POLITECNICO DI TORINO
Repository ISTITUZIONALE

Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners

Original
Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners / Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, R., Pagnani, A.. - In: PLOS ONE. - ISSN 1932-6203. - ELETTRONICO. - 9:(2014), p. e92721. [10.1371/journal.pone.0092721]

Availability:
This version is available at: 11583/2537287 since: 2016-02-19T13:48:08Z

Publisher:
PLOS

Published
DOI:10.1371/journal.pone.0092721

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

07 October 2026

OPEN ACCESS Freely available online

PLOS ONE

Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners

Carlo Baldassi[1,2]◔, Marco Zamparo[1,2]◔, Christoph Feinauer[1], Andrea Procaccini[2], Riccardo Zecchina[1,2], Martin Weigt[3,4], Andrea Pagnani[1,2]*

1 Department of Applied Science and Technology and Center for Computational Sciences, Politecnico di Torino, Torino, Italy, **2** Human Genetics Foundation-Torino, Torino, Italy, **3** Sorbonne Universités, Université Pierre et Marie Curie Paris 06, UMR 7238, Computational and Quantitative Biology, Paris, France, **4** Centre National de la Recherche Scientifique, UMR 7238, Computational and Quantitative Biology, Paris, France

Abstract

In the course of evolution, proteins show a remarkable conservation of their three-dimensional structure and their biological function, leading to strong evolutionary constraints on the sequence variability between homologous proteins. Our method aims at extracting such constraints from rapidly accumulating sequence data, and thereby at inferring protein structure and function from sequence information alone. Recently, global statistical inference methods (e.g. direct-coupling analysis, sparse inverse covariance estimation) have achieved a breakthrough towards this aim, and their predictions have been successfully implemented into tertiary and quaternary protein structure prediction methods. However, due to the discrete nature of the underlying variable (amino-acids), exact inference requires exponential time in the protein length, and efficient approximations are needed for practical applicability. Here we propose a very efficient multivariate Gaussian modeling approach as a variant of direct-coupling analysis: the discrete amino-acid variables are replaced by continuous Gaussian random variables. The resulting statistical inference problem is efficiently and exactly solvable. We show that the quality of inference is comparable or superior to the one achieved by mean-field approximations to inference with discrete variables, as done by direct-coupling analysis. This is true for (*i*) the prediction of residue-residue contacts in proteins, and (*ii*) the identification of protein-protein interaction partner in bacterial signal transduction. An implementation of our multivariate Gaussian approach is available at the website http://areeweb.polito.it/ricerca/cmp/code.

Citation: Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, et al. (2014) Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. PLoS ONE 9(3): e92721. doi:10.1371/journal.pone.0092721

Editor: Kay Hamacher, Technical University Darmstadt, Germany

Received December 5, 2013; **Accepted** February 24, 2014; **Published** March 24, 2014

Copyright: © 2014 Baldassi et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Christoph Feinauer acknowledges funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007–2013/ under REA grant agreement No. 290038. Riccardo Zecchina and Carlo Baldassi acknowledge the European Research Council for grant No. 267915. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: andrea.pagnani@polito.it

◔ These authors contributed equally to this work.

Introduction

One of the most important challenges in modern computational biology is to exploit the wealth of sequence data, accumulating thanks to modern sequencing technology, to extract information and to reach an understanding of complex biological processes. A particular example is the inference of conserved structural and functional properties of proteins from the empirically observed variability of amino-acid sequences in homologous protein families, e.g. via the inference of signals of co-evolution between residues, which may be distant along the sequence, but in contact in the folded protein; cf. [1–6] for a selection of classical works and [7] for a review over recent developments. In the last 5 years, a strong renewed interest in residue co-evolution has been emerging: a number of global statistical inference approaches [8–16] have led to a highly increased precision in predicting residue contacts from sequence information alone. Furthermore, co-evolutionary analysis was found to provide valuable insight on specificity and partner prediction in protein-protein interaction [17,18] in bacterial signal transduction.

Key to this recent progress are *global statistical inference* approaches, like *direct-coupling analysis* (DCA) [8,10] and *sparse inverse covariance estimation* (PSICOV) [12], and the GREMLIN algorithm based on *pseudo-likelihood maximization* [11,16]. DCA is based on the maximum-entropy (MaxEnt) principle [19,20] which naturally leads to statistical models of protein families in terms of so-called Potts models or Markov random fields. Proposed initially more than a decade ago [21,22], it was not until very recently that the first successful MaxEnt approaches to the study of co-evolution were published [8,23]. The main idea behind such global inference techniques is the following: correlations between the amino-acids occurring in two positions in a protein family, i.e. between two columns in the corresponding multiple-sequence alignment (MSA), may result not only from direct co-evolutionary couplings. They may also be generated by a whole network of such couplings. More precisely, if a position i is coupled to a position j, and j is coupled to k, then i and k will also show some correlation

even if they are not coupled. The aim of global methods is to disentangle such direct and indirect effects, and to infer the network of direct co-evolutionary couplings starting from the empirically observed correlations.

In this context, we focus on two different biological problems: the inference of residue-residue contacts and the prediction of interaction partners.

The inference of residue-residue contacts from large MSAs of homologous proteins [8–16] is an important challenge in structural biology. Inferred contacts have been shown to be sufficient to guide the assembly of complexes between proteins of known (or homology modeled) monomer structure [24,25], and to predict the fold of single proteins [26–31], including highlights like large transmembrane proteins [28,31]. In [25], the predicted structure of the auto-phosphorylation complex of a bacterial histidine sensor kinase has been used to repair a non-functional chimeric protein by rationally designed mutagenesis; this structure is also, to the best of our knowledge, the first case of a prediction, which has subsequently been confirmed by experimental X-ray structures [32,33]. The possibility to guide tertiary and quaternary protein structure prediction is an important finding, in light of the experimental effort needed for generating high-resolution structures.

The second problem, concerning molecular determinants of interaction specificity of proteins and the identification of interaction partners [17,18], is a central problem in systems biology. In both cited papers, bacterial two-component signal transduction systems (TCS) were chosen, which constitute a major way by which bacteria sense their environment, and react to it [34]. TCS consist of two proteins, a histidine sensor kinase (SK) and a response regulator protein (RR): the SK senses an extracellular signal, and activates a RR by phosphorylation; the RR typically acts as a transcription factor, thus triggering a transcriptional response to the external signal. The same (homologous) phosphotransfer mechanism is used for several signaling pathways in each bacterium; thus, to produce the correct cellular response to an external signal, interactions have to be highly specific inside each pathway: crosstalk between pathways has to be avoided [35–37]. This evolutionary pressure can be detected by co-evolutionary analysis [17,18]. Results are interesting: statistical couplings inferred by DCA reflect physical interaction mechanisms, with the strongest signal coming from charged amino-acids. They are able to predict interacting SK/RR pairs for so-called orphan proteins (SK and RR proteins without an obvious interaction partner), and the predictions compared favorably to most available experimental results, including the prediction of 7 (out of 8 known) interaction partners of orphan signaling proteins in *Caulobacter crescentus* [18].

In the present study, we describe an alternative approach to co-evolutionary analysis, based on a multivariate Gaussian modeling of the underlying MSA. It can be understood as an approximation to the MaxEnt Potts model in which (*i*) the discreteness constraint is released, i.e. continuous values are allowed for variables representing amino-acids, (*ii*) a Gaussian interaction model is assumed, and (*iii*) a prior distribution is introduced to compensate for the under-sampling of the data. This simplification allows to explicitly determine the model parameters from empirically observed residue correlations. The approach shares many similarities with [12], in which a multivariate Gaussian model is also assumed, and with the mean-field approximation to the discrete DCA model [10], but the simpler structure of the probability distribution makes the model analytically tractable, and allows for an efficient implementation, while still having a prediction accuracy comparable or superior to that of the aforementioned models (see the Results section). The model is briefly described in the next section, and in greater detail in the Materials and Methods section.

A fast, parallel implementation of the multivariate Gaussian modeling approach is provided on http://areeweb.polito.it/ricerca/cmp/code in two different versions, a MATLAB [38] one and a Julia [39] one.

