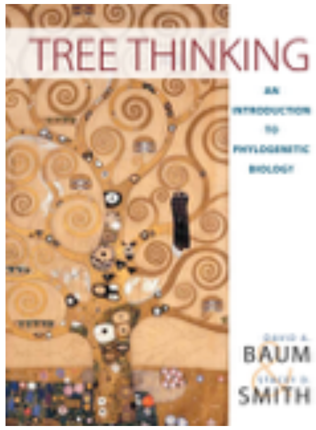
Alignment
Phylogenetic Biology - Week 5

Biology 1425
Professor: Casey Dunn, dunnlab.org
Brown University

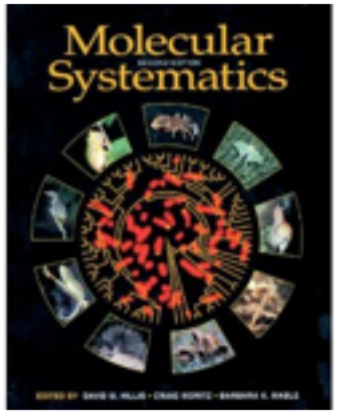
Front matter...

Sources

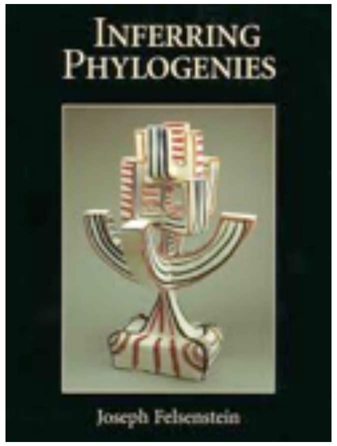
Some non-original content is drawn from:



Baum, D and S. Smith (2012) Tree Thinking: and Introduction to Phylogenetic Biology. Roberts and Company Publishers. ISBN 9781936221165



Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1996). Phylogenetic inference. In: Molecular Systematics, Second Edition. eds: D. M. Hillis, C Moritz, & B. K. Mable. Sinauer Associates. ISBN 9780878932825



Felsenstein, J. (2003) Inferring Phylogenies. Sinauer Associates. ISBN 978-0878931774

Other non-original content is referenced by url.

What is sequence alignment?

The identification of homologous sites in molecular sequence data.

If sequences didn’t evolve and could be observed error free, we could just look for identical sequence regions.

Due to evolution, sequence error, and analysis error, we have to ask - How do we know when the same site in two different sequences is homologous?

Alignment

Reconciles differences that arise from two processes:

- Substitution
- Insertion/deletion (ie, indels)

