Inference of Hopfield-Potts patterns from covariation in protein families: calculation and statistical error bars

2013 J. Phys.: Conf. Ser. 473 012010

(http://iopscience.iop.org/1742-6596/473/1/012010)

Journal of Physics: Conference Series **473** (2013) 012010 doi:10.1088/1742-6596/473/1/012010

Inference of Hopfield-Potts patterns from covariation in protein families: calculation and statistical error bars

Simona Cocco[1], **Rémi Monasson**[2] **and Martin Weigt**[3]

[1] Laboratoire de Physique Statistique de l'Ecole Normale Supérieure - UMR 8550, associé au CNRS et à l'Université Pierre et Marie Curie, 24 rue Lhomond, 75005 Paris,France
[2] Laboratoire de Physique Théorique de l'Ecole Normale Supérieure - UMR 8549, associé au CNRS et à l'Université Pierre et Marie Curie, 24 rue Lhomond, 75005 Paris, France
[3] Laboratoire de Génomique des Microorganismes - UMR 7238, Université Pierre et Marie Curie, 15 rue de l'Ecole de Médecine, 75006 Paris, France

E-mail: `monasson@lpt.ens.fr`

Abstract. We consider the Hopfield-Potts model for the covariation between residues in protein families recently introduced in Cocco, Monasson, Weigt (2013). The patterns of the model are inferred from the data within a new gauge, more symmetric in the residues. We compute the statistical error bars on the pattern components. Results are illustrated on real data for a response regulator receiver domain (Pfam ID PF00072) family.

1. Introduction

A protein family is a set of protein sequences which are evolutionary related. While the amino-acid sequences show great diversity they generally give rise to a common fold. Extracting structural information about this fold from the statistical features of the distribution over the sequences is an important question [1]. Maximum entropy (MaxEnt) modelling [2, 3], which defines least constrained distributions capable of reproducing some statistical features are useful in this regard. When the statistical features of interest are local residue conservation and residue-residue correlations MaxEnt predicts that the sequences are extracted from the equilibrium distribution of a Potts model, whose interaction parameters have to be inferred from the data [4, 5]. This approach, combined with a mean-field approximation for the computation of the parameters, was successful to predict sites which are in contact in the 3-D structure of the protein [6], se also [7, 8, 9] for related approaches, and [10, 11, 12, 13, 14, 15, 16, 17, 18] for applications to use such sequence information to guide tertiary and quaternary structure prediction, and [16] for an example how predictions can be used to rationally design mutagenesis experiments influencing a protein's functionality.

However, given the length of the sequences, $L \sim$ few hundreds, and the number of possible values for each residue on a site, $q = 21$ (20 amino-acids and the gap symbol in the alignment), the number of statistical couplings to be inferred scales as $(Lq)^2$ and can reach millions. To drastically reduce the number of parameters in the model and avoids overfitting we have recently introduced the Hopfield-Potts model [19] for the study of covariation between residues in protein families. The Hopfield-Potts model is a special case of the Potts model in which the coupling

matrix between the residues has a low rank, p, a priori much smaller than Lq. The non-zero modes of the coupling matrix, called patterns, define specific directions in the sequence space. These patterns show some similarities with position-specific scoring matrices, defining ideal sequence motifs which real sequences in the protein family are likely to have, but instead of encoding independent-site amino-acid preferences, they include statistical couplings between sequence positions. Some of the motifs are enforced through 'attractive' patterns assigning positive statistical scores to the sequence aligned along them. In addition, in distinction to the original Hopfield model [20], the model also include 'repulsive' patterns, giving statistical negative scores to all sequences but the ones including a few strongly coevolving residues. Repulsive patterns are very useful, since they are often localized on sites in contact on the 3D fold, see [19] for more details. The Hopfield-Potts model thus offers an efficient way to infer contacts, with much less parameters than the usual Potts model.

In this paper we pursue the study of the Hopfield-Potts along two directions. As one and only one residue can be present on any given site, the Potts model has local gauge invariance, which was broken in a specific way in our previous publication [19]. Here we propose another choice for the gauge, which is more symmetric in the residues, and repeat the calculation of the most likely patterns given the protein multi-sequence alignment. In addition we show how the statistical error bars on the patterns can be computed within the statistical physics framework. Knowledge of the error bars allows us to derive a new criterion for pattern selection and for the number of patterns, p, to be considered. Our analytical results are then illustrated on real genomic data for one protein family.

The paper is organized as follows. In Section 2, we propose a brief reminder on the covariation analysis of protein families with the Maximum Entropy approach leading to the definition of a Potts model for the protein sequences in one family. The Hopfield-Potts model is introduced and the different gauges of the models are discussed in Section 3. In Section 4 we calculate the expressions for the most likely patterns. Error bars are computed in Section 5. Some applications to real protein data are shown in Section 6, while conclusions are given in Section 7. Readers not interested in the details of the statistical mechanics calculations can access the main results for the pattern values and their error bars in Sections 4.1 and 5.1.

2. Reminder on the statistical physics approach to covariation

2.1. Co-evolution and amino-acid correlations

The multi-sequence alignment (MSA) is denoted as $A = \{a_i^m | i = 1, ..., L,\ m = 1, ..., M\}$ with the index i running over the L columns in each row of the alignment (sites), and m over the M sequences, which are the rows in the MSA. The amino-acids a_i^m are assumed to be represented by natural numbers $1, ..., q$ with $q = 21$. This number includes the 20 standard amino acids, and the alignment gap.

In our approach, we do not use the data directly, but we summarize them by the amino-acid occupancies in single MSA columns and pairs of columns,

$$f_i(a) = \frac{1}{M}\sum_{m=1}^{M} \delta_{a,a_i^m} \ , \quad f_{ij}(a,b) = \frac{1}{M}\sum_{m=1}^{M} \delta_{a,a_i^m}\delta_{b,a_j^m} \ , \tag{1}$$

with $i, j = 1, ..., L$ and $a, b = 1, ...q$. Note that, to lower the effect of finite and dependent sampling of phylogenetically related biological sequences, in [6] a pseudocount and a reweighting scheme were used. The latter assigns lower the contribution of very close, almost repeated sequences in the above equations, and to increase the one of isolated ones. We describe how to include pseudocount and reweighting in Section 6.1.

If the two columns i and j were independently occupied by amino-acids, the joint distribution $f_{ij}(a, b)$ would factorize into the product $f_i(a)f_j(b)$. We therefore measure correlations of the

amino-acid occupancies of columns i and j via the covariance matrix

$$C_{ij}(a,b) = f_{ij}(a,b) - f_i(a)f_j(b) \ . \tag{2}$$

Note that finite sampling may introduce a spurious non-zero correlation between two residues, even if those residues do not covary.

2.2. Maximum entropy modeling

The presence of a strong correlation between residues does not entail that they are in direct contact on the 3D fold of the protein. The reason is the following [4, 5]: When i is in contact with j, and j is in contact with k, also i and k will show correlation. It is thus important to distinguish between *direct interaction* and *indirect correlation*, and to infer networks of direct couplings which generate the empirically observed correlations.

Disentangling direct and indirect correlations is only possible, if the statistics of full-length protein sequences is considered. It can be done by constructing a statistical model $P(a_1, ..., a_L)$ describing the probability of observing a particular amino-acid sequence $a_1, ..., a_L$. Due to the limited amount of available data, we require this model to reproduce empirical frequency counts for single MSA columns and column pairs,

$$f_i(a_i) = \sum_{\{a_k|k\neq i\}} P(a_1, ..., a_L) \ , \quad f_{ij}(a_i, a_j) = \sum_{\{a_k|k\neq i,j\}} P(a_1, ..., a_L) \ , \tag{3}$$

i.e. marginal distributions of $P(a_1, ..., a_L)$ are required to coincide with the empirical counts. Beyond this coherence, we aim at the *least constrained* statistical description. To this aim we apply the *maximum-entropy principle* [2, 3], which consists in maximizing the entropy $H[P] = -\sum_{a_1,...,a_L} P(a_1, ..., a_L) \log P(a_1, ..., a_L)$, under the constraints in Eqn. (3). If the constraints are imposed via Lagrange multipliers, we readily find the analytical form

$$P(a_1, ..., a_2) = \frac{1}{\mathcal{Z}(\{e_{ij}(a,b), h_i(a)\})} \exp\left\{\frac{1}{2}\sum_{i,j} e_{ij}(a_i, a_j) + \sum_i h_i(a_i)\right\} \ , \tag{4}$$

where $\mathcal{Z}$ is a normalization constant for P:

$$\mathcal{Z}(\{e_{ij}(a,b), h_i(a)\}) = \sum_{a_1,...,a_L} \exp\left\{\frac{1}{2}\sum_{i,j} e_{ij}(a_i, a_j) + \sum_i h_i(a_i)\right\} \ . \tag{5}$$

Thus, the maximum-entropy model takes the form of a (generalized) q-states Potts model. The parameters $e_{ij}(a,b)$ denote the direct couplings between MSA columns, and the $h_i(a)$ local fields (biases) in single sites. Values for these parameters have to be determined such that Eqs. (3) are satisfied. Applied to amino-acid sequence data and combined with a mean-field calculation of the couplings e, this approach was called *direct coupling analysis* (DCA) [21, 6].

2.3. A Bayesian perspective

The maximum-entropy approach of the last section can be rephrased in a Bayesian framework. Assume the model to be given by Eqn. (4), and assume the sequences in the alignment A to be independently sampled. We thus find the probability of the alignment to be

$$P\left[A|\{e_{ij}(a,b), h_i(a)\}\right] = \prod_{m=1}^{M} P(a_1^m, ..., a_L^m) \ . \tag{6}$$

Plugging in Eqn. (4) we find the log-likelihood of the MSA A,

$$\begin{aligned}\mathcal{L}\left[\{e_{ij}(a,b),h_i(a)\}|A\right] &= \frac{1}{M}\log P\left[A|\{e_{ij}(a,b),h_i(a)\}\right] \\ &= \frac{1}{2}\sum_{i,j}\sum_{a,b}e_{ij}(a,b)f_{ij}(a,b)+\sum_{i,a}h_i(a)f_i(a)-\log\mathcal{Z}(\{e_{ij}(a,b),h_i(a)\})\end{aligned} \tag{7}$$

One can readily see that the interaction and field parameters maximizing $\mathcal{L}$ fulfill Eqs. (3).

