Influence of alignment uncertainty on homology and phylogenetic modeling

Jia-Ming Chang¹, Cedric Notredame²

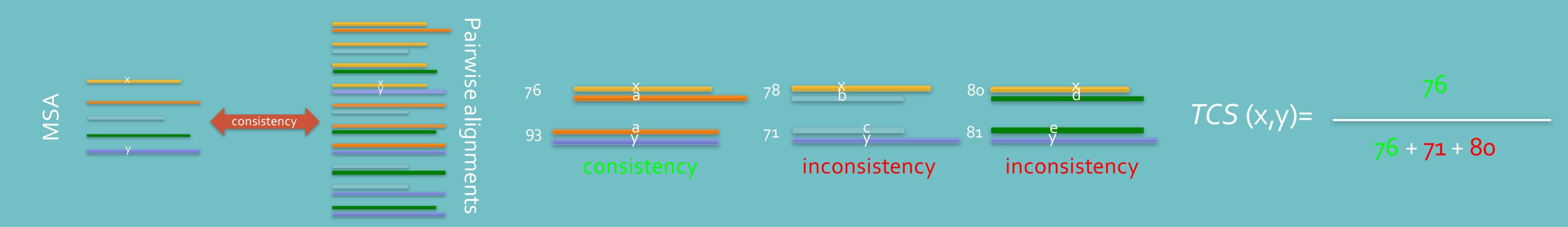
1.Computer Science, National Chengchi University, Taipei, Taiwan; chang.jiaming@gmail.com 2.Bioinformatics and Genomics,Centre for Genomic Regulation (CRG), Barcelona, Spain

Homology extension and sampling approch

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. I would like to share two approaches, homology extension and sampling, on this topic.

Transitive Consistency Score (TCS)

