Correlation-compressed direct-coupling analysis

Chen-Yi Gao,[1,2] Hai-Jun Zhou,[1,2,3,*] and Erik Aurell[4,5,†]

[1] *Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China*
[2] *School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*
[3] *Synergetic Innovation Center for Quantum Effects and Applications, Hunan Normal University, Changsha, Hunan 410081, China*
[4] *Department of Computational Biology, KTH Royal Institute of Technology, 10044 Stockholm, Sweden*
[5] *Department of Applied Physics and Department of Computer Science, Aalto University, 00076 Aalto, Finland*

(Received 4 November 2017; revised manuscript received 8 June 2018; published 11 September 2018)

Learning Ising or Potts models from data has become an important topic in statistical physics and computational biology, with applications to predictions of structural contacts in proteins and other areas of biological data analysis. The corresponding inference problems are challenging since the normalization constant (partition function) of the Ising or Potts distribution cannot be computed efficiently on large instances. Different ways to address this issue have resulted in a substantial amount of methodological literature. In this paper we investigate how these methods could be used on much larger data sets than studied previously. We focus on a central aspect, that in practice these inference problems are almost always severely undersampled, and the operational result is almost always a small set of leading predictions. We therefore explore an approach where the data are prefiltered based on empirical correlations, which can be computed directly even for very large problems. Inference is only used on the much smaller instance in a subsequent step of the analysis. We show that in several relevant model classes such a combined approach gives results of almost the same quality as inference on the whole data set. It can therefore provide a potentially very large computational speedup at the price of only marginal decrease in prediction quality. We also show that the results on whole-genome epistatic couplings that were obtained in a recent computation-intensive study can be retrieved by our approach. The method of this paper hence opens up the possibility to learn parameters describing pairwise dependences among whole genomes in a computationally feasible and expedient manner.

DOI: 10.1103/PhysRevE.98.032407

I. INTRODUCTION

Learning from a multiple sequence alignment has emerged in recent years as an important development in statistical physics and computational biology. A main paradigm has been to use the data to infer parameters of a pairwise model, namely, an Ising model if the data are binary or a Potts model if the data have more than two types. From a statistical point of view these are inference problems in exponential families [1], while from a physical point of view the approach has been called the inverse Ising or Potts problem [2,3] and direct-coupling analysis (DCA) [4,5].

Since DCA was first successfully used to find interprotein contacts in two kinds of protein dimers [4], it has been used to identify various biological interactions, such as native contacts of proteins [3,5–7], nucleotide-nucleotide contacts of RNAs [8], multiple-scale protein-protein interactions [9,10], amino acid–nucleotide interactions in RNA-protein complexes [11], and synergistic effects of mutations (epistasis) not necessarily related to spatial contacts [12–14]. Moreover, an exciting perspective has been opened: Contact prediction of DCA can also be integrated into structure-prediction methodology to approach *in silico* prediction of protein three-dimensional structures from primary sequences (see, e.g., [15–20] for some important contributions, [21] for a highlight, and [22] for a recent review).

To set the stage for the following discussion we introduce basic notions. A set of similar (homologous) sequences can be arranged in a matrix such that each column contains, as often as possible, like symbols. Such a matrix is called a multiple sequence alignment (MSA). The rows in an MSA are the individual sequences, and a column corresponds to a position along the structure, which we will refer to as a locus. In real biological data some sequences typically lack one or several long or short subsequences, which give rise to gaps in the MSA. In nucleotide sequence data gaps are often represented by the International Union of Pure and Applied Chemistry code N (any nucleotide). An MSA is hence represented by an $N \times L$ matrix consisting of N sequences where each, with possible gaps, fits into L loci.

We continue by stressing that the actual use of DCA is comprised of two steps: (1) learning a pairwise model from an MSA and (2) providing a relatively small set of predictions according to the inferred model (for further assessment and use). Hence two challenges are posed. We will refer to the number of parameters in the pairwise model as $\mathcal{P}$ and the number of retained predictions as $\mathcal{K}$.

The first challenge is learning a pairwise model from data. Straightforward implementation of maximum likelihood (ML) estimation is not computationally feasible for the data

[*] zhouhj@itp.ac.cn
[†] eaurell@kth.se

of current interest. Therefore, various approximations to ML have been used in DCA, starting with message passing [4] and mean-field approximation [5]. A different estimator called pseudolikelihood maximization (PLM) was introduced in the DCA field in [3] and is currently considered the best [23]. Given infinite samples, PLM has the important property of statistical consistency, meaning that it will, with probability one, give the same result as ML, provided the data were generated by a model in the exponential family. However, in all successful DCA applications to date, the scenario is very different: The number of samples is typically less or much less than that of the parameters in the pairwise model. These basic theoretical results therefore do not in themselves establish the practical superiority of PLM over other flavors of DCA; such a conclusion instead relies on empirical testing or on other arguments. Direct-coupling analysis was reviewed as a method in biological sequence analysis in [24] and more recently and more extensively from the methodological point of view in [22,23].

For the Ising model, several methods concerning this scenario have been introduced and rigorously analyzed in [25–28]. An assumption made at several places in these contributions, not otherwise made in DCA, is sparsity, i.e., that a fraction of the possible pairwise couplings is zero. A second assumption is use of a criterion common in the machine learning or statistics literature, that the goal of inference is to find the sparsity structure. For reasons we will return to below, this is different from the criterion which has been common in DCA. The first of these methods [25] relies on the same main method of pseudolikelihood maximization, but with a sparsity-promoting ℓ_1 regularization, which is well known to be inferior to ℓ_2 regularization in the context of DCA.

The authors of [25–28] proved strong results of the type that perfect structure recovery is achievable with only $N \sim \ln L$ samples when the interaction graph is sparse and there is a gap between the smallest nonzero coupling and the set of strictly zero couplings. The authors of [27] furthermore proved that the parameters of an Ising model can be learned by the RISE algorithm up to small ℓ_2 error and with high probability from $N \sim \ln L$ independent samples, without assuming a gap (see Theorem 1 therein). An analogous result was later shown also for the PLM algorithm in [28] (in section S1 of the Supplementary Material of that paper). These results, however, assume a bounded maximum interaction strength and bounded degree (β and d in [27,28]) which do not hold for the random power-law (RPL) and Sherrington-Kirkpatrick (SK) models we consider in this paper and which are, in our view, also unlikely to be present in real biological data. Nevertheless, these quite recent developments point to further possibilities in algorithm development and we do not exclude that the RISE and logRISE methods introduced in [27,28] may turn out to be practically competitive or superior to PLM or other DCA methods. Such conclusions would however have to rely on wider empirical testing and are out of the scope of our paper.

The second challenge is how to choose $\mathcal{K}$ to balance correctness and mistakes of predictions when learning cannot be perfect due to limited computation resource and/or limited number of samples. Thus it depends on both the learning scheme of choice and the criterion according to which the inference is evaluated. For *in silico* testing it is clear that comparison can (and should) be made between the inferred parameters and the model parameters from which the data were generated. Due to the high dimensionality of the problem, such comparisons can however be done in many different ways. For the case of protein structure, comparison has been made between inferred parameters and spatial distances in protein structures. It is well known that in this case only a small subset of leading parameters yield good predictions, and we will in the following adopt such a criterion also for *in silico* testing. Not much theoretic analysis has been done to date for such learning criteria, one exception being the regression model built to estimate prediction accuracy of two DCA schemes in [29] and the statistical characteristics of learning in [30].

Let us note that a commonly used rule of thumb in DCA has been to retain about as many predictions as the data dimension L. From the common sense point of view that one should not try to learn more features than one has independent samples, this is inappropriate unless the MSA is a square or a thin matrix, i.e., unless N is at least as large as L. For the genome-scale problems in [14] the MSA was a fat matrix of N about 10^3 and L about 10^5 and the fraction $\mathcal{K}/\mathcal{P}$ of predictions that could be retained, according to the common sense rule, would be less than 10^{-8}. In an extreme extrapolation that the genome of every living human being on earth was accurately sequenced, N would be 10^{10}, while L, if approximately every eighth nucleotide would vary, as has been estimated for the protein-coding part of the human genome [31], would be about 5×10^8. The resulting MSA would be thin, but the number of parameters $\mathcal{P}$ of the Potts model would be very large (about 2×10^{18}) and $\mathcal{K}/\mathcal{P}$ according to the common sense rule would again be not more than about 10^{-8}. From another point of view, both the rule of thumb and the common sense rule would be unnecessarily pessimistic when the scaling $N \sim \ln L$ derived in [25–28] would apply, but as discussed above, we do not know this to be the case for the data we consider.

The issue we address in this work is the following. Assume that DCA is to be scaled up to applications where L is order of 10^6 or larger (the whole human genome would be 5×10^8 or larger). This will not be possible at all using current DCA methods since it is already cumbersome to run even approximate inference methods on the data set whose L is about 10^5 (see [14,32]). However, since inference is only used to retain a very small set of leading predictions, it is conceivable that the problem can be dimensionally reduced before inference. We will introduce a straightforward scheme that makes this possible and show that it works on both *in silico* and real data.

The paper is organized as follows. In Sec. II we review the DCA approach and the PLM computational scheme. In Sec. III we formally introduce correlation-compressed direct-coupling analysis (CCDCA) as an inference procedure. We then test CCDCA on *in silico* data, with the models and principles discussed in Sec. IV and the results presented in Sec. V. In Sec. VI we present an example where epistatic couplings are inferred from a collection of whole-genome sequences of the human pathogen *Streptococcus pneumoniae*. We show that CCDCA finds essentially the same leading couplings as a much more demanding DCA-based method [14] (see also [32]).

We note that the current best-performing versions of DCA for the application in contact prediction of proteins are hybrid schemes that rely also on other information [19,33–35]. Although such schemes can probably be expected to outperform pure DCA also in other applications, we are in this paper only concerned with the performance of DCA and DCA-like procedures alone.

II. BRIEF SUMMARY OF DIRECT-COUPLING ANALYSIS

The basic assumption behind DCA is that samples are drawn from a probabilistic model of the Potts model type

$$P_{\{\boldsymbol{h},\boldsymbol{J}\}}(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\left(\beta\sum_i h_i(\sigma_i) + \beta\sum_{i<j} J_{ij}(\sigma_i,\sigma_j)\right). \quad (1)$$

Here $\boldsymbol{\sigma} \equiv (\sigma_1, \sigma_2, \dots, \sigma_L)$ denotes a configuration of the system and σ_i is the allele (state) of locus i; $\boldsymbol{h} \equiv \{h_i : i \in [1, L]\}$ denotes the set of external fields and $\boldsymbol{J} \equiv \{J_{ij} : 1 \leqslant i < j \leqslant L\}$ denotes the set of pairwise couplings; the parameter β (inverse temperature) is introduced for later convenience and here just sets an overall scale of the energy terms. For later notational convenience, when $i > j$ we define $J_{ij}(\sigma_i, \sigma_j) \equiv J_{ji}(\sigma_j, \sigma_i)$. For simplicity, in the next section and the rest of this section, we give formulas for the Ising model, i.e., the two-state Potts model in the Ising gauge [4,5,36]). Hence, for all allowed i and j, $\sigma_i = \pm 1$, $h_i(\sigma_i) = h_i\sigma_i$, and $J_{ij}(\sigma_i, \sigma_j) = J_{ij}\sigma_i\sigma_j$.

