BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 | Lecture 17

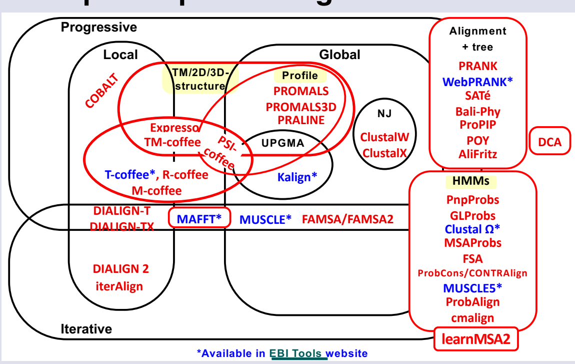
1

Today's topics

- Phylogeny-aware gap placement methods (PRANK, etc.)
- Alignment trimming/filtering
- MSA evaluation
- Conserved domain, pattern, profile
  - Pattern and Profile (PSSM)
- Assignment 8

2

Multiple sequence alignment methods



For reviews: Katoh (2021) and more available on CANVAS

3

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."

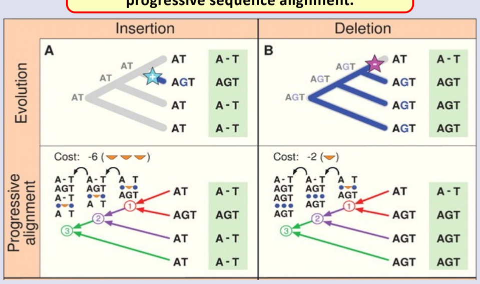
"There is an understandable tendency for users of multiple alignment software to want their residues neatly aligned in blocks and columns. This is fine when such blocks are biologically accurate as will happen in parts of protein alignments. In cases where insertions or deletions have happened in a less 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."

4

PRANK, WebPRANK

Löytynoja & Goldman (2008)

Insertions are more penalized than deletions in progressive sequence alignment.



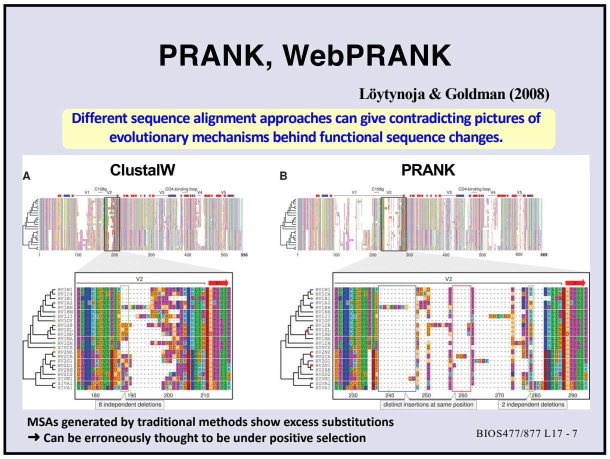
5

PRANK, WebPRANK

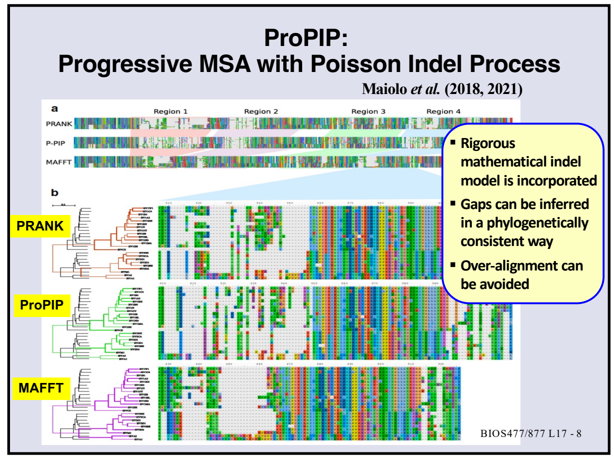
Available at EBI website   Löytynoja & Goldman (2005, 2008, 2010)

- 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.

