Reviews in Computational Biology

Assessing the quality of sequence alignments

Christophe Dessimoz

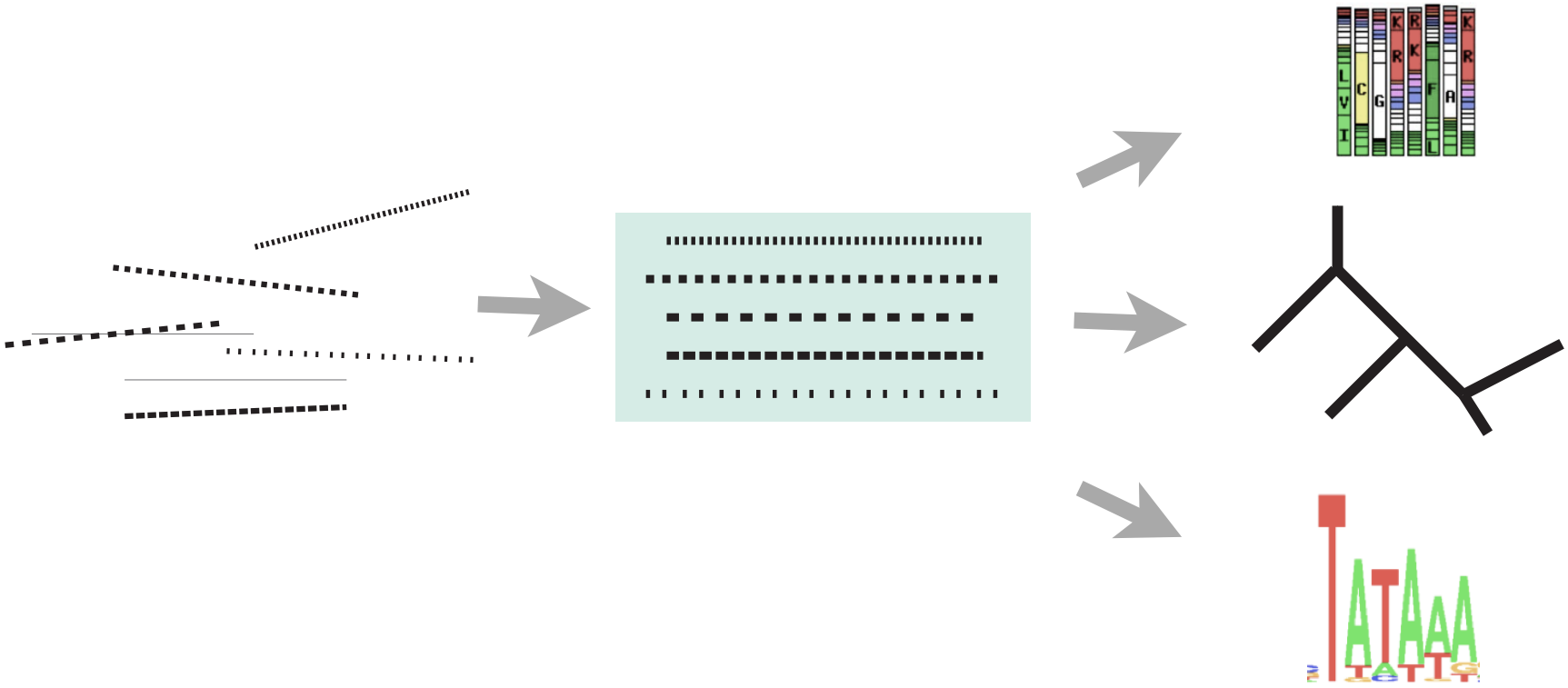
January 2012

“There is nothing so useless as doing efficiently that which should not be done at all.”

Peter F. Druker

Goal of Sequence Alignment

identify homologous residues,
(i.e. residues that share common ancestry)



Recent reviews

BIOINFORMATICS REVIEW Vol. 25 no. 19 2009, pages 2455–2465
doi:10.1093/bioinformatics/btp452

Sequence analysis

Upcoming challenges for multiple sequence alignment methods in the high-throughput era

Carsten Kemena and Cedric Notredame*

Published online 17 July 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 21 7353–7363*
doi:10.1093/nar/gkq625

SURVEY AND SUMMARY

Issues in bioinformatics benchmarking: the case study of multiple sequence alignment

Mohamed Radhouene Aniba[1,2,3,4], Olivier Poch[1,2,3,4] and Julie D. Thompson[1,2,3,4,*]

In this review, I will show that....

despite recent progress, current approaches to alignment benchmarking all have considerable shortcomings

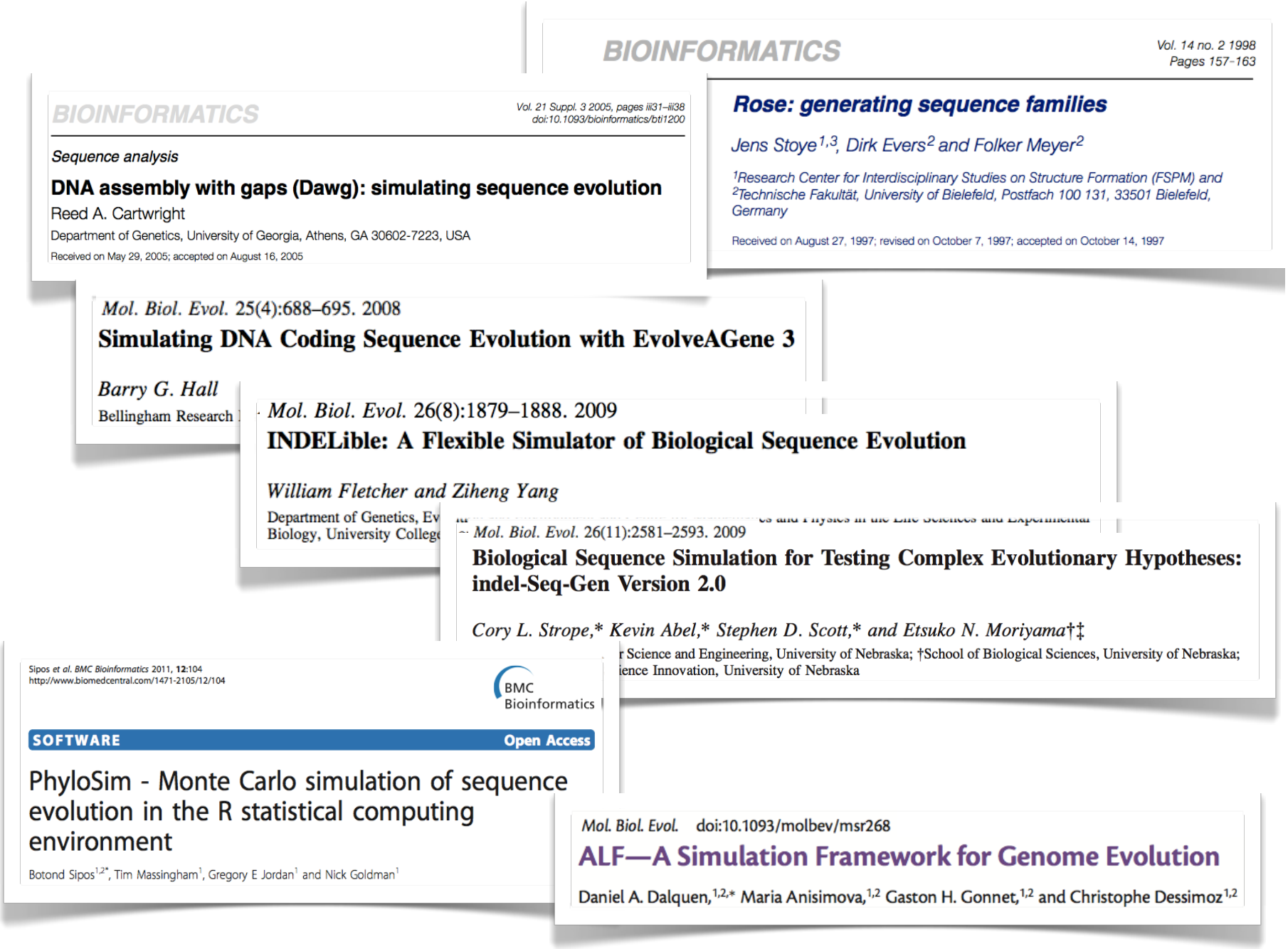
... thus impeding progress in alignment methods.

