2.3 Feature distributions

2.3.1 Known protein structures A small fraction of the known protein structures from the PDB (and their decoy structures) are used only for assessing the accuracy of statistical potentials (Section 2.5). The remaining structures from the PDB are filtered to construct the known protein structure set $\mathbb{K}$, including only structures determined by X-ray crystallography at the resolution better than 2.2 Å and R_{free} better than 25%. Three additional subsets of representative structures were obtained by requiring at most 30%, 60% and 95% sequence identity to any other representative structure, respectively, with preference for structures determined at higher resolutions and with lower R_{free} values. A statistical potential is optimized by choosing among the entire set $\mathbb{K}$ or its three subsets to estimate the feature distributions $p(f^c|\mathcal{Q}_{\mathcal{K}})$.

2.3.2 Calculation of feature distributions The sample for computing this distribution is the set of the individual features of type c in protein set $\mathcal{Q}_{\mathcal{K}}$, where each feature is represented by the distribution of this feature - $p(f^{c(m)}|\mathcal{Q}_{\mathcal{K}})$. The feature distribution $p(f^c|\mathcal{Q}_{\mathcal{K}})$ is the average of these sample distributions. For a distance and an angle, $p(f^{c(m)}|Q_K)$ is approximated by a Gaussian distribution $p'(f^{c(m)}|Q_K)$ with the mean equal to the observed value and the standard deviation computed by the propagation (Neuhauser, 2010) of the uncertainties of individual atomic positions, which in turn are estimated from the atomic isotropic temperature factors (Carugo and Argos, 1999; Schneider, 2000; Cruickshank, 1999). For relative atomic surface accessibility $p(f^{c(m)}|Q_K)$ is approximated using a delta function $p'(f^{c(m)}|Q_K)$ centered at feature $f^{c(m)}$ in K. The approximated feature distribution $p'(f^c|\mathcal{Q}_{\mathcal{K}})$ is then computed from the approximated sample distributions $p'(f^{c(m)}|Q_K)$.

2.4 Bayesian smoothing and smoothing priors

The feature distributions $p'(f^c|\mathcal{Q}_{\mathcal{K}})$ can be noisy when the sample $\mathcal{K}$ is relatively small, as is often the case for the orientation between a pair of covalent bonds (Fig. 3A). Thus, we use Bayesian inference to calculate a smooth feature distribution:

$$p\big(p(f^c|\mathcal{Q}_{\mathcal{K}})|p'(f^c|\mathcal{Q}_{\mathcal{K}})\big) \propto p\big(p'(f^c|\mathcal{Q}_{\mathcal{K}})|p(f^c|\mathcal{Q}_{\mathcal{K}})\big)\cdot p\big(p(f^c|\mathcal{Q}_{\mathcal{K}})\big) \quad (3)$$

where $p(f^c|\mathcal{Q}_{\mathcal{K}})$ is the ideal distribution without noise from an infinitely large set of known structures. Both the likelihood $p\big(p'(f^c|\mathcal{Q}_{\mathcal{K}})|p(f^c|\mathcal{Q}_{\mathcal{K}})\big)$ and the prior $\mathcal{S} \equiv p\big(p(f^c|\mathcal{Q}_{\mathcal{K}})\big)$ are multivariate Gaussian distributions (Rasmussen and Williams, 2005). The smoothness of $p(f^c|\mathcal{Q}_{\mathcal{K}})$ is specified by the prior $\mathcal{S}$; here, the prior is a multivariate Gaussian distribution with a zero mean and a squared exponential covariance function (Mackay, 2003). The characteristic length scale of the covariance function defines the range over which the two points are still correlated (the smoothness of the curve). We set the characteristic length equal to a scale parameter L multiplied by 0.2Å for distance, 10° for angles and 0.1% for atomic surface accessibility. A set of smoothing priors $\mathbb{S}$ is obtained by varying L. Using a scale of 2.0 as an example, the inferred $p(f^c|\mathcal{Q}_{\mathcal{K}})$ is significantly smoother than $p'(f^c|\mathcal{Q}_{\mathcal{K}})$ (Fig. 3B).

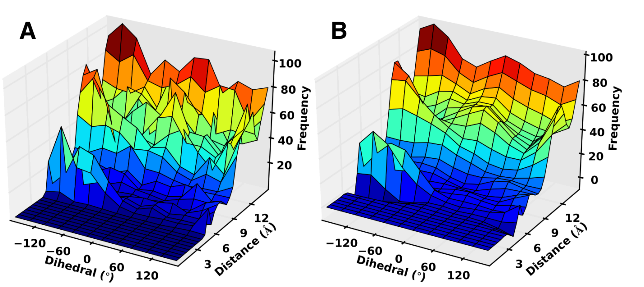

Fig. 3. Distance and dihedral angle joint distribution between alanine N-C_α and alanine O-C, when $\boldsymbol{\alpha_1} \in [\mathbf{60^o}, \mathbf{90^o}]$ and $\boldsymbol{\alpha_2} \in [\mathbf{60^o}, \mathbf{90^o}]$. A) Original distribution. B) Smoothed distribution.

2.5 Decoys and assessment criteria

2.5.1 Learning set for SOAP-PP This set consists of 176 native complex structures in the pairwise protein docking benchmark 4.0 (Hwang *et al.,* 2010) as well as approximately 4,500 decoys for each of the complexes generated using PatchDock (Duhovny *et al.*, 2002).