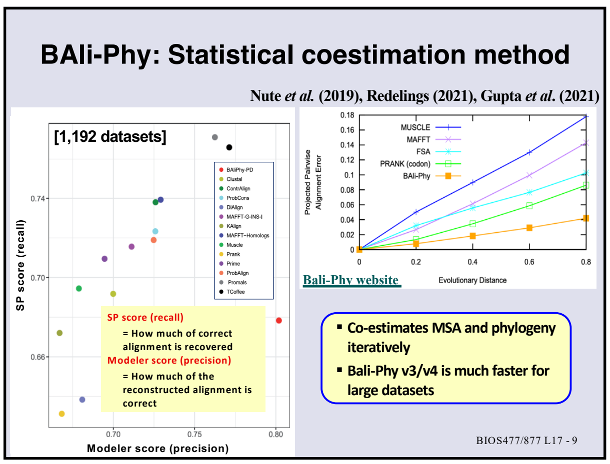
6



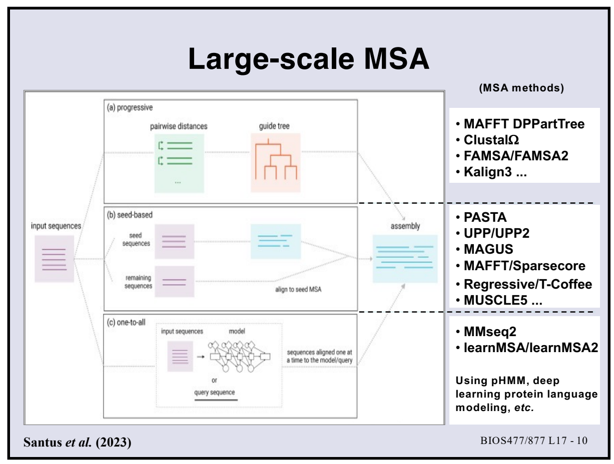
7



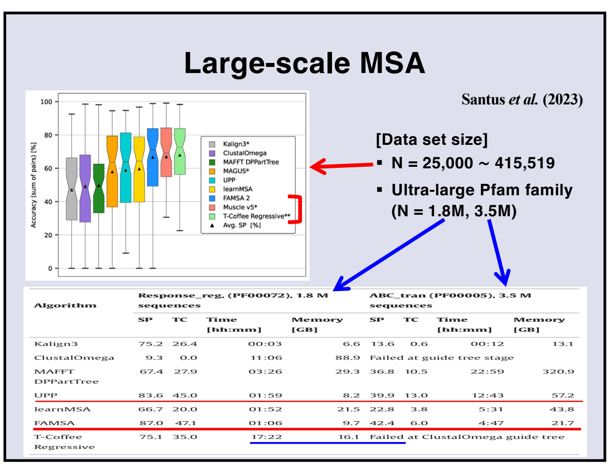
8



9

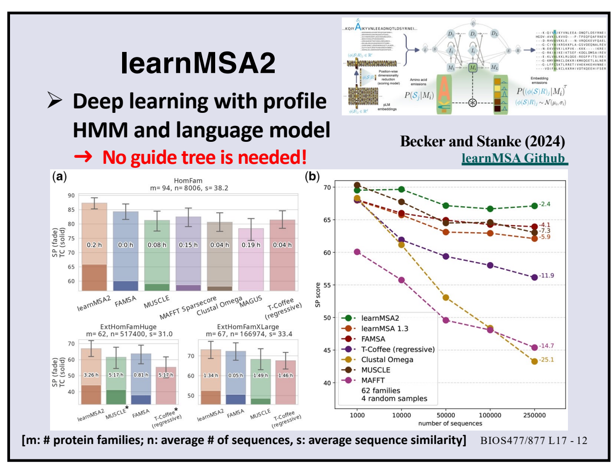


10

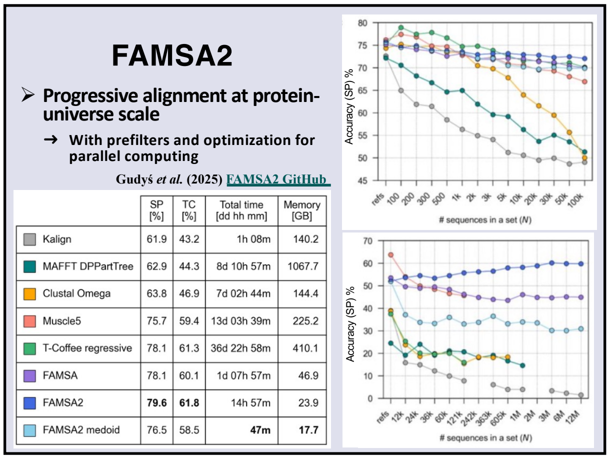


Algorithm	Response_reg. (PF00072), 1.8 M sequences				ABC_tran (PF00005), 3.5 M sequences			
	SP	TC	Time [hh:mm]	Memory [GB]	SP	TC	Time [hh:mm]	Memory [GB]
Kalign3	75.2	26.4	00:03	6.6	13.6	0.6	00:12	13.1
ClustalOmega	9.3	0.0	11:06	88.9	Failed at guide tree stage			
MAFFT DPPartTree	67.4	27.9	03:26	29.3	36.8	10.5	22:59	320.9
UPP	83.6	45.0	01:59	8.2	39.9	13.0	12:43	57.2
learnMSA	66.7	20.0	01:52	21.5	22.8	3.8	5:31	43.8
FAMSA	87.0	47.1	01:06	9.7	42.4	6.0	4:47	21.7
T-Coffee Regressive	75.1	35.0	17:22	16.1	Failed at ClustalOmega guide tree			

