Influence of alignment uncertainty on homology and phylogenetic modeling

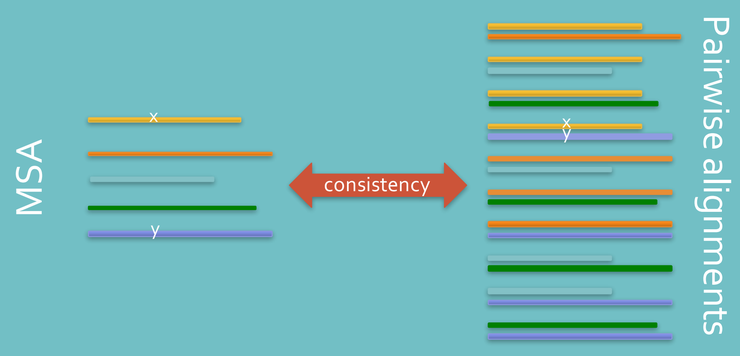
Jia-Ming Chang[1], Cedric Notredame[2]
1.Computer Science, National Chengchi University, Taipei, Taiwan; **chang.jiaming@gmail.com**
2.Bioinformatics and Genomics,Centre for Genomic Regulation (CRG), Barcelona, Spain

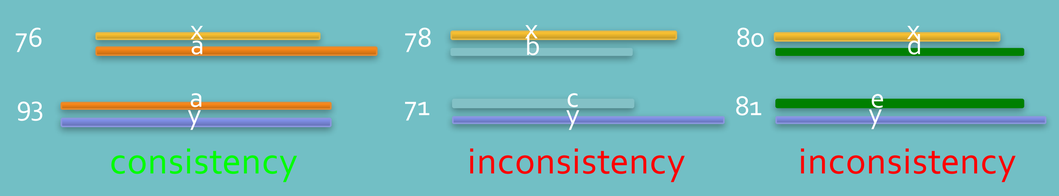
Homology extension and sampling approch

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





$$TCS(x,y) = \frac{76}{76 + 71 + 80}$$

Local evaluation function on homology and evolutionary modeling

Residue level

Col	row	row	TCS
1	1	2	0.762
1	1	3	0.748
1	1	4	0.741
1	2	3	0.651
1	2	4	0.677
1	3	4	0.693
2	1	3	0.562
2	1	4	0.632
2	3	4	0.526
…			

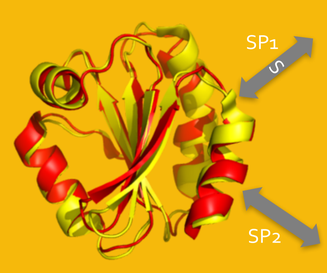
Alignment level

```
T-COFFEE, Version_9.01 (2012-01-27 09:40:38)
Cedric Notredame
CPU_TIME:0 sec.
SCORE=76
*
 BAD AVG GOOD
*
1j46_A    :  74
2lef_A    :  75
1k99_A    :  77
1qqb_     :  72
cons      :  76
```

Column level

```
1j46_A    75------4566---67777777777777777776666--7789999
2lef_A    6--------566---67777777777777777777766--7789999
1k99_A    865454445667---77778888788888888877877--7789999
1qqb_     76------566533356667666666666666666655336789999
cons      64111111345512256677766666677777766665521568999 9
```

