Fast optimization of statistical potentials for structurally constrained phylogenetic models

Cécile Bonnard, Claudia L. Kleinman, Nicolas Rodrigue, Nicolas Lartillot

BMC Evolutionary Biology

BioMed Central

Methodology article

Fast optimization of statistical potentials for structurally constrained phylogenetic models

Cécile Bonnard*[1,2], Claudia L Kleinman[2], Nicolas Rodrigue[3] and Nicolas Lartillot[2]

Address: [1]Département d'Informatique, LIRMM, 161 rue Ada, 34392 Montpellier Cedex 5, France, [2]Département de Biochimie, Université de Montréal, Montréal, Québec, Canada and [3]Department of Biology, University of Ottawa, Ottawa, Ontario, Canada

Email: Cécile Bonnard* - cecile.bonnard@umontreal.ca; Claudia L Kleinman - cl.kleinman@umontreal.ca; Nicolas Rodrigue - nicolas.rodrigue@uottawa.ca; Nicolas Lartillot - nicolas.lartillot@umontreal.ca

* Corresponding author

Published: 9 September 2009

BMC Evolutionary Biology 2009, **9**:227 doi:10.1186/1471-2148-9-227

Received: 9 April 2009
Accepted: 9 September 2009

This article is available from: http://www.biomedcentral.com/1471-2148/9/227

Abstract

Background: Statistical approaches for *protein design* are relevant in the field of molecular evolutionary studies. In recent years, new, so-called structurally constrained (*SC*) models of protein-coding sequence evolution have been proposed, which use statistical potentials to assess sequence-structure compatibility. In a previous work, we defined a statistical framework for optimizing knowledge-based potentials especially suited to SC models. Our method used the maximum likelihood principle and provided what we call the *joint* potentials. However, the method required numerical estimations by the use of computationally heavy *Markov Chain Monte Carlo* sampling algorithms.

Results: Here, we develop an alternative optimization procedure, based on a *leave-one-out* argument coupled to fast gradient descent algorithms. We assess that the leave-one-out potential yields very similar results to the joint approach developed previously, both in terms of the resulting potential parameters, and by Bayes factor evaluation in a phylogenetic context. On the other hand, the leave-one-out approach results in a considerable computational benefit (up to a 1,000 fold decrease in computational time for the optimization procedure).

Conclusion: Due to its computational speed, the optimization method we propose offers an attractive alternative for the design and empirical evaluation of alternative forms of potentials, using large data sets and high-dimensional parameterizations.

Background

Recent advances in computer science and in the acquisition of new genetic sequences from a variety of organisms have opened up a wide spectrum of new possibilities in molecular evolutionary modeling. In particular, codon substitution models explicitly formulated in terms of a balance between mutation and selection constitute an attractive strategy [1-4]. By deriving the substitution process from basic principles of population genetics, their aim is to bridge the gap between population genetics and phylogenetics, and thus to offer a better understanding of the driving forces of the long term evolutionary process. More specifically, these mutation-selection models propose that

the substitution rate from a sequence s to another s' ($R_{ss'}$) depends on the rate of mutation from s to s' ($Q_{ss'}^{mut}$), and on the probability for this mutation to be fixed in the population ($p_{fix}(ss')$):

$$R_{ss'} = Q_{ss'}^{mut} \cdot p_{fix}(ss'). \quad (1)$$

The mutation matrix $Q_{ss'}^{mut}$ depends only on the underlying mutation model, and is generally assumed to be fixed along the lineages and uniform along the sequence. The fixation probability $p_{fix}(ss')$ depends on the particular model chosen.

Among the mutation-selection codon models, we focus on the structurally constrained (SC) models [4-7] which attempt to explicitly link a protein's tertiary structure to the evolution of its sequence. They consider that a protein is under a purifying selection maintaining a stable and constant tertiary structure. Importantly, and unlike most probabilistic models currently used in molecular evolutionary studies, SC models are explicitly site-interdependent, and therefore, require complex Monte Carlo methods to be implemented and applied to empirical data [3,4,8].

In SC models, the fixation probability of a given mutation depends on a score function assessing the adequacy of a sequence s to the tertiary structure of the protein, c. This score should be devised so that the fixation probability is low if the proposed mutation destabilizes the structure or complicates the folding process. Since Anfinsen's experiments [9], the relations between protein structure and sequence have been carefully studied and an intuitive approach consists in relying on first principles of protein thermodynamics, using all-atom force fields (e.g. AMBER [10], CHARMM [11]). However, in our case, the instantaneous rate of substitution ($R_{ss'}$), and thus the structure/sequence score function, have to be computed for each possible nearest neighbor mutant, and for each substitution, along the entire evolutionary tree. Therefore, we need a fast computation of the fixation probability which precludes the use of all-atom force fields.

An attractive alternative is provided by knowledge-based (or statistical) potentials. They mimic the Boltzmann law [12-15] and usually rely on a coarse-grained description of the structure, implicitly integrating out the degrees of freedom of the side chains and thus avoiding the complexity and the computation requirements of all-atom force fields [16-23]. In addition, they are trained empirically from databases of natural proteins. This latter point is of particular interest in evolutionary studies, where we are interested in all aspects of the relations between sequence and structure prevailing in natural sequences, and not only in the specific problem of the thermodynamic stability. In this respect, one expects that learning potentials from native structure-sequence databases using blind machine learning methods will capture all such aspects.

Many statistical potentials have been proposed [12,14,15,19,24,25], either to predict the fold of a given sequence (*protein folding*) or to find a sequence or a set of sequences folding into a given tertiary structure (*protein design*). However, the same potential may not be best-suited to both goals since the spaces of optimization are very different: in the protein folding problem the search is done over the structure space, while in the protein design problem the search is done over the sequence space. The phylogenetic context described here is more akin to a protein design perspective, as the structure of the protein is assumed constant during evolution, representing a constraint under which the sequence is evolving.

Several methods have been developed to train statistical potentials in a protein design perspective [19,24,25]. In a previous work, we introduced a probabilistic framework for protein design purposes based on the maximum likelihood principle [26]. The likelihood we considered was the probability of the sequences S given their native structures C and the model parameters (here, the statistical potential parameters, θ), $P(S|C, \theta)$. This probability was then maximized with respect to the potential parameters (e.g. pairwise contact energy coefficients) by a gradient method. However, the probability $P(S|C, \theta)$ involves a normalizing factor, summing over all possible sequences, which cannot be analytically calculated. We thus had to resort to a Markov Chain Monte Carlo (*MCMC*) numerical procedure: at each step of the gradient descent, we generated a set of sequences by Gibbs sampling, conditional on the current values of the potential. This set of sequences was then used to estimate the gradient. The Gibbs sampling procedure was the limiting step of our algorithm, restricting the set of alternative potentials that we could explore and empirically test. The potentials we obtained using this method are called *joint* potentials hereafter.

Interestingly, Kuhlman and Baker [27] used a *leave-one-out* procedure to estimate a restricted set of parameters of a free physical energy function in order to do protein design. In this procedure, only one site of the protein is changed at a time, while the other positions are kept fixed in their native state. The procedure is thus similar to training a potential to recognize acceptable sequence variants, given the target structure, among all possible point mutants. The leave-one-out criterion seems to give good results. However, it has never been assessed against alternative methods. Here, we adapt the statistical framework

we defined in [26] now using the leave-one-out definition of the likelihood to perform the gradient descent instead of the joint likelihood. We compare the potential parameters obtained by the two methods, and we establish that we can be highly confident in the results obtained using the leave-one-out likelihood. Overall, the leave-one-out procedure allows much faster computations while giving sensibly the same results as the joint one.