3. The Hopfield-Potts model

The Potts model in (4) is defined from $O\big((Lq)^2\big)$ parameters to be inferred from the data. To reduce the risk of overfitting we now introduce the Hopfield-Potts model, a special case of the Potts model, where the number of parameters to be inferred is much lower.

3.1. Definition of the model and dimensional reduction

The main idea is to express the matrix $e_{ij}(a,b)$ in terms of $p \ll Lq$ patterns $\{\xi_i^\mu(a), \mu = 1,...,p\}$, and to write

$$e_{ij}(a,b) = \frac{1}{L}\sum_{\mu=1}^{p}\xi_i^\mu(a)\xi_j^\mu(b) \tag{8}$$

with $i,j = 1,...,L$ being the site indices, and $a,b = 1,...,q$ being amino acids, or Potts states. Note that this matrix, for linearly independent patterns, has rank p, and it depends only on pLq parameters. By defining the log-score of a sequence $(a_1,...,a_L)$ for one pattern ξ^μ as

$$S(a_1,...,a_L|\xi^\mu) = \left[\sum_{i=1}^{L}\xi_i^\mu(a_i)\right]^2 . \tag{9}$$

the probability (4) of that amino-acid sequence can be written as an the exponential of the sum of log-scores along a number p of patterns through :

$$P(a_1,\ldots,a_L) = \frac{1}{\mathcal{Z}}\exp\left\{\frac{1}{2L}\sum_{\mu=1}^{p}S(a_1,\ldots,a_L|\xi^\mu)\right\} \tag{10}$$

An important remark is that patterns are not necessarily real valued. We will find both real and imaginary-valued patterns. The first case correspond to *attractive patterns*, which lead to a positive sign on the log-score $S(a_1,...,a_L|\xi^\mu)$, the second case to *repulsive patterns* which lead to a negative sign on the log-score $S(a_1,...,a_L|\xi^\mu)$. Consequently for the probability $P(a_1,...,a_L)$ to be large, the absolute value of the scores $S(a_1,...,a_L|\boldsymbol{\xi})$ for attractive patterns must be large, whereas for repulsive patterns it must be small (close to zero).

Following closely the derivations for the Ising-case in [22], we will show how mean-field theory can be used to derive an expression for the patterns $\xi_i^\mu(a)$ in terms of the empirical 1- and 2-point frequencies (1).

3.2. Gauge fixing for the Potts states

Due to the presence on each site i of a single amino-acid a in the Potts model defined in Eqn. (4) changes of the couplings $e_{ij}(a,b) \rightarrow e_{ij}(a,b) + g_i(a)$ can be compensated by corresponding changes of the field $h_i(a) \rightarrow h_i(a) - g_i(a)$. In a previous work [19] we removed this gauge invariance by specializing to couplings matrices $e_{ij}(a,b)$ where the q^{th} row ($a = q$) and column

$(b = q)$ are equal to zero for every pairs of site $i < j$. Within the Hopfield-Potts model this amounted to enforce $\xi_i^\mu(q) = 0$ for all sites i and patterns μ.

Hereafter we consider another, natural choice for the gauge, expressed as a linear constraint over the pattern components:

$$\sum_a f_i(a)\ \xi_i^\mu(a) = 0\ , \tag{11}$$

for all sites i and all patterns μ. This gauge is more symmetric over the Potts states than the gauge adopted in [19].

3.3. Gauge fixing in the pattern space

It is important to realize that Hopfield's expression for the couplings in terms of patterns introduces another gauge, in the pattern space. The representation given in Eqn. (8) is indeed not unique. The expression is invariant with respect to a rotation in pattern space, *i.e.* with respect to multiplying all patterns with an orthogonal $p \times p$-matrix $\mathcal{O}$. More precisely, it is easy to check that the couplings $e_{ij}(a, b)$ defined by Eqn. (8) are unchanged under the transformation

$$\xi_i^\mu(a) \to \sum_{\nu=1}^{p} \mathcal{O}^{\mu,\nu}\ \xi_i^\nu(a)\ , \tag{12}$$

for fixed a and i.

To have a unique expression for the patterns, we have to eliminate this arbitrariness in analogy to the gauge fixation discussed in the previous Section for the Potts states. A simple and convenient prescription is to impose that the patterns are orthogonal for the following dot product:

$$\sum_{i,a} f_i(a)\ \xi_i^\mu(a)\ \xi_i^\nu(a) = 0 \qquad (\mu \neq \nu)\ . \tag{13}$$

The number of distinct orthogonality conditions, $p(p-1)/2$, is equal to the number of generators of the p-dimensional orthogonal group[1], entailing that the solution to (13) is generally unique (up to a discrete permutation symmetry of the patterns).

4. Calculation of the Hopfield-Potts patterns

4.1. Expressions of the most likely patterns

We define from the multiple sequence alignment the (Lq)–dimensional Pearson correlation matrix,

$$\Gamma_{ij}(a, b) = \frac{c_{ij}(a, b)}{\sqrt{f_i(a)\, f_j(b)}} \tag{14}$$

These matrix has (at least) L zero eigenvalues, because, on each site i, the probabilities of the $q = 21$ amino-acid or gap symbols sum up to 1:

$$\sum_{a=1}^{q} f_i(a) = 1 \Rightarrow\ \sum_{b=1}^{q} \Gamma_{ij}(a, b)\ \sqrt{f_j(b)} = 0\ , \forall\ i, a. \tag{15}$$

We look for the non-zero eigenmodes:

$$\sum_{j,b} \Gamma_{ij}(a, b)\ v_{jb}^\mu = \lambda^\mu\ v_{ia}^\mu\ , \tag{16}$$

[1] Strictly speaking the gauge group is not the orthogonal group $O(p)$ but the indefinite orthogonal group $O(p_+, p_-)$ where p_+ and p_- are, respectively, the numbers of real- and imaginary-valued patterns. The number of generators of this group is nevertheless equal to $p(p-1)/2$, with $p = p_+ + p_-$.

where the eigenvectors define an orthonormal basis of the $L(q-1)$–dimensional space orthogonal to the null space of Γ:

$$\sum_a v_{ia}^\mu \sqrt{f_i(a)} = 0 \quad \forall i \ , \qquad \sum_{i,a} v_{ia}^\mu \, v_{ia}^\nu = 0 \quad \forall \mu \neq \nu \ , \qquad \sum_{i,a} (v_{ia}^\mu)^2 = L \quad \forall \mu \ . \tag{17}$$

It is easy to check that

$$\sum_{\mu=1}^{L(q-1)} \lambda^\mu = \sum_{i=1}^{L} \sum_{a=1}^{q} \Gamma_{ii}(a,a) = \sum_{i=1}^{L} \sum_{a=1}^{q} \big(1 - f_i(a)\big) = L(q-1) \ . \tag{18}$$

Therefore, the average value of the eigenvalues λ^μ of the matrix Γ (after having removed the L zero modes) is equal to unity.

The statistical mechanics of the inverse Hopfield-Potts model, presented in the next Section, allows us to derive the following expressions for the most likely patterns given the data, *i.e.* the MSA A,

$$\xi_i^\mu(a) = \sqrt{1 - \frac{1}{\lambda^\mu}} \, \frac{v_{i,a}^\mu}{\sqrt{f_i(a)}} \ . \tag{19}$$

By virtue of (17) the gauge conditions (11) are fulfilled. The pattern $\xi^\mu(a)$ is therefore simply written as a function of the eigenvector of the Pearson correlation matrix v^μ with our choice of the gauge. Expression (19) for the patterns in turn defines the Lq–dimensional coupling matrix e through Eqn. (8). Note that the prefactor $\sqrt{1-1/\lambda_\mu}$ is real for $\lambda_\mu > 1$, vanishes for $\lambda_\mu = 1$, and becomes imaginary for $\lambda_\mu < 1$. According to the discussion above, large eigenvalues (> 1) therefore correspond to attractive patterns, and small eigenvalues (< 1) to repulsive patterns. We will discuss how eigenvalues should be selected in Section 4.3.

