

Sequence analysis

ACE: adaptive cluster expansion for maximum entropy graphical model inference

J. P. Barton^{1,2,*}, E. De Leonardis^{3,4}, A. Coucke^{4,5} and S. Cocco^{3,*}

¹Departments of Chemical Engineering and Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, ²Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard, Cambridge, MA 02139, USA, ³Laboratoire de Physique Statistique de L'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure & Université P.&M. Curie, Paris, France, ⁴Computational and Quantitative Biology, UPMC, UMR 7238, Sorbonne Université, Paris, France and ⁵Laboratoire de Physique Théorique de L'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure & Université P.&M. Curie, Paris, France

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 25, 2016; revised on May 15, 2016; accepted on May 18, 2016

Abstract

Motivation: Graphical models are often employed to interpret patterns of correlations observed in data through a network of interactions between the variables. Recently, Ising/Potts models, also known as Markov random fields, have been productively applied to diverse problems in biology, including the prediction of structural contacts from protein sequence data and the description of neural activity patterns. However, inference of such models is a challenging computational problem that cannot be solved exactly. Here, we describe the adaptive cluster expansion (ACE) method to quickly and accurately infer Ising or Potts models based on correlation data. ACE avoids overfitting by constructing a sparse network of interactions sufficient to reproduce the observed correlation data within the statistical error expected due to finite sampling. When convergence of the ACE algorithm is slow, we combine it with a Boltzmann Machine Learning algorithm (BML). We illustrate this method on a variety of biological and artificial datasets and compare it to state-of-the-art approximate methods such as Gaussian and pseudo-likelihood inference.

Results: We show that ACE accurately reproduces the true parameters of the underlying model when they are known, and yields accurate statistical descriptions of both biological and artificial data. Models inferred by ACE more accurately describe the statistics of the data, including both the constrained low-order correlations and unconstrained higher-order correlations, compared to those obtained by faster Gaussian and pseudo-likelihood methods. These alternative approaches can recover the structure of the interaction network but typically not the correct strength of interactions, resulting in less accurate generative models.

Availability and implementation: The ACE source code, user manual and tutorials with the example data and filtered correlations described herein are freely available on GitHub at <https://github.com/johnbarton/ACE>.

Contacts: jpbarton@mit.edu, cocco@lps.ens.fr

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Interpreting patterns of correlations in data is a fundamental problem across scientific disciplines. A common approach to this problem is to infer a simple graphical model that explains the statistics of the data through a network of effective interactions between the variables, which may then be used to generate new predictions (Friedman, 2004). The goal of this approach is to disentangle the direct interactions between variables from their correlations, which arise through a combination of direct and indirect effects. Here, we focus on a particular family of undirected graphical models, referred to as Potts models in the language of statistical physics, which have recently been applied to study a wide variety of biological systems. Applications include inference of the effective connectivity of populations of neurons, and their patterns of firing activity, based on data from multi-electrode recordings (Barton and Cocco, 2013; Cocco et al., 2009; Roudi et al., 2009; Schneidman et al., 2006), and the prediction of protein contact residues (Morcos et al., 2011) and the fitness effects of mutations (Ferguson et al., 2013; Figliuzzi et al., 2016; Mann et al., 2014) based on the analysis of multiple sequence alignments (MSAs).

Unfortunately, the inference of Potts models from data is challenging. The computational time required for naive Potts inference algorithms scales exponentially with the system size, rendering the problem intractable for realistic systems of interest. Various approximations have been employed to combat this problem, including Gaussian and mean-field inference (Kappen and Rodríguez, 1998), perturbative expansions (Nguyen and Berg, 2012; Sessak and Monasson, 2009) and pseudo-likelihood methods (Aurell and Ekeberg, 2012; Ravikumar et al., 2010). These approximate methods can successfully capture the general structure of the network of interactions, recovering, in particular, contact residues in the three-dimensional structure of protein families (Cocco et al., 2013; Ekeberg et al., 2014; Feinauer et al., 2014; Hopf et al., 2012; Marks et al., 2011; Morcos et al., 2011; Sułkowska et al., 2012), but the resulting models typically give a less accurate statistical description of the data (Barton et al., 2014). Alternately, algorithms based on iterative rounds of Monte Carlo simulation (Ackley et al., 1985; Mann et al., 2014; Sutto et al., 2015) are capable of inferring models that accurately reproduce the observed correlations, but they are typically slow to converge.

Here, we describe an extension of the adaptive cluster expansion (ACE) method, originally devised for binary (Ising) variables (Cocco and Monasson, 2011, 2012), to more general (Potts) variables taking multiple categorical values. We also describe new computational methods for faster inference, including a Monte Carlo learning procedure and the optional incorporation of prior knowledge about the structure of the interaction graph. The algorithm has been successfully applied to real data with as many as several hundred variables, including studies of neural activity in the retina and prefrontal cortex (Barton and Cocco, 2013; Cocco and Monasson, 2011, 2012; Tavoni et al., 2015), human immunodeficiency virus (HIV) fitness based on protein MSA data (Barton et al., 2015; Mann et al., 2014), and lattice protein models (Jacquin et al., 2016). Below we illustrate the application of this method to both real and artificial datasets. We show that models inferred by ACE give an excellent reconstruction of the statistics of the data. They also accurately recover, considering sampling limitations, true underlying model parameters when they are known, and can achieve comparable performance to state-of-the-art methods for predicting structural contacts in protein family data. We compare these results to those obtained using other

approximate inference methods, focusing in particular on pseudo-likelihood methods.

1.1 Background

The Potts model emerges naturally in the statistical description of complex systems. Consider a system of N variables described by the configuration $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$, with $x_i \in \{1, 2, \dots, q_i\}$. The number of discrete categories q_i that each variable x_i can take on, which we refer to as states, may depend on the variable index i . For proteins the states correspond to particular amino acids, while for neurons they represent the binary (firing or silent) state of activity. Given a set of measurements of the system, the empirical average over the sampled configurations gives us the $\sum_i q_i$ individual and $\sum_{i < j} q_i q_j$ pairwise frequencies for the different states of each variable in the data. We denote the individual and pairwise frequencies by $p_i(a)$ and $p_{ij}(a, b)$, respectively, where i, j are the index of the variables and a, b are the index of the states. As an example, \mathbf{x} could represent sequences in a MSA, with $p_i(a)$ the frequency of the amino acid labeled by a in column i of the alignment, and $p_{ij}(a, b)$ the frequency of the pair of amino acids a, b in columns i, j .