2.5.2 Testing set for SOAP-PP This set consists of 176 native complex structures in the pairwise protein docking benchmark 4.0 (Hwang *et al.,* 2010) as well as approximately 212,000 decoys for each of the complexes generated using PatchDock (Duhovny *et al.*, 2002) and approximately 54,000 decoys for each of the complexes generated using ZDOCK (Pierce *et al.* 2011).

2.5.3 Assessment criteria for SOAP-PP Each model is assessed for accuracy based on Root-Mean-Square Deviation (RMSD) from the native structure, as used at CAPRI (Lensink *et al*, 2007). A docking model is considered *acceptable* if the ligand C_α RMSD (L-RMSD) after superposition of the receptors is less than 10Å or the interface C_α RMSD (I-RMSD) is less than 4Å. A docking model is of *medium* accuracy if L-RMSD is less than 5Å or I-RMSD is less than 2Å. The success rate for SOAP-PP is the percentage of benchmark cases with at least one medium or acceptable accuracy model in the top N predictions.

2.5.4 Learning set for SOAP-Loop This set consists of 3,838 native loop conformations of 4 to 20 residues as well as approximately 500 decoys for each loop generated using Modeller (Sali and Blundell, 1993; Fiser and Sali, 2003). Loops were extracted from X-ray crystallography structures in the PDB using DSSP (Kabsch and Sander, 1983; Joosten et al., 2011). We only considered protein structures determined at a resolution better than 2Å, R_{free} better than 0.25, and crystallized between pHs 6.5 and 7.5; no pair of source structures had sequence identity higher than 30%. Each loop has only standard residues, no missing non-hydrogen atoms, average atomic surface accessibility between 5% and 60%, no crystal contacts, no clashes with nearby atoms, no contacts with metal ligands, and does not occur in the PLOP loop modeling decoy set (Jacobson et al., 2004).

2.5.5 Testing set for SOAP-Loop This set consists of 833 native loop conformations of 4 to 12 residues as well as approximately 450 decoys for each loop generated using PLOP (Jacobson *et al*, 2004).

2.5.6 Assessment criteria for SOAP-Loop Each model is assessed for accuracy based on its main-chain RMSD to the native conformation, after superposition of all non-loop atoms ($RMSD_{global}$) (Fiser *et al.*, 2000); main-chain atoms include amide nitrogen, C_α, as well as carbonyl carbon and oxygen. SOAP-Loop is assessed by the average $RMSD_{global}$ of the top ranked model for each loop.

2.6 Recovery functions and functional forms

We estimate the recovery function $g(f^c|\mathcal{Q}_{\mathcal{K}})$ by optimizing the accuracy of the corresponding statistical potential on a benchmark of interest. To avoid overfitting, we assume either a single recovery function for all feature types or the same recovery function for a subset of similarly distributed feature types.

The set of recovery function forms $\mathcal{G}_f$ is different for distances, angles, and accessibility: The recovery function for the atomic distance is modeled using one of three functional forms: (1) d^q where d is distance and q is a constant (Zhou and Zhou, 2002); (2) the ideal gas distribution in spheres with varying radii (Shen and Sali, 2006); and (3) spliced cubic splines. For orientation, the recovery function is defined as the product of a recovery function for d, α_1, α_2, and ψ, respectively. The recovery functions for angles α_1, α_2, and dihedral angle ψ are modeled using two different functional forms: (1) the feature distribution calculated using the ideal gas assumption and (2) spliced cubic splines. For the relative atomic surface accessibility, the recovery function form is spliced cubic splines. Control points of cubic splines are defined by their x and y values. When searching for the best cubic spline recovery function, the x values of the control points are either fixed at discrete sampling values or inferred together with the y values.

To optimize the recovery functions, we need to balance minimizing noise and maximizing precision. Thus, for atomic distances, we clustered the distance distributions $p(f^c|\mathcal{Q}_{\mathcal{K}})$ for different atom type pairs using k-mean clustering, and assumed that the pairs of atom types with similar distance distributions have a similar recovery function (Fig. 4).

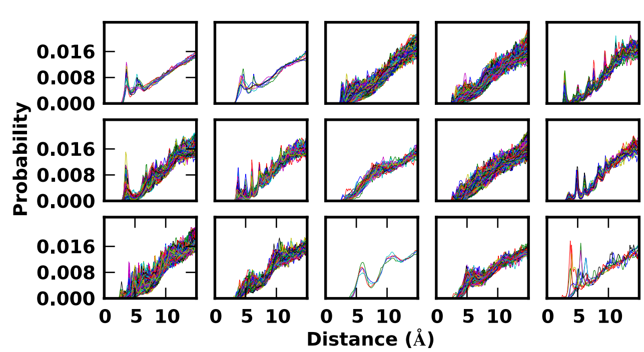

Fig. 4. Distance distributions $p(f^c|\mathcal{Q}_{\mathcal{K}})$ for different atom pairs are clustered into 15 different groups. Each line represents a distance distribution from a pair of atoms of certain types. Each group has 6 to 8401

distributions. During k-mean clustering, the number of clusters was set to 20, resulting in 14 clusters with more than 5 distributions and 6 clusters with less than 5 distributions; the latter 6 clusters are grouped together (bottom right panel).