4.2. Statistical mechanics derivation

We now present the detailed derivation of the results presented above. Our aim is to calculate the partition function $\mathcal{Z}(\{e_{ij}(a,b), h_i(a)\})$, defined in Eqn. (5), in the mean-field approximation. The sum over the sequence can be performed using Hubbard-Stratonovich transformations for each μ,

$$\begin{aligned} \mathcal{Z} &= \int \prod_{\mu=1}^{p} \frac{dx^\mu}{\sqrt{2\pi/L}} \sum_{\{a_i | i=1,\ldots,L\}} \exp\left\{ -\frac{L}{2} \sum_{\mu=1}^{p} (x^\mu)^2 + \sum_{i=1}^{L} \left(h_i(a_i) + \sum_{\mu=1}^{p} x^\mu \xi_i^\mu(a_i) \right) \right\} \qquad (20) \\ &= \int \prod_{\mu=1}^{p} \frac{dx^\mu}{\sqrt{2\pi/L}} \exp\left\{ -\frac{L}{2} \sum_{\mu=1}^{p} (x^\mu)^2 + \sum_{i=1}^{L} \log\left(\sum_{a=1}^{q} \exp\left[h_i(a) + \sum_{\mu=1}^{p} x^\mu \xi_i^\mu(a) \right] \right) \right\} \ . \end{aligned}$$

The leading contribution of the x^μ-integrations can be determined by the saddle-point approximation. The saddle points x_0^μ satisfy the equations

$$x_0^\mu = \frac{1}{L} \sum_{ia} \xi_i^\mu(a) T_i(a) \quad \text{with} \quad T_i(a) = \frac{\exp\left[h_i(a) + \sum_{\mu=1}^{p} x_0^\mu \xi_i^\mu(a) \right]}{\sum_{b=1}^{q} \exp\left[h_i(b) + \sum_{\mu=1}^{p} x_0^\mu \xi_i^\mu(b) \right]} \ . \tag{21}$$

As the next step, we determine the Gaussian corrections to this saddle point. To this aim, we need to calculate the second derivatives (in $x^\mu = x_0^\mu$):

$$\frac{\partial^2}{\partial x^\mu \partial x^\nu} \sum_i \log\left(\sum_a \exp\left[h_i(a) + \sum_{\mu=1}^{p} x^\mu \xi_i^\mu(a) \right] \right) = \sum_{i,a,b} \xi_i^\mu(a) \xi_i^\nu(b) \left[T_i(a) \delta_{a,b} - T_i(a) T_i(b) \right] \ . \tag{22}$$

Carrying out the Gaussian integrations we find the following estimate for the log-likelihood of the MSA, see Eqn. (7),

$$\begin{aligned}\mathcal{L}[\{\xi_i^\mu(a), h_i(a)\}|A] &= \frac{1}{2L}\sum_{\mu,i,j,a,b}\xi_i^\mu(a)\,\xi_j^\mu(b)\,f_{ij}(a,b)+\sum_{ia}h_i(a)\,f_i(a)+\frac{1}{2}\log\det G \\ &+ \frac{L}{2}\sum_{\mu=1}^{p}(x_0^\mu)^2-\sum_{i=1}^{L}\log\left(\sum_{a=1}^{q}\exp\left[h_i(a)+\sum_{\mu=1}^{p}x_0^\mu\,\xi_i^\mu(a)\right]\right)\ , \qquad (23)\end{aligned}$$

up to an irrelevant additive constant. The entries of the matrix G are defined through

$$G_{\mu\nu}=\delta_{\mu\nu}-\frac{1}{L}\sum_i\sum_{a,b}\xi_i^\mu(a)\,\xi_i^\nu(b)\,[T_i(a)\delta_{ab}-T_i(a)T_i(b)]\ . \qquad (24)$$

Maximizing $\mathcal{L}$ over the fields $h_i(a)$ we readily obtain

$$T_i(a)=f_i(a)\ , \qquad (25)$$

where the $T_i(a)$'s have been defined in (21). According to the Potts gauge (11) and the saddle-point equations (21) we have $x_0^\mu=0$ for all μ's. Furthermore the pattern gauge condition (13) implies that G defined above is a diagonal matrix. After some elementary algebra we are left with the simpler expression for the log-likelihood of the data,

$$\begin{aligned}\mathcal{L}[\{\xi_i^\mu(a)\}|A] &= \sum_{ia}f_i(a)\log f_i(a)+\frac{1}{2L}\sum_{\mu,i,j,a,b}\xi_i^\mu(a)\,\xi_j^\mu(b)\,f_{ij}(a,b) \\ &+ \frac{1}{2}\sum_\mu\log\left(1-\frac{1}{L}\sum_{i,a}f_i(a)\,\xi_i^\mu(a)^2\right)\ . \qquad (26)\end{aligned}$$

We find the trivial result that, for $p=0$ (no pattern and, thus, no coupling between the sites), the likelihood is the negative of the sum of all single-column entropies. In the presence of patterns ($p\geq 1$) the values of the latter are found by optimizing $\mathcal{L}$. We write

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\xi_i^\mu(a)} &= \sum_{j,b}f_{ij}(a,b)\,\xi_j^\mu(b)-\frac{1}{G^\mu}\,f_i(a)\,\xi_i^\mu(a) \\ &= \sqrt{f_i(a)}\left(\sum_{j,b}\Gamma_{ij}(a,b)\sqrt{f_j(b)}\,\xi_j^\mu(b)-\frac{1}{G^\mu}\,\sqrt{f_i(a)}\,\xi_i^\mu(a)\right)\ , \qquad (27)\end{aligned}$$

where the entries of matrix Γ have been defined in (14), and

$$G^\mu=1-\frac{1}{L}\sum_{i,a}f_i(a)\,\xi_i^\mu(a)^2\ . \qquad (28)$$

As a consequence the optimal patterns $\xi_i^\mu(a)$ are in one-to-one correspondence with the eigenvectors v_{ia}^μ of Γ. More precisely, for all μ,i,a, $\sqrt{f_i(a)}\,\xi_i^\mu(a)=C^\mu\,v_{ia}^\mu$. The proportionality constant C^μ is such that the eigenvalue λ^μ coincides with $1/G^\mu$. According to Eqn. (28) we find

$$G^\mu=\frac{1}{\lambda^\mu}=1-(C^\mu)^2\ . \qquad (29)$$

Hence C^μ is real-valued if $\lambda^\mu>1$ and imaginary-valued if $\lambda^\mu<1$. The resulting expression for the patterns is given in (19). We check a posteriori that the patterns fulfill the gauge conditions (11) and (13).

4.3. Pattern selection: maximum likelihood criterion
The final expression for the maximal likelihood is therefore

$$\mathcal{L}[A] = \sum_{i,a} f_i(a) \log f_i(a) + \sum_{\mu=1}^{p} \Delta\mathcal{L}(\lambda_\mu) \quad \text{with} \quad \Delta\mathcal{L}(\lambda) = \frac{1}{2}\big(\lambda - 1 - \log\lambda\big) \ . \tag{30}$$

If the number p of patterns is imposed we have to select the p eigenvalues λ_μ giving the largest contributions $\Delta\mathcal{L}(\lambda_\mu)$ to the likelihood. Large contributions arrive from both the largest and the smallest eigenvalues, whereas eigenvalues close to one contribute little. We thus have to determine a threshold value θ for the log-likelihood such that there are exactly p patterns with larger, and $L(q-1)-p$ patterns with smaller (and thus neglected) contributions to the likelihood. This amounts to finding the two positive real solutions $\ell_\pm$ ($\ell_- < \ell_+$) of the equation

$$\Delta\mathcal{L}(\ell_\pm) = \theta \tag{31}$$

and to include only patterns with $\lambda_\mu < \ell_-$ or $\lambda_\mu > \ell_+$. We will call hereafter p_+ and, respectively, p_- the number of eigenvalues λ_μ larger than ℓ_+, respectively smaller than ℓ_-. We obviously have $p_+ + p_- = p$.

5. Statistical error bars on the inferred patterns

5.1. Expressions for the error bars
Let $\mathcal{H}$ be the $(p+1)L(q-1)$–dimensional Hessian matrix of minus the log-likelihood (26), computed in the inferred patterns and fields,

$$\mathcal{H} \equiv \begin{pmatrix} \mathcal{H}_{ij}^{\mu,\nu}(a,b) & \mathcal{H}_{ij}^{\mu,0}(a,b) \\ \mathcal{H}_{ij}^{0,\nu}(a,b) & \mathcal{H}_{ij}^{0,0}(a,b) \end{pmatrix} = \begin{pmatrix} -\frac{\partial^2\mathcal{L}}{\partial\xi_i^\mu(a)\partial\xi_j^\nu(b)} & -\frac{\partial^2\mathcal{L}}{\partial\xi_i^\mu(a)\partial h_j(b)} \\ -\frac{\partial^2\mathcal{L}}{\partial h_i(a)\partial\xi_j^\nu(b)} & -\frac{\partial^2\mathcal{L}}{\partial h_i(a)\partial h_j(b)} \end{pmatrix} . \tag{32}$$

Asymptotically, for a large number M of sequences in the data base, the posterior probability of patterns and fields is Gaussian, with covariance matrix equal to the inverse of M $\mathcal{H}$. Let us call $\delta\xi_i^\mu(a)$ and $\delta h_i(a)$ the deviations of the patterns and of the fields with respect to their most likely values. Then those deviations are normally distributed, with zero averages, and covariances given by

$$\begin{aligned} \langle \delta\xi_i^\mu(a)\ \delta\xi_j^\nu(b)\rangle &= \frac{1}{M}(\mathcal{H}^{-1})_{ij}^{\mu,\nu}(a,b) \ , \quad \langle \delta h_i(a)\ \delta h_j(b)\rangle = \frac{1}{M}(\mathcal{H}^{-1})_{ij}^{0,0}(a,b) \ , \\ \langle \delta\xi_i^\mu(a)\ \delta h_j(b)\rangle &= \frac{1}{M}(\mathcal{H}^{-1})_{ij}^{\mu,0}(a,b) \ , \end{aligned} \tag{33}$$

where $\langle\cdot\rangle$ denotes the average over the posterior distribution. In the following we call error bars on the patterns and on the fields the standard deviations of their components. For instance the error bar on the component (i,a) of pattern μ is given by $\sqrt{\langle\delta\xi_i^\mu(a)^2\rangle} = \sqrt{\frac{1}{M}(\mathcal{H}^{-1})_{ii}^{\mu,\mu}(a,a)}$.

We show in the next Section that the error bars on the pattern components are given by

$$\langle\delta\xi_i^\mu(a)^2\rangle = \frac{1}{M}\left[\frac{(v_{ia}^\mu)^2}{2\lambda_\mu(\lambda_\mu-1)} + \sum_{A\le p\ \&\ A\neq\mu} \frac{(v_{ia}^A)^2(\lambda_\mu-1)\lambda_A}{(\lambda_\mu|\lambda_A-1|+\lambda_A|\lambda_\mu-1|)^2} + \sum_{A>p}\frac{(v_{ia}^A)^2}{\lambda_\mu-\lambda_A}\right] . \tag{34}$$

Note that the second sum in (34) runs over all the eigenvectors of Γ with eigenvalues different from zero and smaller than the ones corresponding to the inferred patterns ($A > p$), while the first sum runs over the top p eigenvectors (except eigenvector μ).