T-Coffee library

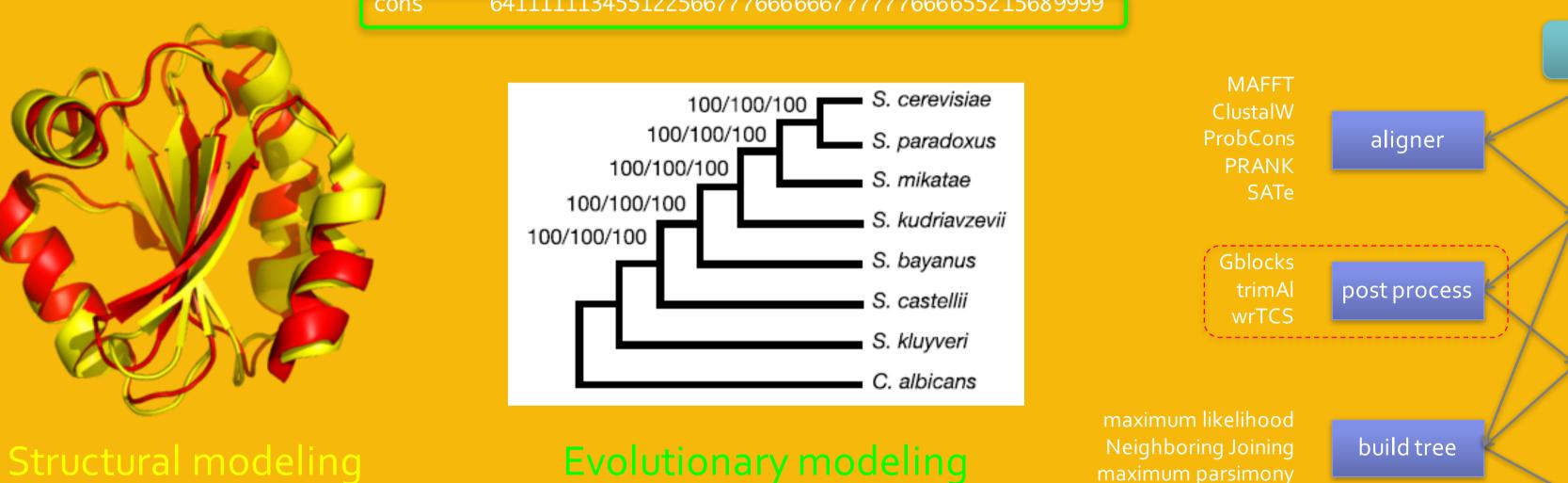


Transitive relation in alignment scene

TCS

Local evaluation function on homology and evolutionary modeling

| | reference alig | nment Guidence/TCS | Guidance = 71.10% | TCS = 83.5% |
|---|---|--|---|--|
| | | | (a) GUIDANCE | (b) TCS |
| esidue level row row TCS | T-COFFEE, Version_9.01 (2012-01-27 09:40:38) Cedric Notredame CPU TIME:0 sec. SCORE=76 Alignmentlevel | SP1 Seq1 SALMLWLSARESIKRENYPD confidence1 Seq2 SAYNIYVSFQRESAKD confidence1 Seqn SAYNIYVSAQRENAKD confidence2 Seq2 SAYNIYVSAQRENAKD confidence2 | A confidences Δ con | |
| $ \begin{array}{cccccccccccccccccccccccccccccccccccc$ | BAD AVG GOOD * 1j46_A : 74 2lef_A : 75 1k99_A : 77 | Seq <i>n</i> SAYNIYVSAQRENAKD SP1-SP2 ? confidence1-confidence2 | | |
| 3 4 0.693 | 1aab_ : 72 | | ∆ Sum–of–Pair Scores | ∆ Sum–of–Pair Scores |
| 1 3 0.562 1 4 0.632 3 4 0.526 | cons : 76 Column level 1j46_A 75456667777777777777777777666667789999 2lef_A 65666777777777777777777777777667789999 1k99_A 8654544456677777888878888888888778777789999 | | | MAFFT – MUSCLE MAFFT – ClustalW MUSCLE – ClustalW |
| | 1aab_ 76566533356667666666666666666666666 | Simulation • 16 tips | RF: average Robinson-Foulds distance respect to TPs: the number of genes whose tree topology is | |



TPs: the number of genes whose tree topology is identical with yeast ToL.

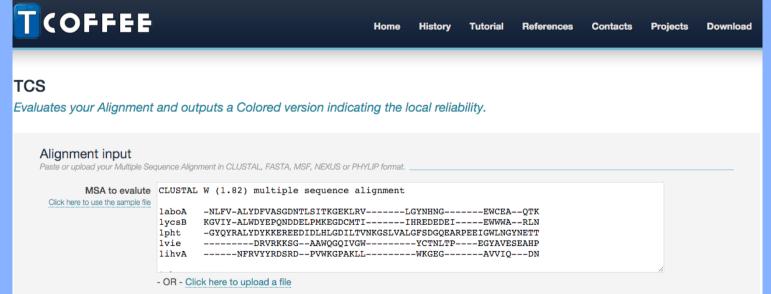
| | Orig | ginal | Gble rela | | | ocks ngent | | nAl oyout | | nAl tplus | TCS re | eplicate |
|----------|------|-------|--------------|-----|------|---------------|------|--------------|------|--------------|--------|----------|
| | RF | TPs | RF | TPs | RF | TPs | RF | TPs | RF | TPs | RF | TPs |
| ClustalW | 0.90 | 643 | 0.99 | 629 | 1.24 | 584 | 0.95 | 628 | 1.31 | 561 | 0.91 | 649 |
| MAFFT | 0.80 | 665 | 0.83 | 653 | 1.26 | 573 | 0.83 | 657 | 1.28 | 562 | 0.76 | 669 |
| Muscle | 0.95 | 639 | 0.91 | 646 | 1.26 | 578 | 0.96 | 633 | 1.29 | 559 | 0.84 | 662 |
| PRANK | 0.79 | 665 | 0.88 | 642 | 1.28 | 565 | 0.84 | 648 | 1.19 | 575 | 0.81 | 662 |
| SATe | 0.86 | 660 | 0.87 | 650 | 1.28 | 578 | 0.85 | 655 | 1.25 | 567 | 0.79 | 666 |
| AVE | 0.86 | 654 | 0.896 | 644 | 1.26 | 575 | 0.88 | 644 | 1.26 | 565 | 0.82 | 661 |

