Statistical Potentials for Improved Structurally Constrained Evolutionary Models

Claudia L. Kleinman,*[,1] Nicolas Rodrigue,[2] Nicolas Lartillot,[1] and Hervé Philippe[1]

[1]Département de Biochimie, Centre Robert Cedergren, Université de Montréal, Montreal, Quebec, Canada

[2]Department of Biology, Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, Ontario, Canada

***Corresponding author:** E-mail: cl.kleinman@umontreal.ca.

Associate editor: Asger Hobolth

Abstract

Assessing the influence of three-dimensional protein structure on sequence evolution is a difficult task, mainly because of the assumption of independence between sites required by probabilistic phylogenetic methods. Recently, models that include an explicit treatment of protein structure and site interdependencies have been developed: a statistical potential (an energy-like scoring system for sequence–structure compatibility) is used to evaluate the probability of fixation of a given mutation, assuming a coarse-grained protein structure that is constant through evolution. Yet, due to the novelty of these models and the small degree of overlap between the fields of structural and evolutionary biology, only simple representations of protein structure have been used so far. In this work, we present new forms of statistical potentials using a probabilistic framework recently developed for evolutionary studies. Terms related to pairwise distance interactions, torsion angles, solvent accessibility, and flexibility of the residues are included in the potentials, so as to study the effects of the main factors known to influence protein structure. The new potentials, with a more detailed representation of the protein structure, yield a better fit than the previously used scoring functions, with pairwise interactions contributing to more than half of this improvement. In a phylogenetic context, however, the structurally constrained models are still outperformed by some of the available site-independent models in terms of fit, possibly indicating that alternatives to coarse-grained statistical potentials should be explored in order to better model structural constraints.

Key words: protein structure, Bayes factor, statistical potentials, maximum likelihood, molecular evolution.

Research article

Introduction

Protein structure has an undeniable role in shaping the evolution of protein-coding sequences. Not only does the function of a protein depend primarily on the spatial arrangement of its atoms, but proper folding is crucial, since misfolded proteins tend to aggregate and cause unspecific cellular toxicity (Bucciantini et al. 2002; Dobson 2003). As a result, over evolutionary time, protein structure changes much more slowly than the associated sequences (Flores et al. 1993; Russell et al. 1997). Despite this obvious role in evolution, the selective constraints imposed for maintaining a certain fold are still poorly characterized. The relationship between the structural importance of a residue and the purifying selection operating on that site is not straightforward, as several complex mechanisms may act simultaneously to accommodate variation. Natural proteins are more robust to random perturbations than expected by chance (Taverna and Goldstein 2002a, 2002b; Shakhnovich et al. 2005). They can accept substitutions at a large proportion of positions by small movements of interacting sites, or subtle shifts in the main chain conformation of spatially distant residues (Williams and Lovell 2009), in addition to compensatory substitutions. Conversely, structural constraints are just one type of the many selective forces operating on sequences, which include maintaining specific function (such as binding and catalysis), folding kinetics, and regulatory constraints at the DNA and RNA level, to name a few. Disentangling the structural constraints from other constraints, from phylogenetic signal, and from stochastic variation is a problem far from being solved.

One of the main difficulties for modeling evolution with explicit treatment of structural constraints is the site interdependencies that the structure implies, which, for computational reasons, are handled by very few phylogenetic methods. Still assuming site independence, several attempts have been made to include an explicit treatment of protein structure (Overington et al. 1990; Wako and Blundell 1994a, 1994b; Koshi and Goldstein 1995; Goldman et al. 1996; Thorne et al. 1996; Lio et al. 1998; Dimmic et al. 2000). In all the cases, in addition to this important assumption, the evolutionary process is described as acting directly on amino acids, which has the shortcoming of confounding mutation and selection. More sophisticated models have been developed recently at the codon level (see Anisimova and Kosiol 2009; Delport et al. 2009 for a review) that permit the modeling of the interplay of mutation, selection, and drift by making an explicit distinction between mutational and selective parameterizations. Among these, the structurally constrained models are of particular interest in our context. A statistical potential (a scoring system for sequence–structure compatibility) is used to evaluate the probability of fixation of a given mutation, assuming a coarse-grained protein structure that is constant through evolution (Parisi and Echave 2001). Robinson et al. (2003) combined this

representation with statistical tools to make evolutionary inferences dealing with site interdependencies (Jensen and Pedersen 2000; Pedersen and Jensen 2001), establishing a model-based framework for assessing the effect of protein tertiary structure on evolution.

Although adding the structural component to a given evolutionary model produces a substantial improvement in model fit (Rodrigue et al. 2006, 2009; Choi et al. 2007), it is not sufficient to outperform state-of-the-art site-independent models of codon substitution (Rodrigue et al. 2009). The oversimplified structural representation used so far in the sequence–structure compatibility scoring functions may play a central role in this issue. Due to the computational costs of the inference methods, a coarse grain representation of the protein is unavoidable; however, substantial improvement could likely be made regarding the form of the potentials in order to test more complex structural hypotheses.

Knowledge-based potentials that yield reliable scoring functions while restricting the conformational search problem have improved over the last several years (Sippl 1993; Miyazawa and Jernigan 1996; Bastolla et al. 2000; Lazaridis and Karplus 2000; Melo et al. 2002; Buchete et al. 2004; Boas and Harbury 2007). They allow for variable levels of detail in describing the specific amino acid interactions and may account for poorly understood physical phenomena, not exclusively related to protein stability (Boas and Harbury 2007). However, the many potential functions developed in the context of protein structure prediction (where, given a sequence, a search is performed in the space of structures) may not be optimal for our purposes because evolutionary studies pose the problem in terms of a protein design perspective: that is, characterizing the set of sequences compatible with a given structure. The several approaches proposed in this direction are either based on lattice models (Chiu and Goldstein 1998; Seno et al. 1998) or at the atomic level (reviewed in Boas and Harbury 2007). Besides implying heavier computational times, this latter representation has the problem of producing sequences too close to the particular native sequence and implying a level of detail more difficult to reconcile with the assumption of a structure constant through evolution.

To overcome these limitations, we have recently developed a maximum likelihood framework for optimizing the parameters of a coarse grain, residue level statistical potential, tailored for evolutionary studies (Kleinman et al. 2006; Bonnard et al. 2009). A pseudo-energy score $E(s,c)$ is defined as a sum of terms related to different structural descriptors, such as pairwise interactions or solvent accessibility. The probability of observing a database of sequences S, given their native conformations C, and the potential parameters θ, $P(S|C,\theta)$, is then maximized by gradient descent methods to obtain an optimal set of parameters. The method guarantees maximal predictive power for a given potential and provides objective ways to selecting models for otherwise seemingly arbitrary definitions of the potentials.

In previous works (Kleinman et al. 2006; Bonnard et al. 2009; Rodrigue et al. 2009), a simple representation of the protein structure was used, consisting in a contact map supplemented with solvent accessibility information. In the present study, we aimed to model some of the the main protein structural features known to affect amino acid propensity: residue interactions, solvent accessibility, backbone conformation, and flexibility of the residues. Residue interactions were described by replacing the binary contact map we previously used by distance-dependent pairwise interactions, the most widely used representation for fold recognition and protein structure prediction (Jones et al. 1992a; Sippl 1993; Jones 1997; Xia et al. 2000). For describing backbone conformation, we focused on modeling torsion angles (Ramachandran et al. 1963; Kocher et al. 1994; Gilis and Rooman 1997, 2001; Melo et al. 2002; Betancourt and Skolnick 2004) or, alternatively, secondary structure conformation. Protein internal flexibility, in turn, critical for biological functions, such as catalysis, allostery, and interaction with other molecules, is a much more difficult feature to capture. Some information on protein dynamics is contained in the atomic displacement parameters (B-factors) of crystal structures, which reflect the fluctuation of atoms around their average position (Artymiuk et al. 1979; Frauenfelder et al. 1979; Sternberg et al. 1979). We included a term based on B-factors into the potentials to assess the relevance of this measure as a surrogate for flexibility at the residue level. A cross-validation (CV) procedure, implicitly penalizing for model dimensionality, is used to evaluate the alternative combinations of these elements.

We will start by describing the derivation and validation of these new representations of the protein structure. Next, we incorporate them into a structurally constrained codon model of sequence evolution and apply it to three protein data sets. We will discuss the selective constraints associated to these structural elements and assess the performance of the new models against current site-independent models of sequence evolution.

Methods

Statistical Potentials

Definition and Optimization

Knowledge-based potentials are scoring functions that encode statistical patterns present in solved protein structures. They are inductive in nature, based on the idea that the propensity of an amino acid in a given site of a protein can be predicted by the observed frequency of that amino acid at other similar structural contexts in other proteins.

