

## Multiple sequence alignment and phylogenetic trees

CRC-CGF Bioinformatics Course

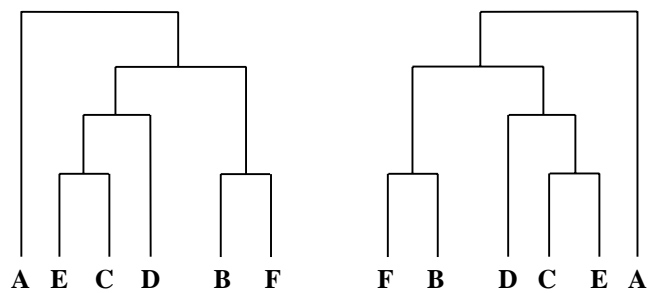
September 9-13, 2002

Terry Speed, WEHI

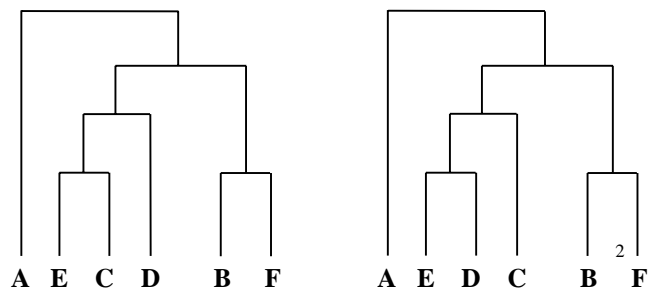
1

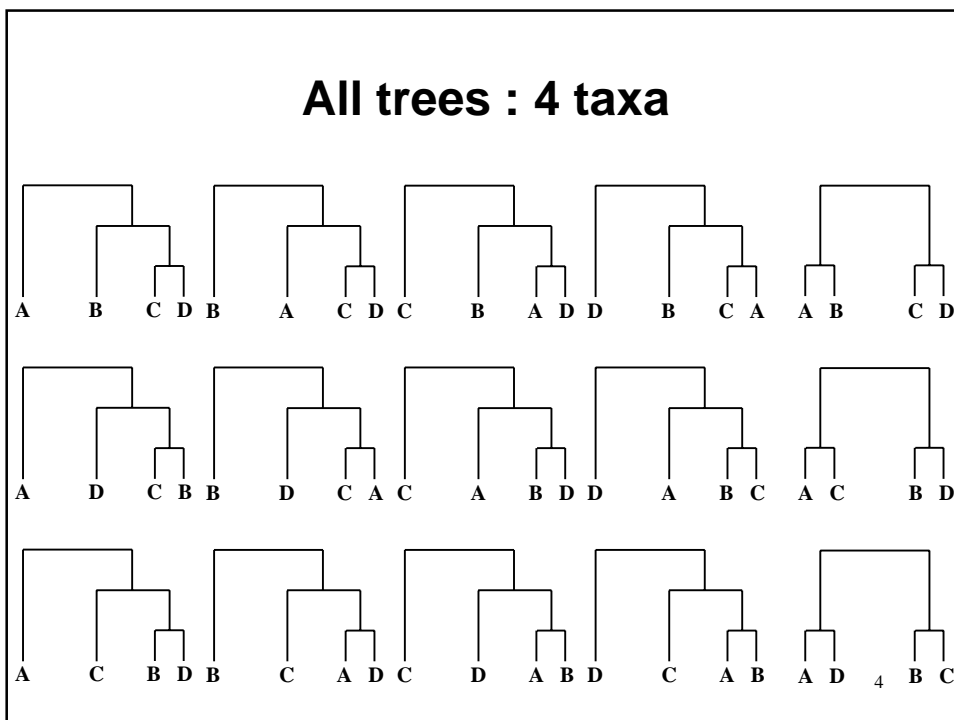
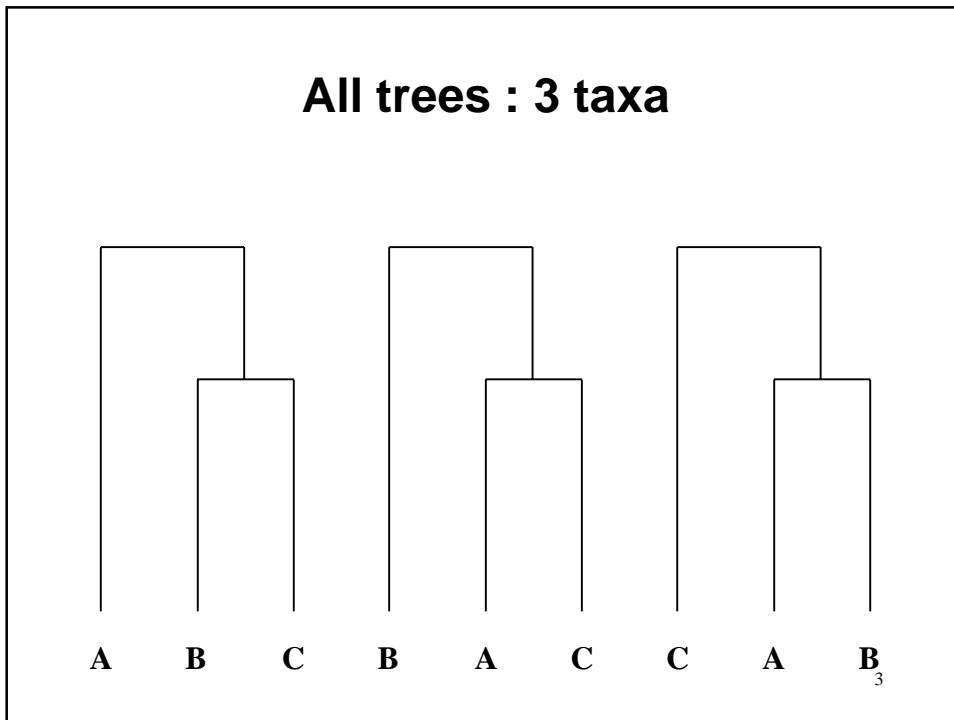
### Tree topologies

Identical:



Not identical:





## All trees : >4 taxa

- In general, for any strictly bifurcating rooted tree with  $n$  species, there are  $(2n-3)! / 2^{n-3}(n-2)!$  different topologies.