Web site

Col

We show that one can identify the most reliable portions TCS h of an MSA, as judged from BAliBASE and PREFAB TCOFFEE structure-based reference alignments. We also show how this measure can be used to improve phylogenetic tree reconstruction using both an established simulated

s TCS http://tcoffee.crg.cat/tcs



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

| SA multiple sequence alignment colored according to TCS scheme | | | TM-Coffee a | |
|--|---|--|---|--|
| If ic. Notredam NRE=373 ND AVG GOOD NOA SB 42 11 40 12 40 14 40 15 15 16 17 18 19 10 10 10 10 10 11 11 12 12 13 14 15 15 16 17 18 10 10 11 12 13 14 15 16 17 18 10 10 10 10 10 10 10 10 <th>on_11.00.d625267 (2016-01-11 15:25:41 - Revision d625267 - Build 507) e LYDFVASGDNTLSITKGEKLRVLGYNHNGEwCEAOTKNGOGWVPSNYITPVN LWDYEPONDDELPMKEGDCMTIIREDDEIEWWWARLNDKEGYVPRNLLGLYP LYDYKKEREEDIDLHGDITTVNKGSLVALGF5DGOEARPEEIGWLNGYNETTGERGDPFOTYVPYLGRKKISP DRVRKKSGAAWOGDIVGW</th> <th></th> <th>MSA The multiple sequent T-COFFEE, Ve Cedric Notree IN HEL OU Dmel_Or42b Dana_Or42b Dere_Or42b Dere_Or42b Dpri_Or42b Dper_Or42b Dper_Or42b Dper_Or42b</th> | on_11.00.d625267 (2016-01-11 15:25:41 - Revision d625267 - Build 507) e LYDFVASGDNTLSITKGEKLRVLGYNHNGEwCEAOTKNGOGWVPSNYITPVN LWDYEPONDDELPMKEGDCMTIIREDDEIEWWWARLNDKEGYVPRNLLGLYP LYDYKKEREEDIDLHGDITTVNKGSLVALGF5DGOEARPEEIGWLNGYNETTGERGDPFOTYVPYLGRKKISP DRVRKKSGAAWOGDIVGW | | MSA The multiple sequent T-COFFEE, Ve Cedric Notree IN HEL OU Dmel_Or42b Dana_Or42b Dere_Or42b Dere_Or42b Dpri_Or42b Dper_Or42b Dper_Or42b Dper_Or42b | |
| esult files | id them all | | Dsim_Or42b Dvir_Or42b Dwil_Or42b Dyak_Or42b Dyak_Or42b | |
| out(s) stem | Input sequences (448 B) Command line (297 B) Log file (36KB) | | Dmel_Or42b Dpse_Or42b Dana_Or42b | |

| ٦ | TM-Coffee ali | gnment result |
|---|--|--|
| | MSA The multiple sequence | alignment result as produced by T-coffee. |
| | T-COFFEE, Vers Cedric Notreda IN HEL OUT | ion_8.99(Thu Feb 17 19:24:49 CET 2011 - Revision 594) me |
| | Dmel_Or42b Dpse_Or42b Dana_Or42b Dere_Or42b Dmoj_Or42b Dmoj_Or42b Dper_Or42b Dsec_Or42b Dsim_Or42b Dvir_Or42b Dwir_Or42b Dyak_Or42b | 1 MVFELIRPAPLTE0KRSRDGCIYLYRAMKFIGWLPPK0GVLRYVYLTWTLMTFVWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEKVGSRDGCIYLFRAMKFIGWLPPKSGVLRYVYLFWTLMTFVWSTTYLPLGFLGS 67 1 MVFKLIRPAPLTEKWGSRDGCIYLFRAMKFIGWLPPKSGVLRYVYLFWTLMTFVWSTTYLPLGFLGS 67 1 MVFQLIRPAPLTEKWGSRDGCIYLFRAMKFIGWLPPKEGVLRYVYLFWTLMTFVWSTTYLPLGFLGS 67 1 MVFQLIRPAPLTEKVGSRDGCIYLYRAMKFIGWLPPKEGULRYVYLFWTLMTFVWSTTYLPLGFLGS 67 1 MVFQLIRPAPLTEQKRSRDGCIYLYRAMKFIGWLPPKEGLLRYVYLFWTLMTFVWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWPPKSGLLRYVYLTWTUTFFWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWPPKSGLRYVYLTWTUTFFWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWPPKSGVRYVYLWTWTUMTFVWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWPPKSGVRYVYLWTWTUMTFVWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWPPKSGVRYVLWTWTUMTFVWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWLPPK0GVLRYVYLWTUMTFVWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWLPPK0GVLRYVYLWTWTUMTFVWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWLPPK0GVLRYYVLWTWTUMTFVWCTTYLPLGFLGS 67 1 MVFKLIRPAPLTEQKRSRDGCIYLYRAMKFIGWLPPK0GVLRYYVLWTWTUMTFVWCTTYLPLGFLGS |
| | Dmel_Or42b Dpse_Or42b Dana_Or42b | 68 YMTQIKSFSPGEFLTSLQVCINAYGSSVKVAITYSMLWRLIKAKNILDQLDLRCTAMEEREKIHLVVARSNHAFLIFTFVVCGYAGSTYL 157 68 YMTQIKSFSPGEFLTSLQVCFNAYGSSVKTAITYSQLWRLIKAKDLLDKLDLRCTSVEEREKIHRVVALSNHAFLIFTCVYCTYAGSTYL 157 68 YMTQIKSFSPGEFLTSLQVCINAYGSSVKVAITYSMLWRLVKAKDLLDQLDLRCTSMEEREKIHRVVARSNHAFLIFTFVYCGYAGSTYL 157 |