The simplest, or maximum entropy (Jaynes, 1982), probabilistic model capable of reproducing the observed frequencies is an exponential distribution, which assigns a probability to every configuration of the system \mathbf{x} :

$$P(\mathbf{x}) = \frac{\exp(-E(\mathbf{x}))}{Z},$$

$$E(\mathbf{x}) = -\sum_{i=1}^N b_i(x_i) - \sum_{i=1}^{N-1} \sum_{j=i+1}^N J_{ij}(x_i, x_j), \quad (1)$$

$$Z = \sum_{\mathbf{x}} \exp(-E(\mathbf{x})).$$

Here, the partition function Z is a normalizing factor which ensures that all probabilities sum to one. In the simple case that all the variables x_i are binary, this model is referred to as an Ising model. More generally, when x_i can take multiple discrete states, this model is referred to as a Potts model. The parameters $b_i(a)$ and $J_{ij}(a, b)$ in the energy function E , called fields and couplings, must be chosen such that variable averages (correlations) in the model match those in the data, i.e.

$$p_i(a) = \sum_{\mathbf{x}} \delta(x_i, a) P(\mathbf{x}),$$

$$p_{ij}(a, b) = \sum_{\mathbf{x}} \delta(x_i, a) \delta(x_j, b) P(\mathbf{x}), \quad (2)$$

where δ is the Kronecker delta function. The problem of finding the parameters $b_i(a)$, $J_{ij}(a, b)$ that satisfy Equation (2) is referred to as the inverse Potts problem. Note that the probability of any configuration remains unchanged under the transformation of the couplings and fields given by $J_{ij}(a, b) \rightarrow J_{ij}(a, b) + K_{ij}(a) + K_{ij}(b)$, $b_i(a) \rightarrow b_i(a) - \sum_{j \neq i} K_{ij}(a)$ for any K . In addition, all the $b_i(a)$ at a site i can be uniformly shifted by a constant with no effect on the probability. This ‘gauge invariance’ reduces the number of free parameters in the Potts model to $q_i - 1$ fields for each site and $(q_i - 1)(q_j - 1)$ couplings for each pair of sites.

Formally, the inverse Potts problem is solved by the set of fields and couplings that maximize the average log-posterior probability of the data or equivalently, those that minimize the cross-entropy between the data and the model

$$S \equiv -\frac{1}{B} \log \mathcal{L}(\mathbf{J}|\mathbf{p}) = S_{\text{Potts}}(\mathbf{J}|\mathbf{p}) - \frac{1}{B} \log P_0(\mathbf{J}), \quad (3)$$

where B is the number of data points in the sample (e.g. the number of sequences in a MSA), and

$$S_{\text{Potts}}(\mathbf{J}|\mathbf{p}) = \log Z - \sum_{i=1}^N \sum_{a=1}^{q_i} b_i(a) p_i(a) - \sum_{i=1}^{N-1} \sum_{j=i+1}^N \sum_{a=1}^{q_i} \sum_{b=1}^{q_j} J_{ij}(a, b) p_{ij}(a, b), \quad (4)$$

and P_0 is a prior distribution for the parameters. Here, for simplicity we have written the set of all individual and pairwise variable frequencies as \mathbf{p} and the set of all fields and couplings as \mathbf{J} . Note that, ignoring the contribution of the prior distribution, the cross-entropy S is equivalent to the entropy of the inferred model satisfying Equation (2). S has a simple interpretation in information theory as it can be written as the sum of the entropy of the data and the Kullback–Leibler divergence of the model with respect to the data (Shannon, 1948), see also the related field of information geometry (Amari, 1987).

The inclusion of a prior distribution helps to avoid overfitting, while also improving convergence. A Gaussian prior distribution for the parameters is a typical choice, which contributes a term

$$\gamma' \sum_{i=1}^N \sum_{a=1}^{q_i} b_i(a)^2 + \gamma \sum_{i=1}^{N-1} \sum_{j=i+1}^N \sum_{a=1}^{q_i} \sum_{b=1}^{q_j} J_{ij}(a, b)^2 \quad (5)$$

to Equation (3). For $\gamma \sim 1/B$ this factor can be thought of as a weakly informative prior (Gelman *et al.*, 2008) whose main purpose is to ensure that solutions of the inverse problem are not infinite due to issues of undersampling (e.g. parameters corresponding to an amino acid that is never observed). Note that this form of the regularization is not invariant under gauge transformations. Thus, the results of the inference including the regularization do have some dependence on the gauge choice. Other forms of regularization are also possible (see Supplementary Materials). Note that the presence of the partition function Z in Equation (4) precludes direct numerical maximization of the posterior when the system size is large, since this requires summing over all $\prod_{i=1}^N q_i$ configurations of the system. Alternate methods of solving the inverse Potts problem involve approximation schemes or rely on computationally costly Monte Carlo simulations, as described above.

2 Methods

2.1 Adaptive cluster expansion

The adaptive cluster expansion (Cocco and Monasson, 2011, 2012) is based on the formal decomposition of the regularized cross-entropy Equation (3) into a sum of contributions from subsets (clusters) of variables $\Gamma = \{i_1, \dots, i_k\}$, $k \leq N$,

$$S = \sum_{\Gamma} \Delta S_{\Gamma}, \quad \Delta S_{\Gamma} = S_{\Gamma} - \sum_{\Gamma' \subset \Gamma} \Delta S_{\Gamma'}, \quad (6)$$

where the sum is over all nonempty subsets of the N variables. The terms ΔS_{Γ} , referred to as cluster entropies, have been recursively defined as the remaining contribution to the subset posterior once all contributions from smaller clusters have been subtracted. Here, S_{Γ} denotes the maximum of Equation (3) restricted only to the variables in Γ . Thus, S_{Γ} depends only on the frequencies $p_i(a)$, $p_{ij}(a, b)$ with $i, j \in \Gamma$. Provided that the number of variables in Γ is small (typically ≤ 20), numerical maximization of the posterior restricted to Γ is tractable. Note that, due to the recursive definition of ΔS_{Γ} ,

the sum over all $2^N - 1$ nonempty and overlapping subsets of the N variables in Equation (6) gives the exact posterior S by construction:

$$\sum_{\Gamma} \Delta S_{\Gamma} = \Delta S_{\Gamma_{\text{tot}}} + \sum_{\Gamma' \subset \Gamma_{\text{tot}}} \Delta S_{\Gamma'} = S_{\Gamma_{\text{tot}}} = S. \quad (7)$$

Here, $\Gamma_{\text{tot}} = \{1, 2, \dots, N\}$ is the set of all variables in the system. The expansion of Equation (6) can be computationally expedient because, practically, it can converge toward S even when only contributions from clusters much smaller than the system size N are considered (see below).