Gaussian Modeling of Multiple Sequence Alignments

This section briefly outlines the prediction procedure coming from our proposed model, and highlights its main distinctive features with respect to other similar methods. A full presentation can be found in the Materials and Methods section, and additional details in File S1.

The input data to our model is the MSA for a large protein-domain family, consisting of M aligned homologous protein sequences of length L. Sequence alignments are formed by the $Q=20$ different amino-acids, and may contain alignment gaps.

As in [12], we consider a multivariate Gaussian model in which each variable represents one of the Q possible amino-acids at a given site, and aim in principle at maximizing the likelihood of the resulting probability distribution given the empirically observed data (in particular, given the observed mean and correlation values, computed according to a reweighting procedure devised to compensate for the sampling bias). Doing so would yield the parameters for the most probable model which produced the observed data, which in turn would provide a synthetic description of the underlying statistical properties of the protein family under investigation. Unfortunately, however, this is typically infeasible, due to under-sampling of the sequence space. A possible approach to overcome this problem, used e.g. in [12], is to introduce a sparsity constraint, in order to reduce the number of degrees of freedom of the model. Here, instead, we propose a Bayesian approach, in which a suitable prior is introduced, and the parameter estimation is then performed over the posterior distribution.

A convenient choice for the prior is the normal-inverse-Wishart (NIW), which, being the conjugate prior of the multivariate Gaussian distribution, provides a NIW posterior. Thus, within this choice, the posterior simply is a data-dependent re-parametrization of the prior: as a result, the problem is analytically tractable, and the computation of relevant quantities can be implemented efficiently. Furthermore, by choosing the parameters for the prior to be as uninformative as possible (i.e. corresponding to uniformly distributed samples), we obtain an expression for the posterior which, interestingly, can be reconciled with the pseudo-count correction of [10]: in the Gaussian framework, the pseudo-count parameter has a natural interpretation as the weight attributed to the prior.

We then estimate the parameters of the model as averages on the posterior distribution, which have a simple analytical expression and can be computed efficiently (in practical terms, the computation amounts to the inversion of a $LQ \times LQ$ matrix). On one hand, this yields an estimate of the strengths of direct interactions between the residues of the alignments, which can be used to predict protein contacts. On the other hand, this allows to build joint models of interacting proteins, which can be used to score candidate interaction partners, simply by computing their likelihood - which can be done very efficiently on a Gaussian model.

The contact prediction between residues relies on the model's inferred interaction strengths (i.e. couplings), which are represented by $Q \times Q$ matrices; in order to rank all possible interactions, we need to compute a single score out of each such matrix. As

mentioned above, these matrices are numerically identical to those obtained in the mean-field approximation of the discrete (Potts) DCA model. We tested two scoring methods: the so-called direct information (DI), introduced in [8], and the Frobenius norm (FN) as computed in [15]. The DI is a measure of the mutual information induced only by the direct couplings, and its expression is model-dependent: in the Gaussian framework it can be computed analytically (see File S1) and yields slightly different results with respect to the Potts model (but with a comparable prediction power, see the Results section). The FN, on the other hand, does not depend on the model, and therefore some of the results which we report here for the contact prediction problem are applicable in the context of the Potts model as well. In our tests, the FN score yielded better results; however, the DI score is gauge-invariant and has a well-defined physical interpretation, and is therefore relevant as a way to assess the predictive power of the model itself.

Results

Residue-residue contact prediction

The aim of the original DCA publication [8] was the identification of inter-protein residue-residue contacts in protein complexes, more precisely in the SK/RR complex in bacterial signal transduction. More recently, global methods for inferring direct co-evolution attacked the problem prediction of intra-domain contacts for large protein domain families [9–16,26]. Thanks to the development of more efficient approximation techniques triggered by the wide availability of single-domain data on databases like Pfam [40], one can now easily undertake co-evolutionary analysis of a large number of protein families on normal desktop computer. To give a comparison, whereas the message-passing algorithm in [8] was limited to alignments with up to about 70 columns at a time (typically requiring some ad-hoc pre-processing of larger alignments to select the 70 potentially most interesting columns), the subsequent approaches easily handle MSA of proteins with up to ten times this number of columns.

In this context, our multivariate Gaussian DCA is particularly efficient: parameter estimation can be done explicitly in one step, and the computation of the relevant coupling measures such as the direct information (DI) and the log-likelihood also uses explicit analytical formulae. The analytical tractability of Gaussian probability distributions results in a major advantage in algorithmic complexity, and therefore in real running time. In the included implementation of the algorithm the largest alignment analyzed (PF00078, $L=214$ residues, $M=126258$ sequences) the DI is obtained in about 20 minutes, whereas a more typical alignment (e.g. PF00089, $L=219$, $M=15894$) is analyzed in less than a minute on a normal @2270 MHz Intel Core i5 M430 CPU on a Linux desktop. With respect to the computational complexity of the algorithm, the sequence reweighting step is $\mathcal{O}(M^2L)$ (since it requires a computation of sequence similarity for all sequence pairs in the MSA), while the model's parameters estimate is $\mathcal{O}(L^3)$ (since it requires to invert a covariance matrix whose size is proportional to L).

Here, we will show that this gain in running time has no detectable cost in terms of predictive power. To this aim, we first studied the prediction of intra-domain contacts (see Fig. 1). From the Pfam database [40], a set of 50 families was selected for which the number of representative sequences is high enough to allow for a meaningful statistical analysis (average length $\langle L\rangle=173.48$ residues, average number of sequences per alignment $\langle M\rangle=32660.2$), cf. the Methods section. For each family, 4 measures were determined: DI in mean-field approximation, DI and Frobenius norm (FN) in the Gaussian model, Average-product-corrected mutual information (MI) as described in [41]. As mentioned above, the FN in the Gaussian model is the same as that computed in the mean-field approximation of the discrete DCA model. Each measure was used to rank residue position pairs (only pairs which are at least 5 positions apart in the chain are considered), and high-ranking pairs are evaluated according to their spatial proximity in exemplary protein structures. A cutoff of 8 Å minimal distance between heavy atoms for contacts was chosen, in agreement with [10] and [42]. The best overall results are obtained with FN, as already noted in [15]; however, it is interesting to note that the Gaussian DI score is comparable to, and even slightly better then the mean-field DI score, which gives an important indication regarding the accuracy of the underlying probabilistic model: this in turn is relevant for subsequent analysis (see next section). Somewhat surprisingly, we also found that the optimal overall value of the pseudo-count parameter is strongly dependent on which scoring function is used: we explored the whole range $(0,1)$ in steps of 0.1, and found that the optimum for the FN score was at 0.8, while for the DI score it was at 0.2.

As a second test we ran on the same data-set a direct comparison between our method's best score, PSICOV [12] and plmDCA [15]. Fig. 2 shows that our method's performance is comparable to that of PSICOV (and even marginally better after the first 50 inferred couplings), and that the two methods are slightly better for the first 10 predicted contacts (with a 100% accuracy on the first contact). At ten predicted contacts, the true positive average is about 95% for all three methods. From ten predicted pairs on, both our method and PSICOV perform slightly worse than plmDCA: at 100 predicted contacts, the true positive rate is about 72% for PSICOV, 77% for the Gaussian model and 80% for plmDCA. A sample of running times for the three methods and different problem sizes, reported in Table 1, shows that our code can be at least an order of magnitude faster then PSICOV, and two orders of magnitude faster then plmDCA. These results suggest that our method is a good candidate for large scale problems of inference of protein contacts.

Visual inspection of the predicted contacts does not reveal any significant bias with respect to the residue position, nor with respect to the sencondary or tertiary structures of the proteins. As an example, in Fig. 3 we show the first 40 predicted contacts (39 out of which are true positives) for the protein familiy PF00069 (Protein kinase domain) using the Gaussian DCA methods with the FN score: the pictures seem to indicate a sparse, fair sampling across the set of all true contacts.

Finally, we have used the SK/RR data set containing 8,998 cognate SK/RR pairs, cf. Methods, to predict inter-protein residue-residue contacts. Results can be compared with those presented in [18], where the original message-passing DCA was applied to the same data-set, and 9 true contact prediction were reported before the first false positive appeared. In Fig. 4, results are shown for mean-field and Gaussian DCA, using the DI score: both methods improve substantially over the message-passing scheme (20 true positive predictions at specificity equal to one), but are highly comparable (with a little but not significant advantage of the Gaussian scheme). Again, we find that the improved efficiency and analytical tractability of Gaussian DCA comes at no cost for the predictive power.