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. We compared TCS with Heads-or-Tails,
- GUIDANCE, Gblocks, and trimAI and found it to lead to
- significantly better estimates of structural accuracy and more accurate phylogenetic trees.

| Note: TCS requires a multiple sequence alignment. If your sequences are *not* aligned, click the the following link to align them with <u>T-Coffee</u> . In the result page click the "Core/TCS" button to return to this page. | Multiple Alignment clustalw_aln file (313 B) fasta_aln file (173 B) phylip file (218 B) score_ascli file (777 B) tcs_weighted file (1KB) tcs_replicate file (50KB) score_html file (7KB) | Dere 0r42b 68 YMTQIKAFSPGEFLTSLQVCINAYGSSVKVAITYSMLWRLIKAKNILDQLDLR Dgri 0r42b 68 YITQIELFSPGEFLTSLQVCINAYGSSVKVAITYSQLWRLIKARELLDKLDVR Dmoj 0r42b 68 YMTQIKSFSPGEFLTSLQVCINAYGSSVKVAITYSQLWRLIKAKDILDKLDLR Dper 0r42b 68 YMTQIKSFSPGEFLTSLQVCINAYGSSVKVAITYSQLWRLIKAKDLLDKLDLR Dsec 0r42b 68 YMTQIKSFSPGEFLTSLQVCINAYGSSVKVAITYSQLWRLIKAKNILDQLDR |
|--|--|---|
| Hide advanced options | Steps to your Dropbox | Dsim [®] Or42b 91 YMTOIKSFSPGEFLTSLOVCINAYGSSVKVAJIYSMLWRLIKAKNILDQLDLR Dvir [®] Or42b 68 YMTMIKLFTPGGFLTSLOVCINAYGSSVKVAVIYTKLSRLIKAKDLLDKHDSR Dwil [®] Or42b 68 YMTQIKLFTPGEFLTSLOVCINAYGSSVKVAIIYSOLWRLIKAKDLLDKMDLR |
| Filter options Use these options to control the filtering intensity of your output alignment and whether you want the filtering to remove columns (recommended) or residues | Phylogenetic reconstruction | 91 *:* *: *:**:************************* |
| Filter o column () residue | Estimate a phylogenetic tree from selected MSAs using the PhyML web server | Dmel_0r42b 158 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYMVMSGAVLQDQLSDSYPLT Dpse_0r42b 158 SSVLSGRPPWQLYNPYIDWRDGRALWLASTLEYVMSGAVLQDQLSDTYPLT Dana 0r42b 158 SSVLSGRPPWQLYNPYIDWRDGRALWASTLEYVMSGAVLQDQLSDTYPLT Dere_0r42b 158 SSVLSGRPPWQLYNPFIDWHDGTAKLWWASTLEYLWSGAVLQDQLSDTYPLT Dere_0r42b 158 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYLMSGAVLQDQLSDTYPLT Dgri_0r42b 158 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYLMAGAVLKDQLSDTYPLT Dmoj_0r42b 158 SSVLSGHPPVQLYNPFIDWHDGLKLWTASTLEYLMSGAVLQDQLSDTYPLT Dmoj_0r42b 158 SSVLSGHPPVQLYNPFIDWHDGLKLWTASTLEYLMSGAVLQDQLSDTYPLT Dmoj_0r42b 158 SSVLSGHPVQLYNPTUWHDGLKLWASTLEYLMSGAVLQDQLSDTYPLT Dmoj_0r42b 158 SSVLSGRPWQLYNPTUWHDRTMLWASTLEYLWSGAVLQDUSDTYPLT Dper_0r42b 158 SSVLSGRPWQLYNPTUWHDRTMLWASTLEYVMSGAVLQDUSDTYPLT |
| Min 4 | Weighted PhyML Send Weighted MSA to PhyML | Dgri-Or42b 158 SSVLSGHPPYOLYNPFLDWHDGTLNMWIVSTLEYLIMAGAVLKDOLSDTYTLV Dmoj-Or42b 158 SSVLSGHPPWOLYNPFIDWHDGLLKLWTASTLEYIVMSGAVLDDQLSDTYPLV Dper-Or42b 158 SSVLSGRPPWOLYNPFIDWHDGRGGLWLASTLEYVVMSGAVLQDQLSDTYPLV Dsec_Or42b 158 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYMVMSGAVLQDQLSDTYPLT |