The above expression is correct when all selected patterns are attractive. Assume now that, say, $p_- \geq 1$ repulsive patterns (corresponding to the smallest p_- eigenvalues) and p_+ attractive patterns are retained. Expression (34) is still valid for an attractive pattern, *i.e.* such that $\mu \leq p_+$, upon changing condition ($A \leq p$ & $A \neq \mu$) into ($A \leq p_+$ & $A \neq \mu$, or $A \geq L(q-1)-p_-$) and condition ($A > p$) into ($p_+ < A < L(q-1) - p_-$). For a repulsive pattern, *i.e.* such that $\mu > L(q-1) - p_-$, formula (34) holds upon changing condition $A \leq p$ into ($A \leq p_+$, or $A \geq L(q-1) - p_-$ & $A \neq \mu$) and condition ($A > p$) into ($p_+ < A < L(q-1) - p_-$). Note that for repulsive patterns the components $\xi_i^\mu(a)$ are imaginary, and the covariances $\langle \delta\xi_i^\mu(a)^2 \rangle$ are negative real numbers.

5.2. Calculation of the Hessian matrix of the log-likelihood

The calculation of the inverse of $\mathcal{H}$ was done in [22] for the Ising case ($q = 2$). We now briefly review how it can be extended to the Potts case, and limit ourselves to the calculation of the error bars on the pattern components only. We start by differentiating the log-likelihood (23) twice with respect to the patterns ξ:

$$\begin{aligned} \mathcal{H}_{ij}^{\mu,\nu}(a,b) &= \frac{\delta^{\mu\nu}}{L}\Big[- C_{ij}(a,b) + \lambda_\mu \delta_{ij}\big(f_i(a)\delta^{ab} - f_i(a)f_i(b)\big) + \frac{\lambda_\mu}{L} f_i(a) f_j(b) \sum_\alpha \lambda_\alpha\, \xi_i^\alpha(a) \xi_j^\alpha(b) \Big] \\ &+ \frac{1}{L^2} \lambda_\mu\, \lambda_\nu\, \xi_i^\nu(a)\, \xi_j^\mu(b)\ . \end{aligned} \tag{35}$$

The sum over the index α above runs over the p selected patterns. We define a set of $p(q-1)L$ vectors, $\vec{e}^{(A,\omega)}$, with A taking values from 1 to $L(q-1)$ and ω from 1 to p, and whose components on the canonical basis are given by

$$\big(\vec{e}^{(A,\omega)}\big)_{i,a}^{\mu} = \delta_{\mu,\omega}\, \frac{v_{i,a}^A}{\sqrt{f_i(a)}}\ . \tag{36}$$

The vectors $\vec{e}^{(A,\omega)}$ form a basis of the subspace orthogonal to the L-dimensional null subspace of Γ. We order the vectors such that the first p vectors, with $A = 1, 2, \ldots, p$ corresponds to the p selected patterns, see Eqn. 19. In the new basis, the entries of the matrix $\tilde{\mathcal{H}}$ defined in (35) are given by:

$$\tilde{\mathcal{H}}_{(A,\omega),(B,\tau)} = \delta_{\omega,\tau}\delta_{A,B}\Big[- \lambda_A + \lambda_\omega + \lambda_\omega(\lambda_A - 1)\delta_{A\leq p}\Big] + \sqrt{\lambda_\omega\lambda_\tau(\lambda_\omega - 1)(\lambda_\tau - 1)}\delta_{\omega,B}\delta_{\tau,A}\ . \tag{37}$$

Here, we have used $\delta_{A\leq p} = 1$ is A is smaller or equal to p, and 0 otherwise. Matrix $\tilde{\mathcal{H}}$ can now be easily diagonalized:

- For $A > p$, $\tilde{\mathcal{H}}_{(A,\omega),(B,\tau)} = (\lambda_\omega - \lambda_A)\delta_{\omega,\tau}\delta_{A,B}$, which is diagonal. Hence, we have already found $((q-1)L - p) \times p$ eigenvalues $\lambda_\omega - \lambda_A$, with $A = p+1, p+2, \ldots, p(q-1)L$ and $\omega = 1, 2, \ldots, p$.
- For $1 \leq A \leq p$, $\tilde{\mathcal{H}}_{(A,\omega),(B,\tau)} = 2\lambda_\tau(\lambda_\tau - 1)\delta_{\omega,\tau}\delta_{A,B}$, which is diagonal. Hence, we have p eigenvalues $2\lambda_\mu(\lambda_\mu - 1)$, with $\mu = 1, \ldots, p$.
- There are $\frac{1}{2}p(p-1)$ eigenvalues $2\lambda_\mu\lambda_\nu - \lambda_\mu - \lambda_\nu$, with $1 \leq \mu < \nu \leq p$. The corresponding (non-normalized) eigenvectors are $\delta_{B,\mu}\delta_{\tau,\nu}\sqrt{1 - 1/\lambda_\nu} + \delta_{B,\nu}\delta_{\tau,\mu}\sqrt{1 - 1/\lambda_\mu}$.
- The remaining $\frac{1}{2}p(p-1)$ eigenvalues vanish. They correspond to rotations of the patterns (in the μ-space) which leave the couplings and the log-likelihood unchanged. Their number coincides with the number of generators of the p-dimensional rotation group, see Section 3.3.

We may now write the expression for the pseudo-inverse of $\tilde{\mathcal{H}}$ in the $\vec{e}$–basis,

$$\begin{aligned}[\tilde{\mathcal{H}}^{-1}]_{(A,\omega),(B,\tau)} &= \frac{\delta^{\omega,\tau}\delta^{A,B}\delta_{A>p}}{\lambda_\omega-\lambda_A}+\delta^{\omega,\tau}\delta^{A,B}(1-\delta^{A,\omega})\delta_{A\leq p}\frac{(\lambda_\omega-1)\lambda_A}{(2\lambda_\omega\lambda_A-\lambda_A-\lambda_\omega)^2} \\ &+ \frac{\delta^{\omega,\tau}\delta^{\omega,A}\delta^{\omega,B}}{2\lambda_\omega(\lambda_\omega-1)}+\delta^{\tau,A}\delta^{\omega,B}(1-\delta^{\tau,\omega})\frac{\sqrt{\lambda_\omega\lambda_\tau(\lambda_\omega-1)(\lambda_\tau-1)}}{(2\lambda_\omega\lambda_\tau-\lambda_\omega-\lambda_\tau)^2}\ .\end{aligned} \tag{38}$$

Using expression (33) for the covariance of the pattern fluctuations we obtain

$$\langle\delta\xi_i^\mu(a)\ \delta\xi_j^\nu(b)\rangle=\frac{L}{M}\ [\tilde{H}^{-1}]_{ij}^{\mu,\nu}(a,b)=\frac{L}{M}\sum_{A,\omega,B,\tau}[\tilde{\mathcal{H}}^{-1}]_{(A,\omega),(B,\tau)}\big(\vec{e}^{\,(A,\omega)}\big)_{i,a}^{\mu}\big(\vec{e}^{\,(B,\tau)}\big)_{j,b}^{\nu}\ . \tag{39}$$

The resulting expression for the diagonal elements of the covariance matrix is given in Eqn. (34).

5.3. Pattern selection: Bayesian criterion

Knowledge of the uncertainties over the patterns allows us to define a Bayesian criterion for pattern selection. Informally speaking, patterns whose components have strong deviations around their most likely values, that is, of the order of the pattern components themselves cannot be considered as reliable and should be discarded. Therefore, for each pattern $\boldsymbol{\xi}$, we consider the ratio of the squared fluctuations to the squared norm of the pattern,

$$\rho(\boldsymbol{\xi})=\frac{\sum_{ia}\langle\big(\delta\xi_i(a)\big)^2\rangle}{\sum_{ia}\big(\xi_i(a)\big)^2}\ . \tag{40}$$

This ratio is positive for attractive and for repulsive patterns. We will decide that the pattern is reliable if the ratio ρ is smaller than some arbitrary error threshold, say, 1 or 2/3 [22]. Exemplary results for one protein family (response regulator domain Pfam ID PF00014) are given in Supplementary Information of [19], Fig. 13. Note that the error bars depend on the error threshold itself, smaller error thresholds lead to increased errors of the selected patterns. As a consequence, pattern selection according to the uncertainty of patterns is a self-consistent criterion, which can be solved in a iterative way.