The probabilistic framework that we summarize below was used to optimize the parameters of different forms of statistical potentials by maximum likelihood, using nonredundant subsets of the Protein Data Bank (PDB) for training (Kleinman et al. 2006; Bonnard et al. 2009). Briefly, for a set of P unrelated proteins, each with a single associated structural conformation c^p and an amino acid sequence s^p of length N^p, let s_i^p be the amino acid at position i. Furthermore, assume that a model, M, consists of a set of structural

contexts parameterized by θ and that the observed frequencies of amino acids in each context can be modeled according to the propensity of each amino acid for that context using a Boltzmann distribution. The probability of obtaining a particular sequence is then (Kleinman et al. 2006) as follows:

$$p(s^p|c^p,\theta,M) = \frac{e^{-G(s^p|c^p,\theta)}}{Y^p}, \quad (1)$$

where $Y^p = \sum_{s'} e^{-G(s'|c^p,\theta)}$ is a normalization factor, taken over all possible sequences s' of length N^p, and $G(s^p|c^p,\theta)$ is the statistical potential. Adopting a Bayesian framework, sampling parameters from their posterior distributions induces substantive computational complications, as the model leads to so-called doubly intractable distributions (Rodrigue et al. 2009). Instead, the parameters of the potential (e.g., the contact energy for a given pair of amino acids) are estimated by directly maximizing the joint probability of the database:

$$p(S|C,\theta) = \prod_p p(s^p|c^p,\theta), \quad (2)$$

which can be seen as a likelihood. In practice, a leave-one-out pseudo-likelihood score function (Bonnard et al. 2009) was used in order to decrease the computational time of optimizations (for details, see supplementary appendix S1, Supplementary Material online).

We will now focus on the definition of the statistical potential $G(s,c)$ (for simplicity, we will omit the superscript p in the notation hereafter). It consists of two terms:

$$G(s|c,\theta) = E(s|c,\theta) - F(s|\theta). \quad (3)$$

The term $F(s|\theta)$ accounts for compositional effects, unrelated to the protein conformation. It cannot be solved analytically (Kleinman et al. 2006). Here, we use an approximation inspired from the random energy model (Shakhnovich and Gutin 1993; Sun et al. 1995; Seno et al. 1998) and write:

$$F(s) = \sum_{a=1}^{20} n_a \mu_a, \quad (4)$$

where n_a is the number of occurrences of amino acid a in the sequence s. The unknown parameters μ_a represent the average propensities toward each amino acid and are obtained in the optimization procedure along with all the other parameters.

$E(s|c,\theta)$, in turn, is the energy score. In our previous works (Kleinman et al. 2006; Bonnard et al. 2009; Rodrigue et al. 2009), $E(s|c,\theta)$ consisted of two terms:

$$E(s,c,\theta) = \sum_{i=1}^{N}\sum_{j=i}^{N} \Delta_{ij}\epsilon_{s_i s_j} + \sum_{i=1}^{N} \alpha_{s_i}^{\nu_i}. \quad (5)$$

The first term is a contact energy: $\Delta_{ij} = 1$ if residues i and j are closer in space than a cutoff distance and 0 otherwise, and ϵ_{ab} defines the contact energy between amino acids a and b. The second term encodes a solvent accessibility energy: for each residue, α_a^ν represents the energy of amino acid a in the solvent accessibility class ν, $a = 1,\ldots,20$, and $\nu = 1,\ldots,V$, where V is the total number of solvent accessibility classes considered.

In what follows, alternative definitions of $E(s|c,\theta)$ are explored, encoding different structural descriptors combined in a linear way:

$$E(s,c) = \lambda_1 E_{\text{Bfactor}}(s,c) + \lambda_2 E_{\text{torsion}}(s,c) + \lambda_3 E_{\text{solv}}(s,c) + \lambda_4 E_{\text{dist}}(s,c) + \lambda_5 E_{\text{ss}}(s,c), \quad (6)$$

where λ_i equals either 0 or 1, depending on whether the term is included or not in the potential under study. Although this linear formulation formally assumes independence between the terms, interactions between these elements do exist during the optimization, so that the parameters must be jointly optimized for each alternative functional form.

Several elements have to be determined a priori, such as the division of the parameter space into discrete classes, thus constituting a part of the model being assessed. The choice between alternative definitions was made based on model fit, measured by CV (see below). Given the computational burden needed to incorporate site interdependencies into evolutionary models, there is a compromise to be considered in some cases, between the accuracy of the structural description and the computational cost of $E(s|c,\theta)$.

Model Comparison and Nomenclature

Alternative definitions of the structural elements considered yield different potentials, which can be interpreted as different models, and evaluated by standard statistical tools of model assessment. Here, once an optimal value of θ is obtained for each potential, the fit of alternative models is assessed by CV, consisting in training the potential on one data set and calculating the log-likelihood score on a different independent data set. More precisely, for each model M,

$$\text{CV}_M = -\ln p(S_T|C_T,\theta_L,M), \quad (7)$$

where S_T and C_T are the sequences and structures of the test set, and θ_L are the parameters optimized on the learning set. The difference with the CV score obtained for a flat potential (μ, only accounting for compositional effects without any structural terms, i.e., $E(s|c) = 0$), normalized by the number of sites on the testing set N_T, is reported:

$$\Delta\text{CV} = \frac{\text{CV}_\mu - \text{CV}_M}{N_T}. \quad (8)$$

We call the potentials obtained by the maximum likelihood framework ML potentials and use the following abbreviations to refer to the structural terms included: dist, distance interactions; cont, contacts; solv, solvent accessibility; Bfactor, flexibility, measured by B-factors; torsion, main chain torsion angles; and ss, secondary structure.

Main Chain Torsion Angles

Backbone conformation can be described by the angle of rotation around the bonds of the main chain atoms, called the torsion angles omega, phi, and psi. To capture the different

conformation tendencies that different amino acids exhibit, we focused on modeling propensities for these angles. Torsion classes for angles phi and psi were defined based on a previously described version of the Ramachandran plot, which is divided into nine discrete classes (Laskowski et al. 1996, supplementary fig. S1, Supplementary Material online). For omega angles, on the other hand, two conformations were considered: *cis* or *trans*.

In this way, the conformation c of the protein includes the observed torsion class vectors T and W. The vector $T = (t_i)$ is the conformation of angles phi and psi associated with each site i, $t_i = 1, \ldots, 9$ and $i = 1, \ldots, N$. The vector $W = (w_i)$, in turn, is the conformation of the angle omega at site i, with w_i being either *cis* or *trans*. The pseudo-energy associated with the three torsion angles has the following form:

$$E_{\text{torsion}}(s,c) = \sum_{i=1}^{N} \tau_{s_i}^{t_i} + \sum_{i=1}^{N} \eta_{s_i}^{w_i}, \quad (9)$$

where τ_a^t is the potential energy of amino acid a with angles phi and psi in conformation t, and η_a^w represents the potential energy for amino acid a with the omega angle in conformation w.

Secondary Structure

As an alternative way of describing local structure, we derived a secondary structure potential:

$$E_{\text{ss}}(s,c) = \sum_{i=1}^{N} \varsigma_{s_i}^{l_i}, \quad (10)$$

where ς_a^l is the energy parameter for amino acid a associated with the secondary structure element l. Secondary structure calculations were performed according to the method of Kabsch and Sanders (1983; Laskowski et al. 1993). The ten elements considered are the following: residue in isolated beta-bridge, extended strand, 3/10 helix, alpha-helix, pi-helix, bend, hydrogen-bonded turn, extension of beta strand, extension of 3/10 helix, and extension of alpha-helix. Alternatively, a simplified definition consisting of only three classes was also tested: alpha-helix, beta strand, and turn.

Flexibility of the Residues

In order to capture some information about flexibility at the residue level, we implemented a potential based on the B-factor value at each site. B-factors were calculated either using alpha-carbons or the average for all the atoms of the residue. Because the experimentally determined B-factor depends on elements such as the overall resolution of the structure, crystal contacts, and on the particular refinement procedures, B-factors from different structures need to be normalized before any comparison. We applied the following normalization:

$$B_i^{\text{norm}} = \frac{B_i - \langle B \rangle}{\sigma_B}, \quad (11)$$

where B_i is the B-factor recorded for residue i. σ_B and $\langle B \rangle$, in turn, are the standard deviation and the mean of B-factors for the given structure.

The energy score associated with B-factors has the form

$$E_{\text{Bfactor}}(s,c) = \sum_{i=1}^{N} \gamma_{s_i}^{g_i}, \quad (12)$$

where γ_a^g represents the potential energy for amino acid a in the B-factor class $g, g = 1, \ldots, G$. To determine the number of classes, G, several potentials were optimized with an increasing number of classes (from 0 to 50) and their fit was assessed by CV. The classes were defined so as to generate G number of equal-sized subsets of amino acids (i.e., G quantiles) when analyzing 1,000 randomly drawn proteins from the PDB.

Solvent Accessibility

Solvent accessibility calculations were performed as described in Kleinman et al. (2006): the accessible surface of a residue is defined as the atomic accessible area when a probe of the radius of a molecule of water is rolled around the Van der Waal's surface of the protein. We used the program Naccess (Hubbard and Thornton 1993) to perform this calculation using the percentage relative to the accessibility in Ala-X-Ala fully extended tripeptide. When using PDB files with multiple chains, solvent accessibility was calculated taking into account all molecules in the structure. The optimal number of classes (in this case, eq. (14)) was determined by deriving potentials with an increasing number of classes and evaluating their fit (Kleinman et al. 2006). We made the assumption that this optimal number of classes does not change when combining different structural terms and verified that this was the case for the final form combining all the terms (data not shown).

Distance-dependent Interactions

The distance potential we implemented represents the separation of a pair of residues (in three-dimensional space) as a discrete variable. An interval $R = [r_{\min}, r_{\max}]$ is defined, where $r_{\min}$ and $r_{\max}$ are, respectively, the minimum and maximum distance between two residues for considering an interaction. The interval is divided into D subintervals (also referred as classes) $r_d = [r_{\min}^d, r_{\max}^d)$, $d = 1, \ldots, D$, such that $r_{\min}^1 = r_{\min}$, $r_{\max}^D = r_{\max}$, and $r_{\max}^{d-1} = r_{\min}^d$.

