BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 14 2013, pages 1817–1818
doi:10.1093/bioinformatics/btt272

Structural bioinformatics

Advance Access publication May 14, 2013

PconsD: ultra rapid, accurate model quality assessment for protein structure prediction

Marcin J. Skwark and Arne Elofsson*

Department of Biochemistry and Biophysics, Science for Life Laboratory, Swedish E-Science Research Center, Stockholm University, Box 1031, 17121 Solna, Sweden

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Clustering methods are often needed for accurately assessing the quality of modeled protein structures. Recent blind evaluation of quality assessment methods in CASP10 showed that there is little difference between many different methods as far as ranking models and selecting best model are concerned. When comparing many models, the computational cost of the model comparison can become significant. Here, we present PconsD, a fast, stream-computing method for distance-driven model quality assessment that runs on consumer hardware. PconsD is at least one order of magnitude faster than other methods of comparable accuracy.

Availability: The source code for PconsD is freely available at http://d.pcons.net/. Supplementary benchmarking data are also available there.

Contact: arne@bioinfo.se

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

Received and revised on April 16, 2013; accepted on May 7, 2013

1 INTRODUCTION

Predicting the 3D structure of protein from its sequence remains one of the yet unsolved problems of molecular biology. Recent improvements in the field allow for constructing more accurate models of protein structures, be it through relying on homologous structures, folding proteins *ab initio* or by combining multiple approaches. Regardless of the approach chosen, there is a need for a method of discriminating good (native-like) models from the mispredicted ones. This is the role of Model Quality Assessment Programs (MQAPs).

The most successful approaches to MQAP problem is the use of clustering methods, as introduced by the Pcons method in CASP5 (Lundstrom *et al.*, 2001). They are based on the premise that among different models of the same protein, the one that is most similar to the others is most likely to be correct. It is widely assumed that if a sufficiently accurate model generation method is used, most of the predictions will tend to cluster near the correct fold (Shortle *et al.*, 1998). Consequently, the largest cluster of protein models is most likely to contain models of native-like fold, and cluster centroid is the most likely candidate for the most accurate structure in the ensemble.

Most of clustering-based MQAPs rely on repeated pairwise structural superposition to find a largest subset of residues superimposable between models in question within a certain threshold. This approach is a computationally relatively expensive process, which renders clustering approaches unfeasible for model ensembles larger than some thousand proteins.

An alternative to rigid superposition is comparing the inter-residue distance matrices of the models. This approach has been successfully used for quality assessment, both in terms of similarity of predictions to the native structure (Ben-David *et al.*, 2009) and predicting the quality of models (McGuffin and Roche, 2010). Authors postulate that this approach is capable of capturing the interactions relevant for protein structure better than rigid $C\alpha$ superposition, as it accounts both for the local proximity of relevant residues (distances close to each other in the sequence space) and the general fold of the protein (long-range distances).

Although avoiding the computational expense of repeated superposition, distance matrix comparison still requires $0.5 \times m \times n^2$ operations for computing the distances and $0.25 \times m^2 \times n^2$ operations for their comparison, when given m models of n residues each.

In this work, we introduce PconsD, a new model quality assessment program, which uses a massively parallel, OpenCL-based streaming approach, which attempts to alleviate the scaling issues and provide an efficient platform for future development in distance-driven quality assessment.

2 APPROACH

PconsD is implemented in Python and OpenCL, a framework for massively parallel data-driven computation, that is capable of execution across heterogeneous platforms, such as CPUs or graphical processing units (GPUs). This work uses commodity GPUs as a platform of choice, but it can be easily ported to other platforms. GPU have been designed originally to handle 3D computer graphics computations, which require a vast amount of relatively simple, mutually independent calculations, based on large input matrices. The problem of computing distance matrices and comparing them is well suited to this computing paradigm.

PconsD operates on sets of models in PDB format, requiring the residue numbering to be consistent among models to maintain short running time. Except of the initial population of an array containing atom coordinates, all the computation happens on GPU, allowing for massively parallel data processing. Program computes similarity score between all the $C\beta$-$C\beta$ (in case of glycines—Cα) atoms in the protein, averages them and outputs the computed similarity score. Here, PconsD relies on linear scoring function with 5 Å cut-off ($\max(1 - \frac{|(d_{i,j}^{m1} - d_{i,j}^{m2})|}{5}, 0)$, but other functions can also be used, see Supplementary Material.

To assess the time performance, four we ran four quality assessment methods—PconsD, ModFoldClustQ (McGuffin and Roche, 2010), Pcons (Larsson *et al.*, 2009) and a naïve predictor based on

*To whom correspondence should be addressed.

pairwise model comparison by TM-scores (Zhang and Skolnick, 2004), on subsets of randomly selected models or varying size (from 10 models to 250 000 models). All the models have been preloaded into memory, and all the softwares were stored on a Linux `tmpfs` random access memory disk, to avoid the effect of i/o on the benchmarking.