5.4. Error bars on the couplings

We now turn to the derivation of the error bars over the couplings, see Eqn. (8),

$$e_{ij}(a,b)=\frac{1}{L}\sum_\mu\xi_i^\mu(a)\,\xi_j^\mu(b)\ , \tag{41}$$

where the index μ takes the values $1,2,\ldots,p_+$ and $L(q-1)-p_-,...,L(q-1)$. Approximating the distribution of the patterns as a Gaussian law with covariance matrix $\mathcal{H}^{-1)}/M$, centered in the inferred patterns $\boldsymbol{\xi}^\mu$, we find the variances of the couplings:

$$\begin{aligned}\langle e_{ij}(a,b)^2\rangle-\langle e_{ij}(a,b)\rangle^2 &= \frac{1}{ML^2}\sum_{\mu,\nu}\Big\{\xi_i^\mu(a)\,\xi_i^\nu(a)(\mathcal{H}^{-1})_{jb,jb}^{\mu\nu}+\xi_j^\mu(b)\,\xi_j^\nu(b)(\mathcal{H}^{-1)})_{ia,ia}^{\mu\nu} \\ &+ \xi_i^\mu(a)\,\xi_j^\nu(b)(\mathcal{H}^{-1})_{ia,jb}^{\mu\nu}+\xi_i^\nu(a)\,\xi_j^\mu(b)(\mathcal{H}^{-1})_{ia,jb}^{\mu\nu} \\ &+ \frac{1}{M}(\mathcal{H}^{-1})_{ia,jb}^{\mu\nu}(\mathcal{H}^{-1})_{jb,ia}^{\mu\nu}+\frac{1}{M}(\mathcal{H}^{-1})_{ia,ia}^{\mu\nu}(\mathcal{H}^{-1})_{jb,jb}^{\mu\nu}\Big\}\ .\end{aligned} \tag{42}$$

Notice that the mean values of the couplings,

$$\langle e_{ij}(a,b)\rangle \;=\; \frac{1}{L}\sum_{\mu}\left(\xi_i^{\mu}(a)\,\xi_j^{\mu}(b)+\frac{1}{M}(\mathcal{H}^{-1})^{\mu\mu}_{ia,jb}\right)\,, \tag{43}$$

are slightly shifted with respect to their typical (most likely) values given by Eqn. (8). The relative error bar on the couplings can be defined through

$$\epsilon=\frac{\sum_{ij,ab}(\langle e_{ij}(a,b)^2\rangle-\langle e_{ij}(a,b)\rangle^2)}{\sum_{ij,ab}\langle e_{ij}(a,b)\rangle^2}\,. \tag{44}$$

The error ϵ is a function of the number p of retained patterns, or, equivalently, of the thresholds $\ell_\pm$ over the selected small/large eigenvalues.

6. Application to protein data

In [19], the Hopfield-Potts approach in its maximum-likelihood formulation has been applied to a number of protein families, in order to establish the connection between the theoretical inference approach, and the information contained in multiple-sequence alignment of large protein families. Here we are going to summarize some of the findings of [19], and we are going to compare them to the Bayesian approach, where patterns are selected according to their relative errors. We will do this concentrating on one single protein family, more specifically the *response regulator* domain (Pfam ID PF00072 [23]) which is given by an alignment of $M = 62,074$ sequences of length $L = 112$. The X-ray crystal structure used to test our results has PDB ID 1nxw [24].

Using this example, we will describe in the following what has to be considered when working with real protein sequence data, and what are the characteristics and the biological meaning of the selected patterns

6.1. Data preprocessing

The theoretical setup in the first part of this article assumes the sequence data to be identically and independently distributed with respect to some statistical model, which is a priori unknown but shall be reconstructed from the data. Real sequence data are, however, phylogenetically related. In addition they are not frequent enough to well sample rare amino-acid combinations, which results in a rank-deficient covariance matrix (beyond the L zero eigenvalues due to the non-independence of the frequency count of $a = q$ from the values for $a = 1, ..., q-1$, cf. the discussion about the gauge invariance of the model). To be able to provide good inference results we follow two heuristic steps of data preprocessing [6]:

- *Pseudocount regularization:* The maximum rank $L(q-1)$ of the matrix is re-established by introducing a so-called pseudocount ν, which formally corresponds to the addition of the (the average over) completely random amino-acid sequences to the MSA.
- *Reweighting:* To reduce the bias due to uneven sampling, we assign a reduced statistical weight to sequences in densely sampled regions in sequence space. More precisely, for each sequence $m \in \{1, ..., M\}$ a weight

  $$w_m = ||\,\{n \in \{1, ..., M\} \mid d_H(n, m) < xL\}\,||^{-1} \tag{45}$$

  is calculated which equals the inverse of the number of sequences n within Hamming distance $d_H(n, m) < xL$ from m, with x being an arbitrary but fixed number from the interval $(0, 1)$. Here, $d_H(n, m)$ denotes the Hamming distance between rows n and m of the MSA. The total weight

  $$M_{eff} = \sum_{m=1}^{M} w_m \tag{46}$$

can be seen as the effective number of independent sequences, and it is used in evaluating the error bars instead of the original sequence number M.

With these two modifications, frequency counts become

$$\begin{aligned} f_i(a) &= \frac{1}{M_{eff}+\nu}\left[\frac{\nu}{q}+\sum_{m=1}^{M} w_m\ \delta_{a,a_i^m}\right] \\ f_{ij}(a,b) &= \frac{1}{M_{eff}+\nu}\left[\frac{\nu}{q^2}+\sum_{m=1}^{M} w_m\ \delta_{a,a_i^m}\ \delta_{b,a_j^m}\right] \end{aligned} \tag{47}$$

Values $\nu = M_{eff}$ and $x = 0.2$ were found to work well in mean-field DCA across many families [6], we use the same values. The modified frequency counts are used to determine covariance and Pearson correlation matrices, the remaining part of the Hopfield-Potts model learning is performed as explained in the previous Sections.

6.2. Spectral density and pattern features

The upper panel of Fig. 1 displays the spectral density (only the $L(q-1) = 2240$ non-zero eigenvalues are shown). It is characterized by a pronounced peak around $\lambda = 1$. Non-zero eigenvalues range from $\lambda_{min} \simeq 0.132$ to $\lambda_{max} \simeq 45.1$.

The middle panel of the same figure shows the contributions of all patterns to the log-likelihood of the Hopfield-Potts model, given by Eqn. (30), vs. their eigenvalues. Maximum-likelihood inference consists of deciding a log-likelihood threshold (i.e. a horizontal line in the plot) such that there are exactly p diamonds above, and $L(q-1)-p$ below this line. Whereas the largest contributions come from the attractive patterns corresponding to large eigenvalues, already for $p = 10$ the first repulsive pattern (smallest eigenvalue) is selected.

The lower panel of Fig. 1 gives the weighted inverse participation ratio (IPR)

$$\mathrm{IPR}(\xi) = \frac{\sum_{ia} f_i(a)^2\,[\xi_i(a)]^4}{\left(\sum_{ia} f_i(a)\,[\xi_i(a)]^2\right)^2} \tag{48}$$

of patterns ξ as a function of the corresponding eigenvalue. The IPR can take values between 1 (for a pattern which has only one single non-zero component) and $1/Lq \simeq 0.00042$ (for a completely distributed pattern of constant entries). This range is in practice reduced due to the gauge, which does not allow a single component to be non-zero, or all components to have the same sign. The presence of the $f_i(a)$ factor in Eqn. (48) allows us to weigh the pattern components in the IPR according to the frequencies of the amino-acids in the alignment. We observe that strongly localized patterns correspond to eigenvalues which are either small (i.e. having a large contribution to the likelihood) or close to one (i.e. not contributing to the likelihood). Attractive patterns are extended.

The upper row of Fig. 2 shows the three most repulsive patterns, which correspond to the smallest eigenvalues. These patterns are found to be very localized in a few sites, which are in contact in the three-dimensional protein fold. It is important to notice that, due to the negative sign of the score function (9) for imaginary-valued patterns, repulsive patterns enforce the presence of particular combination of amino-acids on specific sites in a tight manner. Take for example the first repulsive pattern in Fig. 2: the score function is zero only if both of the two amino-acids, which correspond to the two large (positive and negative) components of the pattern are present, or if none of them are present. The score function is, on the contrary, negative if only one of the above amino-acid is present. As is described in more detail in [19] we

have verified that this strong constraint is due to specific chemical bonds or physical interactions between the two amino-acids.

On the contrary, the three most attractive patterns, displayed in the lower row of Fig. 2, are extended. As discussed in [19], the components of the first attractive pattern are strongly correlated to residue conservation. The other two patterns show strong peaks in the values of a corresponding to the alignment gap symbol. They have the largest entries, in absolute value, close to the two ends of the protein sequence. These two patterns actually represent an artifact of the alignment procedure: It is easier to extend a gap than to open a new one after an aligned amino acid. This leads, to correlations between the occurrence of gaps in close-by positions, in particular at the beginning and at the end of each sequence.

6.3. Errors vs. likelihood

Whereas pattern selection in [19] uses a maximum-likelihood approach, the original theoretical proposal of Hopfield-model learning in the inverse Ising problem [22] uses a Bayesian approach. In this Bayesian approach, error bars for patterns are estimated, and only patterns up to some maximum relative error are considered. As in maximum-likelihood pattern selection, this leads to the inclusion of patterns in both tails of the spectrum, i.e. to attractive and repulsive patterns corresponding to large resp. small eigenvalues. As is explained above, here we have generalized the expressions for the errors on patterns and couplings to the Hopfield-Potts case, in order to compare the two selection criteria.

The results are given in Fig. 3: The main figure contains the error bars of the patterns as a function of the associated eigenvalue. Note that the error bars in Eqn. (34) depend explicitly on the selected patterns, *i.e.* for a given error threshold (or a given pattern number p), the included patterns have to be determined self-consistently. In general, the selection of a smaller number of patterns increases slightly the error bars.

In the insert of Fig. 3, we plot the log-likelihood contribution $\Delta\mathcal{L}(\lambda)$ of patterns against the error bars (determined without error threshold). We find an impressive collapse of the data over several orders of magnitude for both the attractive and the repulsive patterns on a single joint curve. Some minor fluctuations are observed in the tail of the most repulsive patterns.

As a consequence, we find that for all but very small numbers p of selected patterns, the maximum-likelihood and the Bayesian criterion almost coincide. Due to the increased computational complexity of the Bayesian criterion (error bar calculation and self-consistent selection), we therefore stick to the maximum-likelihood criterion of [19].

Figure 3 contains also a parametric plot of the relative error ϵ of the Hopfield-Potts couplings, cf. Eqn. (44), as a function of both ℓ_- (left branch) and ℓ_+ (right branch), the latter being determined via maximum-likelihood selection. A strong accumulation of the individual pattern errors into coupling errors can be observed, underlining the importance of dimensional reduction for obtaining high-quality couplings.