11

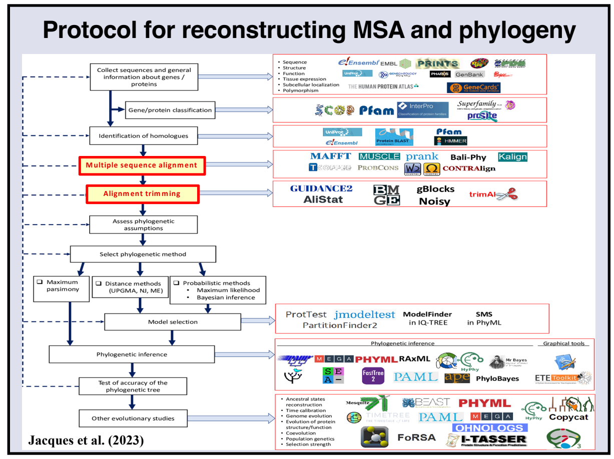


12

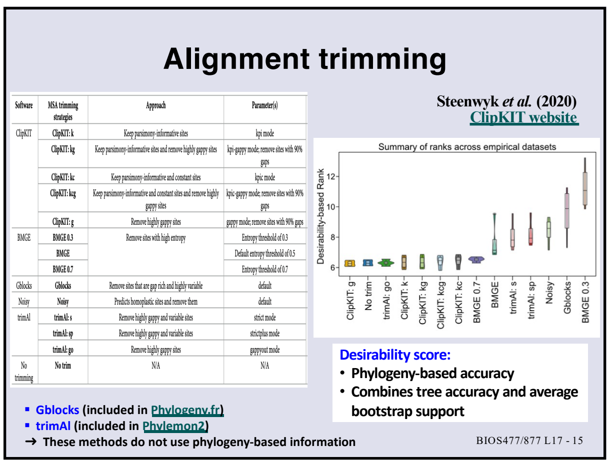


	SP [%]	TC [%]	Total time [dd hh mm]	Memory [GB]
Kalign	61.9	43.2	1h 08m	140.2
MAFFT DPPartTree	62.9	44.3	8d 10h 57m	1067.7
Clustal Omega	63.8	46.9	7d 02h 44m	144.4
Muscle5	75.7	59.4	13d 03h 39m	225.2
T-Coffee regressive	78.1	61.3	36d 22h 58m	410.1
FAMSA	78.1	60.1	1d 07h 57m	46.9
FAMSA2	**79.6**	**61.8**	14h 57m	23.9
FAMSA2 medoid	76.5	58.5	**47m**	**17.7**

13

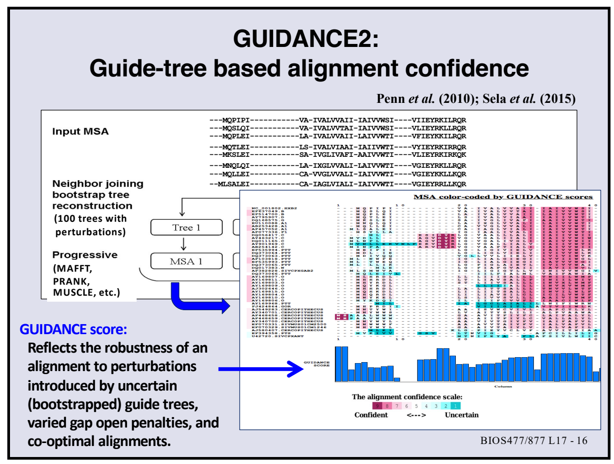


14



Software	MSA trimming strategies	Approach	Parameter(s)
ClipKIT	ClipKIT: k	Keep parsimony-informative sites	kpi mode
	ClipKIT: kg	Keep parsimony-informative sites and remove highly gappy sites	kpi-gappy mode; remove sites with 90% gaps
	ClipKIT: kc	Keep parsimony-informative and constant sites	kpic mode
	ClipKIT: kcg	Keep parsimony-informative and constant sites and remove highly gappy sites	kpic-gappy mode; remove sites with 90% gaps
	ClipKIT: g	Remove highly gappy sites	gappy mode; remove sites with 90% gaps
BMGE	BMGE 0.3	Remove sites with high entropy	Entropy threshold of 0.3
	BMGE		Default entropy threshold of 0.5
	BMGE 0.7		Entropy threshold of 0.7
Gblocks	Gblocks	Remove sites that are gap rich and highly variable	default
Noisy	Noisy	Predicts homoplastic sites and remove them	default
trimAl	trimAl: s	Remove highly gappy and variable sites	strict mode
	trimAl: sp	Remove highly gappy and variable sites	strictplus mode
	trimAl: go	Remove highly gappy sites	gappyout mode
No trimming	No trim	N/A	N/A

15



16

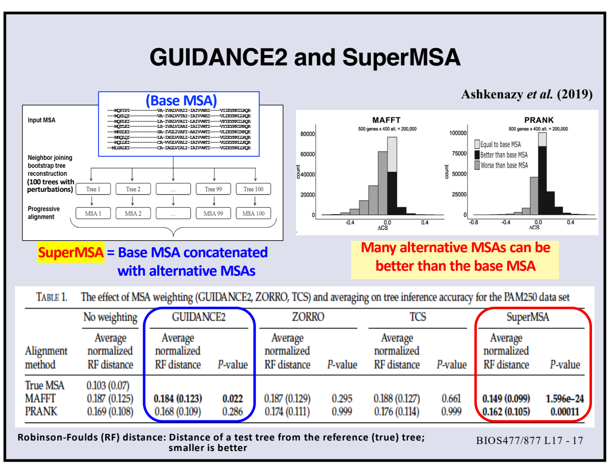
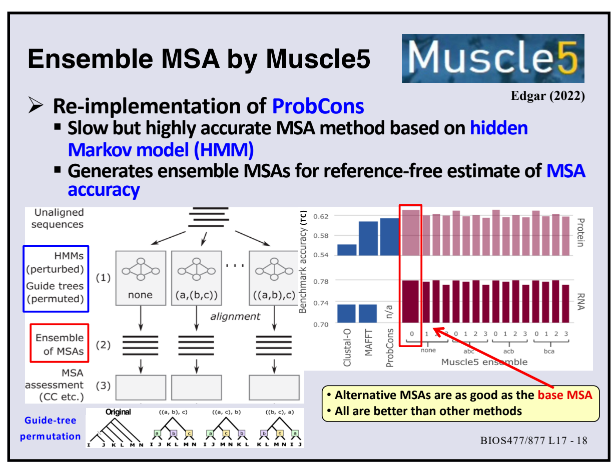


TABLE 1. The effect of MSA weighting (GUIDANCE2, ZORRO, TCS) and averaging on tree inference accuracy for the PAM250 data set

Alignment method	No weighting	GUIDANCE2		ZORRO		TCS		SuperMSA	
	Average normalized RF distance	Average normalized RF distance	P-value	Average normalized RF distance	P-value	Average normalized RF distance	P-value	Average normalized RF distance	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**