| Max 9 | Bootstrap PhyML Send Bootstrap MSA to PhyML | Dsec_Or42b 158 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYMVMSGAVLDQLSDTYPLT Dsim_Or42b 181 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYMVMSGAVLQDQLSDTYPLT Dvir_Or42b 181 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYMVMSGAVLQDQLSDTYPLT Dwil_Or42b 158 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYFVMSGAVLQDQLSDTYPLY Dyak_Or42b 158 SSVLSGRPPWQLYNPFIDWHDGTLKLWVASTLEYFWNSGAVLQDQLSDTYPLY 181 ;;****;***;*** ;***; ;***; ;***;;***; |
| Remove empty columns 🛛 yes | Send results | 181 |
| Output options | Forward this result to other online tools ProtoGene Turning amino acid alignments into bona fide CDS nucleotide alignments | |
| Use these options to control the output formats of your output alignment | MSA hub MyHits: a new interactive resource for protein annotation and domain identification | |
| Weighted ⊘ score_ascii ⊘ tcs_weighted ⊘ tcs_replicate | JalView Open this alignment in the Jalview viewer ESPript ESPript server renders sequence similarities and secondary structure information from aligned sequences | |
| Graphic | Info Some information about this alignment job | |
| Library Computation Your alignment is evaluated by comparison against a collection of alignments. This collection is named a Library. This section lets you control the computation of the library. Different libraries give different score results | Mode: TCS Request ID: 543e1865 Created at: 07 Nov 2016, 06:11 (CET) Elapsed time: 17 sec Expiration at: 17 Nov | |
| Pairwise Methods 🗹 Mproba_pair 💿 Mkalign_msa 💿 Mmafft_msa 💿 Mmuscle_msa 💿 Mlalign_id_pair 💿 Mclustalw_pair 💿 Mfast_pair 💿 Mslow_pair | Replay | |
| | Change some input parameters and resubmit this alignment <u>clicking here</u> . Are you a T-Coffee guru? You may want to use the full featured T-Coffee <u>command line options</u> . | |
| Your email (optional) | Feedback | |
| Submit Reset | Give us feedback about T-coffee web server | |
| T-Coffee Server is hosted by the Centre for Genomic Regulation (CRG) of Barcelona Back to top | G+1 80 Recommend You and 361 others recommend this. For suggestions, questions or any problem send an email to totoffee@googlegroups.com | |

• 32 tips

• 64 tips

Yeasts : 853

MSA

MSA

- 1. 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).
- 2. TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction. Nucleic acids research 43, W3–6 (2015).
- 3. 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).
- 4. Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics 13, S1 (2012).