reference alignment

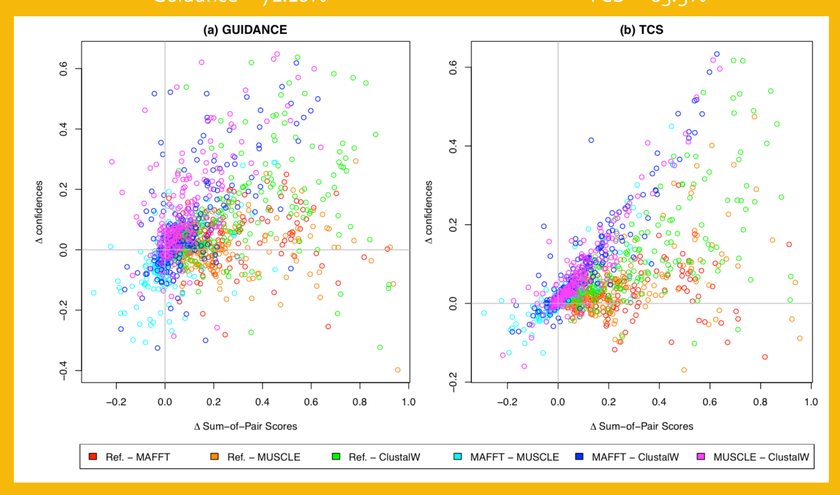


Guidence/TCS

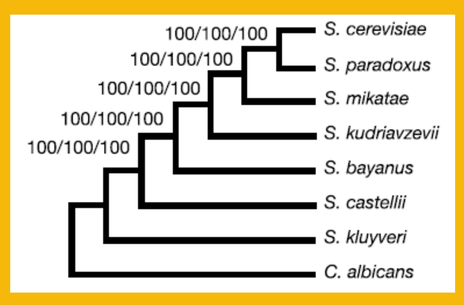
```
Seq1 …SALMLWLSARESIKREN…YPD…
Seq2 …SAYNIYVSFQ----RESA…KD…   confidence1
…
Seqn …SAYNIYVSAQ----RENA…KD…

Seq1 …SALMLWLSARESIKREN…YPD…
Seq2 …SAYNIYVSF----QRESA…KD…   confidence2
…
Seqn …SAYNIYVSA----QRENA…KD…
```

SP1 – SP2 **?** confidence1 – confidence2

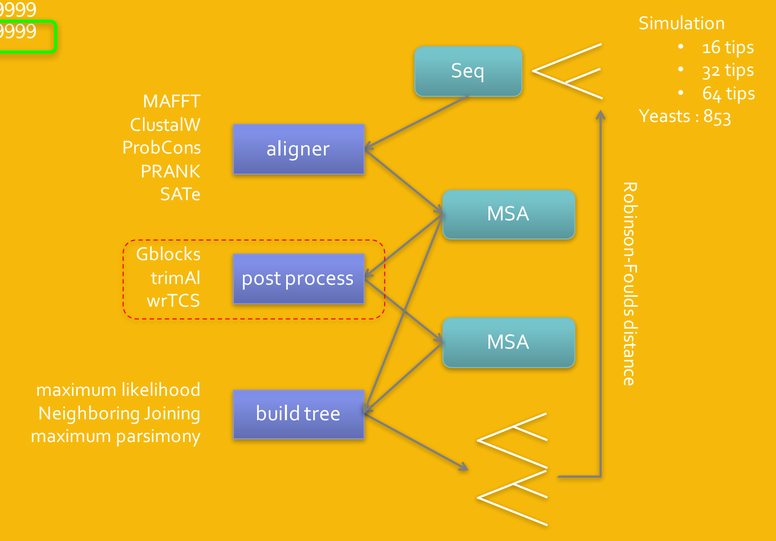
Guidance = 71.10% | TCS = 83.5%



Structural modeling



Evolutionary modeling



RF: average Robinson-Foulds distance respect to Yeast ToL.
TPs: the number of genes whose tree topology is identical with yeast ToL.

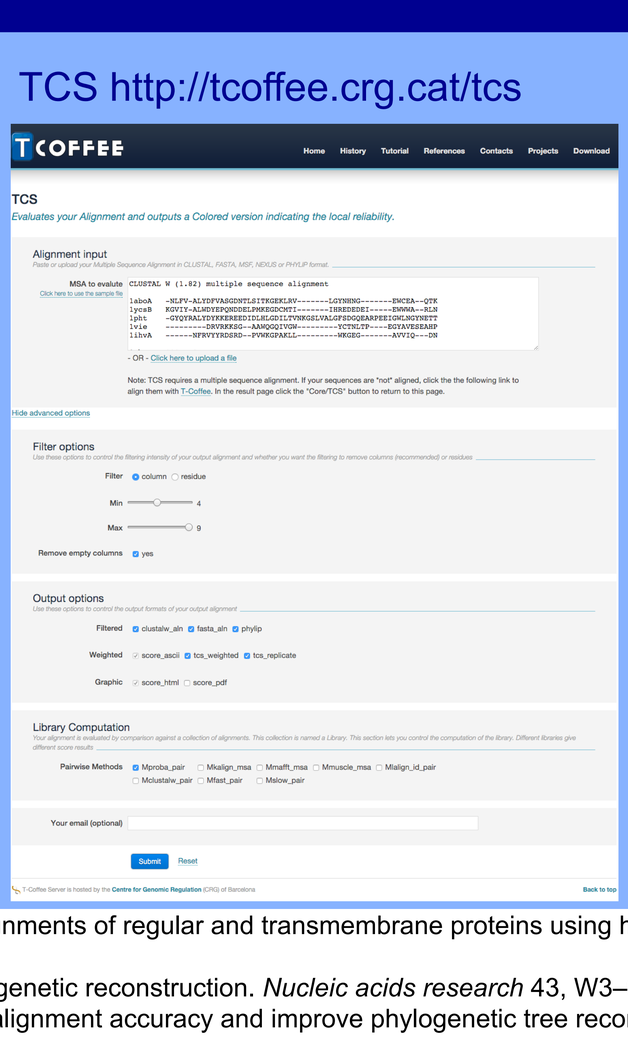
	Original		Gblocks relaxed		Gblocks stringent		trimAl gappyout		trimAl strictplus		TCS replicate	
	RF	TPs	RF	TPs	RF	TPs	RF	TPs	RF	TPs	RF	TPs
ClustalW	**0.90**	643	0.99	629	1.24	584	0.95	628	1.31	561	0.91	**649**
MAFFT	0.80	665	0.83	653	1.26	573	0.83	657	1.28	562	**0.76**	**669**
Muscle	0.95	639	0.91	646	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**

