

**Using PconsC4 and PconsFold to predict protein structure**

Running Title: PconsC4 and PconsFold

Author(s) Claudio Bassot, David Menendez Hurtado, Arne Elofsson

Contact information: Dep. of biochemistry and biophysics and Science for Life Laboratory, Box 1031 171 52 Solna

**Significance Statement**

Nowadays there is a huge gap between the number of proteins sequences that are available and their structural information. Using PconC4 starting from the primary structure of a protein is possible to predict some of its residue-residue contacts in the tertiary structure. The contact prediction and its use as constraints in the 3D protein modelling simplify significantly the folding problem allowing to model the protein structure from its amino acid sequences.

**ABSTRACT**

Although the significant increase in the number of resolved protein structures, for several proteins the structural information is missing. Even though the structural information is codified in the amino acid sequence, its computational prediction is still an unsolved problem. One of the most succeeding method model the protein structure starting the from primary sequence is the contacts prediction starting from a multiple sequence alignment (MSA).

Here we used our contact predictor PconsC4 to generate a list of probable contacts between residues in the primary sequences. These contacts were then used together with the secondary structure prediction as constraints for the CONFOLD folding method. In this way, a 3D protein model can be built starting directly from the primary sequence.

**Keywords: Protein structure prediction, Contact prediction**

**INTRODUCTION**

**Introduction**

The computational determination of the protein structure is still an unsolved problem. Until recently the only available method for large-scale structure prediction was template-based modeling [(Kryshtafovych et al., 2018)](https://paperpile.com/c/tDy6ql/20bh). However, template-based modeling cannot be applied when no homologous structures exist, neither when long insertions are present in the target sequence. Due to these limitations, the development of template-free modeling methods are important. So far the most promising template-free modeling methods are based on residue-residue contact predictions [(Abriata, Tamò, Monastyrskyy, Kryshtafovych, & Dal Peraro, 2018)](https://paperpile.com/c/tDy6ql/PrjY). Contact prediction methods take advantage of the observation that the evolution of residues are constrained by the interactions that they form in the tertiary structure [(Pazos, Helmer-Citterich, Ausiello, & Valencia, 1997)](https://paperpile.com/c/tDy6ql/XxjY).

Here, we present how to use the PconsFold2 pipeline for template-free protein modeling [(Michel, Menéndez Hurtado, Uziela, & Elofsson, 2017)](https://paperpile.com/c/tDy6ql/OpYz). PconsFold2 is based on our recently developed contact predictor PconsC4, the CONFOLD folding method [(Adhikari, Bhattacharya, Cao, & Cheng, 2015)](https://paperpile.com/c/tDy6ql/OOb7) and the model quality assessment programs ProQ3D and Pcons. PconsC4 performs as well, or even better than earlier methods and is also freely available, significantly faster and is distributed as a Python library, so it can be easily adapted to other workflows, and it gives access to predictions at different distance thresholds as well as S-score [(Michel,](https://paperpile.com/c/tDy6ql/lJPn) [Menéndez](https://paperpile.com/c/tDy6ql/OpYz) [Hurtado, & Elofsson, 2018)](https://paperpile.com/c/tDy6ql/lJPn).

Finally, the models are evaluated using a set of model quality estimation methods developed in our lab.

**STRATEGIC PLANNING**

This input to the pipeline described below is a protein sequence. However, some thoughtfulness is required in choosing the sequence to optimize the probability of success. The optimal input sequence is a single domain. Therefore, it might be advantageous to exclude long terminal loops and to divide multi-domain proteins into the domains. However, there exist no best practice that works in each case, therefore, it is often advisable to try several different input sequences.

***BASIC PROTOCOL 1***

**CONTACT PREDICTION AND STRUCTURE GENERATION**

In the PconsFold2 pipeline for contact based modeling the protein structure is generated by CONFOLD using predicted contacts from PconsC4 and secondary structure constraints from PSIPRED [(McGuffin, Bryson, & Jones, 2000)](https://paperpile.com/c/tDy6ql/r2jd). The contacts are predicted by PconsC4 using a multiple sequence alignment (MSA) generated with Jackhmmer [(Eddy, 2011)](https://paperpile.com/c/tDy6ql/bDhH) or HHblits [(Remmert, Biegert, Hauser, & Söding, 2011)](https://paperpile.com/c/tDy6ql/akyR). The pipeline of the protocol is summarized in Figure 1.