Outline

- **Survey of benchmark approaches (and their limitations)**
  - Simulation
  - Consistency
  - Expert review (human appraisal)
  - Empirical indicators
- **Reconciling the various approaches**

1. Simulation

- Start with a random sequence
- Evolve along a tree introducing random insertions, deletions, and mutations
- Doing so, keep track of homology relations (“true MSA”)
- Align resulting sequences and compare them to the true MSA



BIOINFORMATICS

Vol. 14 no. 2 1998
Pages 157–163

Rose: generating sequence families

Jens Stoye[1,3], Dirk Evers[2] and Folker Meyer[2]

[1]Research Center for Interdisciplinary Studies on Structure Formation (FSPM) and [2]Technische Fakultät, University of Bielefeld, Postfach 100 131, 33501 Bielefeld, Germany

Received on August 27, 1997; revised on October 7, 1997; accepted on October 14, 1997

BIOINFORMATICS

Vol. 21 Suppl. 3 2005, pages iii31–iii38
doi:10.1093/bioinformatics/bti1200

Sequence analysis

DNA assembly with gaps (Dawg): simulating sequence evolution

Reed A. Cartwright

Department of Genetics, University of Georgia, Athens, GA 30602-7223, USA

Received on May 29, 2005; accepted on August 16, 2005

Mol. Biol. Evol. 25(4):688–695. 2008

Simulating DNA Coding Sequence Evolution with EvolveAGene 3

Barry G. Hall

Bellingham Research

Mol. Biol. Evol. 26(8):1879–1888. 2009

INDELible: A Flexible Simulator of Biological Sequence Evolution

William Fletcher and Ziheng Yang

Department of Genetics, Ev
Biology, University Colleg

Mol. Biol. Evol. 26(11):2581–2593. 2009

Biological Sequence Simulation for Testing Complex Evolutionary Hypotheses: indel-Seq-Gen Version 2.0

Cory L. Strope, Kevin Abel,* Stephen D. Scott,* and Etsuko N. Moriyama†‡*

r Science and Engineering, University of Nebraska; †School of Biological Sciences, University of Nebraska;
ience Innovation, University of Nebraska

Sipos *et al. BMC Bioinformatics* 2011, **12**:104
http://www.biomedcentral.com/1471-2105/12/104

BMC Bioinformatics

SOFTWARE Open Access

PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment

Botond Sipos[1,2*], Tim Massingham[1], Gregory E Jordan[1] and Nick Goldman[1]

Mol. Biol. Evol. doi:10.1093/molbev/msr268

ALF—A Simulation Framework for Genome Evolution

Daniel A. Dalquen,[1,2,*] Maria Anisimova,[1,2] Gaston H. Gonnet,[1,2] and Christophe Dessimoz[1,2]

Measures of accuracy

Result	Reference
LMGP--	LM-GP
LDRA-V	LDRAV
LF--RR	LF-RR

"True column" (TC) score: 2/6 = 33%
(% of result columns that are present in reference alignment)

"Sum of Pairs" (SP) score: $1/3 * (2/6 + 2/6 + 4/6) = 44\%$
(average % of correct residue pairs, over all sequence pairs)

Blackbox comparisons

IRMbase I/II (Subramanian 2004, 2008)

based on ROSE (Stoye 2004)

Table 7
Results from the IRMbase dataset

	IRM_1_4		IRM_1_8		IRM_1_16		IRM_2_4		IRM_2_8		IRM_2_16		IRM_3_4		IRM_3_8		IRM_3_16	
Align-m	**91.95**	26.47**	85.68*	30.99**	71.42	41.03**	74.04	32.60**	74.64*	40.9**	73.55*	51.97**	69.19**	36.44**	79.03**	52.75**	70.81**	60.32**
ClustalW	1.029**	−0.39**	1.327**	−0.98**	0.0**	−2.28**	3.35**	0.825**	0.33**	−0.34**	0.65**	−1.22**	11.74**	3.635**	2.99**	3.455**	3.068**	2.808**
T-Coffee	73.15	22.53**	86.02	25.88**	77.29	25.74**	66.84*	30.58**	67.54	33.35**	77.52*	36.19**	70.49*	39.21**	74.41**	45.45**	74.76*	45.57**
Dialign-t	88.69	39.10**	**90.69**	47.59**	76.52	52.43**	**88.1**	50.28	**80.86**	**56.03**	79.47	63.42	**83.95**	**57.22**	**86.48**	**66.81**	81.23	**70.83**
Dialign-2	85.29	39.33**	89.55	47.43**	74.58	51.99**	89.51	49.53	78.19	54.16	79.27**	**63.86**	83.76	56.3	85.95	65.37*	77.66	69.01**
FFTNS	40.75**	13.22**	21.14**	14.32**	9.105**	16.65**	41.77**	20.64**	30.95**	22.17**	33.44**	26.38**	43.41**	27.48**	42.89**	32.56**	30.22**	35.85**
FFTNSi	41.06**	13.19**	30.50**	14.89**	37.01**	19.73**	46.30**	22.11**	35.82**	24.06**	44.26**	27.23**	46.79**	28.24**	54.13**	34.65**	55.04**	38.43**
GINSi	47.12**	14.69**	35.13**	18.62**	29.23**	19.06**	35.51**	19.49**	23.11**	18.26**	46.23**	27.57**	51.36**	28.91**	49.35**	33.63**	54.64**	38.59**
FINSi	85.19	25.24**	88.37	27.21**	**78.56**	26.22**	67.37*	30.01**	61.00*	30.78**	69.53	34.38**	58.36**	33.14**	60.69**	39.13**	60.57**	40.53**
MUSCLE	21.54**	8.849**	11.48**	7.72**	2.17**	5.97**	16.41**	10.84**	8.12**	9.026**	8.18**	11.66**	31.24**	19.90**	30.41**	23.53**	16.32**	21.29**
NWNS	36.25**	12.02**	21.55**	13.85**	4.605**	14.47**	31.76**	16.51**	15.33**	15.80**	14.11**	20.17**	41.77**	25.87**	40.53**	29.57**	22.64**	31.97**
NWNSi	36.56**	11.99**	30.62**	14.69**	34.04**	18.63**	39.22**	19.14**	21.45**	17.99**	39.10**	24.50**	44.48**	26.38**	43.82**	30.46**	52.24**	36.37**
PCMA	76.61	25.72**	89.51	29.75**	77.18	31.93**	72.11	34.26**	71.52	39.39**	**84.04**	44.26**	65.92**	38.40**	75.53*	50.82**	**82.7**	53.97**
POA	87.88	**91.34**	74.57**	**81.84**	70.24	**79.09**	26.93**	**52.7**	12.80**	41.12**	5.036**	37.27**	13.49**	39.44**	10.36**	40.85**	3.10**	35.66**
ProbCons	35.88**	14.10**	38.61**	22.15**	33.04**	29.24**	41.15**	21.54**	35.58**	28.15**	44.47**	40.20**	60.50**	34.56**	53.13**	40.41**	47.48**	50.5**
Average	56.59	23.82	52.98	26.4	45	28.66	49.36	27.4	41.15	28.72	46.59	33.85	51.76	33.01	52.65	39.3	48.83	42.11
Std Dev	28.93	21.61	33.7	20.26	31.36	20.56	25.79	14.84	28.27	15.87	29.72	17.74	21.99	13.23	25.48	16.03	27.62	17.58
MAX	91.95	91.34	90.69	81.84	78.56	79.09	89.51	52.7	80.86	56.03	84.04	63.86	83.95	57.22	86.48	66.81	82.7	70.83

Each set differs in the number of ROSE generated motifs (1, 2 or 3) inserted into otherwise random sequences, and the number of such sequences used (4,8 or 16) in each alignment Values in bold denote the highest score within each subset. First score within each cell denotes the Column score, the second value denotes the Shift score. The significance of difference from the most accurate method is indicated by $^{*}(p < 0.05)$ or ** $(p < 0.01)$ (Wilcoxon test).

