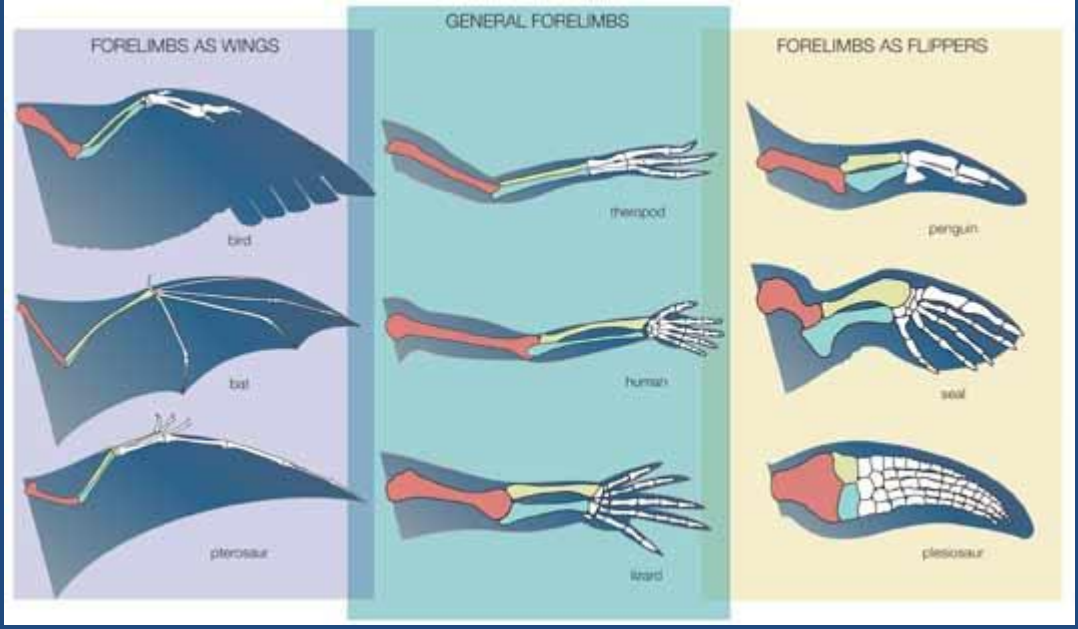
Bioinformatics Tools for

Sequence homology and alignment

Homology

- Similarity between characters due to a common ancestry



Sequence homology

- Similarity between sequences that results from a common ancestor

VLSPAVKWAKVGAHAAGHG

VLSEAVLWAKVEADVAGHG

- Basic assumption:

Sequence homology →
similar structure/function

Sequence alignment

Alignment: Comparing two (pairwise) or more (multiple) sequences. Searching for a series of identical or similar characters in the sequences.

Homology

- Ortholog – homolog with similar function (via speciation)
- Paralog – homolog which arose from gene duplication

Common use:

Orthologs – 2 homologs from **different** species

Paralogs – 2 homologs within the **same** species

How close?

- Rule of thumb:
- Proteins are homologous if 25% identical (length >100)
- DNA sequences are homologous if 70% identical

Twilight zone

- < 20% identity in proteins – may be homologous and may not be….
- (Note that 5% identity will be obtained completely by chance!)

Local vs. Global

- Global alignment – finds the best alig
across the entire two sequences.

```
ADLGAVFALCDRYFQ
||||     ||||| |
ADLGRTQN-CDRYYQ
```

- Local alignment – finds regions of simi
parts of the sequences.

```
ADLG    CDRYFQ
||||    ||||| |
ADLG    CDRYYQ
```

Global alignment: forces alignment in regions which differ

Local alignment will return only **regions** of good alignment

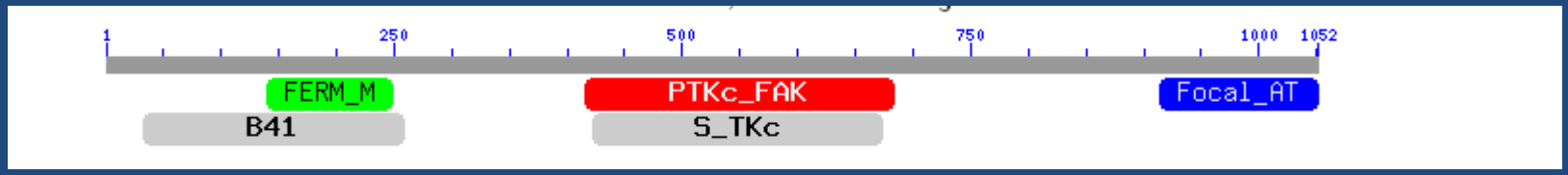
When global and when local?