The cluster entropy contributions are easy to interpret for one- and two-site clusters: neglecting the regularization term the single variable cluster contributions are the entropies of the variables taken as if they were independent, $\Delta S_i \equiv S_i = -\sum_{a=1}^{q_i} p_i \log p_i(a)$. The two variable entropy is $S_{ij} = -\sum_{a=1}^{q_i} \sum_{b=1}^{q_j} p_{ij}(a, b) \log p_{ij}(a, b)$ (see Supplementary Materials for more details). The cluster entropy for a pair of variables is then $\Delta S_{ij} = S_{ij} - S_i - S_j$, which is equivalent to the mutual information. It is zero when $p_{ij}(a, b) = p_i(a) p_j(b)$, i.e. when the two variables are independent. More generally, ΔS_{Γ} is a measure of the interdependence between the variables in the cluster which cannot be accounted for by smaller clusters.

The main idea of this approach is to approximate the cross-entropy (and simultaneously, the parameters that maximize it) by limiting the sum in Equation (6) to a restricted set of clusters Γ that give the most important contributions to it. As shown in (Cocco and Monasson, 2011, 2012), contributions for overlapping clusters sharing the same interaction subgraph partially compensate, and thus summing clusters according to the magnitude of their entropy contribution allows for a faster convergence of Equation (6). Neglecting clusters with small contributions to the cross-entropy also helps to avoid overfitting. As a simple example, for a unidimensional interaction graph in which each variable is only connected to its two nearest neighbors, the expansion can be exactly truncated by summing only clusters of size one and the largest contributing 2-site clusters containing neighboring variables (Cocco and Monasson, 2012; Gori and Trombettoni, 2011; Mastromatteo, 2013).

We define a threshold t on the cross-entropy to separate the significant clusters from those which can be neglected. Starting from a large value of the threshold (typically $t=1$), such that only a few clusters are selected, the algorithm proceeds through two nested iterations. The outer loop is on the value of the threshold t , which is progressively lowered until enough clusters are included to yield a model consistent with the data. The inner loop constructs the set of clusters Γ with contributions to the cross-entropy $|\Delta S_{\Gamma}| > t$ and yields an approximation of the cross-entropy and the model parameters at the threshold t . The algorithm stops at the first value of the threshold t where the inferred model fits the sampled averages and correlations Equation (2) to within the statistical error due to finite sampling (see Section 3.2).

The algorithm for the inner loop, including the selection and summation of individual clusters, is as follows. Given a list L_k of clusters of size k , beginning with the list of all clusters of size $k=2$:

1. For each cluster $\Gamma \in L_k$,
2. Compute S_{Γ} by numerical minimization of Equation (3) restricted to Γ .
3. Record the parameters minimizing Equation (3), called \mathbf{J}_{Γ} .
4. Compute ΔS_{Γ} using Equation (6).
5. Add all clusters $\Gamma \in L_k$ with $|\Delta S_{\Gamma}| > t$ to a new list $L'_k(t)$.
6. Construct a list L_{k+1} of clusters of size $k+1$ from overlapping clusters in $L'_k(t)$.

The rule for constructing new clusters of size $k + 1$ from selected clusters of size k can be lax (such that a new cluster Γ is added provided that any pair of size k subclusters, $\Gamma_1, \Gamma_2 \in L'_k(t)$ and $\Gamma_1 \cup \Gamma_2 = \Gamma$) or strict (such that a new cluster is only added if all of its $k + 1$ subclusters of size k belong to $L'_k(t)$). The above process is then repeated until no new clusters can be constructed.

After the summation of clusters terminates, the approximate value of the parameters minimizing the cross-entropy, given the current value of the threshold, is computed by

$$J(t) = \sum_k \sum_{\Gamma \in L'_k(t)} \Delta J_\Gamma, \quad \Delta J_\Gamma = J_\Gamma - \sum_{\Gamma' \subset \Gamma} \Delta J_{\Gamma'}. \quad (8)$$

Note that this formula generally yields sparse solutions because nonzero couplings are only included in Equation (8) if some clusters containing them have been selected. In this algorithm the dominant contribution to the computational complexity often comes from the evaluation of the partition function Z for large cluster sizes, which requires $\mathcal{O}\left(\prod_{i \in \Gamma} q_i\right)$ operations to compute.

2.2 Compression of the number of Potts states

As mentioned in Section 1.1, the number of states each variable may take on need not be the same for all variables in a system. States with zero (or otherwise very small) probabilities may be observed very infrequently in real, finitely-sampled data, and the relative error on the corresponding correlations due to finite sampling is large.

To limit overfitting and reduce the computational time, the low probability states can be effectively grouped together according to a given compression parameter. Here, we present two heuristic conventions for compressed representations of the data. First, for each variable we can treat explicitly the states observed with probability larger than a cutoff value $p_i(a) > p_o$ while grouping all infrequently observed values into the same state. A natural value for the cutoff is $p_o \sim 1/\sqrt{B}$, such that pair correlations between independent states with frequencies of p_o are at the threshold of detection. Alternatively, we can order the states by their contribution to the total single site entropy S_q and choose a reduced model in which only the first k states are modeled explicitly, with k chosen to capture a certain fraction f of the site entropy. This is achieved by explicitly considering the first k states and grouping the remaining $q - k$ states together, choosing k such that

$$S_k = -\sum_{a=1}^k p(a) \log p(a) - \left(1 - \sum_{a=1}^k p(a)\right) \log \left(1 - \sum_{a=1}^k p(a)\right) \geq f S_q. \quad (9)$$

The frequency of the regrouped Potts state is then the sum of the frequencies of the states which have been regrouped: $p_i(k+1) = \sum_{a=k+1}^q p_i(a)$. Once the reduced model is inferred, one can recover a complete model by modifying the field parameter for the regrouped states, $h_i(a') = h_i(k+1) + \log(p_i(a')/p_i(k+1))$, while keeping the couplings to the value of the regrouped state $J_{ij}(a', b) = J_{ij}(k+1, b)$. For states with zero probabilities in the data, we fix the fields from the regularization alone.

