

Multiple Sequence Alignment Theory & Practice

Peter FitzGerald & Susan Chacko

NCI & CIT

Outline

- Introduction to MSA
 - What is it ?
 - What is it good for ?
 - How do I use it ?
- Software and algorithms
 - The programs
 - How they work
 - Which to use
 - Editing & publishing
- Conclusion & Recommendations
- Multiple Genome Alignment

What is Multiple Sequence Alignment (MSA) ?

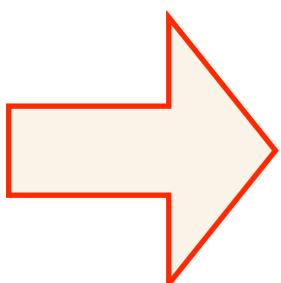
chicken	P L V S S - - - P L R G E A G V L P F Q Q E E Y E K V K R G I V E Q C C H N T C S L Y Q L E N Y C N
xenopus	A L V S G - - - P Q D N E L D G M Q L Q P Q E Y Q K M K R G I V E Q C C H S T C S L F Q L E S Y C N
human	L Q V G Q V E L G G G P G A G S L Q P L A L E G S L Q K R G I V E Q C C T S I C S L Y Q L E N Y C N
monkey	P Q V G Q V E L G G G P G A G S L Q P L A L E G S L Q K R G I V E Q C C T S I C S L Y Q L E N Y C N
dog	L Q V R D V E L A G A P G E G G L Q P L A L E G A L Q K R G I V E Q C C T S I C S L Y Q L E N Y C N
hamster	P Q V A Q L E L G G G P G A D D L Q T L A L E V A Q Q K R G I V D Q C C T S I C S L Y Q L E N Y C N
cow	P Q V G A L E L A G G P G A G G - - - - L E G P P Q K R G I V E Q C C A S V C S L Y Q L E N Y C N
guinea pig	P Q V E Q T E L G M G L G A G G L Q P L A L E M A L Q K R G I V D Q C C T G T C T R H Q L Q S Y C N

* . * * * * : * * * . * : . * * : . * * *

Why do a Multiple Sequence Alignment ?

What's the end goal ?

- Simple sequence comparison
- Conserved *vs.* non-conserved regions
 - proteins - motifs/profiles
 - whole genome - genes, control regions
- Homology (as opposed to similarity)
 - Evolution - phylogeny
 - Structural homology
- Sequence differences
 - Single Nucleotide Polymorphisms (SNPs)



Subsets of Functions

- Multiple Alignment
- Multiple Sequence Editing
- Generating/drawing trees
- Publishing - high quality output
- Structure interface (CN₃D)

Pre-computed MSAs

- DALI/FSSP

<http://www2.ebi.ac.uk/dali/>

- InterPro

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

- PROSITE, PRINTS

<http://us.expasy.org/prosite/>

- CDD, SMART, PFAM, COG

<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>

- VAST

<http://www.ncbi.nlm.nih.gov:80/Structure/VAST/vastsearch.html>

Domain/Profile Construction

- PSI-BLAST

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

- MEME/MAST

http://meme.sdsc.edu/meme/website/intro.html

- BLOCKS

http://www.blocks.fhcrc.org/

- PRATT

http://us.expasy.org/tools/pratt/

- HMMER

http://hmmer.wustl.edu/

Generating an Alignment

- Get the sequences
 - ◆ Reformat them
- Align the sequences
 - ◆ Evaluate the alignment
 - ◆ Realign or modify the alignment
 - ◆ Add/subtract sequence
- Analyze, publish, draw phylogenetic trees, connect to structures

Collecting the Sequences

- Selection of sequences is important
- Most programs will align ***ANYTHING***
- All sequences should be related
- Avoid redundancy
- Diverse set of sequences is best

Sequence Selection

- Common source of sequences is blast output
- Entrez searches
- Many pre-aligned
- Personal sequences

Sequence Format

- Several multiple sequence formats
- Format selection is important for input and output
- Different programs *like* (**need**) different formats
- Reformatting software
 - <http://molbio.info.nih.gov/molbio/gcglite/reformat.html>*
 - <http://genome.nci.nih.gov/tool/reformat.html>*
- Output format determined by next step

Sequence formats

(sequential)

>**chiins** *insulin2.msf*, 107 aa.

BALWIRSLPLLALLVFSGPGETSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ
PLVSS---PLRGEAGVLPFQQEEYEVKRGIVEQCCHNTCSLYQLENYCN

>**xenins** *insulin2.msf*, 106 aa.

BALWMQCLPLVLVLFSTPNTE-ALVNQHLCGSHLVEALYLVCGRGFFYYPKVKRDMEQ
ALVSG---PQDNELDGMQLQPQEYQKMKRGIVEQCCHSTCSLFQLESYCN

>**humins** *insulin2.msf*, 110 aa.

BALWMRLLPPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGPGAGSLQPLALE GSLQKRGIVEQCCTSICSLYQLENYCN

>**monins** *insulin2.msf*, 110 aa.

BALWMRLLPPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
PQVGQVELGGPGAGSLQPLALE GSLQKRGIVEQCCTSICSLYQLENYCN

>**dogins** *insulin2.msf*, 110 aa.

MALWMRLLPPLLALLALWAPAPTRAFFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVED
LQVRDVELAGAPGEGGLQPLALE GALQKRGIVEQCCTSICSLYQLENYCN

>**hamins** *insulin2.msf*, 110 aa.

MTLWMRLLPLLTLLVLWEPNPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRRGVED
PQVAQLELGGPGADDLQT LALEVAQQKRGIVDQCCTSICSLYQLENYCN

>**bovins** *insulin2.msf*, 105 aa.

MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG
PQVGALELAGGPGAGG----LEGPPQKRGIVEQCASVCSLYQLENYCN

>**guiins** *insulin2.msf*, 110 aa.

MALWMHLLTVLALLALWGPNTGQAFVSRHLCGSNLVETLYSVCQDDGFFYIPKDRRELED
PQVEQTELGMGLGAGGLQPLALEMALQKRGIVDQCCTGTCTRHLQSYCN

ClustalW

(interlaced)

CLUSTAL w (1.74) multiple sequence alignment

chiins	BALWIRSLPLLALLVSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ
xenins	BALWMQCLPLVLVLFFSTPNTE-ALVNQHLCGSHLVEALYLVCGDRGFFYYPKVKRDMEQ
humins	BALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
monins	BALWMRLLPLLALLALWGPDPVPAPFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
dogins	MALWMRLLPLLALLALWAPAPTRAFFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVED
hamins	MTLWMRLLPLLTLLVLWEPEPNPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRRGVED
bovins	MALWTRLRPLLALLALWPPPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG
guiins	MALWMHLLTVLALLALWGPNTGQAFVSRHLCGSNLVETLYSVCQDDGFFYIPKDRRELED

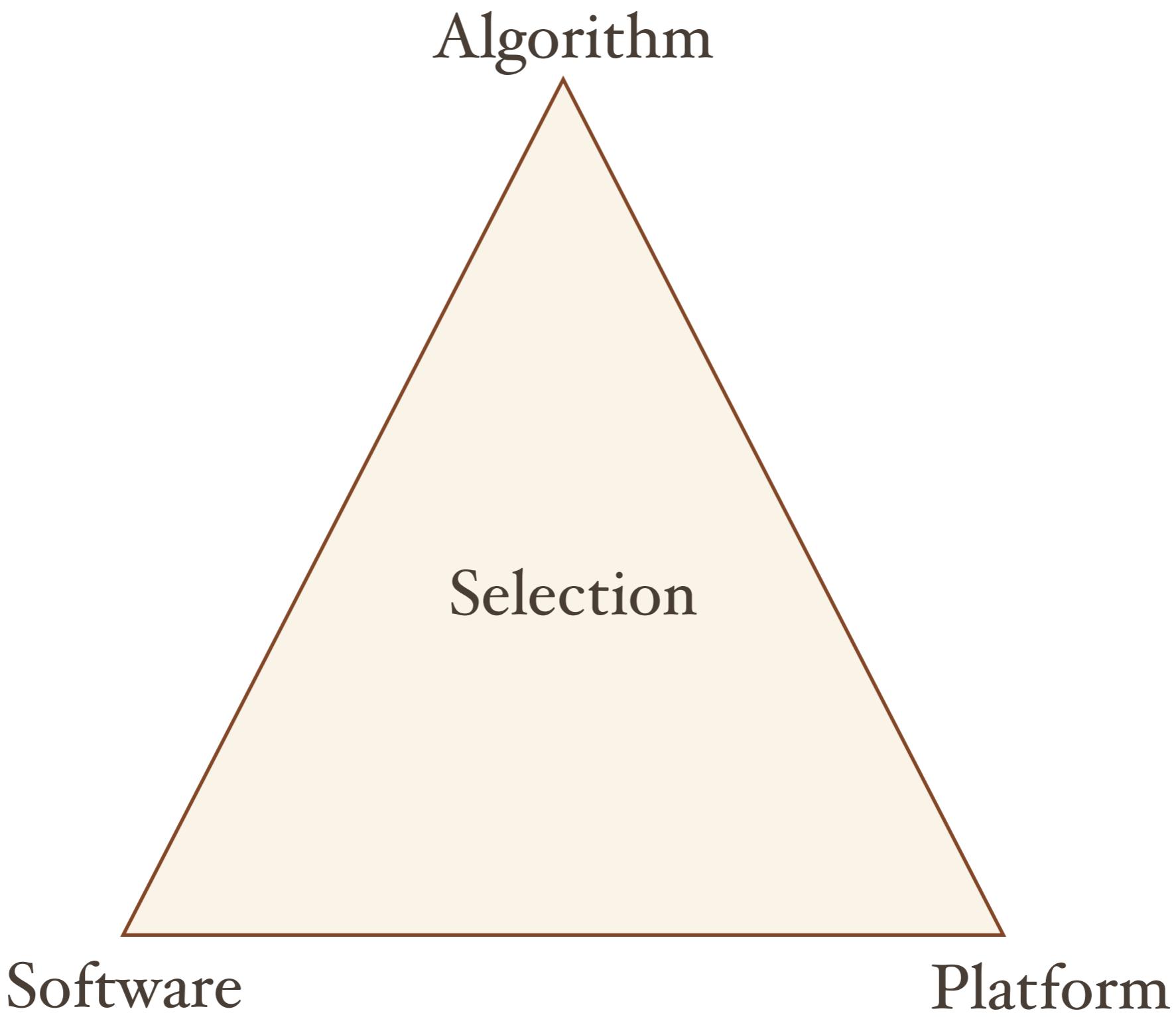
chiins	PLVSS---PLRGEAGVLPFQQEEYEVKRGIVEQCCHNTCSLYQLENYCN
xenins	ALVSG---PQDNELDGMQLQPQEYQKMKRGIVEQCCHSTCSLFQLESYCN
humins	LQVGQVELGGPGAGSLQPLALE GSLQKRGIVEQCCTSICSLYQLENYCN
monins	PQVGQVELGGPGAGSLQPLALE GSLQKRGIVEQCCTSICSLYQLENYCN
dogins	LQVRDVELAGAPGEGGLQPLALE GALQKRGIVEQCCTSICSLYQLENYCN
hamins	PQVAQLELGGPGADDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN
bovins	PQVGALELAGGPGAGG-----LEGPPQKRGIVEQCASVCSLYQLENYCN
guiins	PQVEQTELGMGLGAGGLQPLALE MALQKRGIVDQCCTGTCTRHLQLQSYCN

GCG - MSF - Pileup

PileUp

MSF: 110 Type: P Check: 4380 ..	
Name: chiins oo	Len: 110 Check: 3857 Weight: 0.212
Name: xenins oo	Len: 110 Check: 4552 Weight: 0.050
Name: humins oo	Len: 110 Check: 4867 Weight: 0.050
Name: monins oo	Len: 110 Check: 5690 Weight: 0.080
Name: dogins oo	Len: 110 Check: 3667 Weight: 0.111
Name: hamins oo	Len: 110 Check: 5715 Weight: 0.111
Name: bovins oo	Len: 110 Check: 845 Weight: 0.232
Name: guiins oo	Len: 110 Check: 5187 Weight: 0.100
//	
chiins	BALWIRSLPL LALLVFSGP G TSYAAANQHL CGSHLVEALY LVCGERGFFY
xenins	BALWMQCLPL VLVLFFSTPN TE.ALVNQHL CGSHLVEALY LVCGDRGFFY
humins	BALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY
monins	BALWMRLLPL LALLALWGPD PVPAFVNQHL CGSHLVEALY LVCGERGFFY
dogins	MALWMRLLPL LALLALWAPA PTRAFVNQHL CGSHLVEALY LVCGERGFFY
hamins	MTLWMRLLPL LTLLVLWE PN PAQAFVNQHL CGSHLVEALY LVCGERGFFY
bovins	MALWTRLRPL LALLALWPPP PARAFVNQHL CGSHLVEALY LVCGERGFFY
guiins	MALWMHLLTV LALLALWGPN TGQAFVSRHL CGSNLVETLY SVCQDDGFFY
 chiins	SPKARRDVEQ PLVSS...PL RGEAGVLPF Q QEEYEVKRG IVEQCCHNTC
xenins	YPKVKR DMEQ ALVSG...PQ DNELDGMQLQ POEYQKMKRG IVEQCCHSTC
humins	TPKTRREAED LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC
monins	TPKTRREAED PQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC
dogins	TPKARREVED LQVRDVELAG APGEGGLQPL ALEGALQKRG IVEQCCTSIC
hamins	TPKSRRGVED PQVAQLELGG GPGADDLQTL ALEVAQQKRG IVDQCCTSIC
bovins	TPKARREVEG PQVG ALELAG GPGAGG.... LEGPPQKRG IVEQCCASVC
guiins	IPKDRRELED PQVEQTELGM GLGAGGLQPL ALEMALQKRG IVDQCCTGTC

Generating the Alignment



Platform - selection

Choice dependent on availability, complexity and personal preference



Central server



Web-based



Local computer

Software - selection

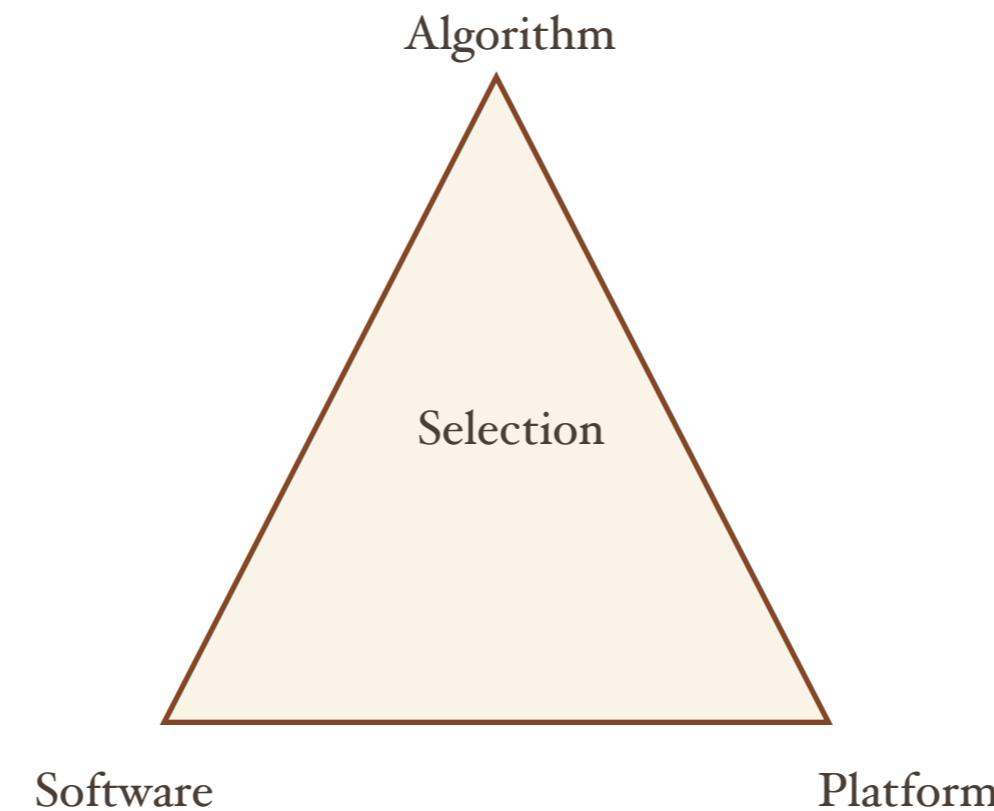
Choice dependent on ease of use and availability

- The best
- What's available
- The easiest to use
- The best output

Algorithm - selection

- The most accurate
- The best for your problem
- What's available
- What you are familiar with

Generating the Alignment



Natural selection of software - driven by ease of use and availability (portability) - determines which programs are used most frequently.