Predicting interactions between proteins in bacterial signal transduction

A typical bacterium uses, on average, about 20 two-component signal transduction systems to sense external signals, and to trigger

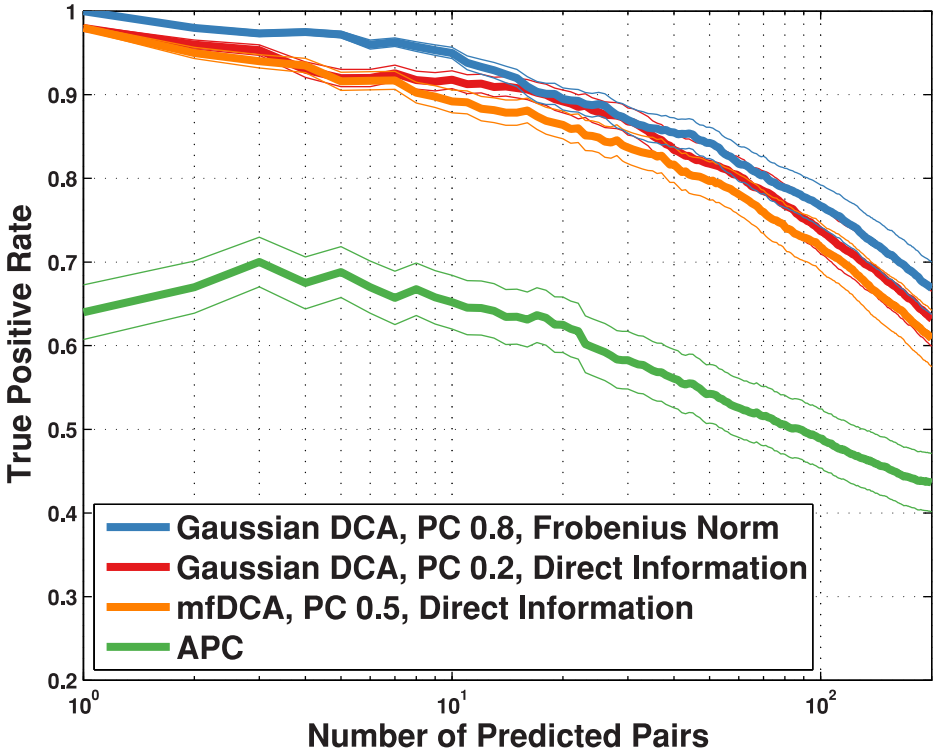

Figure 1. True positive rate plotted against number of predicted pairs. Results are shown for four different different scoring techniques: Frobenius norm (as described in [15], pseudo-count set to 0.8, blue); Gaussian direct information (as described in the text, APC-corrected, pseudo-count set to 0.2, red); mean-field direct information (as described in [10], pseudo-count set to 0.5, orange) and APC-corrected mutual information (as described in [41], green). The true positive rate is an arithmetic mean over 50 Pfam families (see Table 2 for the list); thin lines represent standard deviations.
doi:10.1371/journal.pone.0092721.g001

a specific response. In bacteria living in complex environments, the number of different TCS may even reach 200. While the signals and consequently the mechanisms of signal detection vary strongly from one TCS to another, the internal phosphotransfer mechanism from the SK to the RR, which activates the RR, is widely conserved across bacteria: A majority of the kinase domains of SK belong to the protein domain family *HisKA* (PF00512), all RR to family *Response_reg* (PF00072) [40], cf. the Methods section. Despite their closely related functionality, the interactions in the different pathways have to be highly specific, to induce the correct specific answer for each recognized external signal.

A big fraction of SK and RR genes belonging to the same TCS pathway are co-localized in joint operons; the identification of the correct interaction partner is therefore trivial: such pairs are called cognate SK/RR. However, about 30% of all SK and 55% of all RR are so-called orphan proteins: their genes are isolated from potential interaction partners in the genome. While a large fraction of the RR are expected to be involved in other signal-transduction processes like chemotaxis, for each of the SK at least one target RR is expected to exist. It is a major challenge in systems biology to identify these partners, and to unveil the signaling networks acting in the bacteria. A step in this direction was taken in [17,18], where co-evolutionary information extracted from cognate pairs is used to predict, with some success, orphan interaction partners.

An approach based on message-passing DCA [18] was tested in two well-studied model bacteria, namely *Caulobacter crescentus* (CC) and *Bacillus subtilis* (BS), where several orphan interactions are known experimentally [43–45]. The degree of accuracy of the method can be evinced from figure 4 of [18]: for CC, all known interactions between DivL, PleC, DivJ and CC_1062 with DivK and PleD are correctly reconstructed by the ranking obtained from the co-evolutionary scoring. Only in the case of the pair CenK-CenR, the signal is not sufficiently strong. For BS all the 5 orphan kinases KinA-B-C-D-E are known to interact with the RR Spo0F, which was clearly visible in co-evolutionary analysis in all but the KinB case.

The method proposed here for orphans pairing relies on the Gaussian approximation and on the definition of the score $\mathcal{L}$, cf. Eq. 15 in Methods, which equals the log-odds ratio between the probabilities of two orphan sequences in the interacting model (inferred from cognate SK/RR alignments) and a non-interacting model (inferred independently from the two MSAs of the SK and the RR families). It is worth stressing at this point that all estimates of the likelihood score parameters are learned only on the cognates set. Ranked by $\mathcal{L}$, orphans interactions in CC are shown in Fig. 5. Results are very similar to those mentioned for [18]: known interactions are well reproduced for orphan kinases PleC and DivJ, while for CC_1062 and DivL the signal for an interaction with DivK, though present, is less clear. Finally, predictions for CC_0586 are identical in both studies but neither one is able to identify the CenK-CenR interaction. Fig. 6 shows predictions for orphan interactions in BS: observed interactions between KinA, KinB, KinC, KinD, KinE and Spo0F are manifest. This means that while predictions in CC are slightly less accurate compared to the message-passing strategy, predictions in BS show a greater accuracy.

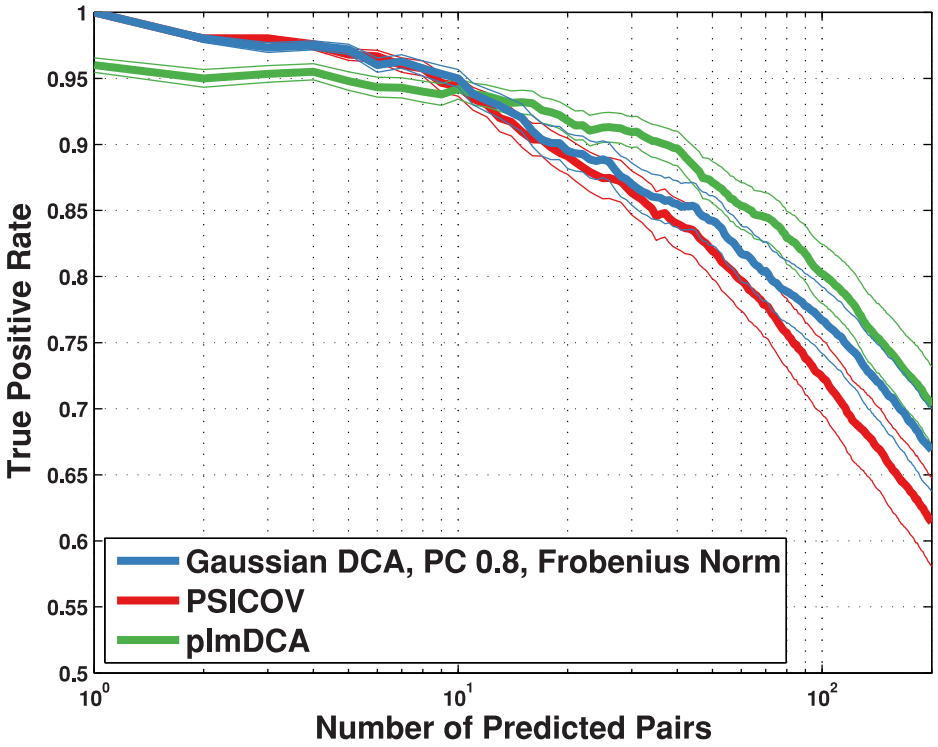

Figure 2. True positive rate plotted against number of predicted pairs. Data for plmDCA [15] (green) and PSICOV version 1.11 [12] (red) was obtained using the code provided by the authors with standard parameters as found in the distributed code, except that PSICOV was run with the -o flag to override the check against insufficient effective number of sequences. The true positive rate is an arithmetic mean over 50 Pfam families (see Table 2 for the list); thin lines represent standard deviations.
doi:10.1371/journal.pone.0092721.g002

Discussion

In this work we have derived a multivariate Gaussian approach to co-evolutionary analysis, whereby we cast the problem of the inference of contacts in MSAs, as well as candidate interacting partners within two MSAs of interacting proteins, into a simple Bayesian formalism, under the hypothesis of normal inverse Wishart distribution of the Gaussian parameters.