<u><math>n</math></u>	<u>#trees</u>
5	105
15	213,458,046,676,875
20	8,200,794,532,637,891,559,375

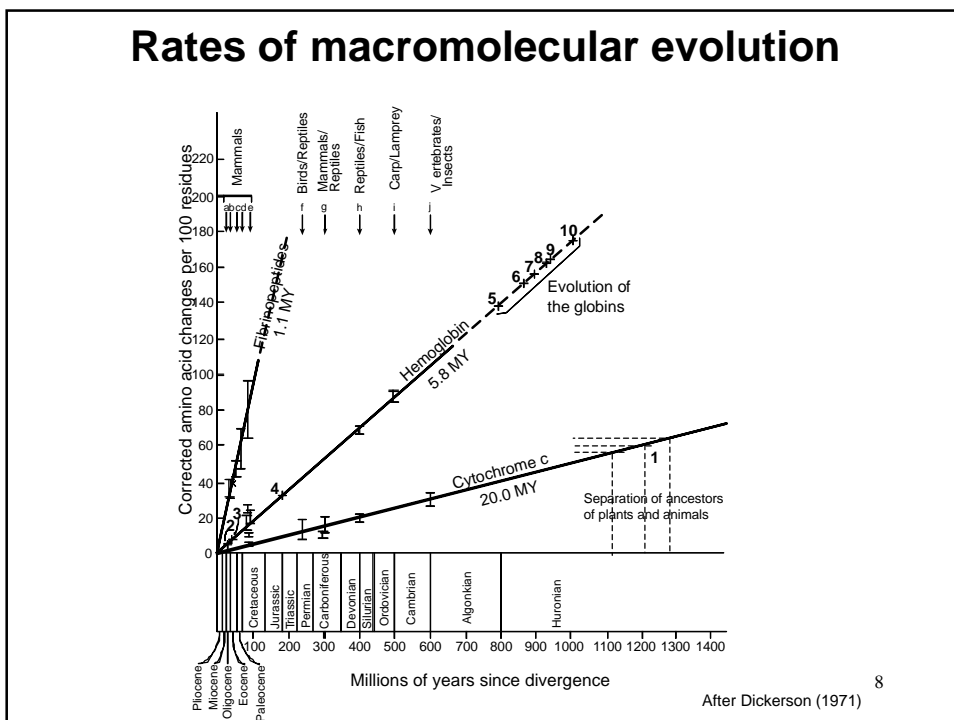
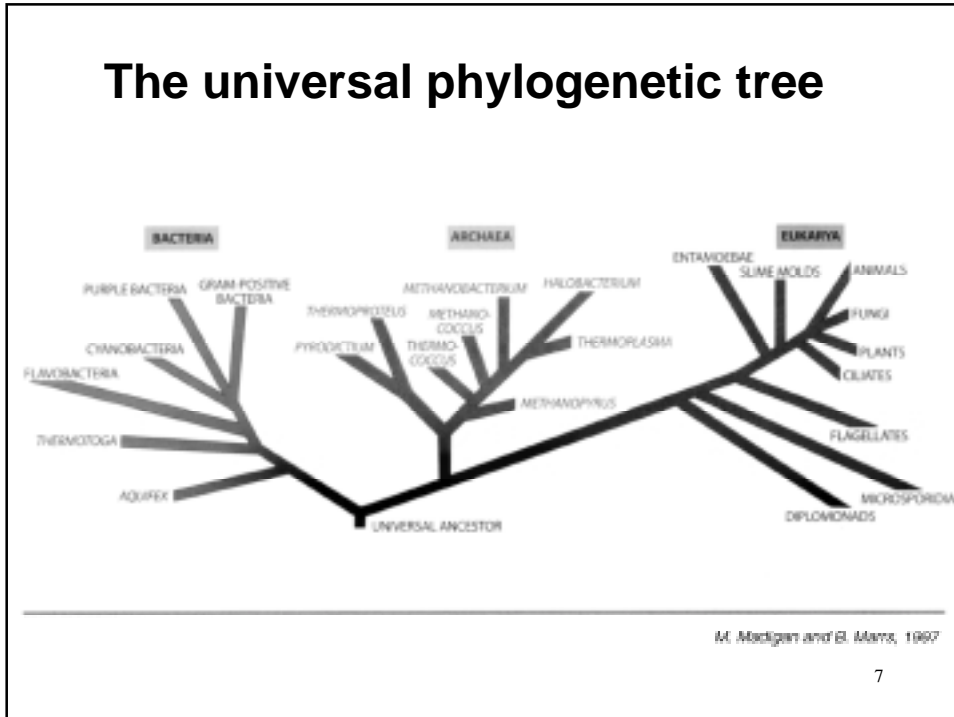
- For unrooted trees, it's only  $(2n-5)! / 2^{n-3}(n-3)!$

5

## Tree reconstruction: some methods

- Distance-based methods
  - UPGMA
  - Transformed distance
  - Neighbor-joining
- Character state-based methods
  - Maximum parsimony
  - Linear invariants
- Maximum Likelihood

6



## Protein evolution: useful distinction

A common mode of protein evolution is by duplication. Depending on the relations between duplication and speciation dates, we have two different types of homologous proteins. Loosely,

