Bioinformatics, 36(7), 2020, 2278–2279
doi: 10.1093/bioinformatics/btz925
Advance Access Publication Date: 18 December 2019
Applications Note

Sequence analysis

MPF–BML: a standalone GUI-based package for maximum entropy model inference

Ahmed A. Quadeer [1], Matthew R. McKay[1,2], John P. Barton[3] and Raymond H. Y. Louie [4,5,*]

[1]Department of Electronic and Computer Engineering, [2]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China, [3]Department of Physics and Astronomy, University of California, Riverside, CA 92521, USA, [4]The Kirby Institute and [5]School of Medical Sciences, University of New South Wales, Sydney, NSW 2052, Australia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 30, 2019; revised on November 3, 2019; editorial decision on November 28, 2019; accepted on December 16, 2019

Abstract

Summary: Learning underlying correlation patterns in data is a central problem across scientific fields. Maximum entropy models present an important class of statistical approaches for addressing this problem. However, accurately and efficiently inferring model parameters are a major challenge, particularly for modern high-dimensional applications such as in biology, for which the number of parameters is enormous. Previously, we developed a statistical method, minimum probability flow–Boltzmann Machine Learning (MPF–BML), for performing fast and accurate inference of maximum entropy model parameters, which was applied to genetic sequence data to estimate the fitness landscape for the surface proteins of human immunodeficiency virus and hepatitis C virus. To facilitate seamless use of MPF–BML and encourage more widespread application to data in diverse fields, we present a standalone cross-platform package of MPF–BML which features an easy-to-use graphical user interface. The package only requires the input data (protein sequence data or data of multiple configurations of a complex system with large number of variables) and returns the maximum entropy model parameters.

Availability and implementation: The MPF–BML software is publicly available under the MIT License at https://github.com/ahmedaq/MPF-BML-GUI.

Contact: rlouie@kirby.unsw.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Interpreting patterns of correlation is a fundamental problem in data analysis. Among different methods for this purpose (e.g. de Juan *et al.*, 2013; Quadeer *et al.*, 2018; Rivoire *et al.*, 2016), maximum entropy models have emerged as a powerful tool and have been widely employed in diverse biological problems. Application of maximum entropy models include inferring fitness landscapes of viral proteins (Ferguson *et al.*, 2013; Louie *et al.*, 2018; Quadeer *et al.*, 2019), predicting protein contacts (Morcos *et al.*, 2011), and representing patterns of neural activity (Ganmor *et al.*, 2011). Inferring the 'maximum entropy parameters' is non-trivial, and the conventional method involving iterative Monte Carlo simulations can be very slow for even moderate system sizes. Faster inference methods have been developed (Ricci-Tersenghi, 2012), (Aurell and Ekeberg, 2012), but these accurately reproduce the original data only in certain limits.

To address this issue, advanced techniques, including adaptive cluster expansion (Barton *et al.*, 2016) and minimum probability flow (MPF) (Sohl-Dickstein *et al.*, 2011), have been developed which are not only fast, but also accurate for general scenarios. The MPF algorithm in particular uses Markov chain theory to reformulate the inference problem to one which is highly computationally efficient while retaining accuracy. The MPF algorithm (Sohl-Dickstein *et al.*, 2011) was extended in Louie *et al.* (2018) to include additional steps to further improve the accuracy using a Boltzmann Machine Learning (BML) approach, to avoid over-fitting, and to allow for non-binary variables. These extensions are required for many datasets, such as protein sequence data. This modified algorithm, referred to as 'MPF–BML', was used to obtain a fitness landscape for highly variable and large-dimensional surface proteins of human immunodeficiency virus (HIV) (Louie *et al.*, 2018) and HCV (Quadeer *et al.*, 2019). Here, to facilitate broadly accessible, seamless use of the MPF–BML algorithm (Louie *et al.*, 2018), we present