***Necessary Resources***

#### **Hardware**

A computer running Linux or some similar operating system

**Software**

* Jackhmmer [(Eddy, 2011)](https://paperpile.com/c/tDy6ql/bDhH) or HHBlits [(Adhikari et al., 2015; Remmert et al., 2011)](https://paperpile.com/c/tDy6ql/OOb7+akyR).
* CONFOLD [(Adhikari et al., 2015; Brunger, 2013)](https://paperpile.com/c/tDy6ql/8yxa+OOb7) and CNS [(Brunger, 2013)](https://paperpile.com/c/tDy6ql/8yxa).
* Python 3 including dependencies
* PconsC4 [(Michel et al., 2018)](https://paperpile.com/c/tDy6ql/lJPn)

**Installations and download**

1. **Jackmmer or HHblits**

Install HMMER3 [(Eddy, 2011)](https://paperpile.com/c/tDy6ql/bDhH) or HHblits [(Adhikari et al., 2015; Remmert et al., 2011)](https://paperpile.com/c/tDy6ql/OOb7+akyR): the two programs are comparable. However the MSAs from HMMER3 give marginally better results but HHblits MSA generation is faster.

Jackhmmer (used for MSA generation) is contained in the HMMER3 package that can be easily installed following the provided instruction <http://hmmer.org/documentation.html>. In addition to Jackhmmer you have to download a protein database to perform the search. Here, we suggest Uniprot90 from <https://www.uniprot.org/downloads>.

Alternatively, HHblits can be installed as described in <https://github.com/soedinglab/hh-suite>. For HHblits we do recommend to use the Unicluster30 database available from <http://gwdu111.gwdg.de/~compbiol/uniclust/2017_10>.

1. **Script for MSA modifications**

PconsC4 supports several formats, including A3M generated by HHblits. When using Jackhmmer, the file must be converted using the library Stockholm reformat: <https://pypi.org/project/stockholm_reformat/>

It can be installed with

> pip3 install -U stockholm\_reformat

1. **Python3.5**

Python vs 3.5 or later is necessary, and the following python packages also need to be installed:

> pip3 install -U tensorflow

> pip3 install numpy Cython pythran

1. **PconsC4 Installation**

PconsC4 tarball can be downloaded from <https://github.com/ElofssonLab/PconsC4/releases/>

Once finished the download PconsC4 can be easily installed with the command:

> pip3 install pconsc4

Scripts for the usage of PconsC4 and for the calculation of the percentage of used contacts can be download from:

https://github.com/ElofssonLab/Contact\_prediction\_scripts

**5.**  **CONFOLD Installation**

Confold can be download from <http://sysbio.rnet.missouri.edu/multicom_toolbox/tools.html#license> The installation of CONFOLD and its required dependency CNS is explained here <https://github.com/multicom-toolbox/CONFOLD>.

Protocol steps—*Step annotations*

To better describe our protocol we present an example taken from the models we generate for the recent CASP13 competition; the *Klebsiella pneumoniae’*s protein B5Y0C2.

1. **Generate the input for PconsC4**

The input file for PconsC4 is a Multiple Sequence Alignment (MSA). The MSA for the target protein can be generated using Jackhmmer or HHblits.

First, it is necessary to create a FASTA file, named “B5Y0C2.fasta” in this example, containing the sequence:

>B5Y0C2, *Klebsiella pneumoniae*, 126 residues

VPEITTAQTIANSVVDAKKFDYLFGKATGNSHTLDRTNQLALEMKRLGVADDINGHAVLAEHFTQATKDSNNIVKKYTDQYGSFEIRESFFIGPSGKATVFESTFEVMKDGSHRFITTIPKNGVTK

The FASTA file is then used as input for Jackhmmer. Below an example of the Jackhmmer running command, were B5Y0C2.fasta is the input sequence and uniref90.fasta the database used for searching. In our example MSA was generated with a more inclusive parameters of 5 iterations and an e value of E 1. However in general we suggest a more conservative approach using 3 iterations and an E-value of 0.1. The step can take several minutes.

> jackhmmer --noali -A B5Y0C2.sto -N 3 -E 0.1 --incE 0.1 --cpu 6 B5Y0C2.fasta uniref90.fasta

Where:

**-A** is the output file

**-N** is the number of iteration

**-E** the reporting threshold and **--incE** the inclusion threshold both expressed in E-value

**--cpu** the number of core in use

Alternatively, HHblits and a corresponding database can be used to generate the multiple sequence alignment.

