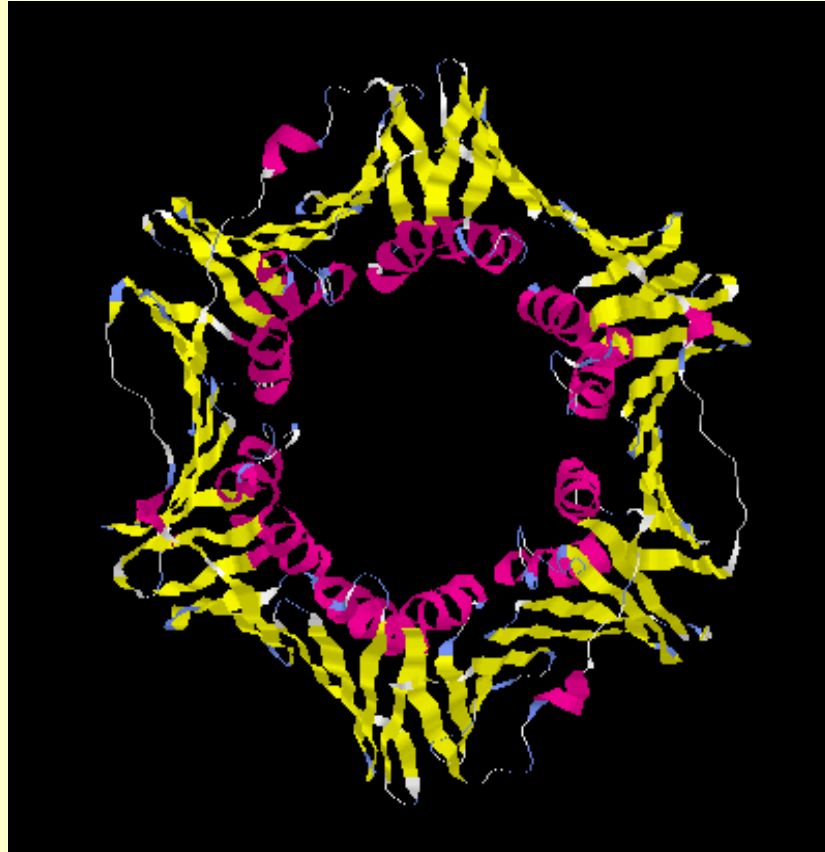


Computational Molecular Biology

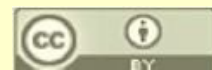
Biochem 218 – BioMedical Informatics 231

<http://biochem218.stanford.edu/>

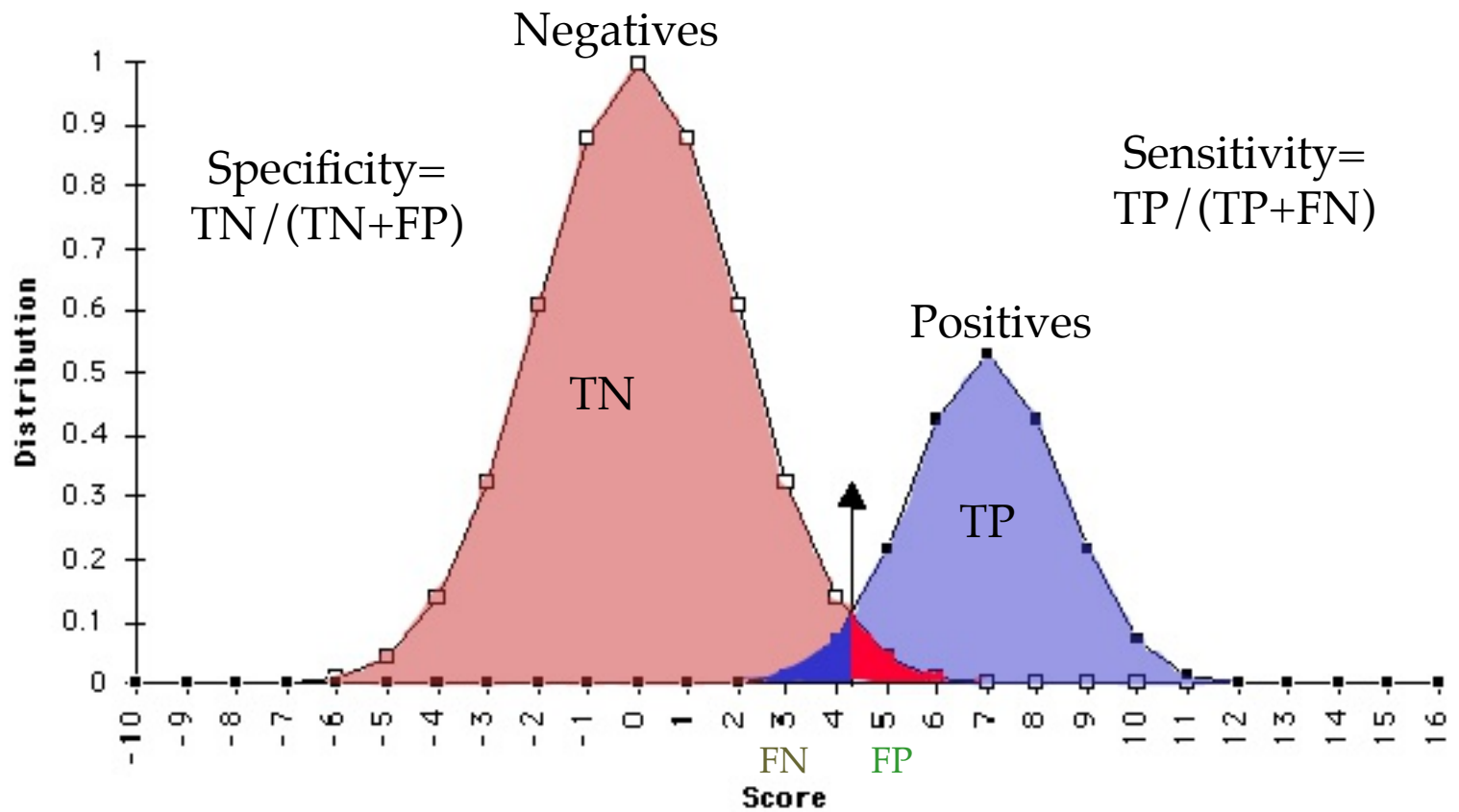
Multiple Sequence Alignment



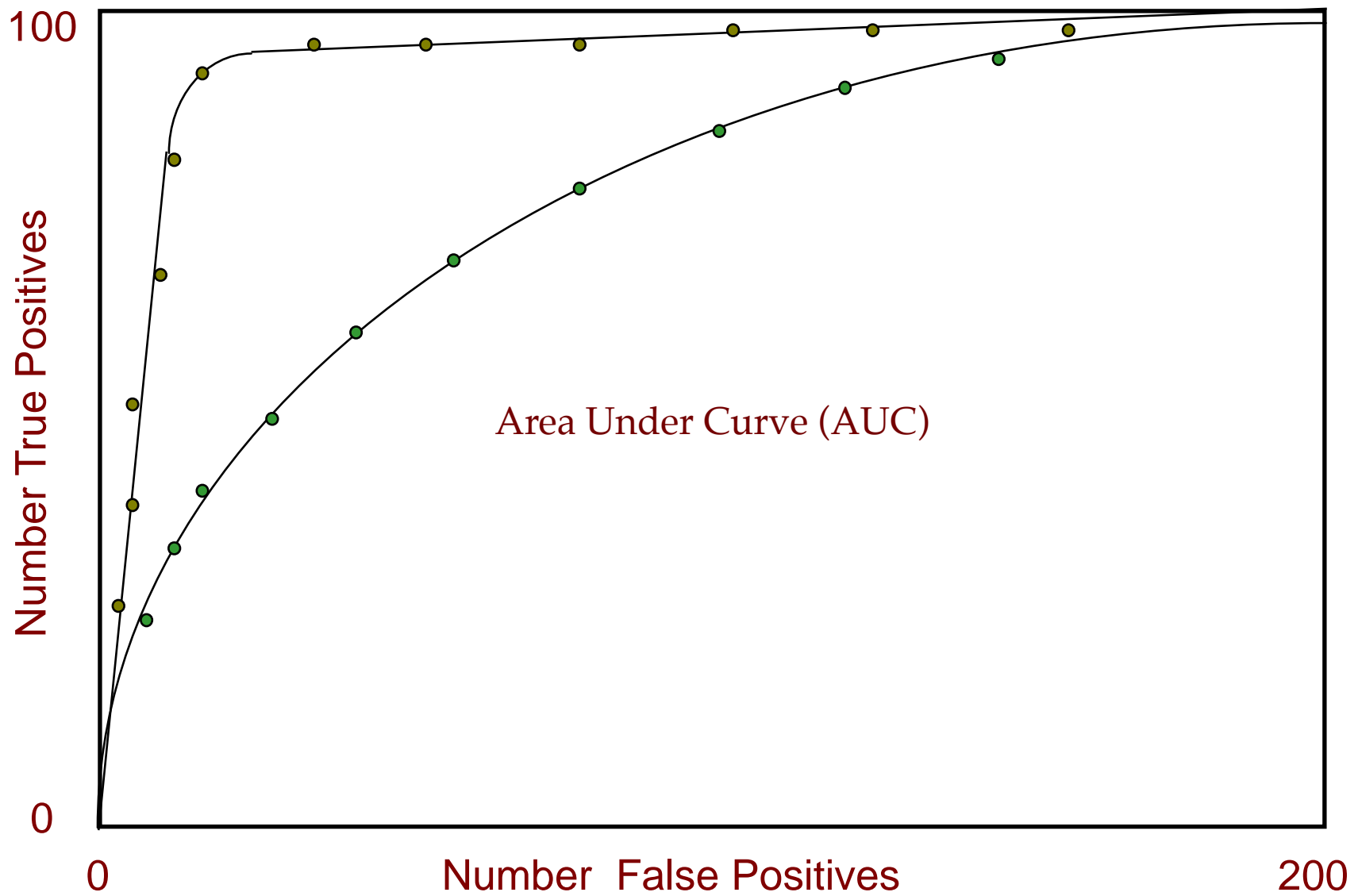
Doug Brutlag
Professor Emeritus
Biochemistry & Medicine (by courtesy)



Evaluation of Search Algorithms



Evaluation of Search Algorithms with Receiver-Operator Characteristic Curve



Pyruvate Dehydrogenase E1 Family (EC 1.2.4.1)

<http://uniprot.org/>

UniProtKB Downloads · Contact · Documentation/Help

Search Blast Align Retrieve ID Mapping *

Search in **Query**

Protein Knowledgebase (UniProtKB) ec:1.2.4.1 AND reviewed:yes AND name:alpha Search Clear Fields »

1 - 25 of 47 results for **ec:1.2.4.1** AND **reviewed:yes** AND **name:alpha** in UniProtKB sorted by **score** descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Customize display [Download...](#)

Page 1 of 2 | Next »

Accession	Entry name	Status	Protein names	Gene names	Organism	Length
P16387	ODPA_YEAST	★	Pyruvate dehydrogenase E1 component subunit alpha, mitochondrial (EC 1.2.4.1) (Pyruvate dehydrogenase complex component E1 alpha) (PDHE1-A)	PDA1 (YER178W)	Saccharomyces cerevisiae (Baker's yeast)	420
P08559	ODPA_HUMAN	★	Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial (EC 1.2.4.1) (PDHE1-A type I)	PDHA1 (PHE1A)	Homo sapiens (Human)	390
P52901	ODPA1_ARATH	★	Pyruvate dehydrogenase E1 component subunit alpha-1, mitochondrial (PDHE1-A) (EC 1.2.4.1)	At1g59900 (F23H11.21)	Arabidopsis thaliana (Mouse-ear cress)	389
Q8H1Y0	ODPA2_ARATH	★	Pyruvate dehydrogenase E1 component subunit alpha-2, mitochondrial (PDHE1-A) (EC 1.2.4.1)	IAR4 (At1g24180) (F3I6.11)	Arabidopsis thaliana (Mouse-ear cress)	393
P29803	ODPAT_HUMAN	★	Pyruvate dehydrogenase E1 component subunit alpha, testis-specific form, mitochondrial (EC 1.2.4.1) (PDHE1-A type II)	PDHA2 (PDHAL)	Homo sapiens (Human)	388
P21873	ODPA_BACST	★	Pyruvate dehydrogenase E1 component subunit alpha (EC 1.2.4.1)	pdhA	Bacillus stearothermophilus (Geobacillus stearothermophilus)	369
P35486	ODPA_MOUSE	★	Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial (EC 1.2.4.1) (PDHE1-A type I)	Pdha1 (Pdha-1)	Mus musculus (Mouse)	390
P26284	ODPA_RAT	★	Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial (EC 1.2.4.1) (PDHE1-A type I)	Pdha1	Rattus norvegicus (Rat)	390

Decypher Home Page

<http://decypher.stanford.edu/>

Algorithm and Feature Index

The following links will take you to specific algorithm pages. [On-line Product Documentation Set and Web Links](#)

Algorithm	Query vs. Database Types
Tera-Blast™ N	DNA to DNA ⓘ
	DNA to DNA ⓘ
Tera-Blast™ P	DNA to Protein ⓘ
	Protein to DNA ⓘ
	Protein to Protein ⓘ
	DNA to DNA ⓘ
GeneDetective™	Genomic DNA to Coding DNA ⓘ
	Coding DNA to Genomic DNA ⓘ
	Genomic DNA to Protein ⓘ
	Protein to Genomic DNA ⓘ
	Genomic DNA to Protein HMM ⓘ
	Protein HMM to Genomic DNA ⓘ
ClustalW	DNA ⓘ
	Protein ⓘ

Algorithm	Query vs. Database Types
Smith-Waterman Standard, Double-Affine	DNA to DNA ⓘ
	DNA to Protein ⓘ
	Protein to Protein ⓘ
	Protein to DNA ⓘ
FrameSearch Symmetric Frame Independent™ for DNA to DNA	DNA to DNA ⓘ
	DNA to Protein ⓘ
	Protein to DNA ⓘ
Hidden Markov Model (HMM)	DNA to Protein HMM ⓘ
	Protein to Protein HMM ⓘ
	Protein HMM to Protein ⓘ
	Protein HMM to DNA ⓘ
HMM FrameSearch	DNA to Protein HMM ⓘ
	Protein HMM to DNA ⓘ
ProfileSearch	DNA to Protein Profile ⓘ
	Protein To Protein Profile ⓘ
	Protein Profile to Protein ⓘ
Profile FrameSearch	Protein Profile to DNA ⓘ
	DNA to Protein Profile ⓘ
	Protein Profile to DNA ⓘ

Decypher Search Input

<http://decypher.stanford.edu/>

DeCypher Smith-Waterman Search
Protein Query vs. Protein Database

Return Results: As:

Protein Query:
 Click *Browse...* to upload your local file, or paste query data into the text box. Use only one query entry if requesting a Graphic reply.
[Use Example Query](#)

```
>PDH E1 HUMAN
MRKMLAAVSRVLSGASQKPA SRVLVASRNFANDATFEIKKCDLHRLEEGPPVTTVLTRED
GLKYYRMMQTVRRMELKADQLYKQKIIRGFCHLCDGQEAACVGLAEPDHLITAYRA
HGFTFTRGLSVREILAEITGRKGGCAKGGKGGSMHMYAKNFYGGNGIVGAQVPLGAGIALA
CKYNGKDEVCLTLYGDGAANQOIF EAYNMAALWKLPCIFICENNRVGMGTSVERAAAST
DYYKRGDFIPGLRVDGMDILCVREATRFAAAAYCRSGKGPILMELQTYRYHGHSMSPGVS
```

Description	Entries	Symbols	Updated
<i>Thermotoga maritima</i> (NCBI)	1858	593368	19-March-2003
<i>Treponema pallidum</i> (NCBI)	1036	356216	19-March-2003
tumora	1	393	19-December-2004
Uniprot-SwissProt: 57.1	514212	180900945	26-January-2010
Uniprot-TREMBL: 57.13	9145906	2958343669	27-January-2010
<i>Ureaplasma urealyticum</i> (614	231000	18-March-2002

Job Options:

Algorithm Variation:	<input type="button" value="Local (Standard)"/>	Score Threshold:	<input type="text" value="1"/>
Filter Query:	<input type="checkbox"/>	Significance Threshold:	<input type="button" value="10"/>
Max Scores:	<input type="text" value="100"/>	Weight Matrix:	<input type="button" value="blosum62"/>
Max Alignments:	<input type="text" value="20"/>	Group Results by Entry:	<input type="checkbox"/>
Identity Symbol:	<input type="button" value="Match Letter"/>		
Show Significance:	<input type="button" value="E-Value"/>		
Gap Open Penalty:	<input type="text" value="12"/>		
Gap Extend Penalty:	<input type="text" value="2"/>		

