M-Coffee: combining multiple sequence alignment methods with T-Coffee

Iain M. Wallace, Orla O’Sullivan, Desmond G. Higgins and Cedric Notredame

Presented by: Jenny Shen
Mentor: Dr. Hugh Nicholas

Objective

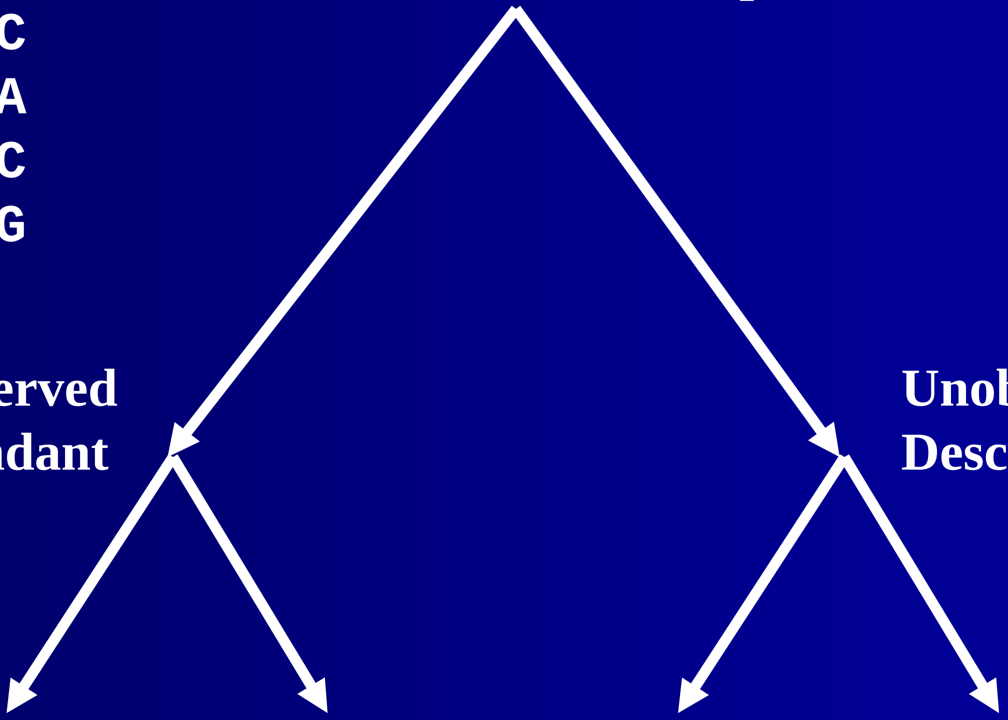
- Using today's sequences to reconstruct the appearance of ancestral sequences – our best guess
- Analogy: In the absence of fossils, using only present day data to construct the appearance of dinosaurs

Background

- Multiple sequence alignments
  - Predict phylogeny
  - Determine structure and function
  - Detect homologues
- Alignment methods
  - Biologically and computationally complex
  - Many programs available
  - Scientist’s dilemma: select best method or combination of methods to produce the most biologically correct alignment

Multiple Sequence Alignments

```
AGCATGATGCGC
AGCCTCATCTCA
AGCCTG...CGC
ACT...ACATTG
```