17

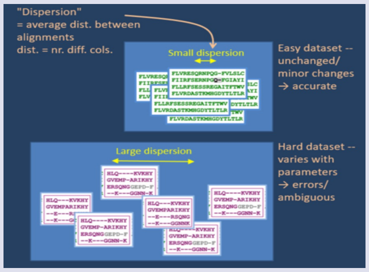


18

Ensemble MSA by Muscle5

"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. "

Edgar (2022)



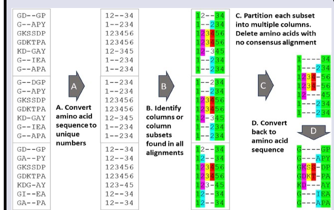
- Easy alignment
  - → Fewer alignment errors
  - → Ensemble MSA less dispersed
- Difficult alignment
  - → More alignment errors
  - → Ensemble MSA more dispersed

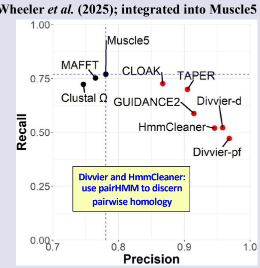
BIOS477/877 L17 - 19

19

Muscle5 + CLOAK: Cleaning on Alignment (K)onsensus

Wheeler *et al.* (2025); integrated into Muscle5





- Likely alignment errors are identified as departure from consensus among alternative MSAs (Muscle5 ensemble MSAs)
  - → Gentle but effective removal of likely alignment errors
- Recall: proportion of aligned pairs in the reference MSA recovered
- Precision: proportion of aligned pairs that are correct

BIOS477/877 L17 - 20

20

How to measure MSA quality

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

Ref	Test	
VA-T	-VAT-	TCS=1/3 =33.3% or 25% (w/ gap col.)
VA-G	-VAG-	
MGTG	M-GTG	SPS=5/9 =55.6%

- 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
- Muscle5: ensemble MSA based accuracy https://www.drive5.com/muscle/
- AlignStat: MSA similarity/dissimilarity https://github.com/TS404/AlignStat
- QuanTest2: secondary structure prediction based http://www.bioinf.ucd.ie/download/QuanTest.tgz (download only)
- Sequence logo: graphical representation of a multiple alignment
  - Weblogo 3: https://weblogo.threeplusone.com/

BIOS477/877 L17 - 21

21

Benchmark alignment database: BAliBASE

Thompson *et al.* (1999); Bahr *et al.* (2001); Thompson *et al.* (2005)
BAliBASE website

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

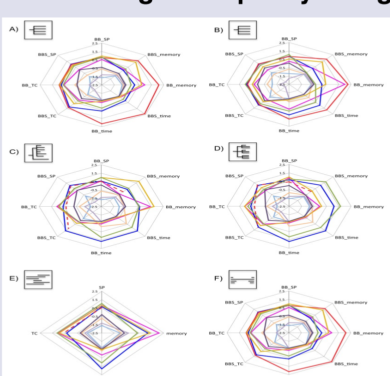
- Reference 1: equidistant sequences with various levels of conservation
- Reference 2: families aligned with a highly divergent "orphan" sequence
- Reference 3: subfamilies with <25% residue identity between groups
- Reference 4: sequences with N/C-terminal extension
- Reference 5: internal insertions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

BIOS477/877 L17 - 22

22

Assessing MSA quality using BAliBASE

Pais *et al.* (2014)



SP score: Sum-of-Pairs score

TC score: Total-Column score

BBS: Short version of BAliBASE

BIOS477/877 L17 - 23

23

Other benchmark alignment databases

- HOMSTRAD: Homologous Structure Alignment Database
  - → Curated database of structure-based alignments for protein families https://homstrad.mizuguchilab.org/homstrad/
- PREFAB: Protein Reference Alignment Benchmark
  - → Automatically generated from structural pairwise alignment expanded with PSI-Blast
  - → Collection of benchmark alignment database is also available (BENCH)

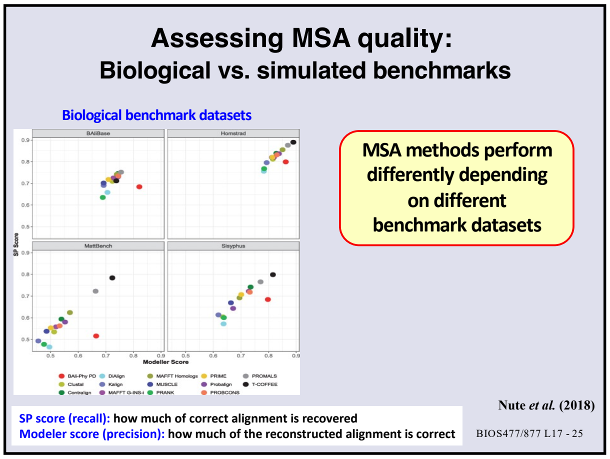
BAliBASE, HOMSTRAD, and PREFAB are all
→ biological benchmark datasets (actual protein sequences)
→ based on protein structural alignment
→ supposed to be highly accurate

[References]