MSA Programs

(a sampling)

Allall

DiAlign

Dali

ComAlign

IterAlign

MSA

Musca

T-Coffee

Pileup(GCG)

PRRP

Blast

Dalign

Clustalw

GA

MAVID

MultAlign

Museqal

ToPLign

POA

SAM

Blocks

DCA

ClustalX

HMMER

MAFFT

MultAlin

Oma

TreeAlign

Praline

SAGA

MSA Programs

(focus)

- MSA
(close-to-) optimal Alignments using the Carrillo-Lipman bound
- ClustalW/ClustalX
the most widely used program for multiple alignment
- T-Coffee
allows the combination of a collection of multiple/pairwise, global or local alignments into a single model
- DiAlign
constructs pairwise and multiple alignments by comparing whole *segments* of the sequences. No gap penalty is used
- POA & MAFFT
POA: partial order alignment, based on a graph representation of an MSA
MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform

Name	Algorithm	URL
MSA	EXACT	http://www.ibc.wustl.edu/ibc/msa.html
DCA (requires MSA)	Exact	http://bibiserv.techfak.uni-bielefeld.de/dca
OMA	Iterative DCA	http://bibiserv.techfak.uni-bielefeld.de/oma
ClustalW, ClustalX	PROGRESSIVE	ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW or clustalX
MultAlign	Progressive	http://www.toulouse.inra.fr/multalin.html
Dialign	CONSISTENCY BASED	http://www.gsf.de/biodv/dalign.html
ComAlign	Consistency-based	http://www.daimi.au.dk/~ocaprani
T-Coffee	CONSISTENCY BASED/PROGRESSIVE	http://igs-server.cnrs-mrs.fr/~cnotred
Praline Iterative/progressive	Iterative/progressive	jhering@nimr.mrc.ac.uk
IterAlign Iterative	Iterative	http://giotto.Stanford.edu/~luciano/iteralign.html
Prrp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/
SAM	Iterative/Stochastic/HMM	rph@cse.ucsc.edu
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/
SAGA	Iterative/Stochastic/GA	http://igs-server.cnrs-mrs.fr/~cnotred
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca

Multiple Sequence Alignment Methods

- Local Alignment-----Global Alignment
- Exact (MSA, DCA)
good for few, short, closely related sequences
- Progressive alignment (ClustalW)
fast, sensitive
- Consistency based method (DiAlign)
better for sequences with large insertions
- Iterative method (HMMER, SAM, HMMs)
slow, sometimes inaccurate ...good for profiles
- Combination methods (T-coffee)
very good but can be slow

Aligning two sequences

(Needleman and Wunsch)

	F	A	S	T
F	.			
A		.		
T				.

10 sequences (100AA)
3 million years and 10 billion gigabytes

Match=1

Mismatch= -1

Gap =-1

FAST

FA-T

	F	A	S	T	
	0	1	-2	-3	-4
F	-1	1	0		
A	-2	0	2		
T	-3				

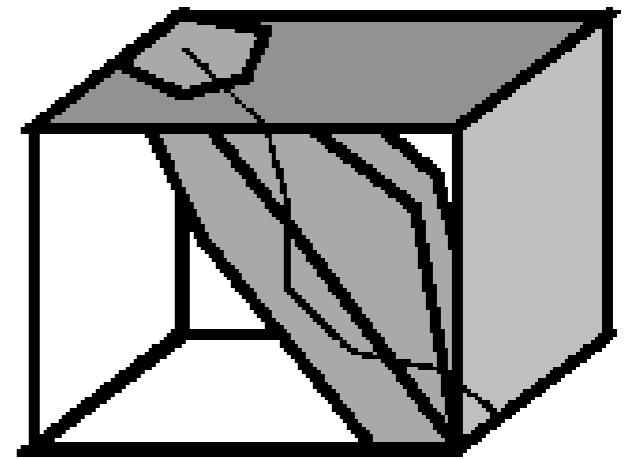
	F	A	S	T	
	0	-1	-2	-3	-4
F	-1	1	0		
A	-2	0	2	1	0
T	-3			1	2

	F	A	S	T
	0	1		
F				
A			2	1
T				2

MSA

An EXACT Alignment

1. Determine the optimal **pairwise alignments**.
2. Perform a **fast multiple sequence alignment** (progressive) and extract the pairwise alignments from this multiple sequence alignment.
3. For each pair of sequences, use the optimal and extracted pairwise alignments to define the **restricted alignment space** defined by the difference in the two alignment scores for this pair of sequences.
4. Project the restricted pairwise alignment spaces into the multidimensional alignment space to define the **restricted hyper-volume of the multidimensional space** to determine the best multiple sequence alignment. The greater the overall sequence similarity, the smaller the restricted alignment space is.
5. Use dynamic programming to compute the **value for all of the cells** within the restricted alignment space.
6. **Backtrack through the restricted alignment space** to recover the best alignment. The result is a minimum distance alignment.



MSA - 4GB memory

limited by length and diversity of sequence

- 20 phospholipase (130 AA)
- 14 (highly diverse) cytochrome C (110)
- 6 (moderately diverse) aspartyl proteases (350)
- 8 (moderately diverse) lipid-binding proteins(480)

ClustalW

A Progressive Alignment

1. **Pairwise Distances** - Perform Needleman-Wunsch (global) alignment on all sequence pairs to find the distance between all pairs of sequences.
2. **Cluster the Pairwise Distances** - Perform a simple clustering to determine which pairs of sequences are closer than others. Using pairwise alignments iteratively one can create phylogenetic relationships, which then allows for the creation of either a UPGMA-constructed guide tree or a Neighbor-Joining guide tree (both rooted trees). These joining trees are based on alignment scores and non-biological rules for creating trees; thus, they should be used cautiously as an evolutionary tree. This step represents a major difference among the various implementations of the PPA and is the part of the algorithm where some of the greatest improvements have occurred.
3. **Align the Sequences Guided by Clustering** - Align the closest sequences in the joining tree together, followed by adding more sequences to the initial alignment. For example, when using an UPGMA guide tree or Neighbor-Joining guide tree, one would align a pair of sequences by starting at the bottom of a branch and successively adding more sequences to the nascent alignment (the nascent alignment defines the range of possibilities for the ancestral sequence).

ClustalW issues

- Choice of input sequences
- Order of sequences in (tree)
- Parameters
weighting, substitution matrix, gap penalties
- Progressive (*once a gap always a gap*)
- Known to miss some conserved residues

T-Coffee

allows the combination of a collection of multiple/pairwise, global or local alignments into a single model

- Pairwise global alignment
- Pairwise local alignment
- Combined above two into a library
- Builds MSA with highest consistency with the library of alignments (progressive assembly)

DiAlign

constructs pairwise and multiple alignments by comparing whole segments of the sequences.

- Alignment of whole segments and not individual amino acids (bases)
- Pair wise comparison > segment pairs (diagonals),
represent local alignments
- Diagonals weighted for likelihood
- Alignment built from consistent diagonals
- No gap penalties
- Independent of sequence order

Meaningfulness

- Is the alignment *correct* ?
- Can I make it *better* ?
- Which programs are *best* ?
- How do you *know* if its correct ?

Is the Alignment *Correct* ?

- What do mean by correct ?
- Mathematically rigorous
- Biologically meaningful
- Operationally useful

Can you make it *better* ?

- Only if you know what you doing !
- Define better ?
- What's the goal ?
- What's the biology ?

Which programs are *best* ?

- No simple answer
- Depends on the particular problem
- Recent objective studies help answer this problem
- Some tools to help compare alignments

How do you *know* it is correct ?

- Methods to evaluate the alignment
- Methods to evaluate the program/algorithm
- Structural information
- Biology

Systematic Comparison of MSA programs

■ BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs

Thompson JD, Plewniak F, Poch O. Bioinformatics. 1999 Jan;15(1):87-8.

■ A comprehensive comparison of multiple sequence alignment programs

JD Thompson, F Plewniak, and O Poch Nucleic Acids Res. 1999 27: 2682-2690.

■ Quality assessment of multiple alignment programs

FEBS Letters Volume 529, Issue 1 , T. Lassmann and E Sonnhammer 2 October 2002, Pages 126-130

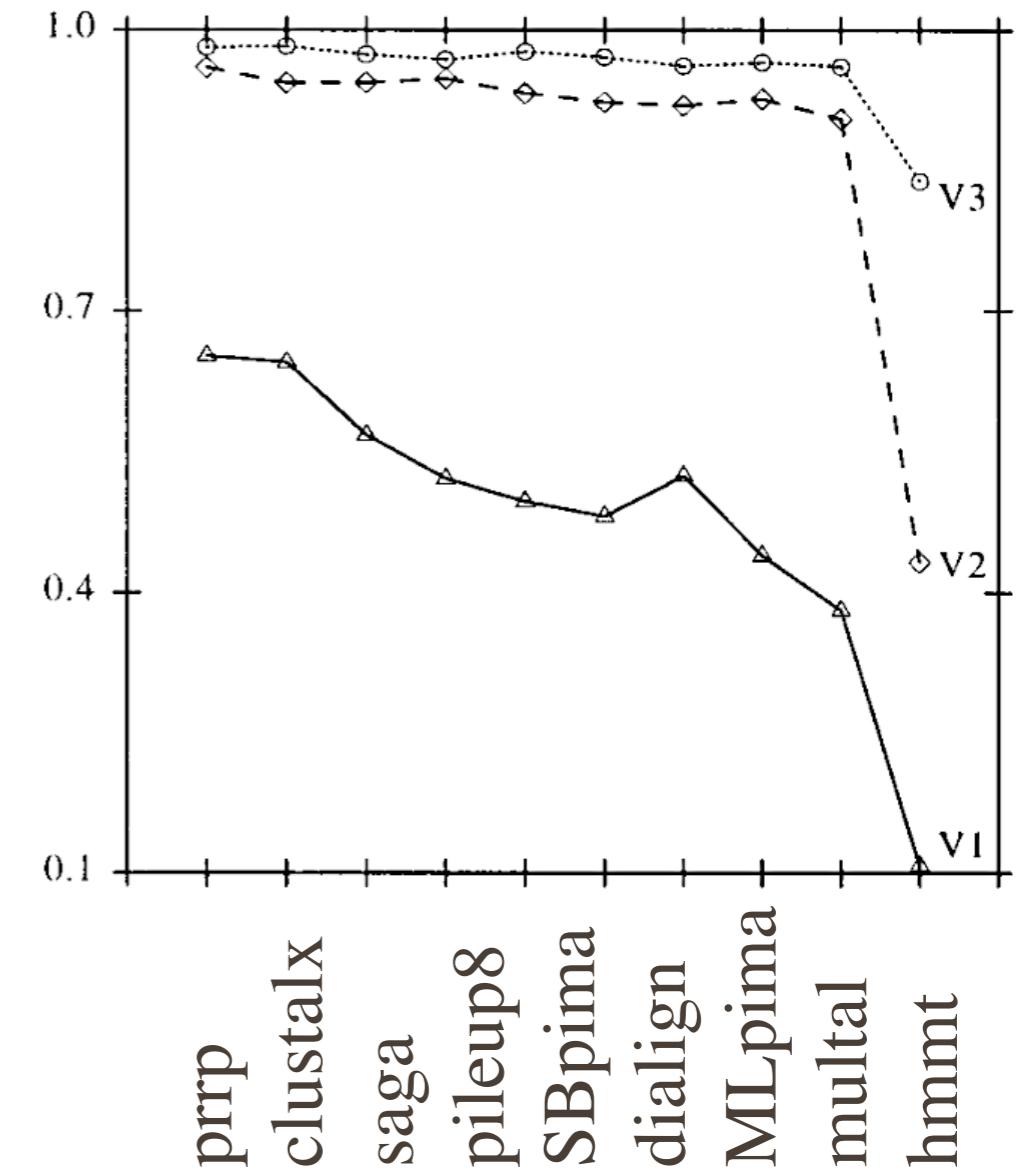
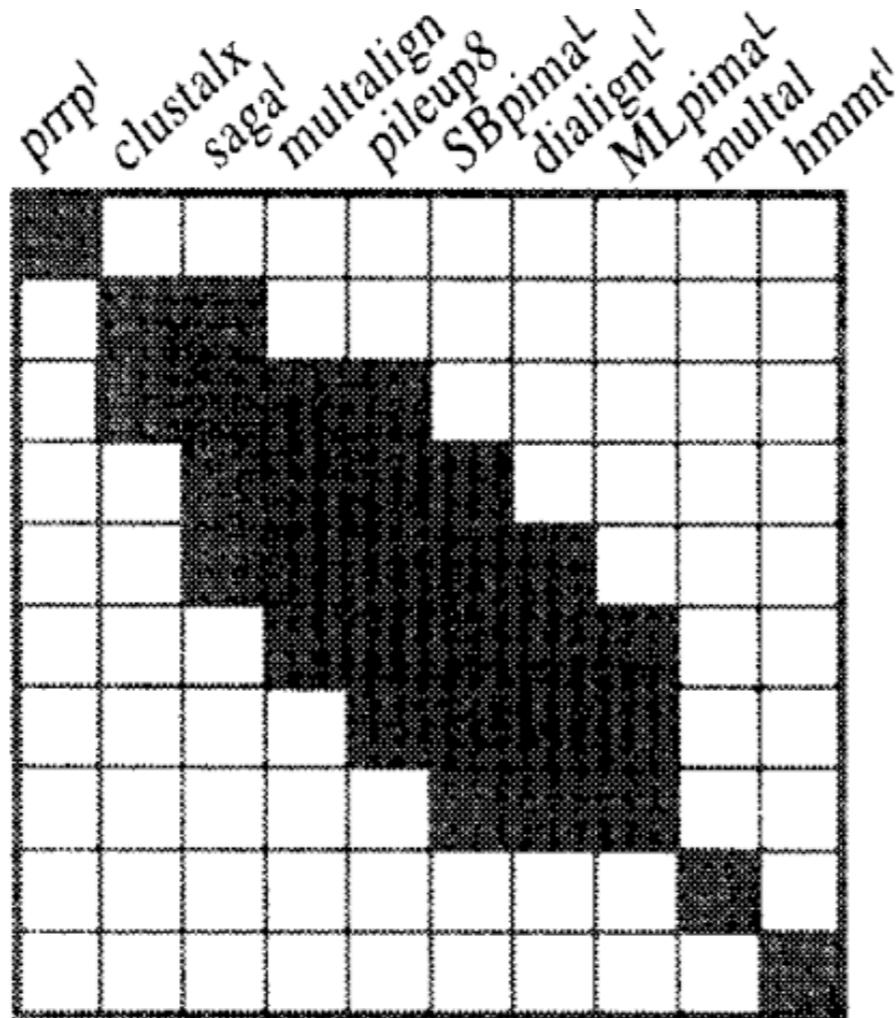
BALiBase -