Results

Likelihood framework

As in [26], we formulate the problem in terms of a probabilistic model, considering a sequence $s = (s_i)_{1..n}$ of length n according to a probability distribution $P(s|c, \theta)$, conditional on the conformation c and on a set of potential parameters θ. The parameters are estimated by maximizing the probability of observing a database of N independent sequence-structure pairs $(\tilde{S}, C)$, with $\tilde{S} = (\tilde{s}^p)_{p=1..N}$, $C = (c^p)_{p=1..N}$. Here, $\tilde{s}^p = (\tilde{s}_i)_{i=1..n_p}^p$ is the p-th native sequence of the dataset, n_p is the lenght of this sequence and c^p is the native conformation associated with $\tilde{s}^p$. In practice, a native sequence-structure pair corresponds to a protein taken from the PDB.

The probability that we want to maximize can be expressed as follows:

$$P(\tilde{S} \mid C, \theta) = \prod_p P(\tilde{s}^p \mid c^p, \theta). \quad (2)$$

As a function of θ, this term can be seen as a likelihood. Hereafter, we define the methodology with one protein, but it can be easily generalized to a set of proteins.

Borrowing from [26], we set:

$$P(s \mid c, \theta) = \frac{e^{-G(s|c,\theta)}}{\sum_{s' \in \mathbb{S}} e^{-G(s'|c,\theta)}} = \frac{e^{-G(s|c,\theta)}}{Y}, \quad (3)$$

where Y is called the *normalization factor*, and $G(s|c, \theta)$ the *inverse potential*, defined as

$$G(s \mid c, \theta) = E(s \mid c, \theta) - F(s), \quad (4)$$

where $E(s|c, \theta)$ is the statistical potential and $F(s)$ is analogous to a free energy term and can be approximated using the *random energy model* [19,28-30]:

$$F(s) = \sum_{1 \le i \le n} \mu_{s_i}, \quad (5)$$

where μ_a, $a = \{1..20\}$ are unknown parameters, analogous to *chemical potentials* [26].

Optimization method

Joint likelihood maximization

In our previous work [26], we defined a score function $\omega(\tilde{s} \mid c, \theta)$ as:

$$\omega(\tilde{s} \mid c, \theta) = -\ln P(\tilde{s} \mid c, \theta) = G(\tilde{s} \mid c, \theta) + \ln Y. \quad (6)$$

This score function should be minimized conditional to θ. Its gradient is:

$$\frac{\partial \omega(\tilde{s}|c,\theta)}{\partial \theta} = \frac{\partial G(\tilde{s}|c,\theta)}{\partial \theta} + \frac{\partial \ln Y}{\partial \theta} = \frac{\partial G(\tilde{s}|c,\theta)}{\partial \theta} - \left\langle \frac{\partial G}{\partial \theta} \right\rangle, \quad (7)$$

where $\langle \cdot \rangle$ stands for the expectation over sequences drawn from the probability defined by eq. 3. Given the size of the sequence space (20^n), this expectation cannot be computed analytically, and therefore, in [26] we used a MCMC method to numerically estimate this expectation.

Leave-one-out likelihood maximization

We define for site i, $i = 1..n$, the leave-one-out probability

$$P_i^l(s_i = a \mid \tilde{s}_{\setminus i}, c, \theta) = P_i^l(s_i = a \mid \forall j \neq i \; s_j = \tilde{s}_j, c, \theta), \quad (8)$$

which is the probability of having an amino acid a at site i, in the context of the native sequence at all other sites ($\forall j \neq i \; s_j = \tilde{s}_j$). This leave-one-out probability can easily be obtained by a normalization over all possible twenty outcomes at site i:

$$P_i^l(s_i = a \mid \tilde{s}_{\setminus i}, c, \theta) = \frac{e^{-G_i(s_i=a|\tilde{s}_{\setminus i},c,\theta)}}{\sum_{k=1}^{20} e^{-G_i(s_i=k|\tilde{s}_{\setminus i},c,\theta)}}. \quad (9)$$

We can write this probability for any amino acid a, and in particular for the native amino acid at site i, $\tilde{s}_i$ i.e. $P_i^l(s_i = \tilde{s}_i \mid \tilde{s}_{\setminus i}, c, \theta)$. Taking the product over all positions $i = 1..n$, and by analogy with our previous definition of likelihood, we define the leave-one-out likelihood:

$$P^l(\tilde{s} \mid \tilde{s}, c, \theta) = \prod_{1 \le i \le n} P_i^l(s_i = \tilde{s}_i \mid \tilde{s}_{\setminus i}, c, \theta). \quad (10)$$

Note that this leave-one-out likelihood is normalized over the sequences, exactly as in the case of eq. 3. Therefore it yields a valid probability distribution over the sequence

space. On the other hand, the probability depends not only on c and θ, but also, in some sense, on the native sequence itself. To make this point explicit, we make $\tilde{s}$ appear on both sides of the conditioning bar.

We define the corresponding scoring function:

$$\omega^l(\tilde{s} \mid \tilde{s}, c, \theta) = -\ln P^l(\tilde{s} \mid \tilde{s}, c, \theta), \tag{11}$$

the gradient of which is immediately obtained (Additional File 1):

$$\frac{\partial \omega^l(\tilde{s}|\tilde{s},c,\theta)}{\partial \theta} = \sum_{i=1..n} \frac{\partial G_i(s_i=\tilde{s}_i|\tilde{s}\backslash_i,c,\theta)}{\partial \theta} - \sum_{i=1..n} \sum_{a=1..20} p_i(a) \frac{\partial G_i(s_i=a|\tilde{s}\backslash_i,c,\theta)}{\partial \theta}. \tag{12}$$

This gradient can be analytically calculated, at each step of a gradient descent. We note that the term corresponding to the normalization factor (the second term in eq. 12) can be seen as an expectation over the leave-one-out probability. It is thus analogous to the expectation appearing in the right hand of eq. 7. However, it is defined on a much more restricted universe ($20 \cdot n$ states, compared to the 20^n states in the case of the joint likelihood).

For implementing both methods, we used a simple form of potential [26], consisting in two terms: one related to contact interactions and the other to the solvent accessibility (see Methods).

Potential optimization

We first run our leave-one-out method on DS_l (see Methods). We consider that the optimization is complete when the overall maximum gradient is smaller than 10^{-2}. This corresponds to a variation of 10^{-6}, at most, in the value of the potential parameters. Using this stopping condition on the dataset DS_l with empirically tuned general steps (e.g for the contact parameters: $\delta^c_{grad} = 10^{-5}$ and for the solvent accessibility parameters: $\delta^a_{grad} = 10^{-4}$), we compare three different gradient descent methods (described in Methods): the simple gradient descent, the inertial gradient descent, and the controlled inertial gradient descent. The values of the parameters stabilized after 14,500 gradient steps for the simplest gradient descent, versus 1,500 gradient steps for the inertial gradient, and 1,200 gradient steps for the controlled inertial gradient. Concerning the last method, if we choose a different general step (e.g. $\delta^a_{grad} = 10^{-3}$ and $\delta^c_{grad} = 10^{-2}$) the procedure automatically reaches the optimal step for that dataset. At the beginning of the optimization procedure, the inertial component of the gradient greatly speeds up the optimization, but is automatically deactivated when the values of the potential parameters are near the optimum, thus avoiding the numerical instabilities usually observed using less adaptive gradient methods.