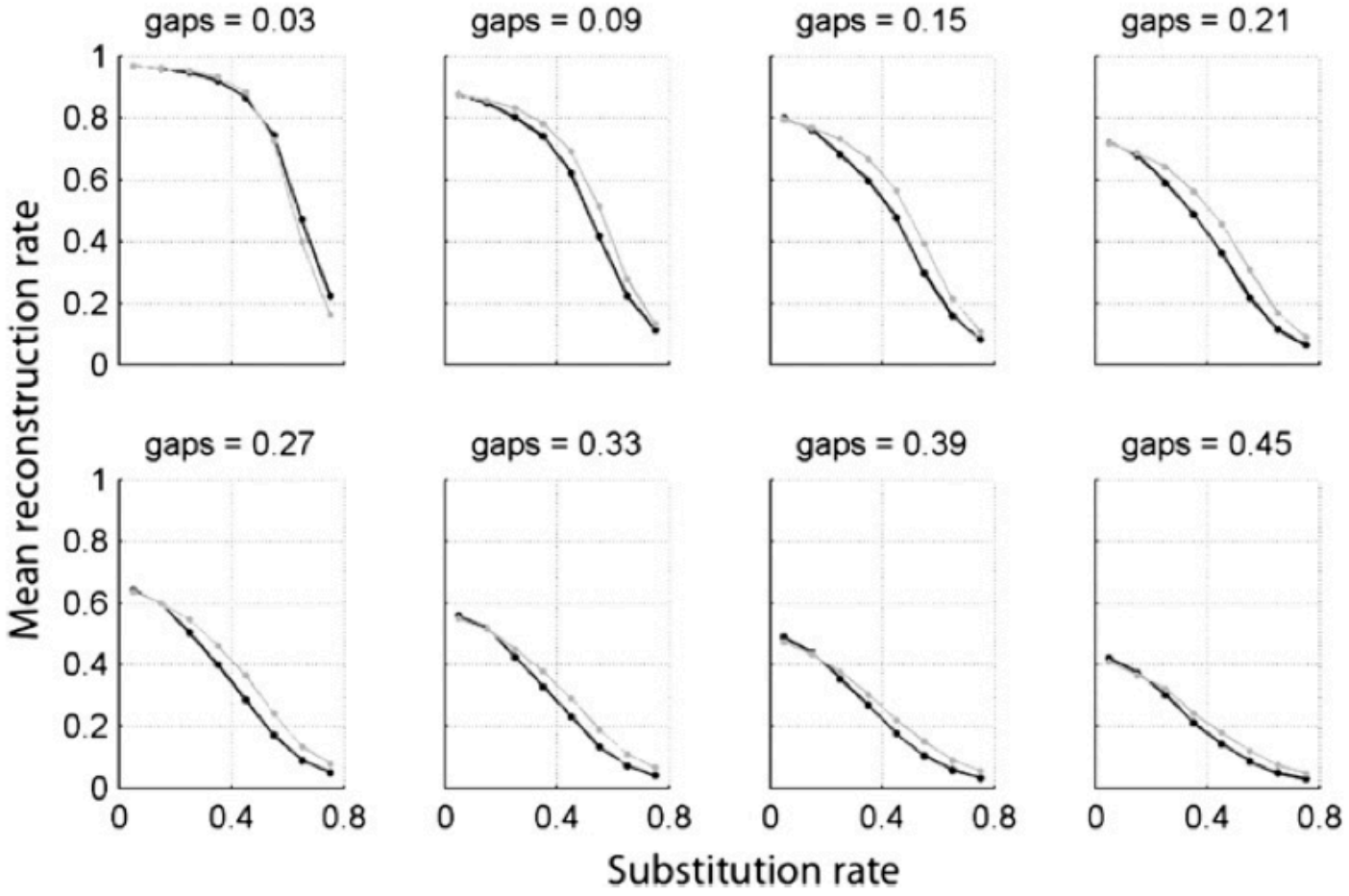
Blackshields et al. 2006

Mechanistic insights

Gene 441 (2009) 141–147

Characterization of pairwise and multiple sequence alignment errors

Giddy Landan *, Dan Graur



“We found a systematic bias towards underestimation of the number of gaps, which leads to the reconstructed MSA being on average shorter than the true one.”

“The quality of the guide-tree was found to affect MSA error levels only marginally.”

“[A]t even moderate evolutionary distances, reconstructed alignments are correct for only about half of their length.”

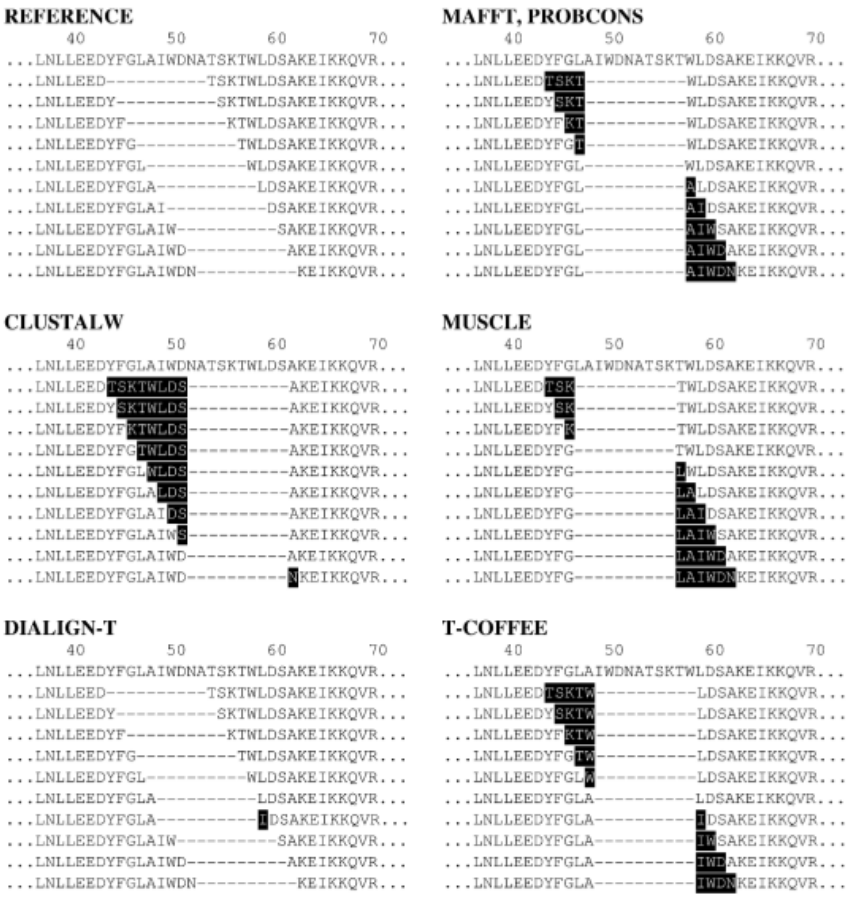
(Note of caution: all results based on one multiple sequence alignment method only - Clustal W. Generalisability?)