2.7 Bayesian inference and model selection

A statistical potential is defined by 4 discrete input variables (the known protein structure subset $\mathcal{K}$, the feature type subset $\mathcal{F}$, the smoothing prior $\mathcal{S}$ and the recovery function form $\mathcal{G}_f$) and a vector of continuous input variables (the recovery function parameters $\mathcal{G}_\theta$). We elected to define the best values for the 4 discrete variables are those that result in the most generalizable statistical potential, as judged by the Bayesian predictive densities (Vehtari and Lampinen, 2002), while the best values for the recovery function parameters are those that result in the most accurate statistical potential, as judged by a given benchmark. Because each of the 5 variables can be sampled at many values, enumeration of all combinations is not computationally feasible. Thus, the search for the best values is carried out in four stages, as follows.

First, irrespective of the final restrained feature $\mathcal{F}$, we begin with the atomic distance and a single recovery function for all atom type pairs. The optimal values of the discrete variables $(\mathcal{F}, \mathcal{K}, \mathcal{S}, \mathcal{G}_f)$ are found by an iterative discrete search:

1) Choose an arbitrary starting value for each variable, out of their possible value sets $\{\mathbb{F}, \mathbb{K}, \mathbb{S}, \mathbb{G}_f\}$ (Table S1 and S2).
2) For each variable, choose the best value and eliminate the worst value in the value set using Bayesian model selection based on Bayesian predictive densities (Vehtari and Lampinen, 2002). The Bayesian predictive density for each value is calculated with other variables fixed at their best previous values:

$$\prod_{\{t,v\}} \int p(D_v|\mathcal{F}, \mathcal{K}, \mathcal{S}, \mathcal{G}_f, \mathcal{G}_\theta) \cdot p(\mathcal{G}_\theta|\mathcal{F}, \mathcal{K},, \mathcal{S}, \mathcal{G}_f, D_t)\, d\mathcal{G}_\theta \qquad (4)$$

where the learning decoys D are randomly separated multiple times into a training set D_t and a validation set D_v, from which the integrals are estimated using Monte Carlo sampling (Evans and Swartz, 2000). $p(\mathcal{G}_\theta|\mathcal{F}, \mathcal{K},, \mathcal{S}, \mathcal{G}_f, D_t)$ is calculated following the Bayes rule:

$$p(\mathcal{G}_\theta|\mathcal{F}, \mathcal{K},, \mathcal{S}, \mathcal{G}_f, D_t) \propto p(D_t|\mathcal{F}, \mathcal{K}, \mathcal{S}, \mathcal{G}_f, \mathcal{G}_\theta) \cdot p(\mathcal{G}_\theta|\mathcal{G}_f) \qquad (5)$$

where the likelihood $p(D_t|\mathcal{F}, \mathcal{K}, \mathcal{S}, \mathcal{G}_f, \mathcal{G}_\theta)$ is a Half-Normal distribution whose corresponding normal distribution has the mean equal to the accuracy of an imaginary statistical potential generating scores that correlate perfectly with the decoy-native RMSD and the standard deviation computed by dividing the mean by the number of the cases in the training set D_t; the prior $p(\mathcal{G}_\theta|\mathcal{G}_f)$ is an informative prior defining a reasonable range for $\mathcal{G}_\theta$.
3) Repeat step 2 until the best values do not change.
4) Repeat 5 times steps 1-3 for different random initial values.
5) Keep the best performing variable values.

Second, keeping the optimal values from the previous step fixed, we find the optimal values for the feature type, smoothing length scale, and the number of spline anchor points using the same 5-step iterative discrete search outlined above.

Third, if the optimal spatial feature selected in the previous step is not orientation, we vary the number of recovery functions and the number of anchor points to optimize their values, again using the 5-step iterative discrete search.

Fourth, using the selected $\{\mathcal{F}, \mathcal{K}, \mathcal{S}, \mathcal{G}_\mathrm{f}\}$, we infer the best recovery function parameter values $\mathcal{G}_\theta$ by maximizing $p(\mathcal{G}_\theta|\mathcal{F}, \mathcal{K},, \mathcal{S}, \mathcal{G}_f, D)$ (Eq. 5). The optimized statistical potential is then calculated (Eq. 2), and assessed on testing decoy sets.

SOAP-PP and SOAP-Loop are available as part of MODELLER (http://salilab.org/modeller). All the training, learning, testing, decoys, benchmark sets, and scripts are available at http://salilab.org/SOAP.

3 RESULTS

3.1 Scoring protein-protein interfaces

SOAP-PP is an atomic statistical potential for assessing a binary protein interface, computed with our Bayesian framework by optimizing its accuracy on the learning set for SOAP-PP (Table **Error! Reference source not found.**).

Using the recovery function parameters optimized for 15 sets of training decoys (each set is randomly selected 50% of the learning set), the average top10 success rate (Section 2.5.3) is 51.7%±0.9% on the sets of training decoys and 46.4%±1.7% on the sets of validation decoys. The relatively small difference between the two success rates likely results from overfitting. To investigate overfitting, we increased the size of the training decoy set from 50% to 67% of the entire learning set of 176 proteins. As a result, the average top10 success rate on the training decoys decreased from 51.7% to 51.3%, but the average success rate on the validation decoys (the remaining 33% of the learning set) increased from 46.4% to 47.5%. This observation suggested that increasing the size of the training set may be an effective way of reducing overfitting (Murphy, 2012). Thus, we optimized SOAP-PP using the entire learning set of 176 proteins as the training set, even though this forces subsequent testing on the training protein sequences. To estimate the resulting overfitting, we calculated 6 optimized statistical potentials, each one of which was based on a training set that included a random subset of ~67% of the learning set. Next, we tested these potentials on two testing sets: the first set consisted only of the training proteins; the second set consisted of the remaining learning proteins. The average top10 success rate for the PatchDock decoys is 51.1% and 48.6% for the first and second test set, respectively; for the ZDOCK decoys, the average top10 success rate is 40.0% and 38.9% for the first and second test set, respectively. Therefore, given that increasing the training set size reduces overfitting as shown above, the accuracy of SOAP-PP estimated based on a completely different testing set is expected to be within 2.5% of the current estimate (below).

