

Sequence comparison (short notes)

David Gilbert

Bioinformatics Research Centre

www.brc.dcs.gla.ac.uk

Department of Computing Science, University of Glasgow

Why compare sequences

How to compare them

Common programs: BLAST & FASTA

Why compare sequences?

- Assume a genome has been sequenced
- We can find out where “putative” genes are by gene-finding (see <http://www.ensembl.org/>)
- → What do such a gene do?
- We can make the protein that it encodes
 - but we can’t easily find out its biological *function*
- So, we can try to find other sequences which are *similar* to it, for which we know the function...

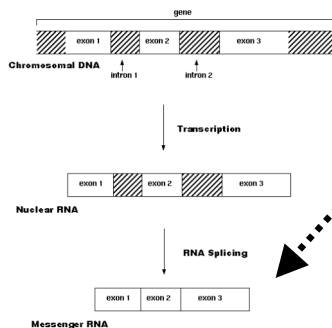
So, what is this sequence similar to?

Amino-acid →
(protein
sequence)

```
MVHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
KVKAHGKKVLGAFSDGLAHLNLDNLKGTFFATLSELHCDKLHVDPENFRLLGNVLCVLAHH
FGKEFTPPVQAAYQKVVAGVANALAHKYH
```

```
acatttgctt ctgacacaac tgtgttcact agcaacctca aacagacacc atggtgcacc
tgactcctga ggagaagtct gcggttactg ccctgtgggg caaggtgaac gtggatgaag
ttggtggtga ggccctgggc aggctgctgg tggcttacc ttggaccag aggttctttg
agtcctttgg ggatctgtcc actcctgatg cagttatggg caaccctaag gtgaaggctc
atggcaagaa agtgctcggg gccttttagtg atggcctggc tcacctggac aacctcaagg
gcacctttgc cactctgagt gagctgcact gtgacaagct gcacgtggat cctgagaact
tcaggctcct gggcaacgtg ctggtctgtg tgctggcca tcactttggc aaagaattca
ccccaccagt gcaggctgcc tatcagaaag tgggtggctg tgtggctaata gccctggccc
acaagtatca ctaagctcgc tttcttgctg tccaatttct attaaagggt cctttgttcc
ctaagtccaa ctactaaact gggggatatt atgaagggcc ttgagcatct ggattctgcc
taataaaaaa catttatttt cattgc
```

Transcription of DNA to Messenger RNA



cDNA (nucleotide sequence)
*Where does the coding start
on this sequence?*

Search using BLAST

<http://www.ncbi.nlm.nih.gov/BLAST/>

or

<http://www.ebi.ac.uk/blastall/>

(c) David Gilbert, 2003 [Sequence
Comparison]

Evolution - basic concepts

- Mutation in DNA a natural evolutionary process
- DNA *replication* errors: (nucleotide)
 - substitutions
 - insertions
 - deletions } **indels**
- Similarity between sequences
 - clue to common evolutionary origin, or
 - clue to common function
- This is a simplistic story: in fact the altered *function* of the expressed protein will determine if the organism will survive to reproduce, and hence pass on [transmit] the altered gene

The genetic code

First Position (5' end)	Second Position								Third Position (3' end)
	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T C A G
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	
	TTA	Leu	TCA	Ser	TAA	<i>Stop</i>	TGA	<i>Stop</i>	
	TTG	Leu	TCG	Ser	TAG	<i>Stop</i>	TGG	Trp	
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T C A G
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T C A G
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	
	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	
	ATG	Met*	ACG	Thr	AAG	Lys	AGG	Arg	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T C A G
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	