Boundless possibilities…

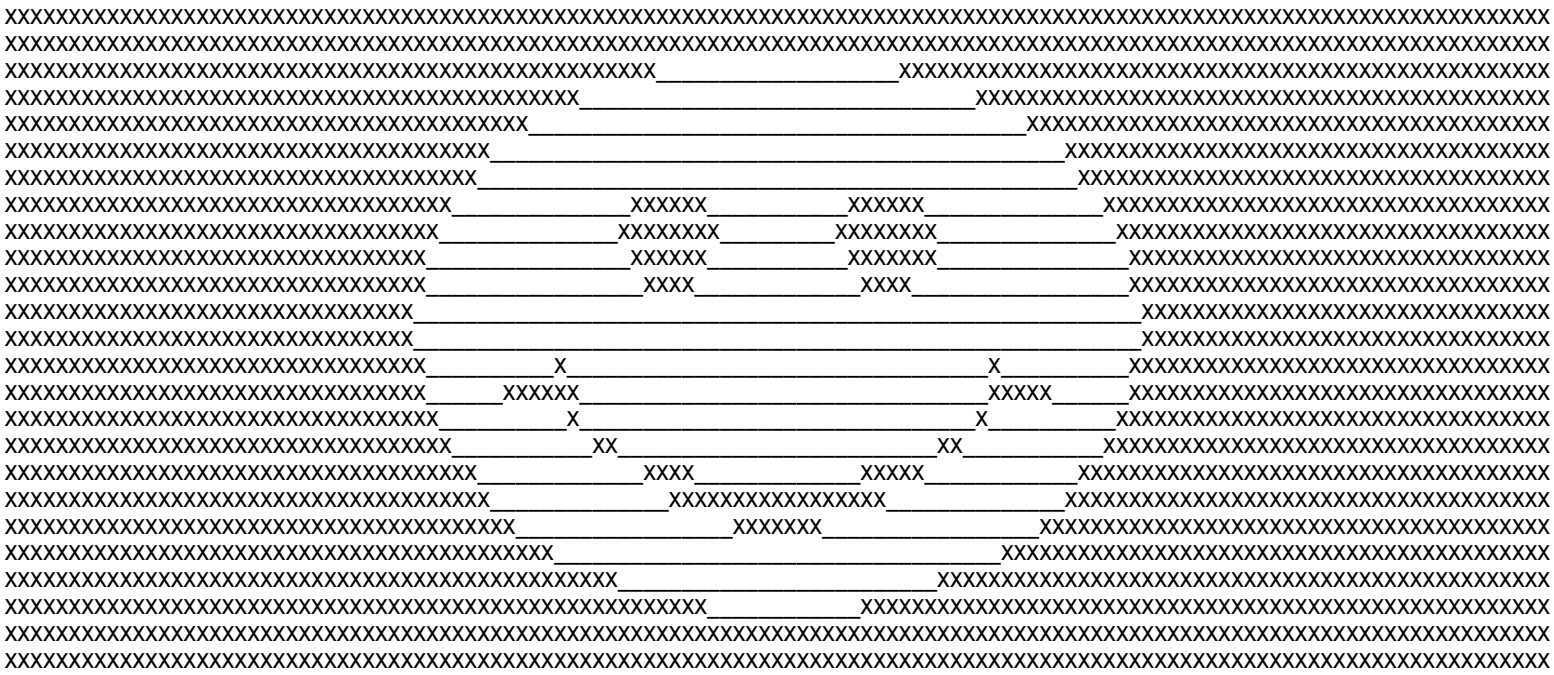
Mol. Biol. Evol. 24(11):2433–2442. 2007

Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments

Tanya Golubchik, Michael J. Wise,† Simon Easteal,‡ and Lars S. Jermiin*



How about this reference alignment?



Sequence artefacts due to technological issues

- **Sequences have artefacts due to technological limitations,** such as sequencing errors, assembly errors, gene models, etc. (Thompson et al. 2011)
- **These are not modelled in current simulation framework**
  (→ research opportunity!)

Simulation: summary

- 👍 all parameters are known exactly
- 👍 ability to test under a wide range of evolutionary conditions

- 👎 conclusion strongly depend on model used to create data
- 👎 a meaningful simulation requires realistic and relevant setup/parameters (e.g. including modelling of seq. artefacts)

2.“Consistency”

- **Warning: not meant here is consistency with pairwise alignment** (a common alignment optimisation criterion)
- **Nor is meant consistency in the statistical sense** (quality of converging to true value as # of datapoint increases)
- **But rather: consistency among alignments produced by different aligners or procedures**
- **Conceptually somewhat analogous to “bootstrapping”**

e.g.

7120–7128 *Nucleic Acids Research, 2005, Vol. 33, No. 22*
doi:10.1093/nar/gki1020

Automatic assessment of alignment quality

Timo Lassmann* and Erik L. L. Sonnhammer

set of columns of alignment a

$$Q_{ab} = \frac{|Q_a \cap Q_b|}{(|Q_a| + |Q_b|)/2}$$

generalisation to >2 underlying aligners:

$$O_{average} = \frac{\sum_{i}^{m-1} \sum_{j=i-1}^{m} O_{ij}}{m(m-1)/2}$$

Methods based on similar ideas include

M-Coffee (Notredame et al. 2000), TrimAl (Capella-Gutiérrez 2009), AQUA (Muller et al. 2010)

Consistency of a method with itself

Mol. Biol. Evol. 24(6):1380–1383. 2007

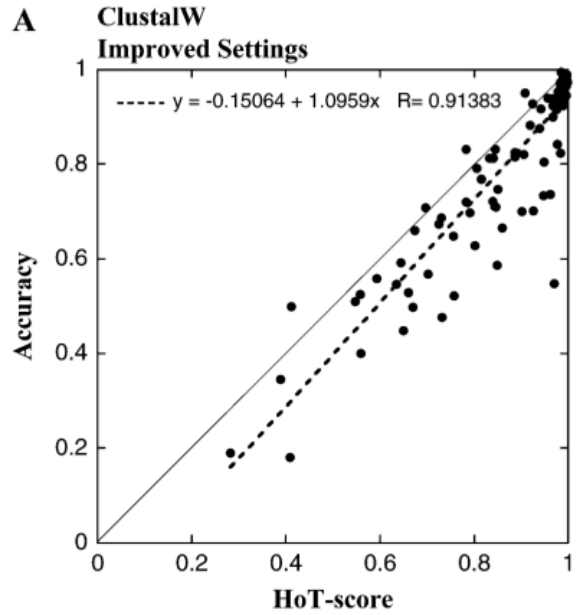
Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments

Giddy Landan and Dan Graur
Department of Biology & Biochemistry, University of Houston

"Thus, [ideally] reversing residue order prior to alignment should yield an exact reversed alignment of that obtained by using the unreversed sequences"

Quis custodiet ipsos custodes?

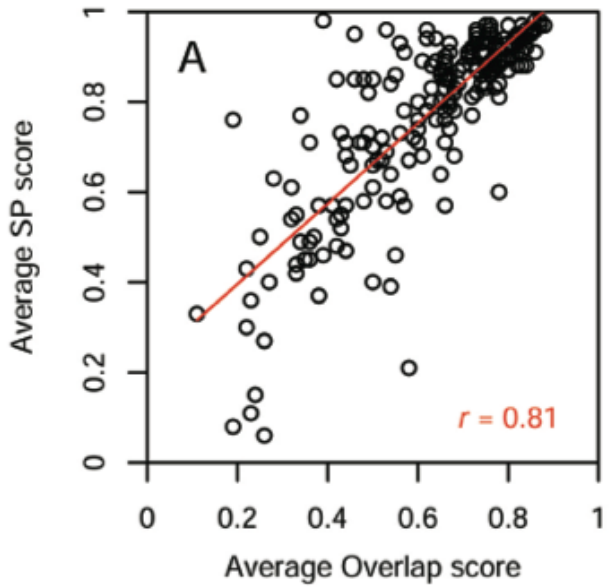
Simulation:



Consistency

Hall BG, "How well does the HoT score reflect sequence alignment accuracy?", MBE 2008

Structure (Balibase)



Consistency

Lassmann & Sonnhammer NAR 2005

Consistency: summary

- 👍 conceptually simple
- 👍 “necessary” condition: good methods are consistent

- 👎 “insufficient” condition: methods can be consistently wrong
- 👎 more generally, difficult to deal with correlation among methods

3. "Expert review"

(a.k.a. eyeballing)

Syst. Biol. 58(1):150–158, 2009

Why Would Phylogeneticists Ignore Computerized Sequence Alignment?