SOAP-PP was assessed on the PatchDock (Schneidman-Duh*ovny et al*, 2012) and ZDOCK decoy sets (Pierce *et al.*, 2011) (Fig. 5). For PatchDock decoys, the top10 success rate of SOAP-PP is 50% (Fig. 5A) compared to 23% for ZDOCK and 27% for FireDock. If only models of medium or better accuracy are considered, the top10 success rate is 40% for SOAP, 17% for ZDOCK, and 23% for FireDock (Fig. 5B). For ZDOCK decoys, the top10 success rate of SOAP-PP is 41% (Fig. 5C) compared to 30% for ZDOCK and 22% for FireDock. If only models of medium or better accuracy are considered, the success rate is 32% for SOAP-PP, 22% for ZDOCK, and 17% for FireDock (Fig. 5D).

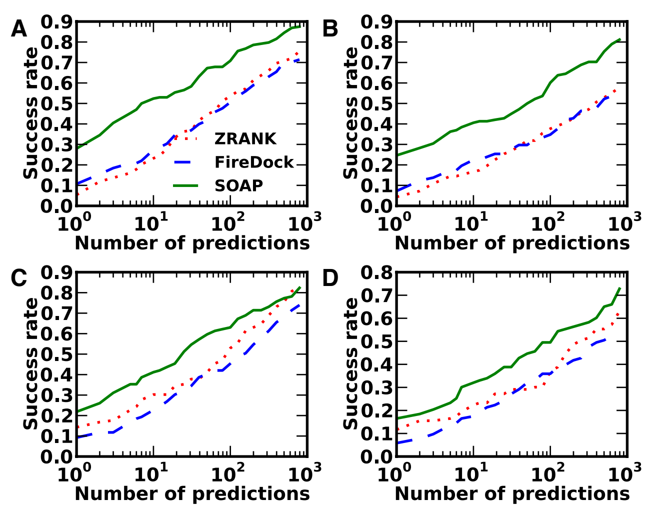

Fig. 5. Success rates of SOAP-PP, ZRANK, and FireDock on the PatchDock and ZDOCK decoy sets. A) Success rates on the PatchDock decoy set, where a success is defined as having an acceptable accuracy structure in the top *N* predictions (x-axis). B) Success rates on the PatchDock decoy set for picking structures with medium accuracy. C) Success rates on the ZDOCK decoy set for picking structures with

acceptable accuracy. D) Success rates on the ZDOCK decoy set for picking structures with medium accuracy.

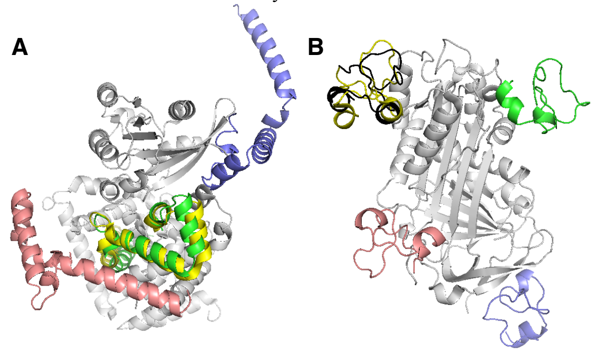

Fig. 6. Comparison of the top ranked, best sampled, and native configurations. A) 2G77. B) 1OC0. The receptor is shown in grey. The ligand is shown in the native configuration (yellow), the best sampled configuration (green for 2G77 and black for 1OC0), and the top ranked configuration by SOAP (green), FireDock (blue), and ZRANK (red).

High accuracy of SOAP-PP can sometimes be attributed to the weaker short-distance repulsion (Fig. 6A) compared to ZRANK (Pierce and Weng, 2007) and FireDock (Andrusier *et al.*, 2007), both of which use a modified van der Waals repulsion term; thus, the clashes of the best sampled structure with a receptor are likely less penalized by SOAP than by ZRANK and FireDock. Although SOAP-PP is more successful than ZRANK and FireDock overall, picking near-native protein-protein complex models out of decoys remains a hard problem (Fig. 5). For some cases, all three scoring functions perform badly, especially when the protein-protein interfaces are small and have poor shape complementarity (Fig. 6B).

3.2 Scoring loops

SOAP-Loop is an atomic statistical potential for assessing protein loop conformations, computed with our Bayesian framework by optimizing its accuracy on the learning set for SOAP-Loop (Table S2).

SOAP-Loop was assessed on the PLOP loop modeling decoy set (Jacobson *et al*, 2004). We compare SOAP-Loop to DOPE (Shen and Sali, 2006), DFIRE (Zhang *et al.*, 2004), Rosetta 3.3 (Simons *et al.*, 1999), and PLOP 25.6 scoring functions (Jacobson *et al.*, 2004) (Fig. 7A). For short loops, SOAP-Loop and Rosetta perform similarly and better than the other tested scoring functions: the main-chain RMSD of SOAP-Loop's top ranked structure is close to that of the best decoy structure. For longer loops, the accuracy differences become larger. SOAP-Loop is still able to pick structures close to the best decoy structures: For 12-residue loops, the average main-chain RMSD of the best scored conformations by SOAP-Loop is 1.5Å, close to the average RMSD of the best decoy conformations (1.2Å) and significantly better than that by DOPE (2.5Å), DFIRE (2.3Å), Rosetta (2.1Å), and PLOP scoring functions (3.0Å). We note that this assessment should not be used to rank the PLOP scoring function, because the decoy set used here was generated with PLOP. Thus, we further compare different scoring functions by their average all-atom RMSD values of the best scored conformations using our learning set for SOAP-Loop (Section 2.5.4; Table S3).