Evolution, DNA->Amino-acids

Triplet code, hence difference between DNA base

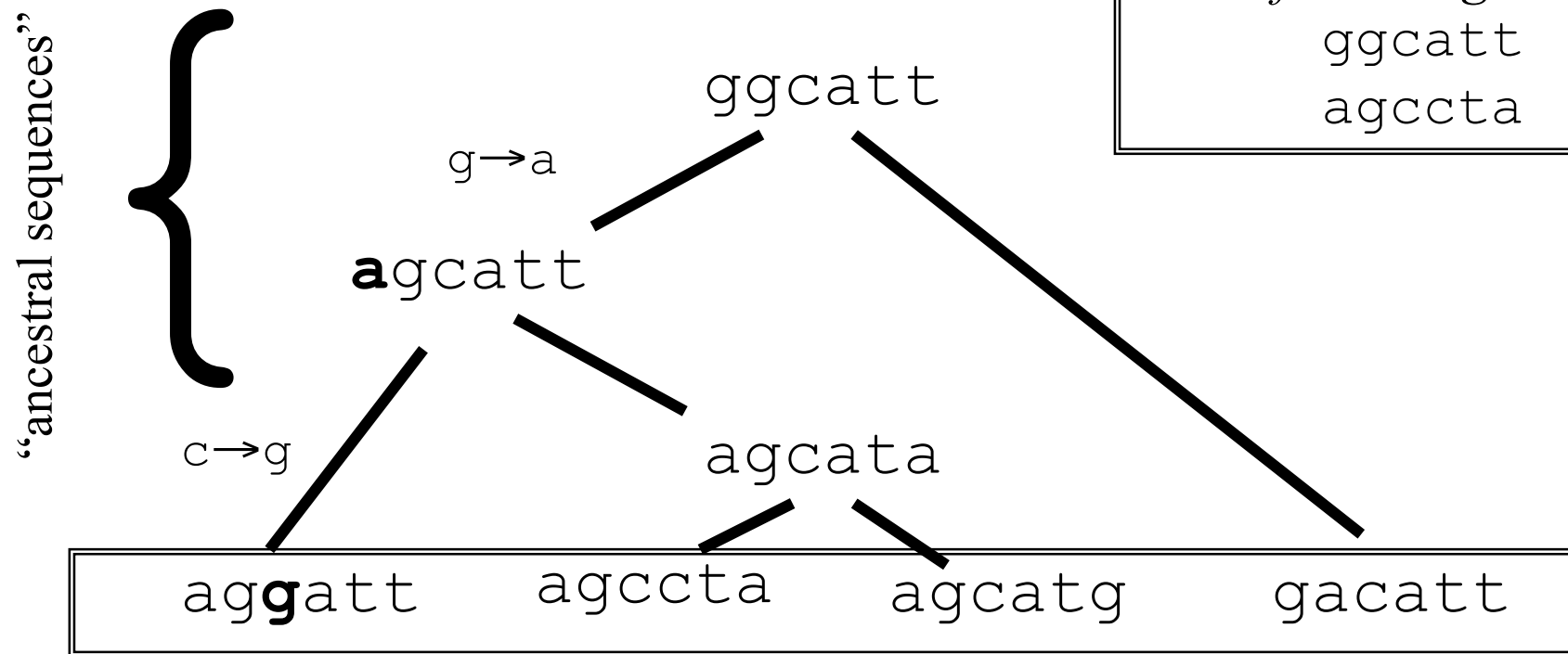
- Substitution: (hence 1 amino-acid changes)
- Insertion / Deletion: “frame shift” (all subsequent amino-acids change)
 - NB, Indels can be in multiples of 3, and hence...

Also

- “Silent mutation” - DNA changes but amino-acid doesn't change - why?
- “Nonsense mutation” - a single DNA base substitution resulting in a stop codon.