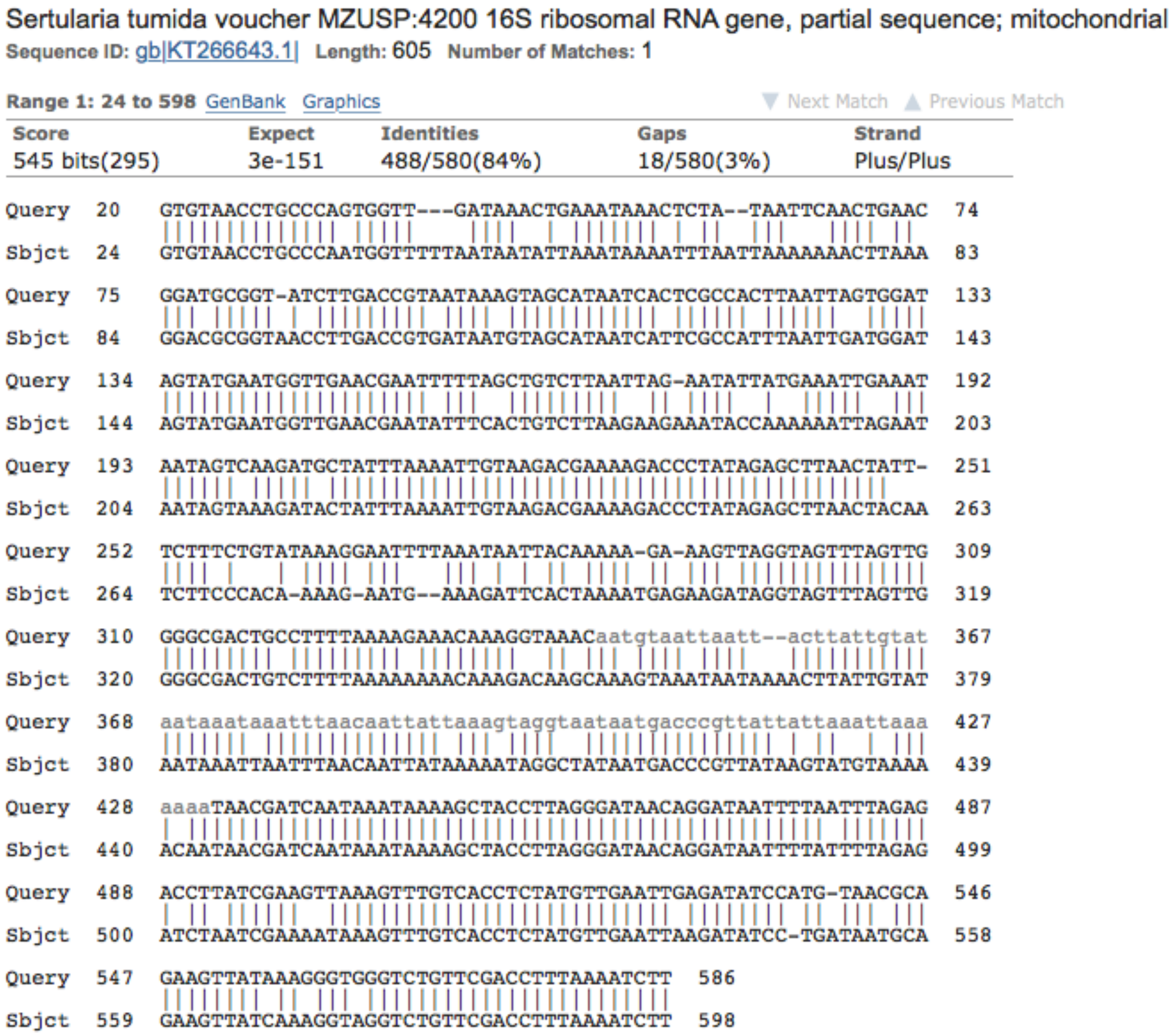
Many applications of alignment

Pairwise sequence alignment (eg blast) to find homologous sequences

Alignment of raw sequence reads to a reference sequence to identify variants

Multiple sequence alignment to build character matrices

Pairwise alignment



Sertularia tumida voucher MZUSP:4200 16S ribosomal RNA gene, partial sequence; mitochondrial

