BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 21 2014, pages 3128–3130
doi:10.1093/bioinformatics/btu500

Sequence analysis

Advance Access publication July 26, 2014

CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations

Stefan Seemayer[1,†], Markus Gruber[1,†] and Johannes Söding[1,2,*]

[1]Gene Center, LMU Munich, Feodor-Lynen-Strasse 25, 81377, Munich and [2]Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Recent breakthroughs in protein residue–residue contact prediction have made reliable *de novo* prediction of protein structures possible. The key was to apply statistical methods that can distinguish direct couplings between pairs of columns in a multiple sequence alignment from merely correlated pairs, i.e. to separate direct from indirect effects. Two classes of such methods exist, either relying on regularized inversion of the covariance matrix or on pseudo-likelihood maximization (PLM). Although PLM-based methods offer clearly higher precision, available tools are not sufficiently optimized and are written in interpreted languages that introduce additional overheads. This impedes the runtime and large-scale contact prediction for larger protein families, multi-domain proteins and protein–protein interactions.

Results: Here we introduce CCMpred, our performance-optimized PLM implementation in C and CUDA C. Using graphics cards in the price range of current six-core processors, CCMpred can predict contacts for typical alignments 35–113 times faster and with the same precision as the most accurate published methods. For users without a CUDA-capable graphics card, CCMpred can also run in a CPU mode that is still 4–14 times faster. Thanks to our speed-ups (http://dictionary.cambridge.org/dictionary/british/speed-up) contacts for typical protein families can be predicted in 15–60 s on a consumer-grade GPU and 1–6 min on a six-core CPU.

Availability and implementation: CCMpred is free and open-source software under the GNU Affero General Public License v3 (or later) available at https://bitbucket.org/soedinglab/ccmpred

Contact: johannes.soeding@mpibpc.mpg.de

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

Received on April 14, 2014; revised on June 24, 2014; accepted on July 17, 2014

1 INTRODUCTION

Evolutionary pressure to maintain a stable protein structure gives rise to correlated mutations between contacting residue pairs. These correlated mutations can be observed in a multiple sequence alignment (MSA) of the protein family and can be used to predict residue–residue contacts. A recent breakthrough was achieved by applying methods from statistics and statistical physics that aim at disentangling direct couplings from mere correlations between MSA columns (Ekeberg *et al.*, 2013; Kamisetty *et al.*, 2013; Marks *et al.*, 2011; Weigt *et al.*, 2009). This has resulted in a boost in contact prediction accuracy, thanks to which it is now possible to reliably predict protein structures using only sequence information if enough homologous sequences are available (Hopf *et al.*, 2012; Marks *et al.*, 2011; Nugent and Jones, 2012). Currently, 30% of Pfam families meet a reasonable criterion of having three homologous sequences per residue in the chain (see Supplementary Section 7 for details).

Modern contact prediction methods differ by their strategy in the disentangling step: the most accurate class of methods (Ekeberg *et al.*, 2013; Kamisetty *et al.*, 2013) such as plmDCA (Ekeberg *et al.*, 2013) and GREMLIN (Kamisetty *et al.*, 2013) learn the direct couplings as parameters of a Markov random field by maximizing its pseudo-likelihood, which has runtime complexity of $O(NL^2)$ where N is the number of homologous sequences in the MSA and L its number of columns. The less accurate methods based on sparse covariance matrix inversion such as PSICOV (Jones *et al.*, 2012) or Mean Field Direct Coupling Analysis use the sequence information only in a preprocessing step, while the main computation in $O(L^3)$ is independent of N. This makes them fast for short alignments (small L) but slow for large alignments. Whereas most protein families used for benchmarking so far have been relatively short, in practice, longer alignments are more relevant, for example, to predict interdomain or even interprotein contacts (Ovchinnikov *et al.*, 2014). Still, existing methods would take ~29 CPU years to complete large-scale studies such as the computation of contact predictions for the 30% of Pfam with sufficient sequence coverage (see Supplementary Section 7).

2 RESULTS

CCMpred implements the approach taken in plmDCA and GREMLIN, which is based on maximizing the pseudo-likelihood of an L_2-regularized Markov random field (see Supplementary Information for details). After successful optimization, the couplings are ranked by the Frobenius norms of the pairwise potentials and the average product correction (Dunn *et al.*, 2008) is applied to compute the final score.