Evolution - related sequences

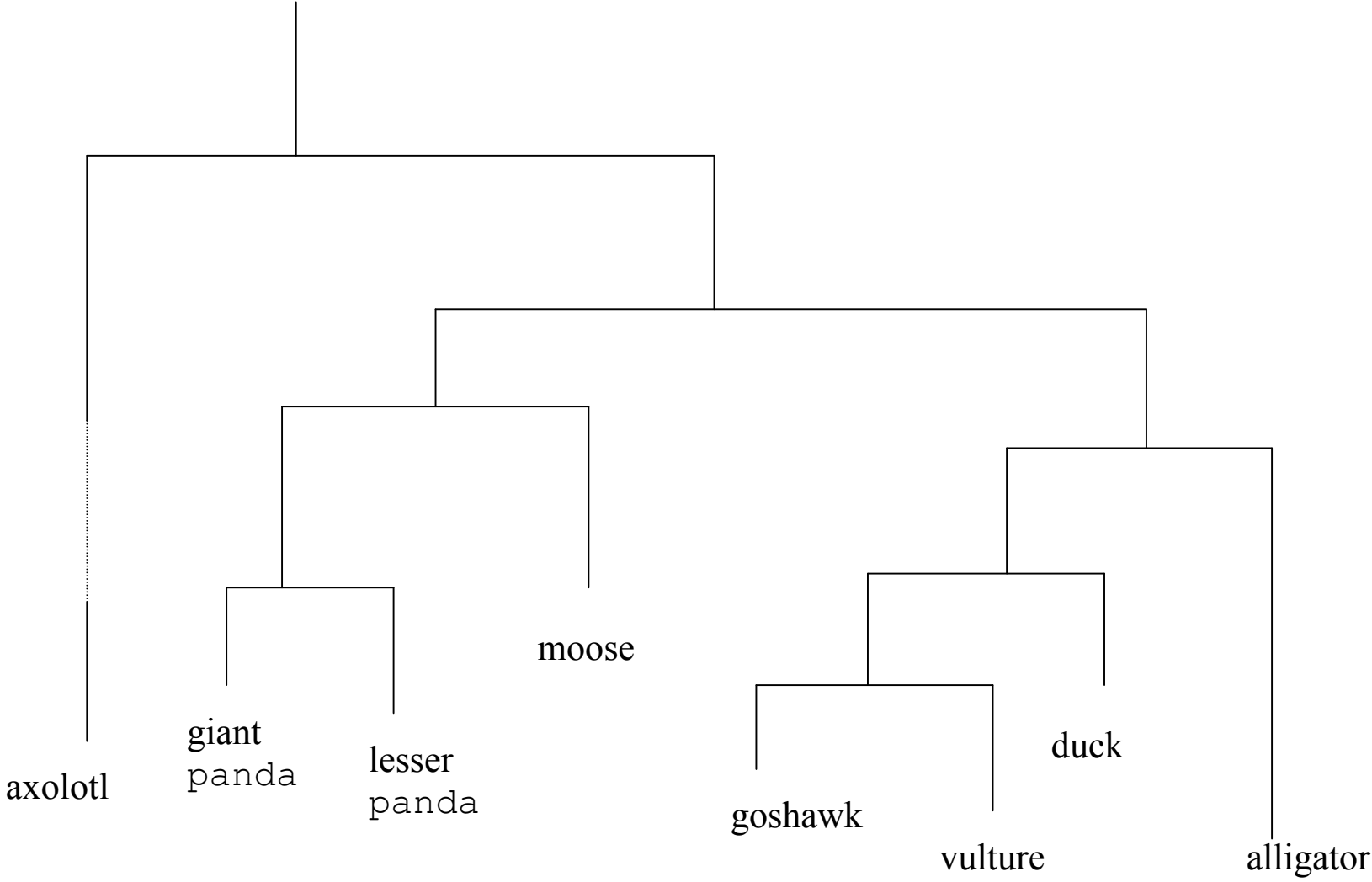
Highlight the other mutations!



