Journal of Computational Physics 276 (2014) 341–356

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences

Magnus Ekeberg [a,b,*,1], Tuomo Hartonen [c,d,1], Erik Aurell [b,c,e]

[a] Engineering Physics Program, KTH Royal Institute of Technology, SE-100 77 Stockholm, Sweden
[b] Department of Computational Biology, AlbaNova University Centre, 106 91 Stockholm, Sweden
[c] Department of Information and Computer Science, Aalto University, PO Box 15400, FI-00076 Aalto, Finland
[d] The Master's Degree Programme in Translational Medicine, Biomedicum Helsinki, FI-00014 University of Helsinki, Finland
[e] Aalto Science Institute, PO Box 15600, FI-00076 Aalto, Finland

ARTICLE INFO

Article history:
Received 24 January 2014
Received in revised form 23 June 2014
Accepted 15 July 2014
Available online 25 July 2014

Keywords:
Protein structure prediction
Contact map
Direct-coupling analysis
Potts model
Pseudolikelihood
Inference

ABSTRACT

Direct-coupling analysis is a group of methods to harvest information about coevolving residues in a protein family by learning a generative model in an exponential family from data. In protein families of realistic size, this learning can only be done approximately, and there is a trade-off between inference precision and computational speed. We here show that an earlier introduced l_2-regularized pseudolikelihood maximization method called plmDCA can be modified as to be easily parallelizable, as well as inherently faster on a single processor, at negligible difference in accuracy. We test the new incarnation of the method on 143 protein family/structure-pairs from the Protein Families database (PFAM), one of the larger tests of this class of algorithms to date.

1. Introduction

A momentous challenge for research, companies, and society at large is how to use better and in novel ways vast swathes of accrued information, often referred to as "Big Data". Such data can be collected and catalogued in many different ways, and then analyzed by different actors, potentially in new fashion to pursue very different objectives than for which the data was originally gathered. In this paper, we report on progress on one important example where data on homologous proteins,[2] collected by many research groups around the world, can be decoded to reveal amino-acid contacts within protein structures to very good accuracy. An existing pseudolikelihood maximization approach currently delivers higher accuracy

Abbreviations: PSP, Protein Structure Prediction; CASP, Critical Assessment of protein Structure Prediction; DCA, Direct-Coupling Analysis; PFAM, Protein Families database; SCOP, Structural Classification of Proteins; mfDCA, mean-field Direct-Coupling Analysis; plmDCA, pseudolikelihood maximization Direct-Coupling Analysis; MSA, Multiple Sequence Alignment; FN, Frobenius Norm; APC, Average Product Correction; CFN, Corrected Frobenius Norm; HMM, Hidden Markov-Model; PDB, Protein Data Bank; UNIPROT, Universal Protein Resource; NMR, Nuclear Magnetic Resonance; SIFTS, Structure Integration with Function, Taxonomy and Sequence; TPR, True-Positive Rate.

* Corresponding author at: Department of Computational Biology, AlbaNova University Centre, 106 91 Stockholm, Sweden.
E-mail address: ekeb@kth.se (M. Ekeberg).

[1] Joint first authors.

[2] In this paper, we use "protein" interchangeably with "protein domain".

http://dx.doi.org/10.1016/j.jcp.2014.07.024
0021-9991/

than other methods, but at the cost of longer running time. We here introduce a new version of this earlier method, and show that it yields predictions with practically identical precision, but with a large computational speed-up.

Protein Structure Prediction (PSP) aims to reap information about the three-dimensional structure of a protein from any suitable data, but in particular from its amino-acid sequence. Advances are regularly evaluated in the framework of CASP (The Critical Assessment of protein Structure Prediction) [1]. Although much progress has been made, the consensus opinion has become that *ab initio* PSP, i.e. predicting the three-dimensional structure of a protein from its amino-acid sequence only, is not feasible. On the other hand, homology PSP, i.e. predictions taking cues from known structures of proteins that are homologous, is often possible, although in many respects remaining an art.

