Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

Structural bioinformatics

pyconsFold: Modelling and docking using distance contact predictions.

Lamb, J [1,2], Elofsson, A [1,2]*

[1] Science for Life Laboratory, Stockholm University, SE-171 21 Solna, Sweden
[2]Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Contact predictions within a protein has recently become a viable method for accurate prediction of protein structure. Using predicted distance distributions has been shown in many cases to be superior to only using a binary contact annotation. Using predicted inter protein distances has also been shown to be able to dock protein chains.
Results: Here, we present pyconsFold, which uses CNS together with distance predictions as restraints for protein modelling. It builds on the popular CONFOLD script (written in Perl) and is rewritten in python both with ease of install, reproducibility and longevity in mind. The package also includes utilities for handling distance predictions from trRosetta and Quality Assessment tools together with a new novel way of using predicted inter protein contacts to simultaneously fold and dock two protein chains.
Availability and implementation pyconsFold is implemented in Python 3 with a strong focus on using as few dependencies as possible for longevity. It is available both as a pip package in Python 3 and as source code on GitHub and is published under the GPLv3 license.
Contact: arne@bioinfo.se
Supplementary information: The full source code and documentation are available from https://github.com/johnlamb/pyconsfold.

1 Introduction

De novo protein modelling has recently seen significant improvements by relying on contact predictions that have been presented in binary format, two residues are either in contact (within 8Å) or not. By running mathematical models such as DCA on multiple seqeunce alignments we can extract possible contacts based on coevolutionary information and when these contacts are filtered using image recognition inspired deep learning we get good enough quality of the contacts to use for structure prediction. In the last few years we have again seen a significant improvement in structure prediction by using predicted distances between residues instead of just contacts(CASP13 Kryshtafovych *et al.* (2019); Yang *et al.* (2020)). Historically, classifications has been easier to achieve (contact or not) in ,achine learning than regression but new advances, particularly in deep learning architecture, has now shown that these new methods can produce distances that results in better models than using just regular binary contacts.

CONFOLDAdhikari *et al.* (2015) is a wrapper around CNS Brünger *et al.* (1998); Brunger (2007) that uses predicted binary contacts together with predicted secondary structure to model proteins. pyconsFold is a re-implementation and extension of CONFOLD that achieves better results using distance predictions and that also expands to allow for more geometric restraints, such as angles predicted by tools such as trRosetta Yang *et al.* (2020).

2 Implementation

pyconsFold is implemented in pyhon 3 as a wrapper around the CNS system used for folding proteins from NMR experimental data. It leverages CNS underlying power by converting predicted distances into geometric restraints. CNS ability to handle multiple chains are also leveraged to achieve the modelling and docking of complexes at the same time by using both inter and intra chain predicted contacts.
pyconsFold mainly runs in one stage as previous work Michel *et al.* (2017)

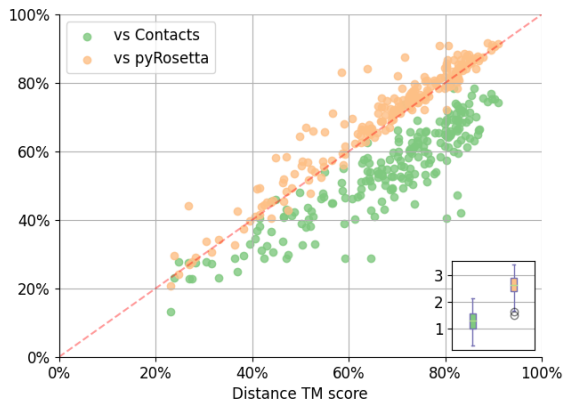

Fig. 1. pyconsFold distance prediction against both contact base (green) and pyRosetta (orange) models. Distance predictions outperforms contact based predictions in almost all cases. pyconsFolds distance predicted models perform almost as well as pyRosetta but is around 20 times (more than 1 order of magnitude) faster per model as the inset shows(log10 scale of per model time in seconds. pyconsFold in green and pyRosetta in orange).

has shown that running two stages only give marginal gains to the TM-score but more than doubles the run time. pyconsFold has three major modes:

2.1 Contact based modelling

Contact based modelling or legacy modelling treats all contacts as binary contacts, they are in contact within 8Å. This mode can simulate CONFOLD by also including secondary structure prediction and setting the gaps parameter to 5, meaning only contacts separated by at least 5 residues will be used. In this mode it is most common to use a numbers threshold for how many contacts to use by the convetion of a multiple of protein length, eg 1.5L would mean to use 1.5 times the protein length number of contacts.

2.2 Distance based modelling

Distance based modelling works similar to classical modelling but in this case, an extra column is expected in the CASP-contacts file which defines the standard error for the predicted distance of the contact. By default this mode also includes contacts within 5 residues from each other and does not require a secondary structure file. Benchmark has even shown that this mode does not produce better models with a secondary structure file.

2.3 Docking modelling

Docking modelling uses CNSs underlying ability to model two protein chains simultaneously. A contact prediction file of both inter and intra contacts needs to be supplied. This can be achieved by running alignment tools for the two chains separately, concatenating these files horizontally and then treating this resulting alignment as one chain and using any contact prediction method to predict contacts. This contact file together with the original fasta files are used as inputs and models the full binary complex.

3 Additional features

For ease of use and reproducibility we have included several extra features and utilities. By default the generated models are ranked by CNS internal NOE energy but Quality Assessment score from pcons Lundström *et al.* (2008) will also be calculated. If a native structure is known and supplied with the tmscore_pdb_file argument, the tmscore Zhang and Skolnick (2007); Xu and Zhang (2010) for each model against the native structure will be calculated. Compiled versions of both pcons and TMscore for unix based x64 systems are packaged together with pyconsFold under the open source Boost license. If your system does not support the built in versions, you can manually install them on your system and as long as they are in your path, will be chosen instead of the built in binaries.

4 Docking

5 Discussion

pyconsFold offers a complete toolkit for ab initio modelling using predicted contact distances and angles. Its focus is on ease of use and reproducibility and is available both as source on github and as an easily installable pip package in Python 3. It comes packaged with QA-programs to rank the generated models and allows transparency for many tweakable parameters for the underlying CNS-system. It also offers a docking protocol where predicted inter chain contacts are used as restraints for docking. It offers a significant increase in accuracy over contact based protocols by using predicted distances. It is also around 20 times faster than pyRosetta per model.

Acknowledgements

This work was supported by grants from the Swedish Research Council (VR-NT 2016-03798 to A.E.), National Science Centre (No. 2012/07/E/NZ1/01900 to J.I.S. and No. 2018/29/N/NZ2/02897 to A.I.J.).

Conflict of Interest Statement

None delared.

References

Adhikari, B., Bhattacharya, D., Cao, R., and Cheng, J. (2015). CONFOLD: Residue-residue contact-guidedab initioprotein folding.

Brunger, A. T. (2007). Version 1.2 of the crystallography and NMR system.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII.

Lundström, J., Rychlewski, L., Bujnicki, J., and Elofsson, A. (2008). Pcons: A neural-network-based consensus predictor that improves fold recognition.

Michel, M., Hurtado, D. M., Uziela, K., and Elofsson, A. (2017). Large-scale structure prediction by improved contact predictions and model quality assessment. *Bioinformatics*, **33**(14), i23–i29.

Xu, J. and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5?

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, **117**(3), 1496–1503.

Zhang, Y. and Skolnick, J. (2007). Scoring function for automated assessment of protein structure template quality.