The distance x_{ij} between a pair of residues i and j is measured using either alpha-carbons, beta-carbons, or the mass centers of the two side chains. The energy term based on this distance has the form

$$E_{\text{dist}}(s,c) = \sum_{i=1}^{N} \sum_{j=i}^{N} \epsilon_{s_i s_j}^{r_{ij}}, \quad (13)$$

where r_{ij} is the distance class such that $x_{ij} \in r_{ij}$, and $\epsilon_{ab}^{r_{ij}}$ defines the interaction energy between amino acids a and b in the distance class r_{ij}.

In order to define the intervals, that is, to specify D and the values of the different thresholds $r^d_{\min}$ and $r^d_{\max}$, a preliminary analysis of the distribution of interactions between pairs of amino acids on 1,000 randomly drawn PDB structures was performed. The region $R = [0\text{Å}, 25\text{Å}]$ was partitioned into equal subintervals of 0.25 Å. Let $f_r(a,b)$ be the frequency of observed interactions between amino acids a and b in the subinterval r, considered symmetrical, that is, $f_r(a,b) = f_r(b,a)$. Let $f_R(a,b)$, on the other hand, be the frequency of interactions for the whole region 0–25 Å. To compare these two distributions, the Kullback–Leibler divergence (KLD) was used:

$$\mathrm{KLD}(f_r, f_R) = \sum_{a=1}^{20}\sum_{b=1}^{20} f_r(a,b)\log\frac{f_r(a,b)}{f_R(a,b)}. \qquad (14)$$

Note that KLD is always positive, and KLD $= 0$ when $f_r(a,b) = f_R(a,b)$.

Sequence Sampling: Site-specific Profiles

Sequences compatible with a given conformation, induced by each one of the potentials, are obtained by Gibbs sampling as described in Kleinman et al. (2006) and displayed graphically as sequence logos. Profiles of natural sequences were generated from multiple sequence alignments obtained from the Consurf-HSSP database (Glaser et al. 2005). Alternatively, sequences were realigned using two programs, with default settings: MUSCLE (Edgar 2004) and FSA (Bradley et al. 2009), producing essentially the same results. All the alignments are available as supplementary material (Supplementary Material online).

Phylogenetic Methods

Evolutionary Model

Evolution of codon sequences is modeled as a Markov process defined in sequence space, fully determined by the matrix of instantaneous rates of change from one sequence (s) to another (s'). Mutation and selection are described as two separate processes by the use of distinct sets of parameters. Following Robinson et al. (2003), selective constraints acting at the phenotype level are modeled by the statistical potential: the influence of the protein structure (a single conformation assumed constant along the entire tree) is represented by the difference in potential energy ΔG, with a parameter $\beta > 0$ modulating the strength of this influence. The parameters of $G(s|c)$ are fixed to the values obtained in the optimization by maximum likelihood described in previous sections. The model also includes an additional parameter ω, modulating nonsynonymous rates without regard to the amino acids involved.

The mutational specification, in turn, consists of two sets of parameters: $\rho = (\rho_{lm})_{1\leqslant l,m\leqslant 4}$ is a set of symmetrical nucleotide exchangeability parameters, with $\sum_{1\leqslant l<m\leqslant 4}\rho_{lm} = 1$, and $\varphi = (\phi_m)_{1\leqslant m\leqslant 4}$ represents a set of global nucleotide equilibrium propensities, where $\sum_{1\leqslant m\leqslant 4}\varphi_m = 1$.

In the complete model considered here, an off-diagonal entry of the Markov generator, corresponding to the instantaneous rate of substitution from s to s', is given by

$$R_{ss'} = \begin{cases} \varrho_{s_{i_c}s'_{i_c}}\varphi_{s'_{i_c}}, & \text{if } \mathcal{A}, \\ \omega\varrho_{s_{i_c}s'_{i_c}}\varphi_{s'_{i_c}}\,\mathrm{e}^{-\beta(G(s')-G(s))}, & \text{if } \mathcal{B}, \\ 0, & \text{otherwise,} \end{cases} \qquad (15)$$

where

$\mathcal{A}$: s and s' differ only at the c^{th} codon position of the i^{th} site and imply a synonymous change; $\mathcal{B}$: s and s' differ only at the c^{th} codon position of the i^{th} site and imply a nonsynonymous change; and where s_{i_c} is the nucleotide at the c^{th} codon position of the ith site of sequence s. Diagonal entries are given by the negative sum of off-diagonal entries in a given row. Note that when $\beta = 0$, the model is similar to the type of codon substitution model proposed by Muse and Gaut (1994).

As described in Rodrigue et al. (2009), the substitution process has a stationary probability given by

$$p(s^o|\theta, M) = \frac{1}{Z}\mathrm{e}^{-2\beta G(s^o)}\prod_{i=1}^{N}\left(\prod_{c=1}^{3}\varphi_{s^o_{i_c}}\right), \qquad (16)$$

where Z is the normalizing factor:

$$Z = \sum_{s}\mathrm{e}^{-2\beta G(s)}\prod_{i=1}^{N}\left(\prod_{c=1}^{3}\varphi_{s_{i_c}}\right), \qquad (17)$$

with the sum being over all 61^N possible sequences.

We used the same priors and nomenclature as described in Rodrigue et al. (2009). We refer to the simplest model based on the mutational parameters only as MG because it is inspired by Muse and Gaut (1994) and write MG-NS to refer to the model with a global nonsynonymous rate factor ω. When using the structurally constrained model based on the statistical potentials, we add the suffix SC, giving MG-SC and MG-NS-SC. Finally, in the model referred as MG-NS$^{\text{DP}}$, heterogeneity among sites is introduced by using a Dirichlet process as the law of the ω_i across sites (Huelsenbeck et al. 2006).

Bayes Factors

Computational tools have been recently developed for sampling parameters from their posterior distribution under site-interdependent codon models and for the estimation of Bayes factors (Rodrigue et al. 2009):

$$B_M = \frac{p(D|c, M)}{p(D|c, M_{\text{ref}})}, \qquad (18)$$

where D represents the data, that is, an alignment of nucleotide sequences related by a phylogenetic tree with a known topology, M is the sequence evolution model being evaluated, and M_{ref} represents the site-independent model used as a reference (in the present case, MG).

Bayes factors are computed using thermodynamic integration or "path sampling," as described in Rodrigue et al. (2009). In the case of the SC models, the procedure consists in sampling parameters using Markov chain Monte Carlo

along a continuous path between M and M_{ref}, through a set of slight changes in the value of β. The result is a curve that represents a numerical evaluation of the fit of the model B_M as a function of β, the factor modulating the strength of the structural term in the evolutionary model (eq. 15). The computations are made in duplicate, with different model-switch orientations, that is, tracing the path from M to M_{ref}, and vice versa, and we display both values obtained from these procedures.

Note that the evolutionary model proposed here imposes the same protein structure (c) to all the sequences in the data set and that the particular native sequence corresponding to this structure (which we call s^c) is present in the alignment. In order to avoid the possible biases introduced by this presence, we can further decompose the marginal likelihood into two factors: one corresponding to the probability of the sequence state s^c and another corresponding to the probability of observing all the other sequences (D^{ϕ}), conditional on s^c:

$$p(D|c,M) = p(D^{\phi}|s^c,c,M)p(s^c|c,M). \quad (19)$$

We then write

$$B_M = \frac{p(D^{\phi}|s^c,c,M)}{p(D^{\phi}|s^c,c,M_{ref})}\frac{p(s^c|c,M)}{p(s^c|c,M_{ref})} \quad (20)$$
$$= \left(B_M^{\phi}\right)\left(B_M^{s^c}\right).$$

Formulated in this way, we are interested in distinct evaluations of two factors:

$$B_M^{\phi} = \frac{p(D^{\phi}|s^c,c,M)}{p(D^{\phi}|s^c,c,M_{ref})} \quad (21)$$

and

$$B_M^{s^c} = \frac{p(s^c|c,M)}{p(s^c|c,M_{ref})}. \quad (22)$$

Given the reversibility of the overall substitution model, the factoring is arbitrary but can be used to contrast contributions to model fit, with, for instance, different leaf sequences taken for stationary probability factors.

The stationary probability factor, given in equation (16), can be computed for any leaf of the tree (Rodrigue et al. 2005), and, in particular, for s^c , making the calculation of the transient factor B_M^{ϕ} straightforward.

Data Sets

Learning Databases

We used proteins culled from the entire PDB according to sequence divergence in order to ensure independence (less than 25% mutual sequence identity) and to structure quality (resolution better than 2.0) (Wang and Dunbrack 2003). After discarding very small chains—less than 90 residues—subsets of 500 randomly drawn proteins were assembled. All data sets are available as supplementary material (Supplementary Material online).

Phylogenetic Data Sets

Three data sets were used. The first, taken from Yang et al. (2000), consists of 17 vertebrate nucleotide sequences of the β-globin gene (144 codons). Structural information was extracted from the PDB file 4HHB. The second one, also from Yang et al. (2000), consists of sequences of the alcohol dehydrogenase (ADH) taken from 23 species of Drosophila (254 codons) and the associated PDB file 1A4U. For both these data sets, we worked under the tree topology used by Yang et al. (2000). The third set consists of 34 calmodulin eukaryotic sequences, with a protein structure defined by the PDB file 1CFD and a tree topology estimated using phyML (Guindon and Gascuel 2003) under the model JTT + F +Γ (Jones et al. 1992b; Yang 1993). All data sets are available as supplementary materials (Supplementary Material online).