> hhblits -i B5Y0C2.fasta -o B5Y0C2.hhblits -oa3m B5Y0C2.hhblits.a3m -all -n 3 -e 0.1 -d uniclust30\_2017\_04

Where**:**

**-i** input file

**-o** output file

**-oa3m** output a3m

**-all** show all sequences in result MSA

**-n** number of iterations

**-e** E-value cutoff for inclusion in result alignment

**-d** databased

1. **MSA modifications**

If using Jackhmmer, the multiple sequence alignment B5Y0C2.sto has to be modified to A3M format before it is used as an input to run PconsC4.

> stockholm\_to\_a3m B5Y0C2.sto B5Y0C2.a3m

1. **MSA evaluation**

The contact prediction is based on the multiple sequence alignment. The optimum MSA contains the highest number of sequences but only within the same family of the target. For this reason, the quality of the MSA can be strongly affected by the E-value choice.

The E-value can be defined as the number of hits that are expected to see by chance when searching in a database of a certain size. In practice, a lower E-value corresponds to a higher significance and setting up a lower E-value for the alignment generation result in a stricter selection of the sequences.

Our default pipeline use an E-value of 0.1 that maximise the number of collected sequences, but for some protein families, a lower E-value can lead to better result.

For assessing the quality of the alignment before to use it in the contact prediction, we perform the following procedure that requires Hmmbuild and Hmmsearch that are included in the HMMER package.

First **Hmmbuild** reads the multiple alignment “B5Y0C2.a3m” and builds a new profile HMM “B5Y0C2.hmm”

> hmmbuild B5Y0C2.hmm B5Y0C2.a3m

Subsequently, **Hmmsearch** calculate the E-value for each sequence. Looking at the E-value of the target sequence is possible to estimate the quality of the alignment.

> hmmsearch -o B5Y0C2.hmmsearch.out -E 10000000 -Z 100000000 B5Y0C2.hmm B5Y0C2.fasta

-o outputfile

-E calculate the E-value of all the sequences with a value below 10000000

-Z calculate the E-value on a fixed number of sequence of 100000000

It has to be noted that it is necessary to set up a defined number of sequences (with the flag -Z) to compare different alignments because the default option calculates the E-value on the number of sequences of the alignment that varies.

Our suggestion is to reduce the E-value cut-off or the number of iterations used in the alignment generation until the E-value of the target sequence reaches at least 1e-20 and the full starting sequence is aligned to the HMM.

1. **Run PconsC4**

PconsC4 can be run as a python module from a python3 script. In the Github repository (<https://github.com/ElofssonLab/Contact_prediction_scripts> ) we do provide an example file, run\_pconsc4.py that can be used. The script can be run as:

> python3 run\_pconsc4.py B5Y0C2.fasta B5Y0C2.a3m B5Y0C2\_contacts.rr B5Y0C2\_ss\_structure.ss

Where:

B5Y0C2\_contacts.rr is the output file in the CASP RR format. The first and the second column of each row in the file correspond to the first and second residues of the contact, and the fifth column is the PconsC4 score (higher is better).

The min\_sep=5 option assures us to save only the contacts between residues separated by a minimum distance of 5 residues, closer contacts would not be relevant for protein modeling.

The script run\_pconsc4.py produce as well the output B5Y0C2\_ss\_structure.ss. It contain the secondary structure prediction for the B5Y0C2.a3m that will be used as input by CONFOLD.

The run\_pconsc4.py script creates a plot of the contact map, as in Figure 2

1. **Run Confold**

The contact prediction in the CASP format (B5Y0C2\_contacts.rr) and the secondary structure prediction (B5Y0C2\_ss\_structure.ss) can be used from CONFOLD for the protein structure modeling.

An example of command to run CONFOLD is:

> confold -seq B5Y0C2.fasta -rr B5Y0C2.a3m.rr -ss B5Y0C2.a3m.ss -o output\_directory -selectrr 2L -stage2 1 -mcount 50

Where:

-selecttr is the number of contacts in proportion to the length of the protein (e.g. 1 = L where L is the length of the protein sequence). We use a -selectrr of 2 assuming that on average each residue form two contacts.

-stage2 is for produce the pdb structure

-mcount is the number of structures produced

In Figure 3 the first ranked PDB model obtained from the CONFOLD contact based modeling of the B5Y0C2 protein is shown.