The major advantage of this method is the very simple structure of the resulting probability distribution, which allows to derive analytical expressions for many relevant quantities (e.g. likelihoods and posterior probabilities). As a result, the computations performed with this model can be very efficient, as demonstrated by the code accompanying this paper.

Furthermore, our tests indicate that the prediction accuracy of residue contacts using the Gaussian model is comparable or superior to that achieved using the mean-field Potts model of [10], or by using the PSICOV method of [12] with default settings; accuracy in pairing interaction partners is comparable to that achieved in [18].

Table 1. Running times in seconds for a representative sample of proteins with varying length (N) and sequences in alignment (M), using different algorithms.

	PF00014	PF00025	PF00026	PF00078
N	53	175	317	214
M	4915	5460	4762	172360
Gaussian DCA (parallel)	0.7	5.3	16.3	534.8
Gaussian DCA (non-parallel)	1.7	12.7	52.1	3583.4
PSICOV	11.7	1141.9	5442.7	10965.1
plmDCA	433.2	6980.7	37364.8	303331.0

Since the Gaussian DCA code is parallelized, we show two series of results, one in which we used 8 cores and one in which we forced the code to run on a single core, for the sake of comparing with the non-parallel code of PSICOV and plmDCA. These benchmarks were taken on a 48-core cluster of 2100.130 MHz AMD Opteron 6172 processors running Linux 3.5.0; PSICOV version 1.11 was used, compiled with gcc 4.7.2 at -O3 optimization level; plmDCA was run with MATLAB version r2011b. Gaussian DCA timings shown are taken using the Julia version of the code, using Julia version 0.2.
doi:10.1371/journal.pone.0092721.t001

The simplicity and tractability of the model also suggests further directions for improvement. For example, the whole posterior distribution of relevant observables such as the DI could be studied and, possibly, used to provide more insight into the kind of predictions presented here (in particular, it could be used to measure the confidence on the predictions). Also, suitably designed, more informative priors (e.g. carrying biologically relevant information) could further enhance the prediction power of the method, although it is not obvious how to set a prior directly on the predicted interaction strengths, whereas with other methods – notably plmDCA [15] and PSICOV [12] – this should be straightforward. Finally, we observe that the log-likelihood score for interaction partners does not require an interaction model to be known in advance: the interaction partners can be identified across the whole families by optimizing the score of the joint alignment as a function of the mapping between potentially interacting partners, thus allowing to infer both the interacting elements and their inter-protein contacts at once.

Materials and Methods

Data

Input data is given as multiple sequence alignments of protein domains. For the first question (inference of residue-residue contacts in protein domains), we directly use MSAs downloaded from the Pfam database version 27.0 [40,46], which are generated by aligning successively sequences to profile hidden Markov models (HMMs) [47] generated from curated seed alignments. We have selected 50 domain families, which were chosen according to the following criteria: (*i*) each family contains at least 2,000 sequences, to provide sufficient statistics for statistical inference; (*ii*) each family has at least one member sequence with an experimentally resolved high-resolution crystal structure available from the Protein Data Bank (PDB) [48], for assessing *a posteriori* the predictive quality of the purely sequence-based inference. The average sequence length of these 50 MSAs is $\langle L \rangle \simeq 173$ residues, the longest sequences are those of family PF00012 whose profile HMM contains 602 residues. The list of included protein domains, together with their PDB structure, is provided in Table 2.

Following [12], we discarded the sequences in which the fraction of gaps was larger then 0.9. However, in [12], an additional pre-processing stage was applied, in which a target sequence is chosen as the one for which prediction of contacts is desired, and all residue positions in the alignment (i.e. columns in the alignment matrix X) where the target sequence alignment has gaps are removed. We did not find this pre-processing step to improve the prediction, for either PSICOV or our model, and therefore all results presented in this work do not include this additional filtering.

For the second question (identification of interaction partners), we have used the data of [18], thus having the possibility to directly compare with previous results. In summary (for details see [18]), this data comes from 769 bacterial genomes, scanned using HMMER2 with the Pfam 22.0 HMMs for the Sensor Kinase (SK) domain *HisKA* (PF00512) and for the Response Regulator domain *Response_reg* (PF00072) [49], resulting in 12,814 SK and 20,368 RR sequences.

A total of 8,998 SK-RR pairs are found to be cognates, i.e. to be coded by genes in common operons, while the rest are so-called orphans. For statistical inference, cognates sequences are concatenated into a single MSA, each line containing exactly one SK and its cognate RR.

A binary representation of MSA

The data we use are MSAs for large protein-domain families. An MSA provides a $M \times L$-dimensional array $A = \left(a_l^m\right)_{l=1,\ldots,L}^{m=1,\ldots,M}$: each row contains one of the M aligned homologous protein sequences of length L. Sequence alignments are formed by the $Q = 20$ different amino-acids, and may contain alignment gaps,

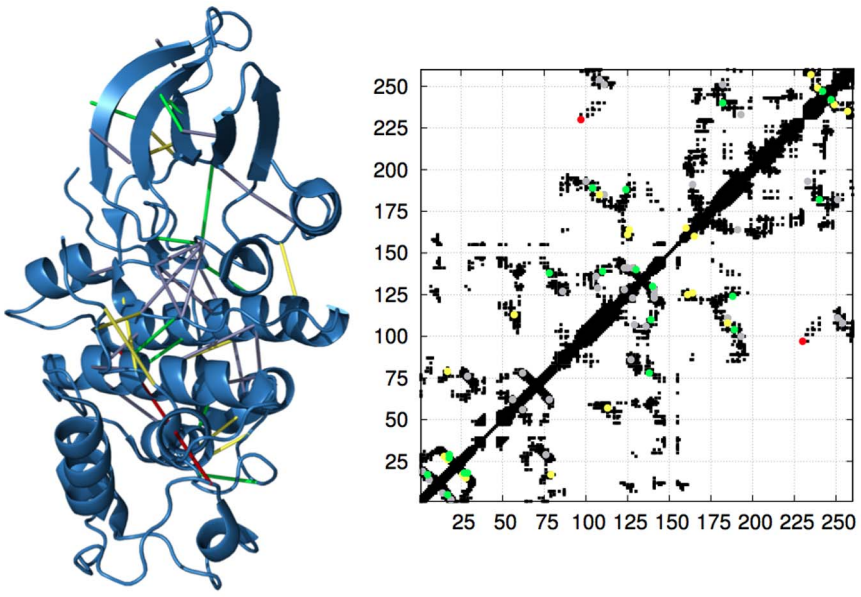

Figure 3. First 40 predicted contacts for the PF00069 family (Protein Kinase domain) with Gaussian DCA, using the same settings as for Fig. 2. The left panel shows the predicted contacts overlaid on the PDB structure *3fz1* (figure produced using the PyMOL software [51]); the right panel shows the predicted pairs overlaid on the contact map (true contacts as obtained by setting the threshold at 8 Å are shown in black). In both panels, the color code is the following: the first 10 predicted contacts are depicted in green, the next 10 contacts in yellow, the last 20 contacts in grey; the only false positive contact (occurring as the 24th predicted pair) is shown in red.
doi:10.1371/journal.pone.0092721.g003

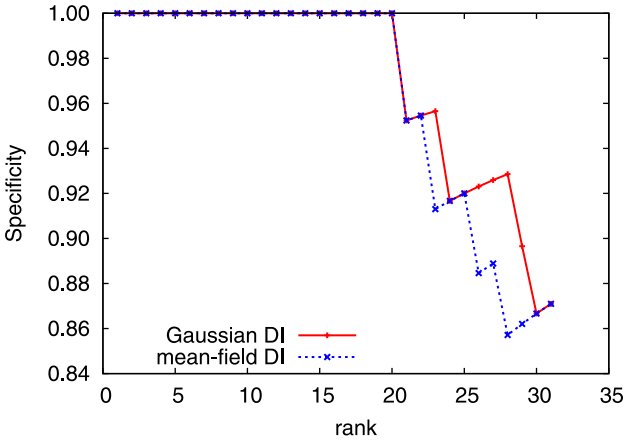

Figure 4. DI-ranking-induced mean true positive rate for predicting *inter-protein* contacts in the SK/RR complex, for both mean-field DCA (blue curve) and multivariate Gaussian DCA (red curve).
doi:10.1371/journal.pone.0092721.g004

and therefore the total alphabet size is $Q+1=21$. For simplicity, we denote amino-acids by numbers $1,\ldots,20$, and the gap by 21.