Given N observed samples $\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}, \dots, \boldsymbol{\sigma}^{(N)}$, the ML estimation means to minimize the objective function

$$\begin{aligned} f(\boldsymbol{h},\boldsymbol{J}) &\equiv -\frac{1}{N}\sum_{n=1}^{N}\ln P_{\{\boldsymbol{h},\boldsymbol{J}\}}(\boldsymbol{\sigma}^{(n)}) \\ &= \ln Z - \beta\sum_i h_i\langle\sigma_i\rangle - \beta\sum_{i<j} J_{ij}\langle\sigma_i\sigma_j\rangle, \end{aligned} \quad (2)$$

where $\langle\cdot\rangle$ means averaging over all the N samples. The optimal values of the parameters $\boldsymbol{h}$ and $\boldsymbol{J}$ are determined by the variational conditions

$$\frac{1}{Z}\frac{\partial Z}{\partial h_i} = \beta\langle\sigma_i\rangle, \quad (3a)$$

$$\frac{1}{Z}\frac{\partial Z}{\partial J_{ij}} = \beta\langle\sigma_i\sigma_j\rangle. \quad (3b)$$

This requires calculating the partition function $Z(\boldsymbol{h},\boldsymbol{J})$ and its first derivatives and therefore renders the straightforward (brute-force or Monte Carlo) implementation of ML only feasible for small systems. Therefore, various approximate implementations have been proposed, reviewed in [22,23].

Besides likelihood, pseudolikelihood can also be used for inference [37,38]; the resultant approach is called pseudolikelihood maximization and is also reviewed in [22,23]. Pseudolikelihood maximization amounts to maximizing conditional probabilities for each locus simultaneously or separately and uses configurations rather than statistics in the inference process. Among pure DCA, PLM is now considered as the best-performing one [33,35].

For the Ising model, the conditional probability of observing one locus σ_i given the observation of all the other loci $\boldsymbol{\sigma}_{\setminus i}$ is

$$P_{\{h_i,\boldsymbol{J}_i\}}(\sigma_i|\boldsymbol{\sigma}_{\setminus i}) = \frac{1}{1+\exp(-2\beta\sigma_i\theta_i)}, \quad (4)$$

where $\boldsymbol{J}_i$ denotes $\{J_{ij} : j \neq i\}$, $\boldsymbol{\sigma}_{\setminus i} \equiv \boldsymbol{\sigma} \setminus \sigma_i$ denotes the state of all the other $(L-1)$ loci, and $\theta_i = h_i + \sum_{j\neq i} J_{ij}\sigma_j$ is the instantaneous field on locus i. The corresponding objective function for locus i in PLM is

$$f_i^{\mathrm{PLM}}(h_i,\boldsymbol{J}_i) = \frac{1}{N}\sum_{n=1}^{N}\ln\left[1+\exp\left(-2\beta\sigma_i^{(n)}\theta_i^{(n)}\right)\right]. \quad (5)$$

We can minimize these L objective functions simultaneously (symmetric PLM [3]) or separately by removing the constraint $J_{ij} = J_{ji}$ (asymmetric PLM [36]). Since the separate optimizations will usually give $J_{ij} \neq J_{ji}$ with finite samples, in [36] the output of asymmetric PLM is taken to be $J_{ij}^{\mathrm{PLM}} \equiv (J_{ij} + J_{ji})/2$. Asymmetric PLM can be easily parallelized and allows for considerable computational speedup as the optimization problems are also smaller; asymmetric PLM gives almost the same accuracy for predictions [36] and we will therefore here also use asymmetric PLM.

Regularization is widely used in DCA literature, partly to avoid overfitting and partly as a heuristic device to make the algorithm return a finite and stable answer. As in [36], the ℓ_2 regularization is used here; the regularization term for f_i^{PLM} reads

$$R_i(h_i,\boldsymbol{J}_i) = \lambda_h h_i^2 + \frac{\lambda_J}{2}\sum_{j\neq i} J_{ij}^2. \quad (6)$$

The factor 2 appears because J_{ij} is present in both f_i^{PLM} and f_j^{PLM}. It has been observed many times that the identity and order of the largest inferred couplings in DCA often do not depend much on regularization; some further examples in this direction are given in Appendix C.

III. DATA COMPRESSION BEFORE INFERENCE

Although pseudolikelihood maximization and other approximate inference methods can handle systems much larger than those for which full maximum likelihood is feasible, they still cannot be applied to very large systems. Therefore, further approximations and/or simplifications are called for. Here we formally introduce correlation-compressed direct-coupling analysis to address this issue. The principle of CCDCA is illustrated in Fig. 1.

A list-based presentation of CCDCA, appropriate for Ising data, is as follows.

(i) Given an $N \times L$ MSA data matrix A, first compute the covariance matrix C with $C_{ij} = \langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle$ being the correlation between the two loci i and j.

(ii) Find the m largest elements (either positive or negative) of the matrix C and then identify the ℓ loci which appear in these m elements. Obviously $\ell \leqslant 2m$.

(iii) Retain these ℓ loci and eliminate all the others. The original MSA matrix A is then reduced to a smaller $N \times \ell$ MSA matrix B. This correlation-compressed matrix B then serves as input for DCA analysis.

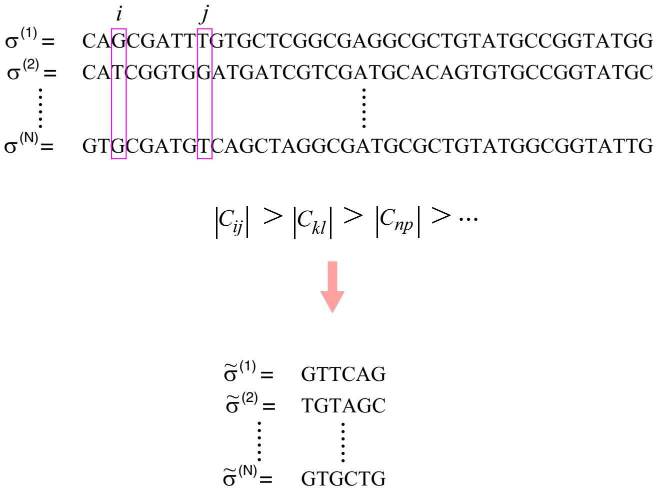

i
j
σ(1) = CAGCGATTTGTGCTCGGCGAGGCGCTGTATGCCGGTATGG
σ(2) = CATCGGTGGATGATCGTCGATGCACAGTGTGCCGGTATGC
σ(N) = GTGCGATGTCAGCTAGGCGATGCGCTGTATGGCGGTATTG
|Cij| > |Ckl| > |Cnp| > ···
σ̃(1) = GTTCAG
σ̃(2) = TGTAGC
σ̃(N) = GTGCTG

FIG. 1. Illustration of CCDCA as applied to nucleotide sequence data. The input MSA data set is an $N \times L$ matrix A, with N the total number of sample sequences ($\boldsymbol{\sigma}^{(\alpha)}$, $\alpha = 1, 2, \ldots, N$) and L the length of each sequence. Each entry is a sequence letter (one of the four nucleotides A, G, C, and T). In the data analyzed in Sec. VI one of the sequence letters can also be N (anything). The covariance matrix C_{ij} between any two loci i and j is then computed by reading the ith and the jth column of matrix A, and the column pair (i, j) is given a score. In Sec. VI the score used is MI. The m column pairs of largest scores are selected. After considering the $\ell \leqslant 2m$ loci involved in these m pairs, the original MSA matrix A is reduced to another $N \times \ell$ MSA matrix B for further analysis by some flavors of DCA.

Our approach is to first reduce the MSA based on measured correlations and only then apply a DCA scheme such as PLM. The idea is hence to take "direct" in the acronym DCA both literally and seriously. We call DCA on the correlation-compressed data CCDCA. When the flavor of DCA is PLM we thus alternatively call the resulting algorithm correlation-compressed pseudolikelihood maximization (CCPLM). This is the flavor assessed and used in the following sections of the paper; two other flavors of DCA (naive mean-field inversion and regularized least squares) are considered in Appendix B.

As a historical note we note that an approach similar to CCDCA was in fact used in the very first papers on DCA [4], but has not been used in the DCA literature since. The motivation in [4] was that the message-passing flavor of DCA used at that time did not scale up well to values of L on the order of 100. This shortcoming was later overcome using other flavors of DCA. Our motivation is in many ways similar in spirit in that no version of DCA currently in use scales up to L of the order of 10^6 or beyond.

IV. TEST SETS AND EVALUATION PROCEDURES

The study of inverse Ising and inverse Potts problems began about a decade ago stimulated by early results in neuroscience [39], before it was widely appreciated that the success of DCA on practical data sets rests on 'both' the inference procedure 'and' the choice to retain only some largest predictions. Controlled tests were done generating data from some known distribution and then checking how well the parameters could be recovered. Most of these tests were done using the root-mean-square (rms) criterion and $\boldsymbol{J}$ matrices generated by some variant of the SK model [40] (see [2] for an early review, [38,41] for two representative examples at the time, and [23] for a recent comprehensive survey).

Neither of these choices is however suitable as to how DCA methods are currently used. The rms criterion includes all predictions in the J matrix rather than just the largest ones, and therefore does not reflect how well a method recovers the leading couplings. In a standard SK model on L spins with Gaussian couplings, the typical values of the couplings scale as $1/\sqrt{L}$ and the largest values follow a Gumbel extreme value distribution with size about $\sqrt{\ln L}/\sqrt{L}$. All the couplings are then of very similar values, so this should in fact be a very challenging case, possibly much more so than realistic data. We will for completeness and back-compatibility also consider this model, but current practice in DCA additionally calls for other model classes and other evaluation criteria, as we will now discuss.

A. Random power-law test model class

As another test model class, relatively simple to describe, we propose the RPL model class as follows.

(a) The magnitudes of the elements of $\boldsymbol{J}$ are generated according to a power-law distribution, with a probability density function $\rho(x) = cx^{-\gamma}$ for x in some interval $x_l \leqslant x \leqslant x_u$. The exponent γ is tunable and c is a normalization constant. If $\gamma > 1$, then $c = (1-\gamma)/[x_u^{1-\gamma} - x_l^{1-\gamma}]$.

(b) The signs of the elements of $\boldsymbol{J}$ can be chosen either all positive (ferromagneticlike model) or randomized (spin-glass-like model).

(c) All elements are generated as independently and identically distributed random variables with the above characteristics. For the Ising model ($q = 2$) this just means that the coupling coefficients J_{ij} are independently and identically distributed random variables as above, while for Potts models ($q > 2$) we take all the elements of the $q \times q$ coupling matrix between any two loci as independently and identically distributed random variables.

Obviously many similar distributions could be considered, e.g., relaxing the biologically questionable assumption that the elements in a $q \times q$ coupling matrix are independent, but such extensions will be left for future work. The essence of the RPL class is that the coupling constants are widely distributed in size.