**Orthologues:** the “same” gene in different organisms; common ancestry goes back to speciation

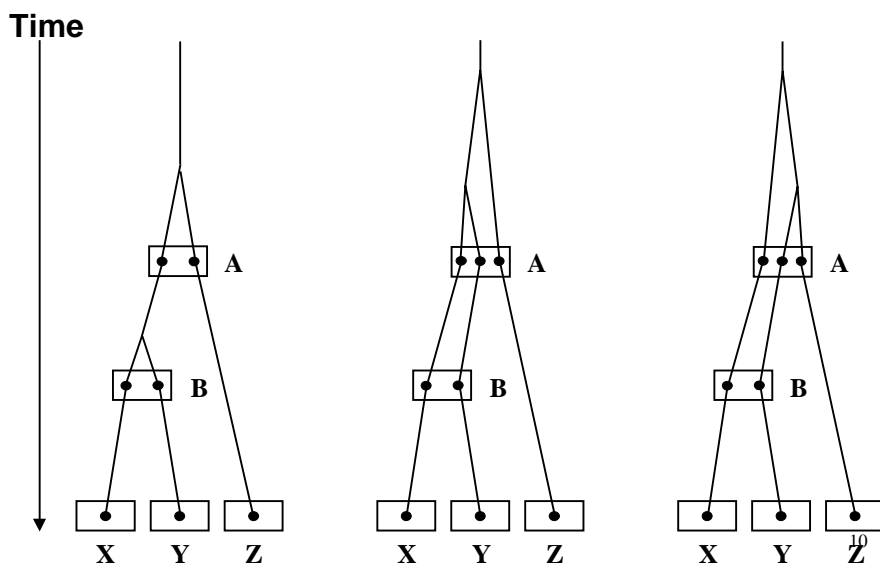
**Paralogues:** different genes in the same organism; common ancestry goes back to a gene duplication.

**Lateral gene transfer** gives another form of homology.

9

## Species and gene trees

(after M. Nei 1987, *Molecular Evolutionary Genetics.*)



### Beta-globins (orthologues)

```

      10          20          30          40
BG-human  M V H L T P E E K S A V T A L W G K V N V D E V G G E A L G R L L V V Y P W T Q
BG-macaque . . . . . N . . . T . . . . .
BG-bovine  - - M . . A . . . . . F . . . . . K . . . . .
BG-platypus . . . . . S G G . . . . . N . . . . . I N . L . . . . .
BG-chicken . . . . . W A . . . . . Q L I . G . . . . . A C A . . . . . I . . . . .
BG-shark   - . . W S E V . L H E I . T T . K S I D K H S L . A K . . . . . A M F I . . . . . T

      50          60          70          80
BG-human  R F F E S F G D L S T P D A V M G N P K V K A H G K K V L G A F S D G L A H L D
BG-macaque . . . . . S . . . . . N . . . . .
BG-bovine  . . . . . A . . . . . N . . . . . D S . . N M K . . . . .
BG-platypus . . . . . A . . . . . S A G . . . . . A . . . . . T S . G A K N . . . . .
BG-chicken . . . . . A . . . . . N . . . . . S T . I L . . . . . M R . . . . . T S . G A V K N . . . . .
BG-shark   . Y . G N L K E F T A C S Y G - - - - . E . A . . . . . T . . L G V A V T . . G

      90          100         110         120
BG-human  N L K G T F A T L S E L H C D K L H V D P E N F R L L G N V L V C V L A H H F G
BG-macaque . . . . . Q . . . . . K . . . . .
BG-bovine  D . . . . . A . . . . . K . . . . . V . . . . . R N . . . . .
BG-platypus D . . . . . K . . . . . N R . . . . . I V . . . . . R . . . . . S
BG-chicken . I . N . . S Q . . . . . . . . . . . D I . I I . . . . . A . . . . . S
BG-shark   D V . S Q . T D . . K K . A E E . . . . . V . S . K . . A K C F . V E . G I L L K

      130         140
BG-human  K E F T P P V Q A A Y Q K V V A G V A N A L A H K Y H
BG-macaque . . . . . Q . . . . .
BG-bovine  . . . . . V L . . D F . . . . . R . . . . .
BG-platypus . D . S . E . . . W . L . S . . H . . G . . . . .
BG-chicken . D . . . E C . . . W . . L . R V . H . . . R . . . . .
BG-shark   D K . A . Q T . . I W E . Y F G V . V D . I S K E . . . . .
    
```

. means same as reference sequence  
- means deletion

11

### Beta-globins: Uncorrected pairwise distances

Distances: between protein sequences. Calculated over: 1 to 147  
Below diagonal: observed number of differences  
Above diagonal: number of differences per 100 amino acids

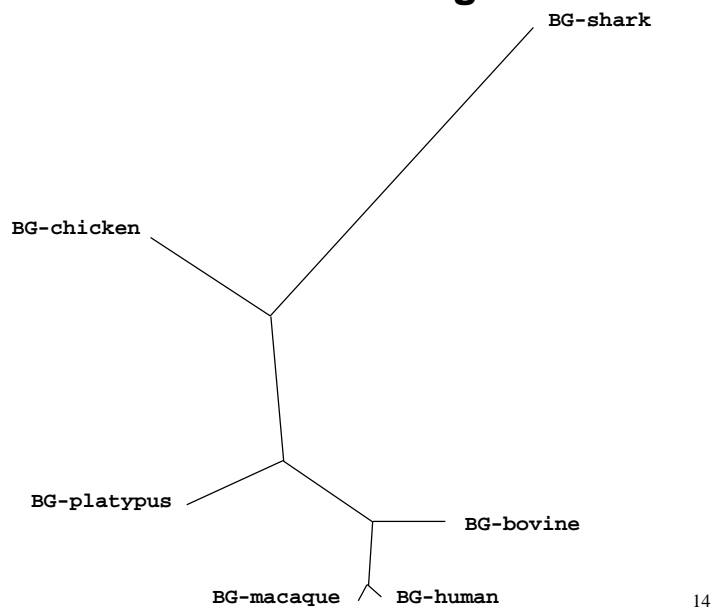
	hum	mac	bov	pla	chi
sha					
hum	----	5	16	23	31
65					
mac	7	----	17	23	30
62					
bov	23	24	----	27	37
65					
pla	34	34	39	----	29
64					
chi	45	44	52	42	----
					12