142 reference sequences

http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE/prog_scores.html

	Residues	<100	200<300	>500
Reference 1	<25% identity	7	8	8
	20-40% identity	10	9	10
	>35% identity	10	10	8
Reference 2	homogenous + outlier	9	8	7
Reference 3	2 distantly related sets	5	3	5
		extensions (<400)	inserts (<100)	
Reference 4/5		12	12	

Program	Rank Sum
prrp	234.0
clustalx	316.5
saga	371.5
pileup8	398.0
SBpima	416.0
dialign	448.5
MLpima	465.5
multal	477.0
hmmt	544.0
	785.0



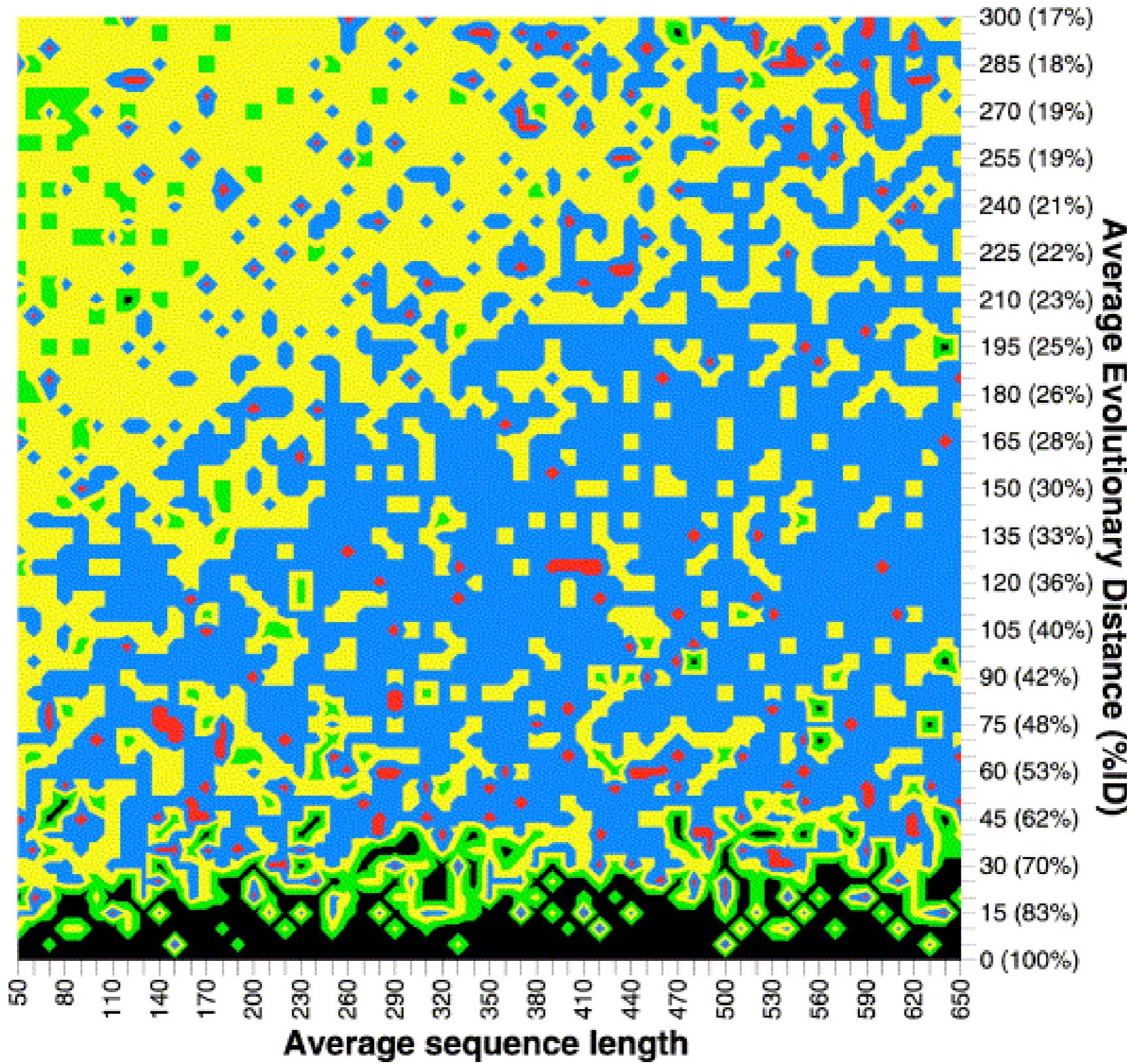


Fig. 1. Color coded matrix showing which method performed best for each pair-combination of conditions: average sequence length (x-axis) and average evolutionary distance (y-axis). The methods are Poa (green), Dialign (yellow), T-Coffee (blue) and ClustalW (red).

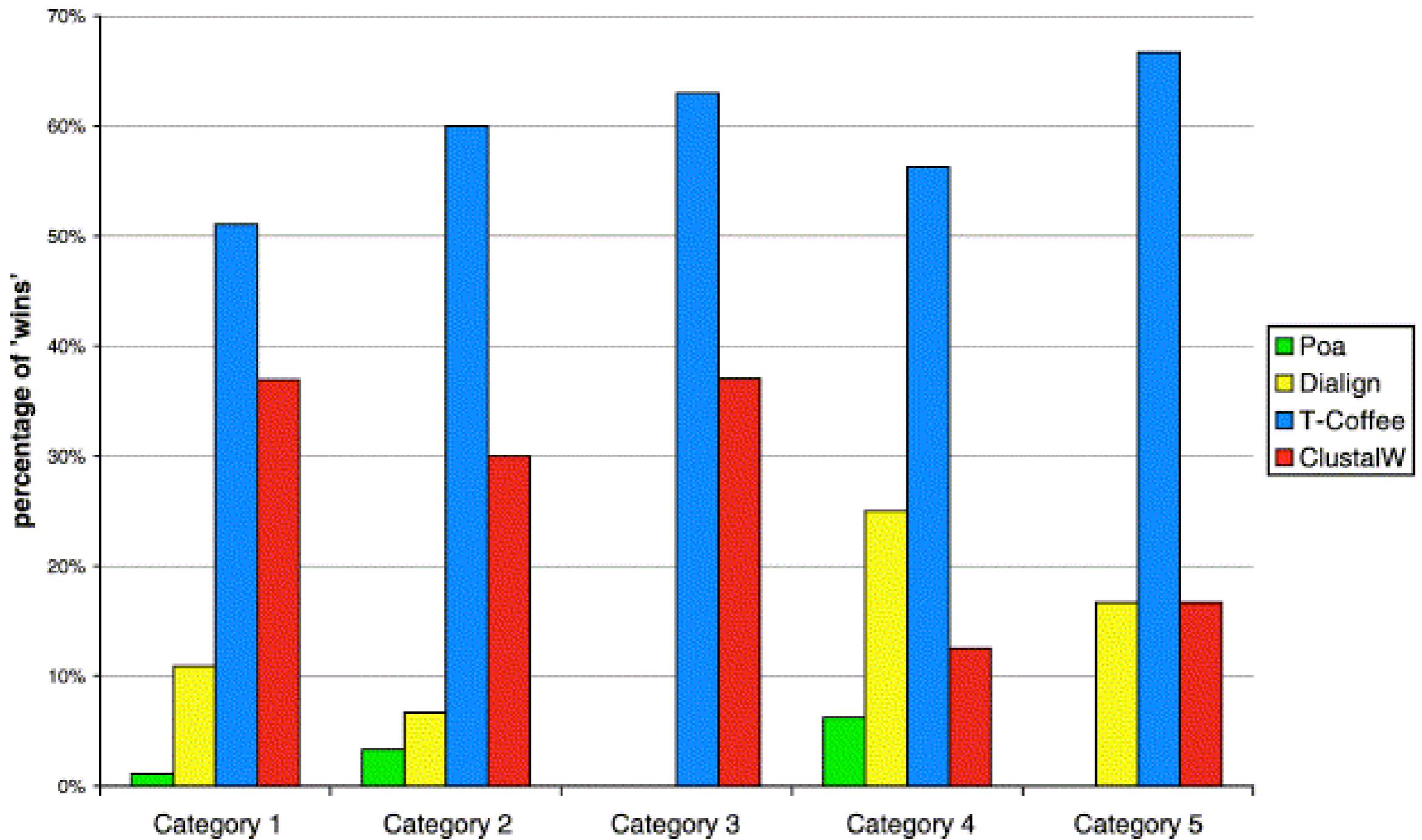
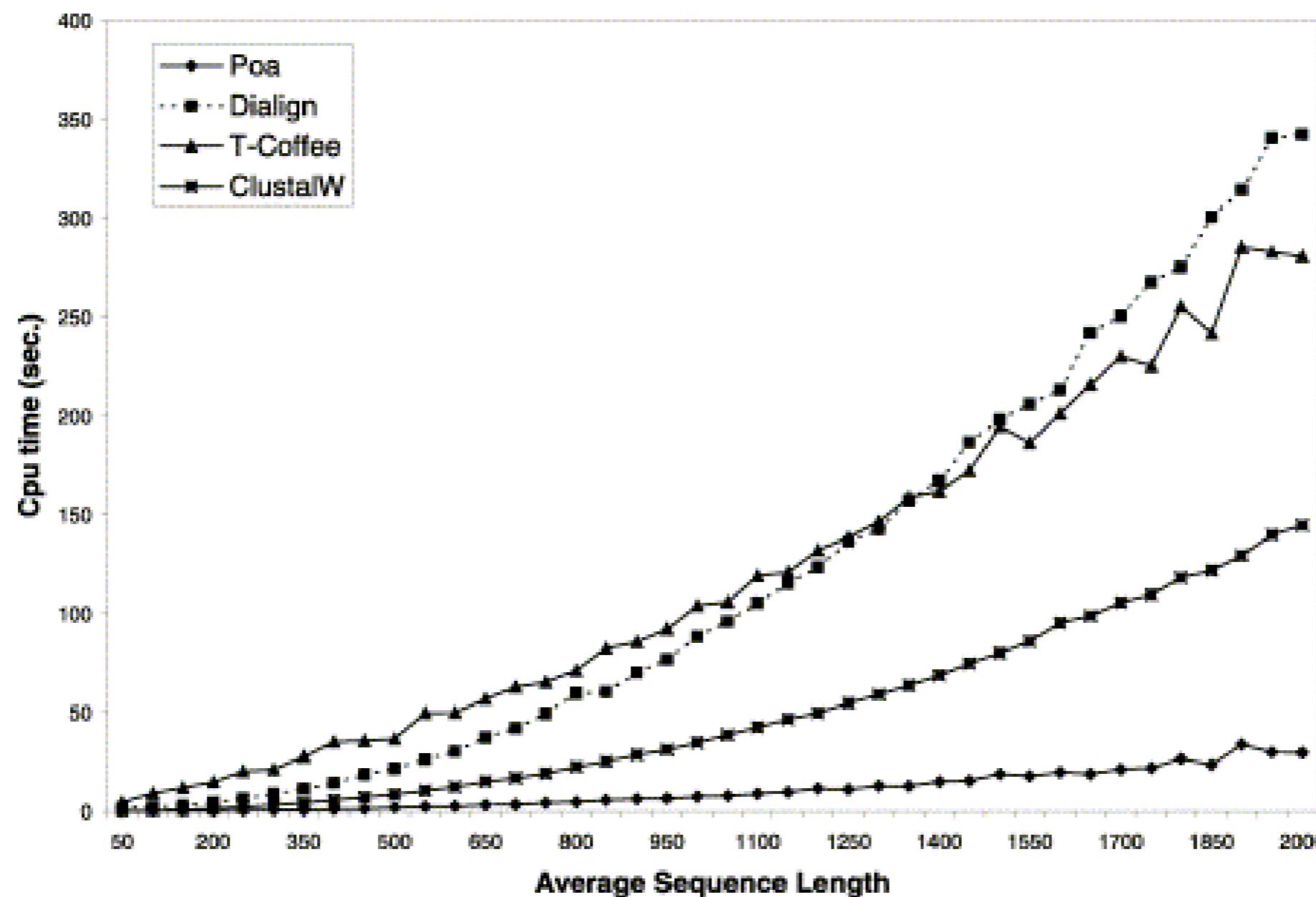


Fig. 2. Results of BALiBASE testing, showing the fraction that each program had the best accuracy (SPS) in each of the five BALiBASE categories.



CPU time consumed by each program to align sets of increasingly long sequences.

The Problem

I53 protein (220AA in length)

- ClustalW - six minutes
- T-Coffee - two days
- MSA - impractical

Recommendations

- **MSA** - for few, short sequences
- **ClustalW** - more versatile, most widely used and only program that *can use multiprocessors*
- **DiAlign** may do better for some
- **T-Coffee** sometimes better than ClustalW, but more computationally expensive
- **POA**, and **MAFFT** new programs which promise speed.