Global alignment

- PTK2 protein tyrosine kinase 2 of human and rhesus monkey

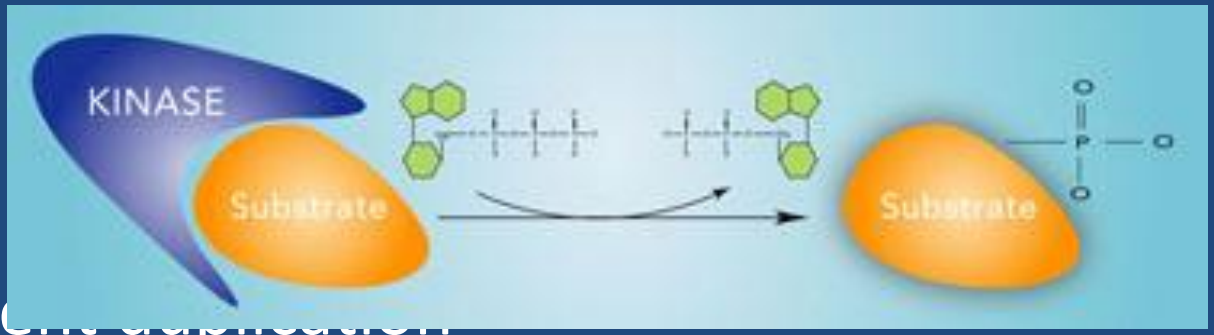
```
human     107 VREKYELAHPPEEWKYELRIRYLPKGFLNQFTEDKPTLNFFYQQVKSDYM  156
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus    151 VREKYELAHPPEEWKYELRIRYLPKGFLNQFTEDKPTLNFFYQQVKSDYM  200

human     157 LEIADQVDQEIALKLGCLEIRRSYWEMRGNALEKKSNYEVLEKDVGLKRF  206
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus    201 LEIADQVDQEIALKLGCLEIRRSYWEMRGNALEKKSNYEVLEKDVGLKRF  250

human     207 FPKSLLDSVKAKTLRKLIQQTFRQFANLNREESILKFFEILSPVYRFDKE  256
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus    251 FPKSLLDSVKAKTLRKLIQQTFRQFANLNREESILKFFEILSPVYRFDKE  300

human     257 CFKCALGSSWIISVELAIGPEEGISYLTDKGCNPTHLADFTQVQTIQYSN  306
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus    301 CFKCALGSSWIISVELAIGPEEGISYLTDKGCNPTHLADFTQVQTIQYSN  350

human     307 SEDKDRKGMLQLKIAGAPEPLTVTAPSLTIAENMADLIDGYCRLVNGTSQ  356
              ||||||||||||||||||||||||||||||||||||||||||||||||.||
rhesus    351 SEDKDRKGMLQLKIAGAPEPLTVTAPSLTIAENMADLIDGYCRLVNGASQ  400

human     357 SFIIRPQKEGERALPSIPKLANSEKQGMRTHAVSVSETDDYAEIIDEEDT  406
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
rhesus    401 SFIIRPQKEGERALPSIPKLANSEKQGMRTHAVSVSETDDYAEIIDEEDT  450

human     407 YTMPSTRDYEIQRERIELGRCIGEGQFGDVHQGIYMSPENPALAVAIKTC  456
              ||||||||||||||||||||||||||||||||||:||||||||||||||||
rhesus    451 YTMPSTRDYEIQRERIELGRCIGEGQFGDVHQGVYMSPENPALAVAIKTC  500

human     457 KNCTSDSVREKFLQEALTMRQFD-HPHIVKLIGVITENPVWIIMELCTLG  505
              ||||||||||||||||.||..| .|..|.|..|...|.:||||::..|.
rhesus    501 KNCTSDSVREKFLQEALTMRQFDSHPHIVKLIGVMTENPVWIIMELCTLG  550
```

Protein tyrosine kinase domain



Protein tyrosine kinase domain

- Human PTK2 and leukocyte tyrosine kinase
- Both function as tyrosine kinases, in completely different contexts