Table 1. Summary of Class Definitions Used for the Various Elements of the Optimized Potentials.

Potential	Definition
$ML_{Bfactor}$: B-Factor	Average for all the atoms in a residue
	Normalized within each protein
	Five equal-sized classes
$ML_{torsion}$: Torsion angles	ϕ,ψ: 9 classes: - A a B b L l p X (supplementary fig. S1, Supplementary Material online)
	ω: *cis trans*
ML_{solv}: Solvent accessibility	14 equal-sized classes (Kleinman et al. 2006)
ML_{dist}: Distance	Interaction center: side chain center
	Range considered: 3–11 Å
	Resolution: 3–7 Å, interval: 0.5 Å
	7–10 Å, interval: 1 Å
	13 classes
ML_{cont}: Contact	Interaction center: side chain center
	Cutoff distance: 6.5 Å
ML_{ss}: Secondary structure	10 classes (see Methods)

Results and Discussion

Definition of Statistical Potentials and Refinement of Structural Descriptors

The probabilistic framework described above was used to optimize the parameters of several forms of statistical potentials based on different structural descriptors. These can be grouped in two types: pairwise interaction descriptors (contact map or distance-based matrix) and a series of site-independent components such as solvent accessibility, torsion angles, secondary structure, and flexibility of the residues (table 1). As described in the Methods, the refinement of the structural descriptors is done by optimizing the alternative potentials and comparing their model fit in CV experiments. We first analyze the site-specific terms, followed by the more complex site-interdependent descriptors.

Site-independent Descriptors

Aiming to capture flexibility at the residue level, we implemented a potential based on B-factor information. This measure was recorded either for the alpha-carbon or as the average for the whole residue and normalized within each protein. A preliminary analysis on a large number of crystal structures shows that the distribution of B-factors is not

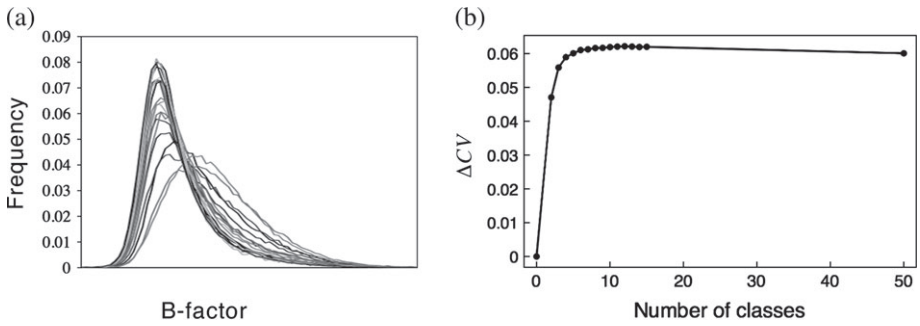

FIG. 1. (a) Distribution of B-factor for the different amino acids in a nonredundant subset of PDB of 1,000 proteins. B-factor was calculated averaging B-factors of all the atoms in the residue and normalized within each protein. (b) Evolution of CV score of the potential as a function of the number of classes.

identical for the different amino acids (fig. 1*a* and supplementary fig. S2*a*, Supplementary Material online), indicating that this is likely an informative element. Moreover, the B-factors of particular regions in proteins seem to be conserved in protein families (Maguid et al. 2006), suggesting that this measure correlates with a biological property. In order to define discrete categories for this feature, we analyzed the evolution of model fit as a function of the number of classes (fig. 1*b*). When the number of classes increases, the fit of the model improves, until the penalization for model dimensionality starts to dominate the score. Not surprisingly, averaging the B-factor for all the atoms in the residue produced a markedly improved model fit compared with the alpha-carbon representation (more than twice the CV score, fig. 1*b* and supplementary fig. S2*a*, Supplementary Material online).

Backbone conformation, in turn, was described using either torsion angles or secondary structure. These two descriptions of the local conformation should in principle be redundant, with dihedral angles encoding richer information than the secondary structure, because they completely specify the position of the backbone. This is indeed reflected in our results. First, the torsion angle potential alone, $ML_{torsion}$, fits the data better (fig. 2). Second, the contribution of the secondary structure term is less important for the combined potential $ML_{torsion,ss}$ (27% improvement with respect to $ML_{torsion}$ in contrast to the 55% expected if the terms were independent) (supplementary table S1, Supplementary Material online). This reflects an important redundancy on the encoded information: for independent terms, one would expect approximately additive contributions to the fit of a combined model; conversely, completely correlated terms would produce a decrease in model fit when combined due to the penalization for model dimensionality. Considering different definitions of secondary structure (see Methods) produced only minor changes in the results (supplementary table S1, Supplementary Material online).

Of all the site-independent descriptors, the solvent potential, based on a discrete measure of the solvent accessible surface for each site, is the term producing the highest value of CV score. The optimal definition of this element was determined previously (Kleinman et al. 2006), in a similar way to the other terms described here, by optimizing the alternative potentials and evaluating their fit. The good performance of this potential is not surprising, given the importance of hydrophobic interactions for stability and folding.

Pairwise Interaction Descriptors

The critical elements defining a distance-based potential are the choice of interacting centers, the range of distances considered, and the clustering of distance into discrete classes. In order to define these elements, we first performed an analysis of the distribution of pairwise interactions in known protein structures. Three interaction center definitions were successively considered: alpha-carbon,

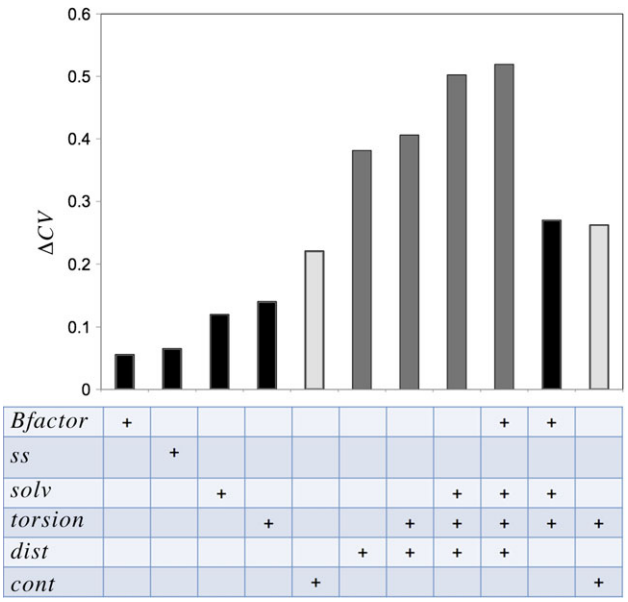

FIG. 2. CV scores for some of the different potentials obtained. The average gain (relative to the CV score obtained with a flat potential, see Methods) for the 2-fold CV experiment is reported. Black bars: site-independent potentials. Dark grey bars: potentials containing distance-based terms. Light grey bars: potentials containing contact terms. The potentials were named according to the structural terms included in the definition: Bfactor, flexibility; ss, secondary structure; torsion, torsion angles; solv, solvent accessibility; cont, contact interactions; and dist, distance interactions.

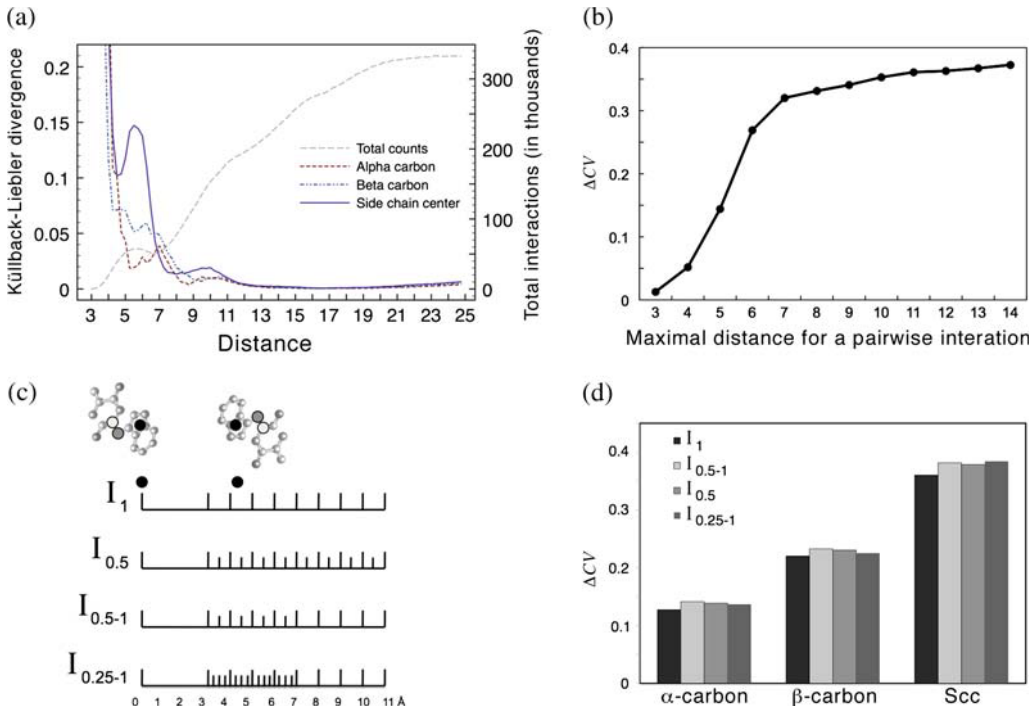

FIG. 3. Distance-based pairwise interactions. (a) The interval 0–25 Åwas divided in windows of 0.25 Å, and the distribution of observed pairwise interactions in each window was compared with the average distribution in the whole region 0–25 Å using the KLD. The total number of interactions, using side chain centers, for each window is shown (dashed grey line). (b) CV score of distance-based potentials as a function of the distance range considered using side chain centers. Distance intervals were partitioned in bins of 1 Å. (c) Graphical representation of the distance classes used in (d). The three interaction centers studied are marked with colored circles: black for side chain center, grey for beta-carbon, and white for alpha-carbon. Windows were defined as follows. The range 3–11 Å was divided in windows of 1 Å(named I_1; 9 classes) or 0.5 Å (named $I_{0.5}$; 17 classes). Alternatively, the resolution was increased only for the interval 3–7 Å, which was divided in windows of 0.5 Å ($I_{0.5-1}$; 13 classes) or in windows of 0.25 Å ($I_{0.25-1}$; 21 classes). (d) CV scores of distance-based potentials as a function of the resolution and the interaction center used.