2.3 Expansion around a reference structure

ACE is a two-fold algorithm: it builds up the interaction graph while also inferring the parameters that reproduce the correlated structure of the data. This expansion can be accelerated if information about the interaction graph is available. It is also possible to expand the cross-entropy around its Gaussian approximation.

- If the list of directly interacting variables is known, one can run the expansion starting from clusters built on the support of the interaction graph. For proteins this procedure can be applied using the real contact map, known from structural information, as the initial list of 2-site clusters. Alternatively, if the contact map is not known, one can use fast inference approaches such as DCA or plmDCA (Ekeberg et al., 2014; Morcos et al., 2011) to obtain a list of initial putative contacts and then refine the expansion from this initial list (Supplementary Materials).
- As shown in (Cocco and Monasson, 2012) for the Ising model, one can analytically calculate the posterior and the parameters that maximize it under the Gaussian approximation with an ad hoc L_2 -norm regularization (where the regularization strength depends on the variable frequencies). It is then possible to perform the cluster expansion around this Gaussian reference model, i.e. the expansion of $S - S_0$, where S_0 is the cross-entropy for a Gaussian model

$$S_0 = \frac{1}{2} \log \det M, \quad M_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}, \quad (10)$$

in the Ising (binary) case. If S_0 is a reasonable approximation of S , then the expansion of $S - S_0$ may converge more rapidly than the expansion of S alone. See Cocco and Monasson (2012) for further details on the expansion of the Gaussian model.

2.4 Refinement with Boltzmann Machine Learning (BML)

In cases where convergence of the cluster algorithm alone is not sufficiently fast, it is often more expedient to use the output set of fields and couplings as starting values for a Boltzmann Machine Learning (BML) routine. In typical cases, provided that the inferred model is not too sparse, this procedure can lead to rapid convergence of the model even when the starting error is large.

Here, we adapted the RPROP algorithm for neural network learning (Riedmiller and Braun, 1993) to the case of Potts models. Given an input set of fields and couplings, we first compute the model correlations $p_i^{MC}(a), p_{ij}^{MC}(a, b)$ through Monte Carlo simulation. The couplings and fields are then updated according to the gradient of the posterior, multiplied by a parameter-specific weight factor

$$\begin{aligned} b_i(a) &\rightarrow b_i(a) - (p_i^{MC}(a) - p_i(a)) w_i(a), \\ J_{ij}(a, b) &\rightarrow J_{ij}(a, b) - (p_{ij}^{MC}(a, b) - p_{ij}(a, b)) w_{ij}(a, b). \end{aligned} \quad (11)$$

Regularization can also be incorporated by adding $2\gamma J_{ij}(a, b)$, or the analogous term for fields, to the gradient. Here, the weights $w_i(a)$ and $w_{ij}(a, b)$ are also updated with each iteration of the algorithm. At each iteration, if the sign of $(p_i^{MC}(a) - p_i(a))$ is the same as in the previous round, $w_i(a) \rightarrow s_+ w_i(a)$, else $w_i(a) \rightarrow s_- w_i(a)$, and similarly for the $w_{ij}(a, b)$. This acceleration of weight parameters allows appropriate step sizes to be chosen adaptively for each coupling and field. To prevent step sizes from becoming too large or too small, the weight parameters are restricted to lie between some w_{\min} and w_{\max} . Typical choices of the weight bounds and update multipliers are $w_{\min} = 10^{-3}$, $w_{\max} = 10$, $s_+ = 1.9$, $s_- = 0.5$. Note that we choose $s_+ < 1/s_-$ so that, if the sign of one of the terms of the gradient continually switches, the corresponding weight decreases.

As with other BML approaches, this procedure is computationally limited by the need to thermalize the system to accurately estimate the model correlations through MC. Each MC step requires a computation of the change in energy due to a change in the configuration \mathbf{x} , requiring $\mathcal{O}(n)$ operations, where n is the typical ‘neighborhood’ size (i.e. number of sites to which another site couples with nonzero $J_{ij}(a,b)$). Future refinements could improve the speed of this routine by implementing, for example, adaptive selection of the number of thermalization steps and more efficient Monte Carlo sampling techniques.

3 Results

3.1 Description of test data and their preprocessing

Here, we apply ACE to five different datasets and test the reconstruction of their statistics. First, we generate artificial data from a Potts model with random fields and couplings, allowing us to test the ability of the algorithm to recover the true model parameters. Second, we infer a Potts model from artificial sequences generated by a $3 \times 3 \times 3$ lattice protein model with large folding probabilities in a given structure. This folding probability (Shakhnovich and Gutin, 1990) contains all-order interactions between amino acids, unlike the Potts model used for the inference, thus serving as an interesting benchmark test (Jacquin *et al.*, 2016). Third, we study trypsin inhibitor protein sequences (Cocco *et al.*, 2013; Ekeberg *et al.*, 2013, 2014; Morcos *et al.*, 2011) to compare structural predictions obtained by ACE to ones obtained using Gaussian (DCA) and pseudo-likelihood (plmDCA) methods. We then test the ability of the algorithm to infer a model describing the HIV nucleocapsid protein p7. Finally, we study multi-electrode recordings of neural activity in the prefrontal cortex of a rat (Peyrache *et al.*, 2009) analyzed in Tavoni *et al.* (2015) to study memory replay.

3.1.1 Potts models on Erdős-Rényi random graphs (ER05)

We consider an example of a Potts model with $q=21$ states, where the network of interactions is described by an Erdős-Rényi random graph with $N=50$ variables. Each edge in the interaction graph is included with probability 0.05. Field and coupling values for interacting pairs of sites are selected from a Gaussian distribution (Supplementary Materials). We compute the correlations through Monte Carlo sampling of $B=10^4$ configurations. In the results shown below we compressed rarely-observed Potts states with $p_i(a) < p_o = 0.05$ and used $\gamma = 1/B = 10^{-4}$, performing the inference in the gauge of the compressed Potts state.