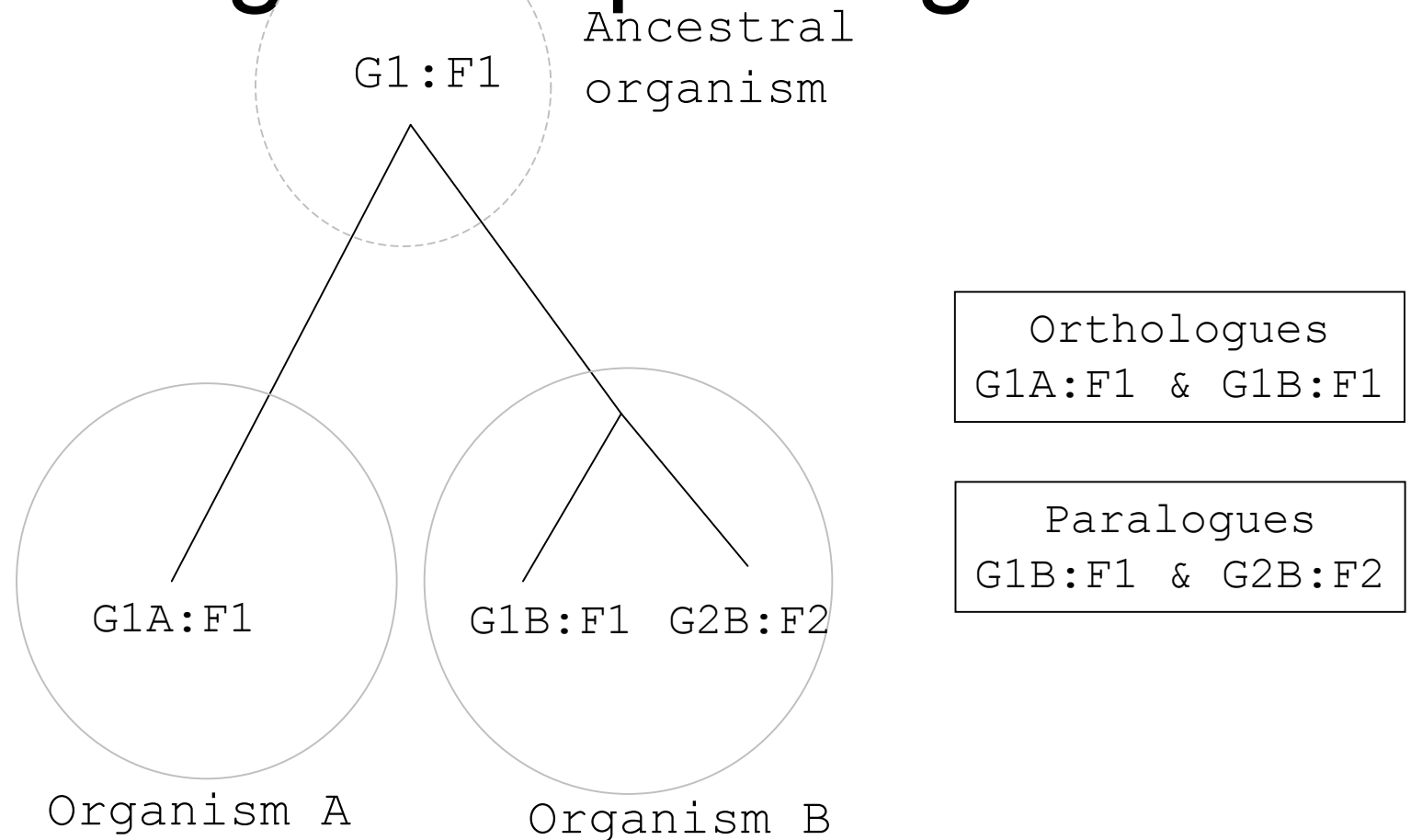
Q: How many changes between 2 sequences?

“living examples”