dha2_yeast_	-----VSEKSQHDDD KAVVDI S ERG RLLN I LADL I E RD RDI LAAIE HLD NGK P F D E A Y
dhac_mouse_	V DKAVKAAR QAF QI GSP WRTMDA SERG C L LNKL ADL M E RD RLL LATME AL N GGK V F A N A Y
dha5_yeast_	V DKAVKAARA AAF -- DNVWSK TS SE QRG IYLSN LLKL IE E E QDT LAALE TLDAGK PYHSNA
dhal_ecoli_	I DRAM SAAR GVFE - RG DW S LSSPA KRKA VL NKLA DL ME AHAEE L ALLE TLDG KPI RHSL
dha2_yeast_	LLDLASVLKE LRYTAGWADKLHGTLRFAITIPTFQD LRF LRYTRHEPVGVCGEIIIPWNIP
dhac_mouse_	LSDLGGCIKALKY CAGWADKI HG----QTIPS DGD I -FT-YTRREPIGVCGQIIIPWNFP
dha5_yeast_	KGD LAQILQLTRYFAGSADKF DKG----ATIPLTFNK-FA-YTLKVPFGVVAQIVPWNYP
dhal_ecoli_	RDDIPGAARAI RWYAE AIDK VYG----EVATTSSHELA-MIVREPVGVI AAI VPWNFP
dha2_yeast_	LLMYIW KIGPALAAGNTVVLKPEELTPLTALT VATL I KEAGFPPGVVN VVSGY GPTAGAA
dhac_mouse_	MLMFIW KIGPALSCGNTVVVKPAEQTPLTALHLAS LIKEAGFPPGVVNIVPGY GPTAGAA
dha5_yeast_	LAMACWKLQGALAAGNTVIIKPAENT SLSLLYFATLIKKAGFPPGVVNIVPGY GSLVGQA
dhal_ecoli_	LLLTCWKLGPAL AAGNSVILKPSEKSP LSAIRLAGLAKEAGL PDGVINVV TGF GHEAGQA
dha2_yeast_	CLSHKDNDKLAFTGSTLVGKVVMKAAA AKSNL KKVTLELGGKSP MIVFIDA-DLDW AVENA
dhac_mouse_	ISSHMDVDKVAFTGSTQVGK LIKEAAGKSNL KRV TLELGGKSP CIVFADA-DLDIAVEFA
dha5_yeast_	LASHMDIDKISFTGSTKVGGFVLEASGQSNLKDVTLECGGKSP ALVFEDA-DLDKAIDWI
dhal_ecoli_	LSRHNDIDAI AFTGST RTGK QLKDAGDSNMK RVWLEAGGKSANIVFADC PDLQQAASAT
dha2_yeast_	HFGVFE NQGQCCIAQS RITV HESIYDEIVERD LEKAKKQ--V LG--NP FESDTRYGPQIL
dhac_mouse_	HHGVFYHQGQCCVAASRI FVEESVYDEFV KRSVERAKKY--V LG--NP LTPG INQGPQID
dha5_yeast_	AAGIFYN SGQNCTANSRVYVQSSIYDKFVEKF KETA KKEWDVAGKF DPFDEKCIVGPVIS
dhal_ecoli_	AAGIFYNQGQVCIAGTRLL EESIADEF LALLKQQAQNW--Q PG--HPLDPATTMGT LID
dha2_yeast_	KIEFD S IPR LINS AKAEG--AKVLCGGGR DDSCV GYYIQPTV FADVTDEM RI AKEE IFGP
dhac_mouse_	KEQHD KILD LIESG KKEG--AKLECGG GRWGN-KGFFVQPTVFSN VTDEM RI AKEE IFGP
dha5_yeast_	STQYD RI KSYIER G KREEK LDMF QTSEFPIGGAKGYFIPPTI FTDV P QTSKLLQDEIFGP
dhal_ecoli_	CAHAD SVHSFIREGE SKG----Q LLLDGRNAG-LAAAIGPTI FVDV DPNA SLSREE IFGP
dha2_yeast_	VITISRFKSVDEAIK RVDNTK YGLAAYVFTK--DKA IRI S AALKAGTVW VNCVHV ASYQI
dhac_mouse_	VQQIMKFKSVDDV I K RANNTT YGLAAGLFTK DLDKA ITVSSALQAGV VV VNCYIMLSAQC
dha5_yeast_	VVVVSKFTNYDDALKL A NDTCYGLASAVFTK DVKKAHMFARDIKAGTVW WINSSN DEDVTV
dhal_ecoli_	VLVVTRFTSEEQALQLANDSQYGLGAAVWTRDLSRAHRMSRRLKAGSVEVNNYNDGDM TV
dha2_yeast_	PFGGNKN SGM GRELGEYGLE-----
dhac_mouse_	PFGGFKMSGN GRELGEH GLYEYTELKT VAMK I S QKNS
dha5_yeast_	PFGGFKMSGI GRELGQSGVDTYLQTKAVHINL SLDN-
dhal_ecoli_	PFGGYKQSGNGRD KSLH ALEKFT ELKT I WI-----

dha2_yeast_	VSEKSQHDDD KAVVDI SERGRLLNILADLIERDRDILAAIEHLDNGKPFDEAY
dhac_mouse_	VDKAVKAAR QAFQIGSPWRTMDA SERGCLLNKLADLMERDRLLLATMEALNGGKVFA
dha5_yeast_	NAYVDKAVKAARA AADF N--VWSKTSSEQRG IYLSNLLK LIEEEQDTLAALETLDAGKPYHSNA
dhal_ecoli_	IDRAMSAARGVFERG-DWSLSSPAKR KAVLNK LADLMEAHAEE LA LETLDGTGKPIRHSL
dha2_yeast_	LLDLASVLKE LRYT AGWADKLHGTLRFAITIPTFQDLRFLRYTRHEPVGVCGEIIPWNIP
dhac_mouse_	LSDLGGCI KALKY CAGWADKIHG-----QTIP--SDGDIFTYTRREPIGVCGQIIIPWNFP
dha5_yeast_	KGD LAQILQLTRYFAGSADKF DKG-----ATIP--LTFNKFAYTLKVPFGVVAQIVPWNYP
dhal_ecoli_	RDDIPGAARAIRWYAEAI DKVYG -----EVAT--TSSHELAMIVREPVGVIAAIVPWNFP
dha2_yeast_	LLMYIW KIGP ALAAGNTVVLKPEEL TPLT ALTVA T LIKEAGFPPGVNVVSGYGP
dhac_mouse_	AAMLMFI WKIGP ALSCGNTVVVKPAEQ TPLT ALHLAS LIKEAGF PPGVNVIVPGYGP
dha5_yeast_	AGAA LAMACWKLQG ALAAGNTV IIKPAENT SLSLLYFATLIK KAGF PPGVNVIVPGYGS
dhal_ecoli_	LVGQALLLTCWKL G PALAAGNSVILK PSEK SP LSAIRL AGLAKEAGLPDGVLNVVTGF
dha2_yeast_	HEAGQACL SHKDNDK LAFTGSTLVGKVVM KA AKSNLKKVTLELGGKSPMIVFIDA-DLDWAVENA
dhac_mouse_	ISSHMDVD KV AFTGSTQVGK L IKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEFA
dha5_yeast_	LASHMDID KISFTG STKVGGFVLEAS Q SNLKDVTLECGGKSPALVFEDA-DLDKAI DWI
dhal_ecoli_	LSRHND DIDAI AFTGSTRT GK QLLK DAGD SNMKRVW LEAGGKSAN IVFADC PD LQQAASAT
dha2_yeast_	HFGVFE NQGQC CIAQS RITV HESIYDEI VERD LEKAK KQ --V LG --NP FESD TRYGPQIL
dhac_mouse_	HHGVFYHQ QCCV AAS RIFV EESVYDEFV KRS VERAK KY --V LG --NP LTPG INQGPQID
dha5_yeast_	AAG IFYNSG QNCTANS RVYV QSS IYDK VEKF KETAK KEWD VAGK FDP DEKC IVGPVIS
dhal_ecoli_	AAG IFYNQG QV CIAG TRL LEES IADEF FLALL K QQAQNW --Q PG --H PLDP ATT MGT LID
dha2_yeast_	KIEFDSIPRLINS AKAEG --AKVLCGGGR DDSCV GYYIQOPTVFADVTDEM RI AKEEIFGP
dhac_mouse_	KEQHD KILD LIES SGK KEG--AK LECGG GR-WGNKGFFV QOPTV FSNVTDEM RI AKEEIFGP
dha5_yeast_	STQYDRIKSYIERG KREE KLDM QTSEF PIGG AKGY FIP PTI F DV PQT SKL Q DEI FGP
dhal_ecoli_	CAHADSVHS FI REG ESKG --Q LLL DGRN--AGLAAA IGPT IFV DV D PN ASLS REE IFGP
dha2_yeast_	VITISRF KSVDEAI KRVDNT KYGLAAYVFTK --DKAIR ISAALK AGTVWVNCVHVASYQI
dhac_mouse_	VQQIMKFK SVDDV I KRAN NTTY GLAAGL FTKDLD KAITV SSALQAGVWWNCYIMLSAQ C
dha5_yeast_	VVVVSKFTNY DDALK LAND TCYGLA SAVFT KDVKK AHM FARDIK AGTVWIN SSN DED VT V
dhal_ecoli_	VLVVTRFT SEEQ AL QL LAND SQYGL AAV WTRD LSRAHR MSRR L KAGS VF VNN YND GDM TV
dha2_yeast_	PFGGNKN SGM GRE LG EY GLE -----
dhac_mouse_	PFGGFKMS GN GRE LG E HGLY EY TEL K VAM K ISQK NS
dha5_yeast_	PFGGFKMS GI GRE LG QSG V D TYL QTKAVH IN LSLD N-
dhal_ecoli_	PFGGY KQSG NG RD KL HALE K TEL K TI WI-----

dha2_yeastVSE KSQHDDDKAV VDISERGRLL NILADLIERD RDILAAIEHL VDKAVKAARQ AFQIGSPWRT MDASERGCLL NKLADLMERD RLLLATMEAL VDKAVKAARA AF DN. WSK TSSEQRGIYL SNLLKLIEEE QDTLAALETI IDRAMSAARG VF ERG. DWSL SSPAKRKAVL NKLADLMEA AELALLETL		dha2_yeastVSE KSQHDDDKAV VDISERGRLL NILADLIERD RDILAAIEHL VDKAVKAARQ AFQIGSPWRT MDASERGCLL NKLADLMERD RLLLATMEAL VDKAVKAARA AF ..DN. WSK TSSEQRGIYL SNLLKLIEEE QDTLAALETI IDRAMSAARG VFE RGDW. SSPAKRKAVL NKLADLMEA AELALLETL
dha2_yeast	DNGKPFD EAY LLD LASVLKE LRYTAGWADK LHGT LRF AIT IPTFQDLRFL NGGKV FANAY LSDLG GCIKA LKYCA GWA DK IHG.....QT IP ..SDGDI DAGKPYHSNA KGDLA QILQL TRYFAGSADK FDKG....AT IP ..LT FNKF DTGKPIRHSL RDDIPGAARA IRWYAEAI DK VYG.....EV AT ..TSSHE L		dha2_yeast	DNGKPFD EAY LLD LASVLKE LRYTAGWADK LHGT LRF AIT IPTFQDLRFL NGGKV FANAY LSDLG GCIKA LKYCA GWA DK IHG.....QT IP PSDGDI.FT DAGKPYHSNA KGDLA QILQL TRYFAGSADK FDKG....AT IP LT FNKF.FA DTGKPIRHSL RDDIPGAARA IRWYAEAI DK VYG.....EV AT TSSHE LA
dha2_yeast	RYTRHEPVGV CGEIIPWNIP LLMYIW KIGP ALAA AGNTVVL KPEELTPLTA TYTRREPIGV CGQIIPWNFP MLMFIW KIGP ALSCG NTVVV KPAEQTPLTA AYTLKVPFGV VAQIVPWNYP LAMACW KLQG ALAA AGNTVII KPAENTSLSL AMIVREPVG V IAAIIVPWNFP LLTCW KLGP ALAA AGNSVIL KPSEK SPLSA		dha2_yeast	RYTRHEPVGV CGEIIPWNIP LLMYIW KIGP ALAA AGNTVVL KPEELTPLTA .YTRREPIGV CGQIIPWNFP MLMFIW KIGP ALSCG NTVVV KPAEQTPLTA .YTLKVPFGV VAQIVPWNYP LAMACW KLQG ALAA AGNTVII KPAENTSLSL .MIVREPVG V IAAIIVPWNFP LLTCW KLGP ALAA AGNSVIL KPSEK SPLSA
dha2_yeast	LTVAT LIKEA GFPPGV VN VV SCY GPTAGAA CLSHKDNDKL AFTG STLVGK LHLAS LIKEA GFPPGV VN IV PGY GPTAGAA ISSHMDVDKV AFTG STQVGK LYFAT LIKKA GFPPGV VN IV PGY GSLVGQA LASHMD IDKI SFTG STKVGG IRLAG LAKEA GLPDGV LN VV TGF GHEAGQA LSRHND IDAI AFTG STRTGK		dha2_yeast	LTVAT LIKEA GFPPGV VN VV SCY GPTAGAA CLSHKDNDKL AFTG STLVGK LHLAS LIKEA GFPPGV VN IV PGY GPTAGAA ISSHMDVDKV AFTG STQVGK LYFAT LIKKA GFPPGV VN IV PGY GSLVGQA LASHMD IDKI SFTG STKVGG IRLAG LAKEA GLPDGV LN VV TGF GHEAGQA LSRHND IDAI AFTG STRTGK
dha2_yeast	VVMKAAAKSN LKKVT LELGG KSPMIV FIDA .DLD WAVENA HFGVFFN QGQ LIKEAAG KSN LKRVT LELGG KSPCIV FADA .DLD IAVEFA HHGV FYHQGQ FVLEAS GQSN LKDVT LECGG KSPALV FEDA .DLD KAIDWI AAGIFY NSGQ QLLKDA GDSN MKRW WLEAGG KSANIV FADC PDLQ QAAASAT AAGIFY NQGQ		dha2_yeast	VVMKAAAKSN LKKVT LELGG KSPMIV FIDA .DLD WAVENA HFGVFFN QGQ LIKEAAG KSN LKRVT LELGG KSPCIV FADA .DLD IAVEFA HHGV FYHQGQ FVLEAS GQSN LKDVT LECGG KSPALV FEDA .DLD KAIDWI AAGIFY NSGQ QLLKDA GDSN MKRW WLEAGG KSANIV FADC PDLQ QAAASAT AAGIFY NQGQ
dha2_yeast	CCIAQS RITV HESIY DEIVE RD LEKAKKQ. .V LG..NPFE SDTRY GPQIL CCVAAS RIFV EESVY DEFVK RS VERAKKY. .V LG..NPLT PGINQ GPQID NCTANS RVVV QSSIY DKFVE KFKET AKKEW DVAGK FDPFD EK CIVGPVIS VCIAGTR LLL EESIA DEFLA LLKQQ AQNW. .QPG..HPLD PATTM GTLID		dha2_yeast	CCIAQS RITV HESIY DEIVE RD LEKAKKQ. .V LG..NPFE SDTRY GPQIL CCVAAS RIFV EESVY DEFVK RS VERAKKY. .V LG..NPLT PGINQ GPQID NCTANS RVVV QSSIY DKFVE KFKET AKKEW DVAGK FDPFD EK CIVGPVIS VCIAGTR LLL EESIA DEFLA LLKQQ AQNW. .QPG..HPLD PATTM GTLID
dha2_yeast	KIEFD SIPRL INSAKA EG.. AKVL CGGGRD DSCV GYYIQP TVFAD VTDEM KEQHD KIDL IESG KKEG.. AKLE CGGGR WGNK GFFVQP TVFS SVTDEM STQYD RIKS IERG KREEKL DMFQT SEFPI GGAK GYFIPP TIFTD VPQTS CAHAD SVHSF IREGES KG.. QLL LDGRN. AGLAA AIGP TIF VDVDPNA		dha2_yeast	KIEFD SIPRL INSAKA EG.. AKVL CGGGRD DSCV GYYIQP TVFAD VTDEM KEQHD KIDL IESG KKEG.. AKLE CGGGR WGNK GFFVQP TVFS SVTDEM STQYD RIKS IERG KREEKL DMFQT SEFPI GGAK GYFIPP TIFTD VPQTS CAHAD SVHSF IREGES KG.. QLL LDGRN. AG.LAAA AIGP TIF VDVDPNA
dha2_yeast	RIAKEE IFGP VITIS RFKSV DEAI KRVDNT KYGLAA YVFT K..DKAIRIS RIAKEE IFGP VQQIM KFKSV DD VIKRANNT TYGLAA GLFT KD LDKAITVS KLLQD EIFGP VVV VSKFTNY DD ALKLANDT CYGLAS AVFT KDV KKAHMFA SLSREE IFGP VL VTRFTSE EQAL QLLANDS QYGLGA AWFT RD LSRAHRMS		dha2_yeast	RIAKEE IFGP VITIS RFKSV DEAI KRVDNT KYGLAA YVFT K..DKAIRIS RIAKEE IFGP VQQIM KFKSV DD VIKRANNT TYGLAA GLFT KD LDKAITVS KLLQD EIFGP VVV VSKFTNY DD ALKLANDT CYGLAS AVFT KDV KKAHMFA SLSREE IFGP VL VTRFTSE EQAL QLLANDS QYGLGA AWFT RD LSRAHRMS
dha2_yeast	AALKACT WV NCVHV ASYQI PFGGN KNSGM GRE LGEYGLE SALQAG VWWV NC YIMLSAQ C PFG GFKMSGN GRE LGEHGLY EY TELKT VAM RDIKACT WV NSSN DEDVTV PFG GFKMSGI GRE LGQSGVD T LQTKAVHI RR LKG SV FV NNYND GDMTV PFG GYKQSGN GRD KSLH ALE KFT ELKT IWI		dha2_yeast	AALKACT WV NCVHV ASYQI PFGGN KNSGM GRE LGEYGLE SALQAG VWWV NC YIMLSAQ C PFG GFKMSGN GRE LGEHGLY EY TELKT VAM RDIKACT WV NSSN DEDVTV PFG GFKMSGI GRE LGQSGVD T LQTKAVHI RR LKG SV FV NNYND GDMTV PFG GYKQSGN GRD KSLH ALE KFT ELKT IWI