Unobserved Ancestral Sequence

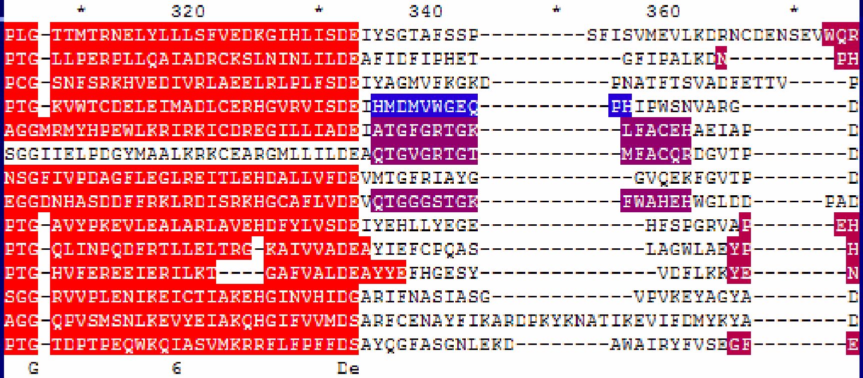
Unobserved Descendant

Unobserved Descendant

AGCATGATGCGC   AGCCTGCGC   ACTACATTG   AGCCTCATCTCA

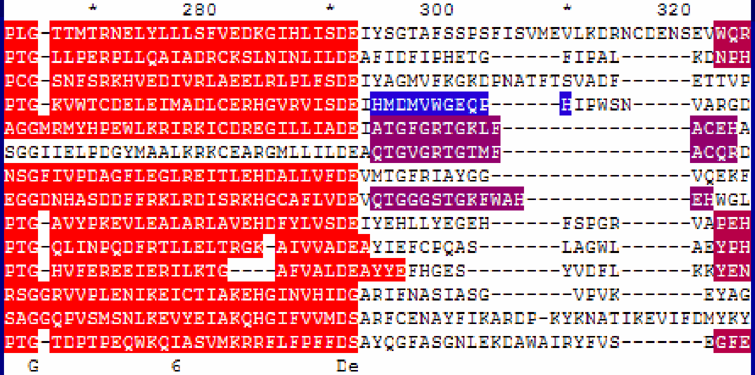
MSA Algorithms

Balibase*



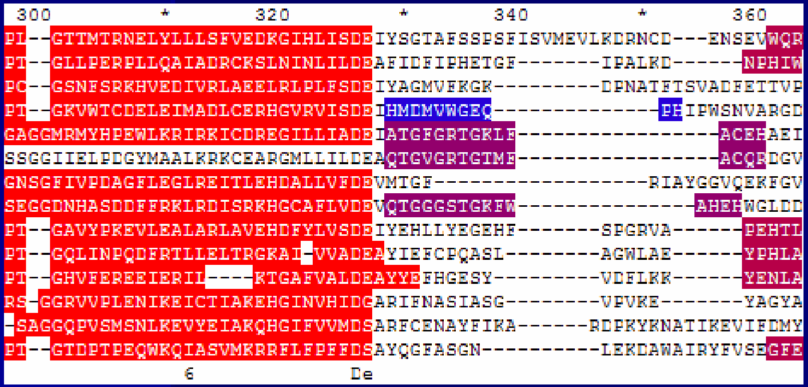
```
      *       320         *       340         *       360         *
PLG-TTMTRNELYLLLSFVEDKGIHLISDEIYSGTAFSSP---------SFISVMEVLKDRNCDENSEVWQR
PTG-LLPERPLLQAIADRCKSLNINLILDEAFIDFIPHET------------GFIPALKDN---------PH
PCG-SNFSRKHVEDIVRLAEELRLPLFSDEIYAGMVFKGKD----------PNATFTSVADFETTV-----P
PTG-KVWTCDELEIMADLCERHGVRVISDEIHMDMVWGEQ-----------PHIPWSNVARG---------D
AGGMRMYHPEWLKRIRKICDREGILLIADEIATGFGRTGK------------LFACEHAEIAP--------D
SGGIIELPDGYMAALKRKCEARGMLIILDEAQTGVGRTGT------------MFACQRDGVTP--------D
NSGFIVPDAGFLEGLREITLEHDALLVFDEVMTGFRIAYG-------------GVQEKFGVTP--------D
EGGDNHASDDFFRKLRDISRKHGCAFLVDEVQTGGGSTGK------------FWAHEHWGLDD------PAD
PTG-AVYPKEVLEALARLAVEHDFYLVSDEIYEHLLYEGE--------------HFSPGRVAP-------EH
PTG-QLINPQDFRTLLELTRG-KAIVVADEAYIEFCPQAS--------------LAGWLAEYP--------H
PTG-HVFEREEIERILKT----GAFVALDEAYYEFHGESY---------------VDFLKKYE--------N
SGG-RVVPLENIKEICTIAKEHGINVHIDGARIFNASIASG------------VPVKEYAGYA--------D
AGG-QPVSMSNLKEVYEIAKQHGIFVVMDSARFCENAYFIKARDPKYKNATIKEVIFDMYKYA--------D
PTG-TDPTPEQWKQIASVMKRRFLFPFFDSAYQGFASGNLEKD--------AWAIRYFVSEGF--------E
  G           6             De
```