Smith-Waterman Search Protein Query vs. Protein Data

	sp P29803 ODPAT_HUMAN	388	sp P29803 ODPAT_HUMAN	Pyruvate dehydrogenase
5.4e-272	sp P29803 ODPAT_HUMAN		Pyruvate dehydrogenase E1 component subunit	
8.7e-234	sp P35486 ODPA_MOUSE		Pyruvate dehydrogenase E1 component subunit	
2.3e-233	sp P08559 ODPA_HUMAN		Pyruvate dehydrogenase E1 component subunit	
3.1e-233	sp A7MB35 ODPA_BOVIN		Pyruvate dehydrogenase E1 component subunit	
1.1e-232	sp P26284 ODPA_RAT		Pyruvate dehydrogenase E1 component subunit a	
2.1e-232	sp A5A6L0 ODPA_PANTR		Pyruvate dehydrogenase E1 component subunit	
2.1e-232	sp Q5R490 ODPA_PONAB		Pyruvate dehydrogenase E1 component subunit	
2.9e-232	sp P29804 ODPA_PIG		Pyruvate dehydrogenase E1 component subunit a	
9.4e-231	sp Q8HXW9 ODPA_MACFA		Pyruvate dehydrogenase E1 component subunit	
8.7e-218	sp P52900 ODPA_SMIMA		Pyruvate dehydrogenase E1 component subunit	
1.4e-211	sp P35487 ODPAT_MOUSE		Pyruvate dehydrogenase E1 component subunit	
9.4e-211	sp Q06437 ODPAT_RAT		Pyruvate dehydrogenase E1 component subunit	
8.7e-146	sp P52899 ODPA_CAEL		Probable pyruvate dehydrogenase E1 componen	
3.6e-135	sp P26268 ODPT_ASCSU		Pyruvate dehydrogenase E1 component subunit	
5.9e-133	sp P26267 ODPA_ASCSU		Pyruvate dehydrogenase E1 component subunit	
5.0e-127	sp Q54C70 ODPA_DICDI		Pyruvate dehydrogenase E1 component subunit	
1.0e-119	sp P16387 ODPA_YEAST		Pyruvate dehydrogenase E1 component subunit	
1.5e-116	sp P52901 ODPA1_ARATH		Pyruvate dehydrogenase E1 component subuni	
5.4e-116	sp Q13366 ODPA_KLULA		Pyruvate dehydrogenase E1 component subunit	
1.4e-115	sp Q10489 ODPA_SCHPO		Pyruvate dehydrogenase E1 component subunit	
5.0e-115	sp Q8H1Y0 ODPA2_ARATH		Pyruvate dehydrogenase E1 component subuni	
1.8e-114	sp P52902 ODPA_PEA		Pyruvate dehydrogenase E1 component subunit a	
3.1e-113	sp P52903 ODPA_SOLTU		Pyruvate dehydrogenase E1 component subunit	
1.9e-111	sp Q9R9N5 ODPA_RHIME		Pyruvate dehydrogenase E1 component subunit	
6.9e-091	sp Q66112 ODPA_ZYMMO		Pyruvate dehydrogenase E1 component subunit	
5.4e-088	sp Q4UKQ6 ODPA_RICFE		Pyruvate dehydrogenase E1 component subunit	
3.1e-085	sp Q92IS3 ODPA_RICCN		Pyruvate dehydrogenase E1 component subunit	
4.3e-085	sp Q1RJK4 ODPA_RICBR		Pyruvate dehydrogenase E1 component subunit	
1.1e-084	sp Q68XA9 ODPA_RICTY		Pyruvate dehydrogenase E1 component subunit	
7.5e-084	sp Q9ZDR4 ODPA_RICPR		Pyruvate dehydrogenase E1 component subunit	
8.1e-069	sp Q1XDM0 ODPA_PORYE		Pyruvate dehydrogenase E1 component subunit	
6.4e-066	sp P51267 ODPA_PORPU		Pyruvate dehydrogenase E1 component subunit	
3.7e-063	sp P27745 ACOA_RALEH		Acetoin: 26-dichlorophenolindophenol oxidore	
4.3e-061	sp Q31404 ACOA_BACSU		Acetoin: 26-dichlorophenolindophenol oxidore	
1.0e-035	sp P37940 ODBA_BACSU		2-oxoisovalerate dehydrogenase subunit alph	
2.3e-033	sp Q72GU1 ODBA_THET2		2-oxoisovalerate dehydrogenase subunit alph	
8.1e-033	sp Q5SLR4 ODBA_THET8		2-oxoisovalerate dehydrogenase subunit alph	
1.5e-032	sp P35485 ODPA_ACHLA		Pyruvate dehydrogenase E1 component subunit	
5.4e-032	sp Q54M22 ODBA_DICDI		2-oxoisovalerate dehydrogenase subunit alph	
1.0e-031	sp P47516 ODPA_MYCGE		Pyruvate dehydrogenase E1 component subunit	
1.9e-031	sp Q5HG21 ODPA_STAAC		Pyruvate dehydrogenase E1 component subunit	
1.9e-031	sp P60089 ODPA_STAAM		Pyruvate dehydrogenase E1 component subunit	
1.9e-031	sp Q820A6 ODPA_STAAN		Pyruvate dehydrogenase E1 component subunit	
1.9e-031	sp Q6GAC1 ODPA_STAAS		Pyruvate dehydrogenase E1 component subunit	

General DNA Similarity Search Principles

- Search both Strands
- Translate ORFs and cDNAs
- Use most sensitive search possible
 - UnGapped BLAST for infinite gap penalty (PCR & CHIP oligos)
 - Gapped BLAST for most searches
 - Smith Waterman or megaBLAST or discontinuous MegaBLAST for cDNA / genome comparisons
 - cDNA => Zero gap-length penalty
 - Consider using transition matrices
 - Ensure that expected value of score is negative
- Examine results with exp. between 0.05 and 10
- Reevaluate results of borderline significance using limited query

General Protein Similarity Search Principles

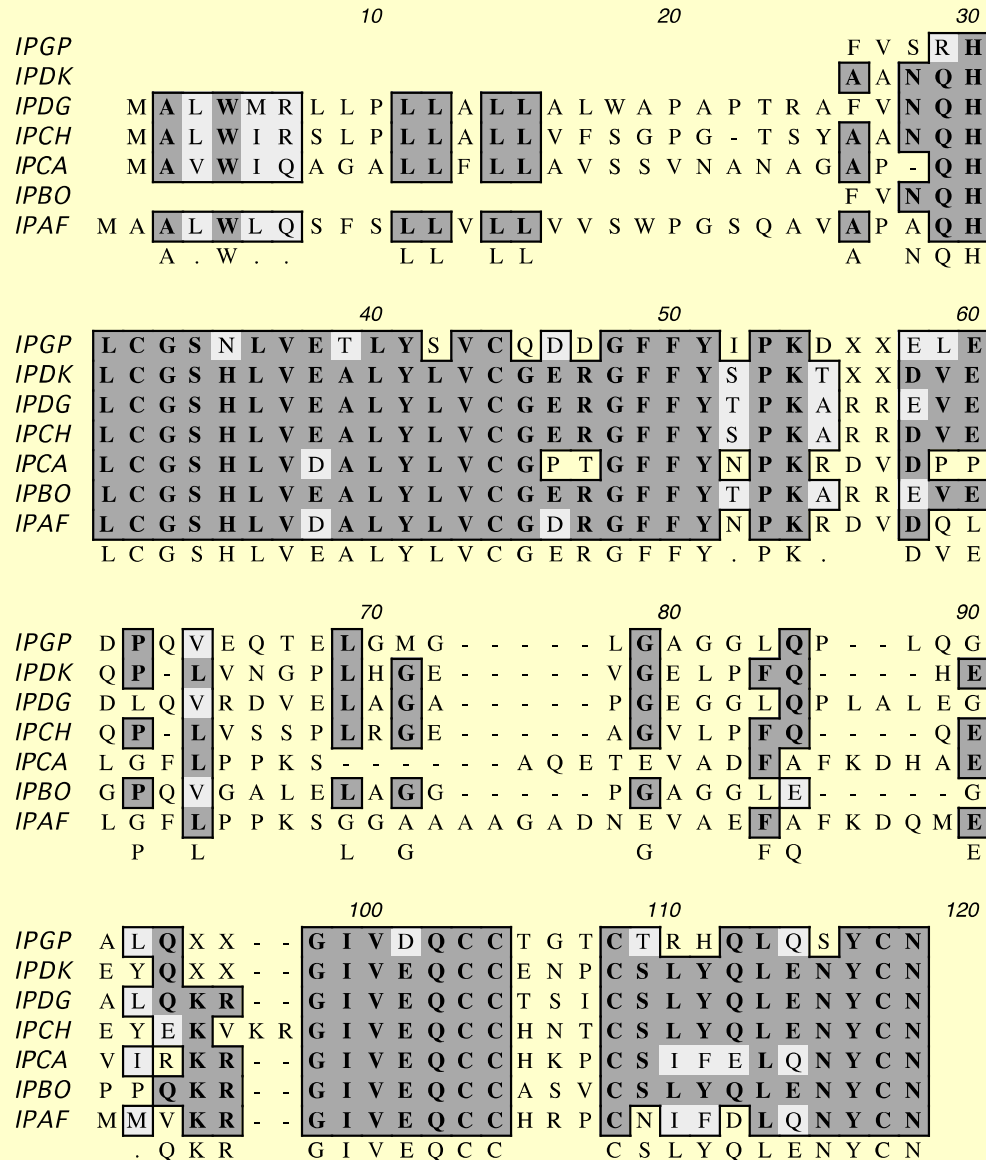
- Chose between local or global search algorithm
- Use most sensitive search algorithm available
 - Original BLAST for no gaps
 - Smith-Waterman for most flexibility
 - Gapped BLAST for well delimited regions
 - PSI-BLAST for families
 - Initially BLOSUM62 and default gap penalties
 - If no significant results, use BLOSUM30 and lower gap penalties
- Examine results between exp. 0.05 and 10 for biological significance
- Beware of long hits or those with unusual amino acid composition
- Reevaluate results of borderline significance using limited query

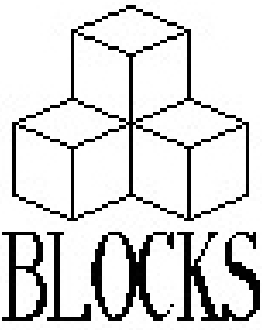
Goals of Multiple Sequence Alignment

- Determine Consensus Sequences
 - Prosite Patterns
- Building Gene Families
 - InterPro, Prints, ProDom, pFAM, DOMO, COGs, KOGs
- Develop Relationships & Phylogenies
 - Clusters, COGs, KOGs, ClusTR
 - Relationships
 - Evolutionary Models
 - UPGMA, Neighbor Joining, Phylip, GrowTree, PAUP
- Model Protein Structures for Threading and Fold Prediction
 - Profiles, Templates, HSSP, FSSP, SwissModel
 - Hidden Markov Models, pFAM, SAM, SuperFamily
 - Network Models, Neural Nets, Bayesian Networks
 - Statistical Models, Generalized Linear Models

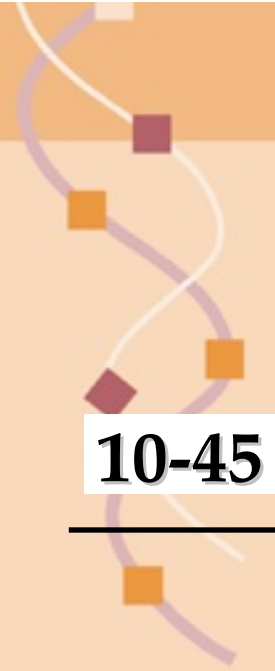
Consensus Sequence From a Multiple Sequence Alignment

ClustalW Insulin Alignments





http://blocks.fhcrc.org/blocks/blockmkr/make_blocks.html



10-45

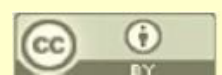
```
NLQGYMLGNP
NFMGYMVGNG
NLKGFLVGNA
NLKGILIGNA
NLKGFAIGNG
NFKGYLVGNG
NLKGFIVGNP
NIKGYIQGNA
NLKGFMIGNA
NLQGYILGNP
NFKGFVGNNA
NLQGYVLGNP
```

25-55

```
PLLLWLNCGGPGCSSIGYGASEEIG
PLVLWFNGGPGCSSVGFAGFEELG
PLMIWLTGGPGCSGLSSFVYEIGP
PLMIWLTGGPGCSGLSTFLYEFGP
PLLLWLSGGPGCSSLTGLLFENGP
PLVLWLNCGGPGCSSVAYGAAEEIG
PVVIWLTGGPGCSSELALFYENGP
PLVIWFNGGPGCSSLGGAFKELGP
PLVIWFNGGPACSSLGGAFLELGP
PLVLWLNCGGPGCSSLYGAFQELGP
PLVLWLNCGGPGCSSIAYGASEEVG
PLTLWLNCGGPGCSSVGGGAFTELG
```

40

```
TVKQWSGYMDYKDS
GVNQYSGYLSVGSN
SFAHYAGYVTVSED
DFAQYAGYVTVDAA
DLGHHAGYYKLPKS
SVESYSGFMTVDAK
GVKSYTGYPYLLANAT
NFKQYSGYINVTGK
NFKSYSGYVDANAN
NFKHYSGFFQVSDN
DFFHYSGYLRAWTD
TVKQYTGYPYLDVEDD
```



Programs

Managers

[Help Topics](#) | [Support](#)

Programs

Comparison

Database Searching

Similarity

Reference

Evolution

Mapping

Pattern Recognition

Primer Selection

Protein Analysis

Nucleic Acid Secondary
Structure

Translation

Utilities

Index

HmmerPfam

?

Identify known hidden Markov model domains.

Input sequences:

Select From:

Sequence	Description	Type	Length	Range
odpat_human.uniprot_sprot	odpat_human	P	388	1..388
odpa_human.uniprot_sprot	odpa_human	P	390	1..390
odpt_ascsu.uniprot_sprot	odpt_ascsu	P	391	1..391
odpa_yeast.uniprot_sprot	odpa_yeast	P	420	1..420
odpa_rat.uniprot_sprot	odpa_rat	P	390	1..390
odpa_mouse.uniprot_sprot	odpa_mouse	P	390	1..390
odpa_caeel.uniprot_sprot	odpa_caeel	P	397	1..397
odpat_rat.uniprot_sprot	odpat_rat	P	391	1..391
odpa_macfa.uniprot_sprot	odpa_macfa	P	390	1..390
odpa_smima.uniprot_sprot	odpa_smima	P	363	1..363
odpa_ponpy.uniprot_sprot	odpa_ponpy	P	390	1..390
odpa_ascsu.uniprot_sprot	odpa_ascsu	P	396	1..396
odpa_pig.uniprot_sprot	odpa_pig	P	389	1..389
odpat_mouse.uniprot_sprot	odpat_mouse	P	391	1..391

Input Parameters:

HMM Database

Number of processors to use

Turn the XNU filtering of target protein sequences on

Limit alignment output of best scoring domains to

 (range 1 thru 1000)

Set E-value cutoff for the per-sequence ranked hit list

 (range 0.0 thru 100.0)

Set bit score cutoff for the per-sequence ranked hit list

 (range -500.0 thru 500.0)

Set E-value cutoff for the per-domain ranked hit list

 (range 0.0 thru 100.0)

Set bit score cutoff for the per-domain ranked hit list

 (range -1000.0 thru 1000.0)

SeqWeb Sequence Profile Search

<http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=hammerpfam>

ProfileSearch



Search a peptide database using a Profile constructed from unaligned peptide sequences.

Input sequences:

Select From:

Sequence	Description	Type	Length	Range
hba_horse.pep	ID HBA_HORSE STANDARD; PRT; 141 AA.	P	141	1 .. 141
lgba_soybn	lgba_soybn	P	143	1 .. 143
hbahuman	hbahuman	P	141	1 .. 141
glb5_petma.pep	ID GLB5_PETMA STANDARD; PRT; 149 AA.	P	149	1 .. 149
lgb1_soybn.pep	- ID LGB1_SOYBN STANDARD; PRT; 143 AA.	P	143	1 .. 143
hba_human	hba_human	P	141	1 .. 141
lgb2_luplu.pep	ID LGB2_LUPLU STANDARD; PRT; 153 AA.	P	153	1 .. 153
hbb_human	hbb_human	P	146	1 .. 146

Input Parameters:

Search set

Limit for the number of alignments (range 0 thru 100000)

The following matrix and gap weight settings apply to the alignment phase of the analysis.

Select a sequence comparison matrix. This matrix determines how matches and mismatches are scored. The default penalties for gap creation and extension are given after each matrix name.

Scoring Matrix

Set gap creation penalty

Set gap extension penalty

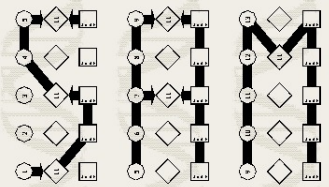
The following gap weight settings apply to the search phase of the analysis.