Here we consider a modified representation, similar to that used in [12], which turns out to be more practical for the multivariate modeling we are going to propose (cf. Fig. 7). The MSA is transformed into a $M\times(Q\cdot L)$-dimensional array $X=\left(x_i^m\right)_{i=1,\ldots,QL}^{m=1,\ldots,M}$ over a binary alphabet $\{0,1\}$. More precisely, each residue position in the original alignment is mapped to Q binary variables, each one associated with one standard amino-acid, taking value one if the amino-acid is present in the alignment, and zero if it is absent; the gap is represented by Q zeros (i.e. no amino-acid is present). Consequently, at most one of the Q variables can be one for a given residue position. For each sequence, the new variables are collected in one row vector, i.e. $x^m_{(l-1)Q+a}=\delta_{a,a_l^m}$ for $l=1,\ldots,L$ and $a=1,\ldots,Q$. The Kronecker symbol $\delta_{a,b}$ equals one for $a=b$, and zero otherwise.

Denoting the row length of X as $N=QL$, we introduce its empirical mean $\overline{x}=(\overline{x}_i)_{i=1,\ldots,N}$ and the empirical covariance matrix $C(X,\mu)=\left(C(X,\mu)_{ij}\right)_{i,j=1,\ldots,N}$ for given mean $\mu=(\mu_i)_{i=1,\ldots,N}$:

$$\overline{x}_i=\frac{1}{M}\sum_{m=1}^{M}x_i^m\,, \tag{1}$$

$$C_{ij}(X,\mu)=\frac{1}{M}\sum_{m=1}^{M}\left(x_i^m-\mu_i\right)\left(x_j^m-\mu_j\right). \tag{2}$$

The empirical covariance is thus $\overline{C}=C(X,\overline{x})$. Note that the entry $\overline{x}_i$, with $i=(l-1)Q+a$, measures the fraction of proteins having amino-acid $a\in\{1,\ldots,Q\}$ at position $l\in\{1,\ldots,L\}$. Similarly, the entry $C_{ij}(X,0)$ of the correlation matrix, with $i=(k-1)Q+a$ and $j=(l-1)Q+b$, is the fraction of proteins which show simultaneously amino-acid a in position k and b in position l.

The Gaussian model. We develop our multivariate Gaussian approach by approximating the binary variables as real-valued variables. Even though the former are highly structured, due to the fact that at most one amino-acid is present in each position of each sequence, we will not enforce these constraints on the model. Instead, we shall rely on the fact that the constraint is present by construction in the input data, and that as a consequence we have, for any residue position l and any two states a and b with $a\neq b$:

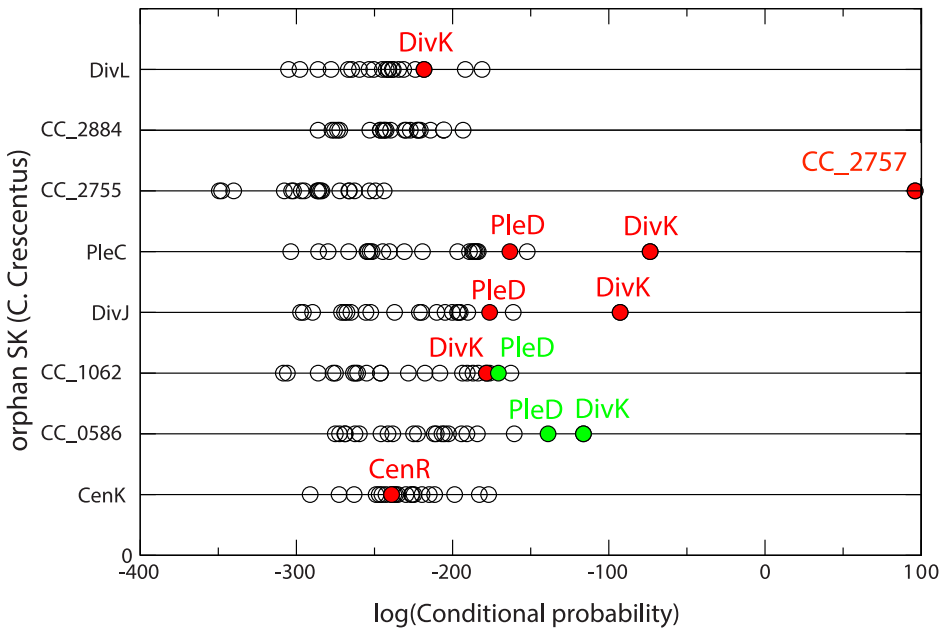

Figure 5. Partner prediction for *Caulobacter crescentus* orphan two-component proteins by the conditional probability method. Experimentally known interaction partners [44,45] are shown in red. Green dots correspond to partner predictions suggested in [18]. As for [18], the overall performance of the algorithm is good, except for the prediction on CenK-CenR interaction.
doi:10.1371/journal.pone.0092721.g005

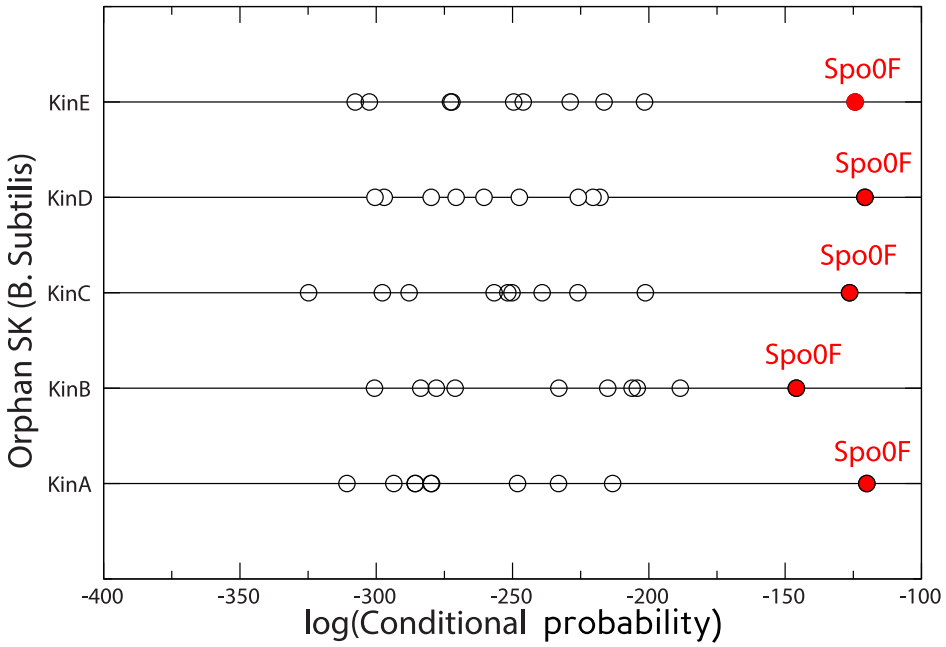

Figure 6. Partner prediction for *Bacillus subtilis* orphan two-component proteins. All 5 orphan kinases, KinA-E, are known to phosphorylate Spo0F, which is displayed in red and is always the maximally scoring protein in the RR set.
doi:10.1371/journal.pone.0092721.g006

$$C_{(l-1)Q+a,(l-1)Q+b} = -\overline{x}_{(l-1)Q+a}\overline{x}_{(l-1)Q+b} \leq 0 \qquad (3)$$

i.e. two different amino-acids at the same site are anti-correlated. Therefore, we shall let the parameter inference machinery work out suitable couplings between different amino-acid values at the same site, which generate these observed anti-correlations.