6.4. Contact prediction by the Hopfield-Potts model

The output of the Hopfield-Potts modeling is a $q \times q$ matrix for each pair (i, j) of sites, i.e. of residue positions in the aligned protein sequences. To predict contacts, we have to say which of these matrices are "large", i.e. stand for strong couplings, and which are small. A single scaler score has to be assigned to each coupling matrix. To this end, we follow [25]. We first introduce the Frobenius norm

$$F_{ij} = \sqrt{\sum_{a,b=1}^{q} \tilde{e}_{ij}(a,b)^2} \tag{49}$$

of the transformed coupling matrices

$$\tilde{e}_{ij}(a,b) = e_{ij}(a,b) - e_{ij}(\cdot,b) - e_{ij}(a,\cdot) + e_{ij}(\cdot,\cdot) \tag{50}$$

with the dot denoting an average over all amino acids and the gap in the concerned position. This transformation leads to a gauge where the sum over each column or row of the coupling matrix for each given (i,j) vanishes; this gauge minimizes the Frobenius norm with respect to the gauge freedom. Intuitively, this gauge puts "as much as possible" of the statistical modeling into the local field parameters, and "as little as necessary" into the couplings. This score is adjusted by the heuristic *average product correction* (APC) [26],

$$F_{ij}^{APC} = F_{ij} - \frac{F_{\cdot j} F_{i\cdot}}{F_{\cdot\cdot}} \ , \tag{51}$$

now with the dot indicating an average over all sites $1, ..., L$.

Once this is done, position pairs $(i,j), 1 \leq i < j \leq L$, are ranked according to their F^{APC} values. The highest-ranking ones are expected to be in contact. Our capacity to actually predict residue contacts is assessed in Fig. 4, where predictions are compared with the native contact map of the response regulator protein. We observe that most 'islands' of tertiary contacts are identified by one or more of the 50 top F^{APC} values. Remarkably the quality of the prediction with only $p = 150$ patterns is very similar to the one obtained from all $p = L(q-1) = 2240$ patterns. Hence most patterns do not actually convey much information about contacts, and can be discarded in the inference.

We show in Fig. 5 the success rate of the prediction, defined as the fraction of true positive contacts among the top x scores F^{APC} as a function of x. Only contacts between far away residues (by more than 4 sites) on the chain are taken into account. We observe that with all $p = L(q-1)$ patterns the first 58 scores correspond to true contacts (black line), while inference with only $p = 150$ patterns make a first false positive prediction after 34 predicted contacts (blue line). Strikingly, however, the Hopfield-Potts model with the $p = 150$ most repulsive patterns only shows remarkable performances and beats the all-pattern model for success rates $\sim 85 - 90\%$, corresponding to $\simeq 130 - 140$ predicted contacts (red line). On the contrary, the Hopfield-Potts with the same number $p = 150$ of attractive patterns only shows much poorer performances (green line).

7. Conclusion

In this paper we have shown, in a detailed way, how the inverse Hopfield-Potts model can be solved within the mean field approximation. In general this inverse model is useful to analyze the correlated activity of a population of variables with the aim of extracting an interaction network between them. The Hopfield-Potts model is expected to be efficient for large data sets, which correspond to networks with many nodes and interactions, and to be robust against undersampling.

We have here applied the inverse Hopfield Potts approach to the correlated mutations found in different sequences of a protein family; the interaction network between the sequence sites allows us to make predictions about the sites in contact in the tridimensional structure of the protein. We have focused on the analysis of the protein family PF00072, which has 112 sites (the variables of our interaction network), each one carrying 20 possible amino-acids plus the gap symbol. This defines a large network with a total of $(112 \times 21)2/2 \simeq 3\,10^6$ interaction parameters to be inferred from a sampled data set made of only 62,000 sequences. The inverse Hopfield-Potts approach offers several interesting features: i) The coupling matrix is expressed as a function of its principal modes, called patterns, and allows for a strong dimensional reduction compared to the usual Potts model; ii) The inferred patterns can be simply related to the eigenvectors of

the Pearson correlation matrix of the data set, see (19); iii) The log-likelihood of the data given the patterns has a simple expression (30) and gives a natural way to select the best patterns; iv) An alternative and equivalent way to select patterns is the evaluation of their statistical error bars, see inset of Fig. 3. Moreover here we have shown how to calculate the statistical errors on the couplings from the statistical errors on the patterns; v) Among the patterns which contribute most to the log-likelihood many low-eigenvalues modes of the Pearson Correlation Matrix are present. The importance of these modes is new with respect to the classical PCA approach in which only large-eigenvalues modes are selected. Low- eigenvalues modes, which we call the repulsive patterns, are often localized on sites in contact on the 3D fold of the protein. Remarkably, for the PF00072 family studied here, the quality of the prediction with only $p = 150$ repulsive patterns is very similar to the one obtained from all $p = L(q-1) = 2240$ patterns as shown in the contact map of Fig. 4. A more exhaustive analysis on 15 protein families can be found in [19]. We have also discussed in detail the two gauge invariances of the Hopfield-Potts model. The first one, typical of the Potts model, results from the conservation of the probability of observing one of the possible symbols on a site, while the second gauge invariance is typical of the Hopfield model, and is due to the invariance of the couplings under rotations in the pattern space. We have chosen here the gauges in a different way with respect to [19]. The choice of the Hopfield gauge is irrelevant as far as couplings are concerned when all $p = L(q-1)$ patterns are selected, but it does affect the couplings when only a limited number of patterns are selected. Moreover the Hopfield gauge choice could a priori change the sites in which the patterns are localized. However, it is interesting to remark that the two choices of the gauge do not alter the main findings, i.e. the localization of repulsive patterns on the same sites and the accurate prediction of the contacts from a limited number of repulsive patterns. It would be interesting to have a better comprehension of how the gauge changes the representation of the patterns and if there is an optimal choice for the gauge.

It would also be useful to have a statistical mechanics justification for the pre-processing of data, through the reweighting procedure and the introduction of a large pseudo-count, and for the use of the average product correction to have a better estimator of contacts. Finally, let us mention that it is in principle possible to calculate corrections to the mean field approximation, as was done in [22]. However, in the case of small pattern amplitudes (of the order $1/\sqrt{L}$), a very good sampling is needed for the corrections to be effective. Calculating corrections to the large and localized pattern components encountered here remains an open problem.

References

[1] de Juan D, Pazos F and Valencia A 2013 *Nature Reviews Genetics* **14** 249
[2] Jaynes E T 1957 *Physical Review Series II* **106** 620
[3] Jaynes E T 1957 *Physical Review Series II* **108** 171
[4] Lapedes A S, Giraud B G, Liu L and Stormo G D 1999 *Lecture Notes-Monograph Series: Statistics in Molecular Biology and Genetics* **33** 236
[5] Weigt M, White R A, Szurmant H, Hoch J A and Hwa T 2009 *Proc. Natl. Acad. Sci. U. S. A.* **106** 67
[6] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D S, Sander C, Zecchina R, Onuchic J N, Hwa T and Weigt M 2011 *Proc. Natl. Acad. Sci. U. S. A.* **108** E1293
[7] Burger L and van Nimwegen E 2010 *PLoS Comput. Biol.* **6** E1000633
[8] Balakrishnan S, Kamisetty H, Carbonell J G, Lee S I and Langmead C J 2011 *Proteins: Struct., Funct., Bioinf.* **79** 1061
[9] Jones D T, Buchan D W A, Cozzetto D and Pontil M 2012 *Bioinformatics* **28** 184
[10] Schug A, Weigt M, Onuchic J N, Hwa T and Szurmant H 2009 *Proc Natl Acad Sci USA* **106** 22124
[11] Marks D S, Colwell L J, Sheridan R P, Hopf T A, Pagnani A, Zecchina R and Sander C 2011 *PLoS ONE* **6** e28766
[12] Sadowski M I, Maksimiak K and Taylor W R 2011 *Computational Biology and Chemistry* **35** 323
[13] Sulkowska J I, F Morcos F, Weigt M, Hwa T and Onuchic J N 2012 *Proc. Natl. Acad. Sci.* **109** 10340
[14] Nugent T and Jones D T 2012 *Proceedings of the National Academy of Sciences* **109** E1540
[15] Hopf T A, Colwell L J, Sheridan R, Rost B, Sander C and Marks D S 2012 *Cell* **149** 1607

[16] Dago A E, Schug A, Procaccini A, Hoch J A, Weigt M and Szurmant H 2012 *Proc Natl Acad Sci USA* **109** 10148
[17] Marks D S, Hopf T A and Sander C 2012 *Nature Biotechnology* **30** 1072
[18] Taylor W R, Hamilton R S and Sadowski M I 2013 *Current Opinion in Structural Biology* –
[19] Cocco S, Monasson R and Weigt M 2013 *PLoS Comput. Biol.* **9** E1003176
[20] Hopfield J J 1982 *Proc. Natl. Acad. Sci.* **79** 2554
[21] Lunt B, Szurmant H, Procaccini A, Hoch J, Hwa T and Weigt M 2010 *Methods in Enzymology* **471** 43
[22] Cocco S, Monasson R and Sessak V 2011 *Physical Review* E **83** 051123
[23] Punta M, Coggill P C, Eberhardt R Y, Mistry J, Tate J G, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer E L L, Eddy S R, Bateman A and Finn R D 2012 *Nucleic Acids Res.* **40** D290
[24] Bent C J, Isaacs N W, Mitchell T J and Riboldi-Tunnicliffe A 2004 *J. Bacteriol.* **186** 2872
[25] Ekeberg M, Lövkvist C, Lan Y, Weigt M and E A 2013 *Phys. Rev.* E **87** 012707
[26] Dunn S D, Wahl L M and Gloor G B 2008 *Bioinformatics* **24** 333