Direct-Coupling Analysis (DCA) belongs to an intermediate level of PSP where predictions are made not from a single amino-acid sequence, but from the set of amino-acid sequences of a family of homologous proteins. The interest of this approach is at least twofold. First, the number of known protein structures grows at a much slower rate than the number of known amino-acid sequences — their ratio today being about $1:550$ — and this can be expected to remain the case for the foreseeable future. Therefore, while today if a protein is a member of a family containing many homologues then very often at least one of the homologues has a known structure, this may be less and less likely to be true in the future. Second, it is of interest to know if the information contained not just in one amino-acid sequence, but in a whole family of sequences — usually evolutionary related and hence subject to the same evolutionary constraints — is sufficient to determine the three-dimensional structure. In fact, it has been known for almost 20 years that the evolutionary history leaves a trace in the correlations between amino acids at different positions along a protein which contains nontrivial information, see e.g. [2–4], but before DCA this information was not fully exploitable. PSP by DCA is thus, apart from its intrinsic scientific interest, also a showcase for Big Data and how it can be exploited to arrive at new useful knowledge checkpoints. For a broader review of coevolution analysis for elucidating protein structures, see e.g. [5].

This paper is organized as follows: in Section 2 we introduce DCA and review and summarize the main approaches used up to now. In Section 3 we then present the pseudolikelihood maximization approach in more detail, first the previous version presented in [6], and then the faster parallel version introduced here. In Section 4 we present the data (and extraction thereof) on which our analysis is based, and in Section 5 we compare the speed and accuracy of the two versions of the pseudolikelihood maximization, followed by extensive experiments on the new version. Finally, in Section 6 we discuss our results. Supplementary Information to this paper gives additional data on proteins used and a family-per-family view of performance.

2. A primer on direct-coupling analysis

Let us represent the amino-acid sequence of a protein as $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_N)$. We assume that we have a Multiple Sequence Alignment (MSA), which is a table $\{\boldsymbol{\sigma}^{(b)}\}_{b=1}^{B}$ of such amino-acid sequences of B proteins that have been aligned to have a common length N. In this work we will limit ourselves to using MSAs obtained from the PFAM database [7,8]. We will discuss how such tables look in Section 4 below and here just observe that each row in the table will represent a protein, and each column a position in the sequence. At row b and position i we hence have a symbol $\sigma_i^{(b)}$ which can be one of the 20 naturally occurring amino acids or a "–", representing a gap in the alignment. For a list of amino acids and the symbols and abbreviations representing them, see Table S1.

The essence of DCA is then to assume that the rows, i.e. our aligned homologous proteins, are independent events drawn from a Potts-model probability distribution,

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \frac{1}{2}\sum_{i,j=1}^{N} J_{ij}(\sigma_i,\sigma_j)\right), \tag{1}$$

and to use the interaction parameters $\mathbf{J}_{ij}$ as predictions of spatial proximity among amino-acid pairs in the protein structure. Interpreting the $\mathbf{J}_{ij}$ this way can be biologically justified as follows: it is well-known that the detrimental effects of a single-site mutation, that alone would impair the function of the protein, can be countered by a compensatory mutation at a nearby site. Consequently, short intra-domain position–position distances can, and do, show up as pairwise couplings among the columns in the table $\{\boldsymbol{\sigma}^{(b)}\}_{b=1}^{B}$.

To avoid trivial overparameterization we will define $\mathbf{J}_{ij}(k,l) = \mathbf{J}_{ji}(l,k)$ if i and j are different and $\mathbf{J}_{ij} = \mathbf{0}$ if $i = j$. The double sum in (1) hence goes over all unordered pairs of distinct positions along the columns in the table, i.e.

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{1\le i<j\le N} J_{ij}(\sigma_i,\sigma_j)\right). \tag{2}$$

Throughout the paper, we will, unless otherwise specified, assume single position-indexes to run across $1 \le i \le N$, pairwise position-indexes to run as $1 \le i < j \le N$, and amino-acid indexes to span $1 \le k \le q$, where $q = 21$ (20 amino acids and one additional state for the alignment gap). Determining the $\mathbf{J}_{ij}$ from the observations $\{\boldsymbol{\sigma}^{(b)}\}_{b=1}^{B}$ is a nontrivial inference problem, since for N large enough the normalization constant Z, the number of terms of which (q^N) grows exponentially with the protein length, cannot be computed efficiently and exactly. Let us note that if we would have a multidimensional Gaussian