Independent runs from different and randomly chosen initial values for the parameters of the leave-one-out potential (θ), lead to the same final values of $\omega^l(\tilde{s} \mid \tilde{s}, c, \theta)$ (fig. 1) and of the potential parameters (fig. 2). These computations were done with the three gradient descent methods, and resulting always in the same final values, which suggests that, in the present case, we do not have local minima in the space of parameters. Similarly, the potential parameters obtained by two independent runs on the same dataset are very similar, indicating that our stopping condition is sufficient to have a good precision in our estimates (Additional file 2). In fig. 1 we have also represented the evolution of some parameters of the potential during optimization. We can see that the values of these parameters oscillate at the beginning of the gradient descent and then reach their optimal values. This behavior is caused by the evolution of the other parameters, as they influence each other during optimization. The complete series of parameter values obtained by our optimization method are presented in the additional file 3.

The contact potentials obtained with the leave-one-out optimization criterion make sense from a biological point of view (fig. 3): as expected, favorable interactions between amino acids in the contact potentials are represented by large negative value (e.g. the Cysteine-Cysteine contact energy, fig. 3), and by large positive value for unfavorable interactions (e.g. the Lysine-Lysine or Lysine-Arginine interactions, which are electrostatically repulsive). Concerning the accessibility potentials, it is important to note that we are working in a protein design context (i.e. we are evaluating the fitness of alternatives amino acids in a given accessibility class). Accordingly, the accessibility potentials have to be interpreted rowwise. If one wants to compare the accessibility potentials between classes for a given amino acid (i.e. in a protein folding perspective), one solution is to remove the logarithm of the frequency of the accessibility classes to each potential (additional file 4). Also, note that there is a lack of identifiability between α and μ, which has been be resolved by including the chemical potentials in the accessibility terms.

Complexity

In our previous work, we had to use a MCMC protocol to numerically evaluate the derivative of the gradient (see.

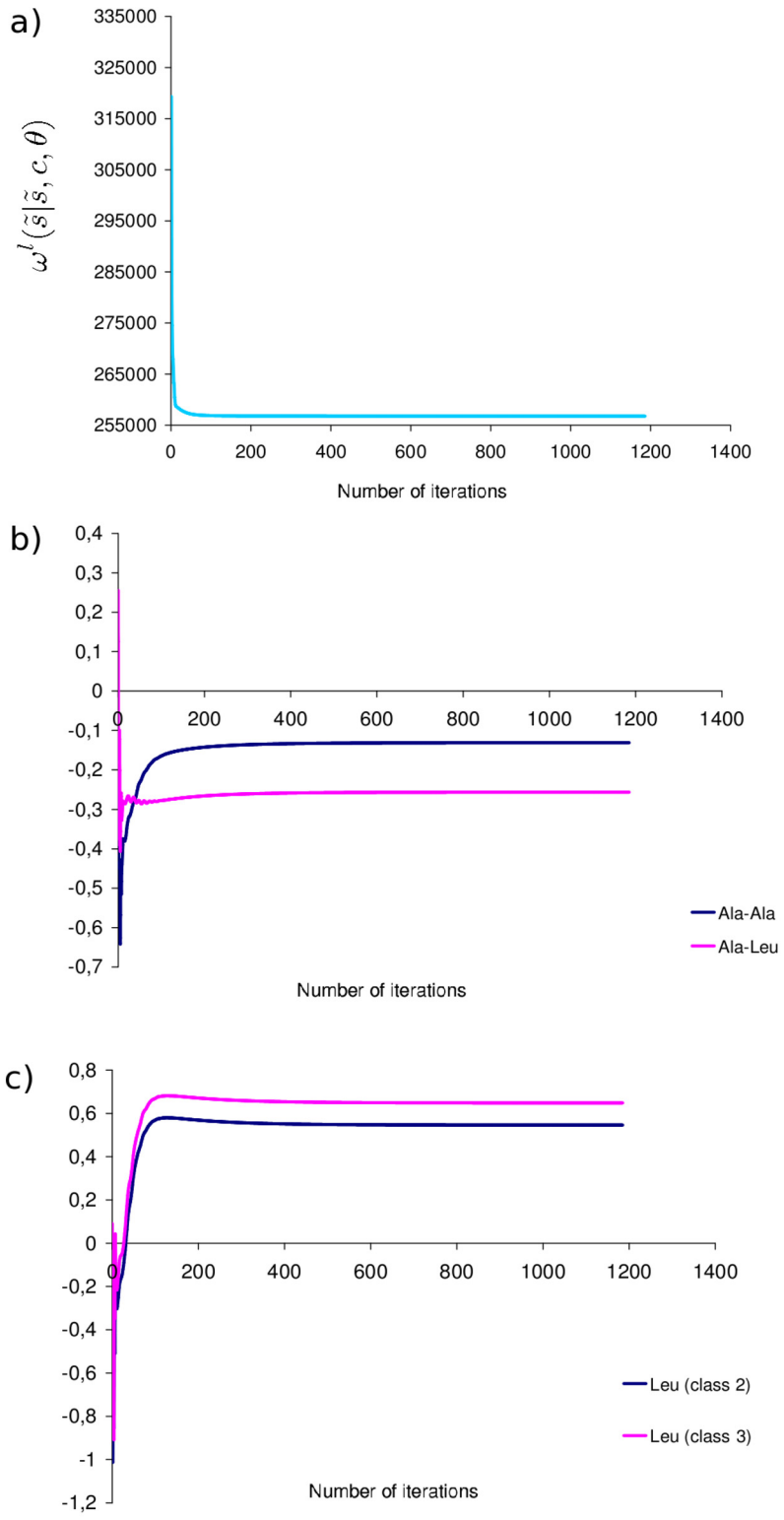

Figure 1

Convergence of the optimization procedure. Evolution of (a) the score function, (b) contact potential parameters and (c) accessibility potential parameters, for the dataset DS_l, using the controlled inertial gradient descent.