B. Sherrington-Kirkpatrick test model class

We also consider the more traditional SK spin-glass model, where conventional DCA methods have been extensively tested *in silico* in the past. In Sec. V we apply CCPLM on SK model data. The coupling constants J_{ij} in this model will be quenched independently and identically distributed Gaussian random variables with mean zero and variance $1/L$. The coupling constants are hence narrowly distributed around zero; there are no exceptionally strong interactions.

C. Evaluation by scatter diagrams

When testing DCA procedures on *in silico* data, a most natural graphical procedure is by scatter diagrams. By this measure the inferred value of an interaction coefficient is given as the ordinate (value on the y axis) plotted against the true value given as the abscissa (value on the x axis). If the inference procedure is accurate, the points will lie along the diagonal ($y = x$). If there are systematic differences between large interactions and other interactions, as there will be in the test cases described below, this will show up as deviations of the data cloud from the diagonal.

D. Evaluation by true positive rate

A general evaluation procedure when using DCA on real data was introduced in [4] and has been used since in most empirical work and DCA applications. It has however not been equally used in testing on model classes and we will therefore introduce it formally.

(i) Generate coupling coefficients or matrices J_{ij} according a preferred scheme, in our case the RPL or SK as in the previous sections.

(ii) Draw N independent samples from the Gibbs-Boltzmann distribution with those model parameters. In practice this has to be done with the Markov-chain Monte Carlo (MCMC) and may have issues with convergence for strongly coupled systems (low temperature). In the tests below we will therefore limit ourselves to weakly coupled systems (high temperature).

(iii) Consider the two lists

$$\mathcal{J}^{\text{true}} = \left|J_{ij}^{\text{true},1}\right| \geqslant \left|J_{ij}^{\text{true},2}\right| \geqslant \cdots \left|J_{ij}^{\text{true},k}\right|,$$
$$\mathcal{J}^{\text{pred}} = \left|J_{ij}^{\text{pred},1}\right| \geqslant \left|J_{ij}^{\text{pred},2}\right| \geqslant \cdots \left|J_{ij}^{\text{pred},k}\right|$$

of the k strongest true interactions and the k strongest predicted interactions, where $|\cdot|$ is a suitable norm. Compare the lists and determine how many elements l they have in common. The true positive rate (TPR) of the k strongest couplings is then defined as

$$\text{TPR}(k) \equiv \frac{l}{k}. \tag{7}$$

E. Evaluation by visualization

In Sec. VI below we consider real data where the true couplings are unknown. In fact, it is then not known if it is a good approximation to assume that the data have been generated from a Potts model or what model class describes the data at all. In a recent paper [14] couplings were inferred by a modified DCA procedure and then discussed from the viewpoint of plausibility and relevance in the light of how the data had been obtained and known facts of *S. pneumoniae* biology. In this work we compare results from CCDCA to those of [14] by a visual procedure where couplings are displayed as lines in a circular plot and the darkness of a line is proportional to the coupling strength. The strongest inferred couplings thus stand out as isolated black lines on a gray background formed by many weaker inferred couplings. The circular plots are produced by CIRCOS [42].

Given a list of scored couplings, evaluation by visualization proceeds as follows.

(a) Partition the whole genome into nonoverlapping windows of size 100 base pairs (bp).

(b) Couplings connected between the same two windows are replaced by a coarse-grained coupling, the score of which is simply the sum of scores of couplings to be replaced. The two end points of coarse-grained coupling are the beginning of the two windows.

F. Evaluation of CCDCA on *in silico* data

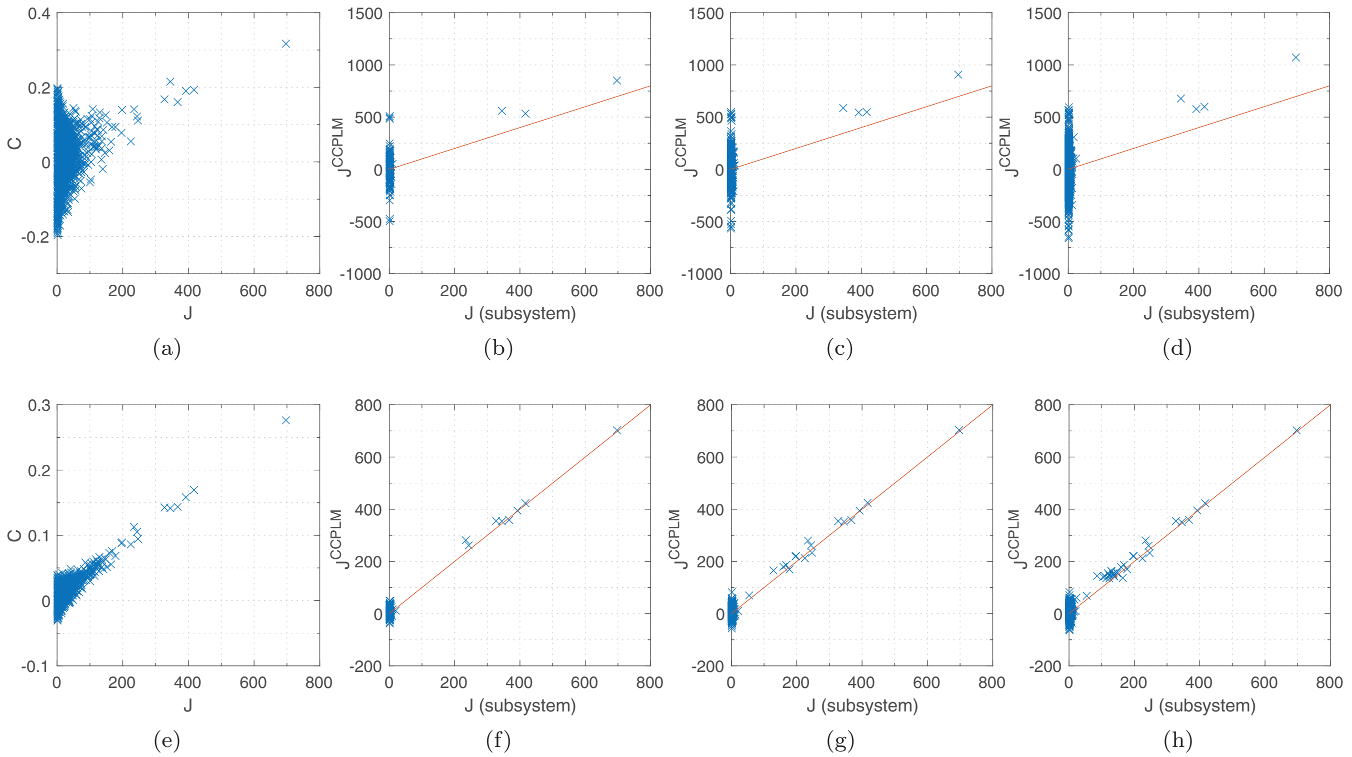
We evaluate our CCDCA method as follows. First we generate coupling coefficients according to model test classes such as RPL or SK (described in Sec. IV) and then we generate N independent samples from the Gibbs-Boltzmann distribution. This yields an $N \times L$ MSA which we call data matrix A. We apply DCA on A to get the values of all couplings and compute a true positive rate $\text{TPR}^{A}(k)$. For this to be feasible L cannot be very large, as discussed above.

The CCDCA method consists in reducing A to a smaller data matrix B on which we run DCA. This leads to a new set of couplings obtained by CCDCA and to new true positive rates $\text{TPR}^{B}(k)$. The evaluation of the data reduction scheme then proceeds by comparing the couplings obtained from DCA and CCDCA in a scatter diagram and by comparing $\text{TPR}^{A}(k)$ to $\text{TPR}^{B}(k)$. Obviously, such an evaluation can only be done on relatively small systems as otherwise we could not run the DCA on the huge matrix A at once. If it works, it would however support the idea to use the same procedure on very large data sets.

V. RESULTS ON *IN SILICO* DATA

In this section we describe the results of CCDCA on the RPL and SK models. The data dimension L is 1024. For RPL we use the power-law exponent $\gamma = 3$, lower cutoff $x_l = 1$, and upper cutoff $x_u = \infty$. For the regularization, we use throughout $\lambda_h = \lambda_J = 0.01$. Other choices of regularization parameters are discussed in Appendix C and are shown to have only small effects on the result of the inference, in agreement with the literature. Further parameter choices are discussed together with the presentations of the results. Some results on the SK model with planted couplings are presented separately in Appendix D.

We schematically show results for the ferromagnetic and spin-glass couplings and for the severely undersampled and slightly undersampled problems and different levels of compression in the CCDCA step. For the ferromagnetic case the signs of all the Ising terms in Eq. (1) are positive, while for the spin-glass case they have random signs. For the ferromagnetic model the critical temperature T_c was estimated to be around 1900, while for the spin-glass model the T_c was estimated to be around 120 (see Appendix A). We note again that β is here not a physical temperature, but only sets an overall scale of the couplings. We here report results only from the high-temperature regime where $T \equiv \beta^{-1}$ is larger than T_c by some margin and so we expect that in all cases considered the MCMC will converge fast enough such that the sampled configurations obey the Gibbs-Boltzmann equilibrium



(a) (b) (c) (d)

(e) (f) (g) (h)