3.1.2 Lattice protein model (LP S_B)

We consider an alignment of 5×10^4 protein sequences with $N=27$ sites, arranged in a $3 \times 3 \times 3$ cube, selected according to their exactly computable (Shakhnovich and Gutin, 1990) folding probability S_B (see (Jacquin *et al.*, 2016), Supplementary Materials). In the results below we have removed amino acids that are never observed (i.e. compression with $p_o=0$), and used the regularization $\gamma = 5/B = 10^{-4}$. Couplings and fields corresponding to the least frequently observed amino acid at each site are gauged to zero.

3.1.3 Trypsin inhibitor protein family (PF00014)

We study an alignment of 4915 sequences downloaded from the PFAM database for the trypsin inhibitor protein family (PF00014, PFAM 28.0 release, May 2015). After removing columns with $> 50\%$ gaps the number of sites is $N=53$. We reweight the contribution of each sequence to the correlations according to its

similarity to other sequences in the alignment, an approach commonly used to attenuate phylogenetic correlations (Morcos *et al.*, 2011). Here, we show results in the consensus gauge after compressing rarely-observed amino acids with $p_i(a) < p_o = 0.05$, using $\gamma = 2/B = 10^{-3}$. Additionally, we note that gaps in the MSA are not generally modeled well in the Potts model representation with pairwise interactions, as they tend to be present in long stretches, especially at the beginning and the end of the alignment (Feinauer *et al.*, 2014). Such stretches of highly correlated gaps slow down the inference procedure because they give rise to large clusters. Here, we have processed the data to replace gaps by random amino acids with the same frequency as observed in the non-gapped sequences. Though this approach obscures the important variability in the sequence lengths in the MSA, it is a simple way to reduce computational problems induced by correlated gaps, valuable for structural prediction.

3.1.4 HIV p7 nucleocapsid protein

The HIV nucleocapsid protein p7 plays an essential role in multiple aspects of viral replication (Freed, 2015). We downloaded a MSA of 4131 p7 sequences from individuals infected by clade B viruses from the Los Alamos National Laboratory HIV sequence database (www.hiv.lanl.gov, accessed October 6, 2014). After removing columns with $> 95\%$ gaps, the remaining number of sites is $N=71$. Here, we do not reweight sequences by similarity, given that they are all in the same phylogenetic cluster, and the regular observation of similar sequences in the HIV population may be indicative of higher fitness (Ferguson *et al.*, 2013; Shekhar *et al.*, 2013). We replaced gaps as described above, compressed rarely-observed amino acids with $f_s = 90\%$, and chose $\gamma \simeq 1/2B = 1.4 \times 10^{-4}$. Inference is performed in the consensus gauge.

3.1.5 Multi-electrode recordings of cortical neurons

We divided a 20 minute recording of the firing activity of 32 cortical neurons into a set of $B = 1.5 \times 10^5$ time bins of 10ms, treating each time window as an observation of the system. During each time window, the variable for each neuron i was assigned $x_i = 1$ if the neuron was active at least once during that time, and zero otherwise. Here, we take $\gamma = 1/B = 6.6 \times 10^{-6}$.

3.2 Convergence of the cluster expansion algorithm

As mentioned in Section 2.1, for each threshold t used to select clusters in the ACE expansion, the model individual $\langle x_i(a) \rangle$ and pairwise $\langle x_{ij}(a,b) \rangle$ frequencies are compared to the data’s frequencies $p_i(a)$ and $p_{ij}(a,b)$. We define a relative error as the ratio between the deviations of the predicted observables from the data, $\delta \langle x_i \rangle = \langle x_i \rangle - p_i$ and $\delta \langle x_{ij} \rangle = \langle x_{ij} \rangle - p_{ij}$, and the expected statistical fluctuations due to finite sampling, $\delta p_i(a) = \sqrt{p_i(a)(1-p_i(a))/B}$, $\delta p_{ij}(a,b) = \sqrt{p_{ij}(a,b)(1-p_{ij}(a,b))/B}$. We define the normalized maximum error as

$$\epsilon_{\max} = \max_{\{i,j,a,b\}} \frac{1}{\sqrt{2 \log(M)}} \left(\frac{|\delta \langle x_i(a) \rangle|}{\delta p_i(a)}, \frac{|\delta \langle x_{ij}(a,b) \rangle|}{\delta p_{ij}(ab)} \right) \quad (12)$$

where M is the total number of one- and two-point correlations.

Figure 1 shows the behavior of ϵ_{\max} and the cross-entropy as a function of the threshold for the five datasets described above. The cross-entropy S approaches a constant value as the threshold is decreased. In all cases except for the lattice protein model, the algorithm converges at $\epsilon_{\max} \sim 1$, when the correlations are reproduced to within the expected error due to finite sampling. The expansion slows dramatically for the lattice protein model at a fairly high value of the threshold due to the large number of states included at each

site in the model (typically $q = 19$). The computational cost of calculating the partition function is a limiting factor as the maximum cluster size increases, corresponding to $K_{\max} = 7$ at the stopping point in Figure 1. At this point BML is needed to refine the parameters inferred through the cluster expansion. Note that, even in cases when the error appears large, convergence of the BML procedure is often rapid because only small changes to the parameters may be necessary to obtain a model that accurately reproduces the correlations.

Convergence of the algorithm can also be more difficult for alignments of long proteins or those with very strong interactions. In such cases one may observe large oscillations in the cross-entropy as a function of the threshold, and large (≥ 10 sites) clusters may appear even at high thresholds. Strong regularization ($\gamma > 1/B$) can help to dampen these oscillations, after which it can be returned to $\approx 1/B$ during the BML procedure.

3.3 Parameters of the ER05 model are recovered by ACE

In Figure 2 we show that the 2×10^4 underlying parameters for the ER05 model corresponding to the explicitly modeled Potts states are accurately recovered by ACE. These states are better sampled and therefore they have smaller statistical uncertainties. In the model inferred by plmDCA, which includes no reduction in the number of states, there are around 10^6 parameters. Those corresponding to the explicitly modeled states are recovered fairly well (with some errors in the fields), but parameters corresponding to compressed states are difficult to infer due to insufficient sampling (see Supplementary Materials for details and analysis of errors in inferred parameters due to finite sampling).