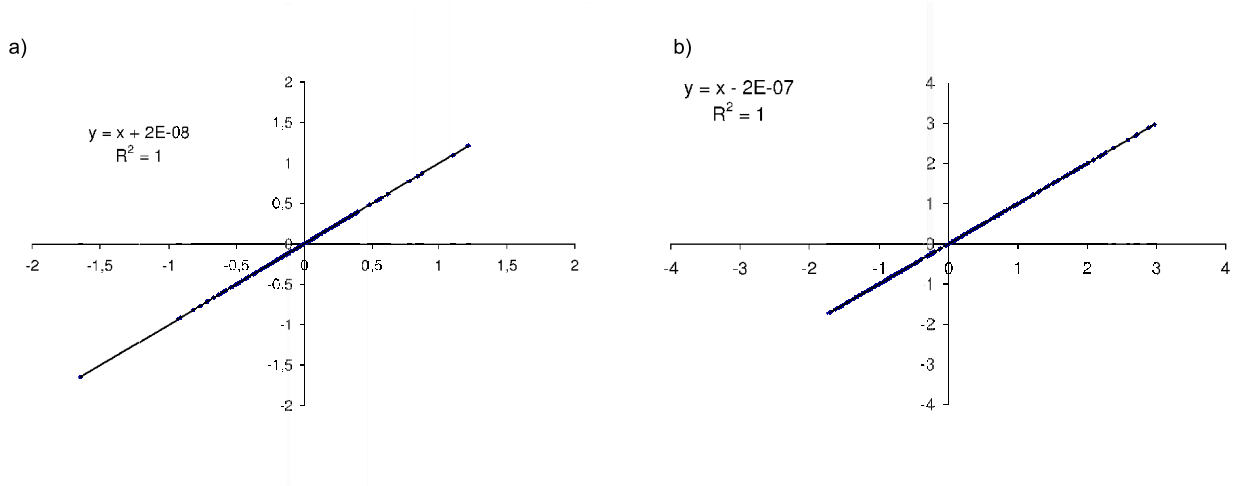

Figure 2
XY comparisons of the leave-one-out potential parameters. XY comparisons of two independent runs on the same dataset DS_I for (a) contact and (b) solvent accessibility potential parameters respectivly.

eq. 7), which was a computationally demanding task. At each step of the gradient descent, we had to sample a set of sequences by Gibbs sampling, under the current values of the parameters, so as to numerically estimate the gradient of the log-likelihood.

To compare the joint and the leave-one-out potentials, we first define an elementary calculation as the evaluation of the *inverse* potential at a particular site i for one particular amino acid a (what we called $G_i(s_i = a | \tilde{s}_{\setminus i}, c, \theta)$, eq. 9). This calculation has to be made in both cases. It is explicitly defined in the leave-one-out procedure (eq. 10), and is implicitly used in the joint context: an elementary step of the Gibbs sampling algorithm consist in computing, at a given site i the leave-one-out probability (eq. 9) for each possible amino-acid at this site, conditional on the rest of the sequence, and to choose the new aminoacid at site i according to these probabilities. Performing such an elementary update for every site in turn corresponds to one Gibbs sampling sweep and represents $20 \cdot n$ elementary computations. A reliable estimate of the joint expectation requires K sweeps (burn in included) and so, for a gradient step, we need $K \cdot n \cdot 20$ elementary calculations (in practice, $K \simeq 1{,}000$).

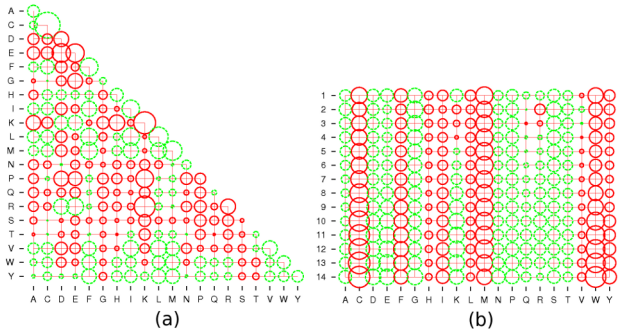

Figure 3
Validation of the potential parameters. Bubble plot representations of (a) contact potential parameters and (b) accessibility potential parameters obtained upon the dataset DS_I. Negative values are plotted in green while positive values are plotted in red.

In the case of the leave-one-out potential, we only have to make the equivalent of one sweep to exactly compute the gradient (eq. 12). Thus, we only need $n \cdot 20$ elementary calculations for a gradient step, which thus represents a 1,000-fold increase in computational speed compared to the joint method. In practice, and after the addition of the acceleration of the gradient descent, it took about one week to have a good estimate when we used the joint method, versus less than fifteen minutes when using the leave-one-out approach.

Potentials are indistinguishable

We applied the two optimization procedures (joint and leave-one-out) to the same dataset DS_I, and found a high correlation between the two resulting potentials (fig. 4). The correlation coefficient R^2 was about 0.96779 for the contact potential parameters and about 0.97374 for the accessibility potential parameters. For comparison, we applied the leave-one-out procedure on the two datasets DS1 and DS2 (see additional file 2) and found a correlation coefficient of 0.9477 for the contact parameters and

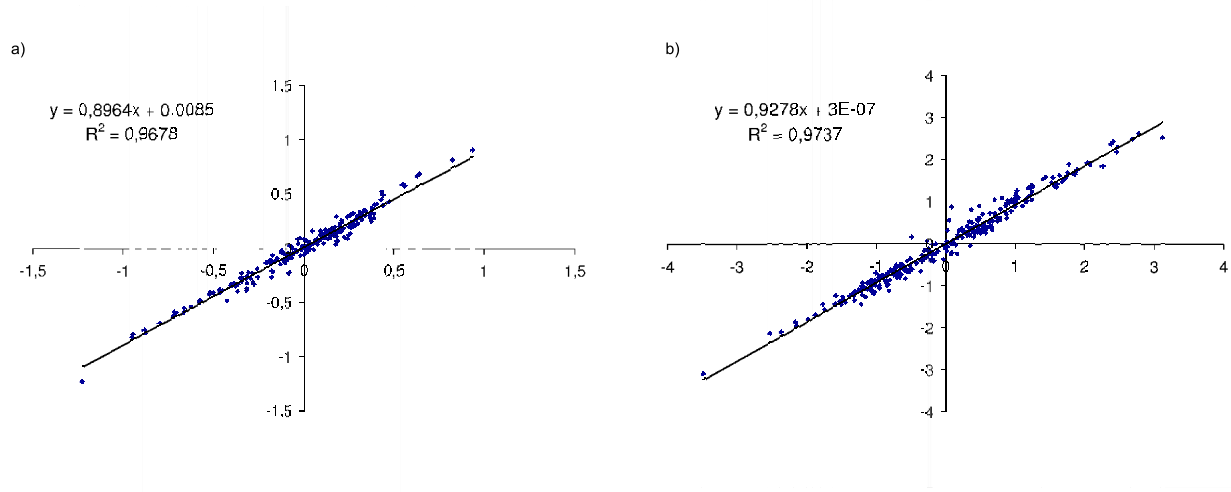

Figure 4
XY comparisons of the leave-one-out and joint potential parameters. XY comparisons between the two potentials (optimized on the same dataset DS_j), with, in X-axis the leave-one-out potential, and in Y-axis the joint potential. (a) represents the correlation between the contact potential parameters, and (b) the correlation between the accessibility potential parameters.

of 0.9596 for the accessibility parameters, indicating that the difference between the joint and the leave-one-out potentials is small compared to the sampling error due to the finite size of the training set. Altogether, the leave-one-out method appears to be a fast and reliable optimization procedure, yielding potentials that are virtually indistinguishable from those obtained under the joint method. As presented in [26], the contact potentials present a correlation (R^2 = 0.6565) with those of Miyazawa and Jernigan [13].

