Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 16

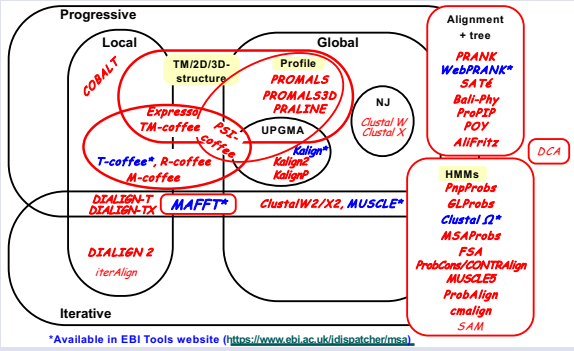
1

TODAY'S TOPICS

- ➢ Multiple Alignment
  - • ClustalΩ, PRANK, etc.
- ➢ Alignment trimming/filtering
- ➢ MSA evaluation
- ➢ Conserved domain, pattern, profile
  - • PROSITE, PRINTS, Profile (PSSM)
- ➢ Assignment 8 (due: April 4)

2

Multiple alignment: Heuristic methods



*Available in EBI Tools website (https://www.ebi.ac.uk/jdispatcher/msa)

For the newest review: Katoh (2021) and more available on CANVAS

3

Clustal Ω http://www.clustal.org/omega/ https://www.ebi.ac.uk/jdispatcher/msa/clustalo

- **Progressive alignment** following the guide tree
- Features a fast method for making "guide trees"
  - ➜ calculates only distances to *n* references (**mBed method**)
  - ➜ scalable for very large datasets
- Alignment is done using **HHalign (a profile hidden Markov model alignment)**
  - ➜ highly accurate alignment
- **Simple iterative refinement**
  - ➜ Alignment is converted to hidden Markov model (HMM)
  - ➜ Realign input sequences against the HMM

Sievers *et al.* (2011, 2018)

4

Clustal Ω http://www.clustal.org/omega/ https://www.ebi.ac.uk/jdispatcher/msa/clustalo

Top-rated methods use HMM or consistency function or both

Table 1 BAliBASE results

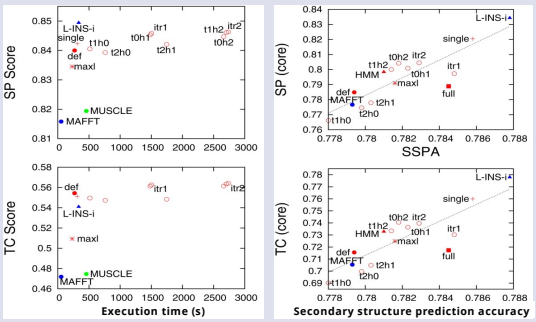
Aligner	Av score (218 families)	BB11 (38 families)	BB12 (44 families)	BB2 (41 families)	BB3 (30 families)	BB4 (49 families)	BB5 (16 families)	Tot time (s)	Consistency
MSAprobs	0.607	0.441	0.865	0.464	0.607	0.622	0.608	12 382.00	Yes
Probaligin	0.589	0.453	0.862	0.439	0.566	0.603	0.549	10 095.20	Yes
MAFFT (auto)	0.588	0.439	0.831	0.450	0.581	0.605	0.591	1475.40	Mostly (203/218)
Probcons	0.558	0.417	0.855	0.406	0.544	0.532	0.573	13 086.30	Yes
Clustal Ω	0.554	0.358	0.789	0.450	0.575	0.579	0.533	539.91	No
T-Coffee	0.551	0.410	0.848	0.402	0.491	0.545	0.587	81 041.50	Yes
Kalign	0.501	0.365	0.790	0.360	0.476	0.504	0.435	21.88	No
MUSCLE	0.475	0.318	0.804	0.350	0.409	0.450	0.460	789.57	No
MAFFT (default)	0.458	0.258	0.749	0.316	0.425	0.480	0.496	68.24	No
FSA	0.419	0.270	0.818	0.187	0.259	0.474	0.398	53 648.10	No
Dialign	0.415	0.265	0.696	0.292	0.312	0.441	0.425	3977.44	No
PRANK	0.376	0.223	0.680	0.257	0.321	0.360	0.356	128 355.00	No
ClustalW	0.374	0.227	0.712	0.220	0.272	0.396	0.308	766.47	No

The figures are total column scores produced using bali score on core columns only. The average score over all families is given in the second column. The results for BAliBASE subgroupings are in columns 3–8. The total run time for all 218 families is given in the second last column. The last column indicates whether the method is consistency based.