Although no testing protein occurs in the learning set, 11 pairs of testing-learning loops have the same sequence. Excluding these 11 loops from the testing set, the average RMSD of the top ranked loop by SOAP-Loop increases insignificantly from 0.895Å to 0.897Å; the average RMSD of the best decoy conformations also increases insignificantly from 0.566Å to 0.567Å.

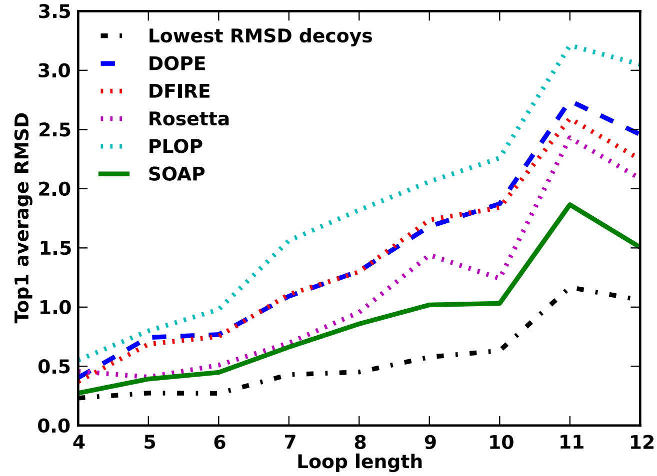

Fig. 7. Accuracy of SOAP-Loop. The average main-chain RMSD of top ranked structures by DOPE, DFIRE, Rosetta, PLOP, and SOAP-Loop on PLOP loop modeling decoys. The average RMSD of the most accurate conformations sampled by PLOP is plotted by a grey dash-dotted line.

The relative success of SOAP is attributed to the scoring of the orientation instead of distance as well as the use of the recovery functions instead of a reference state (Fig. **8**). However, SOAP-Loop still fails to identify the best-sampled conformation in some cases. For a loop in 1CYO, for example, the failure can be attributed to the lack of a sufficiently native conformation among the tested conformations and the absence of significant interactions between the loop and the rest of the protein (**Fig. 9**A). It is also possible that some interactions, such as long-range interactions, are not treated accurately by any scoring function, indicating the need for further development of the theory of statistical potentials.

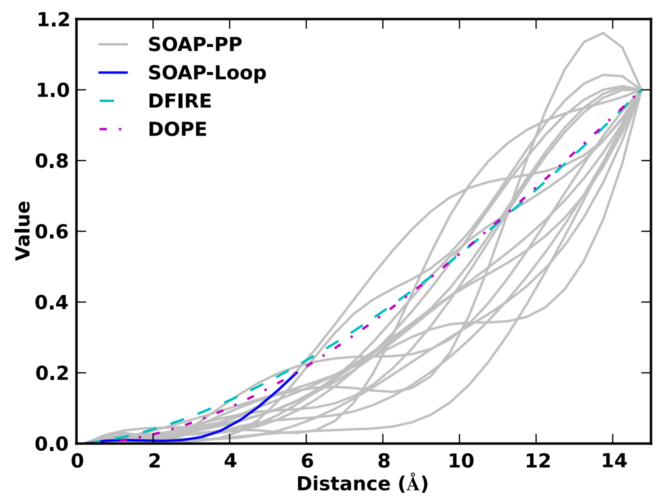

Fig. 8. Recovery functions for SOAP-PP and SOAP-Loop are compared with DOPE and DFIRE's reference states.

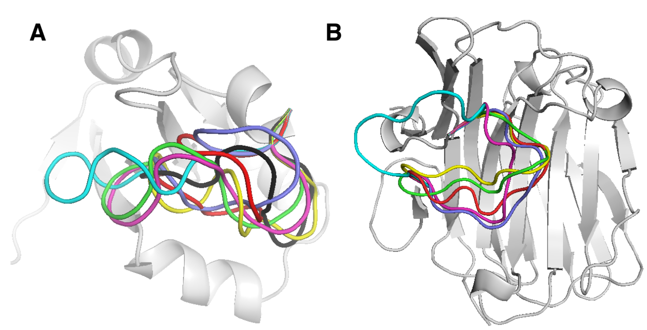

Fig. 9. Comparison of the top ranked, best sampled, and native configurations. A) 1CYO. B) 2AYH. The native structure is shown in light grey. The loop is shown in the native configuration (yellow), the best sampled configuration (black for 1CYO and green for 2AYH), and the top

ranked configuration by SOAP (green), DOPE (blue), DFIRE (red), Rosetta (magenta) and PLOP (light-blue).