The multivariate Gaussian model and the Bayesian inference of its parameters are well-studied subjects in statistics, thus here we only briefly review the main ideas behind our approach, referring to [50] for details. The multivariate Gaussian distribution is parametrized by a mean vector $\mu=(\mu_i)_{i=1,\ldots,N}$ and a covariance matrix $\Sigma=(\Sigma_{ij})_{i,j=1,\ldots,N}$. Its probability density is

$$P(x|\mu,\Sigma)=(2\pi)^{-\frac{N}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right], \quad (4)$$

$|\Sigma|$ being the determinant of Σ, and it turns out that the $Q\times Q$ block

$$e_{kl}(a,b)=-\left(\Sigma^{-1}\right)_{(k-1)Q+a,(l-1)Q+b} \qquad (5)$$

(with $k,l\in\{1,\ldots,L\}$ and $a,b\in\{1,\ldots,Q\}$) plays the role of the direct interaction term in DCA between residues k and l. Assuming for the moment statistical independence of the M different protein sequences in the MSA, the probability of the data X under the model (i.e. the likelihood) reads

$$\begin{aligned}P(X|\mu,\Sigma)&=\prod_{m=1}^{M}P(x^m|\mu,\Sigma)\\&=(2\pi)^{-\frac{NM}{2}}|\Sigma|^{-\frac{M}{2}}\exp\left[-\frac{M}{2}\mathrm{tr}\left(\Sigma^{-1}C(X,\mu)\right)\right],\end{aligned} \qquad (6)$$

with $C(X,\mu)$ given by Eq. 2.

When the empirical covariance $\overline{C}$ is full rank, the likelihood attains its maximum at $\mu=\overline{x}$ and $\Sigma=\overline{C}$, which constitute the parameter estimates within the maximum likelihood approach. However, due to the under-sampling of the sequence space, $\overline{C}$ is typically rank deficient and this inference method is unfeasible. To estimate proper parameters, we make use of a Bayesian inference method, which needs the introduction of a prior distribution over μ and Σ. The required estimate is then computed as the mean of the resulting posterior, which is the parameter distribution conditioned to the data. As we have already mentioned, a convenient prior is the conjugate prior, which gives a posterior with the same structure as the prior but identified by different parameters accounting for the data contribution. The conjugate prior of the multivariate Gaussian distribution is the normal-inverse-Wishart (NIW) distribution. A NIW prior has the form $p(\mu,\Sigma)=p(\mu|\Sigma)p(\Sigma)$, where

$$p(\mu|\Sigma)=(2\pi)^{-\frac{N}{2}}\kappa^{\frac{N}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{\kappa}{2}(\mu-\eta)^T\Sigma^{-1}(\mu-\eta)\right] \quad (7)$$

is a multivariate Gaussian distribution on μ with covariance matrix Σ/κ and prior mean $\eta=(\eta_i)_{i=1,\ldots,N}$. The parameter κ has the meaning of number of prior measurements. The prior on Σ is the inverse-Wishart distribution

$$p(\Sigma)=\frac{1}{Z}|\Sigma|^{-\frac{\nu+N+1}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Lambda\Sigma^{-1}\right)\right], \qquad (8)$$

where Z is a normalizing constant:

$$Z=2^{\frac{\nu N}{2}}\pi^{\frac{N(N-1)}{4}}|\Lambda|^{-\frac{\nu}{2}}\prod_{n=1}^{N}\Gamma\left(\frac{\nu+1-n}{2}\right), \qquad (9)$$

Γ being Euler's Gamma function. The parameters ν and $\Lambda=(\Lambda_{ij})_{i,j=1,\ldots,N}$ are the degree of freedom and the scale matrix,

Table 2. 50 Pfam families used in the benchmarks, together with their associated PDB entries.

Pfam ID	Description	PDB
PF00001	7 transmembrane receptor (rhodopsin family)	1f88, 2rh1
PF00004	ATPase family associated with various cellular activities (AAA)	2p65, 1d2n
PF00006	ATP synthase alpha/beta family, nucleotide-binding domain	2r9v
PF00009	Elongation factor Tu GTP binding domain	1skq, 1xb2
PF00011	Hsp20/alpha crystallin family	2bol
PF00012	Hsp70 protein	2qxl
PF00013	KH domain	1wvn
PF00014	Kunitz/Bovine pancreatic trypsin inhibitor domain	5pti
PF00016	Ribulose bisphosphate carboxylase large chain, catalytic domain	1svd
PF00017	SH2 domain	1o47
PF00018	SH3 domain	2hda, 1shg
PF00025	ADP-ribosylation factor family	1fzq
PF00026	Eukaryotic aspartyl protease	3er5
PF00027	Cyclic nucleotide-binding domain	3fhi
PF00028	Cadherin domain	2o72
PF00032	Cytochrome b(C-terminal)/b6/petD	1zrt
PF00035	Double-stranded RNA binding motif	1o0w
PF00041	Fibronectin type III domain	1bqu
PF00042	Globin	1cp0
PF00043	Glutathione S-transferase, C-terminal domain	6gsu
PF00044	Glyceraldehyde 3-phosphate dehydrogenase, NAD binding domain	1crw
PF00046	Homeobox domain	2vi6
PF00056	Lactate/malate dehydrogenase, NAD binding domain	1a5z
PF00059	Lectin C-type domain	1lit
PF00064	Neuraminidase	1a4g
PF00069	Protein kinase domain	3fz1
PF00071	Ras family	5p21
PF00072	Response regulator receiver domain	1nxw
PF00073	Picornavirus capsid protein	2r06
PF00075	RNase H	1f21
PF00077	Retroviral aspartyl protease	1a94
PF00078	Reverse transcriptase (RNA-dependent DNA polymerase)	1dlo
PF00079	Serpin (serine protease inhibitor)	1lj5
PF00081	Iron/manganese superoxide dismutases, alpha-hairpin domain	3bfr
PF00082	Subtilase family	1p7v
PF00084	Sushi domain (SCR repeat)	1elv
PF00085	Thioredoxin	3gnj
PF00089	Trypsin	3tgi
PF00091	Tubulin/FtsZ family, GTPase domain	2r75
PF00092	Von Willebrand factor type A domain	1atz
PF00102	Protein-tyrosine phosphatase	1pty
PF00104	Ligand-binding domain of nuclear hormone receptor	1a28
PF00105	Zinc finger, C4 type (two domains)	1gdc
PF00106	Short chain dehydrogenase	1a27
PF00107	Zinc-binding dehydrogenase	1a71
PF00108	Thiolase, N-terminal domain	3goa
PF00109	Beta-ketoacyl synthase, N-terminal domain	1ox0
PF00111	2Fe-2S iron-sulfur cluster binding domain	1a70
PF00112	Papain family cysteine protease	1o0e
PF00113	Enolase, C-terminal TIM barrel domain	2al2

doi:10.1371/journal.pone.0092721.t002

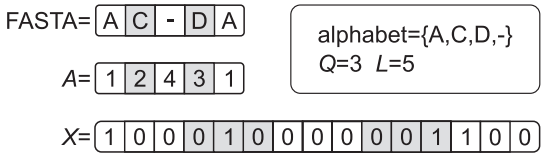

Figure 7. Illustration of the encoding of a sequence from FASTA format to its intermediate numeric representation (matrix A) to its final binarized representation (matrix X). For clarity, we restrict the alphabet to $Q=3$ amino-acids, $\{A,C,D\}$, plus the gap. The alternation of white and gray cell backgrounds helps to track the transformation (e.g. $C\rightarrow 2\rightarrow 010$). Typically, MSAs of protein families are such that in every column (i.e. residue position) there appears a number of distinct residues smaller than or equal to $Q=20$. Here, we did not not consider a restriction of the alphabet to the residues actually occurring, and we used instead the same encoding for all residues.
doi:10.1371/journal.pone.0092721.g007

respectively, shaping the inverse-Wishart distribution. The condition for this distribution to be integrable is $v>N-1$. The posterior $p(\mu,\Sigma|X)$, proportional to $P(X|\mu,\Sigma)\cdot p(\mu,\Sigma)$, is still a NIW distribution, as one can easily verify starting from Eqs. 6, 7 and 8. The posterior distribution $p(\mu,\Sigma|X)$ is characterized by parameters κ', η', v', and Λ' given by the formulae

$$\begin{cases} \kappa' = \kappa + M, \\ \eta' = \dfrac{\kappa}{\kappa+M}\eta + \dfrac{M}{\kappa+M}\overline{x}, \\ v' = v + M, \\ \Lambda' = \Lambda + M\overline{C} + \dfrac{\kappa M}{\kappa+M}(\overline{x}-\eta)(\overline{x}-\eta)^T. \end{cases} \tag{10}$$