ClustalW



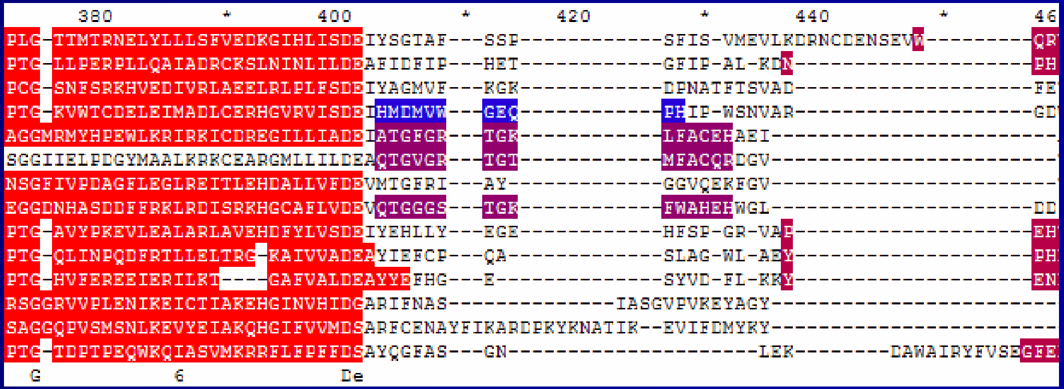
```
       *       280         *       300         *       320
PLG-TTMTRNELYLLLSFVEDKGIHLISDEIYSGTAFSSPSFISVMEVLKDRNCDENSEVWQR
PTG-LLPERPLLQAIADRCKSLNINLILDEAFIDFIPHETG------FIPAL------KDNPH
PCG-SNFSRKHVEDIVRLAEELRLPLFSDEIYAGMVFKGKDPNATFTSVADF------ETTVP
PTG-KVWTCDELEIMADLCERHGVRVISDEIHMDMVWGEQP------HIPWSN-----VARGD
AGGMRMYHPEWLKRIRKICDREGILLIADEIATGFGRTGKLF----------------ACEHA
SGGIIELPDGYMAALKRKCEARGMLIILDEAQTGVGRTGTMF----------------ACQRD
NSGFIVPDAGFLEGLREITLEHDALLVFDEVMTGFRIAYGG-----------------VQEKF
EGGDNHASDDFFRKLRDISRKHGCAFLVDEVQTGGGSTGKFWAH--------------EHWGL
PTG-AVYPKEVLEALARLAVEHDFYLVSDEIYEHLLYEGEH------FSPGR------VAPEH
PTG-QLINPQDFRTLLELTRGK-AIVVADEAYIEFCPQAS-------LAGWL------AEYPH
PTG-HVFEREEIERILKTG----AFVALDEAYYEFHGES--------YVDFL------KKYEN
RSGGRVVPLENIKEICTIAKEHGINVHIDGARIFNASIASG-------VPVK-------EYAG
SAGGQPVSMSNLKEVYEIAKQHGIFVVMDSARFCENAYFIKARDP-KYKNATIKEVIFDMYKY
PTG-TDPTPEQWKQIASVMKRRFLFPFFDSAYQGFASGNLEKDAWAIRYFVS-------EGFE
  G           6              De
```

T-Coffee



```
 300         *       320         *       340         *       360
PL--GTTMTRNELYLLLSFVEDKGIHLISDEIYSGTAFSSPSFISVMEVLKDRNCD---ENSEVWQR
PT--GLLPERPLLQAIADRCKSLNINLILDEAFIDFIPHETGF-------IPALKD------NPHIW
PC--GSNFSRKHVEDIVRLAEELRLPLFSDEIYAGMVFKGK---------DPNATFTSVADFETTVP
PT--GKVWTCDELEIMADLCERHGVRVISDEIHMDMVWGEQ--------------PHIPWSNVARGD
GAGGMRMYHPEWLKRIRKICDREGILLIADEIATGFGRTGKLF-----------------ACEHAEI
SSGGIIELPDGYMAALKRKCEARGMLIILDEAQTGVGRTGTMF-----------------ACQRDGV
GNSGFIVPDAGFLEGLREITLEHDALLVFDEVMTGF------------------RIAYGGVQEKFGV
SEGGDNHASDDFFRKLRDISRKHGCAFLVDEVQTGGGSTGKFW---------------AHEHWGLDD
PT--GAVYPKEVLEALARLAVEHDFYLVSDEIYEHLLYEGEHF-------SPGRVA------PEHTL
PT--GQLINPQDFRTLLELTRGKAI-VVADEAYIEFCPQASL--------AGWLAE------YPHLA
PT--GHVFEREEIERIL----KTGAFVALDEAYYEFHGESY---------VDFLKK------YENLA
RS-GGRVVPLENIKEICTIAKEHGINVHIDGARIFNASIASG--------VPVKE-------YAGYA
-SAGGQPVSMSNLKEVYEIAKQHGIFVVMDSARFCENAYFIKA------RDPKYKNATIKEVIFDMY
PT--GTDPTPEQWKQIASVMKRRFLFPFFDSAYQGFASGN----------LEKDAWAIRYFVSEGFE
               6             De
```