Tree of orthologues based on a set of α -haemoglobins

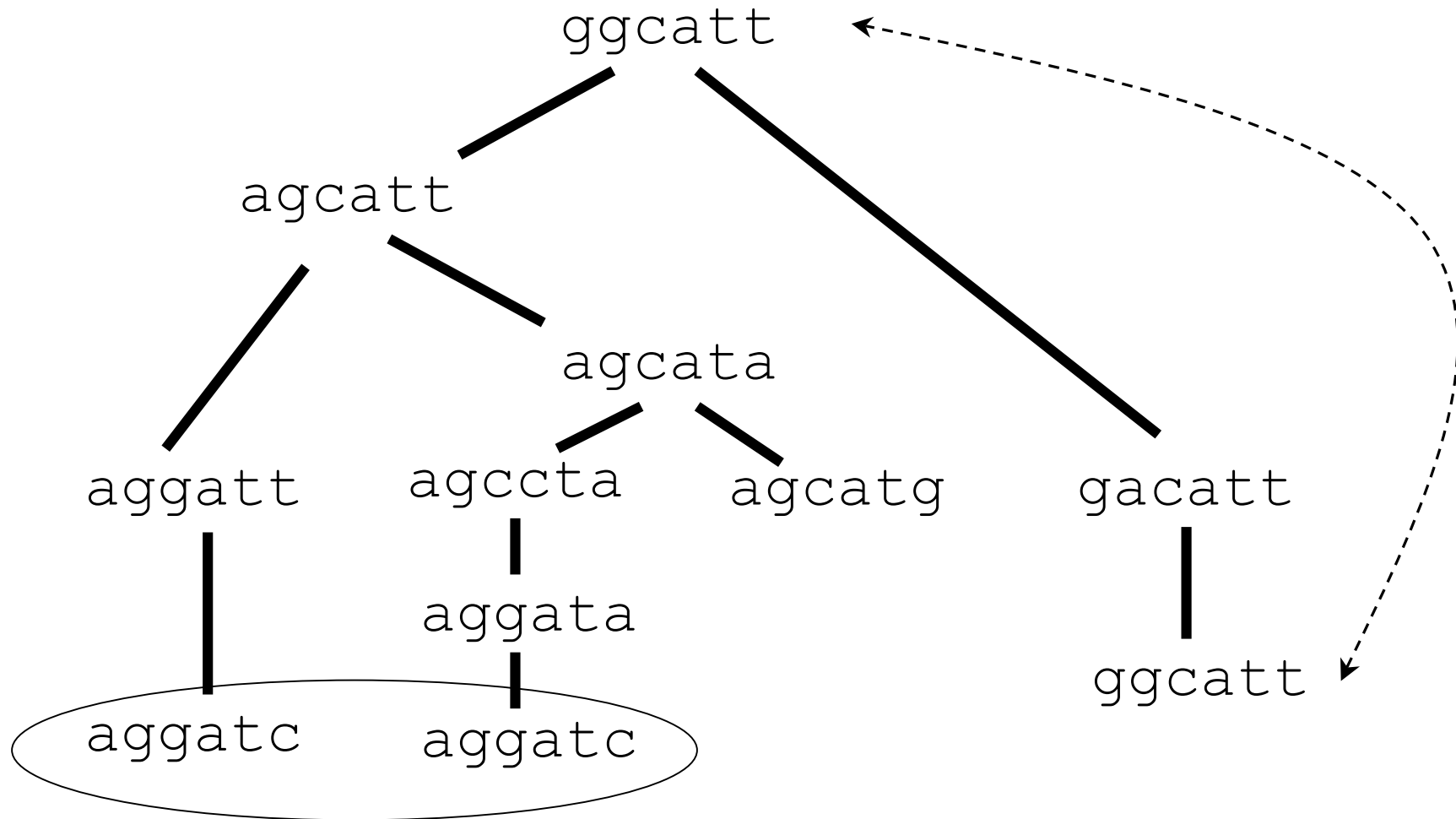


Orthologues & paralogues



Other evolutionary issues

- Convergent evolution: same sequence evolved from different ancestors
- back evolution - mutate to a previous sequence



Edit distance

- Levenshtein 1966
- Minimum number of edit operations to transform 1 string into another
 - insert, delete, (*substitute*) (1 symbol)
- Distance is zero (identical) or positive
- E.g “AIMS” & “AMOS”



(distance=2 for each solution)

Edit distance - Inserts only

Given two strings $V=v_1 \dots v_i$ and $W=w_1 \dots w_j$

$$d(i, j) = \min \begin{cases} d(i-1, j) + 1 & \text{insert in } V \\ d(i, j-1) + 1 & \text{insert in } W \\ d(i-1, j-1) & \text{if } v_i = w_j \text{ match} \end{cases}$$

$$d(i, 0) = i$$

$$d(0, j) = j$$

Naive implementation:
complexity exponential in i and j

? edit distance of strings 's' and 't' : $d('s', 't')$?

? $d(V, V) = ?$

What do we want to know about 2 sequences?

- *Similarity* measure (0 if identical, else >0)

`edit_dist('AIMS', 'AMOS') = 2`

`edit_dist(S,S) = ?`

?edit_dist(S1,S2) always unique?

- *Longest common subsequence (LCS)*: the sequence of nucleotides/amino acids that they have in common.

`LCS('AIMS', 'AMOS') = A.M.S`

`LCS(S,S) = ?`

?LCS(S1,S2) always unique?

- *length of their LCS*

`len(LCS('AIMS', 'AMOS')) = 3`

`len(LCS(S,S)) = ?`

?len(LCS(S1,S2)) always

unique?