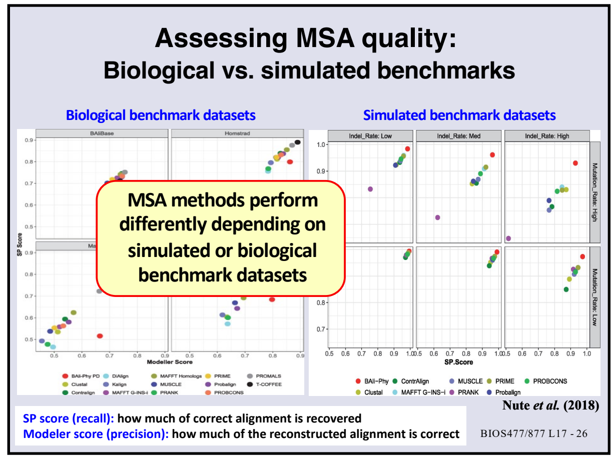
- Edgar (2010) "Quality measures for protein alignment benchmarks" BENCH @ Muscle website
- Iantorno et al. (2014) "Who watches the watchman? An appraisal of benchmarks for multiple sequence alignment."
- Warnow (2021) "Revisiting evaluation of multiple sequence alignment methods"
  - → Discusses the challenges in evaluating MSA methods using biological vs. simulated benchmark datasets

