SMBE 2017

Computation and reproducibility in molecular evolution

POB-306

Influence of alignment uncertainty on homology and phylogenetic modeling

Jia-Ming Chang [1,*], Cedric Notredame [2]

[1]Computer Science, National Chengchi University, Taipei, Taiwan, [2]Bioinformatics and Genomics, Centre for Genomic Regulation (CRG), Barcelona, Spain **chang.jiaming@gmail.com*

Abstract: Most evolutionary analyses or structure modeling are based upon pre-estimated multiple sequence alignment (MSA) models. From a computational point of view, it is too complex to estimate a correct alignment. Hence, increasing or identifying signal inside sequence alignment has intensified over the last few years. During the presentation, I would like to share two approaches, homology extension and sampling, on this topic.

The first part, transmembrane proteins (TMPs) constitute about 20~30% of all protein coding genes. The relative lack of experimental structure has so far made it hard to develop specific alignment methods and the current state of the art (PRALINE™) only manages to recapitulate 50% of the positions in the reference alignments available from the BAliBASE2-ref7. We show how homology extension can be adapted and combined with a consistency based approach in order to significantly improve the multiple sequence alignment of alpha-helical TMPs. TM-Coffee is a special mode of PSI-Coffee able to efficiently align TMPs, while using a reduced reference database for homology extension. Our benchmarking on BAliBASE2-ref7 alpha-helical TMPs shows a significant improvement over the most accurate methods such as MSAProbs, Kalign, PROMALS, MAFFT, ProbCons and PRALINE™. We also estimated the influence of the database used for homology extension and show that highly non-redundant UniRef databases can be used to obtain similar results at a significantly reduced computational cost over full protein databases.

The second part, homology and evolutionary modeling are the most common applications of MSAs. Both are known to be sensitive to the underlying MSA accuracy. In this work, we show how this problem can be partly overcome using the transitive consistency score (TCS), an extended version of the T-Coffee scoring scheme. Using this local evaluation function, we show that one can identify the most reliable portions of an MSA, as judged from BAliBASE and PREFAB structure-based reference alignments. We also show how this measure can be used to improve phylogenetic tree reconstruction using both an established simulated data set and a novel empirical yeast data set. For this purpose, we describe a novel lossless alternative to site filtering that involves overweighting the trustworthy columns. Our approach relies on the T-Coffee framework; it uses libraries of pairwise alignments to evaluate any third party MSA. Pairwise projections can be produced using fast or slow methods, thus allowing a trade-off between speed and accuracy. We compared TCS with Heads-or-Tails, GUIDANCE, Gblocks, and trimAl and found it to lead to significantly better estimates of structural accuracy and more accurate phylogenetic trees.

References:

- PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic acids research* 44, W339–343(2016).
- TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction. *Nucleic acids research* 43, W3–6 (2015).
- TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction.*Molecular biology and evolution* 31, 1625–37 (2014).
- Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. *Bmc Bioinformatics* 13, S1 (2012).

Website:

• PSI/TM-Coffee	http://tcoffee.crg.cat/tmcoffee
• TCS	http://tcoffee.crg.cat/tcs

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3376v1 | CC BY 4.0 Open Access | rec: 27 Oct 2017, publ: 27 Oct 2017