Set maximum gap creation penalty

Set maximum gap extension penalty

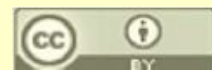
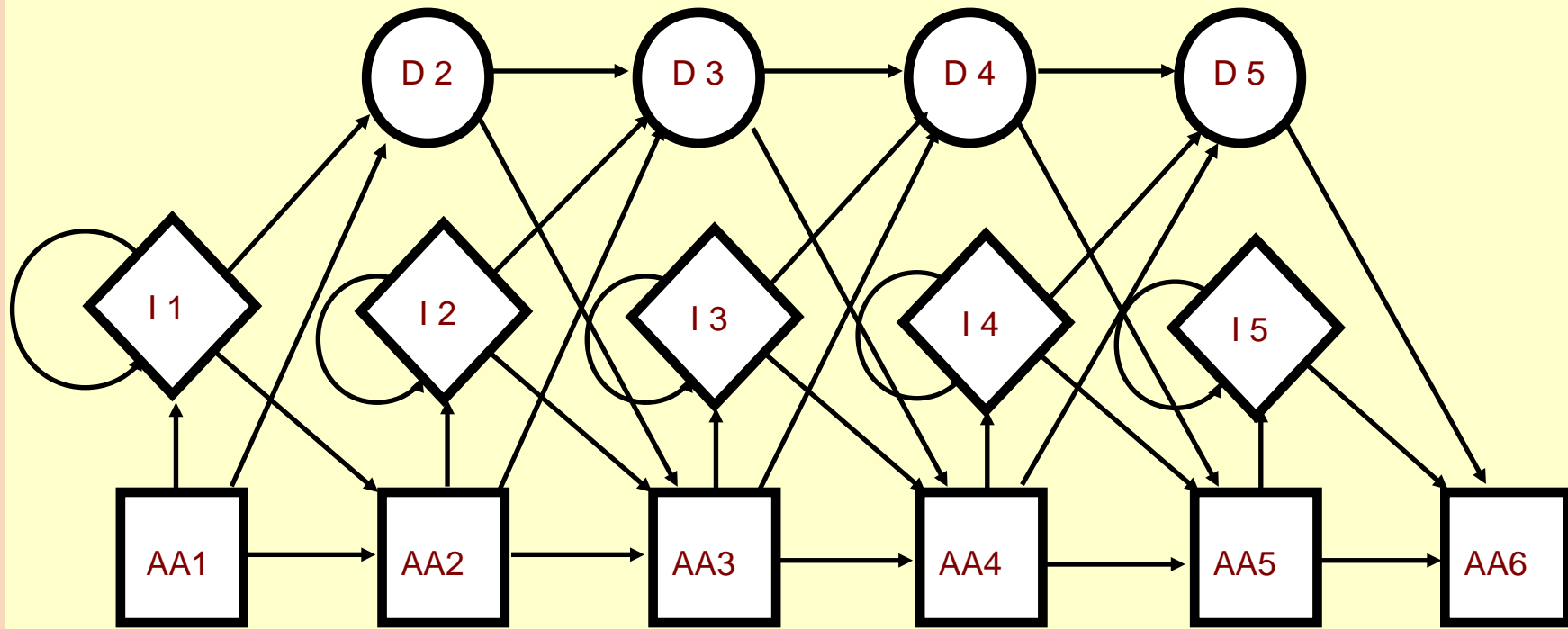
A Z score is the number of standard deviations above background.

Lowest Z score to report in output list (range 0.0 thru 100)

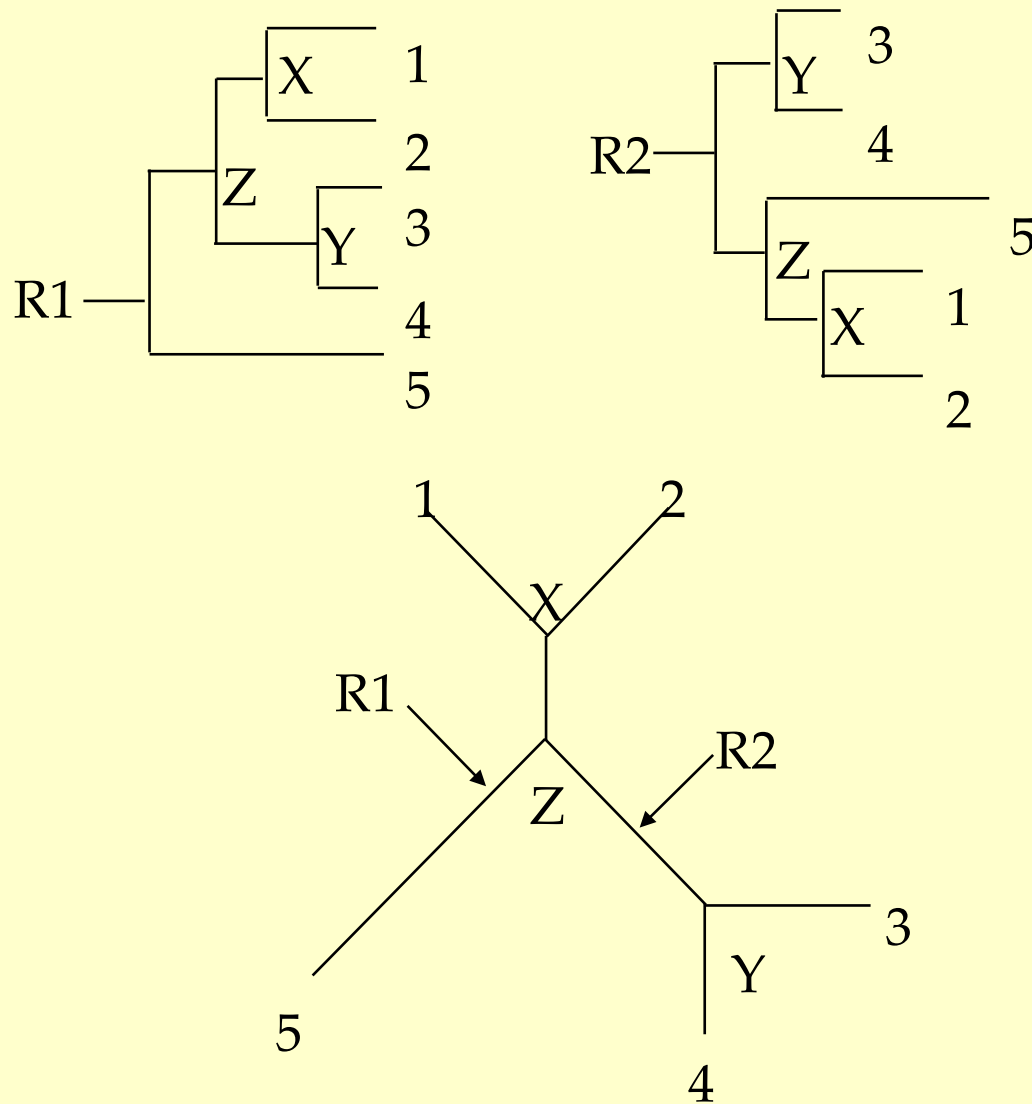


Hidden Markov Models

<http://www.cse.ucsc.edu/research/compbio/sam.html>



Evolutionary Trees

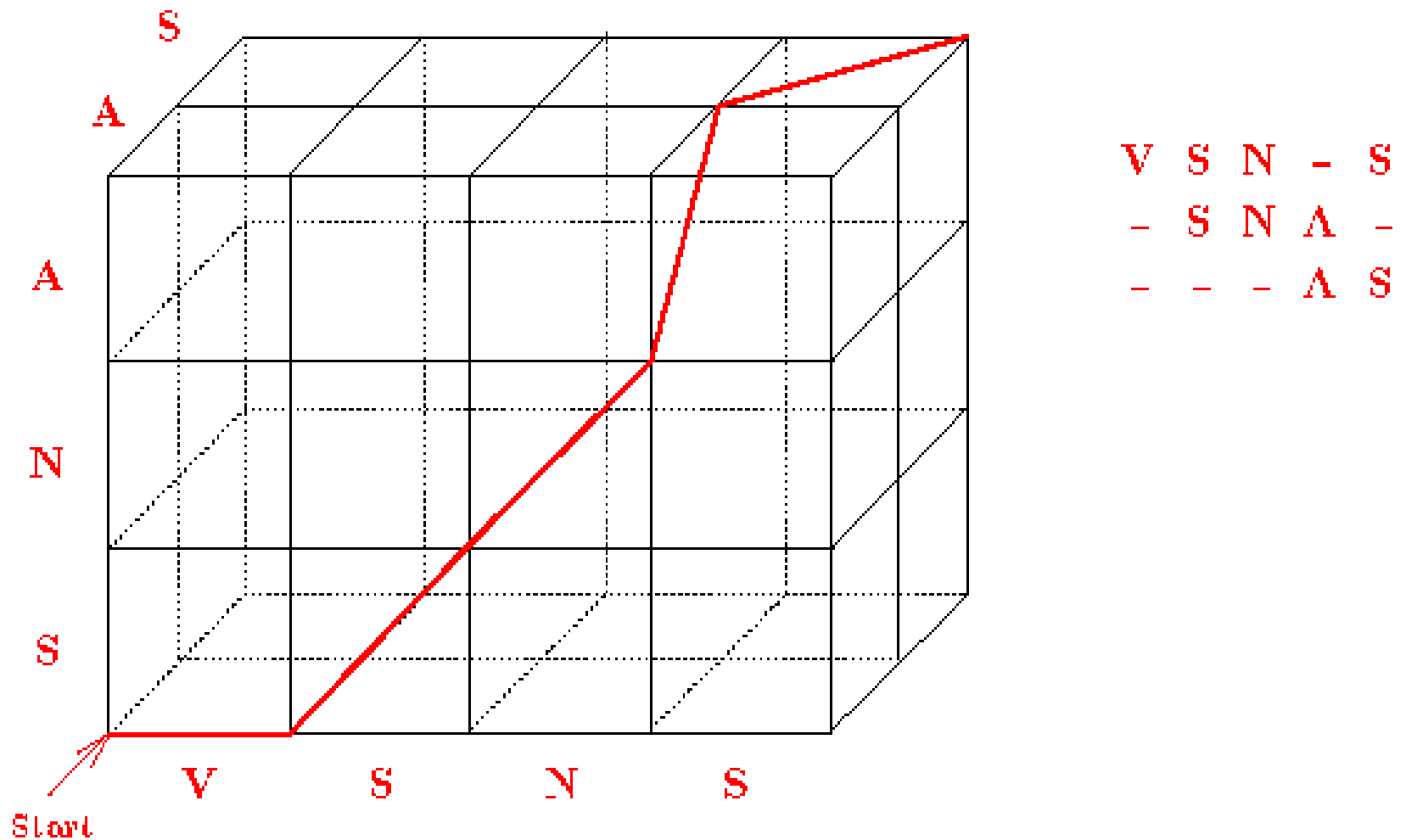


Challenges Aligning Multiple Sequences

- Computational complexity $O(n^k)$ for k sequences n long
- Space requirements $O(n^k)$ for k sequences n long
- Sequence clusters require weighting function
- Weighted alignments tend to overweight erroneous sequences
- Approximations must be used for real world data
 - Linked lists used to find exact words shared between k sequences
 - BLAST can find inexact shared words between k sequences
 - FASTA can be used to do progressive pair-wise alignments
 - HMM Pair models find best overall alignment probabilistically
- Pairwise comparisons followed by Progressive Alignments
- Final alignment is often dependent on order data presented
- Gaps make alignment unnaturally long

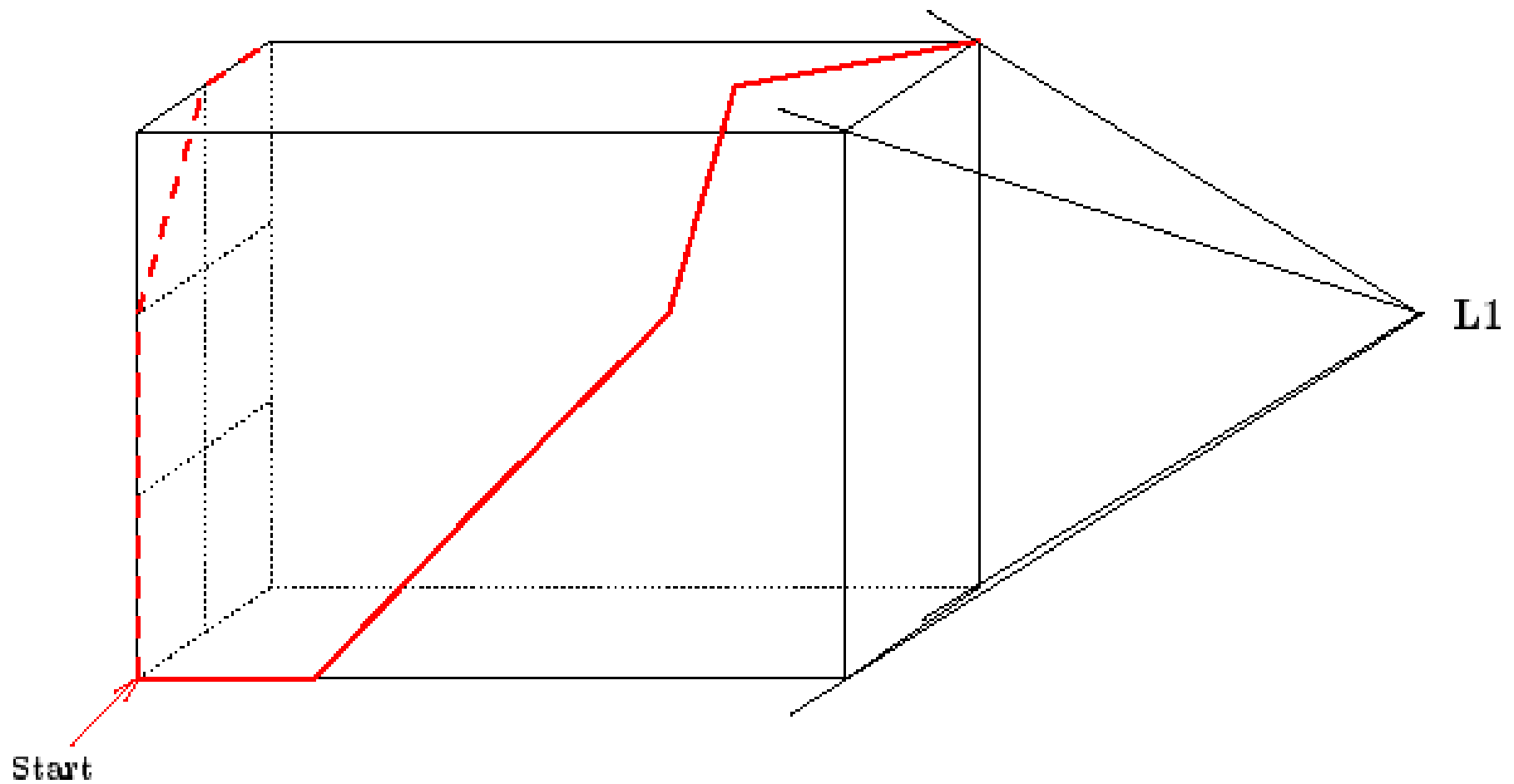
Three Protein Alignment (Murata, Richardson & Sussman)

Figure 1: Alignment Path for 3 Sequences.



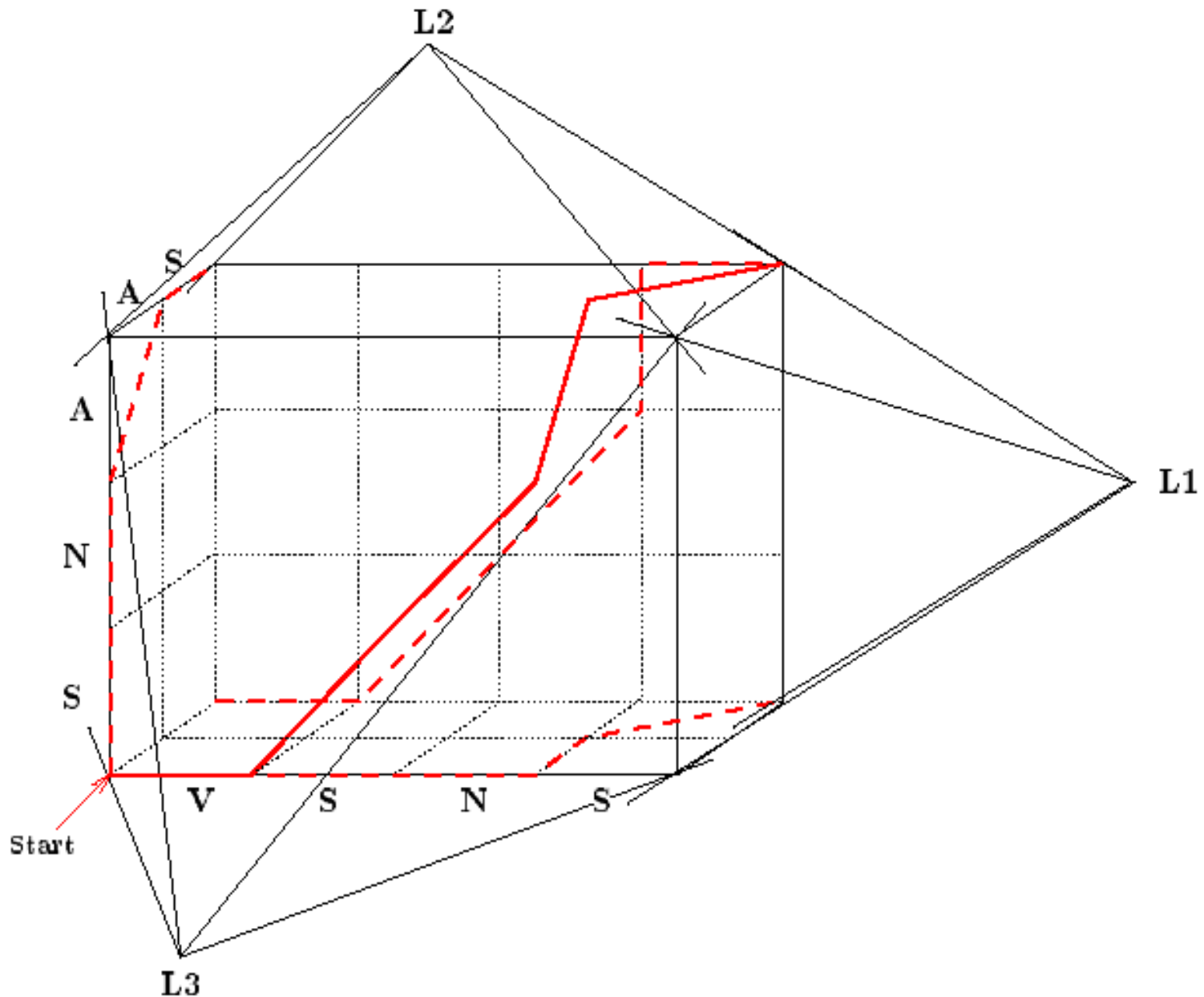
One Pairwise Alignment from the Three-Way Alignment

Figure 2: Projection of the Alignment from the Right-Hand-Side.