Phylogenetic evaluation

In eq. 1, we defined the substitution process of the SC model as a process depending on a mutation rate and a fixation probability. There are many ways the fixation probability could be expressed. Here, we do as in Robinson et al [4] and assume that this probability depends only on the potential difference (ΔG) between the original and the mutated sequences. Let us denote by s_{nuc} and s'_{nuc}, two sequences which differ only by a nucleotide, and s_{aa} and s'_{aa}, the corresponding amino acid sequences (which may be identical due to codon synonymy). Then, the rate of substitution between s and s' is:

$$R_{s_{nuc}s'_{nuc}} = Q^{mut}_{s_{nuc}s'_{nuc}} \cdot e^{-\beta\Delta G_{s_{aa}s'_{aa}}}, \qquad (13)$$

where $Q^{mut}_{s_{nuc}s'_{nuc}}$ is the mutation term depending only on the two sequences s_{nuc} and s'_{nuc}. $\Delta G_{s_{aa}s'_{aa}}$ is the energy difference between s_{aa} and s'_{aa}, and $\beta \geq 0$ can be considered as the strength of the structure-sequence constraint enforced by the model. Thus, a negative (resp. positive) ΔG means that the mutation is more (resp. less) likely to be accepted than a purely neutral (e.g. synonymous) mutation.

Note that the substitution process defined by eq. 13 is reversible and has a stationary distribution defined by:

$$\Pi_s \propto \Pi_0(s_{nuc})e^{-2\beta G(s_{aa})}, \qquad (14)$$

where $\prod_0(s_{nuc})$ is the stationary distribution induced by the pure mutation process ($Q^{mut}_{s_{nuc}s'_{nuc}}$). Given the way our potentials are optimized (see eq. 3 and 9) and assuming that natural sequences are sampled at equilibrium from the process defined by eq. 13, we then expect that the optimal value of β should be close to 0.5. In the following, we explore the entire range $\beta \in [0, 1]$.

We denote by SC^l_β the SC model defined using the leave-one-out potential and SC^j_β the SC model defined using the joint potential; the two models depend on β. Obviously, when $\beta = 0$, $SC^l_0 = SC^j_0 = SC_0$, and the model reduces to a pure mutation model which will be considered as our reference model.

We implemented our potential in the SC model as described in [3] and applied it to the GLOBIN15-144 dataset, with an underlying mutational specification inspired by the codon model in [31] and denoted as MG in [3]. This MCMC framework allows one to obtain a sample of parameter values and substitutional histories along the tree, drawn from the posterior distribution under the $SC^{l}_{0.5}$ model. Such a sample can then be marginalized over quantities of interest. Here, we briefly illustrate the approach by displaying the logo of the reconstructed mammalian ancestor hemoglobin sequence (fig. 5).

Since the leave-one-out procedure can be seen as an approximate but faster training method, compared to the joint method developed previously, we evaluated its impact on model fit via Bayes factors evaluations (see Methods). In this section we consider the three versions of the SC model, SC^{l}_{β}, based on a contact + accessibility leave-one-out potential, SC^{j}_{β}, based on a contact + accessibility joint potential, and SC^{c}_{β} based on a contact only joint potential. As explained in the methods, in the present case, the thermodynamic integration method yields a complete fitness curve (fig. 6) of each model (i.e. a curve representing the Bayes factor of each model against the reference model, as a function of β). In this way, we can readily spot the optimal value of β under each model, and report the Bayes factors under this optimal value (table 1).

As can be seen from fig. 6 and table 1, the models based on the joint and the leave-one-out potentials have a very similar fit across the whole range of value of β that we tested. Interestingly, in all but one cases, the Bayes factor

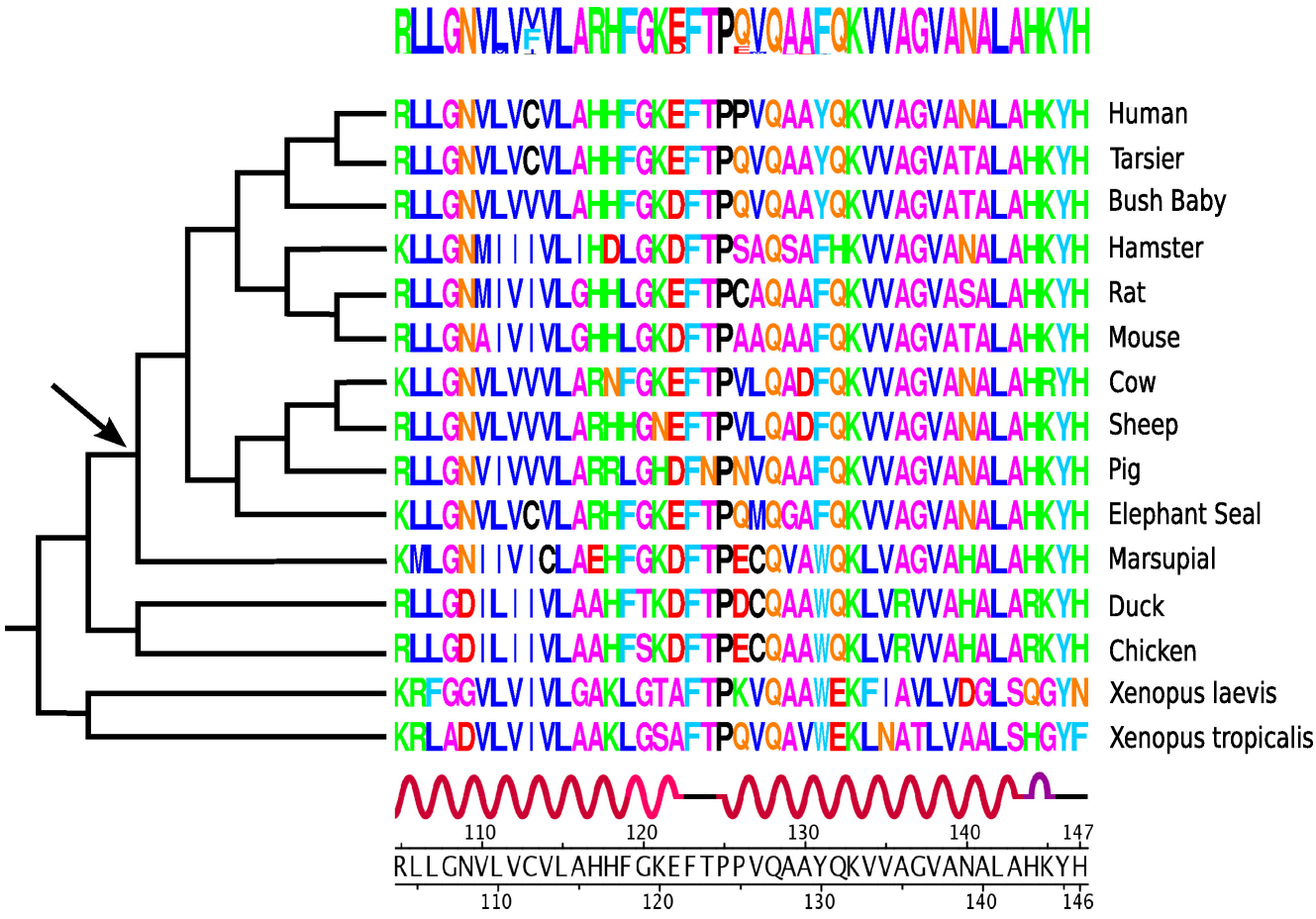

Figure 5
Logo profile of the mammalian ancestral globin sequence. The node is marked by an arrow. The translated sequences of the true alignment are displayed along with the secondary structure of the structure PDB code 4HHBB.