BIOS477/877 L17 - 24

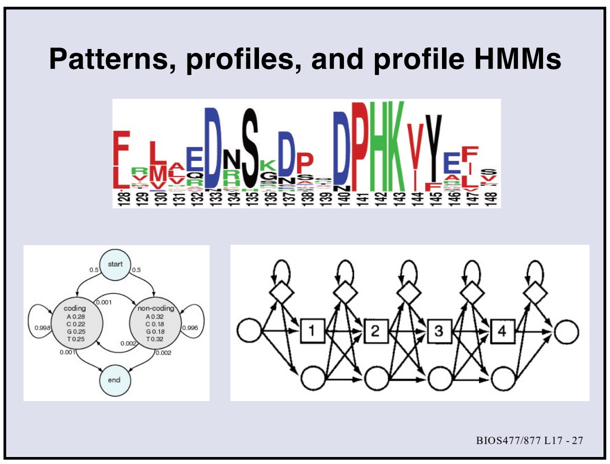
24



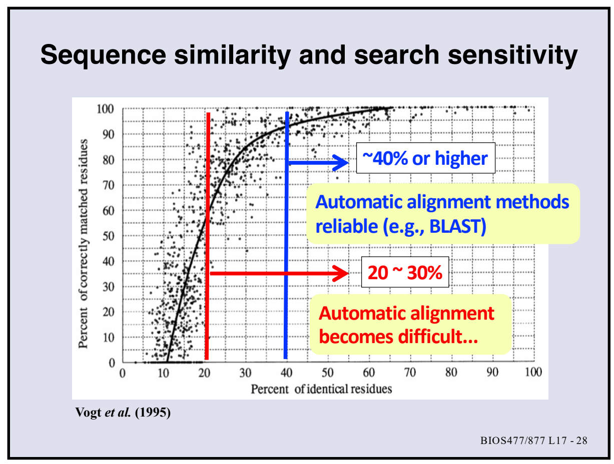
25



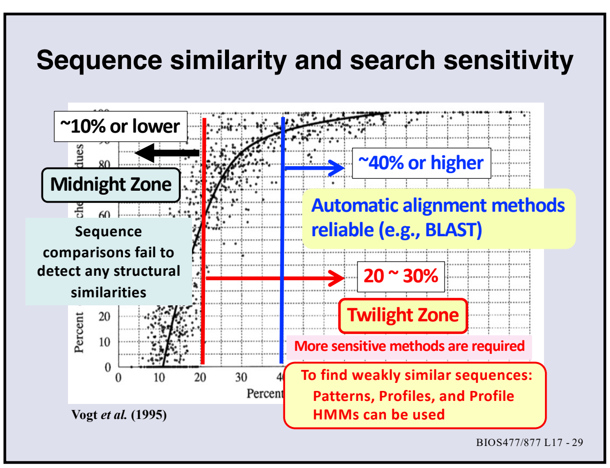
26



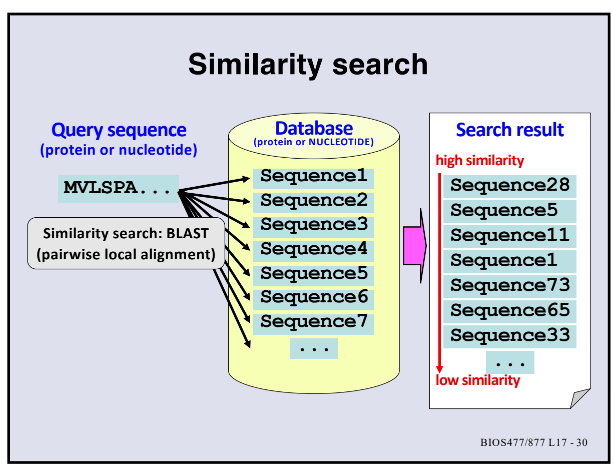
27



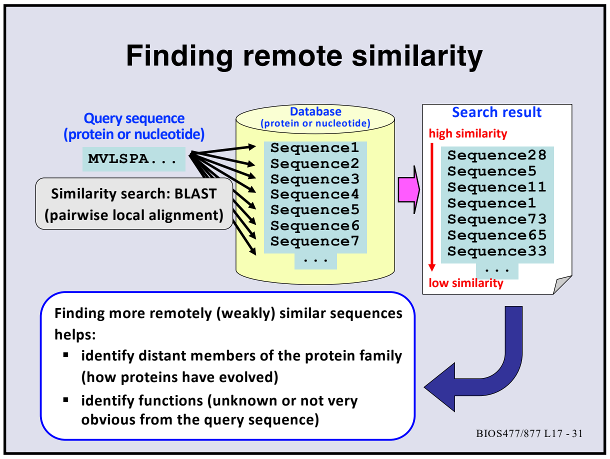
28



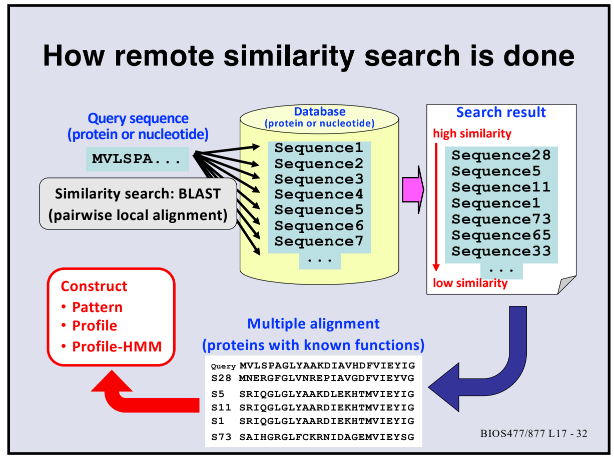
29



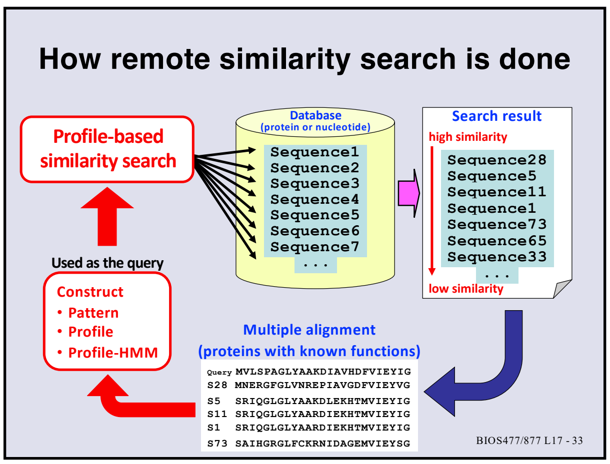
30



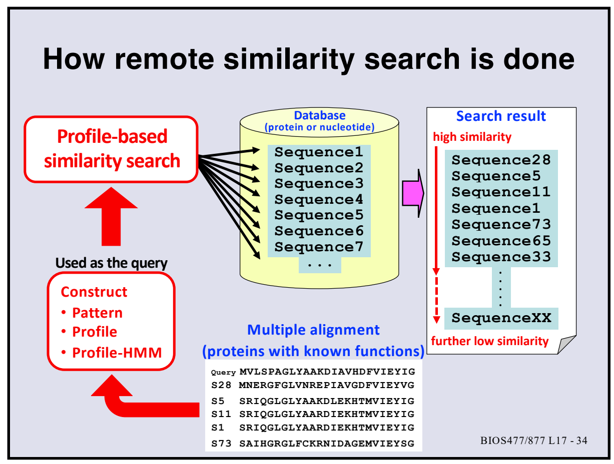
31



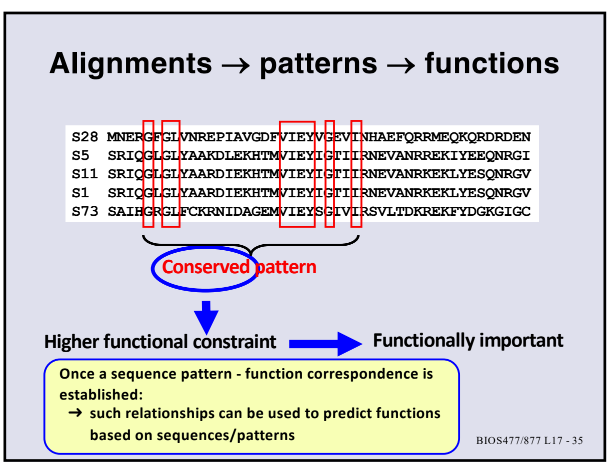
32



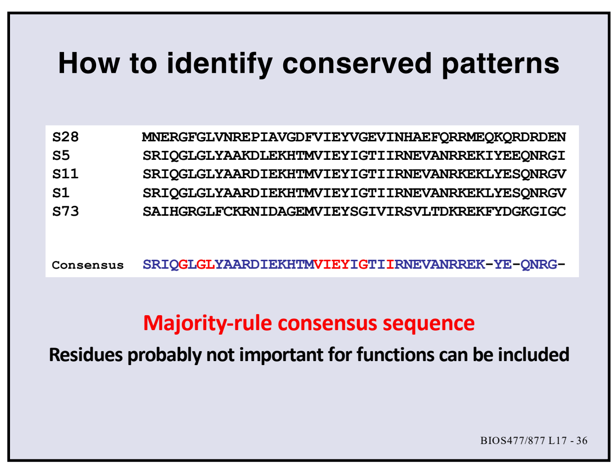
33



34



35



36

Conserved pattern

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

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

Conserved pattern including only identical sites

➜ **Very (too) strict**

BIOS477/877 L17 - 37

37

Reguler expression pattern

```
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)

Regular expression

➜ **More flexile than strict conserved pattern**

BIOS477/877 L17 - 38

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

BIOS477/877 L17 - 39

39

PROSITE Pattern

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)
[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM]