The mean values of μ and Σ under the NIW prior are η and $\Lambda/(v-N-1)$, and, similarly, their expected values under the NIW posterior are η' and $\Lambda'/(v'-N-1)$, respectively. Our estimates of the mean vector and the covariance matrix, that with a slight abuse of notation we shall still denote by μ and Σ for the sake of simplicity, are thus

$$\mu = \eta' = \frac{\kappa}{\kappa+M}\eta + \frac{M}{\kappa+M}\overline{x} \tag{11}$$

and

$$\Sigma = \frac{\Lambda'}{v'-N-1} = \frac{\Lambda + M\overline{C} + \frac{kM}{k+M}(\overline{x}-\eta)^T(\overline{x}-\eta)}{v+M-N-1}. \tag{12}$$

The NIW posterior is maximum at $\mu=\eta'$ and $\Sigma=\Lambda'/(v'+N+1)$, with the consequence that the *maximum a posteriori* estimate would provide the same estimate of μ and an estimate of Σ that only differs from the previous one by a scale factor.

As a first attempt of protein contact prediction by means of the present model, we choose η and Λ to be as uninformative as possible. In particular, since $U=\Lambda/(v-N-1)$ is the prior estimate of Σ, it is natural to set $\eta=(\eta_i)_{i=1,\ldots,N}$ and $U=(U_{ij})_{i,j=1,\ldots,N}$ to the mean and the covariance matrix of uniformly distributed samples. Therefore, we set $\eta_i=1/(Q+1)$ for any i, and U to a block-matrix composed of $L\times L$ blocks of size $Q\times Q$ each, where the out-of-diagonal blocks are uniformly 0:

$$U_{(k-1)Q+a,(l-1)Q+b} = \frac{\delta(k,l)}{Q+1}\left(\delta(a,b) - \frac{1}{Q+1}\right), \tag{13}$$

where $k,l\in\{1,\ldots,L\}$ and $a,b\in\{1,\ldots,Q\}$, and δ is the Kronecker's symbol. Moreover, we choose $v=N+\kappa+1$ in order to reconcile Eq. 12 with the pseudo-count-corrected covariance matrix of [10] with pseudo-count parameter λ. Indeed, identifying λ with $\kappa/(\kappa+M)$, this instance allows us to recast the estimation of Σ as

$$\Sigma = \lambda U + (1-\lambda)\overline{C} + \lambda(1-\lambda)(\overline{x}-\eta)^T(\overline{x}-\eta) \tag{14}$$

and $J=\Sigma^{-1}$ becomes the same as in the mean-field Potts model. Manifestly from here, the effect of the prior is enhanced by values of λ close to 1 while it is negligible when λ approaches 0. Interestingly, the Gaussian framework provides an interpretation of the pseudo-count correction in terms of a prior distribution, which may allow improving the inference issue by exploiting more informative prior choices.

Reweighted frequency counts. The approach outlined in the above sections assumes that the rows of the MSA matrix X, i.e. the different protein sequences, form an independently and identically distributed (i.i.d.) sample, drawn from the model distribution, cf. Eq. 6. For biological sequence data this is not true: there are strong sampling biases due to phylogenetic relations between species, due to the sequencing of different strains of the same species, and due to a non-random selection of sequenced species. The sampling is therefore clustered in sequence space, thereby introducing spurious non-functional correlations, whereas other viable parts of sequence space (in the sense of sequences which would fall into the same protein family) are statistically underrepresented. To partially remove this sampling bias, we use the same re-weighting scheme used in the PSICOV version 1.11 code [12] (which is the same as that used in [8,10], with an additional pre-processing pass to estimate a value for the similarity threshold; see File S1 for details). The procedure can be seen as generalization of the elimination of repeated sequences.

Computing the ranking score. Contact prediction using DCA relies on ranking pairs of residue positions $1\leq k<l\leq L$ according to their direct interaction strength. As mentioned before, two positions interact via a $Q\times Q$ matrix e_{kl} given by Eq. 5. To compare two position pairs kl and $k'l'$, we need to map these matrices to a single scalar quantity. We have tested two different transformations: the first one, following [8], is the so-called direct information (DI), which measures the mutual information induced only by the direct coupling e_{kl} between two positions k and l (for a more precise definition see File S1); the second one, following [15], is the Frobenius norm (FN) of the submatrix obtained by (*i*) changing the gauge of the interaction such that the sum of each row and column is zero, and (*ii*) removing the row and column corresponding to the gap symbol. In our empirical tests (cf. Fig. 1), the FN score can reach a better overall accuracy in residues contacts prediction; the DI score, however, also achieves good results, is gauge-invariant, and has a clear interpretation in terms of the underlying model: it is therefore a useful indicator to compare the Gaussian model with the mean-field approximation to the discrete model. In the multivariate Gaussian setting, the DI can be calculated explicitly, as shown in File S1, thus resulting in a gain in computation time as compared to the mean-field DCA in [10], while achieving similar or better performance (cf. Fig. 1).

We found empirically that both the DI and the FN scores produce slightly better results in the residue contact prediction tests when adjusted via average-product-correction (APC), as described in [41].

Summary of the residue contact prediction steps

To summarize the previous sections, here we list the steps which are taken in order to get from a MSA to the contact prediction:

- clean the MSA by removing inserts and keeping only matched amino acids and deletions;
- remove the sequences for which 90% or more of the entries are gaps;
- assign a weight to each sequence, and compute the reweighted frequency counts $\overline{C}$ and $\overline{x}$ (see Eqs. 1 and 2, and Suporting File S1);
- estimate the correlation matrix Σ by means of Eq. 14;
- compute Σ^{-1}, and divide it in $Q \times Q$ blocks e_{kl} (see Eq. 5);
- for each pair $1 \leq k,l \leq L$, compute a score (DI or FN) from e_{kl}, thus obtaining an $L \times L$ symmetric matrix S (with zero diagonal);
- apply APC to the score matrix (i.e. subtract to each entry S_{kl} the product of the average score over k and the average score over l, divided by the overall score average – the averages are computed excluding the diagonal), and obtain an adjusted score matrix S_{kl}^{APC};
- rank all pairs $1 \leq k < l \leq L$, with $l-k>4$, in descending order according to S_{kl}^{APC}.

A log-likelihood score for protein-protein interaction

In [18], DCA has been used to predict RR interaction partners for orphan SK proteins in bacterial TCS, and to detect crosstalk between different cognate SK/RR pairs. Relying on the improved efficiency of the multivariate Gaussian approach presented here, we can introduce a much clearer but similarly performing definition of a protein-protein interaction score.

This score is based on the existence of a large set of known interaction partners: we collect them in a unified MSA, in which each row contains the concatenation of two interacting protein sequences, and we encode them in a matrix denoted by X_{SKRR}. The encoded MSAs restricted to each of the single protein families are denoted by X_{SK} and X_{RR}. We estimate model parameters Σ_A and μ_A for each of the three alignments X_A, with $A \in \{\mathrm{SK}, \mathrm{RR}, \mathrm{SKRR}\}$. Whereas the parameters for the two alignments of single protein families describe the *intra*-domain co-evolution inside each domain, the parameter matrix Σ_{SKRR}, obtained from the joint MSA, also models the *inter*-protein co-evolution.