Sievers *et al.* (2011)

5

Clustal Ω http://www.clustal.org/omega/ https://www.ebi.ac.uk/jdispatcher/msa/clustalo



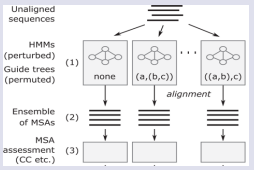
Sievers *et al.* (2018)

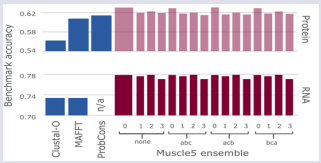
6

Muscle5

Edgar (2022)
http://www.drive5.com/muscle/

- Re-implementation of ProbCons (slow but highly accurate MSA method based on hidden Markov model)
  - Parallelization for rapid alignment
  - Can be used for large datasets
  - Supports both protein and DNA alignments
  - Can generate ensemble MSAs for MSA assessment





- More accurate than Clustal ω, MAFFT, or ProbCons
- Muscle5 ensemble MSAs have equal accuracy
- Provides reference-free estimate of MSA accuracy

7

PRANK, WebPRANK

Mind the gaps: Progress in progressive alignment

D. G. Higgins*, G. Blackshields, and I. M. Wallace
Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland

(2005 *PNAS* commentary)

"CLUSTALW attempts to compensate by using an elaborate scoring scheme to encourage gaps to end up on top of each other. ... results in alignments that are very "block-like"…"

"… there may be a price for this prettiness and detachment from phylogenetic reality. CLUSTALW (and other programs) may be guilty of "overalignment", that is where sequences that should not go together are forced into neat-looking blocks. These overaligned regions may be neat looking but misleading."

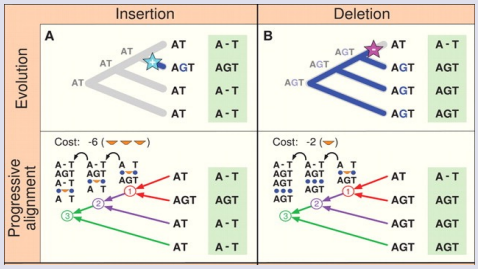
"There is an understandable tendency for users of multiple alignment software to want their residues neatly aligned in blocks and columns. This is fine when such blocks are biologically accurate as will happen in parts of protein alignments. In cases where insertions or deletions have happened in a less organized manner, as will happen in many noncoding DNA sequences and in less organized parts of protein sequences, such block-like alignments may be biologically meaningless. Perhaps we need to reeducate our eyes to see beauty in what actually happened rather than what looks nice on paper."

8

PRANK, WebPRANK

Löytynoja & Goldman (2008)

Insertions are more penalized than deletions in progressive sequence alignment.



9

PRANK, WebPRANK

Löytynoja & Goldman (2005, 2008, 2010)

https://www.ebi.ac.uk/goldman-srv/webprank/
http://wasabiapp.org/software/prank/

➢ PRANK: Probabilistic Alignment Kit

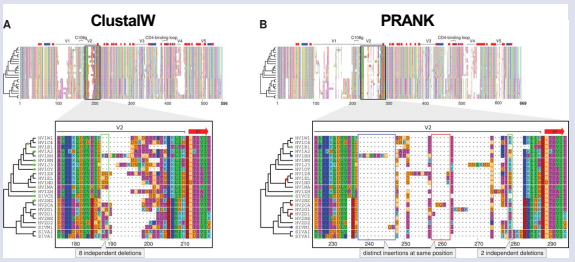
- A probabilistic multiple alignment program for DNA, codon, and amino-acid sequences.
- Treats insertions correctly.
- Avoids over-estimation of the number of deletion events.
- Not meant for the alignment of very diverged protein sequences.

10

PRANK, WebPRANK

Löytynoja & Goldman (2008)

Different sequence alignment approaches can give contradicting pictures of evolutionary mechanisms behind functional sequence changes.

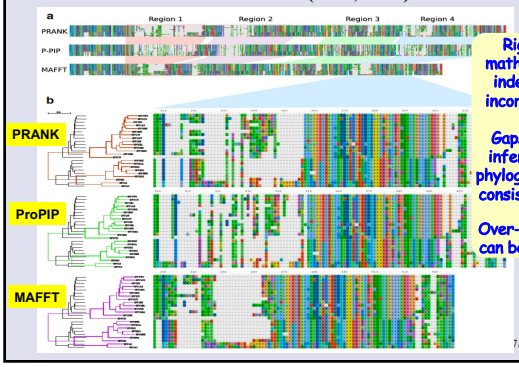


MSAs generated by traditional methods show excess substitutions
→ Can be erroneously thought to be under positive selection

11

ProPIP: Progressive MSA with Poisson Indel Process

Maiolo *et al.* (2018, 2021)



Rigorous mathematical indel model incorporated.

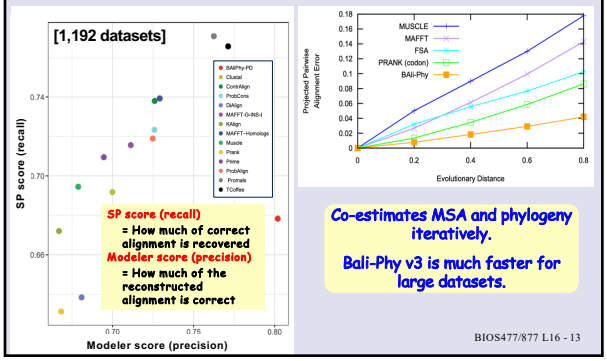
Gaps can be inferred in a phylogenetically consistent way.

Over-alignment can be avoided.