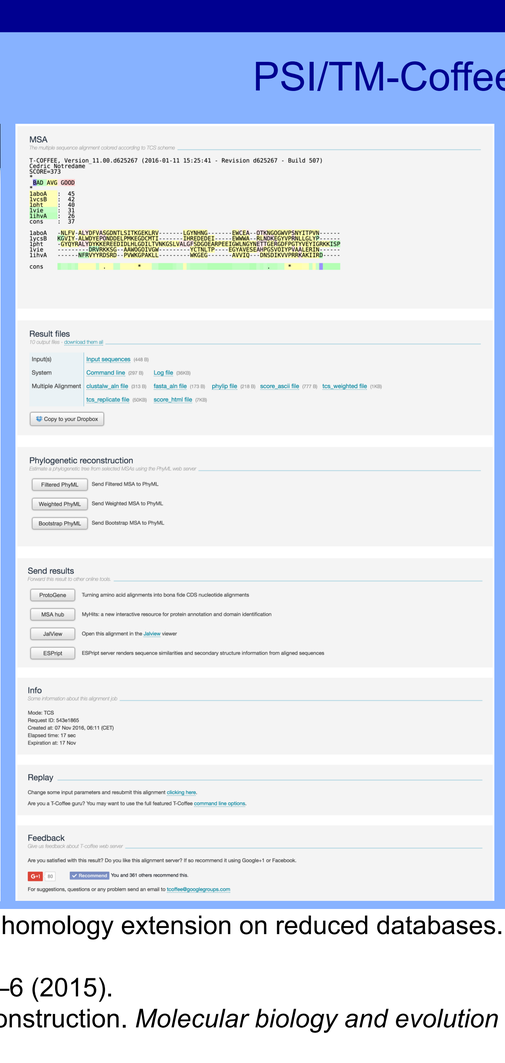
Web site

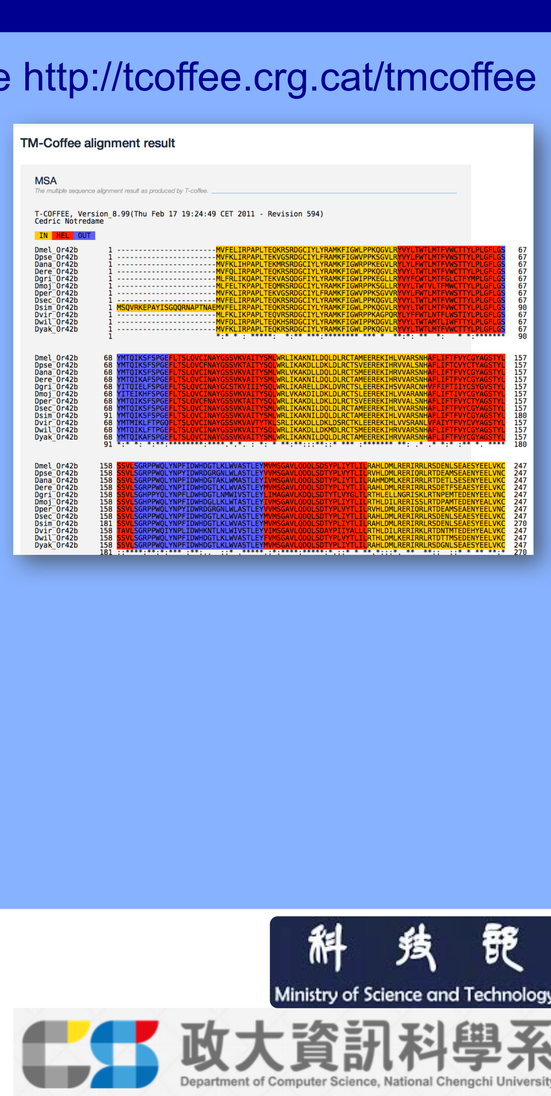
We show that one can identify the most reliable portions of an MSA, as judged from BAliBASE and PREFAB structure-based reference alignments. We also show how this measure can be used to improve phylogenetic tree reconstruction using both an established simulated data set and a novel empirical yeast data set. For this purpose, we describe a novel lossless alternative to site filtering that involves overweighting the trustworthy columns. 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.

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

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







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).

科技部 Ministry of Science and Technology
政大資訊科學系 Department of Computer Science, National Chengchi University