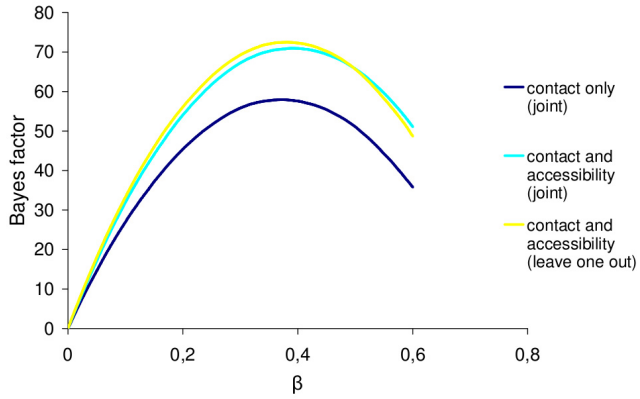

Figure 6

Bayes factor. Curves representing the Bayes factor as a function of β, with SC^{l}_{β} (in yellow), SC^{j}_{β} (in light blue) and SC^{c}_{β} (in dark blue), for the dataset BGLOBIN15-144.

appears to be slightly in favor of the leave-one-out potential, although the differences are not significant. As a point of comparison, we also measured the fit of the contact only potential (joint method), to illustrate that the difference between the joint and the leave-one-out methods is small compared to the differences observed between the alternative forms of statistical potential that we would like to empirically compare (see [26] for an evaluation of the relative contribution of each potential component to the fitness of the model).

Discussion

In a previous work [26], we defined a statistical framework for protein design, using the maximum likelihood principle, with the aim of devising statistical potentials to be used in phylogenetic studies. However, the optimization procedure we introduced at that time requires a MCMC protocol to cope with the proportionality constant entailed by the normalization of the probability over the sequence space. Here, we introduce a different likelihood, which we called leave-one-out, to optimize the potentials. A similar procedure was previously used by Kuhlman and Baker [27], but was not statistically assessed against alternative procedures. We found in this work that the joint and the leave-one-out potentials are virtually indistinguishable, both by direct comparison and by Bayes factor evaluation in a phylogenetic context.

We note that the optimal β for the SC^{l}_{β} model is not 0.5, as one may expect given the way our potentials were normalized (see eq. 3, 6 and 13). Several explanations can be proposed. First, strictly speaking, this expectation is valid under the joint procedure, and not under the leave-one-out procedure. But the very high similarity between the two resulting potentials, and the fact that a similar phenomenon ($\beta \neq 0.5$) can be observed also under a potential optimized using the joint method [3] do not favor this explanation. Alternatively, it may appear at first that this could be due to the fact that the underlying mutation model (the Q^{mut} matrix in eq. 13) was not explicitly taken into account when optimizing the potential (so that the chemical potentials implicitely include a mutational component), whereas our phylogenetic model does involve an explicit mutational process. In this sense, in the phylogenetic analysis, there is a potentially (partially) redundant modeling of mutational features, in having explicit parameters devoted to these, in combination with the use of the SC setting. This might explain the optimal value of β lower than 0.5. The phenomenon may also be the result of model violations, which are very likely to be present given the simple form of the potentials. Finally, it is also likely that the mutation pressure, or the selection strength (represented by β) is not the same for each protein. Accordingly, two possible improvements to the method can thus be proposed here: the first is to optimize the potential while allowing for different values of β for each protein or each family of protein. The second is to cluster proteins into classes, and optimize a potential specifically for each class.

Table 1: The natural logarithm of the Bayes factors.

	ADH23-254	CALM36-444	GLOBIN15-144	LYS25-134
SC^{c}_{β}	[74.748-75.032]	[149.819-149.929]	[57.953-58.135]	[11.5-11.968]
SC^{j}_{β}	[102.666-102.766]	[161.340-161.491]	[70.666-70.948]	[26.287-26.417]
SC^{l}_{β}	[102.977-103.115]	[158.679-158.858]	[72.485-72.872]	[29.545-29.852]
optimal β	[0.387-0.397]	[0.371-0.383]	[0.450-0.498]	[0.179-0.249]

Conclusion

Apart from these two possible improvements, many other directions of research should now be explored: alternative functional forms for the potential should be implemented and empirically tested. Several methods accounting for negative design, through the use of explicit decoys [18] such as the use of a normalized energy gap between a native structure and misfolded structures [32], or using variational methods [19], also deserve further investigation. The supervised learning described here depends on structure-sequence pairs. In the present case, we have used native pairs, but this could be relaxed by taking a set of structures (e.g. obtained by molecular dynamics) as the reference structure or by taking a set of homologous sequences instead of a unique sequence [33]. A more appealing method would consist in doing the optimization directly within the phylogenetic context. Importantly, the fact that the leave-one-out procedure is much faster than the joint method (in the present case, roughly by a factor 1,000), has obvious practical consequences, as it allows a much larger diversity of alternative models and methods to be tested.

Methods

Gradient descent

When performing a gradient descent, several methods can be used. We expose here the three gradient descent methods that we compared. In all cases, the method rely on a cyclical updating of parameter values, where, given the values of parameters at the m^{th} cycle, which we write as $\theta^{(m)}$, the update is given by:

$$\theta^{(m+1)} = \theta^{(m)} - \Delta\theta^{(m+1)}. \quad (15)$$

The increment, $\Delta\theta^{(m+1)}$, is conditional to the scoring function, that we simply denote in this part as $\omega(\theta^{(m)})$.

Fixed step gradient

This is the simplest form of the gradient descent. We write:

$$\Delta\theta^{(m+1)} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m)})}{\partial\theta}, \quad (16)$$

where δ_{grad} is the fixed step of the gradient descent. Even though this formalism is simple, the choice of the step is not trivial. Indeed, if the step is too large, the values of the potential will oscillate around the optimal values. Conversely, if the step is too small, the gradient descent will be too slow.

Inertial gradient

To reduce the optimization time, another method of gradient descent was developed, based on an analogy with the physical phenomenon of inertia.

$$\Delta\theta^{(m+1)} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m)})}{\partial\theta} + \delta_{iner} \cdot \Delta\theta^{(m)}. \quad (17)$$

δ_{iner} is the damping rate of the inertial component, $0 \leq \delta_{iner} < 1$. If $\delta_{iner} = 0$, eq. 17 reduces to the case of the simple gradient. In practice, we set δ_{iner} equal to 0.9.

However, there is a drawback when taking into account the previous variation of the parameters: when the directions of the gradient change, the inertial part of the gradient brings the parameters too far beyond the maximum. In addition, the gradient step δ_{grad} has to be small enough so that the values of the potential do not oscillate around the optimal values, as in the case of the fixed step gradient.