12

BAli-Phy: Statistical coestimation method

http://www.bali-phy.org/; Nute *et al.* (2019), Redelings (2021), Gupta *et al.* (2021)

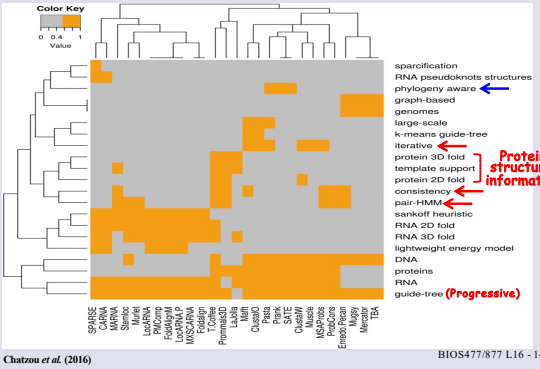


Co-estimates MSA and phylogeny iteratively.

Bali-Phy v3 is much faster for large datasets.

13

Multiple sequence alignment



Chatzou *et al.* (2016)

14

Alignment trimming

TABLE 1. Overview of filtering methods considered in this study

Filtering methods	Type of "undesirable" sites filtered out by the method	Accounts for tree structure?	Uses a substitution matrix or model of evolution?	Adapts parameters for particular data sets?	References
Gblocks	Gap-rich and variable sites	No	No	No	Talavera and Castresana (2007)
TrimAl	Gap-rich and variable sites	No	Yes	Yes	Capella-Gutiérrez et al. (2009)
Noisy	Homoplastic sites	In part	No	No	Dress et al. (2008)
Aliscore	Random-like sites	No	Indirectly	No	Kück et al. (2010)
BMGE	High entropy sites	No	Yes	No	Criscuolo and Gribaldo (2010)
Zorro	Sites with low posterior	Yes	Yes	No	Wu et al. (2012)
Guidance	Sites sensitive to the alignment guide tree	Yes	Indirectly	No	Penn et al. (2010)

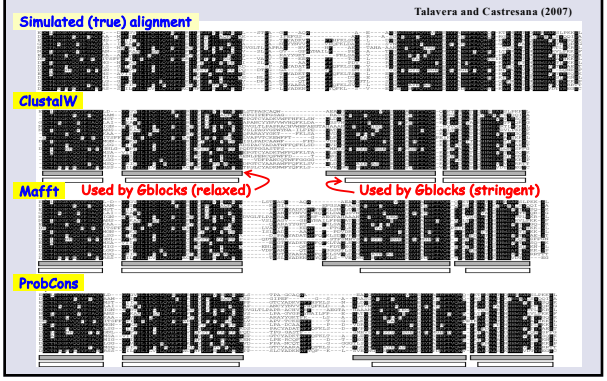
Tan *et al.* (2015)

- **Gblocks: selection of conserved blocks from multiple alignments**
  http://phylogeny.lirmm.fr/phylo_cgi/one_task.cgi?task_type=gblocks
  (included in Phylogeny.fr)
- **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analysis**
  http://trimal.cgenomics.org/trimal
  http://phylemon2.bioinfo.cipf.es (included in Phylemon2)

See also TCS paper by Chang *et al.* (2014)

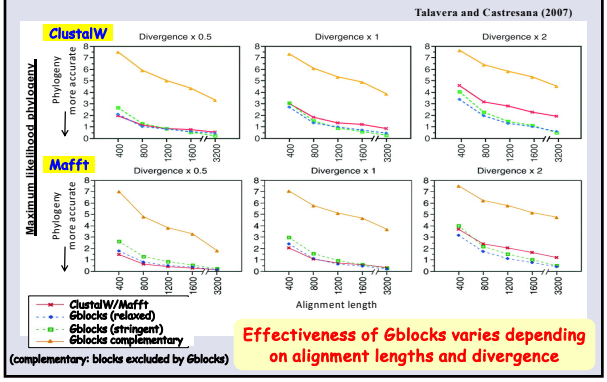
15

Alignment trimming: Gblocks



16

Alignment trimming: Gblocks



Effectiveness of Gblocks varies depending on alignment lengths and divergence

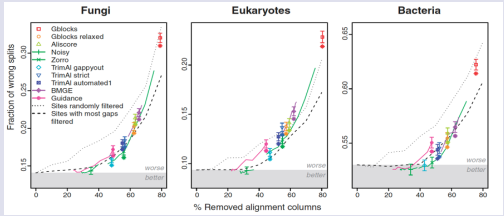
17

Alignment trimming

MSA trimming is used to improve phylogeny reconstruction. However, ...

Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference

Ge Tan[1,2,3], Matthieu Muffato[4], Christian Ledergerber[1], Javier Herrero[4,5], Nick Goldman[4], Manuel Gil[6,7], and Christophe Dessimoz[5,4,*] (2015)



Alignment filtering could worsen phylogenies...

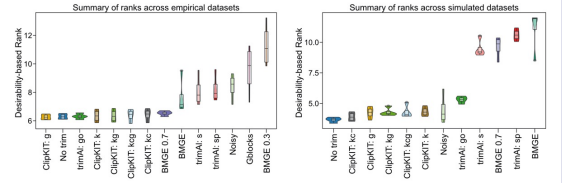
18

Alignment trimming: ClipKIT

Steenwyk et al. (2020) https://clipkit.genomelybio.com

Gblocks, TrimAl: remove phylogenetically uninformative (gap-rich or highly variable) sites

ClipKIT: identify and retain phylogenetically informative (parsimony-informative) sites