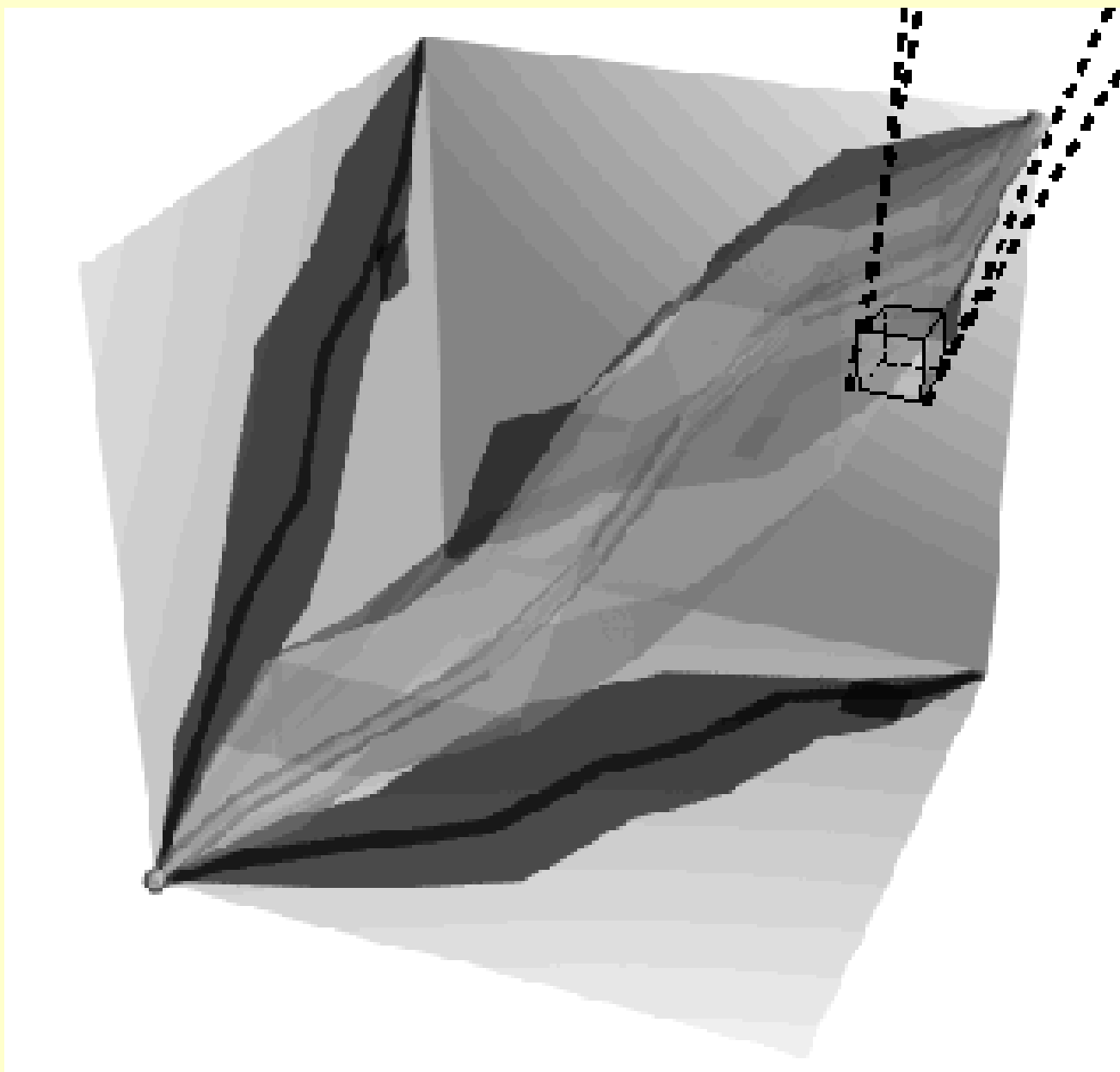
All Pairwise Alignments from the Three-Way Alignment

Figure 3: All 3 Pairwise Projections of the Alignment



Carrillo-Lipman Limits for MSA

<http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html>

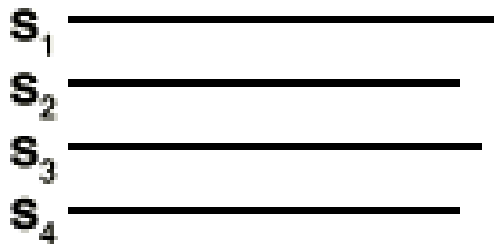


Clustal Progressive Alignment (Step 1)

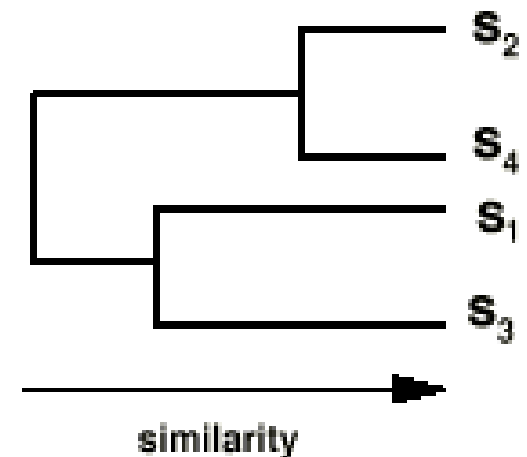
Steps in Multiple Alignment

(A) Pairwise Alignment

Example - 4 sequences S_1 S_2 S_3 S_4

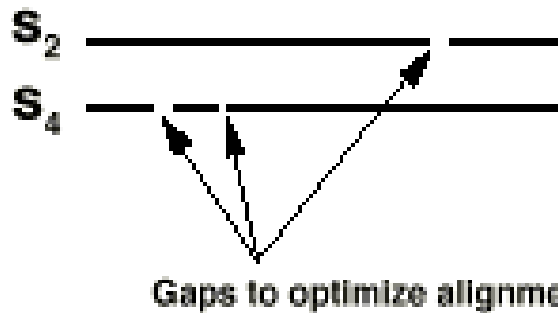


6 pairwise comparisons
then cluster analysis

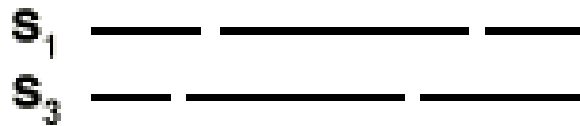


Clustal Progressive Alignment (Step 2)

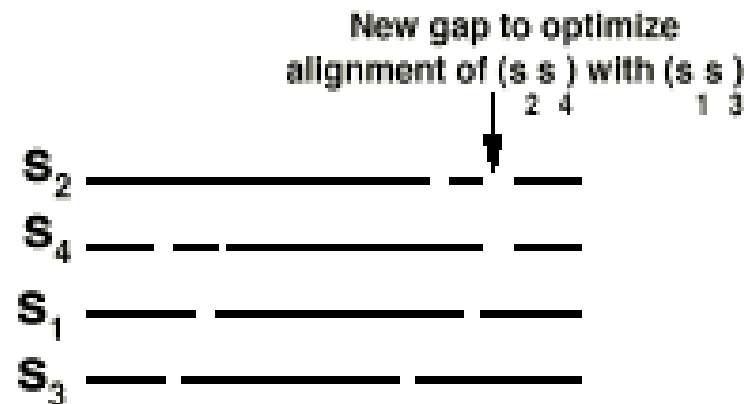
(B) Multiple alignment following the tree from A



align most similar pair



align next most similar pair



align alignments - preserve gaps

Gaps Are Propagated To Make Alignment

SEQ2 : MQQL-DNPYIVRMIGICEAE-SWM

SEQ1 : MGQF-DHPNIIRLEGVVTKSRPVM

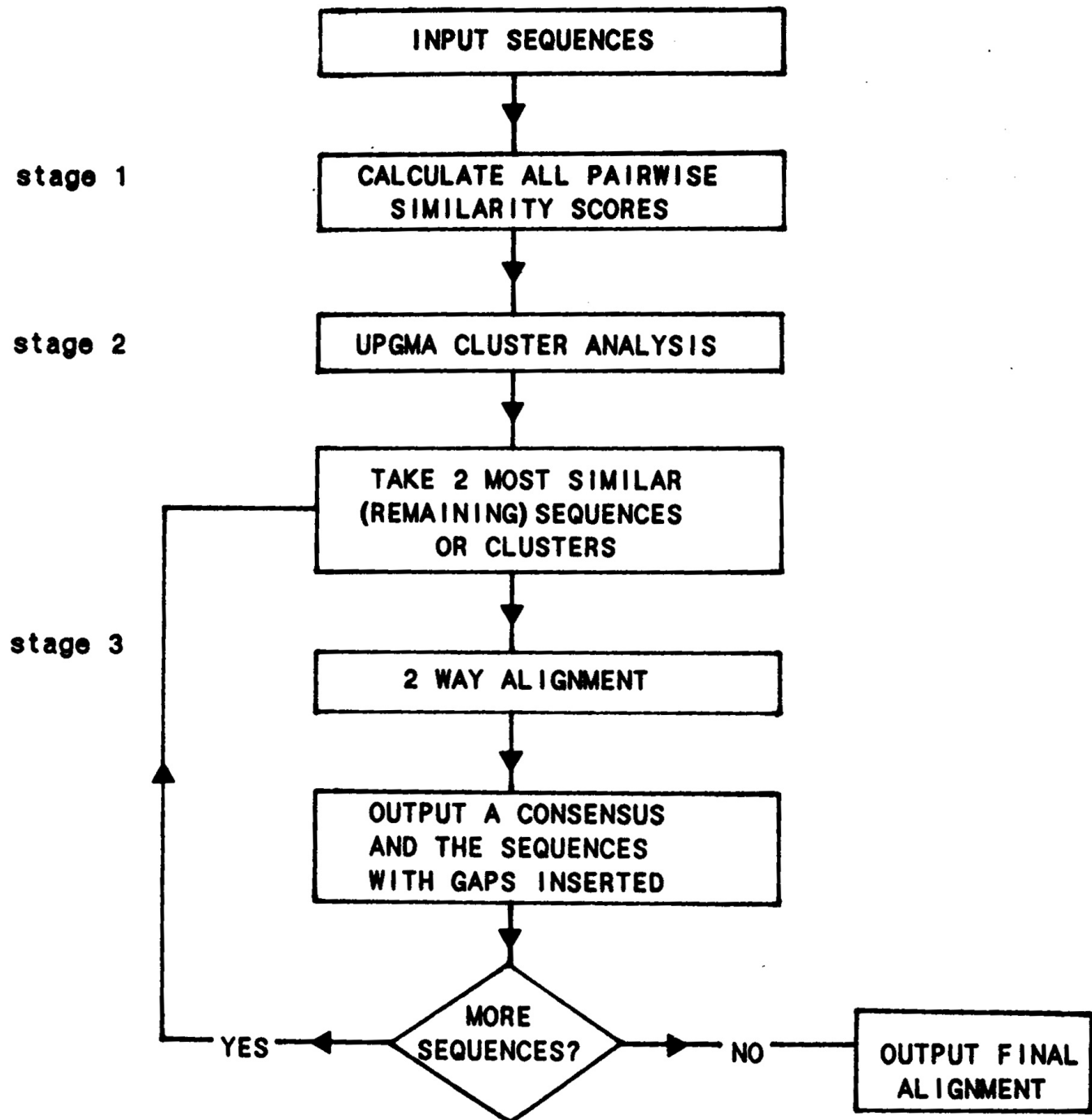
SEQ2 : MQQL-DNPYIVRMIGICEAE-SWM

SEQ3 : MKQL-QHPRLVRLYAVVTQE-PIY

SEQ2 : MQQL-DNPYIVRMIGICEAE-SWM

SEQ4 : MKMIGKHKNIINLLGACTQDGPLY

Clustal Procedure



Clustal Dendrogram

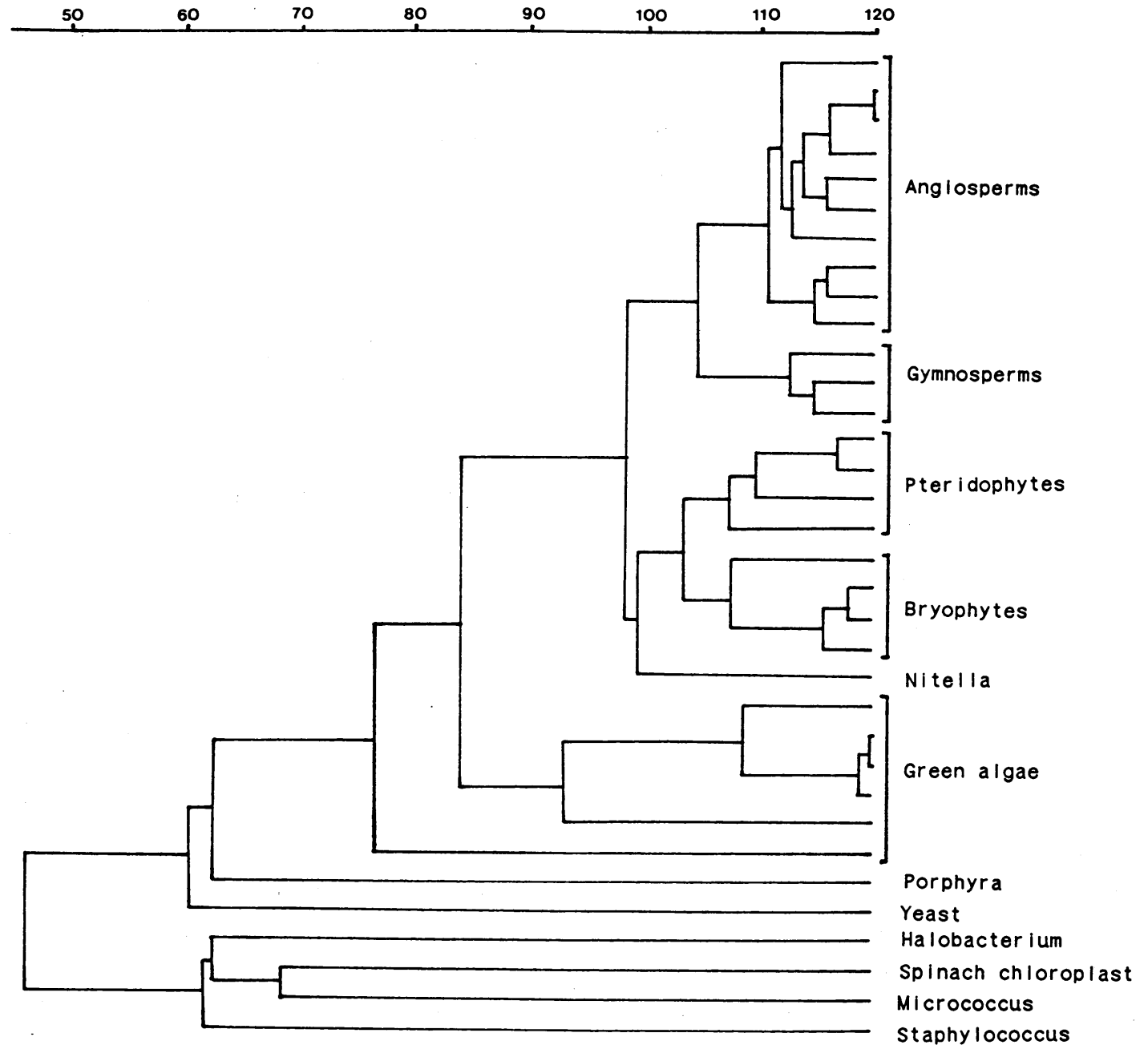


Fig. 4. UPGMA dendrogram of 34 plant, yeast and bacterial 5S RNA sequences. The sequences were taken from Hori et al. (1985) (see RESULTS, section a2). The major plant taxonomic groupings are indicated. The scale across the top margin shows the number of matching nucleotides (after alignment) between two clusters or sequences.