dha2_yeast_	VSEKSQHDDD KAVVDISERGRLLNILADLIERDRDILAAIEHLDNGKPFDEAY
dhac_mouse_	VDKAVKAARQAFQIGSPWRTMDASERGCLLNKLADLMERDRLLLATMEALNGGKFANAY
dha5_yeast_	VDKAVKAARA AAF--DNVWSKTSSEQRGYIYLSNLLKLIEEEQDTLAALETLDAGKPYHSNA
dhal_ecoli_	IDRAMSAARGVFE-RGDWSLSSPAKRKA VLNKLADLMEAHAEEELALLETLDTGKPIRHSL
dha2_yeast_	LLDLASVLKELRYTAGWADKLHG T LRFAITIPTFQDLRFLRYTRHEPVGVCGEIIIPWNIP
dhac_mouse_	LSDLGGCIKALKYCAGWADKIHG-----QTIPSDGDI-FT-YTRREPIGVCGQIIPWNFP
dha5_yeast_	KGDLAQILQLTRYFAGSADKFDKG-----ATIPLTFNK-FA-YTLKVPFGVVQA QIVPWNYP
dhal_ecoli_	RDDIPGAARAIRWYAEAIDKVYG-----EVATTSSHELA-MIVREPVGVIAAIIVPWNFP
dha2_yeast_	LLMYIW KIGPALAAGNTVVLKPEELTPLTALT VATLIKEAGFPPGVNVVSGYGPTAGAA
dhac_mouse_	MLMF IWKIGPALSCGNTVVVKPAEQTPLTALHLAS LIKEAGFPPGVVNIVPGYGPTAGAA
dha5_yeast_	LAMACWKLQGALAAGNTVI IKPAENTSLSLLYFATLIK AGFPPGVVNIVPGYGSLVGQA
dhal_ecoli_	LLLTCWKLGPALAGNSVILK PSEKSP LSAIRLAGLAKEAGLPDGVLNVVTGFGHEAGQA
dha2_yeast_	CLSHKDNDKLAFTGSTLVGKVMKAA AKSNLKKVTLELGGKSPMIVFIDA -DLDWAVENA
dhac_mouse_	ISSHMDVDKVAFTGSTQVGK L IKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEFA
dha5_yeast_	LASHMDIDKISFTGSTKVGGFVLEASGQS NLKDVTLECGGKSPALVFEDA -DLDKAIDWI
dhal_ecoli_	LSRHNDIDAI AFTGSTRTGKQLLK DAGDSNMKRVWLEAGGKSANIVFADC PDLQQ AASAT
dha2_yeast_	HFGVFFNQGQC CIAQSRITVHESIYDEIVERDLEKAKKQ --VLG--NPFE SDTRYGPQIL
dhac_mouse_	HHGVFYHQGQC CCVAASRIFVEESVYDEFVKRSVERAKKY --VLG--NPLTPG INQGPQID
dha5_yeast_	AAGIFYNSGQNCTANSRVYVQSS IYDKFVEKFETAKKEWDVAGKFDPFDEKCIVGPVIS
dhal_ecoli_	AAGIFYNQGQC VIAGTRLL LEESIADEF FLALLKQQAQNW --QPG--HPLDPATT MGT LID
dha2_yeast_	KIEFD SIPRLINSAKAEG --AKVL CGGGRDDSC VGYYIQPTVFADVTDEMRIA EEIFGP
dhac_mouse_	KEQHD KILDLIESGKKEG --AKLE CGGGRWGN -KGFFVQPTVFS SVNTDEMRIA EEIFGP
dha5_yeast_	STQYDRIKSYIERGKREE KLDMFQTSEFPI GGAKGYF IPPTIFTDV PQTSKLLQDEEIFGP
dhal_ecoli_	CAHADSVHS FIREGESKG --QLLDGRNAG-LAAA IGPTIFVDVDPNAS LSREEEIFGP
dha2_yeast_	VITISRF KSVDEAIKRV DNTKYGLAAYVFTK--DKAIRIS AALKAGTVWVNCVHVASYQI
dhac_mouse_	VQQIMKF KSVDV IKRANN NTTYGLAAGLFTKDLDK AITVSSALQAGVV WVNCYIMLSAQC
dha5_yeast_	VVVVSKFTNYDDALK LANDTCYGLASAVFTKDVK KAHM FARDIKAGTVWINSSNDEDVT V
dhal_ecoli_	VLVVTRFTSEE QALQLANDSQYGLGAAVWT RDLSRAHRMSRR LKAGSVFVN NYNDGDMTV
dha2_yeast_	PFGGNKNSGM GRELGEYGLE --
dhac_mouse_	PFGGFKMSG NGRELGEHGLYEYTEL KT VAMKISQKNS
dha5_yeast_	PFGGFKMSG GIGRELGQSGVDTYLQTKAVHINL SLDN-
dhal_ecoli_	PFGGYKQSGNG RDKSLHALEKFTEL KT WI --

dha2_yeast	-----VSEKSQHDDD KAVVDI SERGRLLNILADLIERDRDILAAIEHLDNGKPFDEAY
dhac_mouse	VDKAVKAAR QAFQIGSPWRTMDA SERGCLLNKLADLMERDRLLLATMEALNGGKVFA
dha5_yeast	NAYVDKAVKAARA AAF--DNVWSKTSSEQRG IYLSNLLKLIIEEEQDTLAAL
dhal_ecoli	DAGKPYHSNAIDRAMSAARGVFE-RGDWSLSSPAKRKA VLNKLADLMEAHAEE
dha2_yeast	VL LDLASV LKE RLYTAGWADKLHGTLRFAITIPTFQDLRF LYTRHEPVGVCGEIIIPWNIP
dhac_mouse	LSDLGG CIKALKY CAGWADKIHG-----QTIPSDGDI-FT-YTRREPIGVCGQIIIPWNFP
dha5_yeast	KGD LAQILQLTRYFAGSADKF DKG-----ATIPLTFNK-FA-YTLKVPFGVVAQIVPWNYP
dhal_ecoli	RDDIPGAARAIRWYAEAIDK VYG-----EVATTSSHELA-MIVREPVGVI AAIVPWNFP
dha2_yeast	LLMYIW KIGP ALAAGNTVVLKPEELTPLTALT TVATL IKEAGFPPGVNVVSGYGPTAGAA
dhac_mouse	MLMF IWKIGP ALSCGNTVVVKPAEQTPLTAHLAS LIKEA GFPPGVVNIVPGYGPTAGAA
dha5_yeast	LAMACWKL QGAL AAGNTVIIKPAENT SLSLLY FATLIK KAGF PPGVVNIVPGYGSLVGQA
dhal_ecoli	LLLTCWKL GPA LAAGNSVILKP SEK SEPLSAIRLAG LAKEA GLPDGVLNVVTGFGHEAGQA
dha2_yeast	CLSHKDND KLAFTG STLVGKVVMKAAAKSNLKKVTLELGGKSPMIVFIDA-DLDWAVENA
dhac_mouse	ISSHMDVD KVAFTG STQVGKLIKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEFA
dha5_yeast	LASHMDID KISFTG STKVGGFVLEASGQS NLKDVT LECGGKSPALVFEDA-DLDKAIDWI
dhal_ecoli	LSRHND DIAIAFTG STRTG QOLLKDAGD SNMKRVWLEAGGKSANIVFADC PDLQQ AASAT
dha2_yeast	HFGVFFNQGQC CIAQS RITVHESIYDEIVERD LEKAKK Q--VLG--NPFE SDTRYGPQ IL
dhac_mouse	HHGVFYHQGQC CCVAAS RIFVEESVYDEFVKRSVERAK KY--VLG--NPLTPG INQGPQID
dha5_yeast	AAGIFYN SGQNCT ANSRVYVQSS SIYDKF VEKF KETAKKEWDVAGKF DPFDEKCIVGPVIS
dhal_ecoli	AAGIFYN QGQVCI AGTR LLLEESIA DEF FLALLKQQAQNW--QPG--HPLDPATTMGT LID
dha2_yeast	KIEFD SI PRLINS AKAEG--AKVLC GGGRDD SCVGYYIQPTV FADVTDEM RIAKEE IFGP
dhac_mouse	KEQHD KILD LIES SGKKEG--AKLE CGGGRWGN-KGFFV QPTVFS NTDEM RIAKEE IFGP
dha5_yeast	STQYDRIK SYIERG KREE KLDMFQT SEFP IGGAKGYFIPPT IFTDV PQTS KL LQDE IFGP
dhal_ecoli	CAHADSV VHSFIRE GESKG----QLLDGRNAG-LAAA IGPTIFV DVD DPNAS LS REE IFGP
dha2_yeast	VITISRF KSVDEAI KRVDNT KYGLAAYVFTK--DKA IRIS AALKAGTVWVNCVHV ASYQI
dhac_mouse	VQQIM KFKSVDDV IKRANNT TYGLAAGLFTKDLD KAITVSSALQAGV VWVNCYIMLSAQC
dha5_yeast	VVVVSKFTNYDDALK LANDTCYGLASAVFTKDVK KAHM FARDIKAGTVW INSS NDEDVT V
dhal_ecoli	VLVVTRFT SEEQALQL AND SQYGLGAAVWT RDLSRAHRMS RRLKAGSVFVN NNYND GDM TV
dha2_yeast	PFGGNKN SGM GREL GEYGLE-----
dhac_mouse	PFGGFKMS NGN GREL GEHGLYEYTELK TAMK ISQKNS
dha5_yeast	PFGGFKMS GIG REL GQSGVDT YLQT KA VHIN NSLDN-
dhal_ecoli	PFGGYK QSGNGRDK SLHALE KFTELK TIWI-----

xx:dha2_yeast
 xx:dhac_mouse
 xx:dha5_yeast
 xx:dhal_ecoli

--VSEKSQHDDDK----AVVDISERGRLLNILADLIERDRDILAAIEHLDNGKPFDEAY
 VDKAVKAARQAFQIGSPWRTMDASERGCLLNKLADLMERDRLLLATMEALNGKVFANAY
 VDKAVKAARAADFNV--WSKTSSEQRGIVYLSNLLKLIEEEQDTLAALETLDAGKPYHSNA
 IDRAMSAAARGVFERG-DWSLSSPAKRKAVLNKLADLMEAHAEEELALLETLDTGKPIRHSL

xx:dha2_yeast
 xx:dhac_mouse
 xx:dha5_yeast
 xx:dhal_ecoli

LLDLASVLKEELRYTAGWADKL-HGTLRFAITIPTFQDLRFLRYTRHEPVGVCGEIIIPWN
 LSDLGCGIKALKYCAGWADKI-HG----QTIPSDGD--IFTYTRREPIGVCGQIIIPWN
 KGDLAQILQLTRYFAGSADKFDKG----ATIPLTFN--KFAYTLKVPFGVVAQIVPWN
 RDDIPGAARAIRWYAEAIDKV-YG---EVATTSSHE---LAMIVREPVGVIAAIVPWN

xx:dha2_yeast
 xx:dhac_mouse
 xx:dha5_yeast
 xx:dhal_ecoli

PLLMYIWKIGPALAAGNTVVLKPEELTPLTALTIVATL LIKEAGFPPGVVNVVSGYGPTAGA
 PMLMFIWKIGPALSCGNTVVVKPAEQTPLTALHLASLIKEAGFPPGVVNIVPGYGP
 PLAMACWKLQGALAAGNTVIIKPAENTSLSLYFATLIKAGFPPGVVNIVPGYGS
 PLLTCWKLGPAAGNSVILKPSEKSPLSAIRLAGLAKEAGLPDGVLNVVTGFGHEAGQ

xx:dha2_yeast
 xx:dhac_mouse
 xx:dha5_yeast
 xx:dhal_ecoli

ACLSHKDNDKLAFTGSTLVGVVMKAAAKSNLKKVTLELGGKSPMIVFIDA-DLDWAVEN
 AISSHMDVDKVAFTGSTQVGKLIKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEF
 ALASHMDIDKISFTGSTKVGGFVLEASGQSNLKDVTLECGGKSPALVFEDA-DLDKAIDW
 ALSRHNDIDAIIAFTGSTRTGKQOLLKDAGDSNMKRVWLEAGGKSANIVFADC
 PDLQQAAASA

xx:dha2_yeast
 xx:dhac_mouse
 xx:dha5_yeast
 xx:dhal_ecoli

AHFGVFEHQGCCIAQSRI TVHESIYDE IVERDLEKAKKQV---LGNPFESDTRYGPQI
 AHHGVFYHQGCCVAASRIFVEESVYDEFVKRSVERAKKYV---LGNPLTPGINQGPQI
 IAAGIFYNSGQNCTANSRVYVQSSIYDKFVEKFETAKKEWDVAGKFDPFDEKCIVGPVI
 TAAGIFYNQGQVCIAGTRLLEESIADEFALLKQQA-QNW---QPGHPLDPATTMGT
 LI

xx:dha2_yeast
 xx:dhac_mouse
 xx:dha5_yeast
 xx:dhal_ecoli

LKIEFDsipRLINSAKAEGAKVLCGGGRDD---SCVGYYIQPTVFADVTDEMRIA
 KEEIFDKEQHDKILDLIESGKKEGAKLECGGGRW---GNKGFFVQPTVFS
 NVTDEMRIA
 KEEIF SSTQYDRIKSYIERGKREE-KLDMFQTSEFP
 IGGAKGYFIPPTIIFTDV
 PQTSKLLQDEIF
 DCAHADSVHSFIREGESKG-QLLLDGRNA---GLAAAIGPTIF
 VDVDPNASLSREEIF

xx:dha2_yeast
 xx:dhac_mouse
 xx:dha5_yeast
 xx:dhal_ecoli

GPVITISRFKSVDEAIKRV
 DNTKYGLAAYVFTKD--KAIRISAALKAGTVWVNCVHV
 ASY GPVQQIMKFKSVDDVIKRANNTTYGLAAGLFTK
 DLLDKAITVSSALQAGV
 VVWC
 NYIMLSA GPV
 VVVVSKFTNYDDALKLANDTCYGLAS
 A
 VFTKDV
 KAHMFARDIKAGTVW
 INSSN
 DEDV GPV
 LVVTRFTSEEQALQL
 ANDSQYGLGAAVWT
 RDLSRAHRMS
 RRLKAGS
 VEVNNY
 NDGDM

xx:dha2_yeast
 xx:dhac_mouse
 xx:dha5_yeast
 xx:dhal_ecoli

QIPFGGNKN
 SGMGRELGEYGLE-----
 QC
 PFGGFKMS
 GNGRELGEHGLYEYTELKT
 VAMKIS
 SQKNS
 T
 VPFGGF
 KMSGI
 GRELQSG
 GVDTY
 LQTKAVH
 INLS
 LDN
 T
 VPFGGY
 KQSGN
 GRDK
 SLHA
 LEK
 FTELKT
 IW
 I