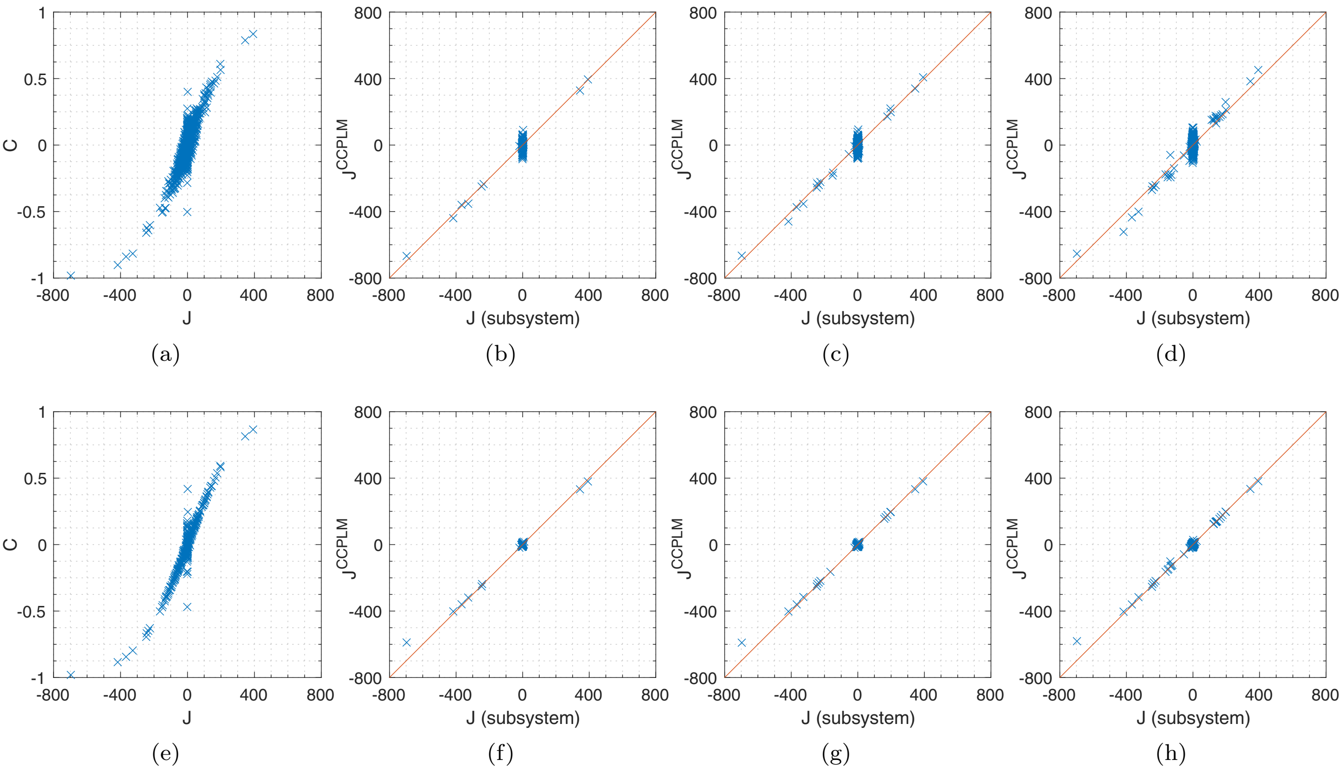
FIG. 2. Scatter diagrams of inferred couplings vs true couplings for ferromagnetic RPL data. The number of spins is $L = 1024$ and the number of samples (obtained at $T = 2500$) is (a)–(d) $N = 0.5L$ and (e)–(h) $N = 16L$. (a) and (e) Scatter diagrams of the covariance elements C_{ij} as the inferred couplings. Also shown are scatter diagrams of the inferred couplings being the CCPLM results at different levels of compression: CCPLM on MSA from the subsystem of size (b) $\ell = 16$ (with $m = 8$), (c) $\ell = 32$ (with $m = 16$), (d) $\ell = 64$ (with $m = 32$), (f) $\ell = 16$ (with $m = 8$), (g) $\ell = 31$ (with $m = 16$), and (h) $\ell = 61$ (with $m = 32$).

distribution. In the results shown here the severely undersampled case has $N = L/2$ configurations, i.e., the MSA is a fat matrix of shape 1:2. The other case, also undersampled, analogously has $N = 16L$; the MSA is a thin matrix of shape 16:1. Both settings are undersampled because N is much less than the total number of model parameters $\mathcal{P}$, which is of order L^2.

Figures 2 and 3 show the performance of CCDCA as scatter diagrams and also the performance of using bare correlations as predictors for the coupling coefficients. We see that CCPLM has similar performance to PLM in identifying the strongest interactions in the system. When the number of sampled configurations is relatively large (i.e., $N = 16L$) the quantitative predictions by CCPLM on the strongest interactions are rather accurate, even though the subsystem contains only very few loci of the original system [Figs. 2(e)–2(h) and 3(e)–3(h)]. When the configurations are severely undersampled (i.e., $N = 0.5L$) there is a high danger of false-positive DCA results (namely, some weak interactions were predicted to be strong); but even in this difficult case the values of the few strongest coupling constants are still predicted relatively accurately by the CCPLM method [Figs. 2(a)–2(d) and 3(a)–3(d)]. The couplings J_{ij} in the RPL instances have quite different values and some of them are very strong (e.g., up to $J_{ij} \approx 700$). The correlations C_{ij} between the strongly interacting loci i and j are then naturally quite strong too. Indeed, for the strongest couplings in the spin-glass case [Figs. 3(a) and 3(e)], the scatter diagram of C_{ij} vs J_{ij} practically falls on a single curve, though not on a straight line. For this reason the strongest covariance coefficients can alone serve as good indicators of the strongest direct interactions in the RPL class. The additional advantage of CCPLM (and full PLM) is that then the strengths of the strongest direct interactions can also be estimated, as one can see from the practically straight lines in Figs. 2(f)–2(h), 3(b)–3(d), and 3(f)–3(h).

Figure 4 displays the same data as true positive rates. For the severely undersampled cases [Figs. 4(a) and 4(c)] CCPLM is basically able to retrieve to the leading (largest) couplings as well as full PLM, while for couplings beyond the compression threshold CCPLM falls below the other curves. Bare correlation analysis works for these instances almost as well as full PLM, a result that can also be deduced from, in particular, Fig. 3(a). Qualitatively the same behavior is also found for the better-sampled data [Figs. 4(b) and 4(d)]. For the better-sampled spin-glass RPL data [Fig. 4(d)] correlations alone are quite good predictors of the identity of the strongest coupled pairs, a result which can also be read off from Fig. 3(e). As discussed above, the actual values of the couplings are less well predicted by bare correlations, with more scatter or more nonlinear deviations away from the diagonal in the scatter diagrams (Figs. 2 and 3).

We then turn to applying CCPLM to the SK spin-glass model. As Fig. 5(a) suggests, the covariance C_{ij} scales roughly linearly with the coupling constant J_{ij} in the high-temperature region, but there is a high degree of dispersion



(a) (b) (c) (d)

(e) (f) (g) (h)

FIG. 3. Same as Fig. 2 but for a spin-glass RPL instance ($L = 1024$ and temperature $T = 300$). The number of samples is (a)–(d) $N = 0.5L$ and (e)–(h) $N = 16L$. The size of the subsystem is (b) $\ell = 16$ (with $m = 8$), (c) $\ell = 29$ (with $m = 16$), (d) $\ell = 57$ (with $m = 32$), (f) $\ell = 16$ (with $m = 8$), (g) $\ell = 31$ (with $m = 16$), and (h) $\ell = 56$ (with $m = 32$).

due to undersampling of equilibrium configurations (here we use $N = 16L$). If the sampled configurations are used by PLM to infer the coupling constants, the qualitative agreement with the true values is better but not perfect [Fig. 5(b)]. Results in this direction were obtained already in the early DCA literature (cf. [38,41]) and have recently been developed further [23,43]. We here apply CCPLM on the subsystem corresponding to the m strongest covariance elements. The inference results for the subsystem of sizes $\ell = 16$ (for $m = 8$), $\ell = 31$ (for $m = 16$), and $\ell = 62$ (for $m = 32$) are shown in Figs. 5(c), 5(d), and 5(e), respectively. The predicted values of the coupling constants in these subsystems are in good agreement with the true values. The CCPLM method therefore is capable of identifying the strongest interactions also in these systems, but the inferred values of the coupling coefficients are less accurate than in the RPL class.

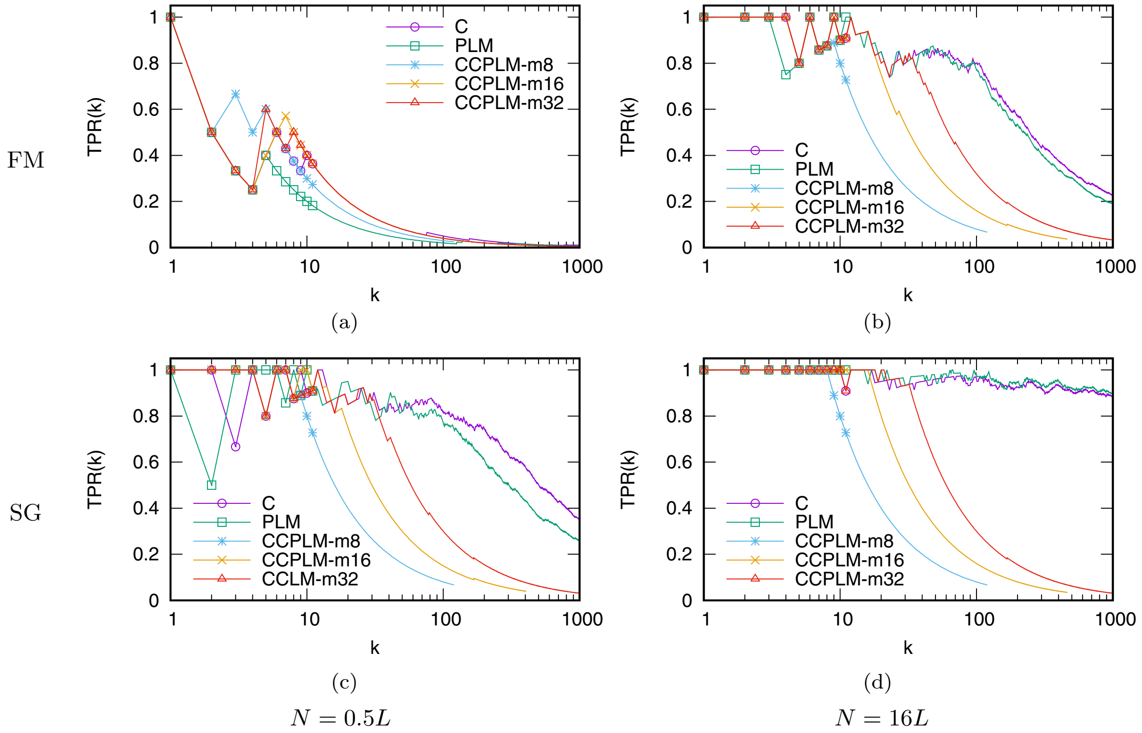
In the main text of the paper we consider only PLM, but as shown in Appendix B, our CCDCA method can also be combined in the same way with other DCA methods such as naive mean-field inversion [44] and regularized least squares [45]. In Appendix B we also discuss the effect of temperature on the inference performance.

VI. RETRIEVAL OF EPISTATIC COUPLINGS FROM WHOLE-GENOME BACTERIAL DATA BY CCDCA

In this section we discuss retrieving couplings on the genome scale from real data. The general biological term for combinatorial effects in fitness is epistasis [46]. All the settings where DCA has been applied can be considered as special cases of epistasis, generated by the physical interactions of residues in a protein or by any other mechanism. As in the DCA literature overall, we here assume that inferred Ising or Potts parameters directly measure epistasis. The phenomenon of correlated variations between loci in data is called linkage disequilibrium [47]. Linkage disequilibrium can be due both to epistasis and to shared ancestry of loci at close enough genomic positions. In the following we will separate long-range couplings, which are unlikely to result from shared ancestry, from short-range couplings, where linkage disequilibrium could be caused by both epistasis and shared ancestry.

In recent years data sets have been obtained on whole genomes of samples from entire bacterial populations. Typically, for these data sets, L is not larger than a few million (size of a bacterial genome) and N is not larger than a few thousand (largest current data set). In practice, genomes in naturally occurring organisms only vary on a subset of all positions, thus the number of varying loci L may reduce to a few hundred thousand. Still, the number of Potts model parameters to describe a distribution over 100 000 loci would be on the order of 10^{10} and to learn such models directly from data is very challenging.