- Ancient duplication

Global alignment of PTK and LTK

Local alignment of PTK and LTK

```
human_ptk2       343 LIDGYCRLVNGTSQSFIIRPQKE----GERALPSIPKLANSEKQGMRTHA    388
                     |:     :|.|.   .|:..||:    .|..||| |:|   |...:||.|
human_LTK        439 LL-----MVCGV---LILVKQKKWQGLQEMRLPS-PEL---ELSKLRTSA    476

human_ptk2       389 VSVSETDDYAEI-IDEEDTYTMPSTRDYEIQRERIELGRCIGEGQFGDVH    437
                     :..:....|.:: :.....:.:|.... |:....:.|.|.:|.|.||:|:
human_LTK        477 IRTAPNPYYCQVGLGPAQSWPLPPGVT-EVSPANVTLLRALGHGAFGEVY    525

human_ptk2       438 QG--IYMSPENPALAVAIKTCKNCTSDSVREKFLQEALTMRQFDHPHIVK    485
                     :|  |.:..::..|.|||||.....|......||.|||.:.:|.|.:||:
human_LTK        526 EGLVIGLPGDSSPLQVAIKTLPELCSPQDELDFLMEALIISKFRHQNIVR    575

human_ptk2       486 LIGV-ITENPVWIIMELCTLGELRSFLQVRKYSLD------LASLILYAY    528
                     .:|: :...|..|::||.:.|:::|||:..:..|.      :..|:..|.
human_LTK        576 CVGLSLRATPRLILLELMSGGDMKSFLRHSRPHLGQPSPLVMRDLLQLAQ    625

human_ptk2       529 QLSTALAYLESKRFVHRDIAARNVLVS---SNDCVKLGDFGLSRYMEDST    575
                     .::....|||...|:|||||||.|:|   .:...|:||||::|.:...::
human_LTK        626 DIAQGCHYLEENHFIHRDIAARNCLLSCAGPSRVAKIGDFGMARDIYRAS    675

human_ptk2       576 YY-KASKGKLPIKWMAPESINFRRFTSASDVWMFGVCMWEILMHGVKPFQ    624
                     || :..:..||:|||.||:.....|||.:|.|.|||.:|||...|..|:.
human_LTK        676 YYRRGDRALLPVKWMPPEAFLEGIFTSKTDSWSFGVLLWEIFSLGYMPYP    725

human_ptk2       625 GVKNNDVIGRIENGERLPMPPNCPPTLYSLMTKCWAYDPSRRPRFTE---    671
                     |..|.:|:..:..|.|:..|..||..:|.:||:||.::|..||.|..
human_LTK        726 GRTNQEVLDFVVGGGRMDPPRGCPGPVYRIMTQCWQHEPELRPSFASILE    775
```

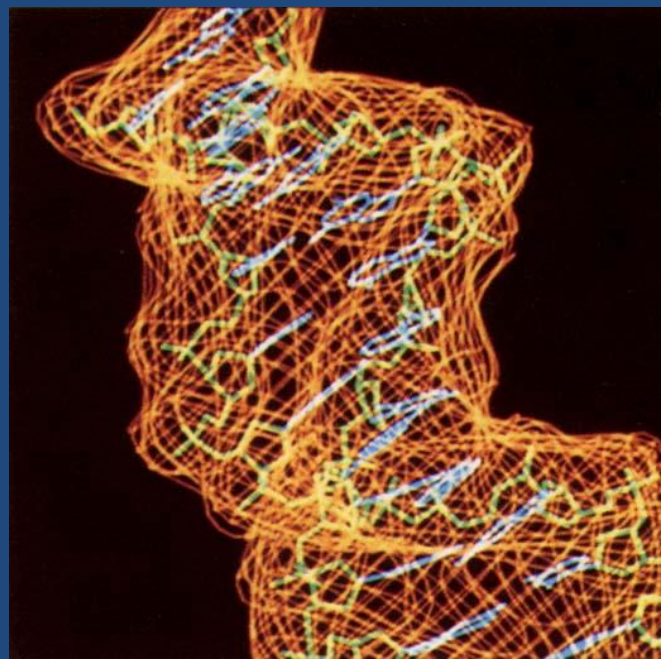
Searching databases

Searching a database

- Using a sequence as the query to find homologous sequences in the database

DNA or protein?

- For coding sequences, we can use the DNA sequence or the protein sequence to search for similar sequences.
- Which is preferable?



Protein is better!

- Selection (and hence conservation) works on the protein level:

```
CTTTCA =   Leu-Ser
TTGAGT =   Leu-Ser
```

Query type

- Nucleotides: a four letter alphabet
- Amino acids: a twenty letter alphabet

- Two random DNA sequences will share on average 25% of identity
- Two random protein sequences will share on average 5% of identity

Conclusions

- Using the amino acid sequence is preferable for homology search

- Why use a nucleotide sequence after all?
- No ORF found, e.g. newly sequenced genome
- No similar protein sequences were found
- Specific DNA databases are available (EST)

Some terminology

- Query sequence - the sequence with which we are searching
- Hit – a sequence found in the database, suspected as homologous

How do we search a database?