Percentage sequence identity

$$\text{Percentage sequence identity} = \frac{\text{number of identical residues} \times 100}{\text{number of residues in smallest sequence}}$$

Can differ if have gaps/no_gaps:
compute for these sequences:

TGCATA
ATCTGAT

-TGCAT-A-
AT-C-TGAT

For each case, what is the

- LCS
- LCS score
- Sequence Identity %

Substitution matrices

	A	C	G	T
A				
C				
G				
T				

- Unitary matrix: match=1, mismatch=0
 - sparse matrix (most elements are 0)
- Poor diagnostic power
 - all identical matches carry identical weighting
- We can enhance scoring potential of weak but biologically significant signals
- Scoring matrices - weight matches for non-identical residues according to observed substitution rates.

PAM 250 matrix

X=0

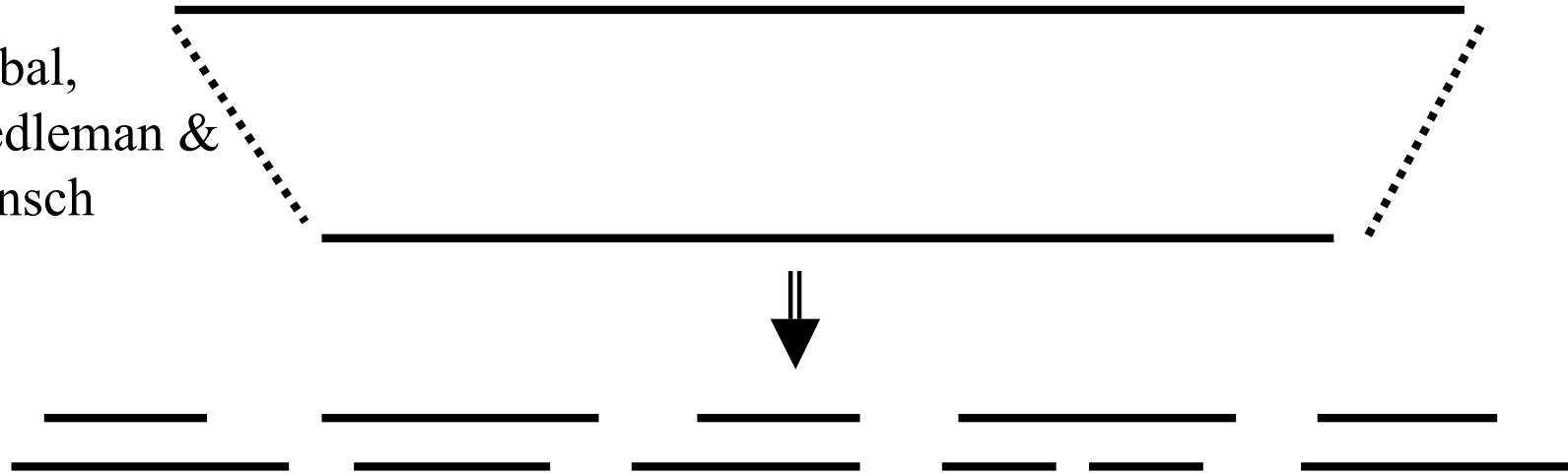
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
W	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Y	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	W	Y