In two recent contributions PLM was used to analyze epistasis in the human pathogen *S. pneumoniae* (the pneumococcus). In the first approach [14] the pneumococcal genome was split into about 1500 chunks. One locus was randomly selected from each chunk and PLM was run on this (much reduced) set, then run again on a new random selection, and so on. A putative interaction was scored by how many times it appeared in the lists of top interactions, each of which was



(a) (b)

(c) (d)

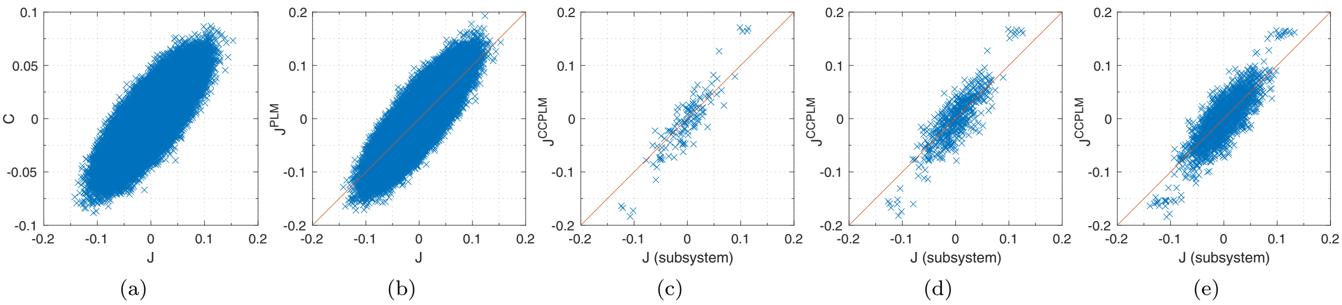
$N = 0.5L$ $N = 16L$

FIG. 4. True positive rates [Eq. (7)] for the random power-law model, with couplings being [(a) and (b)] ferromagnetic (FM) and [(c) and (d)] spin-glass (SG). The data dimension $L = 1024$ and the power-law exponent $\gamma = 3$. Results are shown for sample numbers (a) and (c) $N = 0.5L$ and (b) and (d) $N = 16L$. Temperatures are $T = 2500$ for the ferromagnetic case and $T = 300$ for the spin-glass case, in both cases well above the estimated critical temperature. The elements of the coupling matrix $\boldsymbol{J}$ are ranked according to the covariance matrix C (circles), the PLM predictions on the whole system (squares), or the PLM predictions on the correlation-compressed subsystem constructed using $m = 8$ (stars), $m = 16$ (crosses), and $m = 32$ (triangles) strongest covariance elements.

learned from a random selection. This scheme requires many such samples, in practice several tens of thousands. In the second approach [32] an optimized version of PLM was run on all the loci at once and the inferred Potts model parameters were used as in standard DCA. Both methods yield very similar results, but both also lead to substantial computation time. We will here see how well CCDCA manages on this challenging real-world data set, assuming that the results from [14] can be taken as ground truth. Evaluation will be the visual comparison as described in Sec. IV E.

A. Preparation of data

The data contain the genome alignment for 3156 isolates of *S. pneumoniae* downloaded from the data repository [48]. The nucleic acid codes contained in these data are A, C, G, T, and N (with N meaning complete uncertainty on the type of nucleic acid). The length of sequences is 2 221 315 bp. Thus the data are severely undersampled. A Potts model fitted to these data (with gauge chosen) would have $(q-1)^2 \times L(L-1)/2$ parameters for pairwise interaction and $L(q-1)$ parameters



(a) (b) (c) (d) (e)

FIG. 5. Comparison of PLM and CCPLM on the SK model. There are $L = 1024$ spins and $p = L(L-1)/2$ coupling constants. A total number of $N = 16L$ independent equilibrium configurations are sampled at temperature $T = 2$. (a) Relation between the covariance element C_{ij} and the true coupling constant J_{ij}. (b) Relation between the predicted coupling constant J_{ij}^{PLM} and the true value J_{ij} for the whole system. (c)–(e) Relation between the predicted coupling constant and the true coupling constant for the subsystem of size (c) $\ell = 16$ (obtained by considering the $m = 8$ strongest covariance elements), (d) $\ell = 31$ (for $m = 16$), and (e) $\ell = 62$ (for $m = 32$).

for local fields, where $q = 5$ and $L = 2\,221\,315$, in total about 4×10^{13} parameters.

As said above, genomes in naturally occurring organisms only vary on a subset of all positions. Fixed sites in genome will show little correlations to others, thus will convey little information about epistasis. By removing some frozen sites (discussed below), we can significantly reduce the system size to be considered. Skwark *et al.* [14] implement a filtering procedure, which includes removal of frozen loci, before applying the DCA-derived method. Not only to reduce the system size to be considered, but also to achieve a direct comparison of CCDCA with DCA-derived method used in [14], we first use the same criteria to filter the data, i.e., to remove loci that are not bi-allelic and loci that lack information. For each locus, ignoring N, we denote the most common letter (among A, C, G, and T) as major and the second most common one as minor; when counters of letters are equal, alphabetical order is used. The filtering criteria are as follows.

(i) Remove multi-allelic loci. A locus is considered as multi-allelic when the counter of the third most common letter (among A, C, G, and T) is not zero.

(ii) Remove frozen loci. A locus is considered as frozen when its minor allele frequency (MAF) is less than 0.01. For bi-allelic loci, the MAF is computed by

$$\mathrm{MAF} = \frac{\text{minor}}{\text{major} + \text{minor}}. \tag{8}$$

(iii) Remove loci which have high uncertainty. A locus is considered as highly uncertain when its frequency of the letter N is larger than $500/3156 \approx 0.158$.

Among the 2 221 315 loci, 2 177 096 loci are bi-allelic, out of which 113 237 loci have MAF at least 0.01; moreover, we got 81 506 loci after removing 31 731 highly uncertain ones. By the filtering procedure we reduce the number of states q from 5 to 3: N, major, and minor. In the context of statistical physics, the resulting MSA data set is a collection of 3156 configurations for a $q = 3$ Potts model with 81 506 nodes; by construction major is the most common symbol at all loci and we therefore (trivially) expect to find everywhere the inferred external field favoring the state major.

As a simple correction to sampling bias of biological sequence data, reweighting is widely used in DCA literature [3–5,36]. We also apply it here. After reweighting with threshold $x = 1$ (namely, if $k \geqslant 2$ rows of the $3156 \times 81\,506$ MSA matrix are identical, only one of them is kept while the other $k-1$ rows are eliminated), the number of configurations went from 3156 to 3145, i.e., only a very small change.

B. Results

All results presented in this section have been obtained from the code available at GitHub [49]. Apart from our central computational pipeline (CCPLM) used to obtain the data of Fig. 6, the GitHub repository also holds code to compute correlations (CC) for Fig. 8 and a reimplementation (PLM) of plmDCA [36], which can run on whole bacterial genomes. That code can be run to directly obtain data similar to those in Fig. 7. For back-compatibility, and since our focus in this paper is CCDCA, we show in Fig. 7 instead a different visualization of the data published in [14].

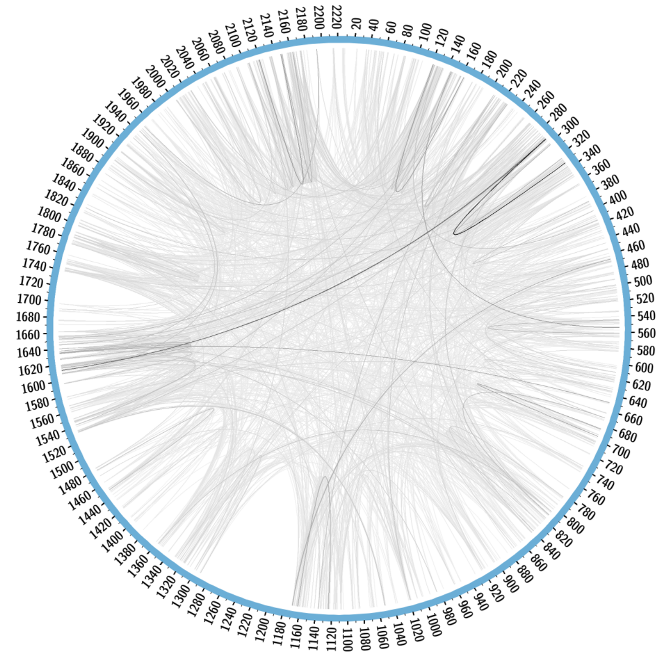


FIG. 6. The 6003 long-range couplings among the 1.2×10^5 strongest ones identified by CCPLM with 3×10^4 largest correlations. Numbers outside the rim indicates genomic positions in units of 1000 bp. The darkness of the lines represents the strength of couplings. The number of loci involved is 9304. Short-range couplings, the distance of which is smaller than 10 000 bp, are not shown here. (See Fig. 11 in Appendix E for the visualization of all 1.2×10^5 strongest couplings.) Positions 293, 332, and 1613 are the genes pbp2x, pbp1a, and pbp2b, respectively.

To quantify correlations of two loci by a scalar, we use as in [4] the mutual information (MI). The first step in CCDCA is hence to find for each pair of loci a real number which is the MI between corresponding two columns in the MSA, then to order the pairs by these numbers in descending order, and then to identify the set of loci which are members in the list of m top-ranking pairs. On this subset of loci (MSA B) we then run DCA. We use as the underlying DCA scheme the asymmetric version of PLM [36] with hyperparameters $\lambda_h = 0.1$ and $\lambda_J = 0.05$. The inferred couplings between loci i and j are scored by a modified Frobenius norm where the state N is not counted, i.e.,

$$\mathcal{S}_{ij} = \sqrt{\sum_{s_i=2}^{3} \sum_{s_j=2}^{3} J_{ij}^2(s_i, s_j)}, \tag{9}$$

where $s_i = 1, 2, 3$ represents the locus i being N, major, and minor, respectively (and so does s_j) and the coupling matrix $J_{ij}(s_i, s_j)$ is in the Ising gauge [3,4]. This scoring scheme is analogous to the plmDCA20 method described in [50] [where only amino-acid residues (nongap states) were included in the scoring], which was there shown to improve the accuracy of contact prediction in a large test set of protein families.

The results obtained by CCPLM from 9304 loci involved in the 3×10^4 strongest correlations are shown in Fig. 6. Numbers outside the rim indicate genomic positions in units of 1000 bp and the numbering goes through the whole S.