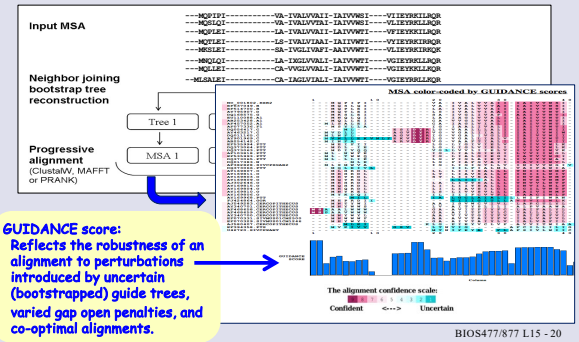


Desirability score (phylogeny based accuracy): combines tree accuracy and average bootstrap support

19

GUIDANCE2: Guide-tree based alignment confidence

https://taux.evolseq.net/guidance/ Penn *et al.* (2010); Sela *et al.* (2015)



GUIDANCE score: Reflects the robustness of an alignment to perturbations introduced by uncertain (bootstrapped) guide trees, varied gap open penalties, and co-optimal alignments.

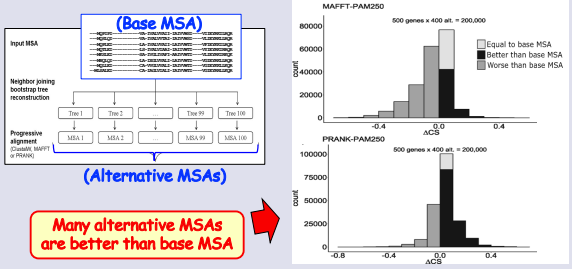
20

GUIDANCE2: Guide-tree based alignment confidence

https://taux.evolseq.net/guidance/

Multiple Sequence Alignment Averaging Improves Phylogeny Reconstruction

Haim Ashkenazy[1], Itamar Sela[2], Eli Levy Karin[1,3], Giddy Landan[4], and Tal Pupko[1,*] (2019)



Many alternative MSAs are better than base MSA

21

GUIDANCE2: Guide-tree based alignment confidence

https://taux.evolseq.net/guidance/

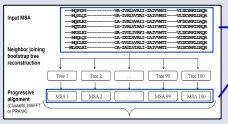
Multiple Sequence Alignment Averaging Improves Phylogeny Reconstruction

Haim Ashkenazy[1], Itamar Sela[2], Eli Levy Karin[1,3], Giddy Landan[4], and Tal Pupko[1,*] (2019)

TABLE 1. The effect of MSA weighting (GUIDANCE2, ZORRO, TCS) and averaging on tree inference accuracy for the PAM250 data set

Alignment method	No weighting Average normalized RF distance	GUIDANCE2 Average normalized RF distance	GUIDANCE2 *P*-value	ZORRO Average normalized RF distance	ZORRO *P*-value	TCS Average normalized RF distance	TCS *P*-value	SuperMSA Average normalized RF distance	SuperMSA *P*-value
True MSA	0.103 (0.07)								
MAFFT	0.187 (0.125)	**0.184 (0.123)**	**0.022**	0.187 (0.129)	0.295	0.188 (0.127)	0.661	**0.149 (0.099)**	**1.596e-24**
PRANK	0.169 (0.108)	0.168 (0.109)	0.286	0.174 (0.111)	0.999	0.176 (0.114)	0.999	**0.162 (0.105)**	**0.00011**

RF distance: tree distance from the correct tree (smaller is better)



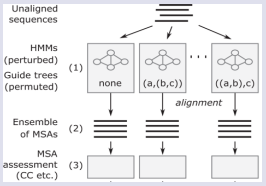
SuperMSA: base MSA concatenated with alternative MSA
Not a single improved MSA
But can be used to improve phylogeny

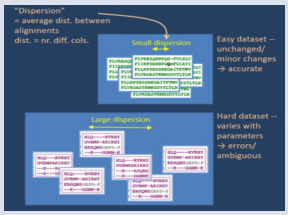
22

Muscle5

Edgar (2022)
http://www.drive5.com/muscle/

"Picking a single best protocol disregards the possibility that the best may not be good enough. … Even if alternative protocols are … less accurate, a thoughtful comparison of the results provides a useful indication of whether the preferred protocol can be trusted. "





Easy alignment
→ fewer alignment errors
→ ensemble MSA less dispersed

Difficult alignment
→ more alignment errors
→ ensemble MSA more dispersed

23

Alignment trimming/filtering: HmmCleaner

Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences (2019)

Sensitivity: true positive prediction/actual positive
Specificity: true negative prediction/actual negative

Table 2 Sensitivity and specificity of filtering software over different error types

sensitivity	prokaryotic MSAs frameshifts	scrambled	insertions	eukaryotic MSAs frameshifts	scrambled	insertions
TP/(TP+FN)						
HmmCleaner	96.95%	96.24%	99.24%	95.99%	96.26%	99.34%
PREQUAL	93.13%	96.77%	99.83%	85.64%	97.34%	99.95%
BMGE	16.00%	19.23%	97.55%	8.57%	12.57%	96.16%
OD-seq	20.11%	25.09%	59.89%	8.08%	9.34%	59.80%
GUIDANCE2	4.90%	5.37%	0.25%	2.33%	3.25%	1.69%

specificity	prokaryotic MSAs frameshifts	scrambled	insertions	eukaryotic MSAs frameshifts	scrambled	insertions
TN/(TN+FP)						
HmmCleaner	87.28%	87.24%	85.64%	94.70%	94.70%	94.63%
PREQUAL	92.80%	92.78%	92.76%	94.77%	94.65%	94.59%
BMGE	91.01%	90.88%	91.10%	97.05%	96.93%	96.94%
OD-seq	94.48%	94.46%	92.55%	95.20%	95.25%	94.90%
GUIDANCE2	99.55%	99.48%	99.85%	98.47%	98.35%	98.51%