In order to decide if two new sequences x_{SK} and x_{RR} interact, we first introduce the sequence x_{SKRR} as the (horizontal) concatenation of x_{SK} with x_{RR}. Next we define a log-odds ratio comparing the probability of these sequences under the joint SKRR-model with the one under the separate models for SK and RR, i.e. we calculate

$$\begin{aligned}\mathcal{L}(x_{\mathrm{SK}},x_{\mathrm{RR}}) &= \log \frac{P(x_{\mathrm{SKRR}}|\Sigma_{\mathrm{SKRR}},\mu_{\mathrm{SKRR}})}{P(x_{\mathrm{SK}}|\Sigma_{\mathrm{SK}},\mu_{\mathrm{SK}})P(x_{\mathrm{RR}}|\Sigma_{\mathrm{RR}},\mu_{\mathrm{RR}})} \\ &= c - \frac{1}{2}(x_{\mathrm{SKRR}}-\mu_{\mathrm{SKRR}})^t \Sigma_{\mathrm{SKRR}}^{-1}(x_{\mathrm{SKRR}}-\mu_{\mathrm{SKRR}}) \\ &+ \frac{1}{2}(x_{\mathrm{SK}}-\mu_{\mathrm{SK}})^t \Sigma_{\mathrm{SK}}^{-1}(x_{\mathrm{SK}}-\mu_{\mathrm{SK}}) \\ &+ \frac{1}{2}(x_{\mathrm{RR}}-\mu_{\mathrm{RR}})^t \Sigma_{\mathrm{RR}}^{-1}(x_{\mathrm{RR}}-\mu_{\mathrm{RR}})\end{aligned} \tag{15}$$

with c being a constant (i.e. not depending on the sequence $x_{\mathrm{SK}},x_{\mathrm{RR}}$) coming from the normalization of the multivariate Gaussians. Intuitively, this score measures to what extent the two sequences are coherent with the model of interacting SK/RR sequences, as compared to a model which assumes them to be just two arbitrary (and thus typically not interacting) SK and RR sequences. In mathematical terms, it can also be seen as the log-odds ratio between the conditional probability of x_{SK} knowing x_{RR}, and the unconditioned probability of x_{SK}.

Supporting Information

File S1
(PDF)

Author Contributions

Conceived and designed the experiments: CB MZ MW A. Pagnani. Performed the experiments: CB MZ CF A. Procaccini A. Pagnani. Analyzed the data: CB CF A. Procaccini A. Pagnani. Contributed reagents/materials/analysis tools: CB MZ CF A. Procaccini MW A. Pagnani. Wrote the paper: CB MZ CF A. Procaccini RZ MW A. Pagnani. Code Implementation: CB MW A. Pagnani.

References

1. Altschuh D, Lesk A, Bloomer A, Klug A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. Journal of Molecular Biology 193: 693–707.
2. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins: Structure, Function and Genetics 18: 309–317.
3. Neher E (1994) How frequent are correlated changes in families of protein sequences? Proceedings of the National Academy of Sciences 91: 98–102.
4. Shindyalov I, Kolchanov N, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Engineering 7: 349–358.
5. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295–299.
6. Fodor AA, Aldrich RW (2004) Inuence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins: Structure, Function, and Bioinformatics 56: 211–221.
7. de Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. Nature Reviews Genetics.
8. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proceedings of the National Academy of Sciences 106: 67–72.
9. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6: e1000633.
10. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108: E1293–E1301.
11. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. Proteins: Struct, Funct, Bioinf 79: 1061.
12. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28: 184.
13. Sreekumar J, ter Braak C, van Ham R, van Dijk A (2011) Correlated mutations via regularized multinomial regression. BMC Bioinformatics 12: 444.
14. Cocco S, Monasson R, Weigt M (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.
15. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. Physical Review E 87: 012707.

16. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residueresidue contact predictions in a sequence- and structure-rich era. Proceedings of the National Academy of Sciences 110: 15674–15679.
17. Burger L, Van Nimwegen E (2008) Accurate prediction of protein–protein interactions from sequence alignments using a bayesian method. Molecular Systems Biology 4: 165.
18. Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M (2011) Dissecting the Specificity of Protein-Protein Interaction in Bacterial Two-Component Signaling: Orphans and Crosstalks. PLoS ONE 6: e19729+.
19. Jaynes ET (1957) Information Theory and Statistical Mechanics. Physical Review Series II 106: 620630.
20. Jaynes ET (1957) Information Theory and Statistical Mechanics II. Physical Review Series II 108: 171190.
21. Lapedes AS, Giraud BG, Liu L, Stormo GD (1999) Correlated mutations in models of protein sequences: Phylogenetic and structural effects. Lecture Notes-Monograph Series: Statistics in Molecular Biology and Genetics 33: pp. 236–256.
22. Lapedes A, Giraud B, Jarzynski C (2012) Using sequence alignments to predict protein structure and stability with high accuracy. arXiv preprint arXiv:12072484.
23. Mora T, Walczak AM, Bialek W, Callan CG (2010) Maximum entropy models for antibody diversity. Proceedings of the National Academy of Sciences 107: 5405–5410.
24. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. Proc Natl Acad Sci USA 106: 22124.
25. Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, et al. (2012) Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. Proceedings of the National Academy of Sciences.
26. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3d structure computed from evolutionary sequence variation. PLoS ONE 6: e28766.
27. Sadowski MI, Maksimiak K, Taylor WR (2011) Direct correlation analysis improves fold recognition. Computational Biology and Chemistry 35: 323–332.
28. Nugent T, Jones DT (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proceedings of the National Academy of Sciences 109: E1540–E1547.
29. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. Proc Natl Acad Sci 109: 10340–10345.
30. Taylor WR, Jones DT, Sadowski MI (2012) Protein topology from predicted residue contacts. Protein Science 21: 299–305.
31. Hopf T, Colwell L, Sheridan R, Rost B, Sander C, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell.
32. Wang C, Sang J, Wang J, Su M, Downey JS, et al. (2013) Mechanistic insights revealed by the crystal structure of a histidine kinase with signal transducer and sensor domains. PLoS Biol 11: e1001493.
33. Diensthuber R, Bommer M, Gleichmann T, Mglich A (2013) Full-length structure of a sensor histidine kinase pinpoints coaxial coiled coils as signal transducers and modulators. Structure 21: 1127–1136.
34. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annual Review of Biochemistry 69: 183–215.
35. Hoch JA, Varughese K (2001) Keeping signals straight in phosphorelay signal transduction. J Bacteriol 183: 4941–4949.
36. Laub MT, Goulian M (2007) Specificity in two-component signal transduction pathways. Annu Rev Genet 41: 121–145.
37. Szurmant H, Hoch JA (2010) Interaction fidelity in two-component signaling. Curr Opin Microbiol 13: 190–197.
38. MATLAB website (nd) Available: http://www.mathworks.it/products/matlab/. Accessed 2014 Feb 27.
39. Julia website (nd) Available: http://julialang.org/. Accessed 2014 Feb 27.
40. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate JG, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290.
41. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the inuence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333–340.
42. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV (2004) To be folded or to be unfolded? Protein Science 13: 2871–2877.
43. Jiang M, Shao W, Perego M, Hoch J (2000) Multiple histidine kinases regulate entry into stationary phase and sporulation in bacillus subtilis. Mol Microbiol 38: 535–542.
44. Ohta N, Newton A (2003) The core dimerization domains of histidine kinases contain recognition specificity for the cognate response regulator. Journal of Bacteriology 185: 4424–4431.
45. Skerker JM, Prasol MS, Perchuk BS, Biondi EG, Laub MT (2005) Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: A system-level analysis. PLoS Biol 3: e334.
46. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The pfam protein families database. Nucleic Acids Research 36: D281–D288.
47. Eddy SR (1998) Profile hidden markov models. Bioinformatics 14: 755–763.
48. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The protein data bank. Nucleic acids research 28: 235–242.
49. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The pfam protein families database. Nucleic Acids Research 36: D281–D288.
50. Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian Data Analysis. Chapman and Hall/CRC.
51. PyMOL website (nd) Available: http://pymol.org/. Accessed 2014 Feb 27.