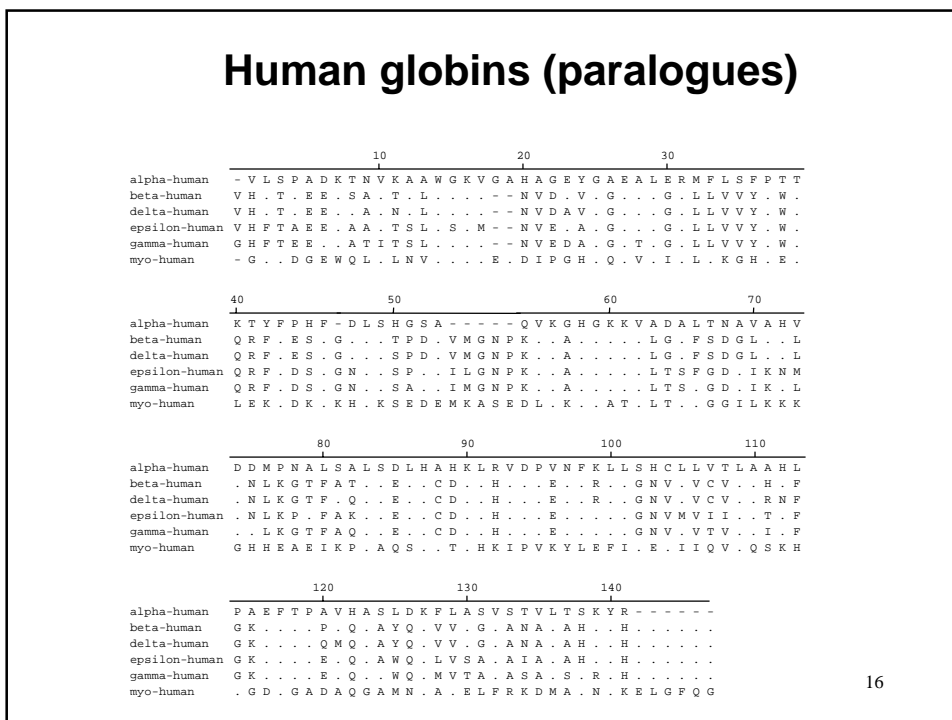
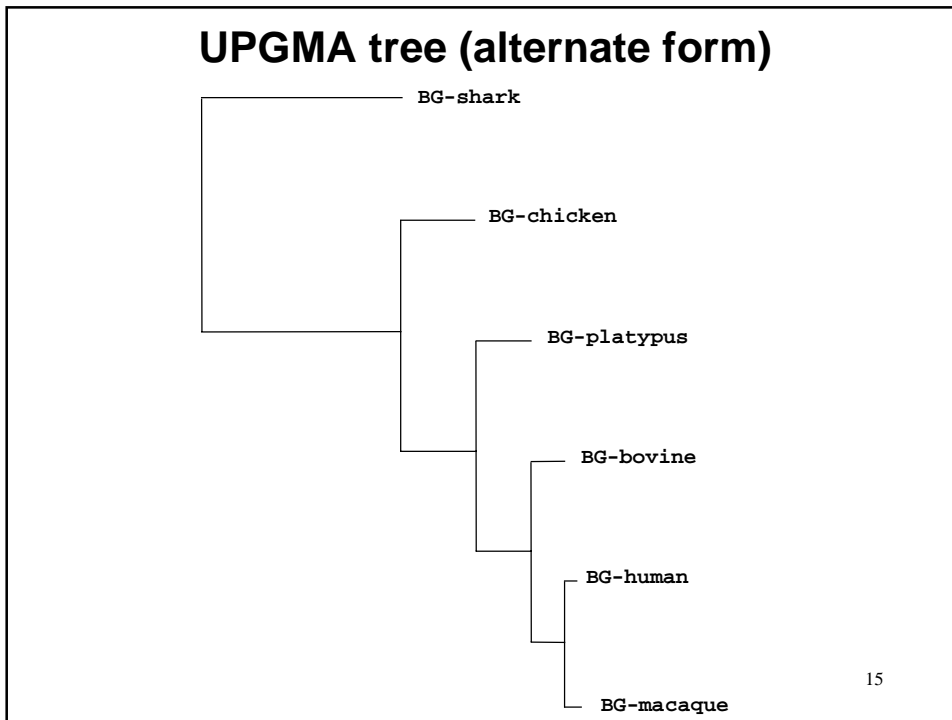
## Beta-globins: Corrected pairwise distances

**Distances:** between protein sequences. **Calculated over:** residues 1 to 147  
**Below diagonal:** observed number of differences  
**Above diagonal:** estimated number of substitutions per 100 amino acids  
**Correction method:** Jukes-Cantor

	hum	mac	bov	pla	chi
sha					
hum	----	5	17	27	
37	108				
mac	7	----	18	27	
36	102				
bov	23	24	----	32	
46	110				
pla	34	34	39	----	
34	106				
chi	45	44	52	42	13
					-