dha2_yeast_	VSEKSQHDDDKA VDKAVKAARQAFQIGSPWRTMDASERGCLLNKLADLMERDRLLATMEALNGGVFANAY	
dhac_mouse_		
dha5_yeast_	VDKAVKAARA DNVWSKTSSEQRGIVYLSNLLKLIEEQDTLAALETLDAGKPYHSNA	
dhal_ecoli	IDRAMSAARGVFE-RGDWSLSSPAKRKAVLNKLADLMEAHAEEELALLETLDTGKPIRHSL	
dha2_yeast_	LLDLASVLKELRYTAGWADKLHGTLRFAITIPTFQDLRFLRYTRHEPVGVCGEIIPWNIP	
dhac_mouse_	LSDLGGCIKALKYCAGWADKIHG-----QTIPSDGDI-FT-YTRREPIGVCGQIIPWNFP	
dha5_yeast_	KGDLAQILQLTRYFAGSADKF DKG-----ATIPLTFNK-FA-YTLKVPFGVVAQIVPWNYP	
dhal_ecoli	RDDIPGAARAIRWYAEAIDKVYG-----EVATTSSHELA-MIVREPVGVIAAIVPWNFP	
dha2_yeast_	LLMYIWKIGPALAAGNTVVLKPEELTPLTALT VATLIKEAGFPPGVNVVSGYGP TAGAA	
dhac_mouse_	MLMFIWKIGPALSCGNTVVVKPAEQTPLTALHLAS LIKEAGFPPGVNVIPGYGP TAGAA	
dha5_yeast_	LAMACWKLQGALAAGNTVII IKPAENTSLSLLYFATLIK KAGFPPGVNVIPGYGSLVGQA	
dhal_ecoli	LLLTCWKLGPAL AAGNSVILKPSEK SPLSAIRLAGLAKEAGLPDGV LNVVTGFGHEAGQA	
dha2_yeast_	CLSHKDNDKLAFTG STLVGKVMKAAKS NLKKVTLELGGKSP MIVFIDA-DLD WAVENA	
dhac_mouse_	ISSHMDVDKVA FTGSTQVGK LIKEAAGK SNLKRV TLELGGKSP CIVFADA-DLD DIAVEFA	
dha5_yeast_	LASHMDIDK ISFTG STKVGGF VLEASG QSNLK DVTLECG GGKSP ALVFEDA-DLD KAIDWI	
dhal_ecoli	LSRHNDIDAI AFTG STRTG KQLLKD AGDSNM KRVW LEAGG KSANIV FADCP DLQQ AASAT	
dha2_yeast_	HFGVFFNQGQC CIAQS RITV HESIY DEIVER DLEKAK KKQ--V LG--NP FESDTRY GPQIL	
dhac_mouse_	HHGVFYHQGQC CCVAAS RIFVEE SYDEFV KRSVER AKKY--V LG--NP LTPGINQ GPQID	
dha5_yeast_	AAGIFYNS GQNCT ANSRVY VQSSIY DKFVE KFETAK KEWDV AGKFD PFDEK CIVGP VIS	
dhal_ecoli	AAGIFYNQG QVCIAG TRLLE ESIADE FLALL KQQAQ NW--Q PG--H PLDPATT MGTLID	
dha2_yeast_	KIEFD SIPRL INSAKA EG--AK VLCGG GRDD SCV GYYIQ PTVFAD VTDEM RIAKEE IFGP	
dhac_mouse_	KEQHD KILD LIESG KKEG--AK LEC GGGR WGN-K GFFV QPTVFS NVTDEM RIAKEE IFGP	
dha5_yeast_	STQYD RIKSYIER GKREE KLD MFQTSE FPI GGAK GYFIP PTIFT DVP QTS KLLQ DEIFGP	
dhal_ecoli	CAHAD SVHS FIRE GESKG--- QLLD GRNAG-L AAAIG PTIFV DVDP NASLS REEIFGP	
dha2_yeast_	VITISRF KSVDEAI KRV DNTKY GLAAY VFTK--D KAIR ISAALK AGTV WVNC HVASY QI	
dhac_mouse_	VQQIM KF KS VDD VIK RANN TTY GLAAG LFT KLD DKAIT VSS ALQAG VVW NCY IMLS AQ	
dha5_yeast_	VVVVSKFT NYDD ALK LAND TCY GLASA VFTK DV KKAH MFARD IKAG TV WIN SSN DED VT	
dhal_ecoli	VLVV TRFT SEE QAL QLAND SQY GLGAA WTR DLS RAHR MSRR LKAG SVF VNNY NDG DMTV	
dha2_yeast_	PFGGN KNSGM GRE LGEY GLE-----	
dhac_mouse_	PFGGF KMSGNG GRE LGEH GLYEY TEL KT VAM KIS QKNS	
dha5_yeast_	PFGGF KMSGIG GRE LGQSG VDTY LQTK AVH INL SLDN-	
dhal_ecoli	PFGGY KQSGN GRDK SLHA LEK FT ELK TIWI-----	

1 60

oth:1bbt3 -----FTNLLDVAEACPTFLRFEGGPYVTTKTDSDRVLAQFDMSLAAKHMSNTFL
oth:1aym3 VKNLIEMCQVDTLIPINSTQSNIGNVSMYTVTLSPOTKLAEEIFAIKVDIASHPLATTI
oth:1bbt2 -----GLETRV---VQAERFFK--THLFDWTSDFGRCHLLELPTDH--KGVY
oth:1aym1 -----EMSVESFLGRSG--CIHESVLDIVDNYNDQSFTKWNINLQEMAQIRRKFEMFTY
oth:1bbt1 -----QHTDVS---FIMDRFVK----VTPQNQ----INILDLMQVPSH---TLV

61 120

oth:1bbt3 AGLAQYYTOYSGTINLHFMFTGPTDAKARYMVAYAPPGMEPPKTPEA----AAHCIHAEW
oth:1aym3 GEIASYFTHWTGSLRFSFMFCGTANTTLKVLLAYTPPGIGKPRSRKE----AMLGTHVVW
oth:1bbt2 GSLTDSYAYMRNGWDVEVTAVGNQFNNGCLLVAMVPELCSIQKRELY---QLTLFPHQFI
oth:1aym1 ARFDSEITMVP-SVAAKDGHIGHIIVMQ---YMYVPPGAPIPTTRDDYAWQSGTNASVFW
oth:1bbt1 GGLLRASTYYFSDLEIAVKHEG--D----LTWVPNGA--PEKALD----NTTNPTAYH

121 180

oth:1bbt3 DTGLNSKFTFSIPYLS--AADYTYTASDVAETTN-VQGWVCL-----FOITHGKAD--
oth:1aym3 DVGLQSTVSLVVPWIS--ASQYRFTTPDTYSSAGYITCWYQTN-----FVVPPNTPN--
oth:1bbt2 NPRTNMTAHITVPFVG--VNRYDQYKVHKPWTLV--VMVVAP-----LTVNTEGAP--
oth:1aym1 QHGQPFPFR-FSLPFLSIASAYYMFYDGYDGDTYKSRYGTVVTNDMGTCSRIVTSEQLHK
oth:1bbt1 KAPLTR---LALPYT---APHRVLATVYNGECXX--XRTLPTS-----FNYGAIKATR-

181 209

oth:1bbt3 ---GDALVVLASAGKDFELRLPVDARAE--
oth:1aym3 ---TAEMLCFVSGCKDFCLRМАRDТDLHK
oth:1bbt2 ---QIKVYANIAPTNVHVAG-EFPS-KE--
oth:1aym1 VKVVTRIYHKAKHTKAWCPR-PPRA-VQ--
oth:1bbt1 ---VTELLYRMKRAETYCPR-PLLA-IH-

ClustalW

ATT1_DROME ATTA_DROME SW_P36193	MQKTSILILA---LFAIAEAVP----TTGP <ins>IRVRRQVLGGSLASNPAGGADARLNLSKGIG</ins> MQKTSILIVALVALFAITEALPSLPTTGP <ins>IRVRRQVLGGSLTSNPAGGADARLDLTKGIG</ins> - MKFTIVFLLLACVFAMAVATP----GKPRP-----YSPR <ins>PTSHPRP</ins> -IRVRR---
ATT1_DROME ATTA_DROME SW_P36193	NPNHNVVGQVFAAGNTQSGPVTTGGTLAYNNAGHGASLTKTHTPGVKDVFQQEAHANLFN NP NHNVVGQVFAAGNTQSGPVTTGGTLAYNNAGHGASLTKTHTPGVKDVFQQEAHANLFN - - EALAIEDHLAQAAIRPPPILPA-----
ATT1_DROME ATTA_DROME SW_P36193	NGRHNLDAKVFASQNKLNGFEFQRNGAGLDYSHINGHGASLTHSNFPFIGQQLGLDGRA NGRHNLDAKVFASQNKLNGFEFQRNGAGLDYSHINGHGASLTHSNFPFIGQQLGLDGRA -----
ATT1_DROME ATTA_DROME SW_P36193	NLWSSPNRATTLDLTGSASKWTSGPFANQKPNFGAGLGLSHHFG NLWSSPNRATTLDLTGSASKWTSGPFANQKPNFGAGLGLSHHFG -----

T-Coffee

ATT1_DROME ATTA_DROME SW_P36193	MQKTSILILAL---FAIAEAVP-----TTG---PIRVRRQVLGGSLASNPAGGADA MQKTSILIVALVALFAITEALPSL-----PTTG---PIRVRRQVLGGSLTSNPAGGADA MKFTIVFLLLA-CVFAMAVATPGKPRPYS PRPTSHPRP <ins>IRVRREAL</ins> -----
ATT1_DROME ATTA_DROME SW_P36193	RLNLSKGIGNPNHNVVGQVFAAGNTQSGPVTTGGTLAYNNAGHGASLTKTHTPGVKDVFQ RLDLTKGIGNPNHNVVGQVFAAGNTQSGPVTTGGTLAYNNAGHGASLTKTHTPGVKDVFQ -----AIEDHLAQAAIRPPPILPA-----
ATT1_DROME ATTA_DROME SW_P36193	QEAHANLFNNGRHNLDAKVFASQNKLNGFEFQRNGAGLDYSHINGHGASLTHSNFPFIG QEAHANLFNNGRHNLDAKVFASQNKLNGFEFQRNGAGLDYSHINGHGASLTHSNFPFIG -----
ATT1_DROME ATTA_DROME SW_P36193	QQLGLDGRANLWSSPNRATTLDLTGSASKWTSGPFANQKPNFGAGLGLSHHFG QQLGLDGRANLWSSPNRATTLDLTGSASKWTSGPFANQKPNFGAGLGLSHHFG -----

T-Coffee

ATT1_DROME	MQKTS I L I L A L - - - F A I A E A V P - - - - - - - - -	TTG - - P I R V R R Q V L G G S L A S N P A G G A D A
ATTA_DROME	MQKTS I L I V A L V A L F A I T E A L P S L - - - - -	PTTG - - P I R V R R Q V L G G S L T S N P A G G A D A
SW_P36193	M K F T I V F L L L A C V F A M A V A T P G K P R P Y S P R P T S H P R P I R V R R E A L - - - - -	
ATT1_DROME	R L N L S K G I G N P N H N V V G Q V F A A G N T Q S G P V T T G G T L A Y N N A G H G A S L T K T H T P G V K D V F Q	
ATTA_DROME	R L D L T K G I G N P N H N V V G Q V F A A G N T Q S G P V T T G G T L A Y N N A G H G A S L T K T H T P G V K D V F Q	
SW_P36193	- - - - - A I E D H L A Q A A I R P P P I L P A - - - - -	
ATT1_DROME	Q E A H A N L F N N G R H N L D A K V F A S Q N K L A N G F E F Q R N G A G L D Y S H I N G H G A S L T H S N F P G I G	
ATTA_DROME	Q E A H A N L F N N G R H N L D A K V F A S Q N K L A N G F E F Q R N G A G L D Y S H I N G H G A S L T H S N F P G I G	
SW_P36193	- - - - -	
ATT1_DROME	Q Q L G L D G R A N L W S S P N R A T T L D L T G S A S K W T S G P F A N Q K P N F G A G L G L S H H F G	
ATTA_DROME	Q Q L G L D G R A N L W S S P N R A T T L D L T G S A S K W T S G P F A N Q K P N F G A G L G L S H H F G	
SW_P36193	- - - - -	

DiAlign

ATT1_DROME	MQKTS I L I L A L F A I A E A V - - - - - - - - -	PTTG - - - - - P I R V R R Q V L G G S L A S N P A G G A D A
ATTA_DROME	MQKTS I L I V A L V A L F A I T E A L P S L P T T G - - - - -	PTTG - - - - - P I R V R R Q V L G G S L T S N P A G G A D A
SW_P36193	- M K F T I V F L L L A C V F A M A V A T P G K P R P Y S P R P T S H P R P I R V R R E A L A - - - - -	
ATT1_DROME	R L N L S K G I G N P N H N V V G Q V F A A G N T Q S G P V T T G G T L A Y N N A G H G A S L T K T H T P G V K D V F Q	
ATTA_DROME	R L D L T K G I G N P N H N V V G Q V F A A G N T Q S G P V T T G G T L A Y N N A G H G A S L T K T H T P G V K D V F Q	
SW_P36193	- - - - - I E D H L A - - - - -	
ATT1_DROME	Q E A H A N L F N N G R H N L D A K V F A S Q N K L A N G F E F Q R N G A G L D Y S H I N G H G A S L T H S N F P G I G	
ATTA_DROME	Q E A H A N L F N N G R H N L D A K V F A S Q N K L A N G F E F Q R N G A G L D Y S H I N G H G A S L T H S N F P G I G	
SW_P36193	Q A A I R P P P I L P A - - - - -	
ATT1_DROME	Q Q L G L D G R A N L W S S P N R A T T L D L T G S A S K W T S G P F A N Q K P N F G A G L G L S H H F G	
ATTA_DROME	Q Q L G L D G R A N L W S S P N R A T T L D L T G S A S K W T S G P F A N Q K P N F G A G L G L S H H F G	
SW_P36193	- - - - -	

MSA Evaluation

- **AltAVisT - A WWW tool for comparison of alternative multiple alignments**
<http://bibiserv.techfak.uni-bielefeld.de/altavist/>