beta-carbon, and the center of mass of side chains. Ideally, in order to maximize the discriminatory power of the potential, distance classes should be defined in such a way that the distribution of interactions for each class is sufficiently different from the average distribution. In order to spot the areas where these distributions are distinctive, we partitioned the interval 0–25 Å into small windows of 0.25 Å and compared the 210 frequency vector of observed pairwise interactions in each window to the average distribution of interactions in the whole range, using the KLD (fig. 3*a*), for 1,000 randomly drawn PDB structures. Note that this is not meant as an optimization procedure but as an heuristic method.

First, note the similarities in the overall shape of the plot for the three interaction centers studied. Windows corresponding to the shortest distances show the highest values of KLD, mainly due to sparse data and not because of a high amount of information in these regions. There is a peak at midrange distances (around 6–7 Å) and a small shoulder at longer distances (around 9–10 Å). Not surprisingly, the value of KLD (which can be interpreted as the amount of relevant information) at these peaks correlates well with the level of detail of the corresponding structural representation. For the alpha-carbon representation, which encodes only information regarding the main chain, KLD is the lowest of the three. Using beta-carbon incorporates, more information about the orientation of the side chains, and consequently, KLD slightly increases. Finally, the highest peaks are found when using side chain centers for defining interactions.

Next, the KLD plot suggests an upper bound for the distances being considered: beyond 12 Å, the distribution in each bin is indistinguishable from the general distribution, until around 21 Å, where the distributions slowly start to diverge again. Not only is this divergence subtle but also including this region would imply an important increase in the cost of the calculation of $E(s,c)$, which is proportional to the number of contacts, approximately scaling with the volume of the sphere considered. It is known that for long distances, interactions are not residue specific and are determined simply by solvation effects and the geometry of the molecule (Jones et al. 1992a), factors that will probably be modeled by other terms of the potential. Given the computational cost of incorporating site interdependencies into evolutionary models, we have a special interest in finding a range with few contacts considered while remaining sufficiently accurate.

To further confirm the effect of the cutoff distance on the resulting potential, we derived several potentials by only varying their range and dividing the resulting interval in bins of 1 Å. The number of classes thus varies in each case, but the resolution and the interaction center are kept constant.

The results obtained using side chain centers are shown in figure 3*b*. The CV score increases markedly when including distances corresponding to the high peak in the KLD plot (6–7 Å). Adding the small peak at 9–10 Å, however, has only a minor effect, indicative of some redundancies in these areas. A cutoff value of 11 Å was used for subsequent analysis: increasing the range beyond such value does not produce a major improvement in the potential performance but has the negative effect of drastically increasing the computational cost to calculate the energy.

Finally, we analyzed the effect of the resolution on the performance of the potentials. A scheme of the bins used is shown in figure 3*c*. The region 0–11 Å was considered. The interval 0–3 was not subdivided, given the small number of interactions it contains. The interval 3–11 Å, in turn, was divided in bins of 1 Å(named I_1) or 0.5 Å($I_{0.5}$). Alternatively, the resolution was increased only for the interval 3–7 Å, divided in bins of 0.5 Å($I_{0.5-1}$) or 0.25 Å($I_{0.25-1}$). Increasing the resolution in the short-distance interval ($I_{0.5-1}$) produces a better fit for all the interaction centers considered (fig. 3*d*). For the potentials that use alpha-carbons or beta-carbons to describe an interaction, this is the optimal resolution obtained. This is not unexpected: potentials using a coarser description of proteins require a lower resolution for optimal performance because overparameterization penalties appear sooner. For all the interaction centers, increasing the resolution in the longer distance interval (7–11 Å, $I_{0.5}$) was also detrimental (with respect to $I_{0.5-1}$, fig. 3*d*, probably due to overparameterization.

In principle, distance classes should be defined by maximizing differences not only with the general distribution of interactions, as we checked before, but also between different classes. We thus tested alternative discrete versions of the interval, not in a linear way, but based on the pairwise comparison of the KLD for all the different bins (supplementary fig. S3, Supplementary Material online). The performance of the potentials defined in this way was similar to the linear definition, suggesting that for this level of structural representation, the resolution is already nearly optimal. No further work was thus done in this direction.

Combining the Potentials

Figure 2 shows the CV scores for the potentials resulting from a linear combination of the terms described so far (table 1). As discussed before, the linear formulation of the combined potential $E(s,c)$ does not imply independence between the terms. Rather, it allows one to test for potential redundancies in the encoded information by checking whether combined model configurations lead to interactions in terms of model fit.

It is worth noting that when considering the potentials separately, the main improvement in model fit is brought about by the distance-based potential. It adds a considerable amount of information to the combination of all the site-independent descriptors and performs better than the contact potential, solvent accessibility, or the combination of both that has been previously used (Kleinman et al. 2006; Rodrigue et al. 2009).

Solvent and pairwise interaction terms are highly correlated, and so the combined potential $ML_{dist,solv}$ has a score merely 5% higher than the distance-based potential ML_{dist} (fig. 2). On the other hand, this score is almost three times higher than the solvent potential alone ML_{solv}, suggesting that most of the information contained in the combined potential comes from the description of pairwise interactions.

Torsion angles, on the other hand, seem to encode orthogonal information to these two terms (fig. 2 and supplementary table S1, Supplementary Material online). This is consistent with the interpretation that they contain implicit information on the local conformation, independent of amino acid interactions, either with other residues or with the solvent.

As for the flexibility information encoded in the B-factor potential, although its inclusion produces a better fit than using a flat potential, this improvement is diluted when combining all the terms (fig. 2 and supplementary table S1, Supplementary Material online). The most plausible cause is a redundancy in the information encoded by the solvent accessibility and the flexibility terms; it is well known that residues in the core of proteins show less flexibility than those located on the surface, and the two measures are somewhat correlated (supplementary fig. S2*c*, Supplementary Material online). A similar behavior is observed for the secondary structure terms; the redundancies in this case are found with the torsion terms (as discussed above) and to a lesser degree with distance and B-factor terms (supplementary table S1, Supplementary Material online).

The aim of this study being to incorporate the main factors affecting the protein structure, we restricted the analysis to a handful of terms whose importance is well established in the structural biology field. The model comparison and analysis of redundancies performed here, on the other hand, is general enough to be easily extended to other structural terms or to terms not explicitly related to structural considerations.

Comparison of Natural and Designed Sequences

Once the parameters of the potentials are optimized, we can perform an analysis in a protein design perspective by generating sequences from $p(s|c,\theta,M)$ by Gibbs sampling (Kleinman et al. 2006). The graphical display of these sampled sequences allows for a qualitative analysis of the properties induced by the different potentials. An illustrative example is shown in figure 4, where the sampled sequences for a thioredoxin protein are contrasted to naturally occurring sequences.

Note that the comparison performed here is not meant as a rigorous test of the performance of the potentials. Designed and naturally occurring sequences are conceptually different: although the former are free to explore the whole space of sequences compatible with the structure, the latter are constrained by their underlying phylogenetic structure. Moreover, because the evolutionary relationship among the sequences is not accounted for when constructing sequence logos, the conservation observed in the natural profile is somewhat distorted by phylogenetic redundancy. Finally,