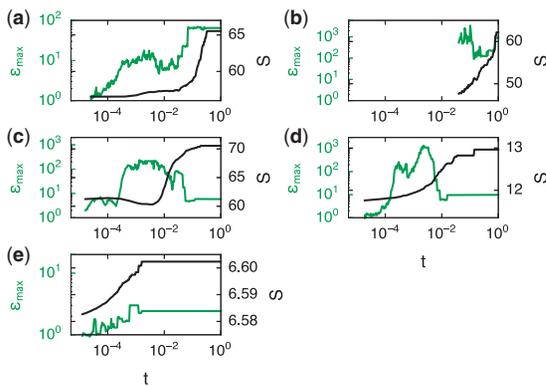


Fig. 1. Convergence of the cluster expansion as a function of the threshold t for (a) ER005, (b) LP S_B , (c) PF00014, (d) HIV p7 and (e) cortical data. As the threshold is lowered, the cross-entropy S approaches a constant value. In all cases except for LP S_B the normalized maximum error ϵ_{\max} reaches 1 through the cluster expansion alone. For LP S_B a Monte Carlo learning procedure is used to refine the inferred parameters and reach $\epsilon_{\max} \approx 1$.

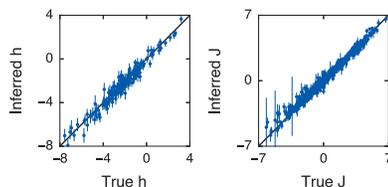


Fig. 2. ACE accurately recovers the true fields h (left) and couplings J (right) corresponding to Potts states with $p_i(a) \geq 0.05$ for the ER05 model. Error bars denote standard deviation in estimated parameters due to finite sampling.

3.4 Statistics of the data are accurately reproduced

Figures 3 and 4 show how the model inferred by ACE reproduces the statistics of the input data. In all cases the model accurately captures the input probabilities and pairwise connected correlations within the expected error due to finite sampling, as anticipated.

We also find that higher order correlations in the data can be accurately reproduced. Figure 4 shows the 3-point connected correlations and the distribution $P(k)$ of Hamming distances k between the sampled configurations and the configuration in which each site takes on the most probable value (i.e. the consensus sequence for proteins). In the neural case the most probable configuration is the silent one and therefore $P(k)$ is the probability to have k active neurons in the same time window. Models inferred by ACE outperform those from plmDCA (Ekeberg et al., 2014), see Figure 3 and Supplementary Materials for higher order statistics.

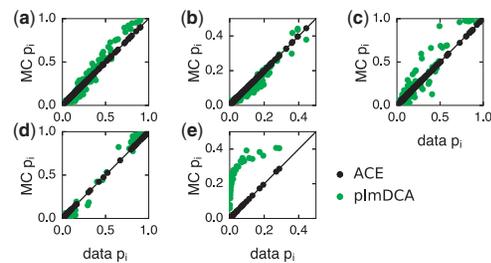


Fig. 3. ACE outperforms plmDCA in recovering the single variable frequencies for models describing (a) ER005, (b) LP S_B , (c) PF00014, (d) HIV p7 and (e) cortical activity. The results for plmDCA are obtained with the regularization $\gamma = 0.01$, which gives better results for the correlations than lower values of the regularization strength (see Supplementary Materials).

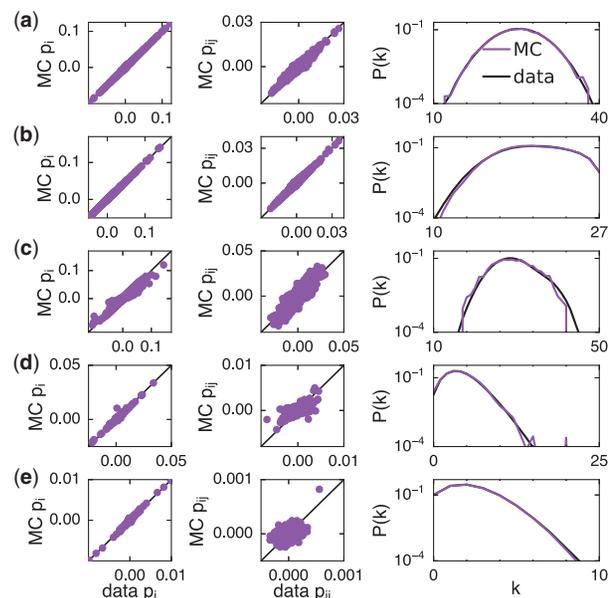


Fig. 4. Fit for models describing (a) ER005, (b) LP S_B , (c) PF00014, (d) HIV p7 and (e) cortical activity. ACE recovers the connected pair correlations $c_{ij}(a, b) = p_{ij}(a, b) - p_i(a)p_j(b)$ (left). The inferred model also successfully captures higher order correlations present in the data, such as the connected three-body correlations (center) and the probability $P(k)$ of observing a configuration with k differences from the consensus configuration (right).

Comparing the distribution of energies E for configurations sampled from the inferred model to the distribution obtained from the original data provides an additional check of statistical consistency. The energy of a configuration is proportional to the logarithm of its probability (in addition, because the entropy S is obtained from the cluster expansion, we can also compute the constant of proportionality). Concordance between the inferred and empirical energy distributions thus indicates that the real data could plausibly be generated from the inferred model. Figure 5 compares the data and model distributions of energies, showing that in most cases they closely overlap. A small discrepancy is introduced in PF00014 because of the reweighting procedure (here, the histogram of the data is normalized by the sequence weights). The energy distribution for the lattice protein model is broader than for the data, though the peak is fit correctly. In contrast with models inferred using ACE, the distribution of energies of the data is less well reproduced with plmDCA (Supplementary Materials). The ability to estimate the probability of a configuration can be useful when comparing the likelihood of a configuration in two different models, for example to decide which family a given protein belongs to.

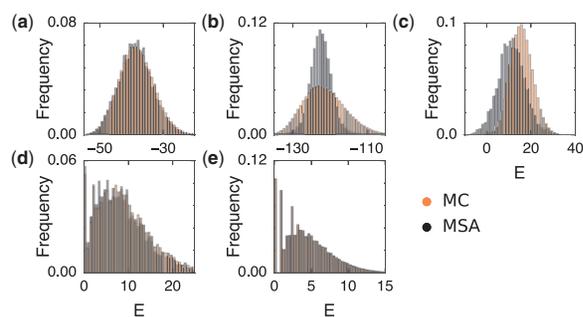


Fig. 5. Histograms of the data (MSA) and model (MC) energy distributions for (a) ER005, (b) LP S_B , (c) PF00014, (d) HIV p7 and (e) cortical activity. Monte Carlo sampling of the inferred Potts model describing each set of data yields a distribution of energies similar to the empirical distribution, a further check on the consistency of the model fit beyond the fitting of correlations (Color version of this figure is available at *Bioinformatics* online.)

