

## Sequence comparison (short notes)

David Gilbert

Bioinformatics Research Centre

[www.brc.dcs.gla.ac.uk](http://www.brc.dcs.gla.ac.uk)

Department of Computing Science, University of Glasgow

Why compare sequences  
How to compare them  
Common programs: BLAST & FASTA

(c) David Gilbert, 2003 [Sequence Comparison]

1

## 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...

(c) David Gilbert, 2003 [Sequence Comparison]

2

## So, what is this sequence similar to?

Amino-acid (protein sequence) → MVHLTPEEKSAVNTALWGKVNVEVGGGALGRLLVYVFWTQRFESFGDLSTPDAVMGNPKVKAHGKKV LGA FSDGLAHLNLDLKGTFATLSELHCDKLVDPENFRLLGNLVLCVLAHHEFGKEFTFPVQAAYQKVVAGVANALAHKYH

```
acatttgctt ctgacacaa cgtgttca ctgacacaa atggtgcacc
tgactcctga ggagaagtct gcggttactg cctctgggg caaggtgaac gtggatgaag
ttggtggtga ggccctgggc aggetgctgg tggctaccc ttggaccag aggttctttg
agtcccttgg ggaactgtcc actcctgatg cagttatgg caaccctaa gtgaaggctc
atggcaagaa agtgcctcgt gcttttagt atggcctggc tcactggac aacctcaagg
gcacctttgc cacactgagt gagctgact gtgacaagct gcactggat cctgagaact
tcaggctcct gggcaactgt ctggtctgtg tctgtggcca tcactttggc aaagaattca
ccccaccagt gcaggctgcc taccagaaag tgggtgctgg tgtggttaat gccctggccc
acaagtatca ctaagctcgt ttctctgtg tccaattctc attaaagtt cctttgtccc
ctaagtcocaa ctactaaact gggggatatt atgaaggccc 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]

3

## Evolution - basic concepts

- Mutation in DNA a natural evolutionary process
- DNA *replication* errors: (nucleotide)
  - substitutions
  - insertions } **indels**
  - deletions }
- 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

(c) David Gilbert, 2003 [Sequence Comparison]

4

## The genetic code

First Position (5' end)	Second Position						Third Position (3' end)
	T	C	A	G	C	G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	Cys
	TTC	Phe	TCC	Ser	TAC	Tyr	Cys
	TTA	Leu	TCA	Ser	TAA	Stop	Stop
	TTG	Leu	TCG	Ser	TAG	Stop	Trp
C	CTT	Leu	CTT	Pro	CAT	His	Arg
	CTC	Leu	CCC	Pro	CAC	His	Arg
	CTA	Leu	CCA	Pro	CAA	Gln	Arg
	CTG	Leu	CCG	Pro	CAG	Gln	Arg
A	ATT	Ile	ACT	Thr	AAT	Asn	Ser
	AIC	Ile	ACC	Thr	AAC	Asn	Ser
	ATA	Ile	ACA	Thr	AAA	Lys	Arg
	ATG	Met*	ACG	Thr	AAG	Lys	Arg
G	GTT	Val	GCT	Ala	GAT	Asp	Gly
	GTC	Val	GCC	Ala	GAC	Asp	Gly
	GTA	Val	GCA	Ala	GAA	Glu	Gly
	GTG	Val	GCG	Ala	GAG	Glu	Gly

(c) David Gilbert, 2003 [Sequence Comparison]

5

## 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

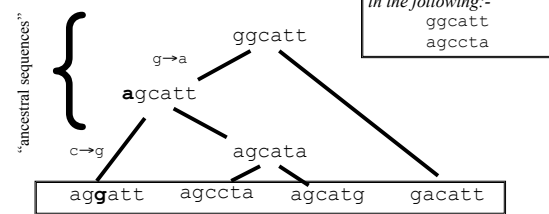