ProbCons



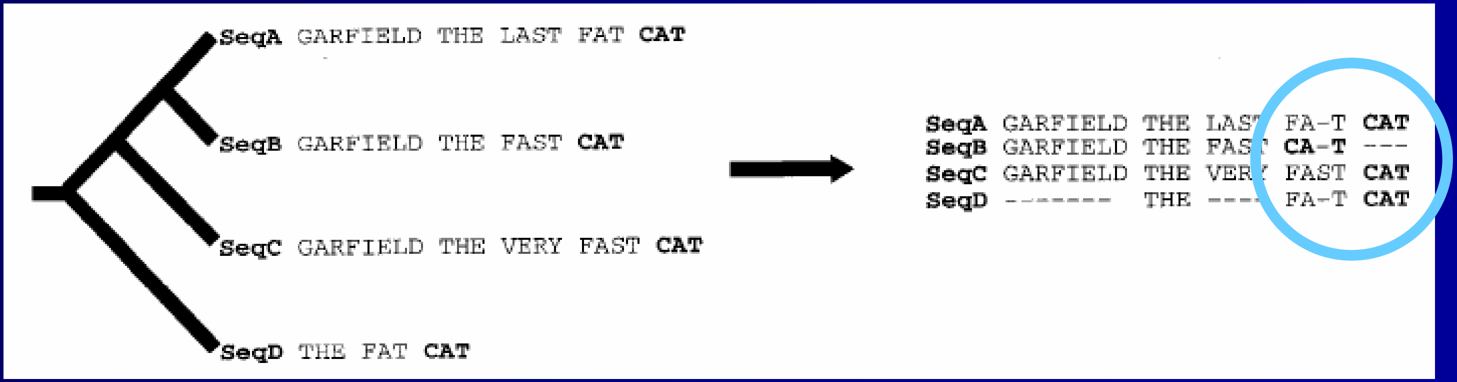
```
      380         *       400         *       420         *       440         *       46
PLG-TTMTRNELYLLLSFVEDKGIHLISDEIYSGTAF---SSP------------SFIS-VMEVLKDRNCDENSEVW---------QR
PTG-LLPERPLLQAIADRCKSLNINLILDEAFIDFIP---HET------------GFIP-AL-KDN--------------------PH
PCG-SNFSRKHVEDIVRLAEELRLPLFSDEIYAGMVF---KGK------------DPNATFTSVAD--------------------FE
PTG-KVWTCDELEIMADLCERHGVRVISDEIHMDMVW---GEQ------------PHIP-WSNVAR--------------------GD
AGGMRMYHPEWLKRIRKICDREGILLIADEIATGFGR---TGK------------LFACEHAEI------------------------
SGGIIELPDGYMAALKRKCEARGMLIILDEAQTGVGR---TGT------------MFACQRDGV------------------------
NSGFIVPDAGFLEGLREITLEHDALLVFDEVMTGFRI---AY-------------GGVQEKFGV------------------------
EGGDNHASDDFFRKLRDISRKHGCAFLVDEVQTGGGS---TGK------------FWAHEHWGL----------------------DD
PTG-AVYPKEVLEALARLAVEHDFYLVSDEIYEHLLY---EGE------------HFSP-GR-VAP--------------------EH
PTG-QLINPQDFRTLLELTRG-KAIVVADEAYIEFCP---QA-------------SLAG-WL-AEY--------------------PH
PTG-HVFEREEIERILKT----GAFVALDEAYYEFHG---E--------------SYVD-FL-KKY--------------------EN
RSGGRVVPLENIKEICTIAKEHGINVHIDGARIFNAS--------------IASGVPVKEYAGY------------------------
SAGGQPVSMSNLKEVYEIAKQHGIFVVMDSARFCENAYFIKARDPKYKNATIK--EVIFDMYKY------------------------
PTG-TDPTPEQWKQIASVMKRRFLFPFFDSAYQGFAS---GN---------------------LEK--------DAWAIRYFVSEGFE
  G           6              De
```

MEME-determined motifs are highlighted in each alignment

Progressive Alignment

- Simultaneous alignment of all sequences is impractical
- Pairwise progressive method
  - Progressively aligns the most similar sequences and successively adds on more
  - Attempts to obtain the best score at every step in the alignment
  - However, optimization is only a local max
  - May not achieve the best overall alignment
  - Propagation of error

Progressive Alignment



```
SeqA GARFIELD THE LAST FAT CAT

SeqB GARFIELD THE FAST CAT

SeqC GARFIELD THE VERY FAST CAT

SeqD THE FAT CAT
```

```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD -------- THE ---- FA-T CAT
```

- The tree indicates the order in which the sequences are aligned
- The world “CAT” is misaligned

Consistency

- Attempt to avoid error by including “look ahead” information in scores
- Considers alignments between all sequence pairs, whether or not they have already been aligned, in each step of progressive alignment
- Correct alignments are more likely to be consistent