1. **Quality assessment of the model**

The final step in the PconsFold2 pipeline is to evaluate the quality of the model produced. This is done by three different methods, PPV, ProQ3D and Pcons.

**6.1 Number of satisfied contacts constraints (PPV)**

For assessing the quality of the models we suggest to take into account the number of satisfied contacts constraints expressed as positive predicted values (PPV). The quality of the model is strongly dependent by the quality of the predicted contacts, if the contacts are mostly predicted wrongly the quality of the model will be low and the contacts used as constraints will often be poorly respected.

We provide in the Github repository (<https://github.com/ElofssonLab/Contact_prediction_scripts> ) a script “ppv.py” for calculating the correct contacts. A typical use is the following:

> python ppv.py -f 2 B5Y0C2.fasta B5Y0C2.a3m.rr B5Y0C2\_pdb\_file

-f has to be the same L value as used in Confold.

Our experience suggests that a model with a ppv lower than 0.4 will rarely be correct [(Michel et al., 2014)](https://paperpile.com/c/tDy6ql/h7YL). The PPV.py script will print three numbers in order: Positive Predicted Values, True Positive, True Negative. For our B5Y0C2 model the PPV is 0.778. It means that the 77.8% of the contacts were successfully satisfied by the model.

**6.2 ProQ3**

A second quality assessment method independent from the predicted contact is ProQ3 [(Uziela, Menéndez Hurtado, Shu, Wallner, & Elofsson, 2017)](https://paperpile.com/c/tDy6ql/MyqD) that make use of the Rosetta energy terms. We suggest using the web server <http://proq3.bioinfo.se/> or to download the docker image from <https://bitbucket.org/ElofssonLab/proq3/>.

In Figure 4 is shown the screenshot of the web server. The model’s PDB file and the FASTA sequence have to be upload on the web page.In our test case, the model gives a ProQ3D score of 58.316, an average value compared with the scores of the models in the CASP11 protein modeling competition [(Uziela, Menéndez Hurtado, Shu, Wallner, & Elofsson, 2018)](https://paperpile.com/c/tDy6ql/jYLL). From our experience the model is acceptable above a score of 50.

**6.3 Pcons**

Finally, model quality estimation can be used by examining the similarity between the models generated by the folding pipeline. This can easily be done by using Pcons [(Lundström, Rychlewski, Bujnicki, & Elofsson, 2001)](https://paperpile.com/c/tDy6ql/OtAy). Pcons can be obtained from <https://github.com/bjornwallner/Pcons> and is easily run using the command:

> pcons -A -d models\_directory

Above a Pcons score of 50 the models can be positively evaluated. If the different methods agree in evaluate positively the model, it can be safely considered correct.

***ALTERNATE PROTOCOL 1***

**Use of PSIpred for the secondary structure prediction**

In our basic protocol we obtain the secondary structure prediction from PconsC4, this is the faster approach. Alternatively is possible use the PSIpred or every other software to predict the secondary structure. In all the case is fondamental convert the output in the CONFOLD format.

***Necessary Resources***

* PSIPRED [(McGuffin et al., 2000)](https://paperpile.com/c/tDy6ql/r2jd).

**PSIPRED Installation**

Predictions can either be obtained from the web server <http://bioinf.cs.ucl.ac.uk/psipred_beta/> or by installing the local version following the instruction in <http://bioinfadmin.cs.ucl.ac.uk/downloads/psipred/README>.

Protocol steps—*Step annotations*

PSIpred can be run (on the web-server or installed locally) starting from the a3m multiple sequence alignment selecting the option for the horizontal output.

The output has to be converted from the horizontal PSipred output to the format accepted by CONFOLD. It looks like a “fasta” formatted file with the letters E, H, C, indicating respectively: beta-sheet, alpha helix, coil. For example:

> PSIpred\_prediction

CCCCCCCCCCCCEEECHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCHHHH

To convert the horizontal PSIpred is possible use a bash script like the following:

> printf '>PSIpred\_prediction \n' > B5Y0C2.a3m.ss | grep 'Pred:' PSIPred\_output | cut -c6- | tr -d '\n ' >> B5Y0C2.a3m.ss

With this command we grep only the secondary prediction line form the horizontal PSIpred prediction, and we add the description '>PSIpred\_prediction’ on the top of it.