Clustal Globin Alignment

- 1 human beta globin
- 2 horse beta globin
- 3 human alpha globin
- 4 horse alpha globin
- 5 cyanohaemoglobin
- 6 whale myoglobin
- 7 leghaemoglobin

```

          A          B          C
VHLTPEEKSAVTALWGKVNVND  EVGGEALGRLLVYPWTQRFESFGDLSTPDAVMGNPK
VQLSAGEEKAAVLALWDKVNNEE  EVGGEALGRLLVYPWTQRFDSFGDLSNPGAVMGNPK
VLSPADKTNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHF  DLSH      GSAQ
VLSAADKTNVKAAWSKVGGHAGEYGAELERMFLGFPTTKTYFPHF  DLSH      GSAQ
PIVDTGTSVAPLSAAEKTIRSAWAPVYSYDYESGVDILVKFFTSTFAAEFFPKFKGLTTADELKKKSAD
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDKFKHLKTEAEMKASED
GALTESQAALVKSSWEETFNANIPKHTHRFFILVLEIAPAAKDLFSSFLKGGTSEVPQNNPE
          *      .      .      *          .      .      *      *      *
    
```

```

          E          F          G          H
VKAHGKKVLGAFSDG  LAHLDNLRKGTFFAT  LSELHCDKLHVDPENFRLLGNVLCVLAHHEFGKEFTTPVQAAAYQKVVAGVANALAHKYH
VKAHGKKVLHSEFGE  VHLDNLRKGTFAA  LSELHCDKLHVDPENFRLLGNVLVVRLARHEFGKDFTPPELQASVYQKVVAGVANALAHKYH
VKGHGKKVADALTNA  VAHVDDMPNALS  LSDLHAHAKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR
VKAHGKKVGDALTNA  VGHLDDLPGALS  LSDLHAHAKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTISKYR
VRWHAERIIDAVIDA  VASMDDEKMSMMDLSGKHAKEFEVDPEYFKVLA AVIADTVAAGD  AGFEKLLRMICILLRSAY
LKKHGVTVLTALGAI  LKKKGHEAELKP  LAQSHATKHKIPKYLEFISEAIIHVLHSRHPGDFGADAQCAMNKALELFRKDIAAKYKELGYQG
LQAHAGKVFKLVEAA  IQLVETGVVAS  DATLKNLGSVHVS  KGVVA  DAHFPVVK  EAILKTIKEV  VGAKWSEELNS  AWTIAYDELAIVIKKEMDDAA
          .      *      .      .      *      *          .      .      .      .      .      .      .
    
```

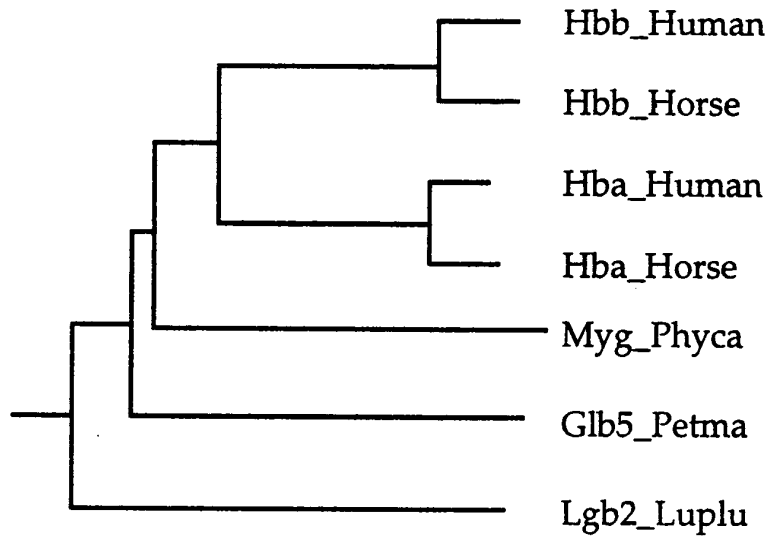
Fig. 3. CLUSTAL-produced multiple alignment of seven globin sequences taken from Lesk and Chothia (1980) (see RESULTS, section 4)

ClustalW Step 1: BLOSUM Distance Matrix

Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

**Pairwise alignment:
Calculate distance matrix**

ClustalW Step 2: Dendrogram



↓

**Rooted Neighbor Joining
tree (guide tree)**

↓

ClustalW Sequence Weighting

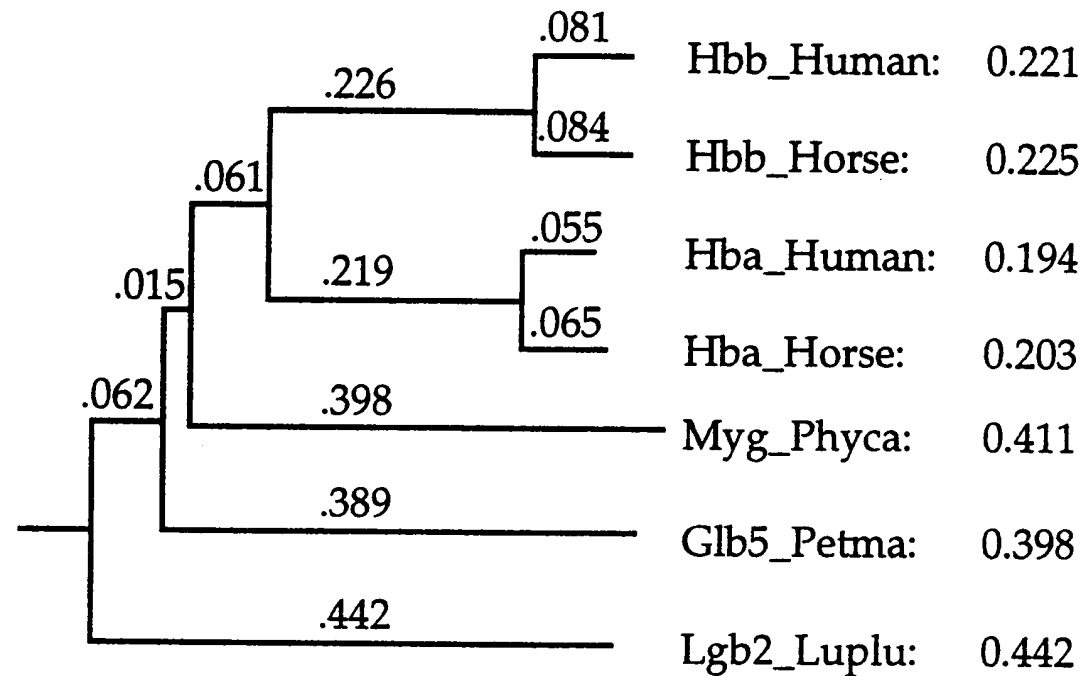


FIG. 3. Sequence weights for the seven globin sequences from Fig. 1. A rooted neighbor-joining tree is shown with branch lengths. The weights are shown for each sequence before normalization (the weights are normalized so as to make the largest equal to 1.0).

ClustalW Residue Specific Gap Penalties

RESIDUE-SPECIFIC GAP OPENING PENALTY FACTORS*

Residue	Penalty	Residue	Penalty
A	1.13	M	1.29
C	1.13	N	0.63
D	0.96	P	0.74
E	1.31	Q	1.07
F	1.20	R	0.72
G	0.61	S	0.76
H	1.00	T	0.89
I	1.32	V	1.25
K	0.96	Y	1.00
L	1.21	W	1.23

Position Specific Gap Penalties

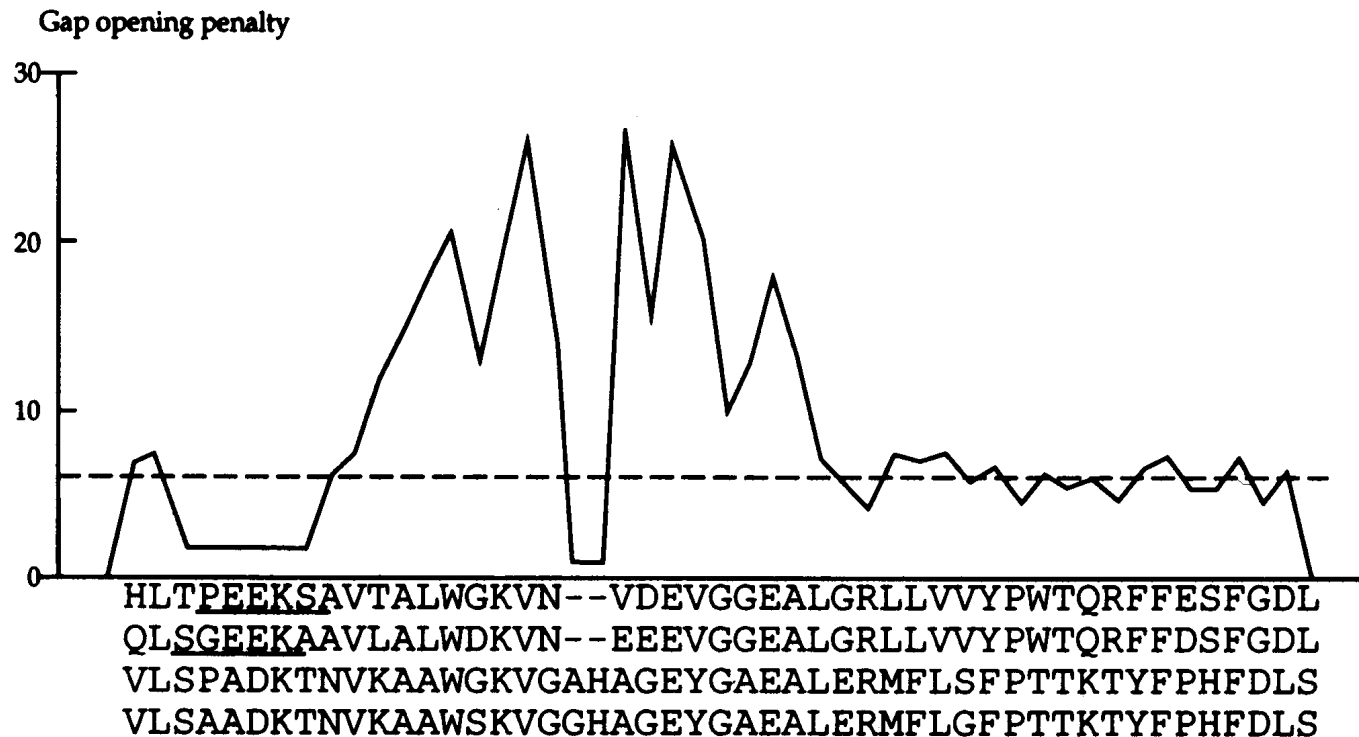


Figure 3. The variation in local gap opening penalty is plotted for a section of alignment. The initial gap opening penalty is indicated by a dotted line. Two hydrophilic stretches are underlined. The lowest penalties correspond to the ends of the alignment, the hydrophilic stretches and the two positions with gaps. The highest values are within 8 residues of the two gap positions. The rest of the variation is caused by the residue specific gap penalties (12).

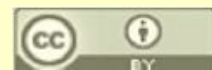
ClustalW Step 3: Progressive Alignment

```
-----VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
-----VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
-----VLSPADKTNVKAANGKVGAGHAGEYGAEALERMF LSFPTTKTYFPHFDLS--
-----VLSAADKTNVKAAWSKVGGHAGEYGAEALERMF LGSFPTTKTYFPHFDLS--
-----VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFD RFKHLKT
PIVDTG SVAPLSAAEKT KIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
-----GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
      * . * * * . *
```