As explained in the Supplementary Information, the task of computing the gradient of the pseudo-likelihood represents an almost ideal use-case for GPUs, as the computations can be run

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

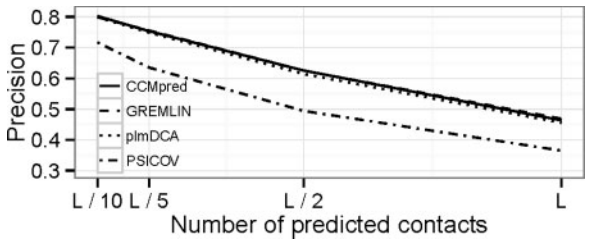

Fig. 1. Precision of contact prediction for increasing numbers of predicted pairs, normalized by length L of the target

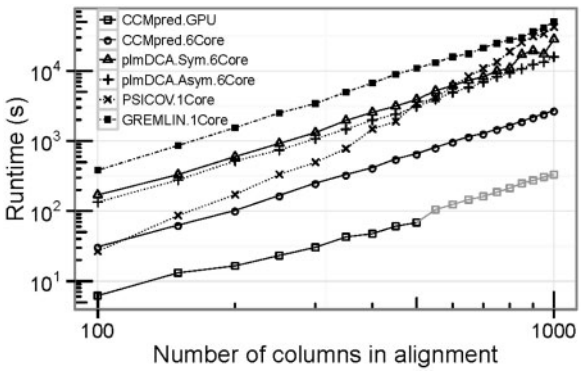

Fig. 2. Total runtimes on MSAs with 3000 sequences.

efficiently in parallel on the thousands of GPU processors with little idling due to memory access limitations.

We compare the runtimes and precisions of CCMpred with two other pseudo-likelihood maximization (PLM)-based tools, plmDCA (plmDCA-symmetric 3, plmDCA-asymmetric 1) and GREMLIN (version 2.01), and with the covariance matrix inversion-based tool PSICOV. The recently published FreeContact (Kaján *et al.*, 2014) software is much faster than PSICOV but clearly less accurate than plmDCA and GREMLIN and was not included here.

2.1 Precision

For benchmarking the precision of contact prediction methods, we use the same set of 150 Pfam families with ≥1000 sequences and high-resolution structures ($\leq 1.9\,\mathring{A}$) with identical input alignments as used in the PSICOV (Jones *et al.*, 2012) method. We rank the list of predicted contacts and determine the fraction of physical contacts (C_β distance $\leq 8\,\mathring{A}$) when selecting increasing numbers of contacts. Figure 1 shows that CCMpred is among the top tools.

2.2 Runtimes

For runtime benchmarks, we generated synthetic MSAs with 3000 sequences and 50, 100, …, 1000 columns (real alignments show similar speedups but exhibit more variance in their runtimes—see Supplementary Fig. S4 for details). Because GPUs and CPUs differ in their numbers of cores, frequency per core, etc., we attempt to make a fair comparison by comparing runtimes for hardware of similar price. We ran the GPU version of CCMpred on an NVIDIA GeForce GTX 780 Ti, all CPU-based methods on an Intel Xeon E5–2620 six-core processor. Alignments with $L > 500$ were run on a Tesla K40 GPU with 12 GB RAM (gray points).

Figure 2 shows the runtime of the methods for increasing alignment length. PSICOV is the fastest CPU method for small L, as its runtime is independent of sequence count. However, for $L \geq 150$, CCMpred becomes faster than PSICOV for alignments with typical numbers of sequences ($N \approx 3000$). At typical alignment lengths of $L = 300$, the CCMpred GPU code is 35 times faster than plmDCA, 113 times faster than GREMLIN and 16 times faster than PSICOV. On the same data, our CPU version is 4.3 times faster than plmDCA, 14 times faster than GREMLIN, 8.3 times slower than our GPU code and 2.0 times faster than PSICOV. For plmDCA and our CPU version, we use all six cores. PSICOV and GREMLIN do not support multithreading and, therefore, ran on a single core. However, even if implementations with perfect scaling existed (dividing runtimes by six), GREMLIN would still not be as fast as the CPU version of CCMpred. Our GPU code would be faster than a parallelized PSICOV at $L > 150$, and our CPU code would be faster at $L > 600$.

3 CONCLUSION

CCMpred is a fast GPU and CPU implementation of a top-performing PLM-based contact prediction approach that runs in a fraction of the time of comparably accurate methods. The speed increase is particularly important for long proteins and large-scale applications. Because CCMpred is free and open-source software, we hope that it also can serve as a basis for further methods development in this field.

ACKNOWLEDGEMENTS

The authors would like to thank Markus Meier for the distribution over SCOP domain lengths and Jessica Andreani, Susann Vorberg, Markus and Armin Meier for helpful comments and discussions.

Funding: This work was funded by the Deutsche Forschungsgemeinschaft (Grants GRK1721 and SFB646) and the Bavarian Center for Molecular Biosystems (BioSysNet). We thank NVIDIA Corporation for donating a Tesla K40 GPU used in this work.

Conflict of interest: none declared.

REFERENCES

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Ekeberg,M. *et al.* (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys. Rev. E*, **87**, 012707.

Hopf,T.A. *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Kaján,L. *et al.* (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics*, **15**, 85.

Kamisetty,H. *et al.* (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl Acad. Sci. USA*, **110**, 15674–15679.

Marks,D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PloS One*, **6**, e28766.

Nugent,T. and Jones,D.T. (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl Acad. Sci. USA*, **109**, E1540–E1547.

Ovchinnikov,S. *et al.* (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, **3**, e02030.

Weigt,M. *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.