## UPGMA tree for beta-globins





### Human globins: uncorrected pairwise distances

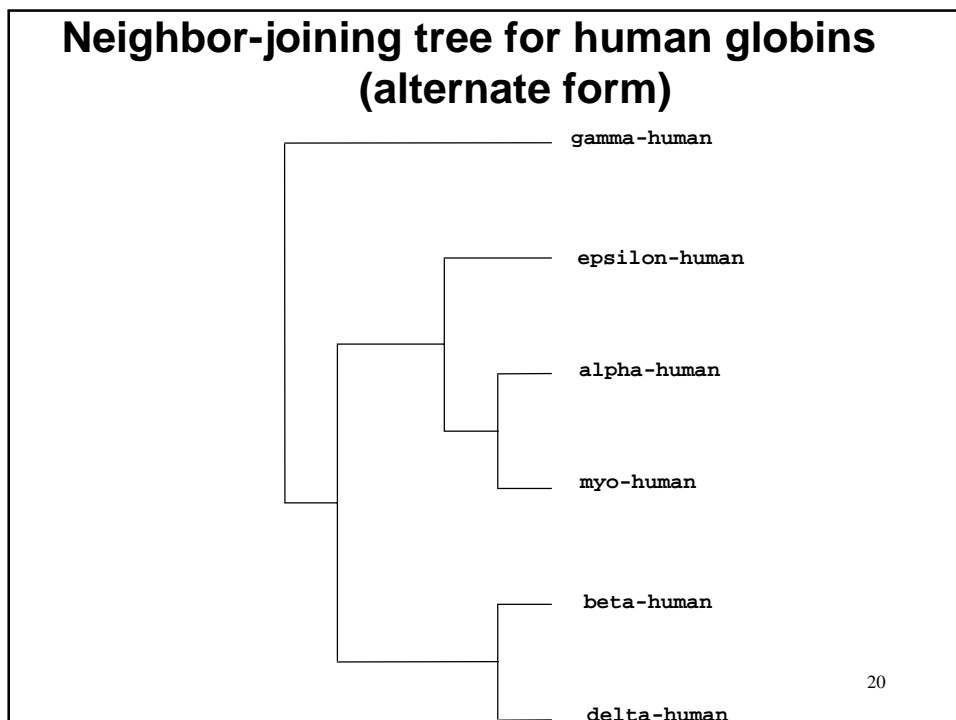
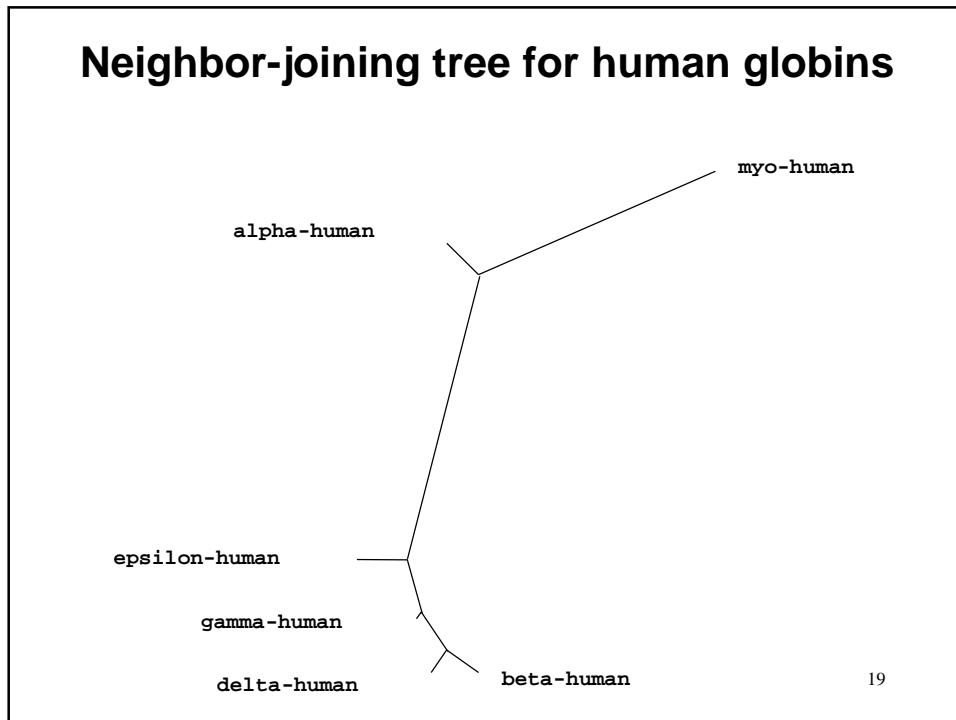
**Distances:** between protein sequences. **Calculated over:** 1 to 154  
**Below diagonal:** observed number of differences  
**Above diagonal:** number of differences per 100 amino acids

	alpha	beta	delta	eps	gamma
myo					
alpha	----	55	55	60	
57	74				
beta	82	----	7	25	
27	75				
delta	82	10	----	27	
29	74				
Eps	89	35	39	----	
20	77				17

### Human globins: Corrected pairwise distances

**Distances:** between protein sequences. **Calculated over:** 1 to 141  
**Below diagonal:** observed number of differences  
**Above diagonal:** estimated number of substitutions per 100 amino acids  
**Correction method:** Jukes-Cantor

	alpha	beta	delta	epsil
gamma	myo			
alpha	----	281	281	281
313	208			
beta	82	----	7	30
31	1000			
delta	82	10	----	34
33	470			
epsil	89	35	39	----
21	402			



## Why multiple alignment?

*The simultaneous alignment of a number of DNA or protein sequences is one of the commonest tasks in bioinformatics.*

Some uses of multiple alignment:

- phylogenetic analysis (inferring a tree, estimating rates of substitution, etc.)
- detection of homology between a newly sequenced gene and an existing gene family
- prediction of protein structure
- demonstration of homology in multi-gene families
- determination of a consensus sequence (e.g., in assembly)

21

## A multiple alignment of globins

	10	20	30	40	50	60
Hbb_Human.pep	-----VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST					
Hbb_Horse.pep	-----VQLSGEKAAVLALWDKVN--EEEVGGALGRLLVVYPWTQRFFDSFGDLSN					
Hba_Human.pep	-----VLSPADKTNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHFDLS--					
Hba_Horse.pep	-----VLSAADKTNVKAWSKVGGHAGEYGAELERMFLGFPTTKTYFPHFDLS--					
Myg_Phyca.pep	-----VLSEGEWQLVHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDKFKHLKT					
Glb5_Petma.pep	PIVDVTGSVAPLSAAEKTIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT					
Lgb2_Luplu.pep	-----GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE					
	*	.	*	.	*	*
Hbb_Human.pep	PDAVMGNPKVKAHGKKVLGAFSDGLAHLA-----NLKGTFAATLSELHCDKLVDPENFRL					
Hbb_Horse.pep	PGAVMGNPKVKAHGKKVLSFGEGVHHLA-----NLKGTFAALSELHCDKLVDPENFRL					
Hba_Human.pep	----HGSAQVKGHGKVKADALTNVAHAVD-----DMPNALSALSDLHAHKLKRVDPVNFKL					
Hba_Horse.pep	----HGSAQVKAHGKKGVDALTLAVGHLD-----DLPGALSALSDLHAHKLKRVDPVNFKL					
Myg_Phyca.pep	EAEMKASEDLKKGHTVTLTALGAILKKGK-----HHEAELKPLAQSHATKHKIPIKYLEF					
Glb5_Petma.pep	ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLGKHAQSFQVDPQYFKV					
Lgb2_Luplu.pep	VP--QNNPELQAHAGKVFKLVEAAIQLVTVVVDATLKNLGSVHVSKG--VADAHFPV					
	..	*	.	.	*	*
Hbb_Human.pep	LGNVLCVLAHHFGKEFTPPVQAAVQKVVAGVANALAHKYH-----					
Hbb_Horse.pep	LGNLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH-----					
Hba_Human.pep	LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR-----					
Hba_Horse.pep	LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTISKYR-----					
Myg_Phyca.pep	ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKIDIAAKYKELGYQG					
Glb5_Petma.pep	LAAVIADTVAAG-----DAGFEKLMSMICILLRSAY-----					
Lgb2_Luplu.pep	VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---					

22

## Extending the pairwise alignment algorithms

- Generally not feasible for more than a small number of sequences (~5), as the necessary computer time and space quickly becomes prohibitive. Computational time grows as  $N^m$ , where  $m$  = number of sequences. For example, for 100 residues from 5 species,  $100^5 = 10,000,000,000$  (*i.e.*, the equivalent of two sequences each 100,000 residues in length.)
- Nor is it wholly desirable to reduce multiple alignment to a mathematical problem similar to that tackled by pairwise alignment algorithms.