Progressive alignment:
Align following the guide tree

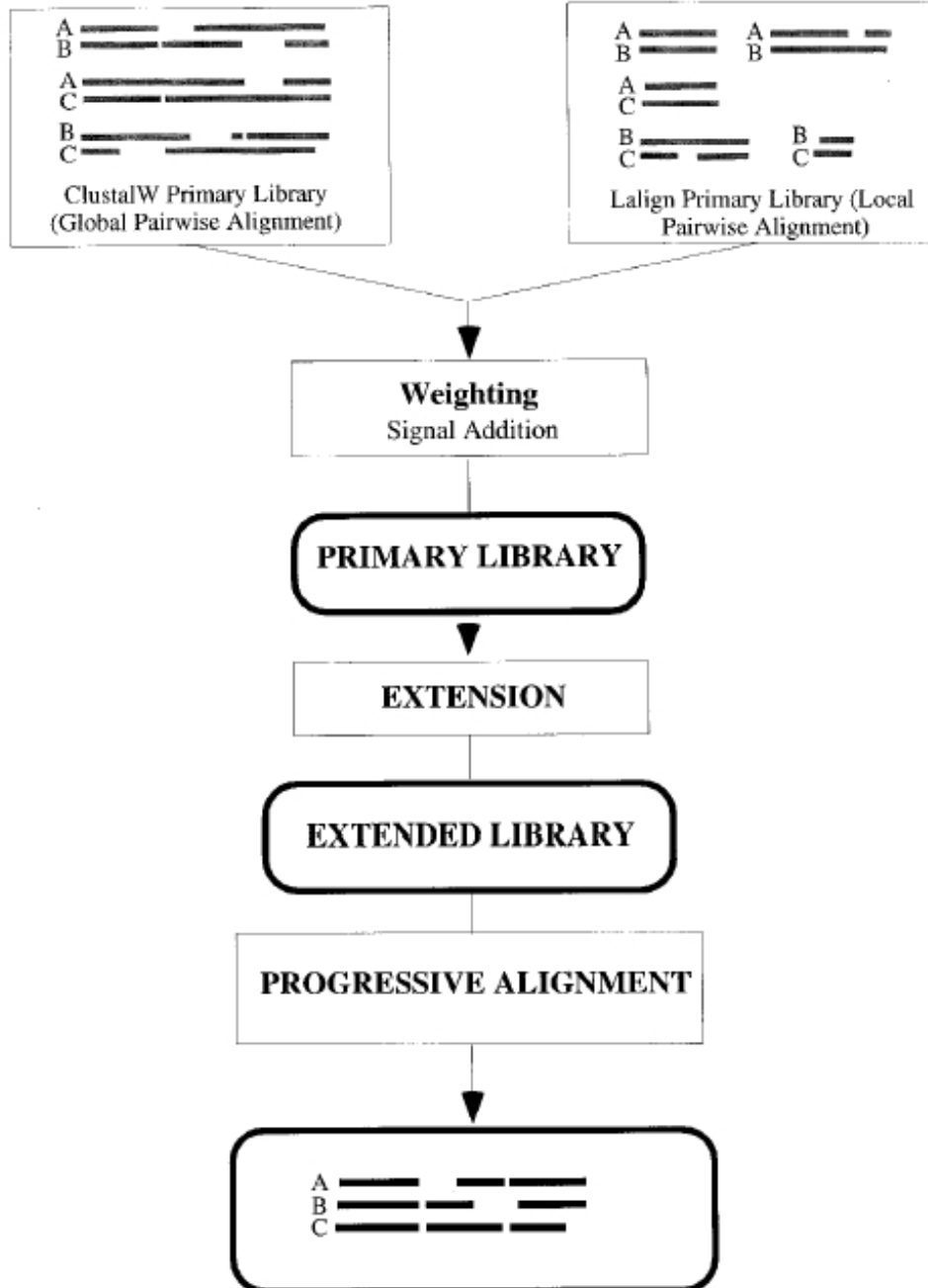
```
PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGT FATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVDELDTL-----DMPNALSALSDLHAHKL RVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLDELDTL-----DLPGALS NLSDLHAHKL RVDPVNFKL
EAEMKASEDLKKHGVTVLTAALGAILKKK-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDL SGGKHA KSFQVDPQYFKV
VP--QNNPELOAHAGKVKLVYEAAIQIQVLTGVVVTDAITLKNLGSVH VSKG-VADAHFPV
.. * . * *
```

```
LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-----
LGNVLVVVLARHFGKDFTPPELQASYQKVVAGVANALAHKYH-----
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR-----
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTISKYR-----
ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LAAVIADTVAAG-----DAGFEKLMSMICILLRSAY-----
VKEAAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```



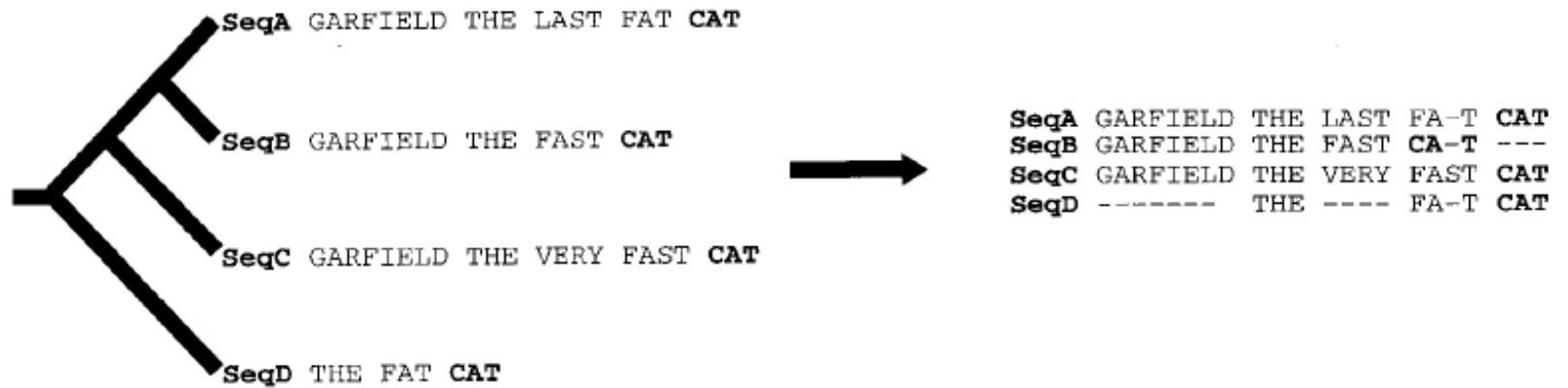
T-Coffee Procedure

T-Coffee: a Method for Sequence Alignment



Regular Progressive Alignment

a) Regular Progressive Alignment Strategy



T-Coffee Primary Alignment Library

b) Primary Library

SeqA	GARFIELD	THE	LAST	FAT	CAT	Prim. Weight = 88	SeqB	GARFIELD	THE	----	FAST	CAT	Prim Weight = 100
SeqB	GARFIELD	THE	FAST	CAT	---		SeqC	GARFIELD	THE	VERY	PAST	CAT	
SeqA	GARFIELD	THE	LAST	FA-T	CAT	Prim. Weight = 77	SeqB	GARFIELD	THE	FAST	CAT	Prim. Weight = 100	
SeqC	GARFIELD	THE	VERY	FAST	CAT		SeqD	-----	THE	FA-T	CAT		
SeqA	GARFIELD	THE	LAST	FAT	CAT	Prim. Weight = 100	SeqC	GARFIELD	THE	VERY	FAST	CAT	Prim. Weight = 100
SeqD	-----	THE	----	FAT	CAT		SeqD	-----	THE	----	FA-T	CAT	

T-Coffee Extended Alignment Library and Progressive Alignment

b) Primary Library

SeqA GARFIELD THE LAST FAT CAT Prim. Weight = 88
 SeqB GARFIELD THE FAST CAT ---

SeqB GARFIELD THE ---- FAST CAT Prim Weight = 100
 SeqC GARFIELD THE VERY FAST CAT

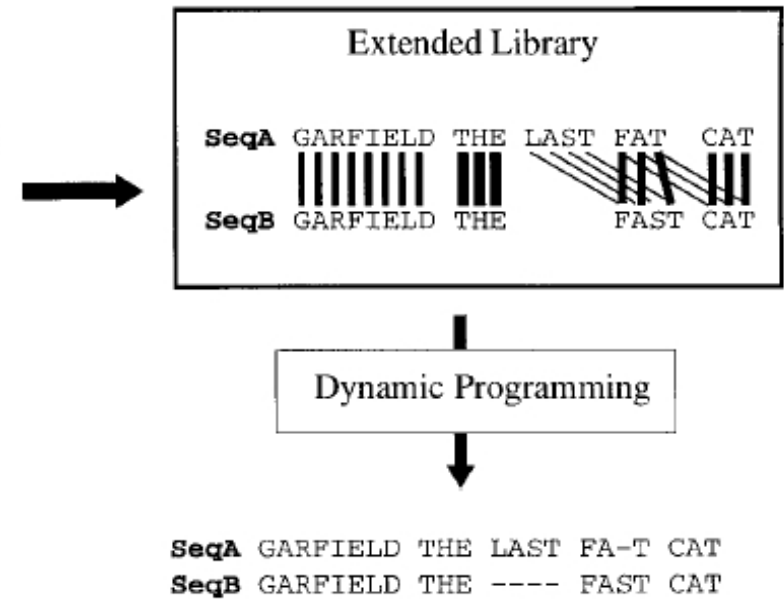
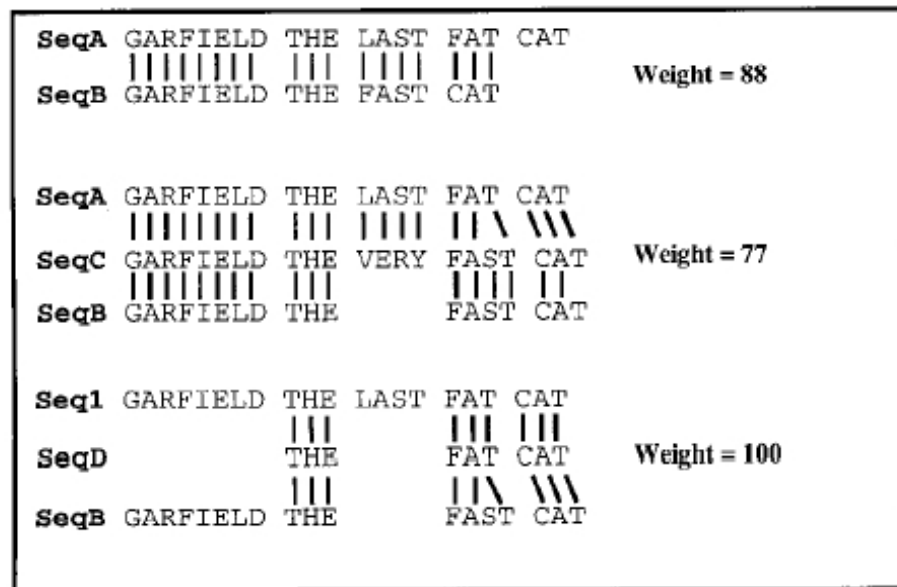
SeqA GARFIELD THE LAST FA-T CAT Prim. Weight = 77
 SeqC GARFIELD THE VERY FAST CAT

SeqB GARFIELD THE FAST CAT Prim. Weight = 100
 SeqD ----- THE FA-T CAT

SeqA GARFIELD THE LAST FAT CAT Prim. Weight = 100
 SeqD ----- THE ---- FAT CAT

SeqC GARFIELD THE VERY FAST CAT Prim. Weight = 100
 SeqD ----- THE ---- FA-T CAT

c) Extended Library for seq1 and seq2



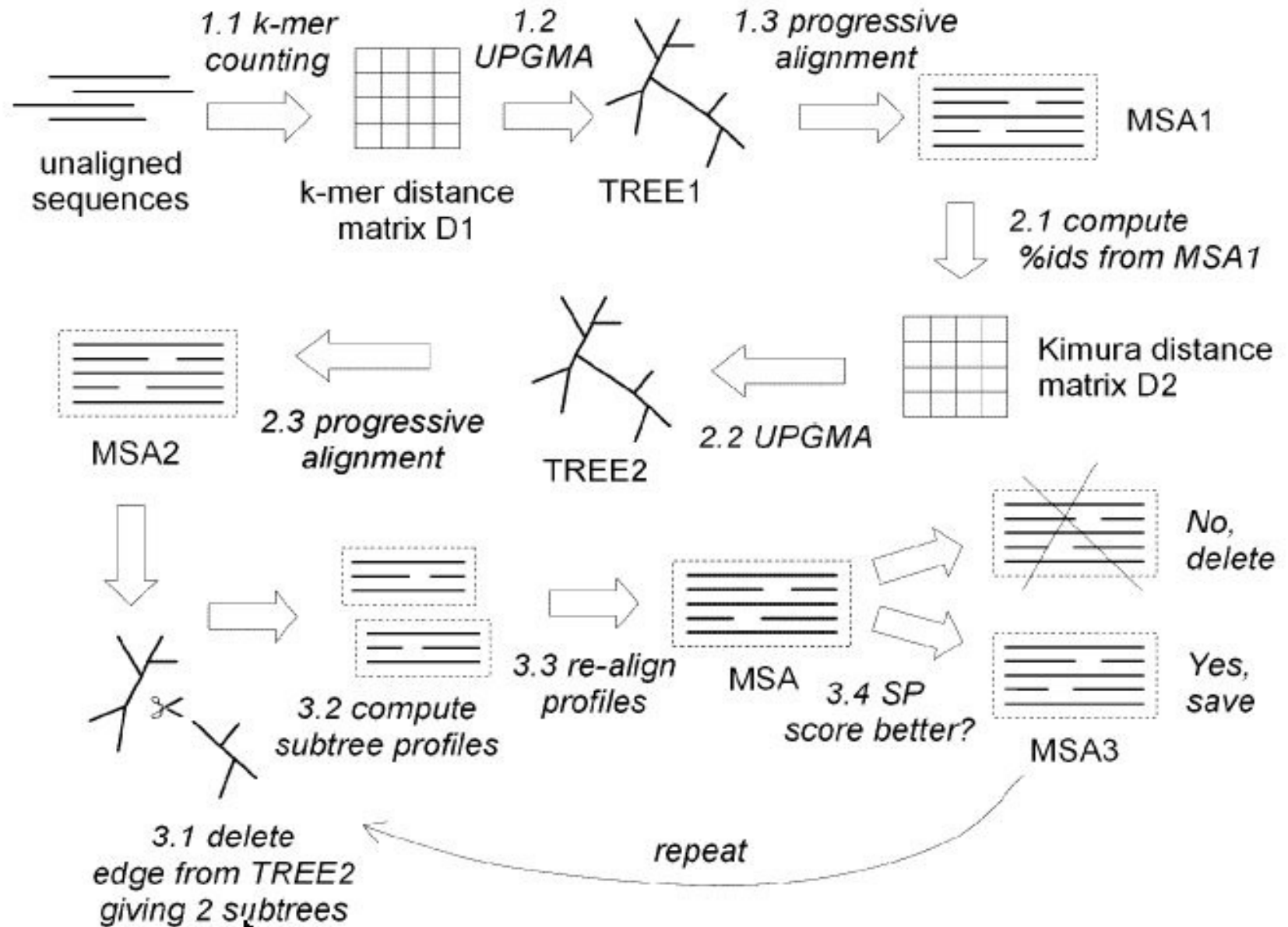
Comparison of T-Coffee to Other MSAs

Table 2. T-Coffee compared with other multiple sequence alignment methods

Method	Cat1 (81)	Cat2 (23)	Cat3 (4)	Cat4 (12)	Cat5 (11)	Total1 (141)
Dialign	71.0	25.2	35.1	74.7	80.4	61.5
ClustalW	78.5	32.2	42.5	65.7	74.3	66.4
Prrp	78.6	32.5	50.2	51.1	82.7	66.4
T-Coffee	<u>80.7</u>	<u>37.3</u>	<u>52.9</u>	<u>83.2</u>	<u>88.7</u>	<u>72.1</u>

MUSCLE

Edgar (2004) NAR 32, 1792-1797



MUSCLE

Edgar (2004) NAR 32, 1792-1797

```
YES_XIPHE  MGCvrSKEaKgPAIKYqpdNsnvvPvSahlgHYGpeptimg
YES_AVISY   -----dKgPAmKYrtdNtpePiSshvsHYGsd
YES_CHICK   -----MGCikSKEdKgPAmKYrtdNtpePiSshvsHYGsd
YES_HUMAN   -----MGCikSKENKsPAiKYrpeNtpePvStsvsHYGae
YES_MOUSE   -----MGCikSKENKsPAiKYtpeNlteP--vSpsasHYG

YES_XIPHE  MGCvrSKEaKgPAIKYqpdNsnvvPvSahlgHYGpeptimg
YES_AVISY   -----dKgPAmKYrtdNtp-ePiSshvsHYGsdssqat
YES_CHICK   MGCikSKEdKgPAmKYrtdNtp-ePiSshvsHYGsdssqat
YES_HUMAN   MGCikSKENKsPAiKYrpeNtp-ePvStsvsHYGaepttvs
YES_MOUSE   MGCikSKENKsPAiKYtpeNlt-ePvSpsasHYGvehatva
```



MUSCLE BaliBase Test

Edgar (2004) NAR 32, 1792-1797

Table 2. BALiBASE Q scores on subsets

Method	Ref1	Ref2	Ref3	Ref4	Ref5
MUSCLE	0.887	0.935	0.823	0.876	0.968
MUSCLE-p	0.871	0.928	0.813	0.857	0.974
T-Coffee	0.866	0.934	0.787	0.917	0.957
NWNSI	0.867	0.923	0.787	0.904	0.963
CLUSTALW	0.861	0.932	0.751	0.823	0.859
FFTNSI	0.838	0.908	0.708	0.793	0.947

ProbCons

Do, et al. Genome Research 15, 330-340

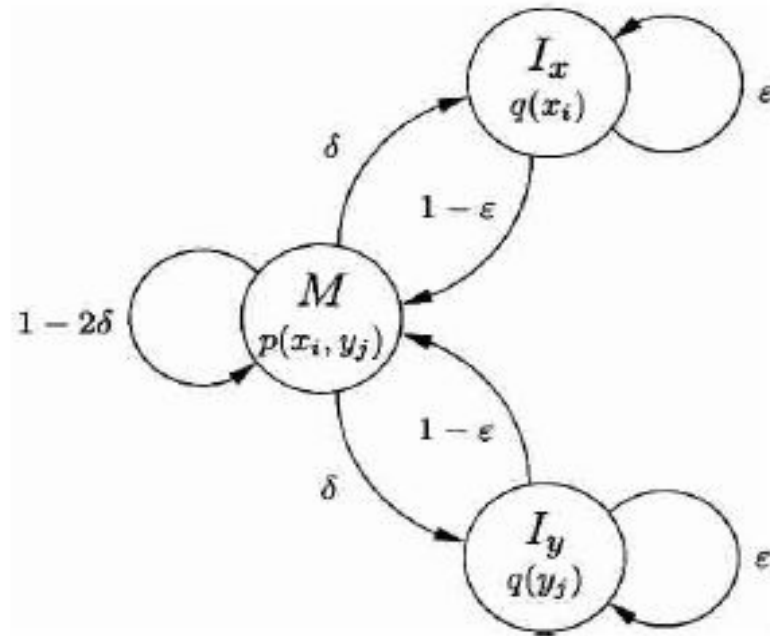


Figure 1. Basic pair-HMM for sequence alignment between two sequences, x and y . State M emits two letters, one from each sequence, and corresponds to the two letters being aligned together. State I_x emits a letter in sequence x that is aligned to a gap, and similarly state I_y emits a letter in sequence y that is aligned to a gap. Finding the most likely alignment according to this model by using the Viterbi algorithm corresponds to applying Needleman–Wunsch with appropriate parameters. The logarithm of the emission probability function $p(.,.)$ at M corresponds to a substitution scoring matrix, while affine gap penalty parameters can be derived from the transition probabilities δ and ϵ (Durbin et al. 1998).

ProbCons

Do, et al. *Genome Research* 15, 330-340

- Step 1: Compute posterior probability matrices of each pair of aligned sequences from the pair-HMM model
- Step 2: Compute expected accuracies of pairwise alignments.
- Step 3: Probabilistic Consistency Transformation
- Step 4: Calculate Guide Tree using UPGMA from measure of similarities of sequence pairs
- Step 5: Progressive alignment
- Step 6: refinement by dividing sequences into two groups and re-align. Repeat multiple times.

ProbCons

Do, et al. Genome Research 15, 330-340

PROBCONS

Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences

RUN

ABOUT

DOWNLOAD

HELP

PROBCONS is an efficient protein multiple sequence alignment program, which has demonstrated a statistically significant improvement in accuracy compared to several leading alignment tools.

BASIC PARAMETERS

E-mail address

E-mail address (again)

Input sequence file

ADDITIONAL OPTIONS

Consistency reps

Iterative refinement reps

Pre-training reps

Output format MFA CLUSTALW

COMPUTE ALIGNMENT

Comments to Chuong Do (chuongdo@cs.stanford.edu)



Table 1. Performance of aligners on the BALiBASE benchmark alignments database

Aligner	Ref 1 (82)		Ref 2 (23)		Ref 3 (12)		Ref 4 (12)		Ref 5 (12)		Overall (141)		Time (mm:ss)
	SP	CS	SP	CS	SP	CS	SP	CS	SP	CS	SP	CS	
Align-m	76.6	n/a	88.4	n/a	68.4	n/a	91.1	n/a	91.7	n/a	80.4	n/a	19:25
DIALIGN	81.1	70.9	89.3	35.9	68.4	34.4	89.7	76.2	94.0	84.3	83.2	63.7	2:53
CLUSTALW	86.1	77.3	93.2	56.8	75.3	46.0	83.4	52.2	85.9	63.8	86.1	68.0	1:07
MAFFT	86.7	78.1	92.4	50.2	78.8	50.4	91.6	72.7	96.3	85.9	88.2	71.4	1:18
T-Coffee	86.6	77.4	93.4	56.1	78.5	48.7	91.8	73.0	95.8	90.3	88.3	72.2	21:31
MUSCLE	88.7	80.8	93.5	56.3	82.5	56.4	87.6	60.9	96.8	90.2	89.6	73.9	1:05
ProbCons	90.1	82.6	94.4	61.3	84.1	61.3	90.1	72.3	97.9	91.9	91.0	77.2	5:32
ProbCons-ext	90.0	82.5	94.2	59.1	84.3	61.1	93.8	81.0	98.1	92.2	91.2	77.6	8:02

Columns show the average sum-of-pairs (SP) and column scores (CS) achieved by each aligner for each of the five BALiBASE references. All scores have been multiplied by 100. The number of sequences in each reference is given in parentheses. Overall numbers for the entire database are reported in addition to the total running time of each aligner for all 141 alignments. The best results in each column are shown in bold.

SeqWeb ClustalW

<http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=clustalw-prot>

ClustalW+



Align several peptide sequences.

Input sequences:

Select From: **Default** | Project | Local File | Clipboard | Database

Sequence	Description	Type	Length	Range
hba_horse.pep	ID HBA_HORSE STANDARD; PRT; 141 AA.	P	141	1 .. 141
lgba_soybn	lgba_soybn	P	143	1 .. 143
hbahuman	hbahuman	P	141	1 .. 141
glb5_petma.pep	ID GLB5_PETMA STANDARD; PRT; 149 AA.	P	149	1 .. 149
lgb1_soybn.pep	- ID LGB1_SOYBN STANDARD; PRT; 143 AA.	P	143	1 .. 143
hba_human	hba_human	P	141	1 .. 141
lgb2_luplu.pep	ID LGB2_LUPLU STANDARD; PRT; 153 AA.	P	153	1 .. 153
hbb_human	hbb_human	P	146	1 .. 146

Refresh

Clear

Input Parameters:

Set pairwise alignment mode	<input checked="" type="radio"/> Slow/Accurate <input type="radio"/> Fast/Approximate	
Pairwise Alignment Parameters (Available for Slow/Accurate alignment mode)	Set gap opening penalty	<input type="text" value="10.0"/> (range 1.0 thru 10.0)
	Set gap extension penalty	<input type="text" value="0.1"/> (range 0.1 thru 10.0)
	Set alignment scoring matrix	<input type="text" value="BLOSUM"/>

SeqWeb ClustalW MSA Parameters

<http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=clustalw-prot>

Multiple Sequence Alignment Parameters

Set gap opening penalty	<input type="text" value="10.0"/>	(range 1.0 thru 10.0)
Set gap extension penalty	<input type="text" value="0.05"/>	(range 0.01 thru 10.0)
Set gap separation penalty range	<input type="text" value="4"/>	
Delay divergent	<input type="text" value="30.0"/>	(1.0 thru 100.0)
Turn on end gap separation penalty	<input type="checkbox"/>	
Matrix contains negative values	<input type="checkbox"/>	
Turn off residue-specific gap penalty	<input type="checkbox"/>	
Turn off hydrophilic residue gaps	<input type="checkbox"/>	
List hydrophilic residues	<input type="text" value="G,P,S,N,D,Q,E,K,R"/>	
Set alignment scoring matrix	<input type="text" value="BLOSUM"/>	
Set output sequences' ordering	<input checked="" type="radio"/> Input order <input type="radio"/> Alignment order	
Sequence range to write [m,n] (starting from m to m+n)	<input type="text"/>	

SeqWeb ClustalW Alignment

<http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=clustalw-prot>

SeqWeb v3.1

Multiple Sequence Alignment Results

[Text View](#)