- **T-Coffee Server**
<http://igs-server.cnrs-mrs.fr/Tcoffee/>

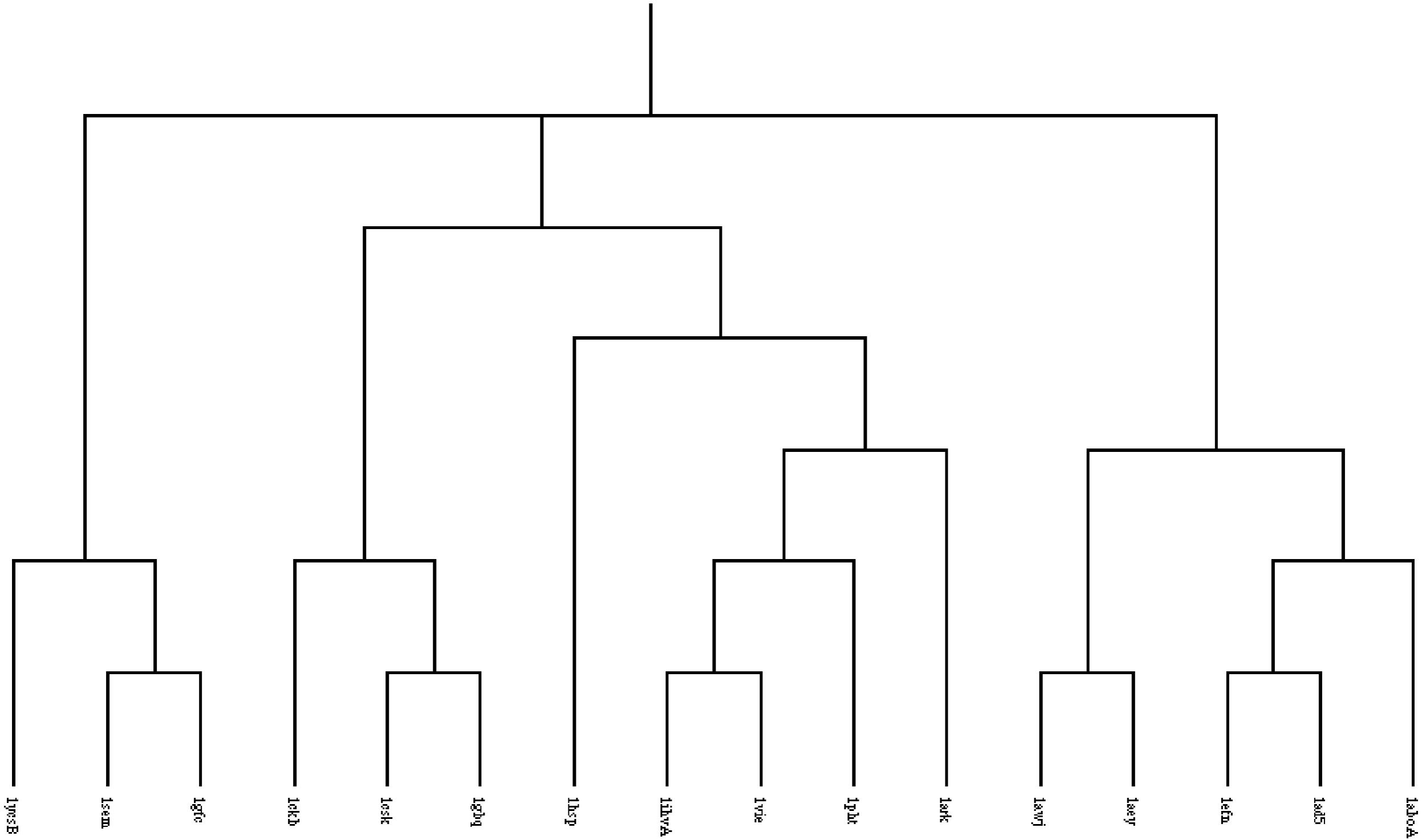
- **BaliScore comparison**
<http://genome.nci.nih.gov/tools/msacomp.html>

MSA hurdles

- Too many sequences
- Repeated sequences are renowned for confusing existing methods
- MSA methods mostly not parallelized and so still require “super computers”
- Combine 3D structural info
- Precomputed families - curated by experts (no need for complete alignment)

Tree - Dendogram

(clustering, not phylogeny)



Tree Viewing/Drawing

- **Phyodendron Phylogenetic tree printer**
<http://iubio.bio.indiana.edu/treeapp/treeprint-form.html>

- **TreeTop** - Phylogenetic Tree Prediction
http://www.genebee.msu.su/services/phtree_reduced.html

- **TreeView** (local view and print)
<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>

- **NJPLOT (ClustalW)**
<ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX>

Pretty Output

■ **Alscript**

<http://www.compbio.dundee.ac.uk/Software/Alscript/alscript.html>

■ **Pretty EMBOSS**

<http://www.emboss.org/>

■ **BOXSHADE**

<http://bioweb.pasteur.fr/seqanal/interfaces/boxshade.html>

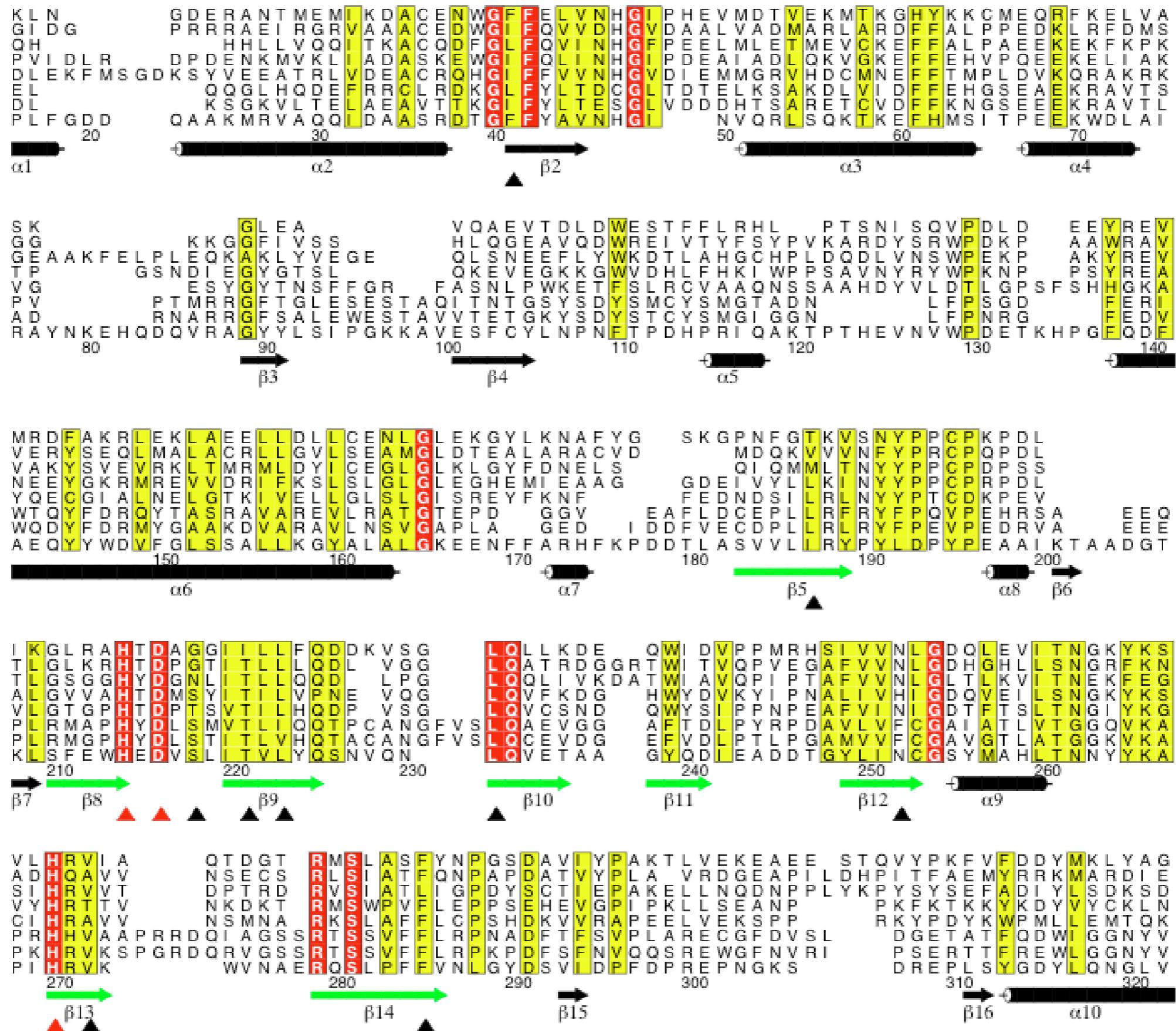
■ **EScript**

<http://prodes.toulouse.inra.fr/EScript/EScript/>

■ **AMAS**

<http://www.compbio.dundee.ac.uk/amas/>

Alscript - Output



Editors

- **JalView (J)**

<http://www.compbio.dundee.ac.uk/Software/JalView/jalview.html>

- **CINEMA (J)**

<http://bioinf.man.ac.uk/dbbrowser/CINEMA2.1/>

- **Seaview (UMP)**

<http://pbil.univ-lyon1.fr/software/seaview.html>

- **MPSA (UM)**

<http://mpsa-pbil.ibcp.fr/>

- **Se-Al (M)**

<http://evolve.zoo.ox.ac.uk/software.html?id=seal>

- **ClustalX (UMP)**

<ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX>

Multiple Genome Alignment

■ MGA

Michael Högl, Stefan Kurtz ,Enno Ohlebusch

Efficient Multiple Genome Alignment Bioinformatics , Vol. 18 (S1): S312-S320, 2002

<http://bibiserv.techfak.uni-bielefeld.de/mga/ref.html>

■ PipMaker and MultiPipMaker

Schwartz S, Elnitski L, Li M, et al.

MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences

NUCLEIC ACIDS RES 31 (13): 3518-3524 JUL 1 2003

<http://bio.cse.psu.edu/pipmaker/>

■ MAVID

Bray N and Pachter L ,MAVID multiple alignment server , Nucleic Acids Research 2003 31: 3525-3526

<http://baboon.math.berkeley.edu/mavid/>

<http://www-gsd.lbl.gov/vista/>

MGA - output

G₁ M G₂ M G₃ MA(MA)x(3)MAM G₄ M G₅ M G₆ M G₇ MAMAM G₈ MA(MA)x(5)MAM G₉ MA(MA)x(87)MAM
G₁₀ MA(MA)x(1)MAM G₁₁ M G₁₂ MA(MA)x(10)MAM G₁₃ MA(MA)x(10)MAM G₁₄ MA(MA)x(12)MAM G₁₅ MAMAM
G₁₆ MAM G₁₇ MA(MA)x(1)MAM G₁₈ MA(MA)x(20)MAM G₁₉ M G₂₀ MA(MA)x(53)MAM G₂₁ MA(MA)x(38)MAM G₂₂
M G₂₃ M G₂₄ M G₂₅ M G₂₆ MA(MA)x(4)MAM G₂₇ MA(MA)x(1)MAM G₂₈ M G₂₉ MA(MA)x(27)MAM G₃₀
MA(MA)x(1)MAM G₃₁ MA(MA)x(12)MAM G₃₂ MA(MA)x(3)MAM G₃₃ MA(MA)x(18)MAM G₃₄ MAM G₃₅
MA(MA)x(6)MAM G₃₆ MA(MA)x(44)MAM G₃₇ MA(MA)x(10)MAM G₃₈ MA(MA)x(1)MAM G₃₉ MA(MA)x(3)MAM G₄₀
MA(MA)x(9)MAM G₄₁ MA(MA)x(19)MAM G₄₂ MA(MA)x(1)MAM G₄₃ MAM G₄₄ MA(MA)x(3)MAM G₄₅
MA(MA)x(42)MAM G₄₆ MA(MA)x(31)MAM G₄₇ MA(MA)x(10)MAM G₄₈ MA(MA)x(1)MAM G₄₉ M G₅₀ MAMAM G₅₁ M
G₅₂ MAMAM G₅₃ M G₅₄ M G₅₅ M G₅₆ M G₅₇ M G₅₈

MA(MA)x(6)MAM G₃₆ MA(MA)x(44)MAM G₃₇ MA(MA)x(10)MAM G₃₈ MA(MA)x(1)MAM G₃₉ MA(MA)x(3)MAM G₄₀
MA(MA)x(9)MAM G₄₁ MA(MA)x(19)MAM G₄₂ MA(MA)x(1)MAM G₄₃ MAM G₄₄ MA(MA)x(3)MAM G₄₅
MA(MA)x(42)MAM G₄₆ MA(MA)x(31)MAM G₄₇ MA(MA)x(10)MAM G₄₈ MA(MA)x(1)MAM G₄₉ M G₅₀ MAMAM G₅₁ M
G₅₂ MAMAM G₅₃ M G₅₄ M G₅₅ M G₅₆ M G₅₇ M G₅₈

Startpositions: 198825 202971

Seq 1: ttctgagttcttgtccacttcttagcctgt-ctgtcccttt-ccttgctactctggatcaacagtacacctcttgaactttgggtgcttggcaa 100
Seq 2:t....gg....c.tt.t.c.....t....c.....t.t....g....gg....g....g....caacag..ca.c....

Seq 1: ---tagttcctct-ccctactcctaattacggca-ggcc-ttagaaaaccataaccttattttaaaggtaaaaaaaaaaaaaagatttaagacaaaagcaa 200
Seq 2: gcc....a....-.c....g.t.t....c....-.g..aat.....g....-.a....-.acc..g....gc--....c.t.g...gg....c.