**COMMENTARY**

**Background Information**

Despite the growing number of resolved protein structures, for many proteins is not available a suitable template for homology modeling. However, in the past few years, a significant improvement in the modeling of these template free targets was performed thanks to the successful use of contact prediction. [(Abriata et al., 2018)](https://paperpile.com/c/tDy6ql/PrjY)

The identification of contacts between residues simplifies the folding problem allowing the identification of the three-dimensional structure of many proteins [(Hopf et al., 2014)](https://paperpile.com/c/tDy6ql/HLxX). Recently high accuracy models for hundreds of proteins family were obtained [(Hopf et al. 2014; Ovchinnikov et al. 2017; Michel et al. 2017)](https://paperpile.com/c/tDy6ql/HLxX+tMXz+OpYz) confirming the utility of contacts prediction based methods for solving unknown protein structures.

An alternative use of the contacts prediction present in literature is to reduce the ambiguity of a low-resolution experimental structure like in case of Cryo-EM ATP synthase A subunit in

[(Zhou et al., 2015)](https://paperpile.com/c/tDy6ql/cRsv).

Our pipeline provides a fast method for template free modeling of protein that can be scaled up for several thousand proteins. Its main limitation is the generation of a sufficiently deep MSA. Indeed in particular for proteins unique to eukaryotes the low number of retrieved sequences could be a limiting factor.

**Critical Parameters**

The PconsFold2 pipeline will only provide a successful model if the predicted contacts are correct. There are several reasons why this might not be the case. In some cases no contacts are reliably predicted. This can be seen in low scores of the contacts and low scores from the model quality evaluations.

Alternatively, in a few cases the MSA can be wrong as it has drifted away to another family. As previously mentioned in these cases, a critical parameters for our pipeline are the E-value threshold and the number of iterations in the multiple sequence alignment. The both the parameters affects the quality of the starting MSA and consequently the quality of the contact prediction but the optimum value change from a protein to another. For this reason, it is essential to perform the MSA evaluation step and in case of poor result for the target sequence repeat the MSA generation with different E-value and iterations number.

Finally, lower quality models can be obtained if the secondary structure prediction is incorrect. Here, alternative methods can be examined.

**Troubleshooting**

**ACKNOWLEDGEMENT**

This work was supported by grants from the Swedish Research Council (VR-NT 2016-03798 to AE) and Swedish e-Science Research Center. The Swedish National Infrastructure provided computational resources for Computing (SNIC).

**FIGURE LEGENDS**

**Figure 1.** Graphical abstract of the protocol. In grey the mandatory steps in with the alternative one. Figure modified from [(Lamb et al., n.d.)](https://paperpile.com/c/tDy6ql/H164)

**Figure 2.** Contact map predicted by PconsC4 for the protein B5Y0C2. On the axis is indicated the progressive numeration of the protein residues while the purple dots correspond to a contact between the residues. The intensity of the colouration varies with the score; higher is the score darker is the colour.

**Figure 3.** Model build by CONFOLD based on the contacts predicted by PconsC4.

**Figure 4.** Screenshot of the ProQ3 web server. On the left the input page where has to upload the PDB file and the FASTA sequence. On the right, the result page of our example with the ProQ3 score highlighted with a red box.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**LITERATURE CITED**

[Abriata, L. A., Tamò, G. E., Monastyrskyy, B., Kryshtafovych, A., & Dal Peraro, M. (2018). Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins*, ***86*** *Suppl 1*, 97–112. https://doi.org/](http://paperpile.com/b/tDy6ql/PrjY)[10.1002/prot.25423](http://dx.doi.org/10.1002/prot.25423)

[Adhikari, B., Bhattacharya, D., Cao, R., & Cheng, J. (2015). CONFOLD: Residue-residue contact-guided ab initio protein folding. *Proteins*, ***83***(8), 1436–1449. https://doi.org/](http://paperpile.com/b/tDy6ql/OOb7)[10.1002/prot.24829](http://dx.doi.org/10.1002/prot.24829)

[Brunger, A. T. (2013). CNS (Crystallography and NMR System). In *Encyclopedia of Biophysics* (pp. 326–327). https://doi.org/](http://paperpile.com/b/tDy6ql/8yxa)[10.1007/978-3-642-16712-6\_318](http://dx.doi.org/10.1007/978-3-642-16712-6_318)

[Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, ***7***(10), e1002195. https://doi.org/](http://paperpile.com/b/tDy6ql/bDhH)[10.1371/journal.pcbi.1002195](http://dx.doi.org/10.1371/journal.pcbi.1002195)

[Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., … Marks, D. S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, *3*. https://doi.org/](http://paperpile.com/b/tDy6ql/HLxX)[10.7554/eLife.03430](http://dx.doi.org/10.7554/eLife.03430)

[Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Moult, J., Schwede, T., & Tramontano, A. (2018). Evaluation of the template-based modeling in CASP12. *Proteins*, ***86*** *Suppl 1*, 321–334. https://doi.org/](http://paperpile.com/b/tDy6ql/20bh)[10.1002/prot.25425](http://dx.doi.org/10.1002/prot.25425)

[Lamb, J., Jarmolinska, A. I., Michel, M., Menéndez Hurtado, D. M., Sulkowska, J. I., & Elofsson, A. (n.d.). PconsFam: An interactive database of structure predictions of Pfam families. *JMB*.](http://paperpile.com/b/tDy6ql/H164)

[Lundström, J., Rychlewski, L., Bujnicki, J., & Elofsson, A. (2001). Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Science: A Publication of the Protein Society*, ***10***(11), 2354–2362. Retrieved from](http://paperpile.com/b/tDy6ql/OtAy) <https://www.ncbi.nlm.nih.gov/pubmed/11604541>

[McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* , *16*(4), 404–405. https://doi.org/](http://paperpile.com/b/tDy6ql/r2jd)[10.1093/bioinformatics/16.4.404](http://dx.doi.org/10.1093/bioinformatics/16.4.404)

[Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S., & Elofsson, A. (2014). PconsFold: improved contact predictions improve protein models. *Bioinformatics* , ***30***(17), i482–i488. https://doi.org/](http://paperpile.com/b/tDy6ql/h7YL)[10.1093/bioinformatics/btu458](http://dx.doi.org/10.1093/bioinformatics/btu458)

[Michel, M., Menéndez Hurtado, D., & Elofsson, A. (2018). PconsC4: fast, accurate, and hassle-free contact predictions. *Bioinformatics* . https://doi.org/](http://paperpile.com/b/tDy6ql/lJPn)[10.1093/bioinformatics/bty1036](http://dx.doi.org/10.1093/bioinformatics/bty1036)

[Michel, M., Menéndez Hurtado, D., Uziela, K., & Elofsson, A. (2017). Large-scale structure prediction by improved contact predictions and model quality assessment. *Bioinformatics* , ***33***(14), i23–i29. https://doi.org/](http://paperpile.com/b/tDy6ql/OpYz)[10.1093/bioinformatics/btx239](http://dx.doi.org/10.1093/bioinformatics/btx239)

[Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., … Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, ***355***(6322), 294–298. https://doi.org/](http://paperpile.com/b/tDy6ql/tMXz)[10.1126/science.aah4043](http://dx.doi.org/10.1126/science.aah4043)

[Pazos, F., Helmer-Citterich, M., Ausiello, G., & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology*, ***271***(4), 511–523. https://doi.org/](http://paperpile.com/b/tDy6ql/XxjY)[10.1006/jmbi.1997.1198](http://dx.doi.org/10.1006/jmbi.1997.1198)

[Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, ***9***(2), 173–175. https://doi.org/](http://paperpile.com/b/tDy6ql/akyR)[10.1038/nmeth.1818](http://dx.doi.org/10.1038/nmeth.1818)

[Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B., & Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics* , ***33***(10), 1578–1580. https://doi.org/](http://paperpile.com/b/tDy6ql/MyqD)[10.1093/bioinformatics/btw819](http://dx.doi.org/10.1093/bioinformatics/btw819)

[Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B., & Elofsson, A. (2018). Improved protein model quality assessments by changing the target function. *Proteins*, ***86***(6), 654–663. https://doi.org/](http://paperpile.com/b/tDy6ql/jYLL)[10.1002/prot.25492](http://dx.doi.org/10.1002/prot.25492)

[Zhou, A., Rohou, A., Schep, D. G., Bason, J. V., Montgomery, M. G., Walker, J. E., … Rubinstein, J. L. (2015). Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM. *eLife*, ***4***, e10180. https://doi.org/](http://paperpile.com/b/tDy6ql/cRsv)[10.7554/eLife.10180](http://dx.doi.org/10.7554/eLife.10180)