T-Coffee methodology

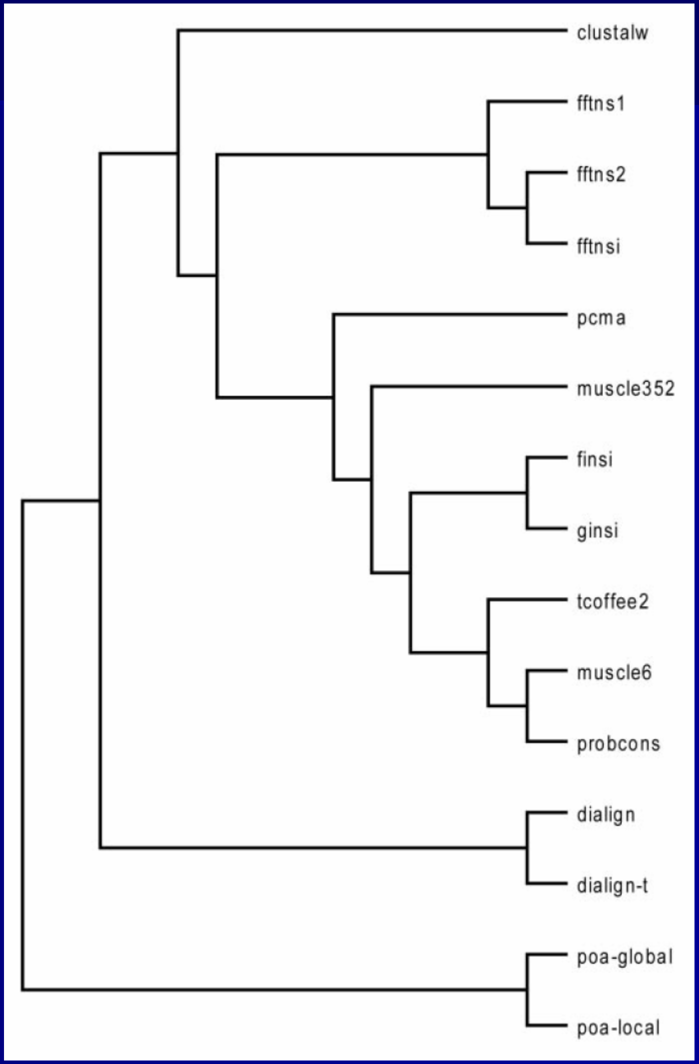
- Consistency-based method
- Pool together ClustalW (global) and Lalign (local) primary libraries
  - Combine information on global and local alignments
  - Generate primary library of alignment information
- Compute MSA from primary library of pairwise alignment information

M-Coffee methodology

- T-Coffee extended to 15 widely used alternative MSA programs from 8 different laboratories
- MSA libraries computed for the same sequences from a variety of algorithms:
  - ClustalW, T-Coffee, ProbCons, PCMA, Muscle, Dialign2, Dialign-T, MAFFT, FFT-NS1, FFT-NS2, FFT-NSI, F-INSI, G-INSI, POA-local, POA-global
- Results of this study compared to reference alignments of benchmark datasets – BaliBase, Prefab, HOMSTRAD

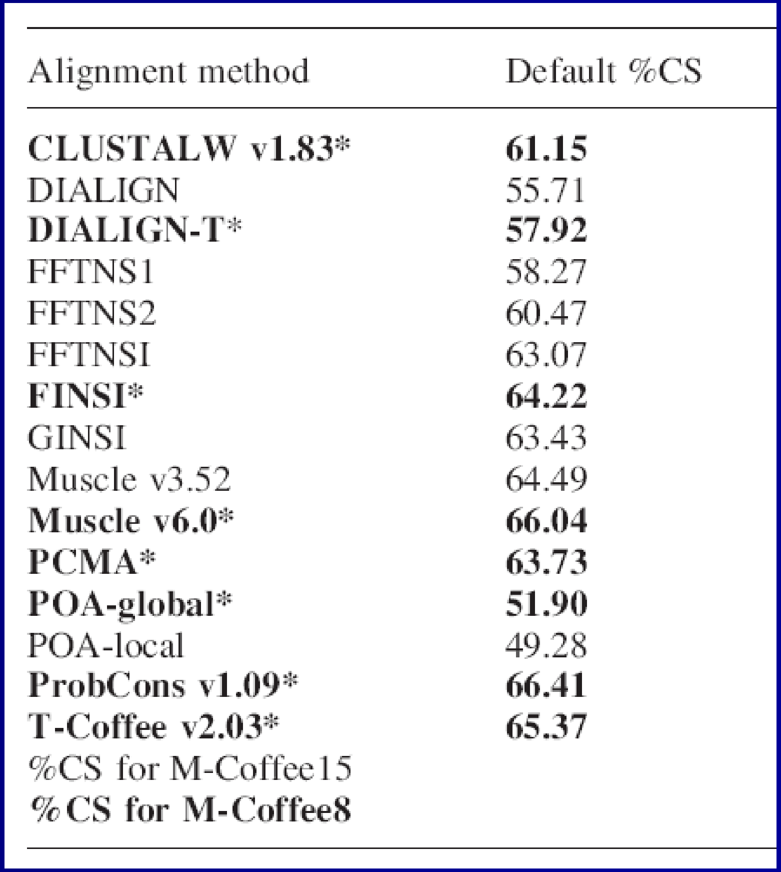
Method Tree

- Visual display of level of similarity between the various methods
- Entire HOMSTRAD dataset aligned with each method
- Distances calculated based on similarity of resulting alignments