Table 3 Detection of positive selection in the presence of primary sequence errors

	Raw	HmmCleaner	PREQUAL	BMGE	TrimAl
MSAs with positive selection in a targeted sequence	8.28%	3.88%	2.67%	7.24%	7.15%
MSAs with positive selection in a targeted sequence with an error	95.69%	7.76%	13.62%	95.00%	94.74%

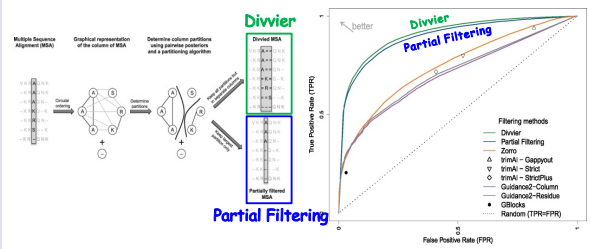
HmmCleaner & PREQUAL: using profile (or pair) HMMs, detect and remove regions with sequencing errors and erroneous annotations (non-homologous regions) from MSAs

24

Alignment trimming/filtering: Divvier

Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments
Raja Hashim Ali, Marcin Bogusz, Simon Whelan
Molecular Biology and Evolution, Volume 36, Issue 10, October 2019, Pages 2340–2351, https://doi.org/10.1093/molbev/msz142

Using a probabilistic model, clusters of characters sharing the homology are identified
➜ Divide them into different columns (divvying)



TPR (sensitivity) = TP/(TP+FN) : true positive prediction/actual positive
FPR (1-specificity) = TN/(TN+FP): true negative prediction/actual negative

25

Multiple alignment quality

➢ How to measure the quality of multiple alignments?

- Sum of pairs score (SPS) and total column score (CS or TCS):
  - ➜ SPS=Proportion of correctly aligned AA pairs
  - ➜ TCS (CS) =Proportion of correctly aligned columns

Ref	Test
VA-T	-VAT-
VA-G	-VAG-
MGTG	M-GTG

TCS=1/3 =33.3% or 25% (w/ gap col.)
SPS=5/9 =55.6%

[Programs available]
bali_score (from BAliBASE website)
qscore: http://drive5.com/qscore SPS, CS, Shift Score, *etc.*
Veralign: https://www.ibi.vu.nl/programs/veralignwww/

- T-coffee consistency based evaluation (TCS, iRMSD-APDB, Strike)
- MUMSA: average overlap score https://msa.sbc.su.se/cgi-bin/msa.cgi
- GUIDANCE2: guide-tree based alignment confidence https://taux.evolseq.net/guidance/
- Muscle5: ensemble MSA based accuracy https://www.drive5.com/muscle/
- AlignStat: MSA similarity/dissimilarity http://alignstat.science.latrobe.edu.au
- QuanTest: secondary structure prediction based
  http://www.bioinf.ucd.ie/download/QuanTest.tgz (download only)
  http://bioinf.ucd.ie/quantest2.tar (download only)
- Sequence logo: graphical representation of a multiple alignment
  Weblogo 3: https://weblogo.threeplusone.com/

26

Benchmark alignment database: BAliBASE

http://www.lbgi.fr/balibase/ (BAliBASE4)
Thompson *et al.* (1999); Thompson *et al.* (2005)

➢ 9 reference alignment sets
➜ can be used to evaluate multiple alignment programs

- Reference 1: equidistant sequences with various levels of conservation
- Reference 2: families aligned with a highly divergent "orphan" sequence
- Reference 3: subfamilies with <25% residue identity between groups
- Reference 4: sequences with N/C-terminal extension
- Reference 5: internal insertions
- References 6, 7, 8: various protein families containing internal repeats, inversions, transmembrane regions, etc.
- References 9: linear motifs
- References 10: mixed

27

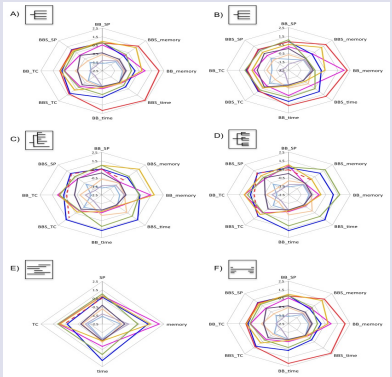
Other benchmark alignment database

- HOMSTRAD: Homologous Structure Alignment Database
  https://homstrad.mizuguchilab.org/homstrad/
  - ➜ A curated database of structure-based alignments for protein families
- PREFAB: Protein Reference Alignment Benchmark
  http://www.drive5.com/muscle/prefab.htm (MUSCLE website)
  - ➜ Automatically generated from structural pairwise alignment expanded with PSI-Blast
  - ➜ Collection of Benchmark alignment database is also available (BENCH)
    http://www.drive5.com/bench
- SABmark: Sequence and Structure Alignment Benchmark
  http://bioinformatics.vub.ac.be/databases/databases.html (no longer available?)
  - ➜ Structural alignments of fold groups
    (Twilight zone set, Superfamily set)

Edgar (2010) "Benchmark collection" (http://drive5.com/bench/)
Iantorno *et al.* (2014) Who watches the watchman? An appraisal of benchmarks for multiple sequence alignment.