```
MSF: 164 Type: P February 07, 2007 21:43 Check: 8543 ..
Name: hba\_horse.pep Len: 164 Check: 1470 Weight: 1.0
Name: lgba\_soybn Len: 164 Check: 52 Weight: 1.0
Name: hbahuman Len: 164 Check: 438 Weight: 1.0
Name: glb5\_petma.pep Len: 164 Check: 974 Weight: 1.0
Name: lgb1\_soybn.pep Len: 164 Check: 405 Weight: 1.0
Name: hba\_human Len: 164 Check: 438 Weight: 1.0
Name: lgb2\_luplu.pep Len: 164 Check: 3153 Weight: 1.0
Name: hbb\_human Len: 164 Check: 1613 Weight: 1.0
```

```
//
1 50
hba_horse.pe .....V LSAADKTNVK AAWSKVGGHA GEYGAEALER MFLGFPTTKT
lgba_soybn .....VA FTEKQDALVS SSFEAFKANI PQYSVVFYTS ILEKAPAAKD
hbahuman .....V LSPADKTNVK AAWGKVGAAH GEYGAEALER MFLSFPTTKT
glb5_petma.p PIVDTGCVAP LSAAEKTIR SAWAPVYSTY ETSQVDILVK FFTSTPAAQE
lgb1_soybn.p .....GA FTEKQDALVS SSFEAFKANI PQYSVVFYTS ILEKAPAAKD
hba_human .....V LSPADKTNVK AAWGKVGAAH GEYGAEALER MFLSFPTTKT
lgb2_luplu.p .....GA LTESQAALVK SSWEFFNANI PKITRFFIL VLEIAPAAKD
hbb_human .....VH LTPEEKSAVT ALWGVN...V DEVGGEALGR LLVVPWTQR

51 100
hba_horse.pe YFPHF.... .DLSHGSAQV KAHGKKVGD A LTLAVGHLD L LPGAALSLSD
lgba_soybn LFSFLANG... .VDPTIPKL TGAEKLFAL VRDAGQLKA SGTVVADAAL
hbahuman YFPHF.... .DLSHGSAQV KAHGKKVGD A LTLAVGHLD L LPGAALSLSD
glb5_petma.p FFPKFKGLTT ADQLKKSADV RWAERIINA VINDAVASMD TEKMSMKLRD
lgb1_soybn.p LFSFLANG... .VDPTIPKL TGAEKLFAL VRDAGQLKT NGTVVADAAL
hba_human YFPHF.... .DLSHGSAQV KAHGKKVGD A LTLAVGHLD L LPGAALSLSD
lgb2_luplu.p LFSFLKGTG. .EVPQNIPEL QAHAGKVFKL VYAAIQLQV TGVVVDATL
hbb_human FFESFGDLST PDAVMGHPKV KAHGKKVGD A FSDGLAHLDN LKGTATLSE

101 150
hba_horse.pe L....HAHK LRVDPVNFKL LSHCLLSTLA VHLPNDFTPA VHASLDKFLS
lgba_soybn G...SVHAQK AVTDP.QFVY VKEALLKTIK AAVGDKWSE LSRAWEVAYD
hbahuman L....HAHK LRVDPVNFKL LSHCLLVTLA AHLPAEFTPA VHASLDKFLA
glb5_petma.p LSG...KHAKS FQVDPQYFKV LAAVIADTVA AGD..... .AGFEKLM
lgb1_soybn.p V...SIHAQK AVTDP.QFVY VKEALLKTIK EAVGGNWSDE LSSAWEVAYD
hba_human L....HAHK LRVDPVNFKL LSHCLLVTLA AHLPAEFTPA VHASLDKFLA
lgb2_luplu.p KNLGSHVSK GVADA.HFPV VKEAILKTIK EVVGAKNSEE LNSAWTIAYD
hbb_human L....HCDK LHVDPENFRL LGNVLVCLVA HIFGKEFTPP VQAAYQKVVA

151 164
hba_horse.pe SVSTVLTISKY R...
lgba_soybn ELAAAIKKA. ....
hbahuman SVSTVLTISKY R...
glb5_petma.p MICILLRSAY ....
lgb1_soybn.p ELAAAIKKA. ....
hba_human SVSTVLTISKY R...
lgb2_luplu.p ELAIVIKKEM NDAA
hbb_human GVANALAIKY H...
```


SeqWeb ClustalW Text Output

<http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=clustalw-prot>

```

!!AA MULTIPLE_ALIGNMENT 1.0
MSF: 164 Type: P February 07, 2007 21:44 Check: 8543 ..
Name: hbahuman Len: 164 Check: 438 Weight: 1.0
Name: hba_human Len: 164 Check: 438 Weight: 1.0
Name: hba_horse.pep Len: 164 Check: 1470 Weight: 1.0
Name: hbb_human Len: 164 Check: 1613 Weight: 1.0
Name: lgba_soybn Len: 164 Check: 52 Weight: 1.0
Name: lgb1_soybn.pep Len: 164 Check: 405 Weight: 1.0
Name: lgb2_luplu.pep Len: 164 Check: 3153 Weight: 1.0
Name: glb5_petma.pep Len: 164 Check: 974 Weight: 1.0

//

hbahuman          1                               50
hbahuman          .....V LSPADKTNVK AAWGKVGGAH GEYGAEALER MFLSFPTTKT
hba_human         .....V LSPADKTNVK AAWGKVGGAH GEYGAEALER MFLSFPTTKT
hba_horse.pep     .....V LSAADKTNVK AAWSKVGGHA GEYGAEALER MFLGFPTTKT
hbb_human         .....VH LTPEEKSAVT ALWGKVN..V DEVGGEALGR LLVVYPWTQR
lgba_soybn        .....VA FTEKQDALVS SSFEAFKANI PQYSVVFYTS ILEKAPAAKD
lgb1_soybn.pep    .....GA FTEKQDALVS SSFEAFKANI PQYSVVFYNS ILEKAPAAKD
lgb2_luplu.pep    .....GA LTESQAALVK SSWEEFNANI PKHTRHFFIL VLEIAPAAKD
glb5_petma.pep    PIVDTGSVAP LSAAEKTKIR SAWAPVYSTY ETSGVDILVK FFTSTPAAQE

hbahuman          51                               100
hbahuman          YFPHF..... .DLSHGSAQV KGHGKKVADA LTNAVAHVDD MPNALSALSD
hba_human         YFPHF..... .DLSHGSAQV KGHGKKVADA LTNAVAHVDD MPNALSALSD
hba_horse.pep     YFPHF..... .DLSHGSAQV KAHGKKVGDA LTLAVGHLDD LPGALSNLSD
hbb_human         FFESFGDLST PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFFATLSE
lgba_soybn        LFSFLANG.. .VDPTNPKL TGHAEKLFAL VRDSAGQLKA SGTVVADAAL
lgb1_soybn.pep    LFSFLANG.. .VDPTNPKL TGHAEKLFAL VRDSAGQLKT NGTVVADAAL
lgb2_luplu.pep    LFSFLKGTGTS .EVPQNNPEL QAHAGKVFKL VYEAIIQLQV TGVVVTDATL
glb5_petma.pep    FFPKFKGLTT ADQLKKSADV RWAERIINA VNDAVASMDD TEKMSMKLRD

hbahuman          101                              150
hbahuman          L.....HAHK LRVDPVNFKL LSHCLLVTLA AHLPAEFTPA VHASLDKFLA
hba_human         L.....HAHK LRVDPVNFKL LSHCLLVTLA AHLPAEFTPA VHASLDKFLA
hba_horse.pep     L.....HAHK LRVDPVNFKL LSHCLLSTLA VHLPNDFTPA VHASLDKFLS
hbb_human         L.....HCDK LHVDPENFRL LGNVLVLCVLA HHFGKEFTPP VQAAYQKVVA
lgba_soybn        G...SVHAQK AVTDP.QFVV VKEALLKTIK AAVGDKWSEDE LSRAWEVAYD
lgb1_soybn.pep    V...SIHAQK AVTDP.QFVV VKEALLKTIK EAVGKNWSEDE LSSAWEVAYD
lgb2_luplu.pep    KNLGSVHVSK GVADA.HFPV VKEAILKTIK EVVGAKWSEE LNSAWTIAYD
glb5_petma.pep    LSG..KHAKS FQVDPQYFKV LAAVIADTVA AGD..... ..AGFEKLSM

hbahuman          151                              164
hbahuman          SVSTVLTSKY R...
hba_human         SVSTVLTSKY R...
hba_horse.pep     SVSTVLTSKY R...
hbb_human         GVANALAHKY H...
lgba_soybn        ELAAAIKKA. ....
lgb1_soybn.pep    ELAAAIKKA. ....
lgb2_luplu.pep    ELAIVIKKEM NDAA
glb5_petma.pep    MICILLRSAY ....

```

Align several peptide sequences.

Input sequences:

Select From:

Default

Project

Local File

Clipboard

Database

Sequence	Description	Type	Length	Range
myg_phyca	myg_phyca	P	153	1 .. 153
glb5_petma.pep	ID GLB5_PETMA STANDARD; PRT; 149 AA.	P	149	1 .. 149
hba_human	hba_human	P	141	1 .. 141
hba_horse.pep	ID HBA_HORSE STANDARD; PRT; 141 AA.	P	141	1 .. 141
hbb_horse.pep	ID HBB_HORSE STANDARD; PRT; 146 AA.	P	146	1 .. 146
lgb1_soybn.pep	- ID LGB1_SOYBN STANDARD; PRT; 143 AA.	P	143	1 .. 143
hbb_human	hbb_human	P	146	1 .. 146

Refresh Clear

Input Parameters:

Select a sequence comparison matrix. This matrix determines how matches and mismatches are scored. The default penalties for gap creation and extension are given after each matrix name.

Scoring Matrix

blosum62

Set gap creation penalty

8

Set gap extension penalty

2

Limit the maximum input sequence range only when needed. Setting a higher limit allows you to align longer sequences while setting a lower limit allows you to add more and longer gaps to each sequence.

Maximum input sequence range

5000 (range 1 thru 7000)

Limit the maximum number of gaps only when needed. Setting a higher limit allows you to add more and longer gaps to each sequence while setting a lower limit allows you to align a greater number of sequences.

Maximum number of gap characters ('.' and '~') added to any sequence

2000 (range 0 thru 7000)

Run

Reset

Symbol comparison table: share_matrix:blosum62.cmp CompCheck: 1102

Gapweight: 8
GapLengthweight: 2

myg_phyca_pileup_15277.txt MSF: 165 Type: P January 28, 2010 10:16 Check: 6593 ..

Name:	Len:	Check:	Weight:
hba_human	165	1231	1.00
hba_horse	165	2167	1.00
hbb_horse	165	9310	1.00
hbb_human	165	208	1.00
glb5_petma	165	2079	1.00
myg_phyca	165	4320	1.00
lgb1_soybn	165	7278	1.00

```
//
      1                               50
hba_human ~~~~~~V lspadktrvk aawgkvgaha geygaealer mflsfpttkt
hba_horse ~~~~~~V LSAADKTNVK AAWSKVGGAH GEYGAEALER MFLGFPTTKT
hbb_horse ~~~~~~VQ LSGEKAQAVL ALWDKV..NE EEVGGEALGR LLVVPWTQR
hbb_human ~~~~~~Vh ltpooksavt alwkv..nv devggealgr llvypwtqr
glb5_petma PIVDTGSVAP LSAAEKTIR SAWAPVYSTY ETSGVDILVK FFTSTPAAQE
myg_phyca ~~~~~~V lsegewqlvl hwakveadv aghgqdlir lfkshpetle
lgb1_soybn ~~~~~~ga ftekqealvs ssfeafkani pqysvfyfns ilekapaakd

      51                               100
hba_human yfphf.dlsh .....gsaqv kghgkkvada ltnavahvdd mpnalsalsd
hba_horse YFPHF.DLSH .....GSAQV KAHGKKVGDA LTLAVGHLDD LPGALSNLSD
hbb_horse FFDSFGDLSN PGAVMGNPKV KAHGKKVLHS FEGEVHHLDN LKGTFAALSE
hbb_human ffesfgdlst pdavmgpkv kahgkkviga fsdglahldn lkgtfatise
glb5_petma FFPKFKGLTT ADQLKKSADV RWAERIINA VNDAVASMDD TEKMSMKLRD
myg_phyca kdrfrkhlkt eaemkasedl kkhgvtvltg lg...ailkk kghheaelkp
lgb1_soybn lfsflan... .gvdptrpk l tghaeklfal vrdsagql.k tngtvvadaa

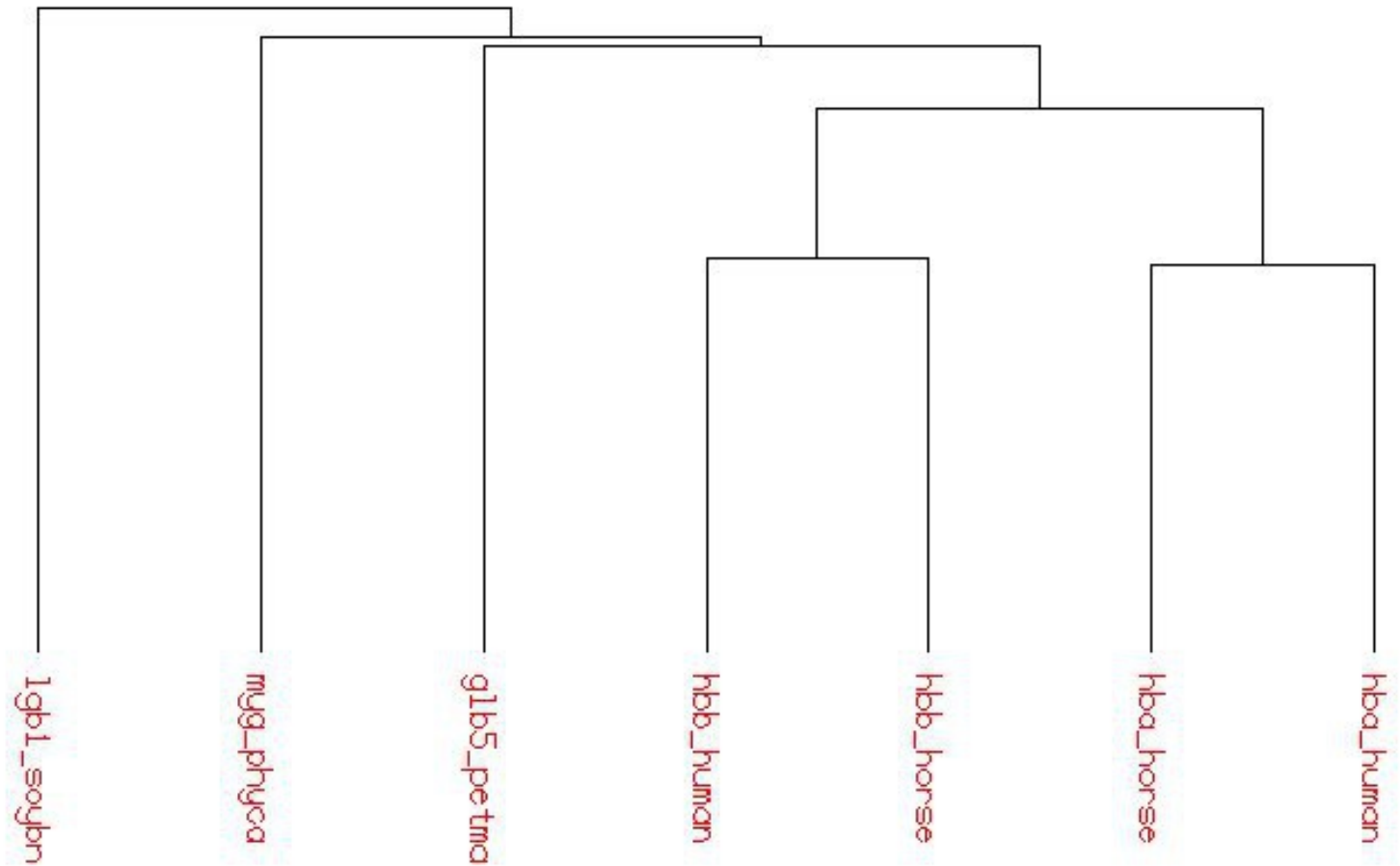
      101                              150
hba_human l...hahklr vdpvnfklls hc1lvtlaah lpaeftpavh asldkflasv
hba_horse L...HAHKLR VDPVNFKLLS HCLLSTLAVH LPNDFTPAVH ASLDKFLSSV
hbb_horse L...HCDKLH VDPENFRLG NVLVVVLARH FGKDFPELQ ASYQKVAVG
hbb_human l...hcdklh vdpnfrllg nvlvcvlahh fgkeftppvq aayqkvavg
glb5_petma LSGKHAQSFQ VDPQYFKVLA AVIADTVA.. .....AGD AGFEKLSMI
myg_phyca laqshatkhh ipikylefis eaiihvlhsh hpgdfgadaq gamnkalelf
lgb1_soybn lvsahaqkav tdpq.fvvvk eallktikea vggnwdsels sawevaydel