Sequence ID: gb|KT266643.1| **Length:** 605 **Number of Matches:** 1

Range 1: 24 to 598 GenBank Graphics

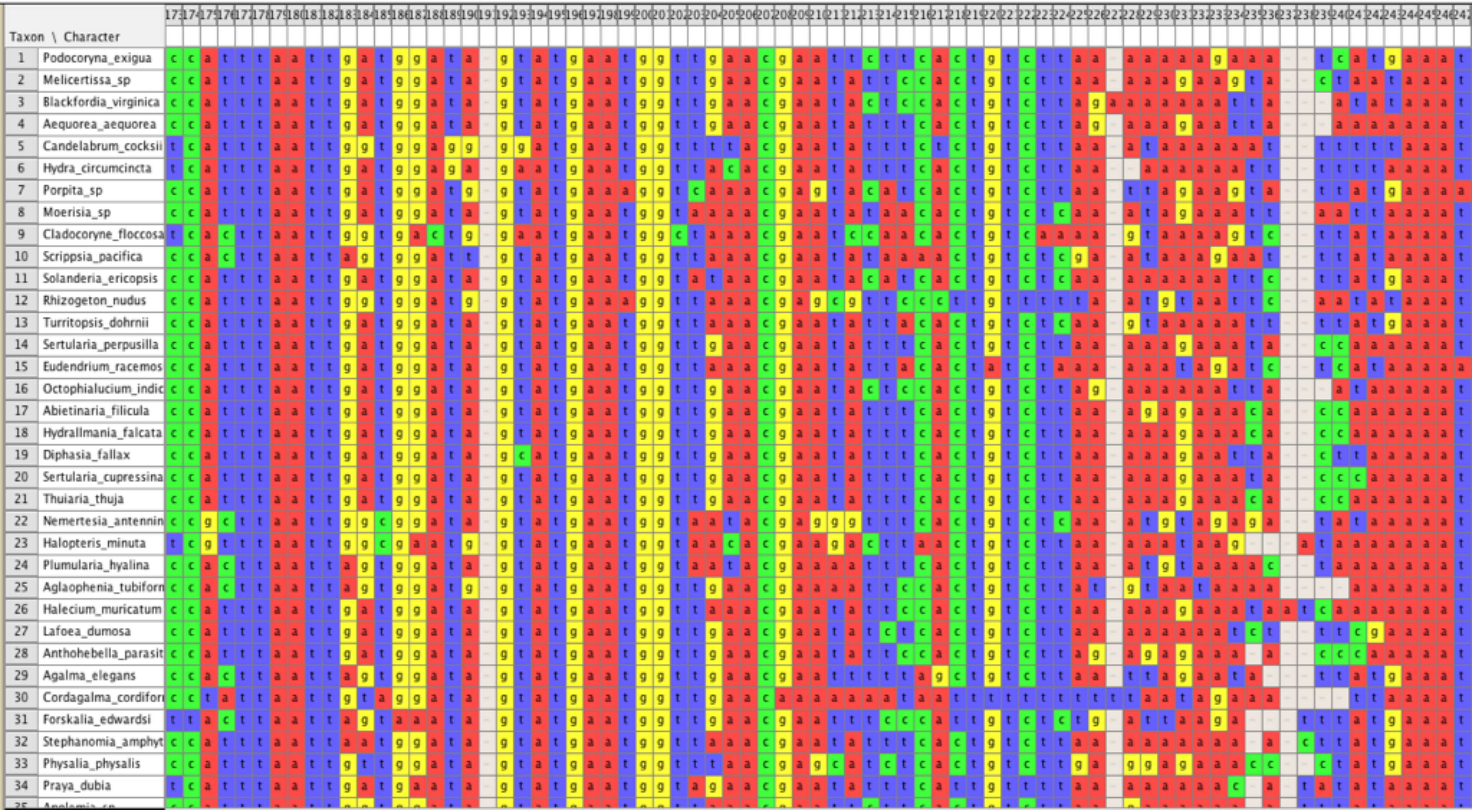
Score	Expect	Identities	Gaps	Strand
545 bits(295)	3e-151	488/580(84%)	18/580(3%)	Plus/Plus

```
Query  20   GTGTAACCTGCCCAGTGGTT---GATAAACTGAAATAAACTCTA--TAATTCAACTGAAC  74
            |||||||||||||| |||||    |||| | ||||||| | ||  |||   |||| ||
Sbjct  24   GTGTAACCTGCCCAATGGTTTTTAATAATATTAAATAAAATTTAATTAAAAAAACTTAAA  83

Query  75   GGATGCGGT-ATCTTGACCGTAATAAAGTAGCATAATCACTCGCCACTTAATTAGTGGAT  133
            ||| ||||| | ||||||||| |||| |||||||||||| |||||| ||||||  |||||
Sbjct  84   GGACGCGGTAACCTTGACCGTGATAATGTAGCATAATCATTCGCCATTTAATTGATGGAT  143

Query  134  AGTATGAATGGTTGAACGAATTTTTAGCTGTCTTAATTAG-AATATTATGAAATTGAAAT  192
            ||||||||||||||||||||| |||  |||||||||  || ||||  |  |||||  |||
Sbjct  144  AGTATGAATGGTTGAACGAATATTTCACTGTCTTAAGAAGAAATACCAAAAAATTAGAAT  203

Query  193  AATAGTCAAGATGCTATTTAAAATTGTAAGACGAAAAGACCCTATAGAGCTTAACTATT-  251
            |||||| ||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  204  AATAGTAAAGATACTATTTAAAATTGTAAGACGAAAAGACCCTATAGAGCTTAACTACAA  263

Query  252  TCTTTCTGTATAAAGGAATTTTAAATAATTACAAAAA-GA-AAGTTAGGTAGTTTAGTTG  309