DAVID A. MORRISON

Department of Parasitology (SWEPAR), National Veterinary Institute and Swedish University of Agricultural Sciences, 751 89 Uppsala, Sweden; E-mail: David.Morrison@bvf.slu.se.

TABLE 1. Characteristics of the multiple alignment procedures reported in 1280 papers published in 26 biology journals during 2007. The percentages do not sum to 100 because any one paper may have included several alignment techniques (e.g., for different genes)

Discipline	No. of papers	Unspecified (%)	Clustal (%)	Second most common (%)	MAFFT or ProbCons (%)	Jump Combined (%)[a]	Modified[b] By eye (%)[d]	Modified[b] Criterion (%)[e]	Manual By eye (%)[f]	Manual Criterion (%)[g]	Start (%)[c]	
General[h]	326	2	51	1 (ProAlign)	1	0	19	14	35	10	4	78%
Systematics[i]	247	2	46	6 (POY)	1	6	24	9	36	7	7	76%
Evolution[j]	222	5	59	4 (Muscle)	1	1	18	5	22	6	3	51%
Molecular biology[k]	232	2	61	7 (Muscle)	2	0	24	9	19	3	7	55%
Microbiology[l]	253	11	66	6 (ARB)	<1	0	9	2	12	4	8	27%
Combined	1280	4	56	1 (POY)[m]	1	1	19	8	26	6	6	59%

[a] Combined alignment + tree building, including POY and BaliPhy.
[b] Excluding the simple statement that gaps were deleted before tree building.
[c] Explicit use of a previous alignment, including those from the ARB project, the Ribosomal Database Project, and the European rRNA Database.
[d] No criterion specified.
[e] Explicitly stated criteria included the use of SOAP, adjustment to match RNA structure/codons/motifs, and minimizing the total number of mutational changes/polymorphisms/variable columns.
[f] Usually occurred for sequences with high similarity, such as within-species comparisons, and chloroplast sequences.
[g] Explicitly stated criteria included match to RNA structure/codons/motifs and recognition of events that cause sequence variation (e.g., repeats, inversions).

Example

Kelchner, Annals of the Missouri Botanical Garden 87:482-498 (2000)

(cited by 338)

EXAMPLE 6A.

1. GGTTAAT **tctat** TCTATCT
2. GGTTAAT **ttaat** TCTATCT
3. GGTTAAT ttaat TCTATCT
4. GGTTAAT ----- TCTATCT
5. GGTTAAT ----- TCTATCT

Alignment of the insertions in Example 6A results in the probably mistaken homology of indels in sequences 2 and 3 with that of sequence 1. The insertion in sequence 1 likely arose from an inserted repeat of the sequence to the right of the gap, TCTAT. This would be a more parsimonious explanation, in terms of total number of mutation events, than to infer a single inserted repeat followed by two adjacent nucleotide substitutions in

...

Expert review: summary

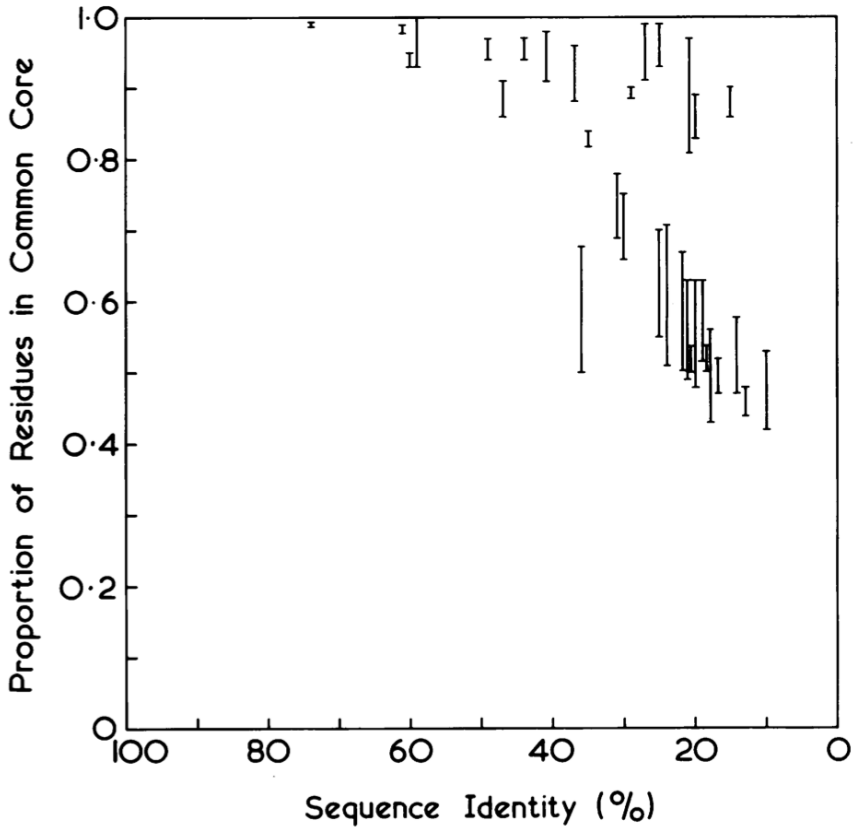
- 👍 based on expert judgement, which builds upon broad array of clues and past experiences
- 👎 lacks reproducibility
- 👎 experts suffer from biases (status quo bias, aesthetic considerations, etc.)

also see Anisimova et al. (Trends Evol Biol, 2011) for forceful arguments against this type of benchmark

4. Empirical approaches

4.1 Structure

Reviewed in detail in
Kemena & Notredame 2010
Edgar 2010
Aniba et al. 2011



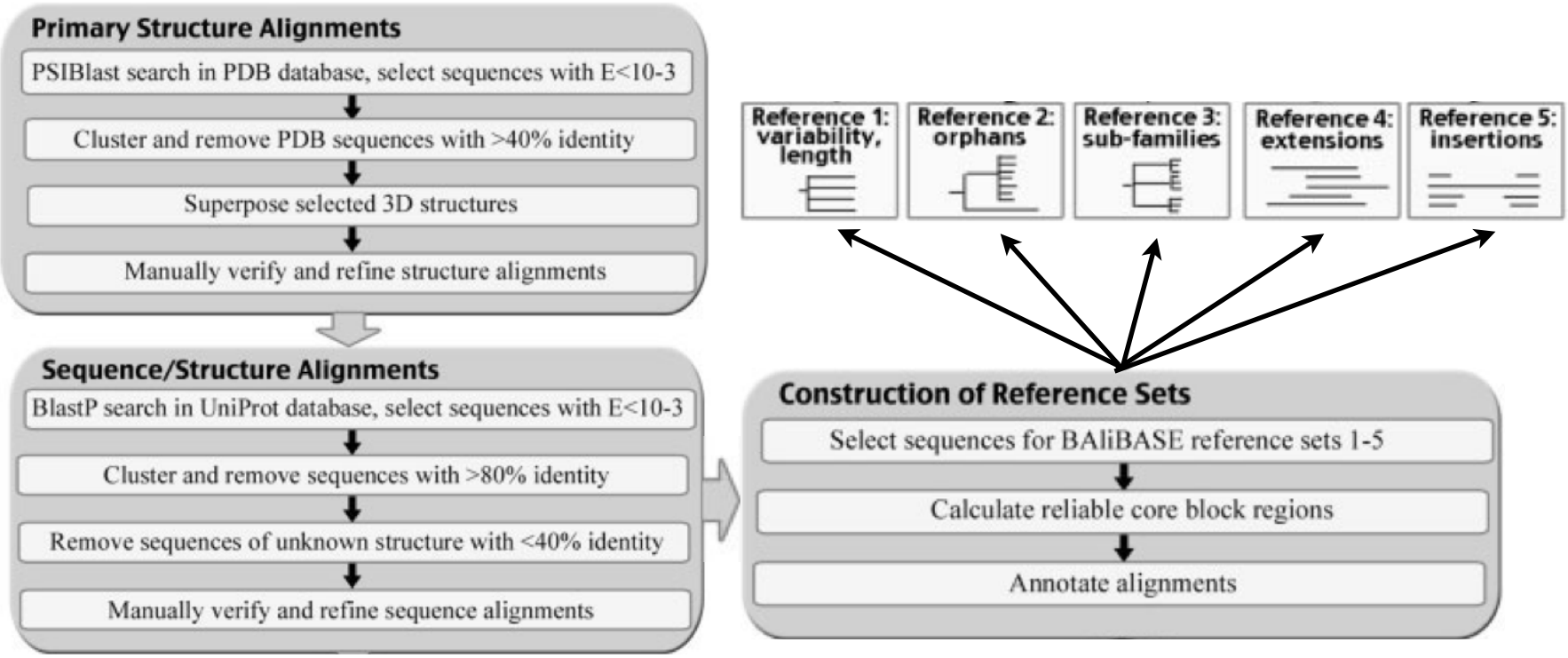
Cyrus Chothia[1] and Arthur M.Lesk[2]