4 DISCUSSION

We developed a Bayesian approach to optimizing statistical potentials, based on probability theory and without recourse to questionable statistical mechanical assumptions and approximations. We also applied this approach to calculate optimized statistical potentials for assessing protein interactions (SOAP-PP) and loops (SOAP-Loop). These two statistical potentials perform better than others in their class. For PatchDock and ZDOCK decoys, the top10 success rate of SOAP-PP is more than 10% higher than that of FireDock and ZRANK (Fig. 5). For 12-residue loops in the PLOP benchmark, the average main-chain RMSD of the best scored conformations by SOAP-Loop is 1.5Å, close to the average RMSD of the best sampled conformations (1.2Å) and significantly better than that from DOPE (2.5Å), DFIRE (2.3Å), Rosetta (2.1Å), and PLOP scoring functions (3.0Å) (Fig. 7). The relative accuracy of SOAP-PP and SOAP-Loop results primarily from normalizing the raw distributions by the recovery functions instead of a reference state, restraining of orientation instead of only distance, and thoroughly optimizing parameter values while avoiding over-fitting.

Next, we discuss three points in turn. First, we describe our recovery functions and compare them to the reference states used for other statistical potentials. Second, we discuss the importance of restraining orientation and using covalent separation as an independent variable. Finally, we conclude by commenting on future improvements of our Bayesian approach and its applications.

4.1 Cubic splines as a recovery function form

A key difference between statistical potentials is the definition of their reference states, which are often derived by assuming that the PDB provides a Boltzmann ensemble of structural features (Sippl *et al.*, 1990). Here, we replace the reference state by data-driven recovery functions, defined self-consistently without recourse to these questionable statistical mechanical assumptions (Finkelstein *et al.*, 1995; Shen and Sali, 2006). In an extreme case, we use cubic splines to compute an optimal recovery functions, relying on Bayesian inference to obtain parameter values that result in the most accurate statistical potential given a benchmark.

The use of splines as recovery functions is motivated by a qualitative analysis of the recovery function (Eq. S2). The distribution $p\left(f^{c(m)}|Q_K\right)$ of a single feature $f^{c(m)}$ is the product of the restraint on $f^{c(m)}$ and an integral involving the restraints on Q_K's other features (i.e., the environment restraint). Then, the recovery function $g(f^c|\mathcal{Q}_{\mathcal{K}})$ is the distribution of feature type c in structure set $\mathcal{K}$ resulting from the environmental restraints alone (Eq. S2). We now discuss three implications of this perspective.

First, if we assume that atoms are placed randomly within the protein shell, a recovery function will be similar to the DFIRE and DOPE reference states based on the ideal gas assumption (Zhou and Zhou, 2002; Shen and Sali, 2006).

Second, using the distance d between atoms A and C in Fig. 2 as an example, the environment restraint on d is a consequence of the restraints on distances between A-D, C-B, and B-D as well as the bonds between A-B and C-D. The restraints on A-D, C-B, and B-D distances have short-range repulsion components. Thus, the environment restraint on the distance A-C will include an effective short-range repulsion. This qualitative analysis is consistent with the observed recovery functions for SOAP-PP and SOAP-Loop, which all have lower values at short distances than the DOPE reference state based on the ideal gas assumption (Fig. 8).

Finally, the recovery functions for different feature types can vary, due to their different environments, as observed for the recovery functions for 15 clusters of atom type pairs used in SOAP-PP (Fig. 8).

Although splines can mimic almost any smooth function given a sufficient number of anchor points, its flexibility could also lead to overfitting; moreover, a large number of anchor points could lead to oscillations (Fig. 8). While our Bayesian model selection method helps with the generalizability of the optimized cubic spline (Vehtari and Lampinen, 2002), it is conceivable that applying Bayesian model selection to a less flexible but appropriate functional form will result in a more accurate and general statistical potential than that based on splines.

4.2 Spatial and sequence features

Our orientation restraints score a spatial relationship between two sets of atoms in more detail than distance restraints alone, and should be particularly useful for scoring spatial relationships between polar atoms, especially for hydrogen bond donors and acceptors. In fact, the relative accuracy of SOAP-Loop can be attributed to the use of orientation and recovery functions instead of distance and reference state, respectively (Table S1). However, using orientation did not result in a better statistical potential for ranking protein interfaces (Table S2). While we may not have found the globally optimal statistical potential for orientation, a more likely reason is insufficient accuracy of the tested conformations produced by rigid docking.

Covalent separation is another important factor affecting the accuracy of the derived statistical potentials. Surprisingly, for ranking protein interfaces, statistical potentials derived from intra-chain non-local atom pairs (bond separation > 9) work better than statistical potentials derived from inter-chain atom pairs (chain separation = 1) (Table S1). A likely reason is that many protein interfaces in the PDB result from crystal contacts that do not reflect interfaces between proteins in solution (Krissinel, 2010; Carugo and Argos, 1997). In the future, a better statistical potential for ranking protein interfaces might be obtained if only true biological interfaces from PDB are used.

4.3 Bayesian inference

Statistical potentials can be derived for many different values of the input variables, with little or no *a priori* reasons to choose one set of values over the others. The Bayesian model selection based on Bayesian predictive densities provides a statistically rigorous way of choosing the values that result in most generalizable statistical potentials (Vehtari and Lampinen, 2002). However, one limitation of this method is that the calculation of predictive densities is computational intensive, often requiring more than tens of thousands of evaluations of the statistical potential on the benchmark. Thus, such calculations are not always practical. Fortunately, increases in the available computer power will enable us to find more accurate statistical potentials in an increasingly larger parameter space in the future. Another approach to improving the search for optimal parameter values is to use physically motivated feature types, functional forms, and allowed value ranges.

In principle, normalizing the feature distributions by recovery functions to obtain a statistical potential (Eq. 2) is not necessary. Instead, we could use parametric (*e.g.*, the mathematical functional forms used in molecular mechanics force fields) or non-parametric functions to represent the statistical potential and directly infer the optimal statistical potential by its accuracy on a benchmark of interest. However, this approach might not provide an accurate statistical potential in practice, due to the large number of parameters whose values would need to be optimized.