            |||| |   | |||| |||   ||| | | || |||| || ||| |||||||||||||||
Sbjct  264  TCTTCCCACA-AAAG-AATG--AAAGATTCACTAAAATGAGAAGATAGGTAGTTTAGTTG  319

Query  310  GGGCGACTGCCTTTTAAAAGAAACAAAGGTAAACaatgtaattaatt--acttattgtat  367
            ||||||||| ||||||||| ||||||||  || ||| |||| ||||   |||||||||||
Sbjct  320  GGGCGACTGTCTTTTAAAAAAAACAAAGACAAGCAAAGTAAATAATAAAACTTATTGTAT  379

Query  368  aataaataaatttaacaattattaaagtaggtaataatgacccgttattattaaattaaa  427
            ||||||| |||||||||||||| ||| ||||  ||||||||||||||| | ||  | |||
Sbjct  380  AATAAATTAATTTAACAATTATAAAAATAGGCTATAATGACCCGTTATAAGTATGTAAAA  439

Query  428  aaaaTAACGATCAATAAATAAAAGCTACCTTAGGGATAACAGGATAATTTTAATTTAGAG  487
            | ||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||
Sbjct  440  ACAATAACGATCAATAAATAAAAGCTACCTTAGGGATAACAGGATAATTTTATTTTAGAG  499

Query  488  ACCTTATCGAAGTTAAAGTTTGTCACCTCTATGTTGAATTGAGATATCCATG-TAACGCA  546
            | || ||||||  ||||||||||||||||||||||||||| |||||||| || ||| |||
Sbjct  500  ATCTAATCGAAAATAAAGTTTGTCACCTCTATGTTGAATTAAGATATCC-TGATAATGCA  558

Query  547  GAAGTTATAAAGGGTGGGTCTGTTCGACCTTTAAAATCTT  586
            |||||||| || ||| ||||||||||||||||||||||||
Sbjct  559  GAAGTTATCAAAGGTAGGTCTGTTCGACCTTTAAAATCTT  598
```