8. Figures and figure captions

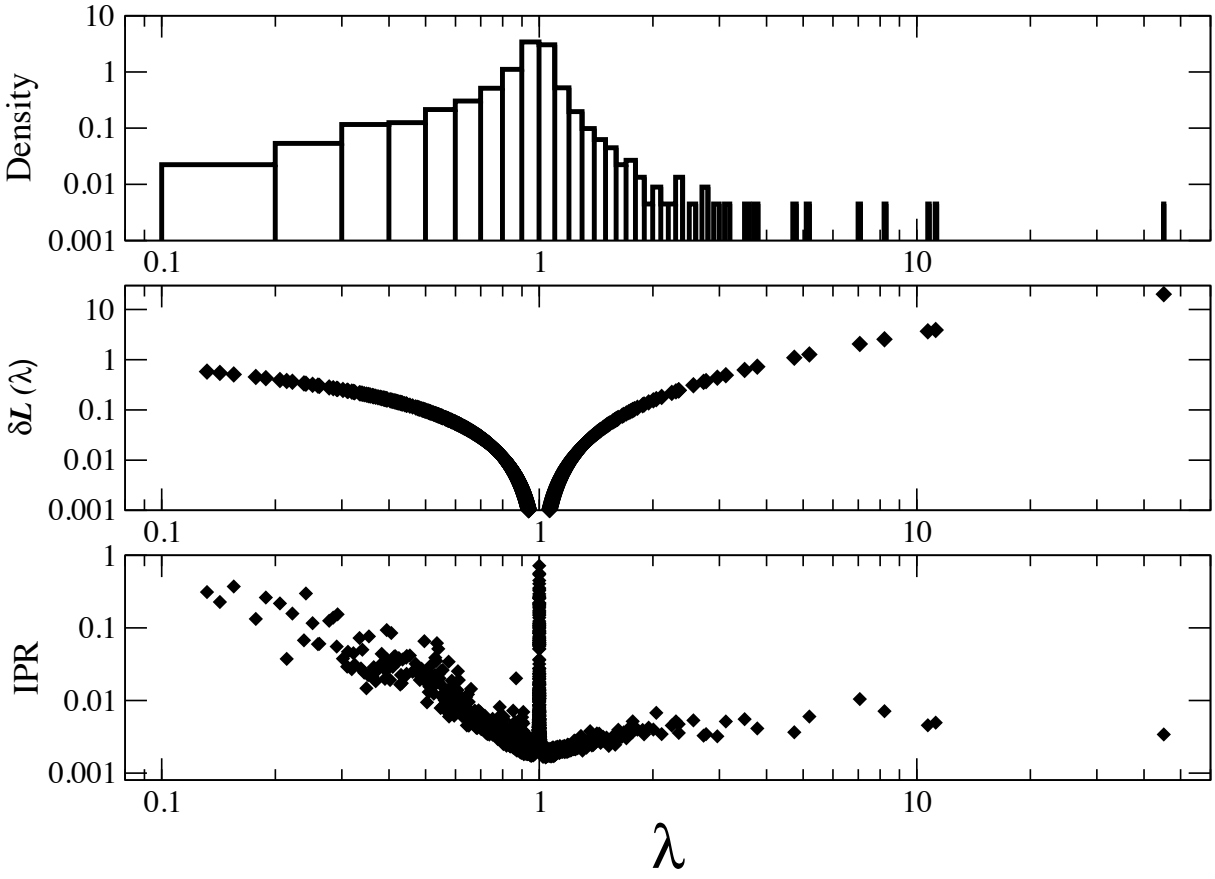

Figure 1. Spectral density (top), contributions to the log-likelihood (middle) and inverse participation ratios of the patterns (bottom) for the family PF00072. See text for description.

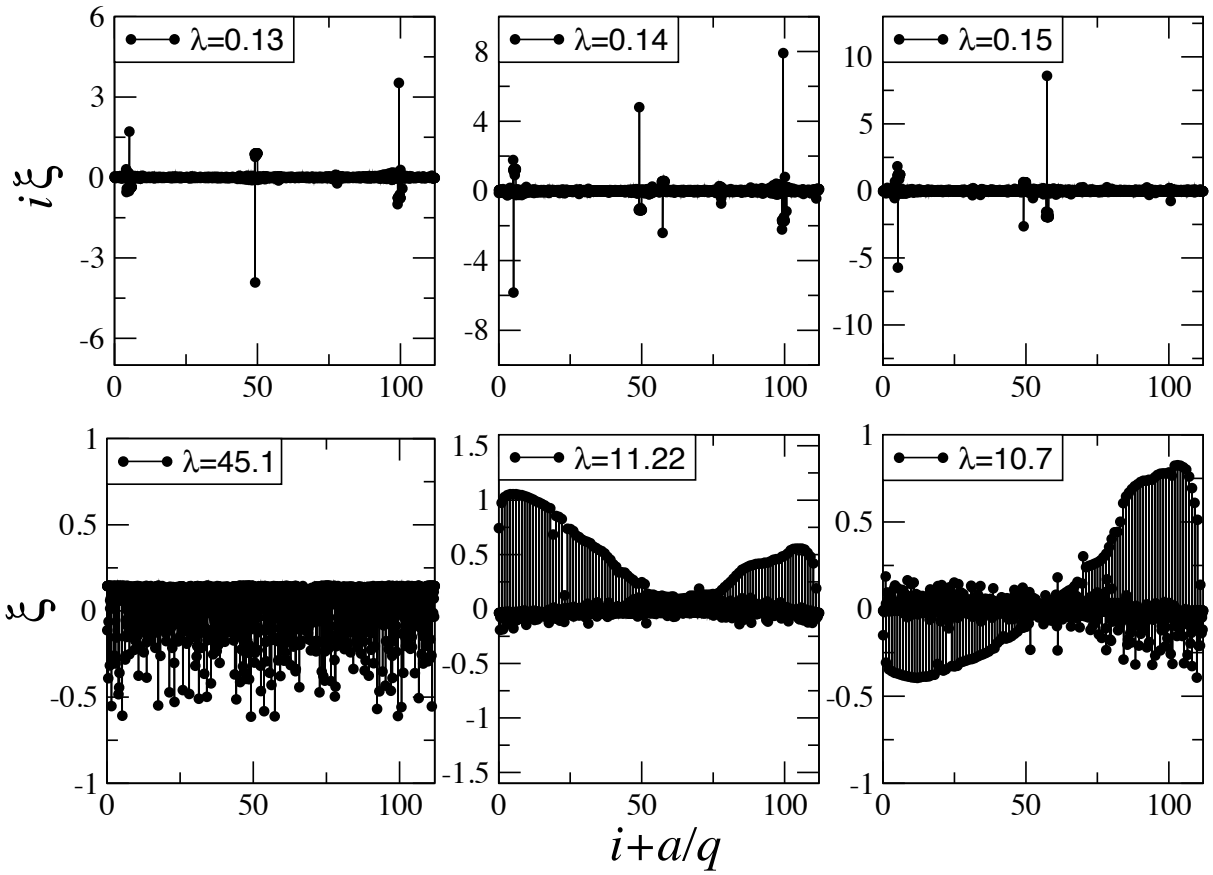

Figure 2. Patterns for the family PF00072. Top: the three patterns corresponding to the three lowest eigenvalues; as pattern components are purely imaginary we plot $i\xi$ instead of ξ. Bottom: the three patterns corresponding to the three largest eigenvalues. The residue coordinate is a rationale number $x = i + a/q$, where i is the site index and a is the amino-acid index.