The EMBO Journal vol.5 no.4 pp.823–826, 1986

PROTEINS: Structure, Function, and Bioinformatics 61:127–136 (2005)

BAliBASE 3.0: Latest Developments of the Multiple Sequence Alignment Benchmark

Julie D. Thompson,[1*] Patrice Koehl,[2] Raymond Ripp,[1] and Olivier Poch[1]



Other structural benchmarks

- Homestrad (Stebbings and Mizugishi 2004)
- OXBench (Raghava et al 2003)
- Prefab (Edgar 2004)
- SABmark (Van Walle et al. 2005)
- Bralibase for ncRNAs (Gardner & al. 2005)

Quis custodiet ipsos custodes?

Published online 4 January 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 7* ***2145–2153***
doi:10.1093/nar/gkp1196

Quality measures for protein alignment benchmarks

Robert C. Edgar*

Table 1. Reference alignment domain and secondary structure agreement scores

Benchmark	Cols	DSS	Ann	CSF	CFLD	ESF	EFLD	ECLS
BALIBASE	Core	28.8	SF	89.4	93.0	11.5	8.1	5.9
	Non-core(U)	69.4	SF	83.9	86.0	22.0	20.4	15.0
	Non-core(T)		SF	90.9	94.1	10.4	7.1	4.2
PREFAB	Ref	28.4	SCOP	96.2	98.5	3.8	1.5	0.5
			CATH	95.4	97.9	4.6	2.1	0.6
OXBENCH	SCR	22.9	SCOP	96.0	99.3	3.9	0.6	0.2
			CATH	96.3	99.2	3.8	0.8	0.0
	Non-SCR	58.7	SCOP	96.5	99.4	3.5	0.6	0.2
			CATH	96.6	99.3	3.4	0.7	0.0

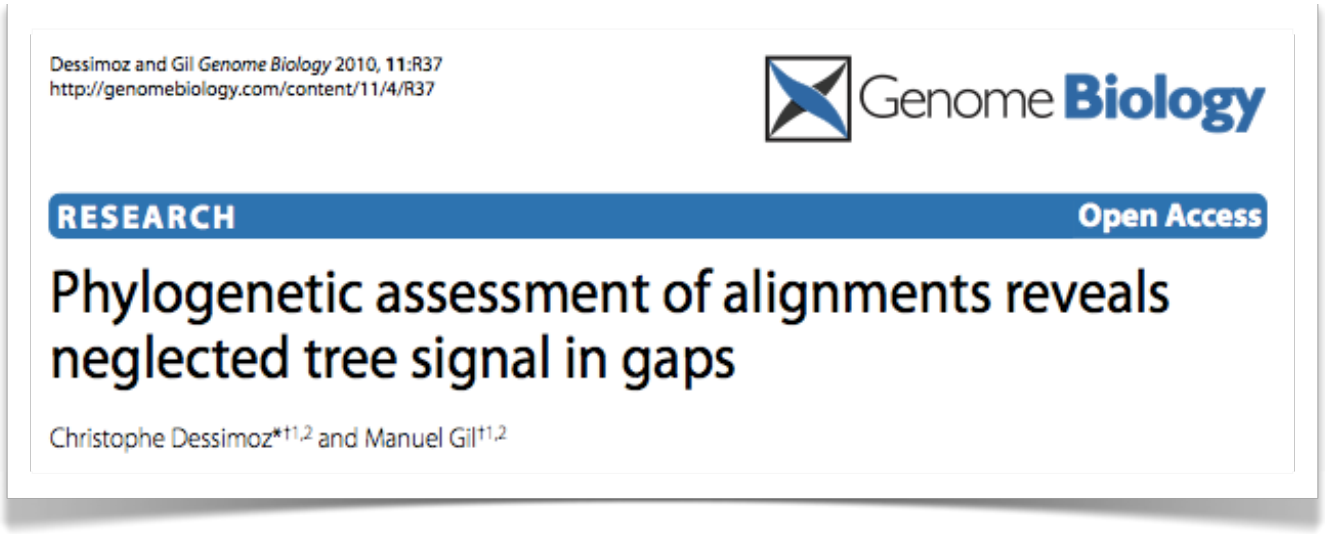
“The present results show that protein alignment assessment is more challenging than generally realized, and skepticism is appropriate for claims that method rankings or advances can be reliably measured by current benchmarks.”

Pros/Cons Structure

- 👍 ability to test real data
- 👍 if interest for structure:
  closely match the biological objective

- 👎 only for structurally conserved regions!
- 👎 small/biased protein sample with known structure
- 👎 map between structural alignment and sequence alignment non-trivial (distance threshold?)