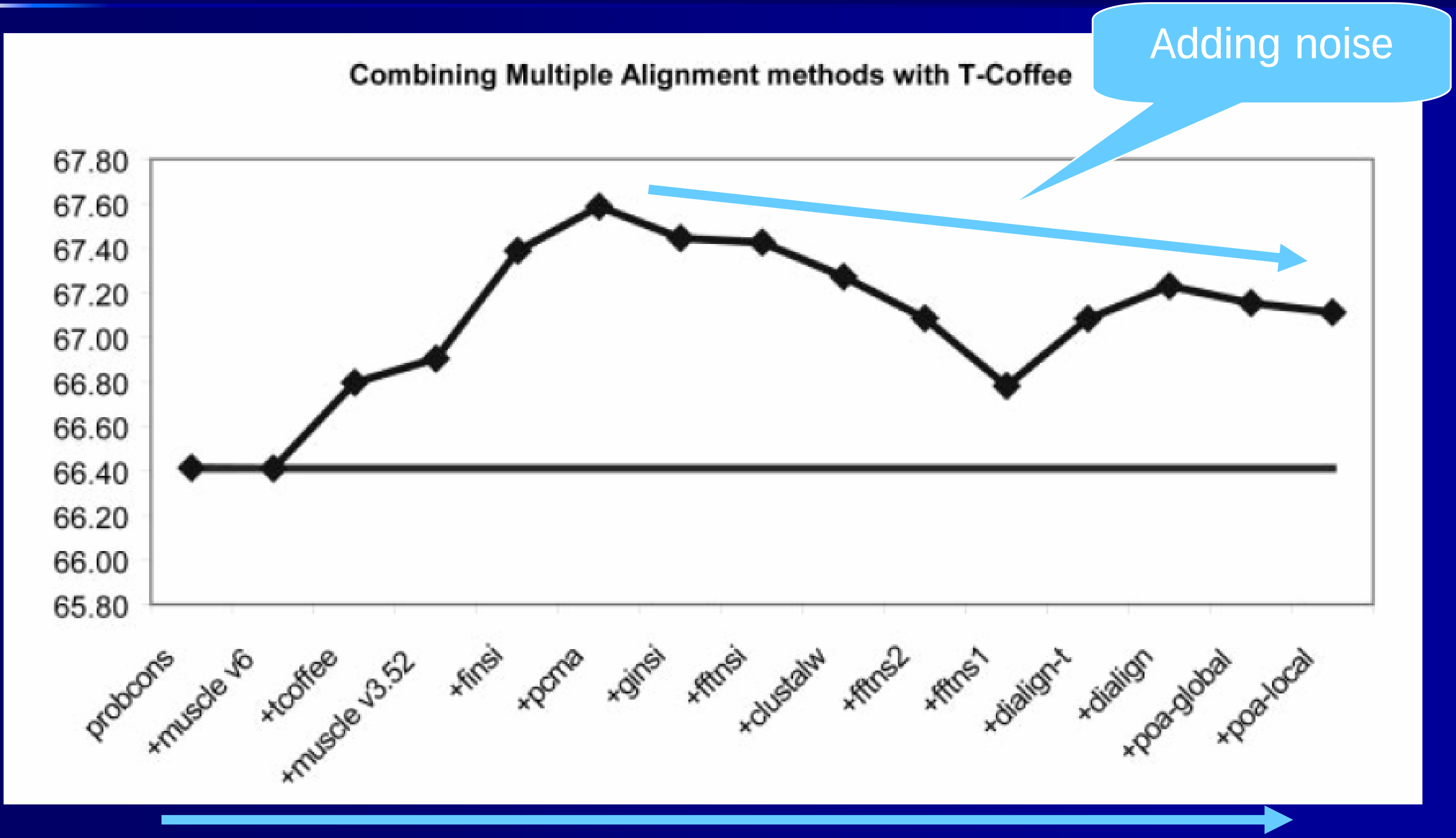
Preliminary Data

- Calculate performance of each individual method in comparison to HOMSTRAD
- Percent of pairwise alignments that are correct