```
5H1A_FUGRU/131-147     SSIlhLcaIALDRYwaI
5H1A_HUMAN/122-138     SSIlhLcaIALDRYwaI
5H1A_MOUSE/122-138     SSIlhLcaIALDRYwaI
5H1A_PANTR/122-138     SSIlhLcaIALDRYwaI
5H1A_RAT/122-138       SSIlhLcaIALDRYwaI
5H1B_CAVPO/134-150     ASImhLcvIALDRYwaI
5H1B_CRIGR/131-147     ASImhLcvIALDRYwaI
5H1B_DIDMA/134-150     ASIlhLcvIALDRYwaI
5H1B_FUGRU/119-135     SSIlhLcvIALDRYwaI
5H1B_HUMAN/135-151     ASIlhLcvIALDRYwaI
5H1B_MOUSE/131-147     ASImhLcvIALDRYwaI
5H1B_HUMAN/135-151     ASIlhLcvIALDRYwaI
5H1B_RABIT/135-151     ASImhLcvIALDRYwaI
5H1B_RAT/131-147       ASImhLcvIALDRYwaI
5H1B_SPAEH/131-147     ASImhLcvIALDRYwaI
5H1D_CANFA/124-140     ASIlhLcvIALDRYwaI
5H1D_CAVPO/124-140     ASIlhLcvIALDRYwaI
5H1D_FUGRU/122-138     ASIlhLcvIALDRYwaI
5H1D_HUMAN/124-140     ASIlhLcvIALDRYwaI
5H1D_MOUSE/121-137     ASIlhLcvIALDRYwaI
5H1D_PIG/44-60         ASIlhLcvIALDRYwaI
5H1D_RABIT/124-140     ASIlhLcvIALDRYwaI
5H1D_RAT/121-137       ASIlhLcvIALDRYwaI
5H1E_HUMAN/108-124     CSIlhLcvIALDRYwaI
5H1E_PANTR/108-124     CSIlhLcvIALDRYwaI
5H1E_PIG/55-71         CSIlhLcvIALDRYwaI
```

BIOS477/877 L17 - 40

40

PROSITE Pattern

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)
[GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R [FYWCSH]-x(2)- [LIVM]

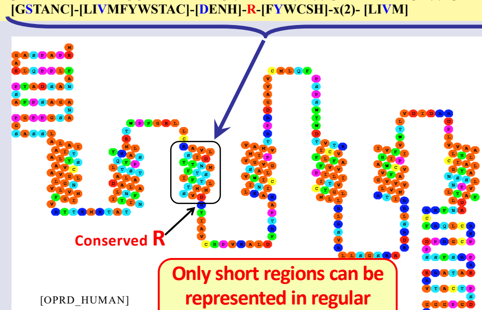
```
5H1A_FUGRU/131-147  →  SSIlhLcaIALDRYwaI
5H1A_HUMAN/122-138     SSIlhLcaIALDRYwaI
5H1A_MOUSE/122-138     SSIlhLcaIALDRYwaI
5H1A_PANTR/122-138     SSIlhLcaIALDRYwaI
5H1A_RAT/122-138       SSIlhLcaIALDRYwaI
5H1B_CAVPO/134-150     ASImhLcvIALDRYwaI
5H1B_CRIGR/131-147     ASImhLcvIALDRYwaI
5H1B_DIDMA/134-150     ASIlhLcvIALDRYwaI
5H1B_FUGRU/119-135     SSIlhLcvIALDRYwaI
5H1B_HUMAN/135-151     ASIlhLcvIALDRYwaI
5H1B_MOUSE/131-147     ASImhLcvIALDRYwaI
5H1B_HUMAN/135-151     ASIlhLcvIALDRYwaI
5H1B_RABIT/135-151     ASImhLcvIALDRYwaI
5H1B_RAT/131-147       ASImhLcvIALDRYwaI
5H1B_SPAEH/131-147     ASImhLcvIALDRYwaI
5H1D_CANFA/124-140     ASIlhLcvIALDRYwaI
5H1D_CAVPO/124-140     ASIlhLcvIALDRYwaI
5H1D_FUGRU/122-138     ASIlhLcvIALDRYwaI
5H1D_HUMAN/124-140     ASIlhLcvIALDRYwaI
5H1D_MOUSE/121-137     ASIlhLcvIALDRYwaI
5H1D_PIG/44-60         ASIlhLcvIALDRYwaI
5H1D_RABIT/124-140     ASIlhLcvIALDRYwaI
5H1D_RAT/121-137       ASIlhLcvIALDRYwaI
5H1E_HUMAN/108-124     CSIlhLcvIALDRYwaI
5H1E_PANTR/108-124     CSIlhLcvIALDRYwaI
5H1E_PIG/55-71         CSIlhLcvIALDRYwaI
```

BIOS477/877 L17 - 41

41

PROSITE Pattern

PROSITE: PS00237 (G-protein coupled receptors family 1 signature)
[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

BIOS477/877 L17 - 42

42

Profile (PSSM)

- Profile: Position Specific Scoring Matrix (PSSM)
  - → Constructed from multiple alignments
    - Short conserved domains (BLOCKS, PRINTS)
    - 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

BIOS477/877 L17 - 43

43

How to build a profile (PSSM)

- EMBOSS "Protein Profile" tools — EMBOSS website
  - PROPHECY: creates profiles from multiple alignment [option]
    - Simple amino acid "frequency"

```
# 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 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 ...
```

(Asx) → Asn & Asp

```
Seq1  RCQAH
Seq2  HCEGH
Seq3  RCEGN
```

BIOS477/877 L17 - 44

44

How to build a profile (PSSM)

- EMBOSS "Protein Profile" tools — EMBOSS website
  - PROPHECY: creates profiles from multiple alignment [option]
    - Gribskov protein profile

```
# 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!