4.2 Phylogenetic tests



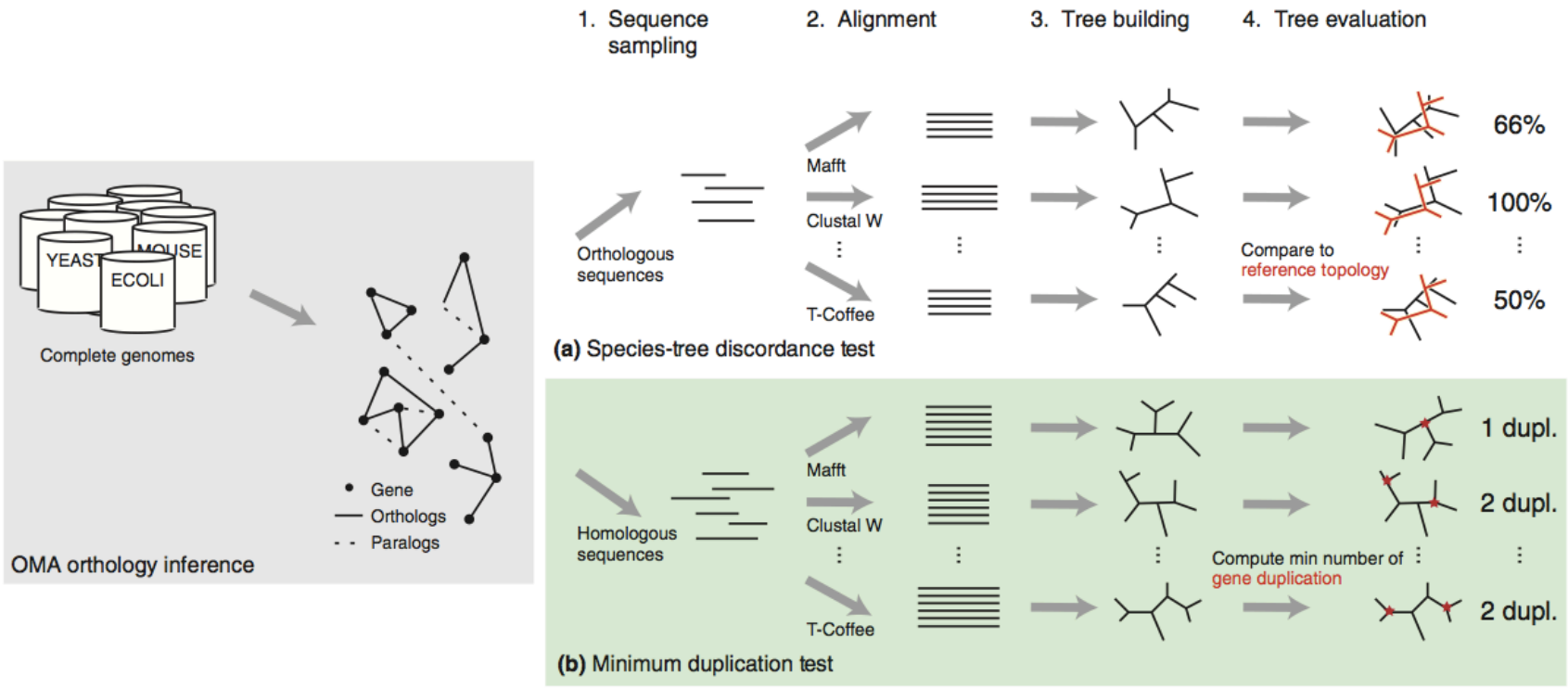
Dessimoz and Gil *Genome Biology* 2010, **11**:R37
http://genomebiology.com/content/11/4/R37

Genome Biology

RESEARCH Open Access

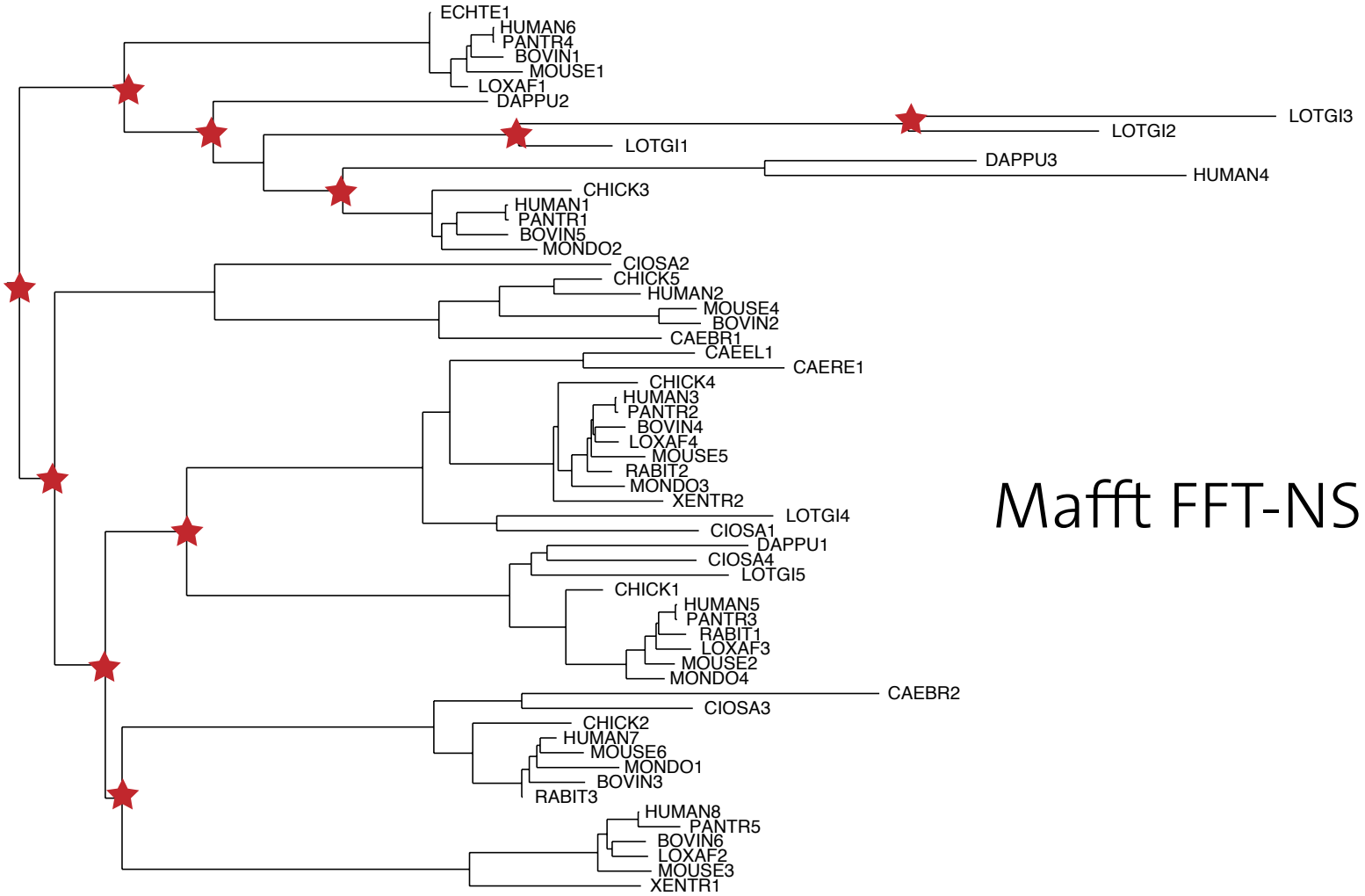
Phylogenetic assessment of alignments reveals neglected tree signal in gaps

Christophe Dessimoz*†[1,2] and Manuel Gil†[1,2]