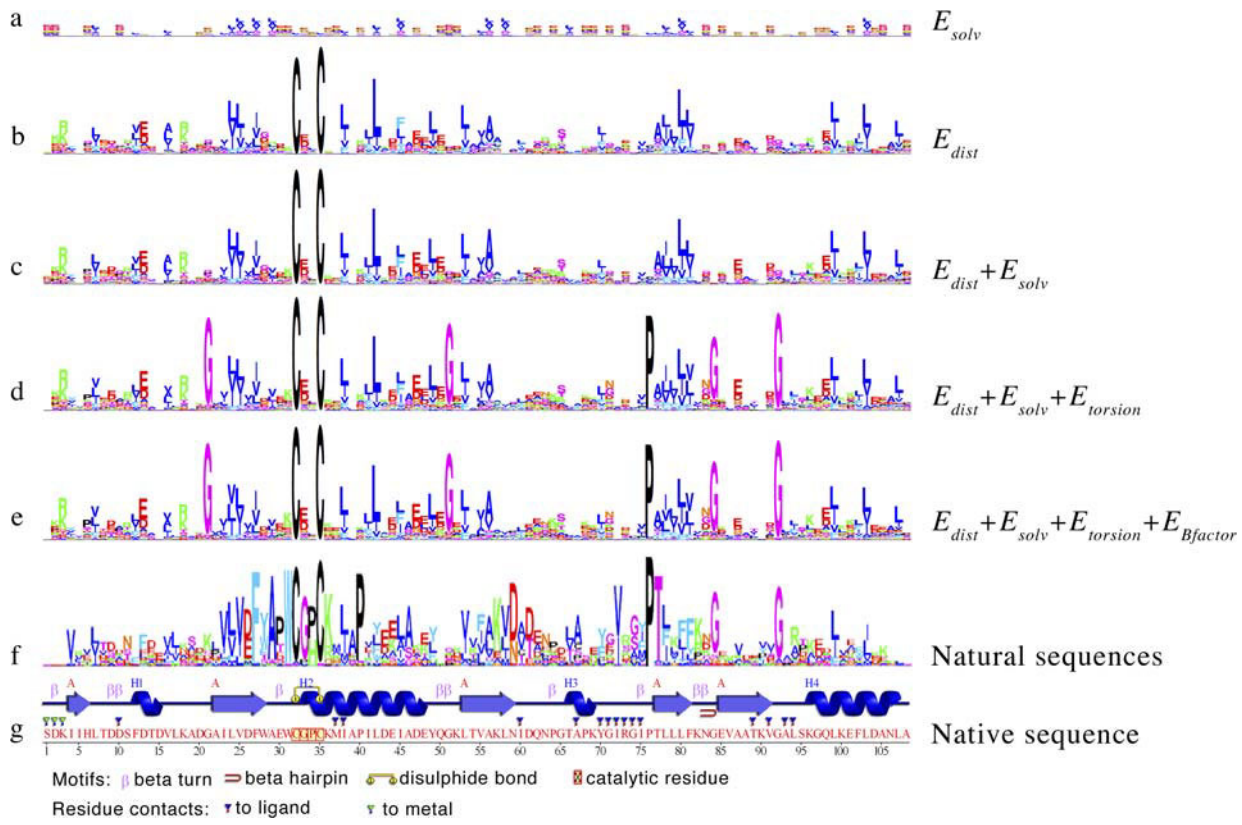

FIG. 4. Sequence logos of site-specific profiles induced on a thiorredoxin (PDB: 2TRX, chain A), using the potentials (a) Ml_{solv}, (b)ML_{dist}, (c) $ML_{dist,solv}$, (d) $Ml_{dist,solv,torsion}$, and (e) $ML_{dist,solv,Bfactor,torsion}$. (f) Profile obtained from a multiple sequence alignment of 162 eukaryotic sequences. (g) Native sequence of the reference protein. Secondary structure representation from PDBsum (Laskowski 2009). A color version of this figure is available as supplementary material, Supplementary Material online.

natural sequences are highly diverged, and so the existence of many potential alignment errors cannot be dismissed.

Globally, designed sequences show a low degree of similarity to natural sequences. Residues that owe their conservation to known specific functional constraints are not predicted at all, as expected, simply because the properties conferring their importance are not being included in the protein structural description. Ligand-binding sites (positions 10, 37, 38, 70–75, 89, 91, 93, 94), or residues in the catalytic site (positions 32–35), fall in this category. Apart from sites with known functional roles, the method fails to predict a number of conserved sites, particularly aromatic residues (positions 12, 27, 28, 31, 49, 81, 102) and specific polar interactions (e.g., Asp26-Lys82, Lys57-Asp61).

Nevertheless, a few general trends are apparent. Regarding the individual structural terms, distance-based potentials ML_{dist} predict very strongly disulfide bonds and tend to predict mainly residue hydrophobicity (supplementary fig. S4, Supplementary Material online). The high redundancy between distance and solvent accessibility potentials suggested by the CV experiments is also apparent here as the sequence logos remain almost unchanged when adding the solvent terms. Several recent studies trying to link evolutionary rate to structural properties point to the solvent accessibility component as one of the main constraints (Goldman et al. 1998; Bustamante et al. 2000; Choi et al. 2006; Conant and Stadler 2009; Franzosa and Xia 2009; Gong et al. 2009). In all the cases, site independence is assumed. However, we can see that a rich description of pairwise interactions like the one presented here suffices to capture most of the information contained in the solvent accessibility terms, suggesting that the solvent exposure would not be in fact the main structural constraint.

A similar effect is observed for the B-factor information: it does not add any qualitatively different information, but it seems instead to modulate the strength of very few predictions (e.g., position 87). Torsion terms, on the other hand, provide new information, changing the predictions for a few key amino acids such as prolines or glycines. In this particular example, thioredoxin has two prolines with very important structural roles. Pro76 is found in *cis* conformation, conserved through evolution and correctly predicted by the potentials including torsion terms. Pro40, on the other hand, produces a bending in a long alpha-helix; the latter feature is not currently modeled by the potentials because the identity and conformation of neighboring sites are not considered when calculating the conformation of a residue, although it is known to affect the Ramachandran basin populations (Zaman et al. 2003). We are considering the inclusion of this feature in future work. As for glycines, potentials with torsion terms predict four of them very strongly; two of which (Gly84 and Gly92) are conserved in the profile of natural sequences, whereas the other two (Gly21 and Gly51) are not. However, this discrepancy is

Table 2. Natural Logarithm of the Bayes Factor and Optimal β for the Models Considered. ω Was Included in the Models Either as a Global Parameter (noted as G) or with a Dirichlet Distribution (noted as DP). Shaded Cells Show Site-independent Models of Sequence Evolution: MG-NS Corresponds to Row 5, and MG-NS^{DP} Corresponds to Row 10. MG Was Used as a Reference Model for the Calculation of Bayes Factors.

		ADH		β-globin	
ω	Potential	Log B_M	β	Log B_M	β
—	ML_{dist}	[145.90:146.01]	[0.383 : 0.390]	[81.99 : 82.16]	[0.312 :0.316]
—	$ML_{dist,solv}$	[162.35:162.82]	[0.390 : 0.392]	[90.00 : 90.03]	[0.325 : 0.327]
—	$ML_{dist,solv,torsion}$	[213.41:214.89]	[0.418:0.419]	[104.88 : 105.69]	[0.327 : 0.328]
—	$ML_{dist,solv,Bfactor,torsion}$	[222.37:222.76]	[0.414 : 0.419]	[114.47 : 114.64]	[0.331 : 0.333]
G	—	[316.3: 319.1]	—	[90.64 : 93.88]	—
G	ML_{dist}	[409.14 : 412.75]	[0.372 : 0.376]	[149.55 : 153.37]	[0.302 : 0.306]
G	$ML_{dist,solv}$	[417.96 : 421.10]	[0.370 : 0.381]	[155.69 : 159.04]	[0.297 : 0.317]
G	$ML_{dist,solv,torsion}$	[453.55 : 457.28]	[0.401 : 0.408]	[168.36 : 172.30]	[0.319 : 0.323]
G	$ML_{dist,solv,Bfactor,torsion}$	[458.32 : 461.92]	[0.397 : 0.399]	[174.73 : 178.90]	[0.325 : 0.326]
DP	—	[413.10 : 419.40]	—	[192.84 : 198.08]	—

easily understood when looking at the actual alignment of natural sequences: both glycines are in fact present in more than one-third of the sequences, but the alignment programs fail to position them properly because they are located in very divergent loops of the protein, where a high number of insertions and deletions are found.

Despite the limitations discussed above, a detailed analysis of the profiles of a particular protein like the one presented here allows for an intuitive visualization of the properties of the different statistical potentials. It spans a broad portion of the sequence space, using a large number of highly diverged sequences, which is more difficult to achieve within a phylogenetic framework.

Assessment in a Phylogenetic Context

Once the parameters of the potentials have been optimized, they can be inserted into a structurally constrained model of sequence evolution and assessed in a Bayesian framework. The log-Bayes factors for two data sets of globular proteins, ADH and β-globin, are shown in table 2. The thermodynamic integration produces a curve representing the log-Bayes factor of each model as a function of β, the factor modulating the strength of the structural term in the evolutionary model (eq. (15)). This allows us, in addition to performing comparisons, to detect the optimal values of β for each model. We will first focus on this measure (table 2). Following the trend we observed using simpler SC models (Rodrigue et al. 2009), we find the optimal β to be positive, consistent with the case where sequences are selected for their compatibility to the structure. Note that the potentials were conceived to maximize a probability similar to the stationary distribution of the site interdependent codon model given in equation (16), although ignoring the contribution of the mutation bias, and with $\beta = 1/2$ (see Rodrigue et al. 2009 for details). The optimal value of β obtained is slightly below this expected value of $1/2$ maybe due to the fact that we are ignoring mutational pressure in the optimization procedure. Note that β-globin shows globally lower values of optimal β. This is probably due to the important structural features of this protein that are not described by the ML potentials considered here: the β-globin structure is greatly influenced by the prosthetic group and by interactions with the other subunits of this oligomeric protein. In any case, for both proteins, models with richer structural description show a progressively higher optimal β: the better the structural representation, the stronger role this term plays in the evolutionary model.

The progression of the Bayes factor values when adding the structural terms one by one, similar to the trend observed before when measuring the fit of native sequence–structure pairs (fig. 2), indicates that the sequence–structure patterns captured by the potentials are also meaningful in an evolutionary context. Once again, pairwise interactions are the most important single component contributing to model fit.