Two issues which are important in discussions of multiple alignment are:

the treatment of gaps: position-specific and/or residue-specific gap penalties are both desirable and feasible, and

the phylogenetic relationship between the sequences (which must exist if they are alignable): it should be exploited.

23

## Progressive alignment

Up until about 1987, multiple alignments would typically be constructed manually, although a few computer methods did exist. Around that time, algorithms based on the idea of **progressive alignment** appeared. In this approach, a pairwise alignment algorithm is used iteratively, first to align the most closely related pair of sequences, then the next most similar one to that pair, and so on.

The rule “once a gap, always a gap” was implemented, on the grounds that the positions and lengths of gaps introduced between more similar pairs of sequences should not be affected by more distantly related ones.

24

## Multiple alignment in 2002

The most widely used progressive alignment algorithm is currently **CLUSTAL W**, and we discuss this in detail below.

There are a number of more specialized procedures based on quite different principles, including the use of hidden Markov models built for protein families. We'll discuss these briefly in the next lecture.

A relatively new and very promising approach uses Markov chain Monte Carlo methods to sample alignments according to certain probabilistic procedures and, by moving randomly around in the huge space of possible alignments, to find good alignments. We'll give references to one of these.

25

## CLUSTAL W

The three basic steps in the **CLUSTAL W** approach are shared by all progressive alignment algorithms:

- A. Calculate a matrix of **pairwise distances** based on pairwise alignments between the sequences
- B. Use the result of A to build a **guide tree**, which is an inferred phylogeny for the sequences
- C. Use the tree from B to guide the **26** of the sequences.

Our discussion comes from the paper Thompson *et al*, 1994.

## A. Calculating the pairwise distances

A pair of sequences is aligned by the usual dynamic programming algorithm, and then a similarity or distance measure for the pair is calculated using the aligned portion (gaps excluded) - for example, percent identity.

**CLUSTAL W** does not correct these distances for multiple substitutions (*e.g.*, by the Jukes-Cantor formula), although other programs do, and it is sometimes an option in different versions of the program.

27

## A. Our globin example

DISTANCES between protein sequences:

Calculated over: 1 to 167  
Correction method: Simple distance (no corrections)  
Distances are: observed number of substitutions per 100 amino acids  
Symmatrix version 1  
Number of matrices: 1

//  
Matrix 1, dimension: 7

Key for column and row indices:

1 hba\_human  
2 hba\_horse  
3 hbb\_human  
4 hbb\_horse  
5 glb5\_petma  
6 myg\_phyca  
7 lgb2\_luplu

Matrix 1: Part 1

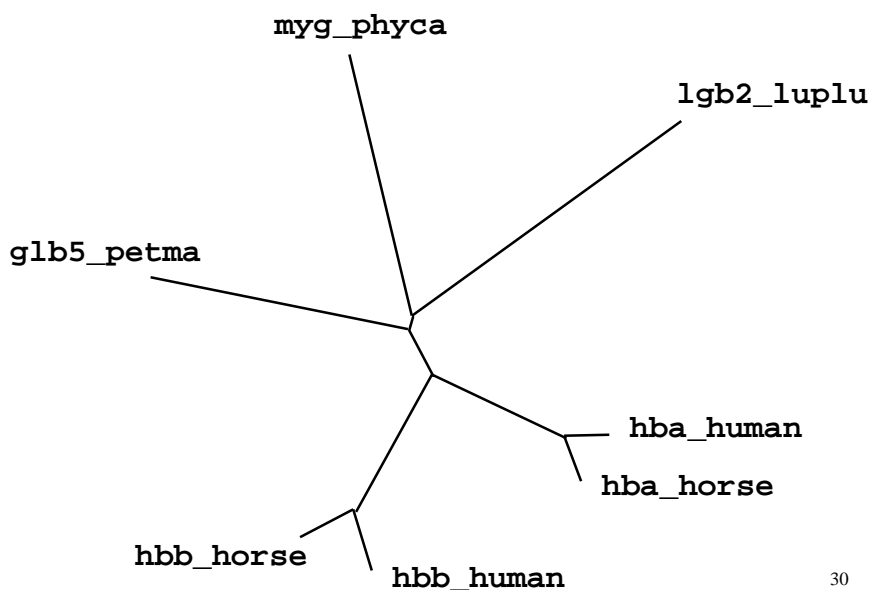
	1	2	3	4	5	6	7	..
1	0.00	12.06	54.68	55.40	64.12	71.74	83.57	
2		0.00	55.40	53.96	64.89	72.46	82.86	
3			0.00	16.44	74.26	73.94	82.52	
4				0.00	75.74	73.94	81.12	
5					0.00	75.91	82.61	
6						0.00	80.95	
7							0.00	