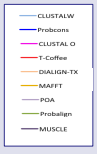
28

Multiple alignment quality



SP score:
Sum-of-Pairs score

TC score:
Total-Column score

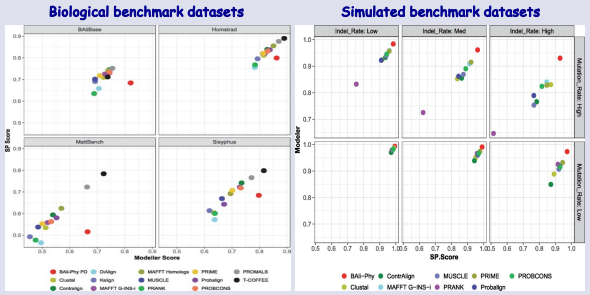


Pais *et al.* (2014)

29

Multiple alignment quality

Biological benchmark datasets
Simulated benchmark datasets

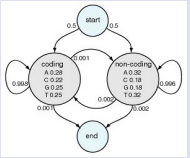


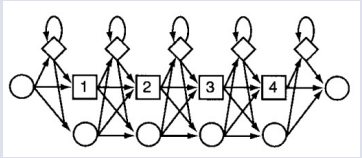
Nute *et al.* (2018)

SP score (recall) = How much of correct alignment is recovered
Modeler score (precision) = How much of the reconstructed alignment is correct

30

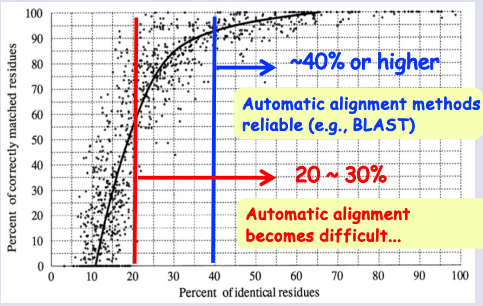
Patterns, profiles, and profile HMMs





31

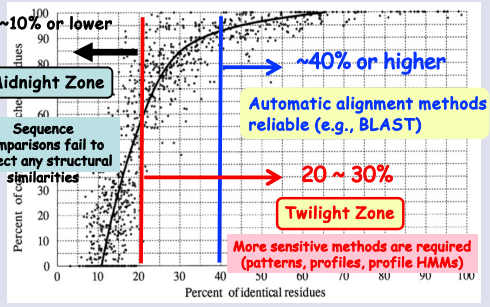
Sequence similarities and search methods



Vogt *et al.* (1995)

32

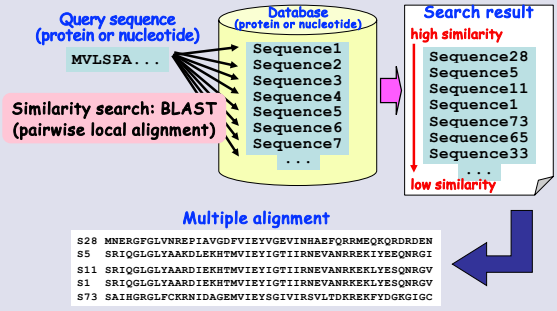
Sequence similarities and search methods



Vogt *et al.* (1995)

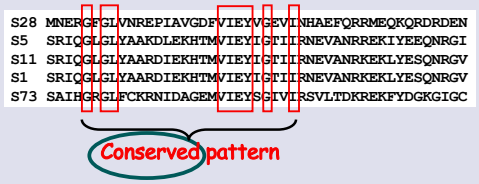
33

Alignments → patterns → functions



34

Alignments → patterns → functions



Higher functional constraint → Functionally important

Once a sequence pattern - function correspondence is established, we can use such relationships to predict functions based on sequences

35

How to identify conserved patterns

```
S28       MNERGFGLVNREPIAVGDFVIEYVGEVINHAEFQRRMEQKQRDRDEN
S5        SRIQGLGLYAAKDLEKHTMVIEYIGTIIRNEVANRREKIYEEQNRGI
S11       SRIQGLGLYAARDIEKHTMVIEYIGTIIRNEVANRKEKLYESQNRGV
S1        SRIQGLGLYAARDIEKHTMVIEYIGTIIRNEVANRKEKLYESQNRGV
S73       SAIHGRGLFCKRNIDAGEMVIEYSGIVIRSVLTDKREKFYDGKGIGC

Consensus SRIQGLGLYAARDIEKHTMVIEYIGTIIRNEVANRREK-YE-QNRG-
```