Controlled inertial gradient

To avoid these two drawbacks, we define here a controlled inertial gradient descent formalism. Specifically, let us define:

$$\Delta\theta^{*} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m-1)})}{\partial\theta} + \delta_{iner} \cdot \Delta\theta^{(m)}, \quad (18)$$

$$\Delta\theta^{\bullet} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m-1)})}{\partial\theta}. \quad (19)$$

The decision procedure can thus be described as follows (see additional file 5). First, we test if the addition of $\Delta\theta^{*}$ (derivative component and inertial component) to the actual values of parameters $\theta^{(m)}$ gives a higher likelihood than $\theta^{(m)}$. If it does, then the step corresponds to a classical step of the inertial gradient descent. Otherwise, the algorithm tests if the addition to $\theta^{(m)}$ of the derivative component ($\Delta\theta^{\bullet}$) only gives a higher likelihood than the actual values. If it does, the step corresponds to a classical gradient descent. Otherwise, we retry a simple gradient descent with a smaller δ_{grad}.

The above procedure has two advantages. The first is the speed-up offered by the inertial component, when its addition has a positive influence on the likelihood. The second advantage is that the last part of the algorithm automates the search for an optimal value of the steps, and avoids both oscillations of θ around the optimum, and a slow gradient descent.

Statistical potentials

We used the same statistical potential function as in our previous work [26]. The (pseudo) energy score consists of two terms:

$$E(s \mid c) = \sum_{1 \leq i \leq j \leq n} \Delta_{ij}\varepsilon_{s_i s_j} + \sum_{1 \leq i \leq n} \alpha_{s_i}^{v_i}. \quad (20)$$

The first term represents the contact free energy (defined between sidechain centers): $\Delta_{ij} = 1$ if i and j are closer than the cutoff distance (here 6.5 Å), and ε_{ab} represents the contact potential between amino acids a and b. The second term represents the accessibility free energy: v_i is the accessibility class of the site i and α_a^d is the solvent accessibility potential of the amino acid a when placed into the accessibility class d ($d = \{1..D\}$), where D is the number of accessibility classes.

We use the *random energy model* principle to approximate $F(s)$ (eq. 5), so that the inverse potential becomes:

$$G(s \mid c, \theta) = \sum_{1 \le i \le j \le n} \Delta_{ij} \varepsilon_{s_i s_j} + \sum_{1 \le i \le n} \alpha_{s_i}^{v_i} + \sum_{1 \le i \le n} \mu_{s_i}. \tag{21}$$

As in our previous work we fix the constraints:

$$\sum_{1 \le a \le 20} \mu_a = 0, \tag{22}$$

$$\sum_{1 \le a \le 20} \sum_{1 \le b \le 20} \varepsilon_{ab} = 0, \tag{23}$$

$$\sum_{1 \le a \le 20} \alpha_a^d = 0, d = \{1..D\}, \tag{24}$$

since $G(s|c, \theta)$ is invariant under the following transformations $\mu'_a = \mu_a + J_1$, $\varepsilon'_{ab} = \varepsilon_{ab} + J_2$ and $\alpha'^d_a = \alpha_a^d + J_3$. However, there is an additional lack of identifiability between a and μ, which can be resolved by including the chemical potentials in the accessibility terms. Indeed, the μ_a terms can be seen as an additive constant to each accessibility term for a given accessibility class (see additional file 6). In the present case, our final inverse potential is therefore:

$$G(s \mid c) = \sum_{1 \le i \le j \le n} \Delta_{ij} \varepsilon_{s_i s_j} + \sum_{1 \le i \le n} \alpha_{s_i}^{v_i}, \tag{25}$$

and our set of parameters for the statistical potential will thus consist of:

$$\theta = \{\varepsilon_{ab}, \alpha_a^d\}, \quad 1 \le a \le 20, \quad 1 \le b \le 20, \quad d = \{1..D\}. \tag{26}$$

Bayes factor evaluation

In a Bayesian statistical framework the method of choice for comparing models is to compute Bayes factors. The Bayes factor between two models is defined as the ratio of their respective marginal likelihood. The case $B(SC_0, SC_\beta^l) > 1$ (resp. $B(SC_0, SC_\beta^l) < 1$) is considered as an evidence in favor of (resp. against) the SC_β^l model. We write the Bayes factor between SC_0 and SC_β^l as:

$$B(SC_0, SC_\beta^1) = \frac{P(A|SC_\beta^1)}{P(A|SC_0)}, \tag{27}$$

where A corresponds to the data, composed by an alignment of coding nucleotide sequences and a topology and

$$P(A \mid SC_\beta^l) = \int_\theta P(A \mid \theta) P(\theta) d\theta. \tag{28}$$

Here we compute Bayes factors by thermodynamic integration (or *path sampling*) as described in [3]. The procedure consists in sampling along a continuous path between SC_0 and SC_β^l through a set of slight changes in the value of β. In fact, this procedure provides a complete curve representing the fit of the model as a function of β. Sampling from $\beta = 0$ to $\beta = \beta_{max}$ and from $\beta = \beta_{max}$ to $\beta = 0$ gives two different curves for the logarithm Bayes factor, which we used as an internal check of the reliability of the method (not shown).

Datasets

Optimization datasets

The datasets are made of proteins (structure-sequence pairs) culled from the PDB, with less than 25% of mutual sequence identity and a resolution better than 2 Å [34]. This sequence homology percentage and the size of the database avoid possible bias that could be induced by related proteins. To compare the joint and leave-one-out potentials, we used the dataset on which we previously estimated the joint potentials, DS_j. This dataset is made of 441 proteins and 98,155 sites [26]. We also consider a dataset DS_l (made of 3,363 proteins and 835,717 sites) which was split into two subsets: $DS1$ (1,691 proteins and 419,208 sites), and $DS2$ (1,672 proteins and 416,509 sites). To determine the accessibility classes, we first compute the solvent accessibility area using Naccess 2.1 [35] and partitioned the resulting values into classes [26].

Phylogenetic Datasets

The SC model was applied to 4 distinct multiple sequence alignments: GLOBIN15-144, LYSIN25-134, ADH23-254 and CALM33-444. GLOBIN15-144 is made of 15 vertebrates sequences of the β-globin gene (taken from the original dataset from [36]), with a protein structure defined by the

PDB file 4HHB and a tree topology estimated using Phylobayes 3.1c [37] (which is consistent with the tree topology described in [38]). LYSIN25-134 is made of 25 Abalone sperm lysin sequences [39], with a protein structure defined by the PDB file 1LYS and the tree topology previously defined by [39]. ADH23-254 is made of 23 alcohol dehydrogenase sequences taken form Drosophila [36], with a protein structure defined by the PDB file 1A4U and the tree topology previously defined by [36]. CALM36-444 is made of 36 calmodulin sequences taken from eukaryotes, with a protein structure defined by the PDB file 1CFD and the tree topology estimated using phyML [40] under the model JTT + F + Γ [41,42].

Authors' contributions

CB implemented the leave-one-out and gradient descent methods described here and performed the run of all the experiments. CLK implemented the data pre-processing methods. NR implemented the phylogenetic framework. NL set up the theoretical framework and directed the overall project. All the authors co-wrote the manuscript and approved the final manuscript.