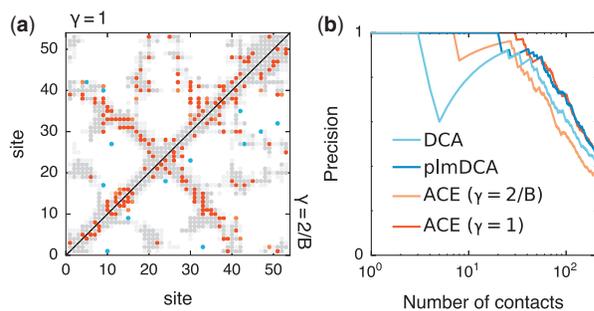


Fig. 6. (a) Contact map for PF00014 inferred by ACE. Here, we show the top 100 predicted contacts, with true predictions in orange and false predictions in blue. Other contact residues in the crystal structure are shown in gray. For true positives and other contact residues, close contacts ($<6\text{\AA}$) are darkly shaded and further contacts ($<8\text{\AA}$) are lightly shaded. The upper and lower triangular parts of the contact map give predictions for the inferred model with strong regularization/no compression ($\gamma=1$) and weak regularization/high compression ($\gamma=2/B$), respectively. (b) Precision (number of true predictions divided by the total number of predictions) as a function of the number of contact predictions for close contact residues that are widely separated on the protein backbone ($i-j > 4$). Results using ACE compare favorably with those from DCA (Morcos *et al.*, 2011) and are competitive with those from plmDCA (Ekeberg *et al.*, 2014) (Color version of this figure is available at *Bioinformatics* online.)

3.5 ACE accurately infers structural contacts for PF00014

In Figure 6, we use the inferred couplings to predict pairs of residues that are in contact in the folded protein structure for PF00014, and we compare results from ACE to the standard contact prediction methods DCA (Morcos *et al.*, 2011) and plmDCA (Ekeberg *et al.*, 2014). In this case the pairs of sites for which the Frobenius norm of the couplings is largest, including the average product correction (APC, see (Dunn *et al.*, 2008)), are predicted to be most likely to be in contact. We define contact residues to be those that are within 6\AA of each other in the folded structure of the protein, and we exclude trivial contact pairs along the protein backbone ($i-j \leq 4$).

The accuracy of contact predictions with ACE can be increased by decreasing the compression ($p_o=0$) and using a large regularization ($\gamma=1$), in the same spirit as the strong regularization employed in typical DCA and plmDCA approaches. Here, we gauged the parameters for the least frequently observed amino acids to zero and computed the Frobenius norm of the couplings in the zero sum gauge (as is typical in DCA). The couplings are then strongly damped by regularization and the cluster expansion converges for maximal cluster sizes much smaller than those needed in the case with weaker regularization. Figure 6b shows that the precision in this case is competitive with the one obtained from plmDCA, and the prediction of the first ~ 30 contacts is slightly better for ACE. However, in this case we note that because of the small values of the couplings the generative properties of the inferred model are lost (see Supplementary Materials for the statistical fit of the model).

4 Discussion

Potts models have been successfully applied to study a variety of biological systems. However, the computational difficulty of the inverse Potts problem, i.e. the inference of a Potts model from correlation data, has presented a barrier to their use. Here, we presented ACE, a flexible, easy-to-use method for solving the inverse Potts problem, which can be applied to analyze a wide variety of real and synthetic data. We also provide tools for automatically generating correlation data from multiple sequence alignments (MSA), making the analysis of this type of data even more accessible.

Here, we have adapted the complexity of the inferred Potts models to the level of the sampling in the data. This is achieved by regrouping less frequently observed Potts states into a unique state (according to a threshold on entropy or frequency), then by a sparse inference procedure that omits interactions that are unnecessary for reproducing the statistics of the data to within the error bounds due to finite sampling. On artificial data we verified that compression of the number of Potts states allows for a faster and more precise inference of the uncompressed model parameters while reducing overfitting. The methods of compression that we describe here can also be applied to other inference methods (including, for example, the DCA and plmDCA approaches discussed above), a topic of future study. In addition, as described above ACE yields sparser models when sampling is poor, leading to more robust inference.

This method allows for the simple construction of models from various types of data, which can then be used to predict the evolution of experimental systems and their response to perturbations. Previous work has demonstrated promising applications of such models in a variety of different biological contexts. In neuroscience, the analysis of multi-electrode recordings has led to models that identify cell assemblies, which are thought of as basic units of neural computation and memory (Hebb, 1949; Peyrache *et al.*, 2009; Tavoni *et al.*, 2015).

Studies of MSAs of protein families allows for the prediction of pairs of residues in contact in the folded protein structure, giving insights on the protein structure from sequence information alone. Classical protein folding algorithms can be then used to refine the structure from contact predictions (Hopf *et al.*, 2012; Marks *et al.*, 2011; Sułkowska *et al.*, 2012). Potts models have also been used to describe the mutational landscape of viral and bacterial proteins, where they provide information about the effects of mutations on protein function, which could potentially be exploited to improve vaccine design and drug treatment (Barton *et al.*, 2015; Ferguson *et al.*, 2013; Figliuzzi *et al.*, 2016; Mann *et al.*, 2014).

In the present work, we have compared ACE with standard maximum entropy inference methods based on Gaussian and pseudo-likelihood approximations. These methods are particularly fast and adapted to find structural contacts and use, respectively, large pseudocounts and regularizations. Inference with ACE is generally slower than mean-field and pseudo-likelihood approaches. However, it allows for the accurate inference of underlying model parameters (when they are known), and for the construction of good generative models of the data when using a Bayesian value of the regularization strength ($\gamma \approx 1/B$). In analogy with DCA and plmDCA, when using ACE with little compression (e.g. $p_o = 0$) and strong regularization the contact prediction obtained using traditional contact estimators is improved while the generative power of the inferred model is degraded. A recent work has also shown that a BML algorithm can be used to give a good generative model predicting the structure and functional dynamics of proteins (Sutto *et al.*, 2015). Running such algorithms from a good initial guess of parameters, such as those obtained by ACE as shown here, could help to accelerate the inference procedure.