Alignment method	Default %CS
CLUSTALW v1.83*	**61.15**
DIALIGN	55.71
DIALIGN-T*	**57.92**
FFTNS1	58.27
FFTNS2	60.47
FFTNSI	63.07
FINSI*	**64.22**
GINSI	63.43
Muscle v3.52	64.49
Muscle v6.0*	**66.04**
PCMA*	**63.73**
POA-global*	**51.90**
POA-local	49.28
ProbCons v1.09*	**66.41**
T-Coffee v2.03*	**65.37**
%CS for M-Coffee15	
%CS for M-Coffee8	

Combining MSA Methods



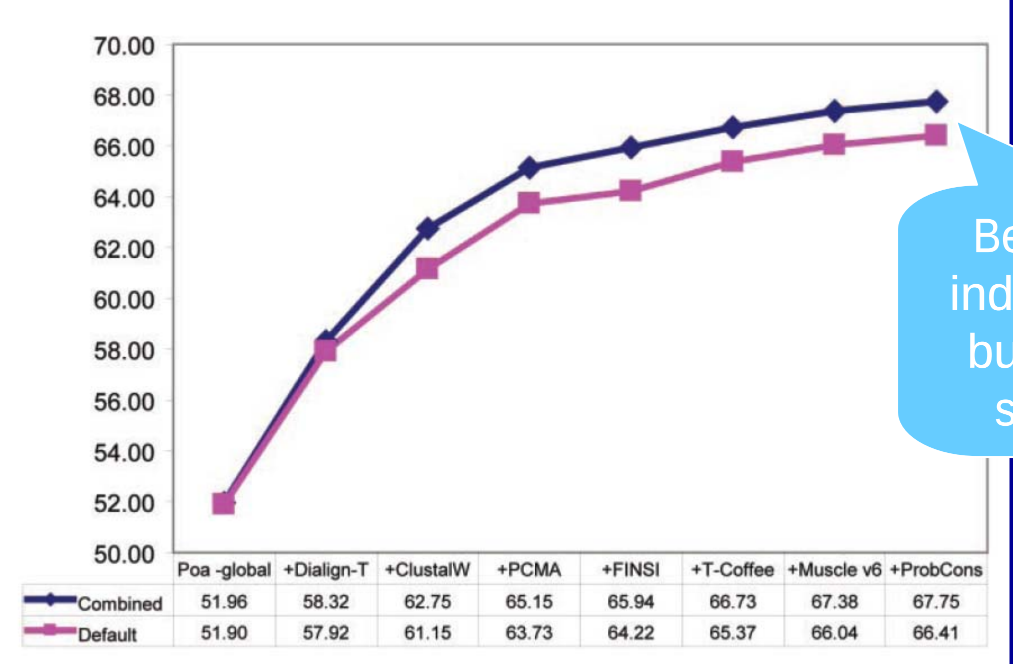
Order of addition

Method Weighing

- Methods developed by the same laboratory tend to be highly correlated because of arbitrary code settings
- Four different schemes used to generate weights for each of the alignment methods
  - Variance/Covariance, Altschul Carrillo Lipman, Thompson Higgins Gibson, and Accuracy
- Results:
  - Weighing failed to significantly outperform un-weighed combination of all methods
  - One method per developer for most accurate results – eight methods selected, called M-Coffee8

M-Coffee8

- Outperforms any of the constituent methods



	Poa -global	+Dialign-T	+ClustalW	+PCMA	+FINSI	+T-Coffee	+Muscle v6	+ProbCons
Combined	51.96	58.32	62.75	65.15	65.94	66.73	67.38	67.75
Default	51.90	57.92	61.15	63.73	64.22	65.37	66.04	66.41

