*Bioinformatics*
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Genome analysis

# pyconsFold: Ab initio 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

**Summary:** Contact predictions within a protein are most commonly used in tertiary model prediction of protein structure. Using predicted distance distributions has been shown in many cases to be superior to only using a binary contact annotation, either in contact or not. pyconsFold builds on the CNS NMR modelling program Brünger *et al.* (1998); Brunger (2007) to use contact predictions as restraints for use in modelling. It builds on the popular CONFOLD-perl Adhikari *et al.* (2015) script 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 Yang *et al.* (2020) and Quality Assessment tools together with a new novel way of using predicted interprotein contacts to dock two separate protein chains. In contrast to most other folding protocols in use today that uses different approaches of deep learning, pyconsFold builds around the CNS NMR system that is mainly used to fold proteins from NMR spectroscopy data and where the full folding of predicted contacts is the same as the folding when using experimental data.
**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 Gnu GPL v3 license.
**Contact:** arne@bioinfo.se
**Supplementary information:** The full source code together with further documentation and examples are available from https://github.com/johnlamb/pyconsfold.

## 1 pyconsFold

Protein modelling has long relied on contact predictions that have been presented in binary format, two residues are either in contact (within 8Å) or not. More and more methods are now leveraging full contact distance predictions and this shows (CASP13 Kryshtafovych *et al.* (2019); Yang *et al.* (2020)) a higher accuracy of the models than if binary contacts were used. CONFOLD is a wrapper around CNS that uses predicted binary contacts together with predicted secondary structure to model proteins. pyconsFold is a reimplementation 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 and for fast iteration of models and parameters.

pyconsFold has three major modes:

### 1.1 Contact based modelling

Contact based modelling works by treating contacts as an upper limit to distances. CONFOLD style modelling can be replicated using this mode with contact prediction and secondary structure prediction as input and setting the gaps parameter to 5(only use contacts that are at least 5 residues away from each other).

### 1.2 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 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.

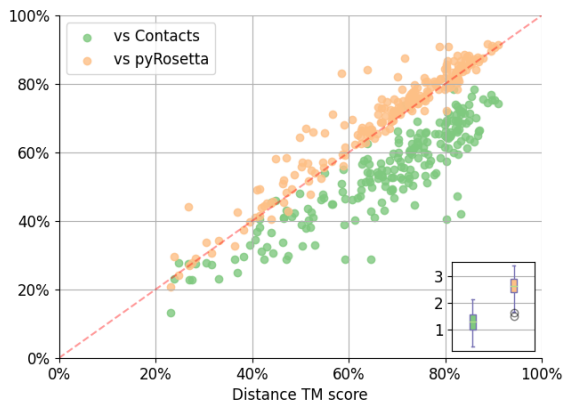


**Fig. 1.** pyconsFold distance prediction against both contact base (green) and pyRosetta (orange) models. Distance predictions outperforms contact based predictions in almost all cases. pyconsFolds distance predicted models perform almost as well as pyRosetta but is around 20 times (more than 1 order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

### 1.3 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.

## 2 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, wil be chosen instead of the built in binaries. l

## 3 Conclusion

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.

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.