Read alignment

https://vimeo.com/120429438

Multiple sequence alignment



	Taxon \ Character	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247
1	Podocoryna_exigua	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	t	c	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	a	a	g	a	a	a	-	-	t	c	a	t	g	a	a	a	t
2	Melicertissa_sp	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	c	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	g	a	a	g	t	a	-	-	c	t	a	a	t	a	a	a	t
3	Blackfordia_virginica	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	c	t	c	c	a	c	t	g	t	c	t	t	a	g	a	a	a	a	a	a	a	t	t	a	-	-	-	a	t	a	t	a	a	a	t
4	Aequorea_aequorea	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	g	-	a	a	a	g	a	a	t	t	a	-	-	-	a	a	a	a	a	a	a	t
5	Candelabrum_cocksii	t	c	a	t	t	t	a	a	t	t	g	g	t	g	g	a	g	g	-	g	g	a	t	g	a	a	t	g	g	t	t	t	t	a	c	g	a	a	t	a	t	t	t	c	t	c	t	g	t	c	t	t	a	a	-	a	t	a	a	a	a	a	a	t	-	-	t	t	t	t	t	a	a	a	t
6	Hydra_circumcincta	t	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	g	a	-	g	a	a	t	g	a	a	t	g	g	t	t	a	c	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	-	a	a	a	a	a	a	t	t	-	-	t	t	t	t	a	a	a	a	t
7	Porpita_sp	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	g	-	g	t	a	t	g	a	a	a	g	g	t	c	a	a	a	c	g	a	g	t	a	c	a	t	c	a	c	t	g	t	c	t	t	a	a	-	t	t	a	g	a	a	g	t	a	-	-	t	t	a	t	g	a	a	a	a
8	Moerisia_sp	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	a	a	a	a	c	g	a	a	t	a	t	a	a	c	a	c	t	g	t	c	t	c	a	a	-	a	t	a	g	a	a	a	t	t	-	-	a	a	t	t	a	a	a	a	t
9	Cladocoryne_floccosa	t	c	a	c	t	t	a	a	t	t	g	g	t	g	a	c	t	g	-	g	a	a	t	g	a	a	t	g	g	c	t	a	a	a	c	g	a	a	t	c	c	a	a	c	a	c	t	g	t	c	a	a	a	a	-	g	t	a	a	a	a	g	t	c	-	-	t	t	a	t	a	a	a	a	t
10	Scrippsia_pacifica	c	c	a	c	t	t	a	a	t	t	a	g	t	g	g	a	t	t	-	g	t	a	t	g	a	a	t	g	g	t	t	a	a	a	c	g	a	a	t	a	t	a	a	a	a	c	t	g	t	c	t	c	g	a	-	a	t	a	a	a	g	a	a	t	-	-	t	t	a	t	a	a	a	a	t
11	Solanderia_ericopsis	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	a	t	a	a	c	g	a	a	t	a	c	a	t	c	a	c	t	g	t	c	t	c	a	a	-	a	a	a	a	a	a	t	t	c	-	-	t	t	a	t	g	a	a	a	t
12	Rhizogeton_nudus	c	c	a	t	t	t	a	a	t	t	g	g	t	g	g	a	t	g	-	g	t	a	t	g	a	a	a	g	g	t	t	a	a	a	c	g	a	g	c	g	t	t	c	c	c	t	t	g	t	t	t	t	t	a	-	a	t	g	t	a	a	t	t	c	-	-	a	a	t	a	t	a	a	a	t
13	Turritopsis_dohrnii	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	a	a	a	c	g	a	a	t	a	t	t	a	c	a	c	t	g	t	c	t	c	a	a	-	g	t	a	a	a	a	a	t	t	-	-	t	t	a	t	g	a	a	a	t
14	Sertularia_perpusilla	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	g	a	a	a	t	a	-	-	c	c	a	a	a	a	a	a	t
15	Eudendrium_racemos	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	a	a	a	c	g	a	a	t	a	t	t	a	c	a	c	t	a	t	c	t	a	a	a	-	a	a	a	t	a	g	a	t	c	-	-	t	c	a	t	a	a	a	a	a
16	Octophialucium_indic	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	c	t	c	c	a	c	t	g	t	c	t	t	a	g	-	a	a	a	a	a	a	t	t	a	-	-	-	a	t	a	a	a	a	a	t
17	Abietinaria_filicula	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	g	a	g	a	a	a	c	a	-	-	c	c	a	a	a	a	a	a	t