28

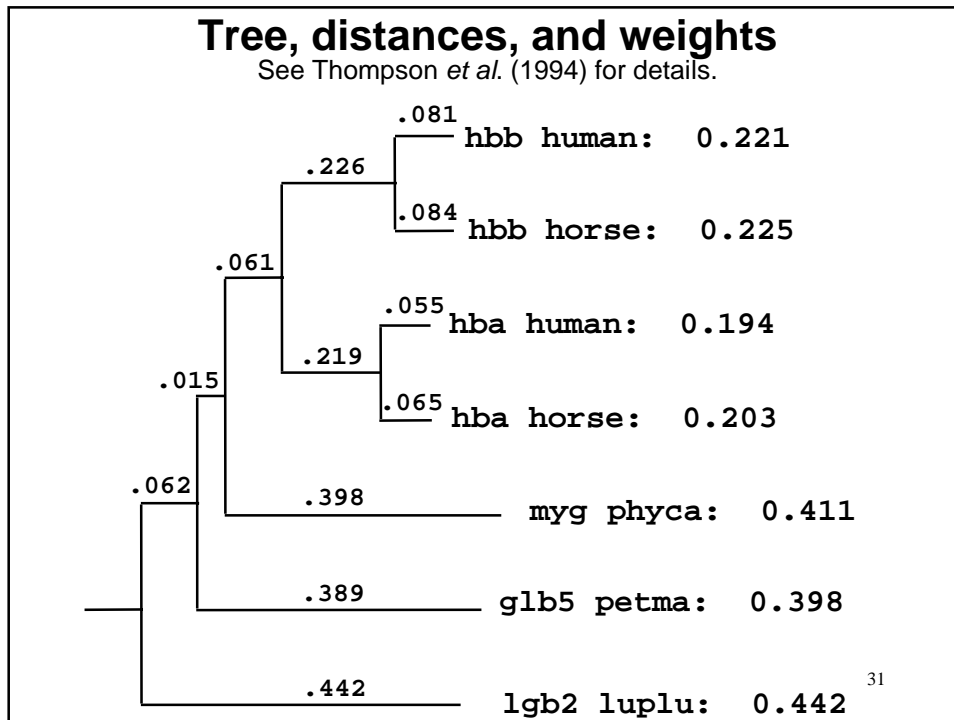
## B. Building the guide tree.

- There are many ways of building a tree from a matrix of pairwise distances. **CLUSTAL W** uses the *neighbour-joining* (NJ) method, which is the most favoured approach these days. Earlier versions of **CLUSTAL** used the *unweighted pair group method using arithmetic averages* (UPGMA), and this is still used in some programs.
- A *root* of the tree is then determined by the so-called *mid-point method* (giving equal means for the branch lengths on either side of the root).
- The **W** in **CLUSTAL W** stands for **Weights**, an important feature of this program. These are calculated in a straightforward way. They correct for unequal sampling at different evolutionary distances. <sup>29</sup>

## NJ tree for our globin examples



30



### C. Progressive alignment.

The basic idea is to use a series of pairwise alignments to align larger and larger groups of sequences, following the branching order of the guide tree. We proceed from the tips of the rooted tree towards the root.

In our globin example, we align in the following order:

- a) human and horse  $\beta$ -globin;
- b) human and horse  $\alpha$ -globin;
- c) the two  $\alpha$ -globins and the two  $\beta$ -globins;
- d) myoglobin and the haemoglobins;
- e) cyanohaemoglobin and the combined haemoglobin, myoglobin group;
- f) leghaemoglobin and the rest.

32

### C. Progressive alignment, cont.

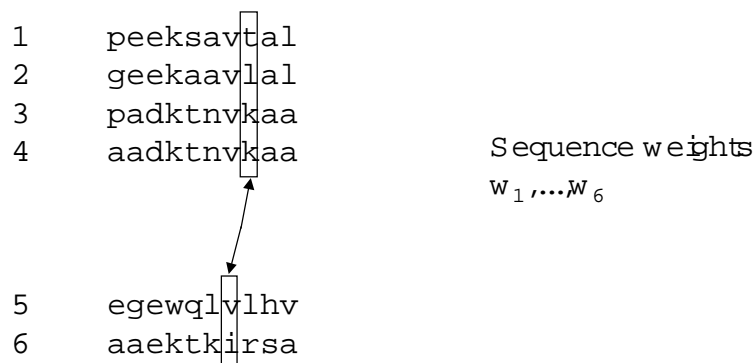
At each stage a full dynamic programming algorithm is used, with a residue scoring matrix (e.g., a PAM or a BLOSUM matrix) and gap opening and extension penalties.

Each step consists of aligning two existing alignments. Scores at a position are averages of all pairwise scores for residues in the two sets of sequences using matrices with only positive values. Gap vs. residue scores zero. Sequence weights are used at this stage. See next slide.

Gaps that are present in older alignments remain fixed. New gaps introduced at each stage initially get full opening and extension penalties, even if inside old gap positions. This gets modified.

33

### Scoring an alignment of two partial alignments



$$\text{Score: } \frac{1}{8} [M(t, v)w_1w_5 + M(t, i)w_1w_6 + \dots + M(k, i)w_4w_6]$$

34

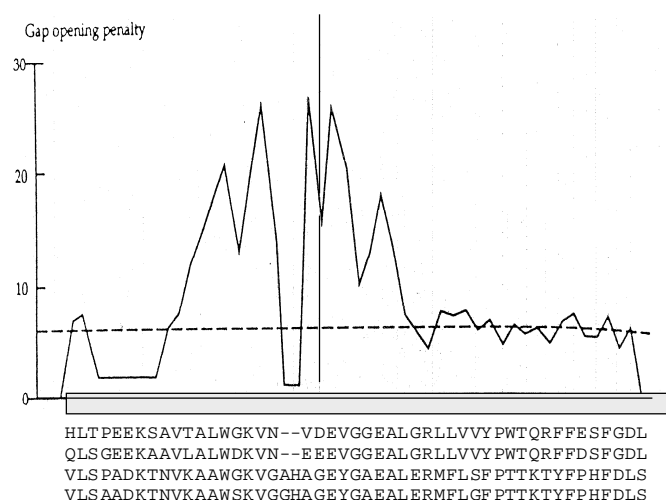
### C. Progressive alignment: gaps

**CLUSTAL W** has quite a sophisticated treatment of gaps, incorporating into opening and extension penalties a dependence on a) weight matrix, b) sequence similarity, c) sequence length, d) difference in sequence length, e) position of gaps (see figure), f) residues at gaps.

Regarding e) and f), the motivation is as follows: if one knew the positions of all secondary structure elements ( $\alpha$ -helices,  $\beta$ -strands) in all or some of the sequences, one could increase the gap penalties *inside* and decrease *outside* them, forcing gaps to occur most often in loop regions, which is what is observed in alignments of sequences with known 3-D structure.