- “Silent mutation” - DNA changes but amino-acid doesn’t change - why?
- “Nonsense mutation” - a single DNA base substitution resulting in a stop codon.

(c) David Gilbert, 2003 [Sequence Comparison]

6

## Evolution - related sequences

Highlight the other mutations!

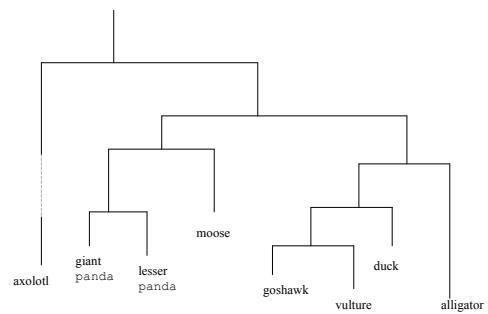


Q: How many changes between 2 sequences? "living examples"

(c) David Gilbert, 2003 [Sequence Comparison]

7

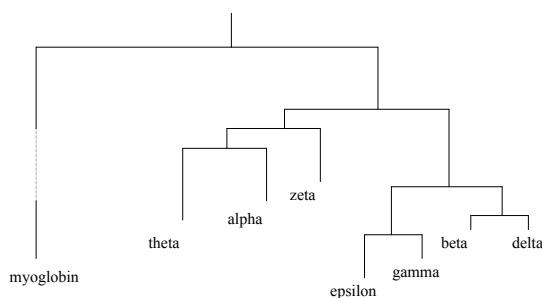
## Tree of orthologues based on a set of $\alpha$ -haemoglobins



(c) David Gilbert, 2003 [Sequence Comparison]

8

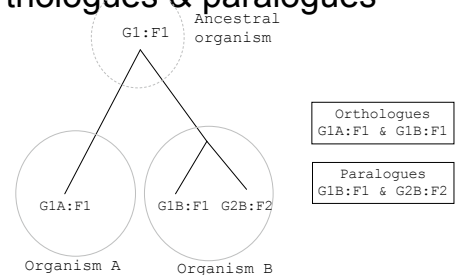
## Tree of paralogues: human haemoglobins and myoglobin



(c) David Gilbert, 2003 [Sequence Comparison]

9

## Orthologues & paralogues

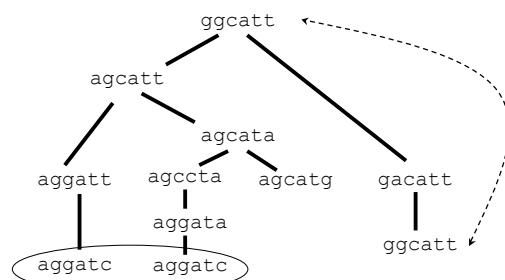


(c) David Gilbert, 2003 [Sequence Comparison]

10

## Other evolutionary issues

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

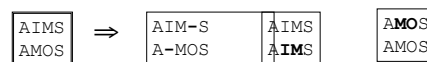


(c) David Gilbert, 2003 [Sequence Comparison]

11

## 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)

(c) David Gilbert, 2003 [Sequence Comparison]

12

## 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':  $d('s', 't')$ ?

?  $d(V, V) = ?$

(c) David Gilbert, 2003 [Sequence Comparison]

13

## What do we want to know about 2 sequences?

- **Similarity** measure (0 if identical, else >0)  
 $\text{edit\_dist}('AIMS', 'AMOS') = 2$   
 $\text{edit\_dist}(S, S) = ?$  ?*edit\_dist(S1, S2) always unique?*
- **Longest common subsequence (LCS)**: the sequence of nucleotides/amino acids that they have in common.  
 $\text{LCS}('AIMS', 'AMOS') = A.M.S$   
 $\text{LCS}(S, S) = ?$  ?*LCS(S1, S2) always unique?*
- **length of their LCS**  $\text{len}(\text{LCS}('AIMS', 'AMOS')) = 3$   
 $\text{len}(\text{LCS}(S, S)) = ?$  ?*len(LCS(S1, S2)) always unique?*

(c) David Gilbert, 2003 [Sequence Comparison]

14

## 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 %

(c) David Gilbert, 2003 [Sequence Comparison]

15

## 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.

(c) David Gilbert, 2003 [Sequence Comparison]

16

## PAM 250 matrix

```

X=0
C 12
S 0 2
T -2 1 3
F -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 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 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
    