18	Hydrallmania_falcata	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	g	a	a	a	c	a	-	-	c	c	a	a	a	a	a	a	t
19	Diphasia_fallax	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	c	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	g	a	a	t	t	a	-	-	c	t	t	a	a	a	a	a	t
20	Sertularia_cupressina	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	g	a	a	a	t	a	-	-	c	c	c	a	a	a	a	a	t
21	Thuiaria_thuja	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	g	a	a	a	c	a	-	-	c	c	a	a	a	a	a	a	t
22	Nemertesia_antennin	c	c	g	c	t	t	a	a	t	t	g	g	c	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	a	a	t	a	c	g	a	g	g	g	t	t	t	c	a	c	t	g	t	c	t	c	a	a	-	a	t	g	t	a	g	a	g	a	-	-	t	a	t	a	a	a	a	a	t
23	Halopteris_minuta	t	c	g	t	t	t	a	a	t	t	g	g	c	g	a	a	t	g	-	g	t	a	t	g	a	a	t	g	g	t	a	a	c	a	c	g	a	a	g	a	c	t	t	a	a	c	t	g	t	c	t	t	a	a	-	a	a	a	t	a	a	g	-	-	-	a	t	a	a	a	a	a	a	a	t
24	Plumularia_hyalina	c	c	a	c	t	t	a	a	t	t	a	g	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	a	a	t	a	c	g	a	a	a	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	t	g	t	a	a	a	a	c	-	-	t	a	a	a	a	a	a	a	t
25	Aglaophenia_tubiforn	c	c	a	c	t	t	a	a	t	t	a	g	t	g	g	a	t	g	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	a	a	t	t	c	c	a	c	t	g	t	c	t	t	a	t	-	g	t	a	a	t	a	a	a	a	-	-	-	-	a	a	a	a	a	a	t
26	Halecium_muricatum	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	a	a	a	c	g	a	a	t	a	t	t	c	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	g	a	a	a	t	a	a	t	c	a	a	a	a	a	a	a	t
27	Lafoea_dumosa	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	c	t	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	a	a	a	t	c	t	-	-	t	t	c	g	a	a	a	a	t
28	Anthohebella_parasit	c	c	a	t	t	t	a	a	t	t	g	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	a	t	t	c	c	a	c	t	g	t	c	t	t	a	g	-	a	g	a	g	a	a	a	-	a	-	-	c	c	c	a	a	a	a	a	t
29	Agalma_elegans	c	c	a	c	t	t	a	a	t	t	a	g	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	t	t	t	t	a	g	c	t	g	t	c	t	t	a	a	-	t	t	a	g	a	a	t	a	-	-	-	t	t	a	t	g	a	a	a	t
30	Cordagalma_cordiforr	c	c	t	a	t	t	a	a	t	t	g	t	a	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	a	a	a	a	a	a	a	t	a	a	t	t	t	t	t	t	t	t	t	t	t	a	a	t	a	g	a	a	a	-	-	-	-	t	t	a	a	a	a	t
31	Forskalia_edwardsi	t	t	a	c	t	t	a	a	t	t	a	g	t	a	a	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	g	a	a	c	g	a	a	t	t	t	c	c	c	a	t	t	g	t	c	t	c	t	g	-	a	t	t	a	a	g	a	-	-	-	t	t	t	a	t	g	a	a	a	t
32	Stephanomia_amphyt	c	c	a	t	t	t	a	a	t	t	a	a	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	a	a	a	c	g	a	a	t	a	t	t	t	c	a	c	t	g	t	c	t	t	a	a	-	a	a	a	a	a	a	a	-	a	-	c	t	t	a	t	g	a	a	a	t
33	Physalia_physalis	c	c	a	t	t	t	a	a	t	t	g	t	t	g	g	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	t	t	a	a	c	g	a	g	c	a	t	c	t	c	a	c	t	g	t	c	t	t	g	a	-	g	g	a	g	a	a	a	c	c	-	-	c	t	a	t	g	a	a	a	t
34	Praya_dubia	t	c	a	t	t	t	a	a	t	t	g	a	t	g	a	a	t	a	-	g	t	a	t	g	a	a	t	g	g	t	a	g	a	a	c	g	a	a	t	a	t	t	t	c	a	t	t	g	t	t	t	t	a	a	-	a	a	a	a	a	a	c	-	a	-	t	a	t	a	t	a	a	a	a	t