An additional advantage of ACE is that it evaluates the entropy of the Potts model corresponding to a given set of data. For protein sequence data, this entropy gives a measure of the variability of the sequences in the same protein family, and can be used to predict site-dependent variability and robustness with respect to mutations (Barton *et al.*, 2016). We have now successfully applied the method to protein sequences of a few hundred amino acids in length collected from phylogenetically distant organisms, or longer sequences (up to 500 amino acids) for more phylogenetically related and less variable HIV proteins.

Acknowledgements

This work originates from the development of ACE in the Ising case in collaboration with R. Monasson, to whom we are grateful for many helpful discussions. We also thank U. Ferrari for his contribution to the development of the adaptive Monte Carlo sampling to evaluate the reconstruction errors in the inference, D. Murakowski for his contribution to the development of the partition function expansion, and H. Jacquin for useful discussions.

Funding

J.B. is funded by the Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard. A.C. is funded by the Institut des Systèmes Complexes (ISC-PIF) and the Région Île-de-France. S.C. is funded by ANR-13-BS04-0012-01 (Coevstat).

Conflict of Interest: none declared.

References

Ackley, D. *et al.* (1985) A learning algorithm for Boltzmann machines. *Cognit. Sci.*, **9**, 147–169.

- Amari, S. (1987) Differential geometrical theory of statistics. *IMS Monograph vol. 10, Differential Geometry in Statistical Inference*, pp. 20–94.
- Aurell, E. and Ekeberg, M. (2012) Inverse Ising inference using all the data. *Phys. Rev. Lett.*, **108**, 090201.
- Barton, J. and Cocco, S. (2013) Ising models for neural activity inferred via selective cluster expansion: structural and coding properties. *J. Stat. Mech.: Theory Exp.*, **2013**, P03002.
- Barton, J.P. *et al.* (2014) Large pseudocounts and L2-norm penalties are necessary for the mean-field inference of Ising and Potts models. *Phys. Rev. E*, **90**, 012132.
- Barton, J.P. *et al.* (2015) Scaling laws describe memories of host–pathogen riposte in the HIV population. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 1965–1970.
- Barton, J.P. *et al.* (2016) On the entropy of protein families. *J. Stat. Phys.*, **162**, 1–27.
- Cocco, S. and Monasson, R. (2011) Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Phys. Rev. Lett.*, **106**, 090601.
- Cocco, S. and Monasson, R. (2012) Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. *J. Stat. Phys.*, **147**, 252–314.
- Cocco, S. *et al.* (2009) Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 14058–14062.
- Cocco, S. *et al.* (2013) From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. *PLoS Comput. Biol.*, **9**, e1003176.
- Dunn, S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.
- Ekeberg, M. *et al.* (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E*, **87**, 012707.
- Ekeberg, M. *et al.* (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.*, **276**, 341–356.
- Feinauer, C. *et al.* (2014) Improving contact prediction along three dimensions. *PLoS Comput. Biol.*, **10**, e1003847.
- Ferguson, A.L. *et al.* (2013) Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, **38**, 606–617.
- Figliuzzi, M. *et al.* (2016) Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.*, **33**, 268–280.
- Freed, E.O. (2015) HIV-1 assembly, release and maturation. *Nat. Rev. Microbiol.*, **13**, 484–496.
- Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
- Gelman, A. *et al.* (2008) A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, **2**, 1360–1383.
- Gori, G. and Trombettoni, A. (2011) The inverse Ising problem for one-dimensional chains with arbitrary finite-range couplings. *J. Stat. Mech.: Theory Exp.*, **2011**, P10021.
- Hebb, D.O. (1949). *The Organization of Behavior: A Neurophysiological Approach*. John Wiley & Sons, New York.
- Hopf, T.A. *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.
- Jacquin, H. *et al.* (2016) Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. *PLoS Comput. Biol.*, **12**, e1004889.
- Jaynes, E.T. (1982) On the rationale of maximum-entropy methods. *Proc. IEEE*, **70**, 939–952.
- Kappen, H.J. and Rodríguez, F.B. (1998) Efficient learning in Boltzmann machines using linear response theory. *Neural Comput.*, **10**, 1137–1156.
- Mann, J.K. *et al.* (2014) The fitness landscape of HIV-1 Gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput. Biol.*, **10**, e1003776.
- Marks, D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.
- Mastromatteo, I. (2013) Beyond inverse Ising model: structure of the analytical solution. *J. Stat. Phys.*, **150**, 658–670.

- Morcos, F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, E1293–E1301.
- Nguyen, H.C. and Berg, J. (2012) Bethe–Peierls approximation and the inverse Ising problem. *J. Stat. Mech.: Theory Exp.*, **2012**, P03004.
- Peyrache, A. *et al.* (2009) Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat. Neurosci.*, **12**, 919–926.
- Ravikumar, P. *et al.* (2010) High-dimensional Ising model selection using l_1 -regularized logistic regression. *Ann. Stat.*, **38**, 1287–1319.
- Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster back-propagation learning: The rprop algorithm. In: *IEEE International Conference on Neural Networks, 1993*, IEEE, pp. 586–591.
- Roudi, Y. *et al.* (2009) Ising model for neural data: Model quality and approximate methods for extracting functional connectivity. *Phys. Rev. E*, **79**, 051915.
- Schneidman, E. *et al.* (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, **440**, 1007–1012.
- Sessak, V. and Monasson, R. (2009) Small-correlation expansions for the inverse Ising problem. *J. Phys. A: Math. Theor.*, **42**, 055001.
- Shakhnovich, E. and Gutin, A. (1990) Enumeration of all compact conformations of copolymers with random sequence of links. *J. Chem. Phys.*, **93**, 5967–5971.
- Shannon, C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
- Shekhar, K. *et al.* (2013) Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Phys. Rev. E*, **88**, 062705.
- Sulkowska, J.I. *et al.* (2012) Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 10340–10345.
- Sutto, L. *et al.* (2015) From residue coevolution to protein conformational ensembles and functional dynamics. *Proc. Natl. Acad. Sci.*, **112**, 201508584.
- Tavoni, G. *et al.* (2015). Inferred model of the prefrontal cortex activity unveils cell assemblies and memory replay. *bioRxiv* 10.1101/028316.