BIOS477/877 L17 - 45

45

Gribskov profile

(Gribskov et al. 1987)

- Used in PROSITE profile
- Weighted scoring matrix for each position
  - → Generated from:
    - Multiple alignment
    - Scoring matrix (*e.g.*, BLOSUM62): $Y(a,b)$
  - → For amino acid a at position p,

$$M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$$

$Y(a,b)$: Value in the scoring matrix for AA pair a and b
$W(p,b)$: Frequency of amino acid b in position p

BIOS477/877 L17 - 46

46

Gribskov profile

(Gribskov et al. 1987)

$$M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$$ for amino acid a at position p

Multiple alignment

```
Seq1  RCQAH
Seq2  HCEGH
Seq3  RCEGN
```

$W(1,b)$: frequency of each AA at pos. 1

$$M(1,A) = \sum_{b=1}^{20} W(1,b) * Y(A,b)$$

For amino acid A at position 1

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2

```
1 -2.33 0.00 -2.33  1.33 ...  8.33 ... 11.67 ...
2  3.00 0.00 15.00 -5.00 ... -1.00 ... -3.00 ...
3  2.67 0.00 -6.00  8.67 ...  4.67 ...  1.33 ...
4  9.00 0.00  2.33  5.00 ... -1.67 ... -3.00 ...
5  0.00 0.00 -1.67  4.67 ... 11.67 ...  3.67 ...
     A    B     C     D  ...    H  ...    R  ...
```

(Asx)

BIOS477/877 L17 - 47

47

Gribskov profile

(Gribskov et al. 1987)

$$M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$$ for amino acid a at position p

Multiple alignment

```
Seq1  RCQAH
Seq2  HCEGH
Seq3  RCEGN
```

$$M(1,A) = \sum_{b=1}^{20} W(1,b) * Y(A,b)$$

$W(1,A) = 0$
Frequency of Ala (b=1) at pos. 1

$Y(A,A) = 4$ (for b=1)

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2

$M(1,A) = 0 * 4 + W(1,R)*Y(A,R) +$
$W(1,N)*Y(A,N) + W(1,D)*Y(A,D) +$
$... + W(1,H)*Y(A,H) + W(1,I)*Y(A,I) +$
... [sum up for all 20 amino acids]

```
1 -2.33 0.00 -2.33  1.33 ...  8.33 ... 11.67 ...
2  3.00 0.00 15.00 -5.00 ... -1.00 ... -3.00 ...
3  2.67 0.00 -6.00  8.67 ...  4.67 ...  1.33 ...
4  9.00 0.00  2.33  5.00 ... -1.67 ... -3.00 ...
5  0.00 0.00 -1.67  4.67 ... 11.67 ...  3.67 ...
     A    B     C     D  ...    H  ...    R  ...
```

BIOS477/877 L17 - 48

48

Gribskov profile

$$M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$$ for amino acid *a* at position *p*

(Gribskov et al. 1987)

Multiple alignment

```
Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN
```

$M(1,A) = \sum_{b=1}^{20} W(1,b) * Y(A,b)$

$W(1,R) = 2/3$
Frequency of Arg (b=2) at pos. 1

$Y(A,R) = -1$ for b=2

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2

$M(1,A) = 0 * 4 + 0.67 * (-1) + 0 * (-2) + 0 * (-2) + ... + 0.33 * (-2) + 0 * (-1) + ...$ [sum up for all 20 amino acids]

1	-2.33	0.00	-2.33	1.33	...	8.33	...	11.67	...
2	3.00	0.00	15.00	-5.00	...	-1.00	...	-3.00	...
3	2.67	0.00	-6.00	8.67	...	4.67	...	1.33	...
4	9.00	0.00	2.33	5.00	...	-1.67	...	-3.00	...
5	0.00	0.00	-1.67	4.67	...	11.67	...	3.67	...
	A	B	C	D	...	H	...	R	...

BIOS477/877 L17 - 49

49

Gribskov profile

$$M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$$ for amino acid *a* at position *p*

(Gribskov et al. 1987)

But many amino acids have 0 counts! Should we just ignore them?

Multiple alignment

```
Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN
```

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2

$M(1,A) = 0 * 4 + 0.67 * (-1) + 0 * (-2) + 0 * (-2) + ... + 0.33 * (-2) + 0 * (-1) + ...$ [sum up for all 20 amino acids]

1	-2.33	0.00	-2.33	1.33	...	8.33	...	11.67	...
2	3.00	0.00	15.00	-5.00	...	-1.00	...	-3.00	...
3	2.67	0.00	-6.00	8.67	...	4.67	...	1.33	...
4	9.00	0.00	2.33	5.00	...	-1.67	...	-3.00	...
5	0.00	0.00	-1.67	4.67	...	11.67	...	3.67	...
	A	B	C	D	...	H	...	R	...

BIOS477/877 L17 - 50

50

Gribskov profile

$$M(p,a) = \sum_{b=1}^{20} W(p,b) * Y(a,b)$$ for amino acid *a* at position *p*

(Gribskov et al. 1987)

Instead of 0, a very small weight (e.g., 0.025/#seq) can be used.

Multiple alignment

```
Seq1 RCQAH
Seq2 HCEGH
Seq3 RCEGN
```

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2

Why not 0?

$M(1,A) = 0.008*4 + 0.67*(-1) + 0.008*(-2) + 0.008*(-2) + ... + 0.33*(-2) + 0.008*(-1) + ...$ [sum up for all 20 amino acids]

1	-2.33	0.00	-2.33	1.33	...	8.33	...	11.67	...
2	3.00	0.00	15.00	-5.00	...	-1.00	...	-3.00	...
3	2.67	0.00	-6.00	8.67	...	4.67	...	1.33	...
4	9.00	0.00	2.33	5.00	...	-1.67	...	-3.00	...
5	0.00	0.00	-1.67	4.67	...	11.67	...	3.67	...
	A	B	C	D	...	H	...	R	...

BIOS477/877 L17 - 51

51