Although improving the description of the evolutionary process when contrasted to the MG model, the performance of the SC models remains altogether weak. MG-NS, a site-independent model with only one global parameter modeling selection (ω), has a comparable performance (better in one case, worse in the other). Combining the structural specifications with the MG-NS model increases the model fit, though in a less important way than when adding them to a pure MG model. This is similar to what had been observed before (Rodrigue et al. 2009), which we interpret as a consequence of the overlap in the two approaches—ω and the SC settings—of modeling the purifying selection. Note, however, that despite this overlap, the combined MG-NS-SC model displays a fit that is in the order of MG-NS^{DP} (a site-independent model allowing heterogeneity of ω across sites), which the simpler SC models failed to attain before (Rodrigue et al. 2009). This suggests that the structural components of the model are explaining, if not the average nonsynonymous rate of substitution, a part of the heterogeneity of nonsynonymous rates across sites.

The mechanistic formulation of this approach allows for a simple interpretation of certain model violations. As an example, we analyzed a third protein, calmodulin, for which simple general rules of protein structure may not apply. Calmodulin acts as an intermediary protein that reacts to calcium levels and relays signals to numerous proteins. For this purpose, calmodulin undergoes major conformational changes (Hoeflich and Ikura 2002). As such, this type of

Table 3. Natural Logarithm of the Bayes Factor and Optimal β for the Models Considered, Considering Separately the Native Sequence (s^c) and All the Other Sequences in the Alignment (D^ϕ). See Methods for Details. MG Was Used as a Reference Model for the Calculation of Bayes Factors.

	Data set			
Potential	$B_M^{s^c}$	B_M^{ϕ}	β^{s^c}	β^{ϕ}
ADH				
ML_{dist}	[76.84:76.86]	[69.85 : 69.92]	[0.413 : 0.417]	[0.356 : 0.360]
$ML_{dist,\ solv}$	[85.27:85.40]	[77.33 : 77.86]	[0.410 : 0.414]	[0.372 : 0.373]
$ML_{dist,solv,torsion}$	[106.92:107.28]	[106.49 : 107.61]	[0.416 : 0.418]	[0.420 : 0.420]
$ML_{dist,solv,Bfactor,torsion}$	[110.83:110.99]	[111.40 : 111.92]	[0.418 : 0.419]	[0.412 : 0.419]
β-globin				
ML_{dist}	[47.03 : 47.04]	[39.36 : 39.51]	[0.410 : 0.432]	[0.261 : 0.266]
$ML_{dist,solv}$	[50.01 : 50.07]	[43.54 : 43.62]	[0.411 : 0.412]	[0.276 : 0.276]
$ML_{dist,solv,torsion}$	[54.72 : 54.76]	[52.70 : 53.69]	[0.393 : 0.402]	[0.288 : 0.298]
$ML_{dist,solv,Bfactor,torsion}$	[59.48 : 59.55]	[58.19 : 57.89]	[0.408 : 0.415]	[0.292 : 0.292]

protein may not be well represented in the PDB. When applying the SC models, we observe a progressive increase in model fit (supplementary fig. S5, Supplementary Material online). However, this improvement is almost negligible compared with the fit of MG-NS, which is five times higher. Consistently, neither layering the SC settings with the parameter ω, nor modeling heterogeneous ω parameters across sites with the MG-NS^{DP} model improve significantly the fit (less than 10% improvement). Because the global selective pressure in the present case is known to be unrelated to maintaining a single rigid native structure, the detailed description of the amino acid interactions is not surprisingly meaningless in an evolutionary perspective.

Transient Properties of the SC Models

We also explored one additional aspect regarding the assessment of the SC models in this framework. Given our supervised learning procedure for optimizing the potentials, there is a risk of a bias toward the native sequence, that is, the sequence that was used to obtain the crystallographic structure, a risk that increases with the level of detail in the structural description (Kuhlman and Baker 2000). However, we are looking for a scoring function that predicts not only this native sequence s^c but also more general sequence features that could be accepted by evolution under the particular structural constraints of c.

We can probably be confident that the coarse-grained modeling adopted here prevents such an overfitting, but this can be addressed quantitatively based on the following argument. We can decompose the Bayes factor into two factors (eqs. (20–22)):

$$B_M = (B_M^{\phi})(B_M^{s^c}).$$

The factor B_M^{ϕ}, which we call the "transient" factor, measures the ability of the model to generalize beyond the native sequence and predict new sequences related to the native one by their evolutionary history. The "stationary" factor $B_M^{s^c}$, in turn, corresponds to the fit of the model on the native sequence itself. The results are reported in table 3. Note that both factors progress in the same order for the different potentials and that the transient factor B_M^{ϕ} increases faster when enriching the SC model. This implies that the structural specification is modeling meaningful selective constraints and not merely describing too faithfully the relation between the native sequence and its structure.

Finally, note that the stationary factor represents an important contribution to the total Bayes factor, which may indicate that much of the model fit is obtained by explaining the native sequence. Although it is true that, given that the model is time reversible, the marginal likelihood is invariant to the choice of s^c, the transient and stationary factors individually are not. In order to assess the role of the native sequence in this contribution, we repeated the experiment but considering all the sequences of the alignment, one at a time, as s^c (fig. 5). We can see that the actual native sequence is not the one displaying the best stationary fit, indicating once again that the SC models are not merely predicting the native sequence. Changing s^c for other

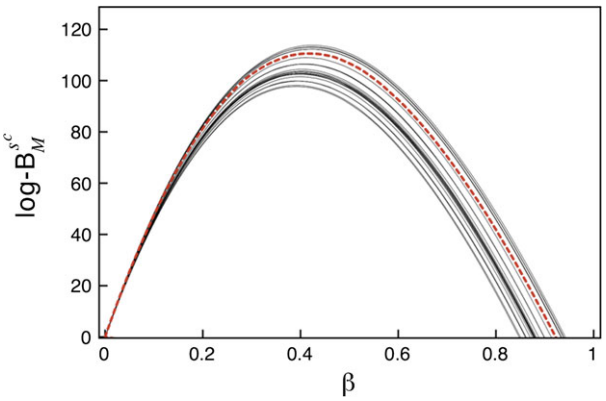

FIG. 5. Trace plots representing the stationary factor $B_M^{s^c}$ as a function of β, the factor modulating the strength of the structural term in the evolutionary model. The computation was performed on the ADH data set, using the potential combining torsion angles, solvent accessibility, pairwise interactions, and B-factors ($ML_{dist,solv,Bfactor,torsion}$). In each curve, a different sequence from the alignment is taken as s^c. The dashed line corresponds to the case where the native sequence is taken as s^c.

sequences produces relatively minor changes in the overall behavior of the plots for the two proteins tested (supplementary figs. S6 and S7, Supplementary Material online), suggesting that what is at stake here is a transient–stationary distinction rather than a native–non native one. The potentials have been optimized in a stationary state, without considerations related to the transient aspects of the evolutionary model; model violations may thus be more evident in the description of transient properties of the evolutionary process. A wide range of codon substitution models, presenting the same associated stationary distribution to the one used here, but different transient forms, could be explored to further investigate this question (Thorne et al. 2007).

Conclusion and Perspectives

The main motivation behind this work is to incorporate explicit protein structure information in an evolutionary context using a unified model-based statistical framework to assess the relevance of this information. To what extent are the factors known to affect protein structure—in vitro, in isolation and controlled laboratory conditions—shaping the evolution of protein sequences? Can we disentangle structural constraints from other selective forces? To address these questions, we derived statistical potentials with rich structural descriptions, optimized for evolutionary studies. We incorporated them into a structurally constrained model of sequence evolution and evaluated them in a Bayesian framework.

We found that including detailed information on the protein structure improves the description of the evolutionary process. However, the performance of the potentials remains relatively weak compared with the most sophisticated site-independent models of evolution. Further improvements could be made regarding the specific form of the energy function, including terms related to interactions in torsion angles among successive positions along the chain (Betancourt and Skolnick 2004), side chain–backbone interactions (Buchete et al. 2004), or considering sequence separation ranges for distance interactions (Sippl 1993). The modeling of flexibility, in particular, needs significant improvement. Even though B-factors have been previously used as an approximation of protein flexibility (Schlessinger and Rost 2005; Yuan et al. 2005), our results do not support this role. The coarse-grained representation of the structure provides an indirect way of allowing flexibility, but given its importance for protein function, an explicit modeling of this feature would be desirable. Other measures of protein dynamics could be explored, for example, considering several conformations for each sequence in the learning database, each one representing different protein states, or homologous structures. In a different direction, refinements of the optimization procedure, which has not been modified here, should be considered, such as elements of negative design (Bolon et al. 2005), by the use of explicit decoy structures or better approximations than the random energy model.

In any case, structural constraints represent only a fraction of the total selective constraint operating on sequences (Drummond et al. 2006; Pal et al. 2006; Drummond and Wilke 2008). As shown by the logos of natural sequences, relatively few positions are strongly conserved, suggesting that the critical interactions for maintaining the overall structure may be relatively sparse. This has also been proven experimentally: a statistical function capturing coevolution in a sequence alignment, specifying very few key positions, suffices to produce correctly folded proteins in vitro (Suel et al. 2002; Socolich et al. 2005). Because Bayes factors are a global measure of how well all aspects of the data are explained by the model, if there are only a handful of positions constrained by the structure, the improvement in model fit will be minor.