Our method for smoothing feature distributions is a generalization of the two related methods used in calculating statistical potentials (Sippl, 1990) and homology restraints (Sali and Blundell, 1993). Both methods are equivalent to our Bayesian smoothing method with a diagonal covariance matrix as the smoothing prior. Their prior distribution is equivalent to the mean of our prior $\mathcal{S}$, while the weights on their prior distributions are defined by the standard deviation in our covariance matrix.

In conclusion, our Bayesian framework can be applied to derive an optimized statistical potential for many other kinds of modeling problems for which sample structures are available, thus affording better leverage of the experimentally determined protein structures. Examples include membrane protein topology and complexes of proteins with small molecules or peptides.

ACKNOWLEDGEMENTS

Funding: This work was supported by NIH grants (GM071790 and GM093342; R01 GM054762 to A.S.).

REFERENCES

Andrusier,N. *et al.* (2007) FireDock: fast interaction refinement in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, **69**, 139–159.

Audie,J. and Swanson,J. (2012) Recent work in the development and application of protein-peptide docking. *Future medicinal chemistry*, **4**, 1619–44.

Baker,D. and Sali,A (2001) Protein structure prediction and structural genomics. *Science (New York, N.Y.)*, **294**, 93–6.

Benkert,P. *et al.* (2008) QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*, **71**, 261–77.

Betancourt,M R and Thirumalai,D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein science : a publication of the Protein Society*, **8**, 361–9.

Betancourt,Marcos R and Skolnick,Jeffrey (2004) Local propensities and statistical potentials of backbone dihedral angles in proteins. *Journal of molecular biology*, **342**, 635–49.

Brenke,R. *et al.* (2012) Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics (Oxford, England)*, **28**, 2608–14.

Carugo,O. and Argos,P. (1997) Protein-protein crystal-packing contacts. *Protein science : a publication of the Protein Society*, **6**, 2261–3.

Carugo,O. and Argos,P. (1999) Reliability of atomic displacement parameters in protein crystal structures. *Acta crystallographica. Section D, Biological crystallography*, **55**, 473–8.

Chuang,G.-Y. *et al.* (2008) DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophysical journal*, **95**, 4217–27.

Colovos,C. and Yeates,T.O. (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein science : a publication of the Protein Society*, **2**, 1511–9.

Cossio,P. *et al.* (2012) A simple and efficient statistical potential for scoring ensembles of protein structures. *Scientific Reports*, **2**.

Cruickshank,D.W. (1999) Remarks about protein structure precision. *Acta crystallographica. Section D, Biological crystallography*, **55**, 583–601.

Dehouck,Y. *et al.* (2006) A new generation of statistical potentials for proteins. *Biophysical journal*, **90**, 4010–7.

Dill,K.A. and MacCallum,J.L. (2012) The protein-folding problem, 50 years on. *Science (New York, N.Y.)*, **338**, 1042–6.

Ding,X.-M. *et al.* (2010) Computational prediction of DNA-protein interactions: a review. *Current computer-aided drug design*, **6**, 197–206.

Evans,M. and Swartz,T. (2000) Approximating Integrals Via Monte Carlo and Deterministic Methods Oxford University Press, USA.

Fan,H. *et al.* (2011) Statistical potential for modeling and ranking of protein-ligand interactions. *Journal of chemical information and modeling*, **51**, 3078–92.

Fan,H. *et al.* (2012) Virtual ligand screening against comparative protein structure models. *Methods in molecular biology (Clifton, N.J.)*, **819**, 105–26.

Fan,H. and Sali,Andrej Virtual Ligand Screening Using Comparative Models of Proteins. *Structure*, 1–32.

Fernández-Recio,J. and Sternberg,Michael J. E. (2010) The 4th meeting on the Critical Assessment of Predicted Interaction (CAPRI) held at the Mare Nostrum, Barcelona. *Proteins: Structure, Function, and Bioinformatics*, **78**, 3065–3066.

Ferrada,E. *et al.* (2007) A knowledge-based potential with an accurate description of local interactions improves discrimination between native and near-native protein conformations. *Cell biochemistry and biophysics*, **49**, 111–24.

Finkelstein,A. V *et al.* (1995) Why do protein architectures have Boltzmann-like statistics? *Proteins*, **23**, 142–50.

Fiser,A *et al.* (2000) Modeling of loops in protein structures. *Protein science : a publication of the Protein Society*, **9**, 1753–73.

Gao,M. and Skolnick,Jeffrey (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic acids research*, **36**, 3978–92.

Gatchell,D.W. *et al.* (2000) Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins*, **41**, 518–34.

Hendlich,M. *et al.* (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *Journal of molecular biology*, **216**, 167–80.

Huang,S.-Y. and Zou,X. (2010) Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *Journal of chemical information and modeling*, **50**, 262–73.

Jacobson,M.P. *et al.* (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins*, **55**, 351–67.

Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of molecular biology*, **287**, 797–815.

Keasar,C. and Levitt,Michael (2003) A Novel Approach to Decoy Set Generation: Designing a Physical Energy Function Having Local Minima with Native Structure Characteristics. *Journal of Molecular Biology*, **329**, 159–174.

Kocher,J.P. *et al.* (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *Journal of molecular biology*, **235**, 1598–613.

Kouranov,A. *et al.* (2006) The RCSB PDB information portal for structural genomics. *Nucleic acids research*, **34**, D302–5.