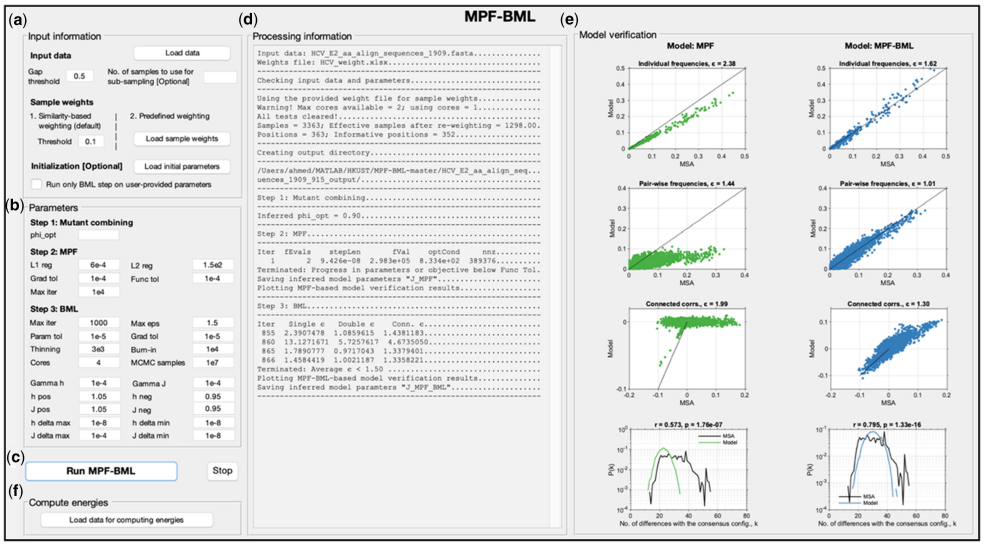

Fig. 1. The MPF–BML GUI. (**a**) Input information required to run MPF–BML; (**b**) parameters of MPF–BML; (**c**) main execution button to run and stop MPF–BML; (**d**) display showing information about the provided data and progress of MPF–BML execution; (**e**) plots to verify the inferred MPF and MPF–BML models and (**f**) button to calculate energies corresponding to new input samples using the inferred model parameters. Results are shown for HCV E2 protein (Quadeer *et al.*, 2019) as an example

a standalone graphical user interface (GUI)-based package of this method for fast and accurate inference of maximum entropy parameters.

2 Description of MPF–BML

MPF–BML is distributed as a standalone cross-platform application available for Mac, Linux and Microsoft Windows. MATLAB runtime libraries necessary to run the software are automatically downloaded during the software installation. We use data for HCV surface protein E2 as a test case (Fig. 1), reproducing results of Quadeer *et al.* (2019), but MPF–BML can be applied to data in diverse fields (see Supplementary Table S1 and Figs S1–S6 for results on other datasets).

Running the installed file opens the MPF–BML GUI (Fig. 1). The 'Input information' panel (Fig. 1a) allows the user to provide inputs to run the MPF–BML method. The mandatory input data is the multiple sequence alignment of the protein or categorical data (rows and columns represent samples and variables, respectively). Other inputs are optional and described in Supplementary Text S1.

The 'Parameters' panel (Fig. 1b) allows the user to set values of the parameters involved in the three steps of the MPF–BML method (see Louie *et al.*, 2018 and Supplementary Text S2 for details). A brief description of parameters is provided as tooltip to assist the user. After providing input information and setting the parameters, the user can run the method by clicking on the 'Run MPF–BML' button (Fig. 1c). A 'Stop' button is also provided to stop execution in case the algorithm is not converging. Progress of the execution process is displayed in the 'Processing information' panel (Fig. 1d).

The inferred models are validated (Fig. 1e) by comparing the individual (p_i) and pair-wise frequencies (p_{ij}) as well as connected correlations ($p_{ij} - p_i p_j$) of configurations in the data with those from the inferred model. The statistical error in each case is quantified using the measure ε (see Louie *et al.*, 2018 for details). Although not used for training the model, the probability $P(k)$ of observing a configuration with k differences with the consensus configuration obtained using the inferred model is also compared to further demonstrate the accuracy of the inferred model.

Finally, the inferred maximum entropy parameters using MPF and MPF–BML are saved in tab-delimited file (see Supplementary Table S2). Contacts predicted using the inferred couplings (Supplementary Text S3) are also provided in a tab-delimited file. Once a model has been inferred, panel 'Compute energies' (Fig. 1f) appears for the user to calculate energies corresponding to new samples using the inferred model parameters.

Funding

M.R.M., R.H.Y.L. and A.A.Q. were supported by the General Research Fund of the Hong Kong Research Grants Council (RGC) (grant numbers 16207915 and 16204519).

Conflict of Interest: none declared.

References

Aurell,E. and Ekeberg,M. (2012) Inverse Ising inference using all the data. *Phys. Rev. Lett.*, **108**, 090201.

Barton,J.P. *et al.* (2016) ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics*, **32**, 3089–3097.

Ferguson,A.L. *et al.* (2013) Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, **38**, 606–617.

Ganmor,E. *et al.* (2011) Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proc. Natl. Acad. Sci. USA*, **108**, 9679–9684.

de Juan,D. *et al.* (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.

Louie,R.H.Y. *et al.* (2018) Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl. Acad. Sci. USA*, **115**, E564–E573.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA*, **108**, E1293–E1301.

Quadeer,A.A. *et al.* (2018) Co-evolution networks of HIV/HCV are modular with direct association to structure and function. *PLoS Comput. Biol.*, **14**, e1006409.

Quadeer,A.A. *et al.* (2019) Identifying immunologically-vulnerable regions of the HCV E2 glycoprotein and broadly neutralizing antibodies that target them. *Nat. Commun.*, **10**, 2073.

Ricci-Tersenghi,F. (2012) The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *J. Stat. Mech. Theory Exp.*, **2012**, P08015.

Rivoire,O. *et al.* (2016) Evolution-based functional decomposition of proteins. *PLoS Comput. Biol.*, **12**, e1004817.

Sohl-Dickstein,J. *et al.* (2011) New method for parameter estimation in probabilistic models: minimum probability flow. *Phys. Rev. Lett.*, **107**, 220601.