More importantly, there is an intrinsic limitation of the modeling approach used here. Statistical potentials are designed to capture general trends of amino acid propensities for average proteins, well represented in the learning data set. However, as illustrated by the example of calmodulin, and to a lesser extent by β-globin, each protein structure has features critical for its function, folding, and stability, which may be too particular to be accessible by estimating propensities over a large number of cases. Estimating the parameters for specific protein families, or, better yet, inferring them directly within the phylogenetic framework, along with the other parameters of the evolutionary model, may serve to overcome this limitation. In a more ambitious direction, more physically based representations and energy functions could be used to model protein structure instead of relying on statistical potentials. This approach will certainly be computationally demanding, thus limiting the amount of data that can be analyzed, but it may prove to be a more direct and robust way to characterize structural constraints.

All in all, the quantitative analysis performed in this study, combining a mechanistic approach to modeling evolution with model-based statistical inference, may now be applied to study less well-characterized particular proteins to answer more specific biological questions. In a different perspective, this framework can be extended naturally to handle other aspects of protein structure affecting sequence evolution, such as folding constraints, interactions with other proteins, or yet other phenotypic features, not exclusively related to the native conformation.

Supplementary Material

Supplementary figures S1–S7 and Table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Yaoqing Shen and Béatrice Roure for critical comments on the article, and José M. Delfino, Javier Santos, Cécile Bonnard, and Jean-Christophe Grenier for helpful discussions. Henner Brinkmann kindly assisted with the phylogenetics data sets. This work was supported by the Natural Sciences and Engineering Research Council of Canada,

the biT fellowships for excellence (a Canadian Institutes of Health Research strategic training program grant in bioinformatics), the Robert Cedergren Centre for Bioinformatics and Genomics, and the Canadian Research Chair Program. We also thank the Réseau Québécois de Calcul de Haute Performance for computational resources.

References

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol.* 26:255–271.

Artymiuk PJ, Blake CC, Grace DE, Oatley SJ, Phillips DC, Sternberg MJ. 1979. Crystallographic studies of the dynamic properties of lysozyme. *Nature* 280:563–568.

Bastolla U, Vendruscolo M, Knapp EW. 2000. A statistical mechanical method to optimize energy functions for protein folding. *Proc Natl Acad Sci U S A.* 97:3977–3981.

Betancourt MR, Skolnick J. 2004. Local propensities and statistical potentials of backbone dihedral angles in proteins. *J Mol Biol.* 342:635–649.

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

Bolon DN, Grant RA, Baker TA, Sauer RT. 2005. Specificity versus stability in computational protein design. *Proc Natl Acad Sci U S A.* 102:12724–12729.

Bonnard C, Kleinman CL, Rodrigue N, Lartillot N. 2009. Fast optimization of statistical potentials for structurally constrained phylogenetic models. *BMC Evol Biol.* 9:227.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS Comput Biol.* 5:e1000392.

Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416:507–511.

Buchete NV, Straub JE, Thirumalai D. 2004. Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci.* 13:862–874.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Mol Biol Evol.* 17:301–308.

Chiu TL, Goldstein RA. 1998. Optimizing potentials for the inverse protein folding problem. *Protein Eng.* 11:749–752.

Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. *Mol Biol Evol.* 24:1769–1782.

Choi SS, Vallender EJ, Lahn BT. 2006. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. *Mol Biol Evol.* 23:2131–2133.

Conant GC, Stadler PF. 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol Biol Evol.* 26:1155–1161.

Delport W, Scheffler K, Seoighe C. 2009. Models of coding sequence evolution. *Brief Bioinform.* 10:97–109.

Dimmic MW, Mindell DP, Goldstein RA. 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac Symp Biocomput.* 18–29.

Dobson CM. 2003. Protein folding and misfolding. *Nature* 426: 884–890.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol.* 23: 327–337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Flores TP, Orengo CA, Moss DS, Thornton JM. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* 2:1811–1826.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol.* 26: 2387–2395.

Frauenfelder H, Petsko GA, Tsernoglou D. 1979. Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* 280:558–563.

Gilis D, Rooman M. 1997. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J Mol Biol.* 272:276–290.

Gilis D, Rooman M. 2001. Identification and ab initio simulations of early folding units in proteins. *Proteins* 42:164–176.

Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N. 2005. The consurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins* 58:610–617.

Goldman N, Thorne JL, Jones DT. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J Mol Biol.* 263:196–208.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149:445–458.

Gong S, Worth CL, Bickerton GR, Lee S, Tanramluk D, Blundell TL. 2009. Structural and functional restraints in the evolution of protein families and superfamilies. *Biochem Soc Trans.* 37: 727–733.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Hoeflich KP, Ikura M. 2002. Calmodulin in action: diversity in target recognition and activation mechanisms. *Cell* 108:739–742.

Hubbard SJ, Thornton JM. 1993. Naccess. London: Department of Biochemistry and Molecular Biology, University College.

Huelsenbeck JP, Jain S, Frost SW, Pond SL. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci U S A.* 103:6263–6268.

Jensen JL, Pedersen AK. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv Appl Prob.* 32:499–517.

Jones DT. 1997. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins* Suppl 1:185–191.

Jones DT, Taylor WR, Thornton JM. 1992a. A new approach to protein fold recognition. *Nature* 358:86–89.

Jones DT, Taylor WR, Thornton JM. 1992b. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot N. 2006. A maximum likelihood framework for protein design. *BMC Bioinformatics* 7:326.

Kocher, JPA., Rooman MJ, Wodak SJ. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol.* 235:1598–1613.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. *Protein Eng.* 8:641–645.

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

Laskowski RA. 2009. PDBsum new things. *Nucleic Acids Res.* 37: D355–D359.

Laskowski RA, Macarthur MW, Moss DS, Thornton JM. 1993. PROCHECK—a program to check the stereochemical quality of protein structures. *J Appl Crystallogr.* 26:283–291.

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR.* 8:477–486.

Lazaridis T, Karplus M. 2000. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol.* 10:139–145.

Lio P, Goldman N, Thorne JL, Jones DT. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* 14:726–733.

Maguid S, Fernandez-Alberti S, Parisi G, Echave J. 2006. Evolutionary conservation of protein backbone flexibility. *J Mol Evol.* 63: 448–457.

Melo F, Sanchez R, Sali A. 2002. Statistical potentials for fold assessment. *Protein Sci.* 11:430–448.

Miyazawa S, Jernigan RL. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol.* 256:623–644.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol.* 11:715–724.

Overington J, Johnson MS, Sali A, Blundell TL. 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc Biol Sci.* 241:132–145.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet.* 7:337–348.

Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol.* 18:750–756.

Pedersen AM, Jensen JL. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol Biol Evol.* 18:763–776.

Ramachandran GN, Ramakrishnan C, Sasisekharan V. 1963. Stereochemistry of polypeptide chain configurations. *J Mol Biol.* 7:95–99.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol.* 20:1692–1704.

Rodrigue N, Kleinman CL, Philippe H, Lartillot N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol Biol Evol.* 26:1663–1676.

Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347:207–217.

Rodrigue N, Philippe H, Lartillot N. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol.* 23:1762–1775.

Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. 1997. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J Mol Biol.* 269:423–439.

Schlessinger A, Rost B. 2005. Protein flexibility and rigidity predicted from sequence. *Proteins* 61:115–126.

Seno F, Micheletti C, Maritan A, Banavar JR. 1998. Variational approach to protein design and extraction of interaction potentials. *Phys Rev Lett.* 81:2172–2175.

Sippl MJ. 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des.* 7: 473–501.

Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E. 2005. Protein structure and evolutionary history determine sequence space topology. *Genome Res.* 15:385–392.

Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci U S A.* 90: 7195–7199.

Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. 2005. Evolutionary information for specifying a protein fold. *Nature* 437:512–518.

Sternberg MJ, Grace DE, Phillips DC. 1979. Dynamic information from protein crystallography. An analysis of temperature factors from refinement of the hen egg-white lysozyme structure. *J Mol Biol.* 130:231–252.

Suel GM, Lockless SW, Wall MA, Ranganathan R. 2002. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol.* 10:59–69.

Sun S, Brem R, Chan HS, Dill KA. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.* 8:1205–1213.

Taverna DM, Goldstein RA. 2002a. Why are proteins marginally stable? *Proteins* 46:105–109.

Taverna DM, Goldstein RA. 2002b. Why are proteins so robust to site mutations? *J Mol Biol.* 315:479–484.

Thorne JL, Choi SC, Yu J, Higgs PG, Kishino H. 2007. Population genetics without intraspecific data. *Mol Biol Evol.* 24:1667–1677.

Thorne JL, Goldman N, Jones DT. 1996. Combining protein evolution and secondary structure. *Mol Biol Evol.* 13:666–673.

Wako H, Blundell TL. 1994a. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. i. Solvent accessibility classes. *J Mol Biol.* 238:682–692.

Wako H, Blundell TL. 1994b. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. ii. Secondary structures. *J Mol Biol.* 238:693–708.

Wang G, Dunbrack RL Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics* 19:1589–1591.

Williams SG, Lovell SC. 2009. The effect of sequence evolution on protein structural divergence. *Mol Biol Evol.* 26:1055–1065.

Xia Y, Huang ES, Levitt M, Samudrala R. 2000. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol.* 300:171–185.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol.* 10:1396–1401.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Yuan Z, Bailey TL, Teasdale RD. 2005. Prediction of protein B-factor profiles. *Proteins* 58:905–912.

Zaman MH, Shen MY, Berry RS, Freed KF, Sosnick TR. 2003. Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the flory isolated-pair hypothesis for peptides. *J Mol Biol.* 331:693–711.