Dessimoz & Gil 2010

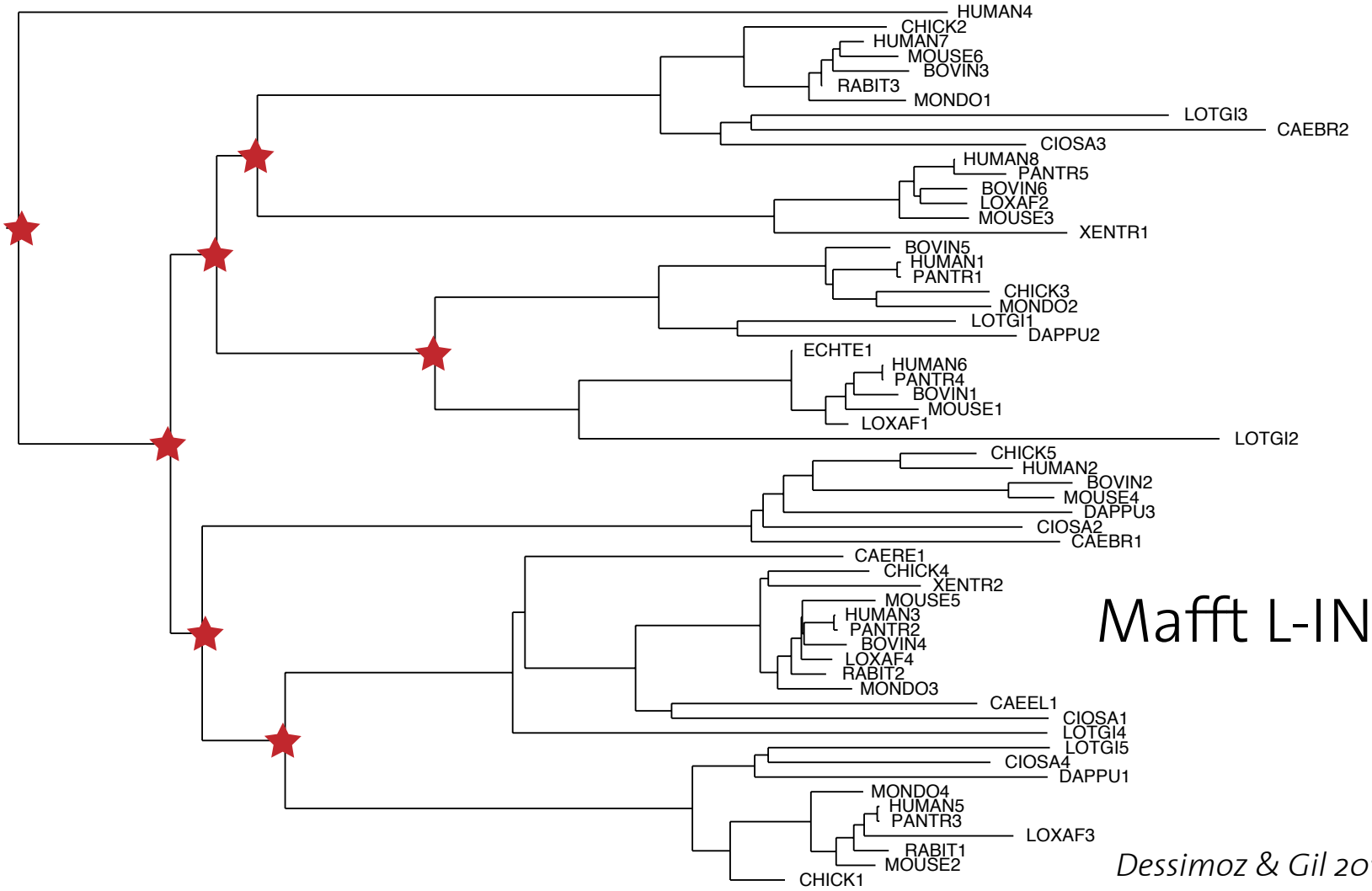
E.g. GDP-fucose transporter



Mafft FFT-NS-2

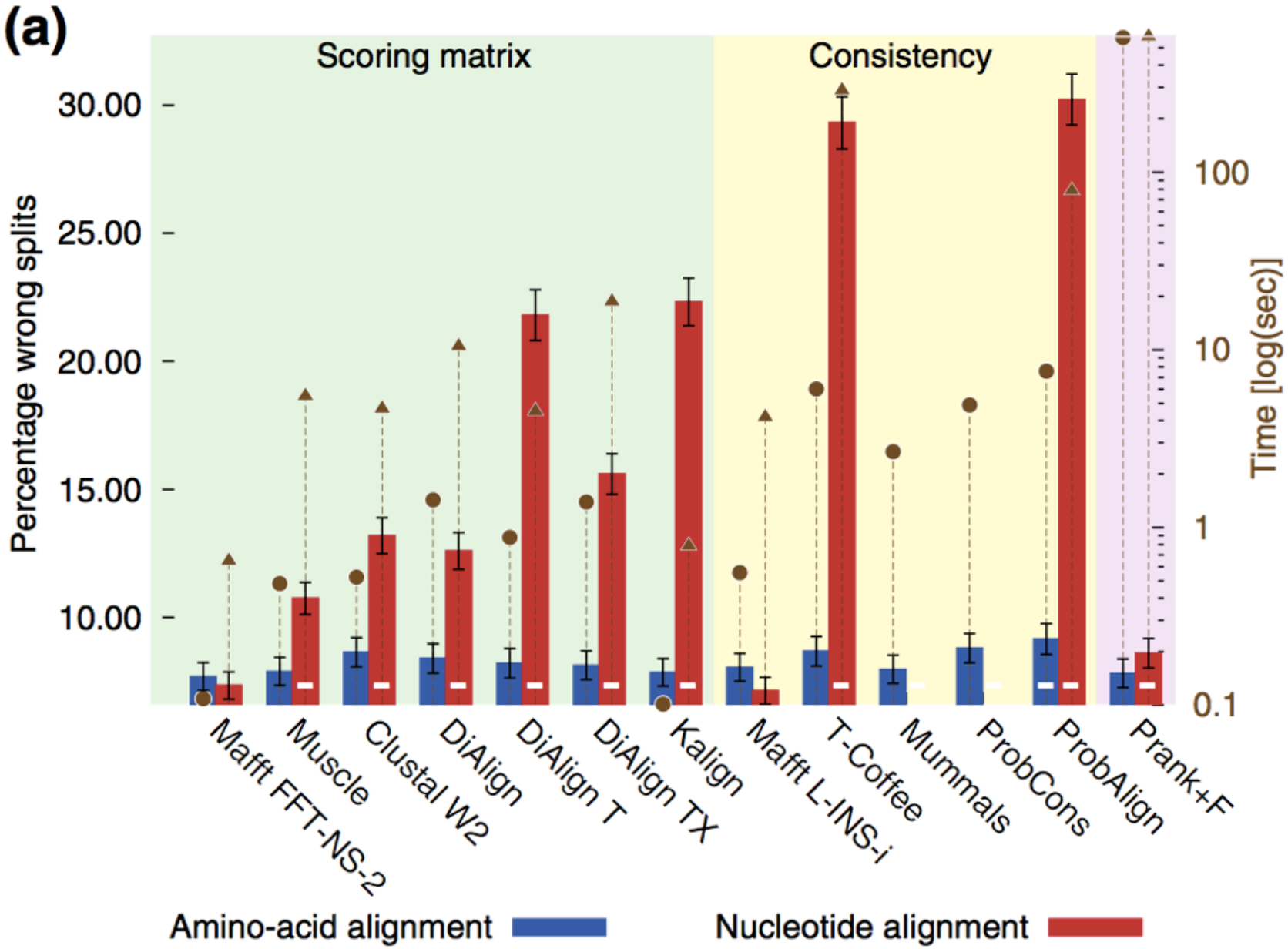
Dessimoz & Gil 2010

E.g. GDP-fucose transporter



Mafft L-INS-i

Dessimoz & Gil 2010



Dessimoz & Gil 2010

Pros/Cons Phylogenetic tests

- 👍 are able to test broad sample of real data
- 👍 if interest for tree building:
  closely match the biological objective
- 👎 cannot provide a quality statement about a single alignment
- 👎 provide limited insights for MSA applications other than tree building (e.g. detecting sites under positive selection, build profile, predict structurally conserved sites, etc..)

One step back: what are ideal empirical indicators?

- **Surrogate is (highly) correlated with objective**
- **Surrogate is not "used" by any of the aligner**

→ meta-methods (e.g. Notredame et al. 2000, Lassmann & Sonnhammer 2005, etc.) cannot be reliably assessed with consistency
→ HoT can be easily "gamed"
→ 3-D coffee (Poirot et al. 2005) cannot be reliably assessed with structural benchmarks

Reconciling the various approaches

Structure vs Simulation?

using six available datasets. The main trend uncovered by this analysis is that all the empirical reference datasets tend to yield similar results, quite significantly distinct from those measured on artificial datasets such as IRMbase (Subramanian *et al.*, 2005,

Kemena & Notredame 2010

Phylogeny vs Structure?

tion of gap regions. Indeed, our results show that consistency-based alignment methods, which score best in structural benchmarks, do not yield significantly better trees than their scoring matrix-based counterparts. Our

Dessimoz & Gil 2010

What do?

- **Be very careful with generalisation:**
  - Important parameters might include: gappy regions (near neutral) vs conserved core (strong purifying selection); size of alignment; divergence; etc.
  - e.g. Landan & Graur 2009: “In this study we use ClustalW as the standard in MSA reconstruction.”
- **Be aware of the assumptions and biases of each method**
  - Simulation strategy
  - Data underlying empirical tests
- **Exploit differences to gain understanding!**