Krissinel,E. (2010) Crystal contacts as nature's docking solutions. *Journal of computational chemistry*, **31**, 133–43.

Lensink,M.F. *et al.* (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins*, **69**, 704–18.

Li,Y. *et al.* (2013) Building a Knowledge-Based Statistical Potential by Capturing High-Order Inter-residue Interactions and its Applications in Protein Secondary Structure Assessment. *Journal of chemical information and modeling*, **53**, 500–8.

Liu,Shiyong and Vakser,I.A. (2011) DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC bioinformatics*, **12**, 280.

Liu,Y. and Gong,H. (2012) Using the unfolded state as the reference state improves the performance of statistical potentials. *Biophysical journal*, **103**, 1950–9.

Lu,H. and Skolnick,J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, **44**, 223–32.

Lu,M. *et al.* (2008) OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of molecular biology*, **376**, 288–301.

Mackay,D.J.C. (2003) Information Theory , Inference , and Learning Algorithms.

Mandell,D.J. *et al.* (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods*, **6**, 551–2.

McConkey,B.J. *et al.* (2003) Discrimination of native protein structures using atom-atom contact scoring. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 3215–20.

Melo,F and Feytmans,E. (1997) Novel knowledge-based mean force potential at atomic level. *Journal of molecular biology*, **267**, 207–22.

Melo,Francisco *et al.* (2002) Statistical potentials for fold assessment. *Protein science : a publication of the Protein Society*, **11**, 430–48.

Miyazawa,S. and Jernigan,R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology*, **256**, 623–44.

Moult,John *et al.* (2011) Critical assessment of methods of protein structure prediction (CASP)--round IX. *Proteins*, **79 Suppl 1**, 1–5.

Murphy,K.P. (2012) Machine Learning: A Probabilistic Perspective.

Neuhauser,C. (2010) Calculus For Biology and Medicine (3rd Edition) (Calculus for Life Sciences Series) Pearson.

Park,B. and Levitt,M (1996) Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *Journal of molecular biology*, **258**, 367–92.

Pierce,B. and Weng,Z. (2007) ZRANK : Reranking Protein Docking Predictions With an Optimized Energy Function. *Bioinformatics*, **1086**, 1078–1086.

Pierce,B.G. *et al.* (2011) Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS one*, **6**, e24657.

Qiu,J. and Elber,R. (2005) Atomically detailed potentials to recognize native and approximate protein structures. *Proteins*, **61**, 44–55.

Rajgaria,R. *et al.* (2008) Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins*, **70**, 950–70.

Rasmussen,C.E. and Williams,C.K.I. (2005) Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series) The MIT Press.

Rata,I. a *et al.* (2010) Backbone statistical potential from local sequence-structure interactions in protein loops. *The journal of physical chemistry. B*, **114**, 1859–69.

Reva,B.A. *et al.* (1997) Residue-residue mean-force potentials for protein structure recognition. *Protein engineering*, **10**, 865–76.

Rojnuckarin,A. and Subramaniam,S. (1999) Knowledge-based interaction potentials for proteins. *Proteins*, **36**, 54–67.

Rykunov,D. and Fiser,Andras (2010) New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC bioinformatics*, **11**, 128.

Sali,A and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, **234**, 779–815.

Samudrala,R and Moult,J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of molecular biology*, **275**, 895–916.

Schneider,T.R. (2000) Objective comparison of protein structures: error-scaled difference distance matrices. *Acta crystallographica. Section D, Biological crystallography*, **56**, 714–21.

Schneidman-Duhovny,D. *et al.* (2012) A Method for Integrative Structure Determination of Protein-Protein Complexes. *Bioinformatics (Oxford, England)*.

Shen,M. and Sali,Andrej (2006) Statistical potential for assessment and prediction of protein structures. 2507–2524.

Simons,K.T. *et al.* (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology*, **268**, 209–25.

Simons,K.T. *et al.* (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82–95.

Sippl,M.J. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *Journal of computer-aided molecular design*, **7**, 473–501.

Sippl,M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of molecular biology*, **213**, 859–83.

Skolnick,Jeffrey *et al.* (2013) Are predicted protein structures of any value for binding site prediction and virtual ligand screening? *Current opinion in structural biology*, **null**.

Summa,C.M. *et al.* (2005) An atomic environment potential for use in protein structure prediction. *Journal of molecular biology*, **352**, 986–1001.

Vehtari,A. and Lampinen,J. (2002) Bayesian model assessment and comparison using cross-validation predictive densities. *Neural computation*, **14**, 2439–68.

Wang,K. *et al.* (2004) Improved protein structure selection using decoy-dependent discriminatory functions. *BMC structural biology*, **4**, 8.

Wass,M.N. *et al.* (2011) Challenges for the prediction of macromolecular interactions. *Current opinion in structural biology*, **21**, 382–90.

Wodak,Shoshana J and Méndez,R. (2004) Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Current opinion in structural biology*, **14**, 242–9.

Zhang,C.H.I. *et al.* (2004) Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Society*, 391–399.

Zhang,J. and Zhang,Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one*, **5**, e15386.

Zhao,F. and Xu,J. (2012) A position-specific distance-dependent statistical potential for protein structure and functional study. *Structure (London, England : 1993)*, **20**, 1118–26.

Zhou,H. and Skolnick,Jeffrey GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction.

Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein science : a publication of the Protein Society*, **11**, 2714–26.

Zhu,J. *et al.* (2008) Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins*, **72**, 1171–88.