Global and local alignment

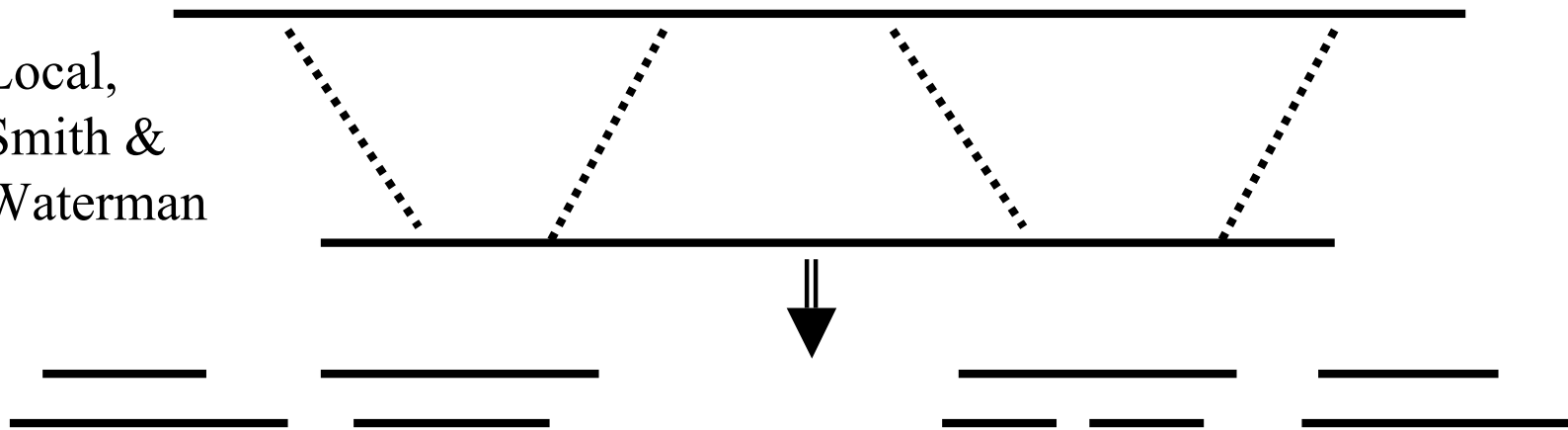
- Global alignment - as per dynamic programming solution
 - Needleman & Wunsch algorithm (1970)
- Local alignment - find local regions from each string which are similar:
 - Corresponds to shorter, localised paths in the matrix.
 - Justification - biological functional sites localised to short conserved regions (no indels/mutations).
 - Smith-Waterman algorithm (1981)

Global vs local alignment

Global,
Needleman &
Wunsch



Local,
Smith &
Waterman



FASTA (Lipman & Pearson 1985)

- Local alignments - tries to find paths of regional similarity, rather than trying to find the best alignment between 2 sequences.
- Alignments can contain gaps.
- Rapid
- Heuristic - not guaranteed to find the best alignment between 2 sequences; it may miss matches.
 - uses a strategy which is expected to find most matches, but sacrifices complete sensitivity in order to gain speed.
- A substitution matrix is used during all phases of protein searches (BLASTP, BLASTX, TBLASTN)

BLAST - Basic Local Alignment Tool

Altschul et al 1990

- Given 2 sequences:
 - Segment pair - pair of subsequences of the same length forming an ungapped alignment
 - Computes all segment pairs
 - If there is a MSP maximal segment pair (highest score of all pairs for 1 comparison) above some cutoff score C and C is “significant” then report hit
 - Also reports those sequences where the score of $MSP < C$, but several segment pairs in combination which are significant.
 - Reports score from highest scoring pairs & probability scores [E values] (expected by chance).
- <http://www.ncbi.nlm.nih.gov/BLAST/>
- <http://www.ebi.ac.uk/blastall/>

BLAST - Edited results (EMBL)

Database: embl: 958,670 sequences; 2,466,994,978 total letters

Sequences producing significant alignments:	Score (bits)	E Value
EM_HUM:HSBGL1 V00497 Human messenger RNA for beta-globin.	1241	0.0
EM_HUM:AF181989 AF181989 Homo sapiens hemoglobin beta subuni...	1116	0.0
EM_HUM:HSHEMOB M25113 Human sickle beta-hemoglobin mRNA.	1100	0.0
EM_PAT:I32884 I32884 Sequence 9 from patent US 5589367.	910	0.0
EM_HUM:HS202231 U20223 Human thalassemia beta globin gene, c...	416	e-114
EM_OM:AGHBD M19061 Spider monkey (A.geoffroyi) delta-globin ...	369	1e-99
EM_OM:CPHBB5CP J00330 monkey (c.polykomos) beta-globin gene;...	367	4e-99
EM_OM:PPHBD M21825 Orangutan delta globin gene, complete cds.	347	4e-93
EM_OM:CPHBDPSC J00335 Monkey (colobus) delta-globin pseudoge...	297	3e-78
EM_OM:LMHBB M15734 Lemur (brown) beta-globin gene, complete ...	270	7e-70
EM_OM:TSHBD J04428 T.syrichta delta-globin gene, complete cds.	266	1e-68
EM_OM:OCU60902 U60902 Otolemur crassicaudatus epsilon-, gamm...	266	1e-68
EM_OM:LEBGLOB Y00347 Lepus europaeus adult beta-globin gene	266	1e-68
EM_OM:GCDELGLB M61740 G.crassicaudatus beta globin gene, com...	266	1e-68
EM_OM:MOHBDPS J00332 monkey (anubis) silent delta-globin gene.	262	2e-67
EM_OM:TSHBB J04429 T.syrichta beta globin gene, complete cds.	260	7e-67
EM_PAT:A34698 A34698 Synthetic pSXBeta+ sequence	258	3e-66
EM_OM:OCBGLO V00882 Rabbit (O. cuniculus) gene for beta-globin.	250	7e-64
EM_OM:BTBG M63453 Bovine Beta globin gene and globin (PSI-3)...	220	6e-55