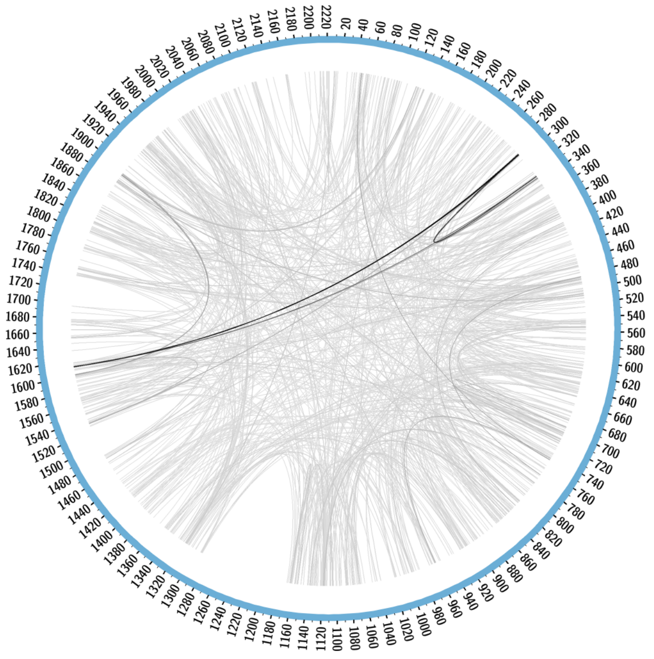


FIG. 7. The 5199 long-range strong couplings identified in [14] depicted with the same visualization procedure as for Fig. 6. Besides the links among genes pbp2x, pbp1a, and pbp2b, one can also identify the link to dyr (position 1530) and the triad of interactions involving divIVA (position 1600), pspA (position 120), and a site upstream of gene ply (position 1890).

pneumoniae genome presented in the data. Lines connecting genomic positions indicate the 6003 long-range couplings among the 1.2×10^5 strongest ones identified by CCPLM with 3×10^4 largest correlations. The darkness of lines represents the strength of couplings. Here we only show the long-range couplings, the distance of which is at least 10 000 bp (Fig. 11 in Appendix E shows the visualization including short-range ones). The details of visualization are described above in Sec. IV E.

As a comparison, the results reported in [14] are revisualized in Fig. 7 with the same procedure. The interactions between genes pbp2b and pbp2x as well as between pbp2x and pbp1a are immediately visible in both figures. It is also possible to identify other links discussed in [14] (see the caption to Fig. 7) as well as the characteristic absence of couplings involving loci at positions 1170–1290.

As another comparison, the 1.2×10^5 strongest correlations of the 9304-locus subsystem are also visualized in Fig. 8, where the short-range ones are not shown again. The link between genes pbp2x and pbp1a is also visible here. However, in Fig. 8 there are many other links not identified as significant in Figs. 6 and 7. They are presumably spurious because they do not stand out among correlations of the whole system (Fig. 12 in Appendix E). As noted in Sec. III, correlations can be caused by both direct and propagated couplings. By applying DCA, we can further identify direct couplings among strong correlations.

To demonstrate the robustness of CCDCA with the number of top correlations used, we also show the results of CCDCA with 1×10^4 largest correlations in Appendix E. The appearance of figures is quite similar.

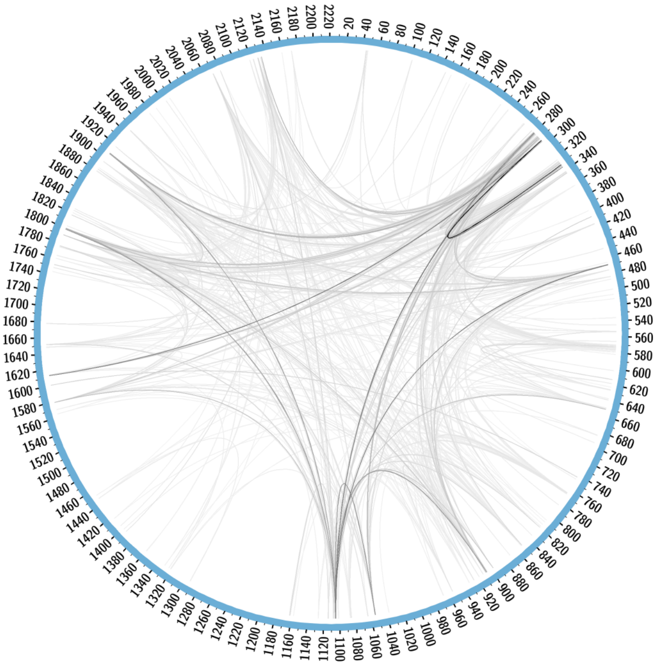


FIG. 8. The 19 224 long-range correlations among the 1.2×10^5 strongest ones of the subsystem from which the results shown in Fig. 6 are obtained.

In conclusion, the agreement between results obtained by CCDCA and the DCA-derived method in [14] should be deemed fair, especially given that CCDCA here represents a very significant simplification of the computational task. Although some results from [14] (Fig. 7) can also be identified directly from correlations (Fig. 8), the agreement between CCDCA and DCA is better overall.

VII. DISCUSSION

We have in this work introduced correlation-compressed direct-coupling analysis as a convenient method to detect the strongest direct interactions from data sets (MSAs) so large that direct application of DCA is cumbersome or not feasible. We have validated this method on synthetic data sampled from the random power-law model and standard Sherrington-Kirkpatrick model, as well as (in Appendix D) the SK model with some additional planted large couplings. Results are good to very good for all cases tested. We have also shown that CCDCA allows one to recover, at very low computational overhead, the results on whole-genome bacterial population-wide sequence data obtained in [14].

A large amount of work on inference from biological data has been done in neuroscience and related fields (see, e.g., [51–53]). Usually some nodes are then hidden from experimental measurement. This poses challenges different from those in the study of sequence data, where all nodes can be observed, but where the system size can be very large. The latter is the scenario we have addressed here.

We have in this work so far not given detailed performance measures since the components of CCDCA either are standard (computation of covariance matrices) or have been amply documented in the earlier literature (using PLM on a correlation-compressed MSAs). For the data sizes tested in the present paper the main computational bottleneck of CCDCA is to compute the covariance matrix based on the empirical data. Since the calculations of correlations are independent of each other, this task can be easily parallelized. For the MSA data set after filtering in Sec. VI the total time used to compute all the correlations by a MATLAB implementation available at GitHub [49] was about half an hour on a 56-core server with four Intel Xeon E7-4850 v3 processors, which translates to about 30 core hours. The runtime memory used is about 70 GB when storing all correlations in memory. In practical applications this task can be further simplified by maintaining a running list of the m strongest correlations and discarding all the other elements (or several running lists for parallelized implementation). By comparison, the results in [14] required approximately 500 000 core hours. Results from PLM, the reimplementation of plmDCA which we provide in [49], required on the order of 20 000 core hours, while the results from another and more code-optimized reimplementation of plmDCA presented in [32] required on the order of 10 000 core hours. The CCDCA thus transforms DCA with single-nucleotide resolution on the genome scale from something that requires a sizable compute cluster to something that can be done in a reasonable time on a stand-alone desktop computer.

A theoretically and conceptually interesting point which we leave to future work is a more detailed comparison of the distributions of the couplings obtained from correlation analysis, from PLM, and from CCPLM.

In summary we have demonstrated a means of application of DCA-like methods to very large data sets of biological interest by using intelligent preprocessing to reduce computational costs by a large factor.

ACKNOWLEDGMENTS

This research was supported by the Chinese Academy of Sciences (CAS) Presidents International Fellowship Initiative Grant No. 2016VMA002 (E.A.), by the Academy of Finland through its Centre of Excellence COIN through Grant No. 251170 (E.A.), by the National Science Foundation of China through Grants No. 11747601 and No. 11421063, and by Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-SYS018. We thank Pan Zhang for inspiring discussion and Marcin Skwark and Martin Weigt for discussions and valuable comments on the manuscript. The numerical computation was carried out partly at the HPC cluster of ITP-CAS.

APPENDIX A: TEMPERATURE CHOICES

Since the power-law distribution involved in RPL models is not bounded above and the exponent is 3, the variance is not finite. Therefore, we cannot rescale couplings with respect to system size, as in the Sherrington-Kirkpatrick model, to make the free energy intensive. As a result, the phase diagram can only be determined instance by instance. In Fig. 9 we show the phase diagrams of instances used for ferromagnetic and spin-glass RPL models. For each instance, we explore three temperatures: One is apparently lower than the transition point, one is close, and one is apparently higher. Accordingly, the temperatures chosen for the ferromagnetic RPL instance are 1500, 2000, and 2500; for the spin-glass RPL instance, we choose $T = 100$, 120, and 300.

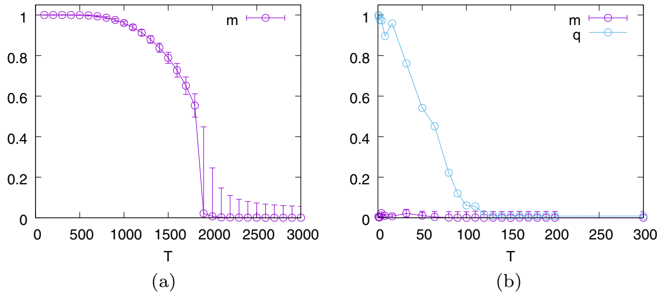


FIG. 9. Phase diagrams of RPL instances used: (a) FM RPL and (b) SG RPL. Here $m \equiv \frac{1}{L}\sum_{i=1}^{L} m_i$ denotes the system's magnetization and $q \equiv \frac{1}{L}\sum_{i=1}^{L} |m_i|$ denotes the schematic order parameter for a spin-glass model, where $m_i = \frac{1}{N}\sum_{b=1}^{N} \sigma_i^b$.

APPENDIX B: DETAILED COMPARISONS OF RPL DATA

In Figs. S1–S12 of [54] we present detailed comparisons between DCA and CCDCA on RPL data. Figures S1–S6 show comparisons on few-sample data ($N = 0.5L$), whereas Figs. S7–S12 show comparisons on many-sample data ($N = 16L$).

Besides the DCA flavor used in the main text, PLM, two more DCA flavors are considered here: naive mean-field inversion (NMFI) [44] and regularized least squares (RLS) [45]. The NMFI gives couplings by the formula

$$J_{ij}^{\text{NMFI}} = -\frac{1}{\beta}(\boldsymbol{C}^{-1})_{ij}. \tag{B1}$$

The NMFI is not applicable when the rank of the data matrix is not high enough, i.e., when the covariance matrix is not invertible. The RLS gives couplings by the formula

$$J_{ij}^{\text{RLS}} = -\frac{1}{\beta}[(\boldsymbol{C}^2 + \lambda\mathbf{1})^{-1}\boldsymbol{C}]_{ij}. \tag{B2}$$

With positive λ, this formula avoids the conditioning problem of inverting the covariance matrix. The RLS can also be considered as an ℓ_2-regularized NMFI. Here $\lambda = \lambda_J/2 = 0.005$.

Due to the difficulty of sampling according to Gibbs-Boltzmann distribution at low temperature, some many-sample data of the spin-glass RPL instance do not contain enough independent samples and thus are not feasible for NMFI. So comparisons of NMFI and CCNMFI are only performed on all many-sample data of the ferromagnetic RPL instance and high-temperature many-sample data of the spin-glass instance, as shown in Figs. S7–S9 and S12. For data which are not feasible for NMFI, the results of RLS are also bad, especially on the data of spin-glass RPL instance.