Majority-Rule Consensus Sequence

Residues probably not important for functions are included

36

How to identify conserved patterns

```
S28     MNERGFGLVNREPIAVGDFVIEYVGEVINHAEFQRRMEQKQRDRDEN
S5      SRIQGLGLYAAKDLEKHTMVIEYIGTIIRNEVANRREKIYEEQNRGI
S11     SRIQGLGLYAARDIEKHTMVIEYIGTIIRNEVANRKEKLYESQNRGV
S1      SRIQGLGLYAARDIEKHTMVIEYIGTIIRNEVANRKEKLYESQNRGV
S73     SAIHGRGLFCKRNIDAGEMVIEYSGIVIRSVLTDKREKFYDGKGIGC

Pattern      GxGLxxxxxxxxxxVIEYxGxxI
                                (x:any amino acid)
```

Conserved pattern including only identical sites
Very (too) strict

37

How to identify conserved patterns

```
S28     MNERGFGLVNREPIAVGDFVIEYVGEVINHAEFQRRMEQKQRDRDEN
S5      SRIQGLGLYAAKDLEKHTMVIEYIGTIIRNEVANRREKIYEEQNRGI
S11     SRIQGLGLYAARDIEKHTMVIEYIGTIIRNEVANRKEKLYESQNRGV
S1      SRIQGLGLYAARDIEKHTMVIEYIGTIIRNEVANRKEKLYESQNRGV
S73     SAIHGRGLFCKRNIDAGEMVIEYSGIVIRSVLTDKREKFYDGKGIGC
```

G-[FLR]-G-L-X10-[FM]-V-I-E-Y-[VIS]-G-[ETI]-[VI]-I
(10 any amino acids)

Conserved pattern with more flexibility

Regular expression

38

PROSITE Pattern Database

prosite Database of protein domains, families and functional sites

➜ consists of biologically significant sites, patterns, and profiles
https://prosite.expasy.org/

- PROSITE pattern syntax is described in:
https://prosite.expasy.org/prosuser.html - meth1

39

PROSITE: PS00237
[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-
[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM]

```
5H1A_FUGRU/131-147      SSIlh caIALDRYwaI
5H1A_HUMAN/122-138      SSIlh caIALDRYwaI
5H1A_MOUSE/122-138      SSIlh caIALDRYwaI
5H1A_PANTR/122-138      SSIlh caIALDRYwaI
5H1A_RAT/122-138        SSIlh caIALDRYwaI
5H1B_CAVPO/134-150      ASImh cvIALDRYwaI
5H1B_CRIGR/131-147      ASImh cvIALDRYwaI
5H1B_DIDMA/134-150      ASIlh cvIALDRYwaI
5H1B_FUGRU/119-135      SSIlh cvIALDRYwaI
5H1B_HUMAN/135-151      ASIlh cvIALDRYwaI
5H1B_MOUSE/131-147      ASImh cvIALDRYwaI
5H1B_HUMAN/135-151      ASIlh cvIALDRYwaI
5H1B_RABIT/135-151      ASImh cvIALDRYwaI
5H1B_RAT/131-147        ASImh cvIALDRYwaI
5H1B_SPAEH/131-147      ASImh cvIALDRYwaI
5H1D_CANFA/124-140      ASIlh cvIALDRYwaI
5H1D_CAVPO/124-140      ASIlh cvIALDRYwaI
5H1D_FUGRU/122-138      ASIlh cvIALDRYwaI
5H1D_HUMAN/124-140      ASIlh cvIALDRYwaI
5H1D_MOUSE/121-137      ASIlh cvIALDRYwaI
5H1D_PIG/44-60          ASIlh cvIALDRYwaI
5H1D_RABIT/124-140      ASIlh cvIALDRYwaI
5H1D_RAT/121-137        ASIlh cvIALDRYwaI
5H1E_HUMAN/108-124      CSIlh cvIALDRYwaI
5H1E_PANTR/108-124      CSIlh cvIALDRYwaI
```

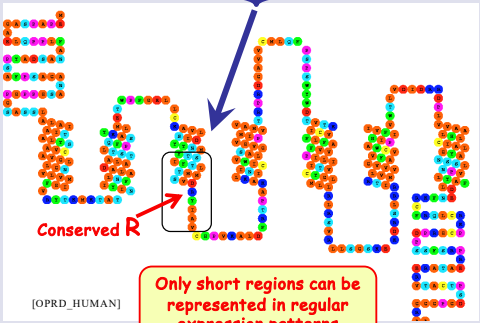

40

PROSITE: PS00237
[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-
[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM]

```
5H1A_FUGRU/131-147      SSIlh caIALDRYwaI
5H1A_HUMAN/122-138      SSIlh caIALDRYwaI
5H1A_MOUSE/122-138      SSIlh caIALDRYwaI
5H1A_PANTR/122-138      SSIlh caIALDRYwaI
5H1A_RAT/122-138        SSIlh caIALDRYwaI
5H1B_CAVPO/134-150      ASImh cvIALDRYwaI
5H1B_CRIGR/131-147      ASImh cvIALDRYwaI
5H1B_DIDMA/134-150      ASIlh cvIALDRYwaI
5H1B_FUGRU/119-135      SSIlh cvIALDRYwaI
5H1B_HUMAN/135-151      ASIlh cvIALDRYwaI
5H1B_MOUSE/131-147      ASImh cvIALDRYwaI
5H1B_HUMAN/135-151      ASIlh cvIALDRYwaI
5H1B_RABIT/135-151      ASImh cvIALDRYwaI
5H1B_RAT/131-147        ASImh cvIALDRYwaI
5H1B_SPAEH/131-147      ASImh cvIALDRYwaI
5H1D_CANFA/124-140      ASIlh cvIALDRYwaI
5H1D_CAVPO/124-140      ASIlh cvIALDRYwaI
5H1D_FUGRU/122-138      ASIlh cvIALDRYwaI
5H1D_HUMAN/124-140      ASIlh cvIALDRYwaI
5H1D_MOUSE/121-137      ASIlh cvIALDRYwaI
5H1D_PIG/44-60          ASIlh cvIALDRYwaI
5H1D_RABIT/124-140      ASIlh cvIALDRYwaI
5H1D_RAT/121-137        ASIlh cvIALDRYwaI
5H1E_HUMAN/108-124      CSIlh cvIALDRYwaI
5H1E_PANTR/108-124      CSIlh cvIALDRYwaI
```