```

(c) David Gilbert, 2003 [Sequence Comparison]

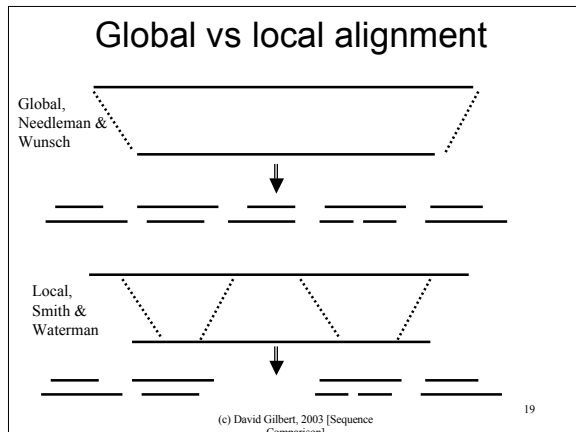
17

## 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)

(c) David Gilbert, 2003 [Sequence Comparison]

18



## 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)

(c) David Gilbert, 2003 [Sequence Comparison] 20

## 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/>

(c) David Gilbert, 2003 [Sequence Comparison] 21

## BLAST - Edited results (EMBL)

Database: emb1: 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.syrichtha delta-globin gene, complete cds.	266	1e-68
EM_OM:OCU60902 U60902 Otolemur crassicaudatus epsilon-, gamm...	266	1e-68
EM_OM:LEBGLB Y00347 Lepus europaeus adult beta-globin gene	266	1e-68
EM_OM:OCBGLB 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.syrichtha beta globin gene, complete cds.	260	7e-67
EM_PAT:A34698 A34698 Synthetic pSXBeta+ sequence	258	3e-66
EM_OM:OCBGL0 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

(c) David Gilbert, 2003 [Sequence Comparison] 22

## BLAST - Edited results (Swiss-prot)

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

Sequences producing significant alignments:	Score (bits)	E Value
SW:HBB HUMAN P02023 HEMOGLOBIN BETA CHAIN. (human)	306	2e-83
SW:HBB GORGO P02024 HEMOGLOBIN BETA CHAIN. (gorilla)	305	4e-83
<b>SW:HBB2 PANLE P18988 HEMOGLOBIN BETA-2 CHAIN. (lion)</b>	<b>302</b>	<b>3e-82</b>
SW:HBB HYLLA P02025 HEMOGLOBIN BETA CHAIN. (gibbon)	300	8e-82
SW:HBB PREEN P02032 HEMOGLOBIN BETA CHAIN. (Hanuman 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 MANSF 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 PANFO P04244 HEMOGLOBIN BETA CHAIN. (Bison)	251	5e-67
SW:HBB BISBO P09422 HEMOGLOBIN BETA CHAIN. (Leopard)	251	5e-67

(c) David Gilbert, 2003 [Sequence Comparison] 23

## 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 VHLTPEERSAVTALWGKVVNDEVGGEALGRLLVYVFWTQRFFSFGDLSTFDVAVMGNPKV 61
      VHLT EERS V+ LWGKVVNDEVGGEALGRLL+YVFWTQRFF+SFGLSTFDVAVM N KV
Sbjct: 1 VHLTAEEKSLVSLGKVVNDEVGGEALGRLLIVYVFWTQRFFSFGDLSTFDVAVMSNAKV 60

Query: 62 KAHGKVLGAFSDGLAHLNLDLNGTFAFLSELHCKDLHVDPENFRLLGNVLCVLAHHPGK 121
      KAHGKVL +FSDGL +LDNLKGTFA LSELHCKDLHVDPENF+LLGNVLCVLAHHPGK
Sbjct: 61 KAHGKVLNLSFSDGLKLNLDLNGTFAFLSELHCKDLHVDPENFKLLGNVLCVLAHHPGK 120

Query: 122 EFTFPVQAAYQKVVAGVANALAHKYH 147
      EFTP VQAAYQKVVAGVANALAHKYH
Sbjct: 121 EFTFPVQAAYQKVVAGVANALAHKYH 146
```

(c) David Gilbert, 2003 [Sequence Comparison] 24