Comparing Figs. S1–S6 with Figs. S7–S12, we can find that a few samples are usually not enough to identify even the largest coupling unless the temperature is high. In Fig. S1b all TPR curves fall to zero because temperature is low and samples are few: The quality of data is so poor that many correlations paired with small couplings are stronger than correlations paired with large couplings; with more samples, as shown in Fig. S7, all methods are able to identify leading couplings with more samples. The oscillation of TPR curves indicates that couplings of similar strength are ranked in the wrong order, e.g., in Fig. S7b TPR obtained by the C method (ordering couplings by correlations) drops at $k = 2$ because the method ranks the third largest couplings above the second largest one. In Fig. S4b the hill-like TPR curve obtained by the C method indicates that the method cannot identify the largest coupling but finds some other large ones; similar hill-like TPR curves appear in Figs. S5, S10, and S11.

According to Figs. S1–S12, we can conclude that leading couplings can be identified by CCDCA when they can be identified by DCA no matter the DCA flavor is PLM, RLS, or NMFI.

APPENDIX C: DEPENDENCE OF RESULTS IN THE MAIN TEXT ON THE REGULARIZATION STRENGTH

In the main text, the regularization strength of PLM $\lambda_h = \lambda_J = 0.01$. In Figs. S13–S16 of [54], we provide comparisons of PLM and CCPLM with three choices of regularization coefficient ($\lambda_h = \lambda_J \in \{0.1, 0.01, 0.001\}$); the conclusion that CCDCA has the similar ability to DCA concerning identifying leading couplings is not changed.

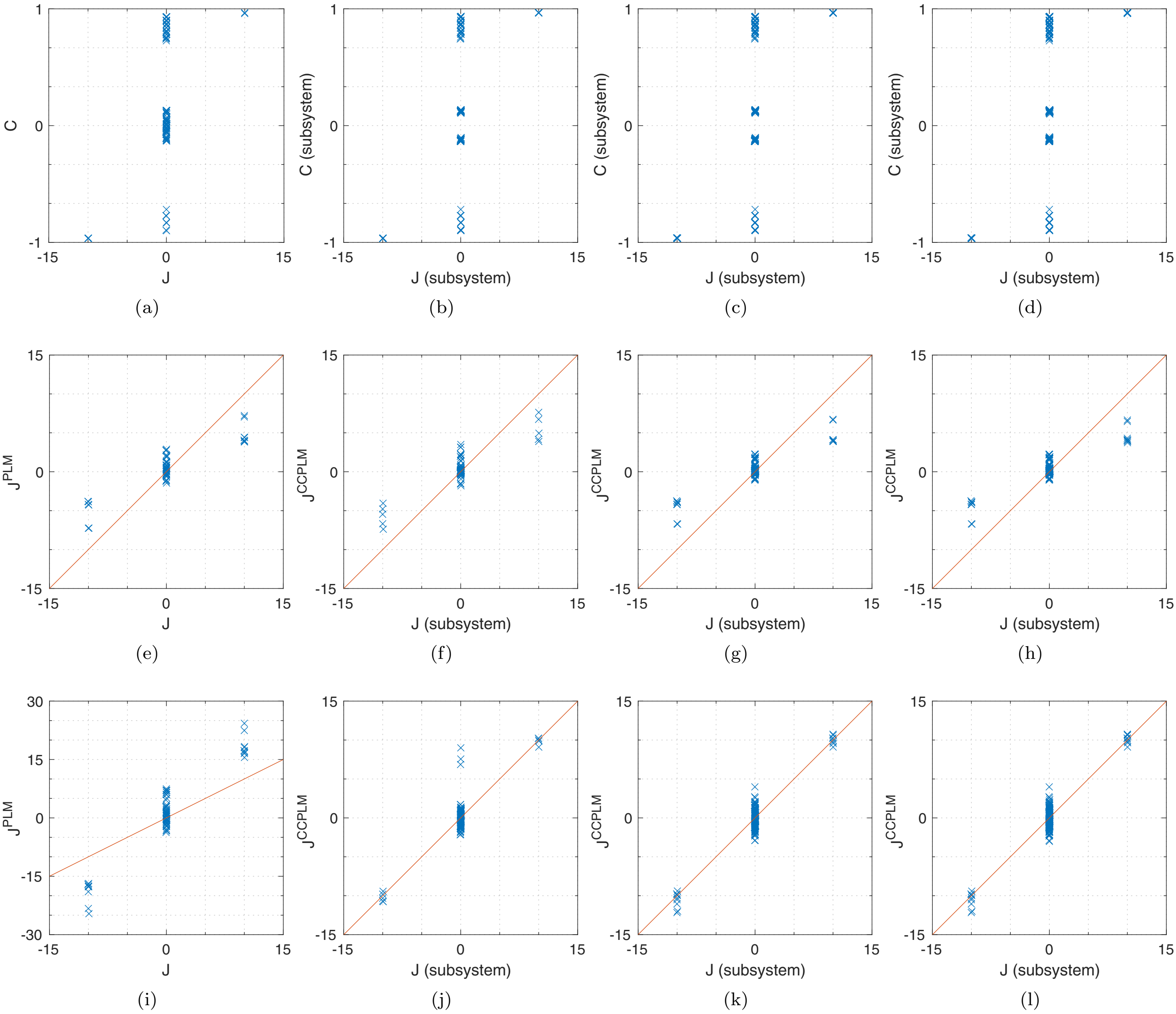


FIG. 10. The (a), (e), and (i) DCA and (b)–(d), (f)–(h), and (j)–(l) CCDCA on a chain SK model containing $L = 1024$ spins and $p = L(L-1)/2$ couplings J_{ij}. A total number of $N = 16L$ equilibrium spin configurations are sampled from this model at temperature $T = 5$. (a)–(d) Relation between coupling J_{ij} and covariance element C_{ij}. (e)–(h) and (i)–(l) Relation between the predicted coupling $J_{ij}^{\rm PLM}$ and the true coupling J_{ij}; the results are obtained with regularization parameter (e)–(h) $\lambda = 10^{-2}$ and (i)–(l) $\lambda = 10^{-5}$. The analysis is of (a), (e), and (i) the whole system and the subsystems with (b), (f), and (j) $\ell = 15$ (according to the $m = 8$ strongest covariance elements); (c), (g), and (k) $\ell = 19$ (for $m = 16$); and (d), (h), and (l) $\ell = 20$ (for $m = 32$) spins.

APPENDIX D: CCDCA ON THE SK MODEL WITH PLANTED COUPLINGS

In this appendix we consider a simply modified SK model with planted couplings in the form of two chains of strongly interacting loci. The test will then be to see how well CCDCA can recover the planted couplings. The system is constructed as follows.

(a) Generate a graph for the SK model. The number of spins is $L = 1024$; each of $\{J_{ij}\}$ is an independently and identically distributed Gaussian random variable with mean zero and variance L^{-1}.

(b) Add one ferromagnetic chain of length 10 by modifying nine coupling constants as $J_{ij} \leftarrow J_{ij} + 10$ for $(i, j) \in \{(1, 2), (2, 3), \ldots, (9, 10)\}$.

(c) Add one antiferromagnetic chain of length 10 by modifying nine coupling constants $J_{ij} \leftarrow J_{ij} - 10$ for $(i, j) \in \{(11, 12), (12, 13), \ldots, (19, 20)\}$.

For the conventional SK model, the critical temperature $T_c = 1$. Since we modify only 18 of $1024 \times 1023/2 \approx 5 \times 10^5$ couplings in the system, when the temperature is much higher than T_c, most of the spins in the system are only weakly coupled, except those in the two chains. The spins in the two chains are strongly correlated even if they are not directly coupled with each other [Fig. 10(a)]. We can perform DCA analysis on the N sampled equilibrium configurations through PLM. This method assigns a value to each of the $\mathcal{P} = L(L-1)/2$ coupling constants. As demonstrated in Figs. 10(e) and 10(i), the performance of this method is relatively good even when the number of sampled configurations N is much smaller than the total number of parameters $\mathcal{P}$.

In the case of undersampling ($N \ll \mathcal{P}$) the aim is not so much to infer all the coupling constants but to identify the most significant interactions. For this latter task we can construct a subsystem by retaining only the spins involved in the strongest correlations. As demonstrated in Fig. 10 (second, third, and fourth columns), the CCPLM works fine for this problem instance. It is able to distinguish the true interactions even if the subsystem only contains from 15 to 20 spins.

APPENDIX E: ADDITIONAL RESULTS ON THE WHOLE-GENOME DATA SET

Here we present additional results on the real data.

First, as a supplement to Fig. 6, the visualization including short-range couplings is shown in Fig. 11, which visualizes all 1.2×10^5 strongest couplings identified by CCPLM with 3×10^4 largest correlations. This figure is almost the same as Fig. 6 except that short-range couplings are not filtered out but are depicted near the rim. As before, lines represent coarse-grained couplings and their darkness represents the strength. Here not only long-range couplings appear lighter, but also their difference in darkness becomes vague: Some short-range couplings are stronger than all long-range ones and thus limit the range of darkness for long-range ones. Since the length of genes is on the order of 10^4 bp and the focus of this work is epistasis, we rule out couplings shorter than 10^4 bp to make long-range strong couplings stand out.

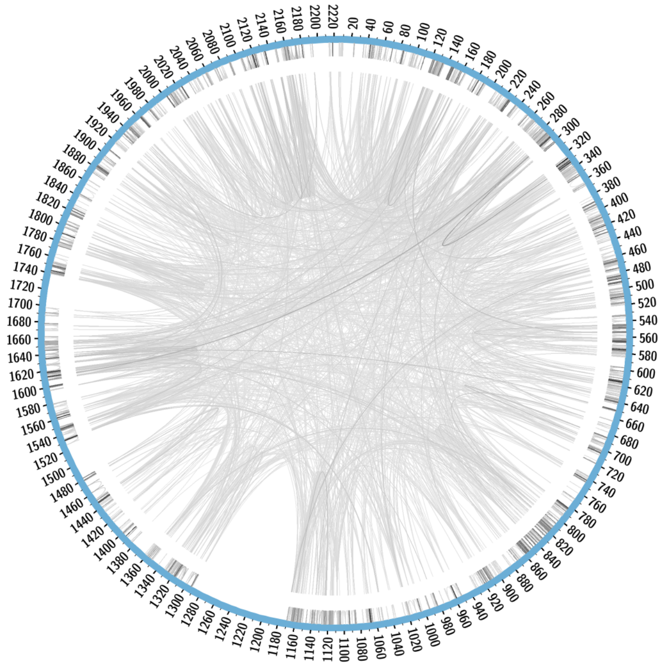


FIG. 11. The 1.2×10^5 strongest couplings identified by CCPLM with the 3×10^4 largest correlations. This figure supplements Fig. 6 by showing short-range couplings as well as long-range ones. Short-range couplings are depicted near the rim and long-range couplings are depicted inside; there is a visible margin between them.