      151                              165
hba_human stvltskyr~ ~~~~~~
hba_horse STVLTSKYR~ ~~~~~~
hbb_horse ANALAHKYH~ ~~~~~~
hbb_human analahkyh~ ~~~~~~
glb5_petma CILLRSAY~ ~~~~~~
myg_phyca rkdiaakyke lgyqq
lgb1_soybn aaaiikka~ ~~~~~~
```

SeqWeb Pileup Dendrogram

<http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=pileup-prot>

Multiple Sequence Alignment Dendrogram January 28, 2010 10:16



SeqWeb Pretty Input

<http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=pretty-prot>

Pretty

Align several peptide sequences and calculate a consensus.

Input sequences:

Select From:

Sequence	Description	Type	Length	Range
myg_phyca	myg_phyca	P	153	1..153
gib5_petma.pep	ID GLB5_PETMA STANDARD; PRT; 149 AA.	P	149	1..149
hba_human	hba_human	P	141	1..141
hba_horse.pep	ID HBA_HORSE STANDARD; PRT; 141 AA.	P	141	1..141
hbb_horse.pep	ID HBB_HORSE STANDARD; PRT; 146 AA.	P	146	1..146
lgb1_soybn.pep	- ID LGB1_SOYBN STANDARD; PRT; 143 AA.	P	143	1..143
hbb_human	hbb_human	P	146	1..146

Input Parameters:

Select a sequence comparison matrix. This matrix determines how matches and mismatches are scored. The default penalties for gap creation and extension are given after each matrix name.

[Scoring Matrix](#)

[Set gap creation penalty](#)

[Set gap extension penalty](#)

Limit the maximum input sequence range only when needed. Setting a higher limit allows you to align longer sequences while setting a lower limit allows you to add more and longer gaps to each sequence.

[Maximum input sequence range](#)

(range 1 thru 7000)

Limit the maximum number of gaps only when needed. Setting a higher limit allows you to add more and longer gaps to each sequence while setting a lower limit allows you to align a greater number of sequences.

[Maximum number of gap characters \('.' and '~'\) added to any sequence](#)

(range 0 thru 7000)

[Display consensus only at completely conserved positions in the alignment](#)

[At each column in the alignment:](#)

show positions agreeing with the consensus in upper case

display alignment only at positions that disagree with the consensus

none of the above

are identical

are similar

are somewhat similar

[Count residues toward consensus that](#)[Minimum number of votes required for a consensus](#)

Pretty Results

Seq

Plurality: 2.00 Threshold: 4 AveWeight 1.00 AveMatch 2.78 AvMismatch

Consensus Sequence

Symbol comparison table: share_matrix:blosum62.cmp CompCheck: 1102

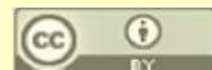
Gapweight: 8
GapLengthweight: 2

Pileup MSF: 165 Type: P January 28, 2010 10:22 Check: 6593 ..

Name: hba_human	Len: 165	Check: 1231	weight: 1.00
Name: hba_horse.pep	Len: 165	Check: 2167	weight: 1.00
Name: hbb_horse.pep	Len: 165	Check: 9310	weight: 1.00
Name: hbb_human	Len: 165	Check: 208	weight: 1.00
Name: glb5_petma.pep	Len: 165	Check: 2079	weight: 1.00
Name: myg_phyca	Len: 165	Check: 4320	weight: 1.00
Name: lgb1_soybn.pep	Len: 165	Check: 7278	weight: 1.00



//

hba_human	1	~~~~~v	lspadktnvk	aawgkvgaha	geygaealer	mflsfpttkt	50
hba_horse		~~~~~v	LSAADKTNVK	AAWSKVGGHA	GEYGAEALER	MFLGFPTTKT	
hbb_horse		~~~~~VQ	LSGEEKAAVL	ALWDKV..NE	EEVGGEALGR	LLVVYPWTQR	
hbb_human		~~~~~vh	ltpEEKsavt	alwgkv..nv	devggealgr	llvvyPwtqr	
glb5_petma		PIVDTGSVAP	LSAAEKTkir	SAWAPVYSTY	ETSGVDILVK	FFTSTPAAQE	
myg_phyca		~~~~~v	lsegewqlvl	hwakveadv	aghgqdlir	lfkshpetle	
lgb1_soybn		~~~~~ga	ftekqealvs	ssfeafkani	pqysvfvfyns	ilekapaakd	
Consensus		-----w	LS-AEKT-V-	AAW-KVGAN-	-EYG-EAL-R	LF-S-P-T--	
hba_human	51	yfphf.dlshgsaqv	kghgkkvada	ltnavaHvdd	mpnalsalsd	100
hba_horse		YFPHF.DLSHGSAQV	KAHGKKVGDA	LTLAVGHLDD	LPGALSNSLD	
hbb_horse		FFDSFGDLSN	PGAVMGNPKV	KAHGKKVLHS	FGEVHHLDN	LKGTFAALSE	
hbb_human		ffesfgdlst	pdavmgnpkv	kahgkkvlga	fsdglahldn	lkgtfatlse	
glb5_petma		FFPKFKGLTT	ADQLKKSADV	RWHAERIINA	VNDAVASMDD	TEKMSMKLRD	
myg_phyca		kfdrfkhlkt	eaemkasedl	kkhgvtvlt	lg...ailkk	kghheaelkp	
lgb1_soybn		lfsflan...	.gvdptnpkl	tghaeklfal	vrdsagql.k	tngtvvadaa	
Consensus		FFP-F-DLST	P-AV-GS-KV	KAHGKKVLDA	L-DAVAHLDD	L-GT-AALSD	
hba_human	101	l...hahklr	vdpvnfklls	hcllvtlaah	lpaeftpavh	aslDKflasv	150
hba_horse		L...HAHKLR	VDPVNFKLLS	HCLLSTLAVH	LPNDFTPAVH	ASLDKFLSSV	
hbb_horse		L...HCDKLH	VDPENFRLLG	NVLVWVLARH	FGKDFTPELQ	ASYQKVAVG	
hbb_human		l...hcdklh	vdpenfrllg	nlvcvlahh	fgkeftppvq	aayqkvavg	
glb5_petma		LSGKHAKSfq	VDPQYFKVLA	AVIADTVA..AGD	AGFEKLMsMI	
myg_phyca		laqshatkHk	ipikylefis	eaiihvlhsr	hpgdfgadaq	gamnkalelf	
lgb1_soybn		lvsahaqkav	tdpq.fvvvk	eallktikea	vpgnwsdels	sawekaydel	
Consensus		L---HA-KL-	VDP-NFKLLS	-VLLVTLA-H	---DFTPAVQ	A---K-LA-V	
hba_human	151	stvltskyr~	~~~~~	165			
hba_horse		STVLTSKYR~	~~~~~				
hbb_horse		ANALAHKYH~	~~~~~				
hbb_human		analahkyh~	~~~~~				
glb5_petma		CILLRSAY~	~~~~~				
myg_phyca		rkdiaakyke	lgyqq				
lgb1_soybn		aaaiikka~	~~~~~				
Consensus		A-ALASKY--	-----				



Decypher ClustalW Input

<http://decypher.stanford.edu/>

ClustalW Search on DeCypher
 Protein Sequences

?	Job Description:	<input type="text" value="ClustalW Search on DeCypher Protein Sequences"/>
?	E-mail Address:	<input type="text"/>
?	Return Results:	To your web browser ▾ As: <input type="text" value="Web Page"/> ▾

Protein Query Set:
Click *Browse...* to upload your local file, or paste query data into the text box.

[Use Example Query](#)

?	Job Options:	<input type="button" value="Hide"/>
?	KTuple Size:	<input type="text" value="1"/>
?	Window Size:	<input type="text" value="5"/>
?	Pairwise Gap Penalty:	<input type="text" value="3"/>
?	Gap Open Penalty:	<input type="text" value="10"/>
?	Gap Extend Penalty:	<input type="text" value="0.05"/>
?	Matrix:	<input type="text" value="BLOSUM"/> ▾
?	Hydrophilic Residues:	<input type="text" value="GPSNDQEKR"/>
?	Residue-Specific Gaps:	<input type="text" value="ON"/> ▾
?	Hydrophilic Gaps:	<input type="text" value="ON"/> ▾
?	Negative Matrix Values:	<input type="text" value="OFF"/> ▾

?	Gap Separation Distance:	<input type="text" value="8"/>
?	End Gaps:	<input type="text" value="OFF"/> ▾
?	Transition Weight:	<input type="text" value="0.5"/>
?	Top Diagonals:	<input type="text" value="5"/>
?	Score:	<input type="text" value="PERCENTAGE"/> ▾
?	Show Sequence #s:	<input type="text" value="OFF"/> ▾
?	Output Order:	<input type="text" value="ALIGNED"/> ▾

Decypher ClustalW Results

<http://decypher.stanford.edu/>

[Home Page](#)

Build HMM Model | Click once!

Results for Job CGI_Temp2623732f208
ClustalW Search on DeCypher Protein Sequences

[View dendrogram](#)

```

GLB5_PETMA      PIVDTGGSVAPLSAAEKTIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
HBA_HORSE      -----VLSAADKTNVKAAWSKGVGHAGEYGAEALERMFLGFPTTKTYFPHF-DLSH
HBA_HUMAN      -----VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH
HBB_HORSE      -----VQLSGEEKAAVLALWDKV--NEEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
HBB_HUMAN      -----VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLST
LGB2_LUPLU     -----GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
MYG_PHYCA      -----VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
  
```

```

GLB5_PETMA      ADQLKKSADVRWHAERIINAVNDAVASMDDTEK--MSMKLRDLSGKHAKSFQVDPQYFKV
HBA_HORSE      -----GSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
HBA_HUMAN      -----GSAQVKGHGKKVADALTNAVAHVDD-----MPNALSLSDLHAHKLRVDPVNFKL
HBB_HORSE      PGAVMGNPKVKAHGKKVLHSFGEGVHLLDN-----LKGTFALSELHCDKLHVDPENFRL
HBB_HUMAN      PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
LGB2_LUPLU     VPQ--NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSK-GVADAHFPV
MYG_PHYCA      EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
  
```

```

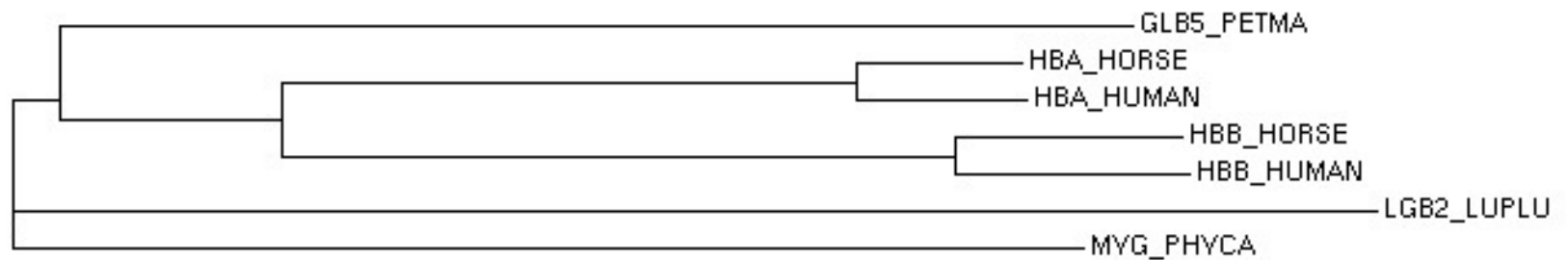
GLB5_PETMA      LAAVIADTVAAG-----DAGFEKLMSMICILLRSAY-----
HBA_HORSE      LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR-----
HBA_HUMAN      LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSVTVLTSKYR-----
HBB_HORSE      LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH-----
HBB_HUMAN      LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-----
LGB2_LUPLU     VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
MYG_PHYCA      ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
  
```

Decypher ClustalW Results

<http://decypher.stanford.edu/>

Dendrogram

[Return to top](#)



ClustalW @ EBI Input

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

EBI > Tools > Sequence Analysis > ClustalW

ClustalW

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. [New users, please read the FAQ.](#)

>> Download Software   

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
<input type="text"/>	Sequence	interactive	full	single
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def
OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Enter or Paste a set of Sequences in any supported format:

Help

Upload a file:

Browse...

Run

Reset

ClustalW @ EBI Results

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

Alignment

CLUSTAL W (1.83) multiple sequence alignment

```

HBA_HORSE      -----VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPFDLS-- 49
HBA_HUMAN      -----VLSPADKTNVKAAWGKVGAGHAGEYGAEALERMFLSFPTTKTYFPFDLS-- 49
HBB_HORSE      -----VQLSGEKA AVLALWDKVN--EEVVGGEALGRLLVVYPWTQRFFDSFGDLSN 50
HBB_HUMAN      -----VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST 50
GLB5_PETMA     PIVDTGVSAPLSAAEKTIRSAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT 60
MYG_PHYCA     -----VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFD RFKHLKT 51
LGB2_LUPLU     -----GALTESQAALVKSSWEEFNANIPKHTRFFILVLEIAPAADLFSFLKGTSE 52
                * : : * . : .. * : * :

HBA_HORSE      ----HGSAQVKAHGKKVGDALTLAVGHLDL-----LPGALS NLSDLHAHKLRVDPVNFKL 100
HBA_HUMAN      ----HGSAQVKGHGKKVADAL TNAVAHVDD-----MPNALSALS DLHAHKLRVDPVNFKL 100
HBB_HORSE      PGAVMGNPKVKAHGKKVLH SFGEGVHHLDN-----LKGTF AALSELHCDKLHVDPENFRL 105
HBB_HUMAN      PDAVMGNPKVKAHGKKV LGA FSDGLAHLDN-----LKGTF ATLSELHCDKLHVDPENFRL 105
GLB5_PETMA     ADQLKKSADVRWHAERI INAVNDAVASMDDT--EKMSMKLRDL S GKHAKSFQVDPQYFKV 118
MYG_PHYCA     EAEMKASEDLKKHGVT VLTALGAILKKKGH-----HEAELKPLAQ SHATKHKIPIKYLEF 106
LGB2_LUPLU     VP--QNNPELQAHAGKV FKL VYEAAIQLVQVTGVVVT DATLKNLGSVHVSKGVAD-AHFPV 109
                . .:: * . : . : * . * . : .

HBA_HORSE      LSHCLLSTLAVHLPNDFT PAVHASLDKFLSSVSTVLTSKYR----- 141
HBA_HUMAN      LSHCLLVTLAAHLPAEFT PAVHASLDKFLASVSTVLTSKYR----- 141
HBB_HORSE      LGNVLVVVLARHFGKDF TP ELQASYQKV VAGVANALAHKYH----- 146
HBB_HUMAN      LGNVLVCVLAHHFGKE FT PPVQAAYQKV VAGVANALAHKYH----- 146
GLB5_PETMA     LAAVIADTVAAG-----DAGFEKLSMICILLRSAY----- 149
MYG_PHYCA     ISEAIHVLHSRHPGDFG ADAQGAMNKALELFRKDIAAKYKELGYQG 153
LGB2_LUPLU     VKEAILKTIKEVVGAKW SEELNSAWTTIAYDELAIVIKKEMNDAA--- 153
                : : .. ... . :

```

PLEASE NOTE: Showing colors on large alignments is slow.