3 RESULTS

PconsD has been blindly benchmarked in model quality assessment category of CASP10 experiment. According to CASP assessors, PconsD is at least as good a model classifier as other clustering MQAP methods (Pcons Lundstrom *et al.*, 2001) and comparable with compound methods [methods combining multiple approaches to quality assessment, such as Pcomb (Larsson *et al.*, 2009), ModFoldclust2 (Roche *et al.*, 2011), MUFOLD-QA (Wang *et al.*, 2011)], both as far as distinguishing good models from bad ones, and as far as selecting best model in the ensemble are concerned, see Table 1. PconsD selected more good models (with less than two GDT-TS (Zemla *et al.*, 1999) points loss to the best model in the ensemble) than Pcons (one of state-of-art clustering methods) or Pcomb (the best MQAP in CASP10 in terms of picking best models). It also picked fewer poor models than Pcons (>10 GDT-TS points loss to the best model in ensemble). The slightly worse performance in terms of average δ GDT-TS is due to prediction targets on which clustering did not work sufficiently well (i.e. 'difficult', free-modeling targets). PconsD seemed to be slightly less suitable for ranking models, when considering the Pearson correlation coefficient with superposition-based metrics, such as GDT-TS, but it still performs better in this category than MUFOLD-QA.

The main strength of PconsD lies in its outstandingly short prediction time. The prediction performance and scaling have been assessed on a set of up to 250 000 models (55 GB of PDB files) of PTS EIIA type-2 domain from *Escherichia coli* (pdbid: 1a3a:A), produced by *ab initio* folding protocol implemented by ROSETTA (Bradley *et al.*, 2005). The benchmark results show that PconsD for non-trivial task sizes is at least an order of magnitude faster than Pcons, which is already almost two orders of magnitude faster than the naïve clustering using TM-score, thanks to aggressive optimization (Wallner and Elofsson, 2007), Figure 1.

PconsD scales linearly with the amount of models, until it needs to perform problem domain partitioning, on which moment it starts scaling quadratically. On the Nvidia GTX 560 used for testing PconsD, the running time with 10^8 contacts (5000 models of 143 residues each, 1000 of 300 residues and so forth) is <1 min.

Table 1. Quality assessment performance in CASP10

	PconsD	Pcons.net	Pcomb	MFC2	MUF
Pearson r	0.90	0.92	0.91	0.93	0.88
Δ GDT-TS	5.21	5.45	4.78	4.79	4.89
Δ GDT-TS ∈ [0, 2]	0.30	0.23	0.26	0.33	0.34
Classification (MCC)	0.84	0.85	0.84	0.85	0.82

Note: Official results from CASP10. Pearson r: correlation between method score and GDT-TS. delta GDT-TS: average GDT-TS loss that is difference in GDT-TS between top-ranked model and best model in model ensemble. Next row specify the fraction of targets for which the selected model had a negligible (<2) GDT-TS loss. Classification shows Matthew's correlation coefficient for discriminating between correct and incorrect models. Comparison is done between two pure clustering methods (PconsD and Pcons.net) and three compound methods (Pcomb, MFC2: ModFoldclust2, MUF: MUFOLD-QA).

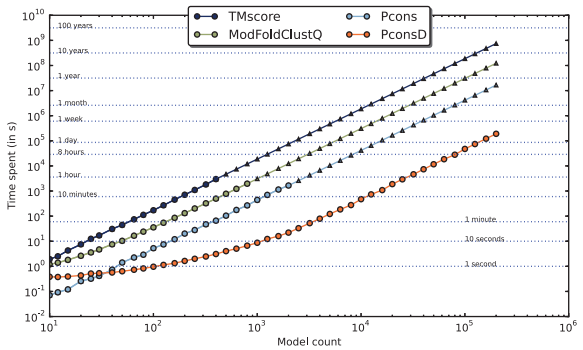

Fig. 1. Running times for the quality assessment methods mentioned in the text. The time is measured for a pairwise comparison of X number of models of a 143 AA long protein (pdb: 1A3A). Circles: measured prediction times, Triangles: extrapolated times

ACKNOWLEDGEMENT

The benchmarking results were kindly provided to us by Andriy Kryshtafovych.

Funding: Swedish Research Council (VR-NT 2009-5072, VR-M 2010-3555), SSF (the Foundation for Strategic Research) and Vinnova through the Vinnova-JSP program, the EU 7'Th Framework Program by support to the EDICT project (FP7-HEALTH-F4-2007-201924). Computational resources were provided by KFI and SNIC. M.J.S. has been funded by TranSys, a Marie Curie ITN (No FP7-PEOPLE-2007-ITN-215524).

Conflict of Interest: none declared.

REFERENCES

Ben-David,M. *et al.* (2009) Assessment of CASP8 structure predictions for template free targets. *Proteins*, **77** (**Suppl. 9**), 50–65.

Bradley,P. *et al.* (2005) Toward high-resolution de novo structure prediction for small proteins. *Science*, **309**, 1868–1871.

Larsson,P. *et al.* (2009) Assessment of global and local model quality in CASP8 using Pcons and ProQ. *Proteins*, **77** (**Suppl. 9**), 167–172.

Lundstrom,J. *et al.* (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.*, **10**, 2354–2362.

McGuffin,L.J. and Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics*, **26**, 182–188.

Roche,D.B. *et al.* (2011) The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. *Nucleic Acids Res.*, **39** (**Suppl. 2**), W171–W176.

Shortle,D. *et al.* (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci. USA.*, **95**, 11158–11162.

Wallner,B. and Elofsson,A. (2007) Prediction of global and local model quality in CASP7 using Pcons and ProQ. *Proteins*, **69** (**Suppl. 8**), 184–193.

Wang,Q. *et al.* (2011) MUFOLD-WQA: A new selective consensus method for quality assessment in protein structure prediction. *Proteins*, **79** (**Suppl. 10**), 185–195.

Zemla,A. *et al.* (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, (**Suppl. 3**), 22–29.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.