Second, in Fig. 12 we show the 2423 long-range correlations among the 1.2×10^5 strongest ones of the whole genome after filtering. Comparing with Fig. 8, we can tell that some correlations, which are not strong in the whole genome, will stand out in the correlation-compressed subsystem. Then DCA justifies its ability to distinguish direct couplings from

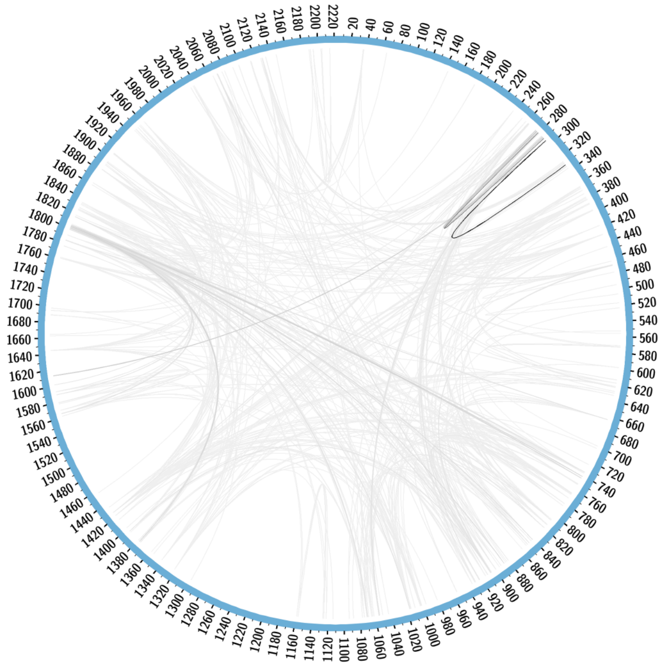


FIG. 12. The 2423 long-range correlations among the 1.2×10^5 strongest ones of the whole genome after filtering.

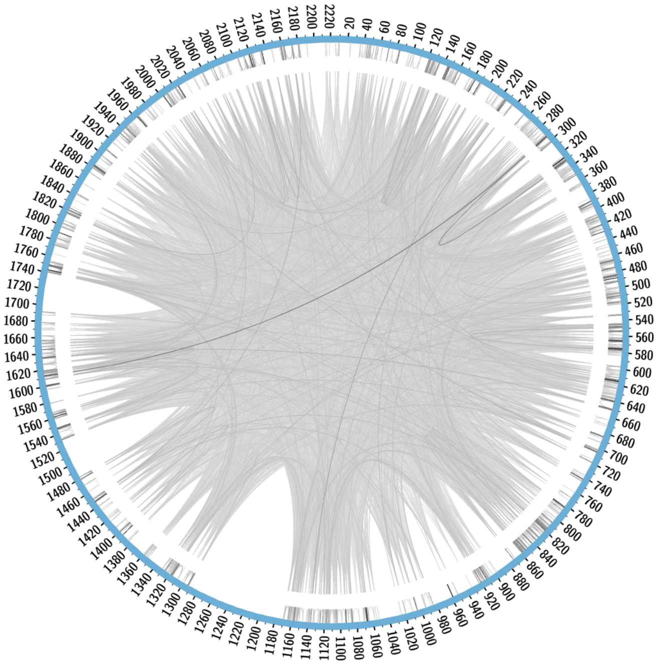


FIG. 13. The 1.2×10^5 strongest couplings identified by CCPLM with the 10^4 largest correlations.

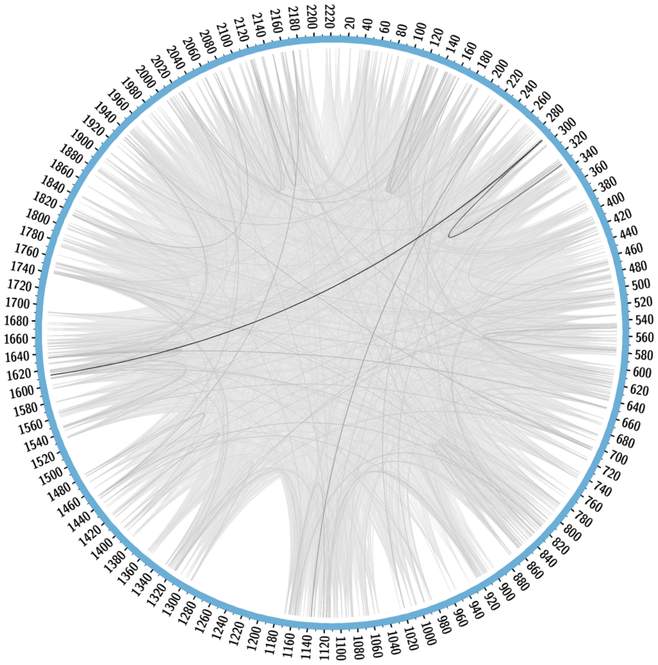


FIG. 14. The 52 790 long-range couplings among the 1.2×10^5 strongest ones identified by CCPLM with the 10^4 largest correlations.

propagated ones by filtering out these seemingly important correlations, as shown in Fig. 6.

Third, as a demonstration of CCDCA being robust to the number of correlations used, the results of CCDCA with the 10^4 largest correlations are shown in Figs. 13 and 14, which show visualization with and without short-range couplings, respectively. The lines between genes pbp2x and pbp1a as well as those between pbp2x and pbp2b stand out in both figures, as in Fig. 6.

[1] M. J. Wainwright and M. I. Jordan, Found. Trends Mach. Learn. **1**, 1 (2008).

[2] Y. Roudi, E. Aurell, and J. A. Hertz, Front. Comput. Neurosci. **3**, 1 (2009).

[3] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Phys. Rev. E **87**, 012707 (2013).

[4] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. USA **106**, 67 (2009).

[5] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).

[6] L. Burger and E. van Nimwegen, PLoS Comput. Biol. **6**, e1000633 (2010).

[7] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, Bioinformatics **28**, 184 (2012).

[8] E. De Leonardis, B. Lutz, S. Ratz, C. Simona, R. Monasson, M. Weigt, and A. Schug, Nucleic Acids Res. **43**, 10444 (2015).

[9] T. Gueudré, C. Baldassi, M. Zamparo, M. Weigt, and A. Pagnani, Proc. Natl. Acad. Sci. USA **113**, 12186 (2016).

[10] G. Uguzzoni, S. John Lovis, F. Oteri, A. Schug, H. Szurmant, and M. Weigt, Proc. Natl. Acad. Sci. USA **114**, E2662 (2017).

[11] C. Weinreb, A. J. Riesselman, J. B. Ingraham, T. Gross, C. Sander, and D. S. Marks, Cell **165**, 963 (2016).

[12] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, Mol. Biol. Evol. **33**, 268 (2016).

[13] T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Scharfe, M. Springer, C. Sander, and D. S. Marks, Nat. Biotechnol. **35**, 128 (2017).

[14] M. J. Skwark, N. J. Croucher, S. Puranen, C. Chewapreecha, M. Pesonen, Y. Y. Xu, P. Turner, S. R. Harris, S. B. Beres, J. M. Musser, J. Parkhill, S. D. Bentley, E. Aurell, and J. Corander, PLoS Genet. **13**, e1006508 (2017).

[15] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks, Cell **149**, 1607 (2012).

[16] S. Hayat, C. Sander, D. S. Marks, and A. Elofsson, Proc. Natl. Acad. Sci. USA **112**, 5413 (2015).

[17] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, Science **355**, 294 (2017).

[18] S. Ovchinnikov, H. Park, D. E. Kim, F. DiMaio, and D. Baker, Proteins **86**, 113 (2018).

[19] M. Michel, M. J. Skwark, D. Menéndez Hurtado, M. Ekeberg, and A. Elofsson, Bioinformatics **33**, 2859 (2017).

[20] M. Michel, D. Menéndez Hurtado, K. Uziela, and A. Elofsson, Bioinformatics **33**, i23 (2017).

[21] J. Söding, Science **355**, 248 (2017).

[22] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Rep. Prog. Phys. **81**, 032601 (2018).

[23] H. C. Nguyen, R. Zecchina, and J. Berg, Adv. Phys. **66**, 197 (2017).

[24] R. R. Stein, D. S. Marks, and C. Sander, PLoS Comput. Biol. **11**, e1004182 (2015).

[25] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, Ann. Stat. **38**, 1287 (2010).

[26] G. Bresler, *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, *STOC '15* (ACM, New York, 2015), pp. 771–782.

[27] M. Vuffray, S. Misra, A. Lokhov, and M. Chertkov, Adv. Neural Inf. Process. Syst. **29**, 2595 (2016).

[28] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov, Sci. Adv. **4**, e1700791 (2018).

[29] P. P. Wozniak, B. M. Konopka, J. Xu, G. Vriend, and M. Kotulska, Bioinformatics **33**, 3405 (2017).

[30] Y. Xu, E. Aurell, J. Corander, and Y. Kabashima, arXiv:1704.01459v1.

[31] Exome Aggregation Consortium, Nature (London) **536**, 285 (2016).

[32] S. Puranen, M. Pesonen, J. Pensar, Y. Y. Xu, J. A. Lees, S. D. Bentley, N. J. Croucher, and J. Corander, Microb. Genom. **4**, 1 (2018).

[33] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner, Bioinformatics **31**, 999 (2015).

[34] V. Golkov, M. J. Skwark, A. Golkov, A. Dosovitskiy, T. Brox, J. Meiler, and D. Cremers, Adv. Neural Inf. Process. Syst. **29**, 4222 (2016).

[35] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, PLoS Comput. Biol. **13**, e1005324 (2017).

[36] M. Ekeberg, T. Hartonen, and E. Aurell, J. Comput. Phys. **276**, 341 (2014).

[37] J. Besag, Statistician **24**, 179 (1975).

[38] E. Aurell and M. Ekeberg, Phys. Rev. Lett. **108**, 090201 (2012).

[39] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).

[40] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).

[41] V. Sessak and R. Monasson, J. Phys. A: Math. Theor. **42**, 055001 (2009).

[42] M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, Genome Res. **19**, 1639 (2009).

[43] J. Berg, J. Stat. Mech. (**2017**) 083402.

[44] H. J. Kappen and F. B. Rodríguez, Neural Comput. **10**, 1137 (1998).

[45] M. Andreatta, S. Laplagne, S. C. Li, and S. Smale, arXiv:1311.1301v3.

[46] H. J. Cordell, Hum. Mol. Genet. **11**, 2463 (2002).

[47] M. Slatkin, Nat. Rev. Genet. **9**, 477 (2008).

[48] M. J. Skwark, N. J. Croucher, S. Puranen, C. Chewapreecha, M. Pesonen, Y. Y. Xu, P. Turner, S. R. Harris, S. B. Beres, J. M. Musser, J. Parkhill, S. D. Bentley, E. Aurell, and J. Corander (2017), Data from: Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. Dryad Digital Repository. https://doi.org/10.5061/dryad.gd14g.

[49] https://github.com/gaochenyi/CC-PLM

[50] C. Feinauer, M. J. Skwark, A. Pagnani, and E. Aurell, PLoS Comput. Biol. **10**, e1003847 (2014).

[51] Y. Roudi, S. Nirenberg, and P. E. Latham, PLoS Comput. Biol. **5**, 1 (2009).

[52] Y. Roudi and G. Taylor, Curr. Opin. Neurobiol. **35**, 110 (2015).

[53] B. Bravi and P. Sollich, J. Stat. Mech. (**2017**) 063404.

[54] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.98.032407 for figures showing the results discussed in Appendixes B and C.