Additional material

Additional file 1

Derivatives of the potential parameters.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S1.pdf]

Additional file 2

XY-comparison of the leave-one-out potentials estimated from two independent datasets: (a) and (b) two independent runs on DS*1 (X-axis) and* DS*2 (Y-axis) for contact and accessibility potentials respectivly.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S2.eps]

Additional file 3

Contact potentials and solvent accessibility potentials written in an alphabetical order.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S3.txt]

Additional file 4

Bubble plot of the solvant accessibility potential where we remove from each potential the corresponding natural logarithm frequency of the accessibility class.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S4.eps]

Additional file 5

Controlled inertial gradient algorithm.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S5.pdf]

Additional file 6

Inclusion of μ_a in the accessibility terms.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-227-S6.pdf]

Acknowledgements

The authors are grateful to the three anonymous referees for their useful comments on the manuscript. CB was financially supported by the french Centre National de la Recherche Scientifique (CNRS), the Région Languedoc-Roussillon and the Université de Montréal, CLK by NSERC, CIHR and the Université de Montréal, NR by NSERC, and NL by the Université de Montréal, NSERC and the CNRS.

References

1. Halpern AL, Bruno WJ: **Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies.** *Mol Biol Evol* 1998, **15(7):**910-917.
2. Yang Z, Nielsen R: **Mutation-Selection models of codon substitution and their use to estimate selective strengths on codon usage.** *Mol Biol Evol* 2008, **25(3):**568-579.
3. Rodrigue N, Kleinman CL, Philippe H, Lartillot N: **Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons.** *Molecular Biology and Evolution* 2009, **26(7):**1663-1676.
4. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL: **Protein evolution with dependence among codons due to tertiary structure.** *Molecular Biology and Evolution* 2003, **20(10):**1692-1704.
5. Rodrigue N, Lartillot N, Bryant D, Philippe H: **Site interdependence attributed to tertiary structure in protein evolution.** *Gene* 2005, **347(2):**207-217.
6. Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL: **Quantifying the impact of protein tertiary Structure on molecular evolution.** *Mol Biol Evol* 2007, **24(8):**1769-1782.
7. Parisi G, Echave J: **Structural constraints and emergence of sequence patterns in protein evolution.** *Mol Biol Evol* 2001, **18(5):**750-756.
8. Choi SC, Redelings BD, Thorne JL: **Basing population genetic inferences and models of molecular evolution upon desired stationary distribution of DNA or protein sequences.** *Phil trans R Soc B* 2008, **363:**3931-3939.
9. Anfinsen CB: **Principles that govern the folding of protein chains.** *Science* 1973, **181:**223-230.
10. Case D, A Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Crowley M, Walker RC, Zhang W, Merz KM, Wang B, Hayik S, Roitberg A, Seabra G, Kolossváry I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews DH, Seetin MG, Sagui C, Babin V, A KP: *AMBER 10* University of California, San Francisco; 2008.
11. MacKerel AD Jr, Brooks CL III, Nilsson L, Roux B, Won Y, Karplus M: **CHARMM: The Energy Function and Its Parameterization with an Overview of the Program.** In *The Encyclopedia of Computational Chemistry Volume 1*. Edited by: v Schleyer RP, et al. John Wiley & Sons: Chichester; 1998:271-277.
12. Miyazawa S, Jernigan RL: **Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation.** *Macromolecules* 1985, **18:**534-552.
13. Miyazawa S, Jernigan RL: **Residue-residue potentials with a favorable contact pair term and an unfavorable hight packing density term, for simulation and threading.** *Journal of molecular biology* 1996, **256(3):**623-644.

14. Sippl MJ: **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* 1993, **7:**473-501.
15. Solis AD, Rackovsky S: **Improvement of statistical potentials and threading score functions using information maximixation.** *Proteins* 2006, **62(4):**892-908.
16. Tozzini V: **Coarse-grained model for proteins.** *Current opinion in structural biology* 2005, **15:**144-150.
17. Seno F, Vendruscolo M, Maritan A, Banavar JR: **Optimal protein design procedures.** *Physical review letter* 1996, **77(9):**1901-1904.
18. Deutsch JM, Kurowski T: **New algorithm for protein design.** *Physical review letter* 1996, **76:**323-326.
19. Seno F, Micheletti M, Maritan A, Banavar JR: **Variational approach to protein design and extraction of interactional potentials.** *Physical review letter* 1998, **81:**2172-2175.
20. Rossi A, Maritan A, Micheletti C: **A novel iterative strategy for protein design.** *Journal of Chemical physics* 2000, **112(4):**2050-2055.
21. Rossi A, Micheletti C, Seno F, Maritan A: **A self-consistent knowledge-based approach to protein design.** *Biophysical journal* 2001, **80(1):**480-490.
22. Moult J: **Comparison of database potentials and molecular mechanics force fields.** *Current opinion in structural biology* 1997, **7(2):**194-1999.
23. Mendes J, Guerois R, Serrano L: **Energy estimation in protein design.** *Current opinion in structural biology* 2002, **12(4):**441-446.
24. Bowie JU, Luthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure.** *Science* 1991, **253(5016):**164-170.
25. Chiu TL, Goldstein RA: **Optimizing potentials for the inverse protein folding problem.** *Protein engineering* 1998, **11:**749-752.
26. Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot N: **A maximum likelihood framework for protein design.** *BMC Bioinformatics* 2006, **7:**326-343.
27. Kuhlman B, Baker D: **Native protein sequences are close to optimal for their structure.** *PNAS* 2000, **97(19):**10383-10388.
28. Shakhnovich E, Gutin A: **Engineering of stable and fast-folding sequences of model proteins.** *Proceedings Natl Academy of sciences USA* 1993, **90(15):**7195-7199.
29. Sun S, Brem R, Chan R, Dill K: **Designing amino acid sequences to fold with good hydrophobic cores.** *Protein Engineering* 1995, **8(12):**1205-1213.
30. Pande VS, Grosberg AY, Tanaka T: **Statistical mechanics of simple model of protein folding and design.** *Biophysical journal* 1997, **73(6):**3192-3210.
31. Muse SV, Gaut BS: **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions, with applications to cholorplast genome.** *Mol Biol Evol* 1994, **11:**715-724.
32. Bastolla U, Porto M, Roman HE, Vendruscolo M: **A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank.** *BMC Evolutionary Biology* 2006, **6:**43.
33. Panjkovich A, Melo F, Marti-Renom M: **Evolutionary potentials: structure specific knowledge-based potentials exploiting the evolutionary record of sequence homologs.** *Genome Biology* 2008, **9:**R68.
34. Wang G, Dunbrack RLJ: **PISCES: a protein sequence culling server.** *Bioinformatics* 2003, **19(12):**1589-1591.
35. Hubbard SJ, Thornton JM: **NACCESS.** Computer Program, Department of Biochemistry and Molecular Biology, University College London; 1993.
36. Yang Z, Nielsen R, Goldman N, K P: **Codon substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155:**431-449.
37. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3. A Bayesian software for phylogenetic reconstruction and molecular dating.** *Bioinformatics* 2009, **25(17):**2286-2288.
38. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS: **Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenics.** *Science* 2001, **294(5550):**2348-2351.
39. Yang Z, Swanson WJ, Vacquier VD: **Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites.** *Molecular Biology and Evolution* 2000, **17:**1446-1455.
40. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Systematic Biology* 2003, **52(5):**696-704.
41. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *CABIOS* 1992, **8:**275-282.
42. Yang Z: **Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites.** *Mol Biol Evol* 1993, **10:**1396-1401.