BLAST - Edited results (Swiss-prot)

Database: swissprot: 86,593 sequences; 31,411,157 total letters

Sequences producing significant alignments:					Score	E	
					(bits)	Value	
SW:HBB_HUMAN	P02023	HEMOGLOBIN	BETA	CHAIN.	(human)	306	2e-83
SW:HBB_GORGO	P02024	HEMOGLOBIN	BETA	CHAIN.	(gorilla)	305	4e-83
SW:HBB2_PANLE	P18988	HEMOGLOBIN	BETA-2	CHAIN.	(lion)	302	3e-82
SW:HBB_HYLLA	P02025	HEMOGLOBIN	BETA	CHAIN.	(gibbon)	300	8e-82
SW:HBB_PREEN	P02032	HEMOGLOBIN	BETA	CHAIN.	(Hanumam langur)	298	5e-81
SW:HBB_COLPO	P19885	HEMOGLOBIN	BETA	CHAIN.	(Colobus)	295	3e-80
SW:HBB_CERAE	P02028	HEMOGLOBIN	BETA	CHAIN.	(Green monkey)	295	3e-80
SW:HBB_MACFU	P02027	HEMOGLOBIN	BETA	CHAIN.	(Japanese macaque)	293	2e-79
SW:HBB_CALAR	P18985	HEMOGLOBIN	BETA	CHAIN.	(Marmoset)	292	2e-79
SW:HBB_ATEGE	P02034	HEMOGLOBIN	BETA	CHAIN.	(Spider monkey)	292	2e-79
SW:HBB_MANSP	P08259	HEMOGLOBIN	BETA	CHAIN.	(Mandrill)	291	4e-79
...							
SW:HBB1_RAT	P02091	HEMOGLOBIN	BETA	CHAIN,	(Rat)	255	4e-68
SW:HBB_ERIEU	P02059	HEMOGLOBIN	BETA	CHAIN.	(Hedgehog)	252	2e-67
SW:HBB_PANPO	P04244	HEMOGLOBIN	BETA	CHAIN.	(Bison)	251	5e-67
SW:HBB_BISBO	P09422	HEMOGLOBIN	BETA	CHAIN.	(Leopard)	251	5e-67

Blast alignment

Blast output

>SW:HBB_CANFA P02056 HEMOGLOBIN BETA CHAIN.

Length = 146

Score = 276 bits (698), Expect = 2e-74

Identities = 131/146 (89%), Positives = 137/146 (93%)

```
Query: 2   VHLTPEEKSAVTALWGKVVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 61
          VHLT EEKS V+ LWGKVVDEVGGEALGRLL+VYPWTQRFF+SFGDLSTPDAVM N KV
Sbjct: 1   VHLTAEKSLVSGLWGKVVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAKV 60

Query: 62  KAHGKKVLGAFSDGLAHL DNLKGTFA TLSELHCDKLHVDPENFRLLGNVLCVLAHHFGK 121
          KAHGKKVL +FSDGL +LDNLKGTFA LSELHCDKLHVDPENF+LLGNVLCVLAHHFGK
Sbjct: 61  KAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPENFKLLGNVLCVLAHHFGK 120

Query: 122 EFTPPVQAAYQKVVAGVANALAHKYH 147
          EFTP VQAAYQKVVAGVANALAHKYH
Sbjct: 121 EFTPQVQAAYQKVVAGVANALAHKYH 146
```