For further details, see Thompson *et al.*, **NAR 22** 1994, p.4673 or **Methods in Enzymology 266**, 1996, article 22<sup>35</sup>

### Position- and residue-specific gap opening penalties



36

### Final CLUSTALW alignment (obtained using eclustalw)

```

10      20      30      40      50      60
Hbb_Human.pep  -----VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPTWQRFFESFGDLSL
Hbb_Horse.pep  -----VQLSGEKAAVLALWDKVN--EEEVGGGEALGRLLVVYPTWQRFFDSFGDLSN
Hba_Human.pep  -----VLSPADKTNVKAAWGKVGAGHAGEYGAEALERMFLSFPTTKTYFPHFDLS--
Hba_Horse.pep  -----VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLS--
Myg_Phyca.pep  -----VLSGEGEWQLVHLVWAKVEADVAGHGQDILIRLRFKSHPETLEKFFDRFKHLKT
Glb5_Petma.pep PIVDTGVSAPLSAAEKTKIRSAWAPVYSTIYETSGVDILVKFFTSIPAAQEFPPKFKGLTT
Lgb2_Luplu.pep -----GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
          *      *      *      *      *      *
          .      .      .      .      .      .

Hbb_Human.pep  PDAVMGNPKVKAHGKKV L GAFSDGLAHLD-----NLKGTFAATLSELHCDK LHVDPENFRL
Hbb_Horse.pep  PGAVMGNPKVKAHGKKV LHSFGEGVHHL D-----NLKGTFAALSELHCDK LHVDPENFRL
Hba_Human.pep  ---HGSAQVKGHGKKVADAL TNAV AHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
Hba_Horse.pep  ---HGSAQVKAHGKKVGDAL TLA VGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
Myg_Phyca.pep  EAEMKASEDLKKGHTVLTALGAILKKKG-----HHEAELKPLAQSHATKKIKPKYLEF
Glb5_Petma.pep ADQLKKSADVRWHAERI INAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
Lgb2_Luplu.pep VP--QNNPELQAHAGKVF KLVYEAALQLQVTGVVVT D AT LKNLGSVHVS KG-VADAHFPV
          *      *      *      *      *      *
          .      .      .      .      .      .

Hbb_Human.pep  LGNVLVCVLAH HFGKEFTPPVQAAYQKVVAGVANALAHKYH-----
Hbb_Horse.pep  LGNVLVVVLA RHF GKDFTP ELQAS YQKVVAGVANALAHKYH-----
Hba_Human.pep  LSHCLLVTLAAHLPAEF TPAVHASLDKFLASVSTVLT SKYR-----
Hba_Horse.pep  LSHCLLSTLAVHL PNDFTPAVHASLDKFLSSVSTVLT SKYR-----
Myg_Phyca.pep  ISEAI IHVLHSRHPGDFGADAQGAMNKALELFRKDI AAKYKELGYQG
Glb5_Petma.pep LAAVIADTVAAG-----DAGFEKLM SMICILLRSAY-----
Lgb2_Luplu.pep VK EALIKTIKEVIGAKWSEELMSAWT TAYDELAIVIKKEMNDAA---

```

**7  $\alpha$ -helices**

## Markov chain Monte Carlo alignment

This procedure is currently only available for gap-free local alignments, or blocks of gap-free local alignments. A good source of software for this and related tasks is:

<http://www.wadsworth.org/resnres/bioinfo/>

## Web-based multiple sequence alignment

- ClustalW
    - [www2.ebi.ac.uk/clustalw/](http://www2.ebi.ac.uk/clustalw/)
    - [dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html](http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html)
    - [www.clustalw.genome.ad.jp/](http://www.clustalw.genome.ad.jp/)
    - [bioweb.pasteur.fr/intro-uk.html](http://bioweb.pasteur.fr/intro-uk.html)
    - [pbil.ibcp.fr](http://pbil.ibcp.fr)
    - [transfac.gbf.de/programs.html](http://transfac.gbf.de/programs.html)
    - [www.bionavigator.com](http://www.bionavigator.com)
  - PileUp
    - [helix.nih.gov/newhelix](http://helix.nih.gov/newhelix)
    - [www.hgmp.mrc.ac.uk/](http://www.hgmp.mrc.ac.uk/)
    - [bcf.arl.arizona.edu/gcg.html](http://bcf.arl.arizona.edu/gcg.html)
    - [www.bionavigator.com](http://www.bionavigator.com)
  - Dialign
    - [genomatix.gsf.de/](http://genomatix.gsf.de/)
    - [bibiserv.techfak.uni-bielefeld.de/](http://bibiserv.techfak.uni-bielefeld.de/)
    - [bioweb.pasteur.fr/intro-uk.html](http://bioweb.pasteur.fr/intro-uk.html)
    - [www.hgmp.mrc.ac.uk/](http://www.hgmp.mrc.ac.uk/)
  - Match-box
    - [www.fundp.ac.be/sciences/biologie/bms/matchbox\\_submit.html](http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html)
- For reviews: G. J. Gaskell, **BioTechniques** 2000, **29**:60, and  
[www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/welcome.html](http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/welcome.html)

39

## Comparing multiple sequence alignment programs

Even below the 10-20% identity twilight zone, the best programs correctly align 47% of residues on average

Iterative algorithms are superior, but with a large trade-off in use of computational resources

Global generally performs better than local

**No single 'best' program exists**

For reviews, see:

P. Briffeuil *et al.*, **Bioinformatics** 1998, **14**:357

J. D. Thompson *et al.*, **NAR** 1999, **27**:2682

40

## **Multiple sequence alignment editors**

- EditSeq/MegAlign - Lasergene - Mac or MS-Windows
- DNA Strider - Macintosh
- Seq-AL - Macintosh
- ASAD - Excel - Macintosh or MS-Windows
- BioEdit - MS-Windows
- Genedoc - MS-Windows
- SeqPup - Mac. MS-Windows, X-Windows
- For a review of these:  
<http://www.wehi.edu.au/bioweb/KeithsStuff/seqeditors.html>

41