Better than any individual method but by relatively small amount

Addition in order of increasing performance

M-Coffee8

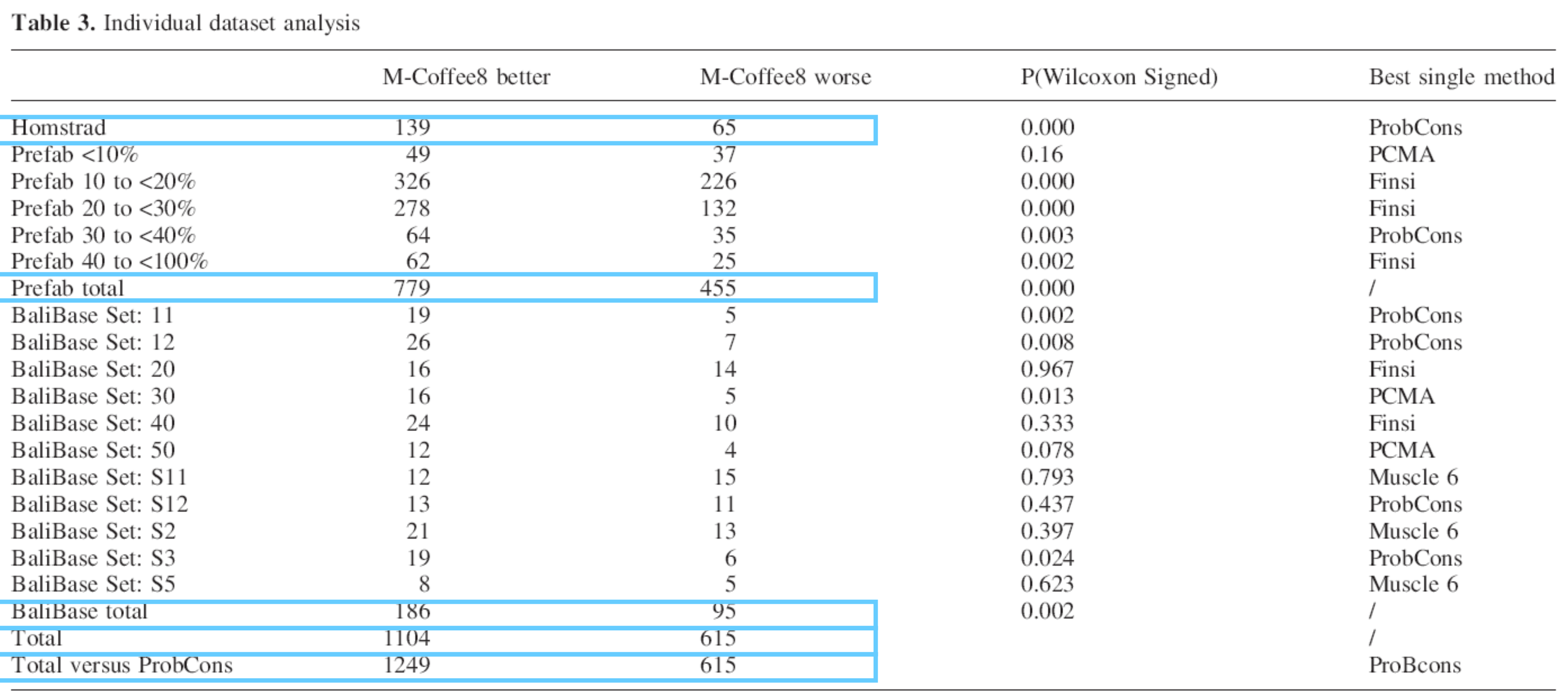


Table 3. Individual dataset analysis

	M-Coffee8 better	M-Coffee8 worse	P(Wilcoxon Signed)	Best single method
Homstrad	139	65	0.000	ProbCons
Prefab <10%	49	37	0.16	PCMA
Prefab 10 to <20%	326	226	0.000	Finsi
Prefab 20 to <30%	278	132	0.000	Finsi
Prefab 30 to <40%	64	35	0.003	ProbCons
Prefab 40 to <100%	62	25	0.002	Finsi
Prefab total	779	455	0.000	/
BaliBase Set: 11	19	5	0.002	ProbCons
BaliBase Set: 12	26	7	0.008	ProbCons
BaliBase Set: 20	16	14	0.967	Finsi
BaliBase Set: 30	16	5	0.013	PCMA
BaliBase Set: 40	24	10	0.333	Finsi
BaliBase Set: 50	12	4	0.078	PCMA
BaliBase Set: S11	12	15	0.793	Muscle 6
BaliBase Set: S12	13	11	0.437	ProbCons
BaliBase Set: S2	21	13	0.397	Muscle 6
BaliBase Set: S3	19	6	0.024	ProbCons
BaliBase Set: S5	8	5	0.623	Muscle 6
BaliBase total	186	95	0.002	/
Total	1104	615		/
Total versus ProbCons	1249	615		ProBcons

Conclusions

- M-Coffee alignments
  - On average 1-3% more accurate than that obtained from best individual method
  - Nearly twice as likely to deliver best MSA
- ProbCons usually the best individual method
- Caveat: Because generating MSAs libraries is very time-consuming, the gain is not always worth the time invested; may be better off using Probcons

References

- Lassmann, T., and Sonnhammer, E.L.L. (2005) Automatic assessment of alignment quality. *Nucleic Acids Research* 33: 7120-7128.
- Notredame, C., Higgins, D.G., and Heringa, J. (2000) T-Coffee: A novel method for multiple sequence alignments." *J. Mol. Bio.* 302: 205-217.
- Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee . *Nucleic Acids Research* 34: 1692-1699.

Questions?

"Mr. Osborne, may I be excused? My brain is full."