Multiple sequence alignment

Common models used in phylogenetic inference (eg GTR) accommodate substation, but not insertion/ deletion

Most phylogenetic programs therefore don’t infer homology, they assume that each column is a set of homologous sequences

They treat gaps introduced by indels as missing data

Multiple sequence alignment

Need an aligner upstream of phylogenetic inference that infers which sequence differences are due to substitution and which are due to insertion/deletion

Partitions sites according to inferred mechanism: changes due to indels are put in separate columns, each column contains sites that are hypothesized to be only due to substitution

Multiple sequence alignment

Many MSA tools are available:

mafft
clustalw
muscle
t-coffee

Multiple sequence alignment

In general, they work by:

1. Defining a set of penalties for site differences and the introduction of gaps

2. Heuristically searching for an alignment that minimizes these penalties

Multiple sequence alignment

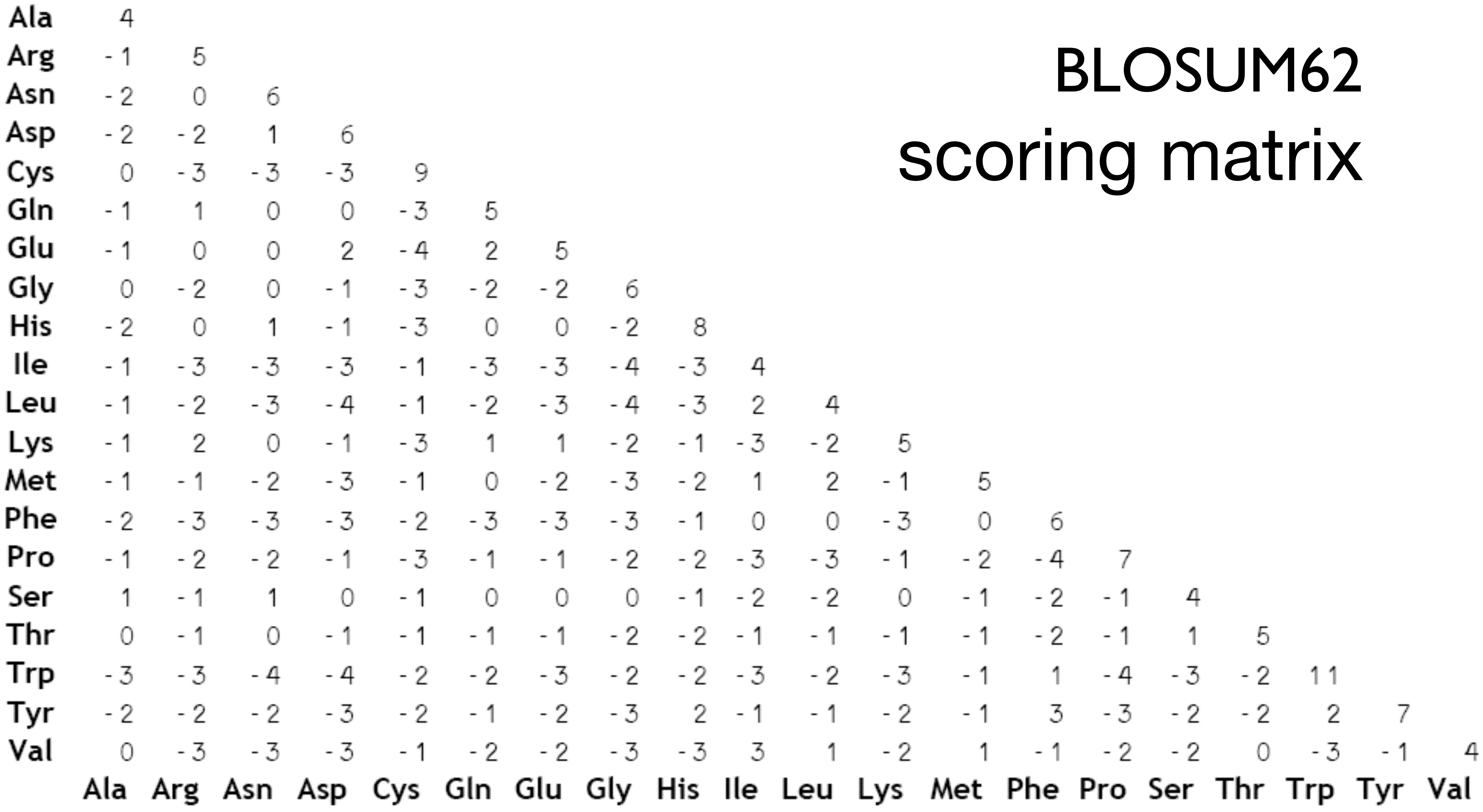
The scoring matrix is used to evaluate site differences.

Explains how surprised we should be to see a particular substitution that leads to a difference between homologous sequences.

Related to site substitution models.

Multiple sequence alignment

BLOSUM62
scoring matrix



Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

https://commons.wikimedia.org/wiki/File:BLOSUM62.gif

Multiple sequence alignment

Different types of gap penalties:

Gap opening penalty - The cost of creating a gap of one site where there was no gap

Gap extension penalty - The cost of adding gaps adjacent to an existing gap.