41

PROSITE: PS00237
[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-
[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM]



Only short regions can be represented in regular expression patterns

42

PRINTS

A compendium of protein fingerprints

PRINTS entry: 5HT1ARECEPTR

5HT1ARECEPTR1
GQGNNTTASQEPFGTGG
GQGNNTTTSLEPFGTGG
GQGNNTTSPPAPFETGG

5HT1ARECEPTR2
GNVTSISDVTFSYQVITS
GNDTGLSNVTFSYQVITS
GNTTGISDVTVSYQVITS

5HT1ARECEPTR3
RTPEDRSDPDACTISK
RAPEDRSNPNECTISK
RTPEDRSDPDACTISK

5HT1ARECEPTR4
FRIRKTVRKVEKKGAGTSLG
FRIRKTVKKVEKKGAGTSFG
FRIRKTVKKVEKTGADTRHG

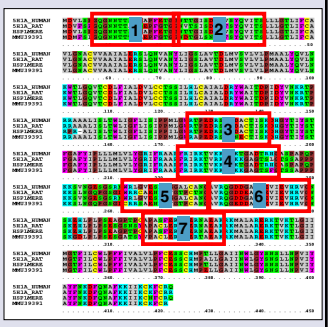
5HT1ARECEPTR5
WRRCAENRAVGTPCTNG
CRRSAENRAVGTPCANG
WRLGVESKAGGALCANG

5HT1ARECEPTR6
AVRQGDDEATLEVIEVHRVG
AVRQGEDDATLEVIEVHRVG
AVRQGDDGAALEVIEVHRVG

5HT1ARECEPTR7
APACLERKNERNAEAK
VPACLERKNERTAEAK
APASFERKNERNAEAK

7 motifs

Fingerprint:
➜ a group of conserved motifs used to characterize a protein family
➜ Part of InterPro



43

Profile (PSSM)

- Profile: Position Specific Scoring Matrix (PSSM)
  - ➜ Constructed from multiple alignments
    - Short conserved domains (PRINTS, BLOCKS)
    - Protein families (PROSITE)
    - Results of similarity search (PSI-BLAST)
- More flexible than simple patterns
  - ➜ PSSM lists amino acid frequencies for each alignment position
- Profiles (PSSMs) can be used for database search to identify remote similarities

44

Profile: Position Specific Scoring Matrix

- EMBOSS: http://bio.biomedicine.gu.se/emboss/ (Protein Profiles)
  - PROPHECY: creates profiles from multiple alignment
  - PROFIT: scans a sequence/database with a profile
  - PROPHET: aligns a profile with sequence(s)

```
# Columns are amino acid counts A->Z
# Rows are alignment positions 1->n
Simple
Name            mymatrix
Length          5
Maximum score   11
Thresh          75
Consensus       RCEGH
1 0  0  0  0  0  0  1  0  ... 0  ... 0  2  ...
2 0  0  3  0  0  0  0  0  0  ... 0  ... 0  0  ...
3 0  0  0  0  2  0  0  0  0  ... 0  ... 1  0  ...
4 1  0  0  0  0  0  2  0  0  ... 0  ... 0  0  ...
5 0  0  0  0  0  0  0  2  0  ... 1  ... 0  0  ...
  A  B  C  D  E  F  G  H  I  ... N  ... Q  R  ...
```

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

Simple amino acid frequency

(Asx) → Asn & Asp

45

Profile: Position Specific Scoring Matrix

- EMBOSS: http://bio.biomedicine.gu.se/emboss/ (Protein Profiles)
  - PROPHECY: creates profiles from multiple alignment
  - PROFIT: scans a sequence/database with a profile
  - PROPHET: aligns a profile with sequence(s)

```
# Gribskov Protein Profile
# Columns are amino acids A->Z
# Last column is indel penalty
# Rows are alignment positions 1->n
...
Consensus       RCEGH
1 -2.33 0.00 -2.33  1.33  1.33 -3.67 -2.67  8.33 ... 11.67
2  3.00 0.00 15.00 -5.00 -6.00 -1.00  2.00 -1.00 ... -3.00
3  2.67 0.00 -6.00  8.67 12.00 -6.67  4.00  4.67 ...  1.33
4  9.00 0.00  2.33  5.00  4.33 -5.67 12.00 -1.67 ... -3.00
5  0.00 0.00 -1.67  4.67  4.33 -2.33  0.00 11.67 ...  3.67
     A    B    C     D     E     F     G     H   ...  R
          (Asx)
```

Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN

Non-zero values are used even for the AAs that are not found.

46