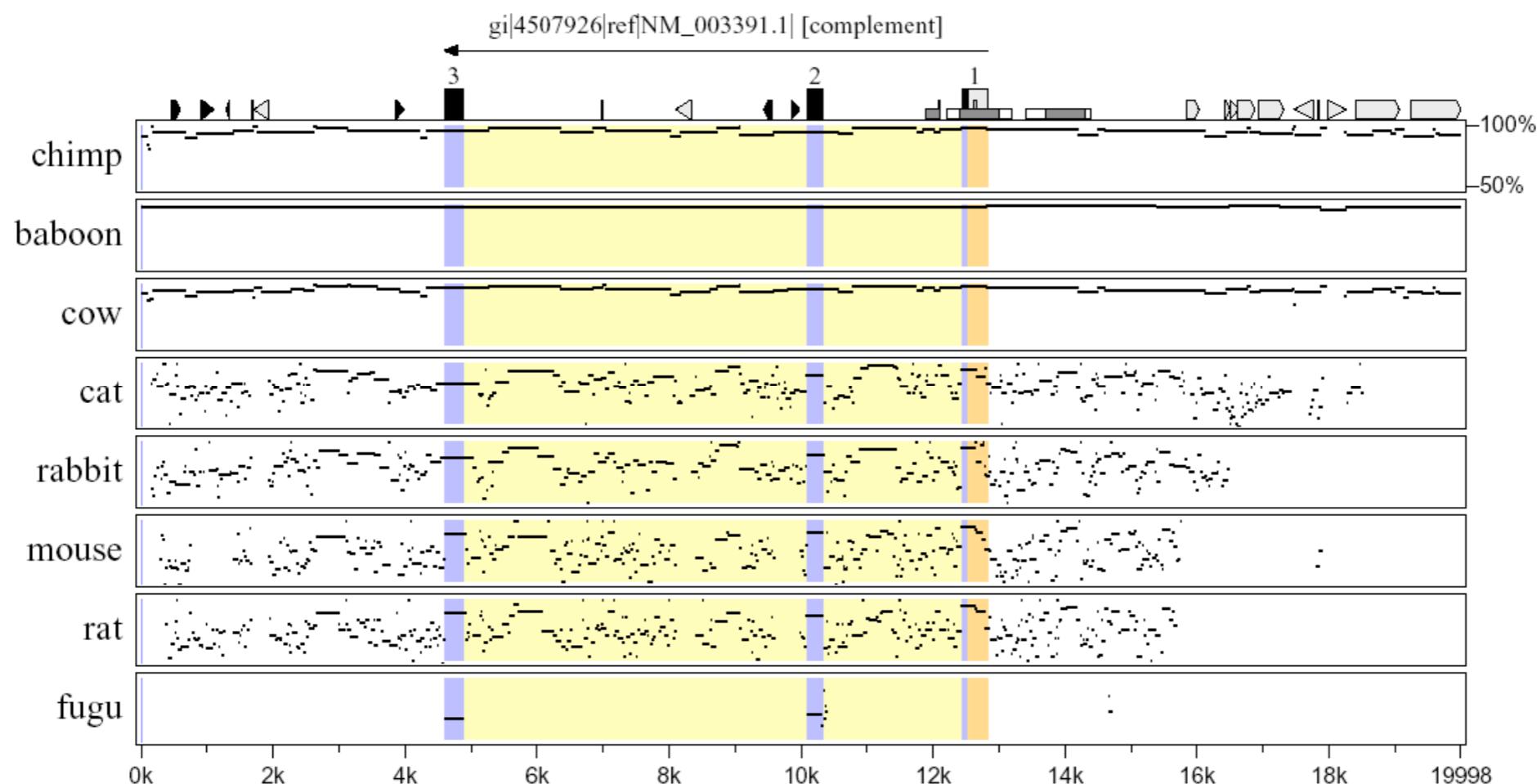
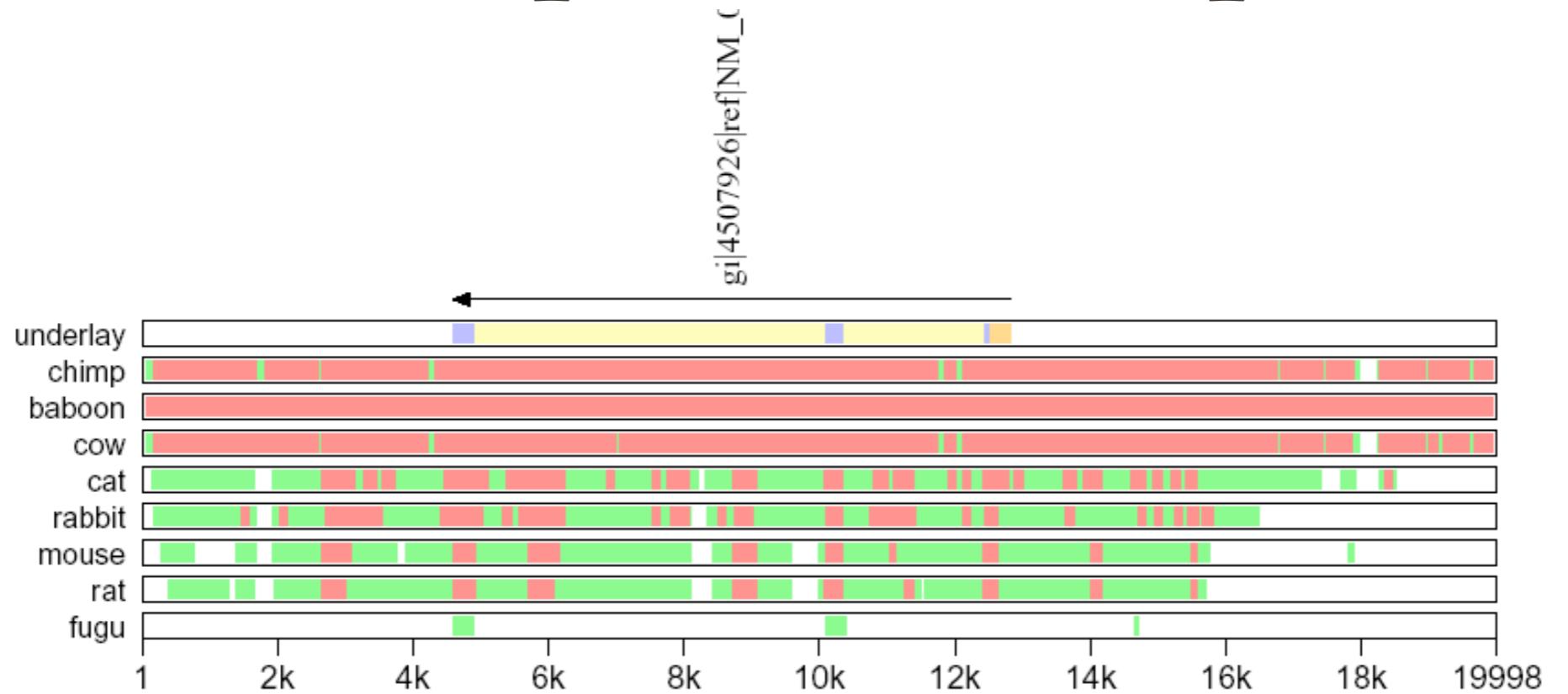
Seq 1: ggggctt-gg-gtgccttccttatgga--ct-t--aggcctggtaacatct-gttctggccacttagaggccttgtgtctatttcttgcagg 300
Seq 2: ...t..cg..c.....gca...cg..g.ac.....a.c..g..t.g.....tg..----a---g--c..g.c----gaa

Seq 1: gtttgcaaggaggacgttg-tttagttccaaacaggtgaggattgcac--actagcaaacacatgagaagaaggcggaggaattggagaaaaataa 400
Seq 2: .-..g.g.....t.c.g....ca....c.ct....c....a..g.ag....g.gg.gt----t....-t..a.t.ca..c..g....c.

Seq 1: aaagaatgcagcagccaggttag-caggaacgttaagacggta-cggagaacagcaaaggcttggaaagcaagccgcgtggagaaggaa--g---aact 500
Seq 2:cag..a.a..gg.....cc....accg....caca.ga..t..-gg.....t..ctc.g....tt....aa....-....ggg.aca....

Seq 1: gtgctgaggtagttgtgacaacccaggctgatggatgtaaacaccaaacccttggctggctcgctcagcgggcttggagct 600
Seq 2:-.....c...cc..gg..g.....tg.....c.....a..t..c.....a....g.c....

MultiPipMaker - output



MAVID/VISTA - output

MAVID zoo

Alignment 1
Seqs: human/chimp
Criteria: 75%, 100 bp
Regions: 16

Alignment 2
Seqs: human/baboon
Criteria: 75%, 100 bp
Regions: 26

Alignment 3
Seqs: human/cat
Criteria: 75%, 100 bp
Regions: 82

Alignment 4
Seqs: human/dog
Criteria: 75%, 100 bp
Regions: 62

Alignment 5
Seqs: human/cow
Criteria: 75%, 100 bp
Regions: 73

Alignment 6
Seqs: human/pig
Criteria: 75%, 100 bp
Regions: 73

Alignment 7
Seqs: human/mouse
Criteria: 75%, 100 bp
Regions: 21

Alignment 8
Seqs: human/rat
Criteria: 75%, 100 bp
Regions: 19

Alignment 9
Seqs: human/chicken
Criteria: 75%, 100 bp
Regions: 3

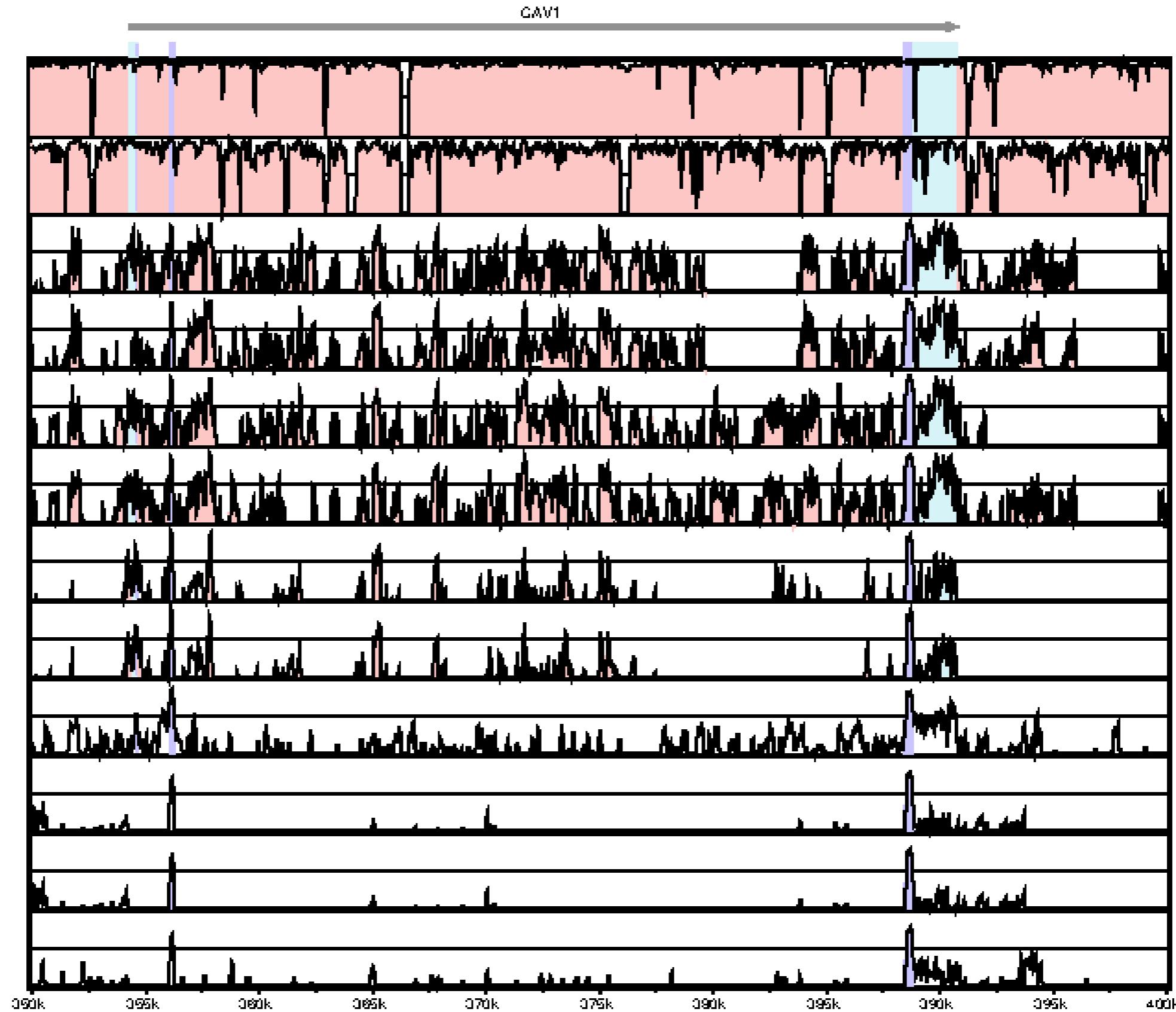
Alignment 10
Seqs: human/fugu
Criteria: 75%, 100 bp
Regions: 1

Alignment 11
Seqs: human/tetra
Criteria: 75%, 100 bp
Regions: 2

Alignment 12
Seqs: human/zebrafish
Criteria: 75%, 100 bp
Regions: 1

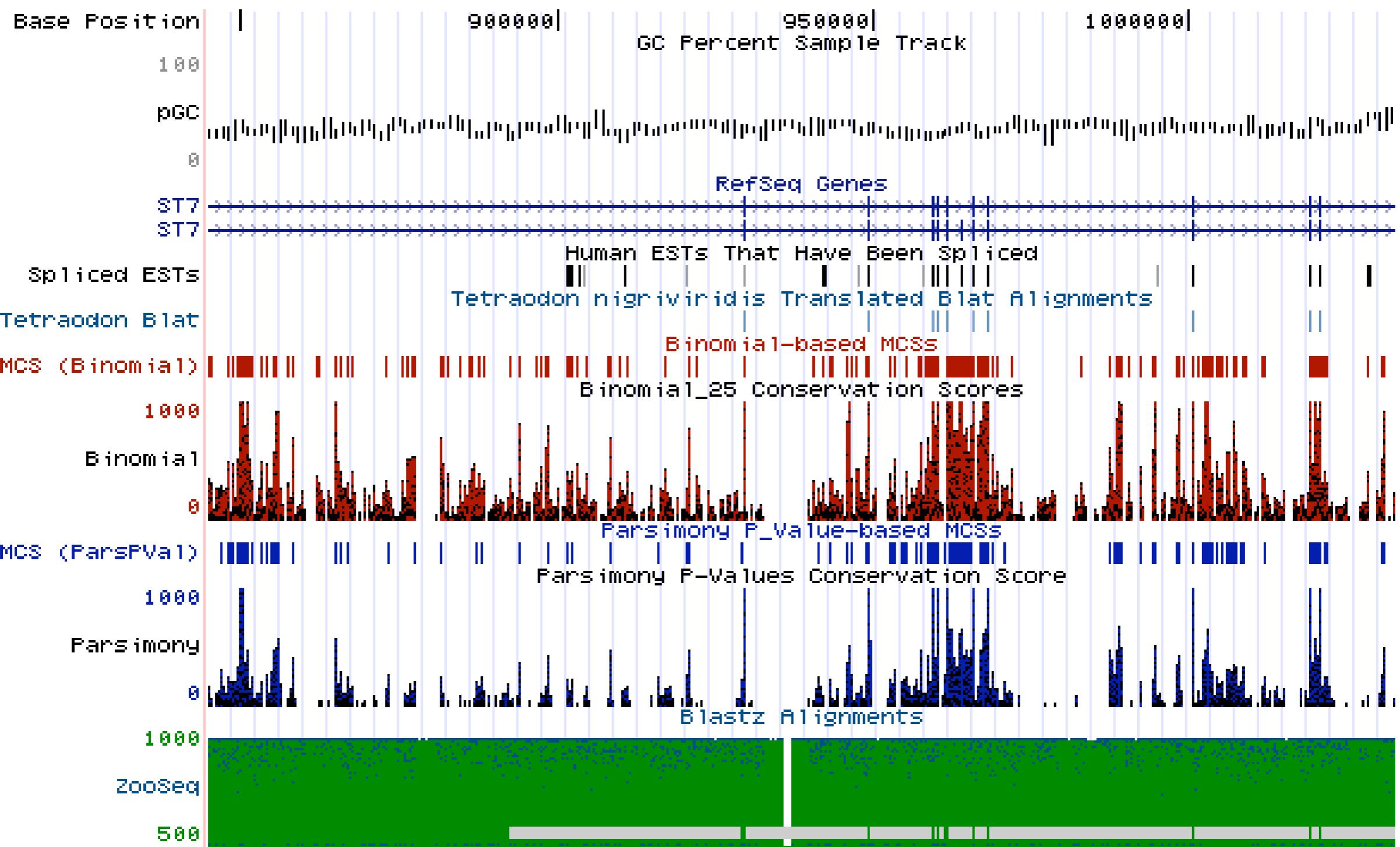
X-axis: human
Resolution: 25
Window size: 100 bp

gene
exon
UTR
CNS



Genomic Targets for Comparative Sequencing

<http://genome.ucsc.edu/>



References

■ **BALiBASE: a benchmark alignment database for the evaluation of multiple alignment programs**

Thompson JD, Plewniak F, Poch O. Bioinformatics. 1999 Jan;15(1):87-8.

■ **A comprehensive comparison of multiple sequence alignment programs**

JD Thompson, F Plewniak, and O Poch Nucleic Acids Res. 1999 27: 2682-2690.

■ **Quality assessment of multiple alignment programs**

FEBS Letters Volume 529, Issue 1 , T. Lassmann and E Sonnhammer 2 October 2002, Pages 126-130

■ **Recent progress in multiple sequence alignment: a survey.**

Notredame C. Pharmacogenomics. 2002 Jan;3(1):131-44. Review.

■ **Strategies for multiple sequences alignment**

HB Nicholas Jr, AJ Ropelewski and DW Deerfield II, BioTechniques 32:572-591

This talk URLs

- <http://genome.nci.nih.gov/talks/msa.html>
- <http://helix.nih.gov/talks/>