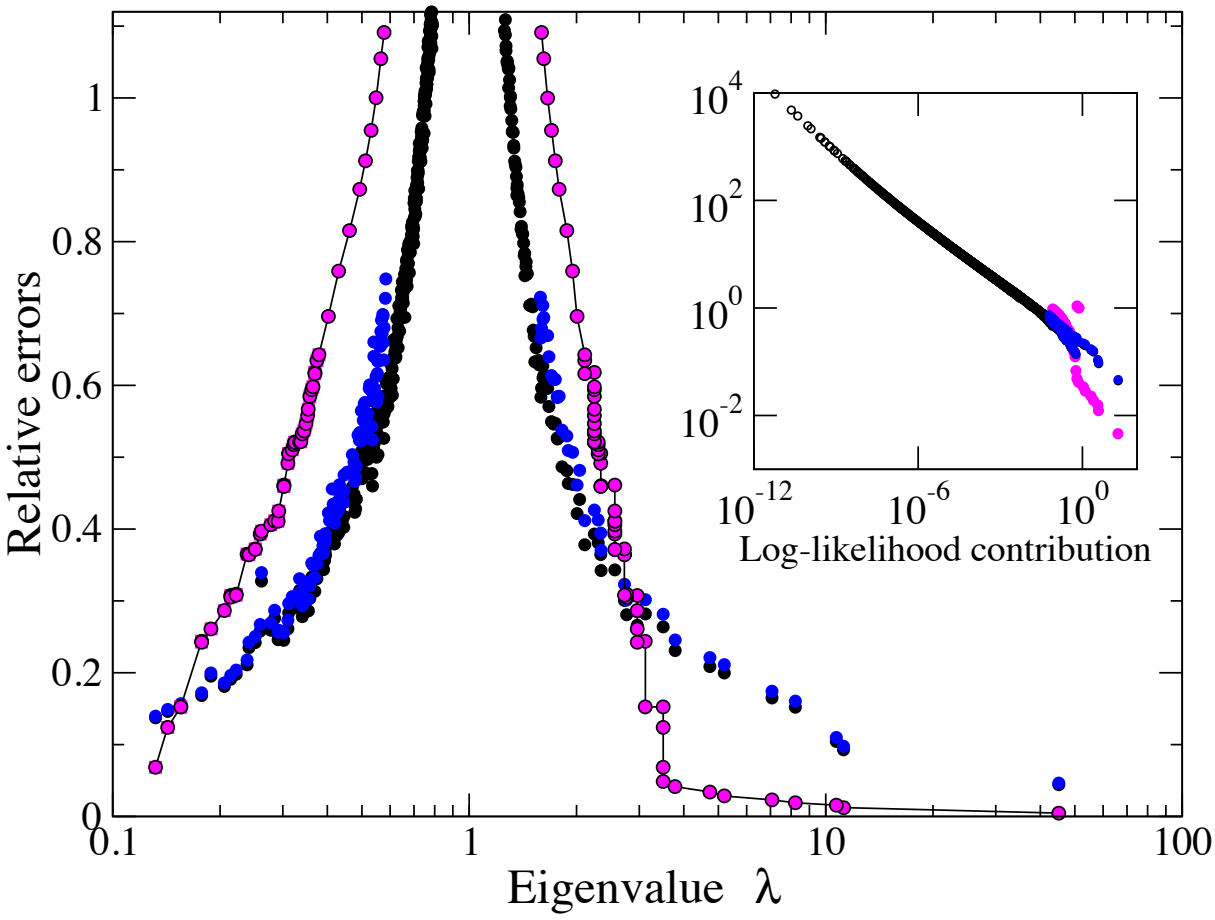

Figure 3. Bayesian pattern selection for the family PF00072. Relative error $\sqrt{\rho^{\mu}}$, see Eqn. (40), as a function of the eigenvalues for $p = 150$ selected patterns according to the log-likelihood contributions (blue dots) and for all $p = 2240$ patterns (black dots). The number M of sequences in formula (44) for the error is replaced with the number of independent sequences, $M_{eff} = 29408$, estimated from the reweighting procedure. Magenta dots give the relative error $\sqrt{\epsilon}$ of the Hopfield-Potts couplings as a function of the eigenvalue ℓ_- of the maximal selected repulsive pattern (left branch) and of the eigenvalue ℓ_+ minimal selected attractive pattern (right branch). The relative error on the couplings is displayed up to a maximum of $p = 120$ selected patterns. The inset shows a parametric plot of the error for each pattern as a function of the pattern contribution $\Delta\mathcal{L}(\lambda)$; same color code as for the main figure. The relative error on the couplings as a function of $\Delta\mathcal{L}(\lambda)$ is also given in the inset (magenta dots) up to 120 selected patterns.

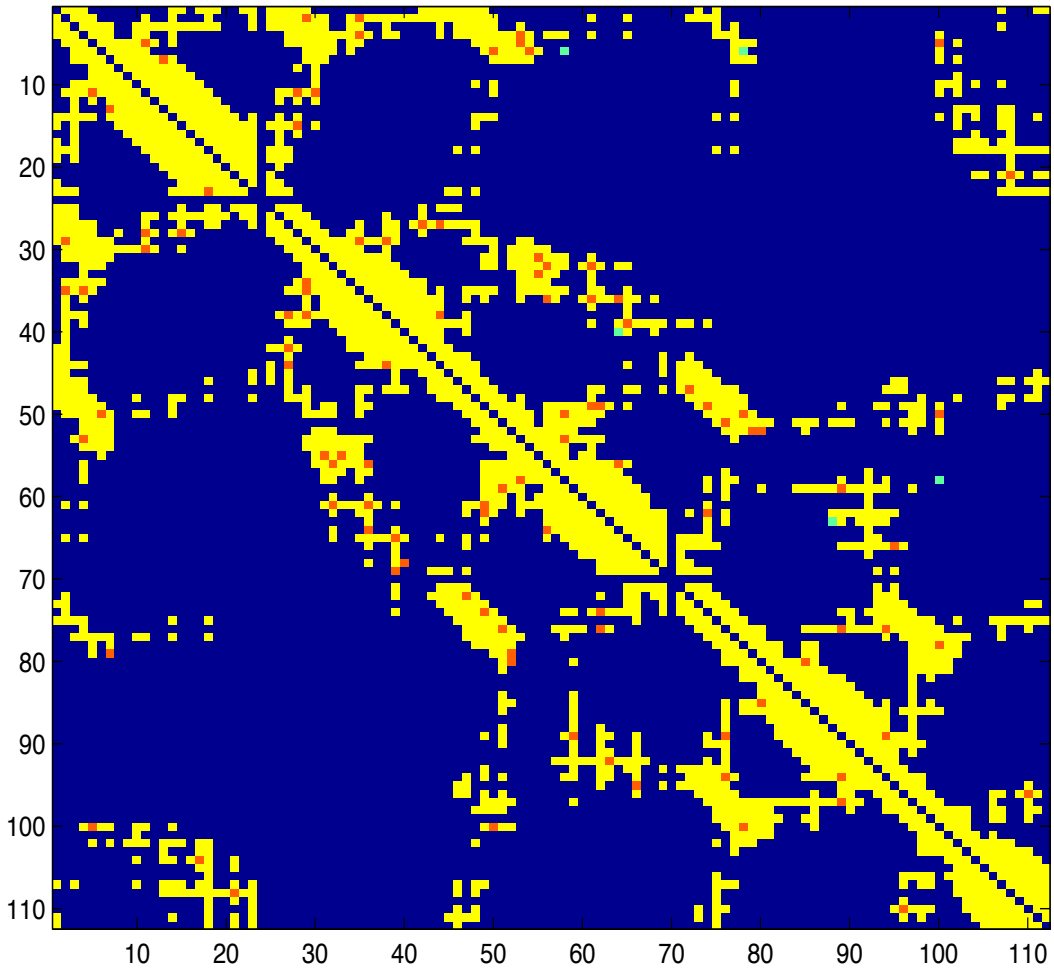

Figure 4. Contact map of the family PF00072. Yellow squares show residues i and j in contact (distance between Calpha smaller than 8Å); For the sake of clarity we have removed contacts between residues nearby on the chain, *i.e.* such that $|i-j| \leq 4$. Red and green dots are contact predictions from the Hopfield-Potts model with $p = 150$ patterns (top right corner) and with all $P = L(q-1) = 2240$ patterns (bottom left corner). Predictions are made from the top 50 Frobenius norms, red dots indicate true positives and green dots locate false positives.

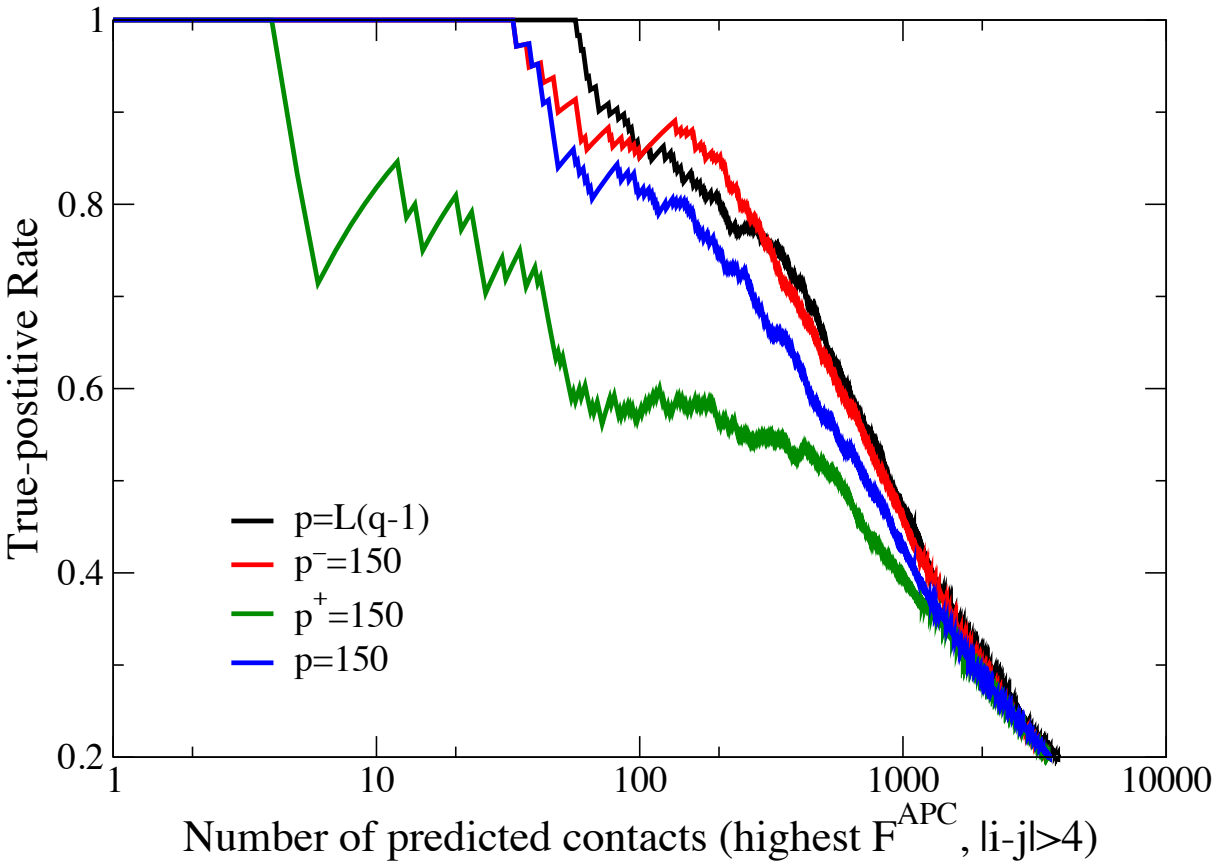

Figure 5. True positive rates for the family PF00072. Fraction of true contacts as a function of the number of top scores F^{APC} retained for the $p = 150$ patterns with largest contributions to the log-likelihood ($p_+ = 39, p_- = 111$, blue curve), for the most 150 repulsive patterns (red curve), for the most 150 attractive patterns (green curve), and for all $p = 2240$ patterns (black curve). Only contacts between sites i, j such that $|i - j| > 4$ on the chain are considered to compute the fraction of true positive.