- Assume we perform pairwise alignment of the query against all the sequences in the database
- Exact pairwise alignment is $O(mn) \approx O(n^2)$
  (m – length of sequence 1,
  n – length of sequence 2)

How much time will it take?

- $O(n^2)$ computations per search.
- Assume n=200, so we have 40,000 computations per search
- Size of database - ~60 million entries
- 2.4×10^{12} computations for each sequence search we perform!
- Assume each computation takes 10^{-6} seconds → 24,000 seconds ≈ 6.66 hours for each sequence search
- 150,000 searches (at least!!) are performed per day

Conclusion

- Using the exact comparison pairwise alignment algorithm between query and all DB entries – too slow

Heuristic

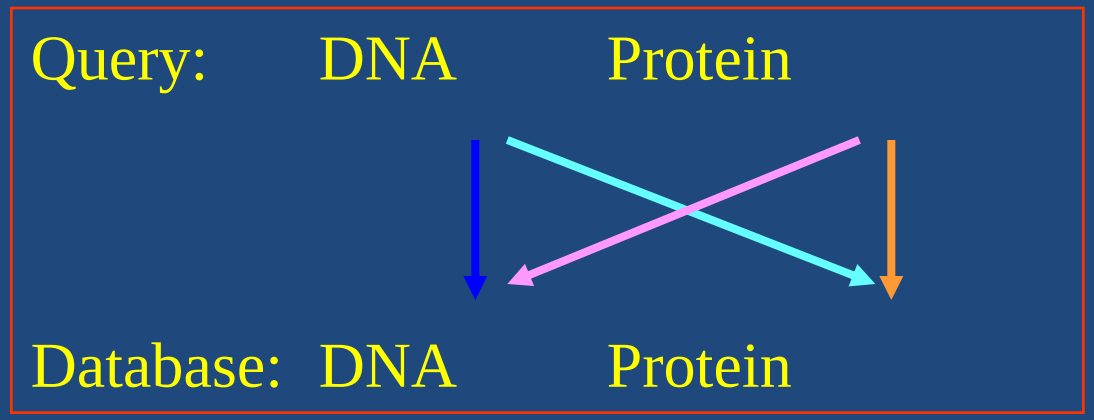
- Definition: a heuristic is a design to solve a problem that does not provide exact solution (but is not too bad) and reduces the time complexity of the exact solution

BLAST

- BLAST - Basic Local Alignment and Search Tool
- A heuristic for searching a database for similar sequences

DNA or Protein

- All types of searches are possible.



blastn – nuc vs. nuc
blastp – prot vs. prot
blastx – translated query vs. protein database
tblastn – protein vs. translated nuc. DB
tblastx – translated query vs. translated database

Translated databases:

trEMBL
genPept

BLAST - underlying hypothesis

- The underlying hypothesis: when two sequences are similar there are short ungapped regions of high similarity between the two
- The heuristic:

1. Discard irrelevant sequences
2. Perform exact **local** alignment with remaining sequences

How do we discard irrelevant sequences quickly?

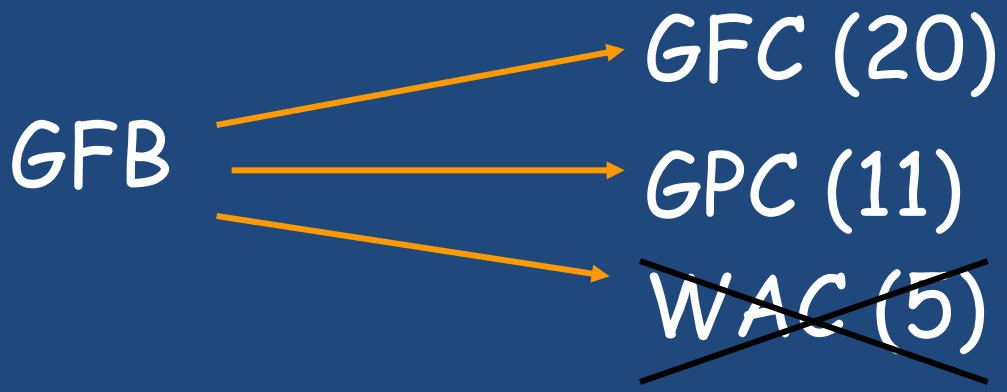
- Divide the **database** into words of length w (w = 3 for protein and w = 7 for DNA)
- Save the words in a look-up table that can be searched quickly

BLAST: discarding sequences

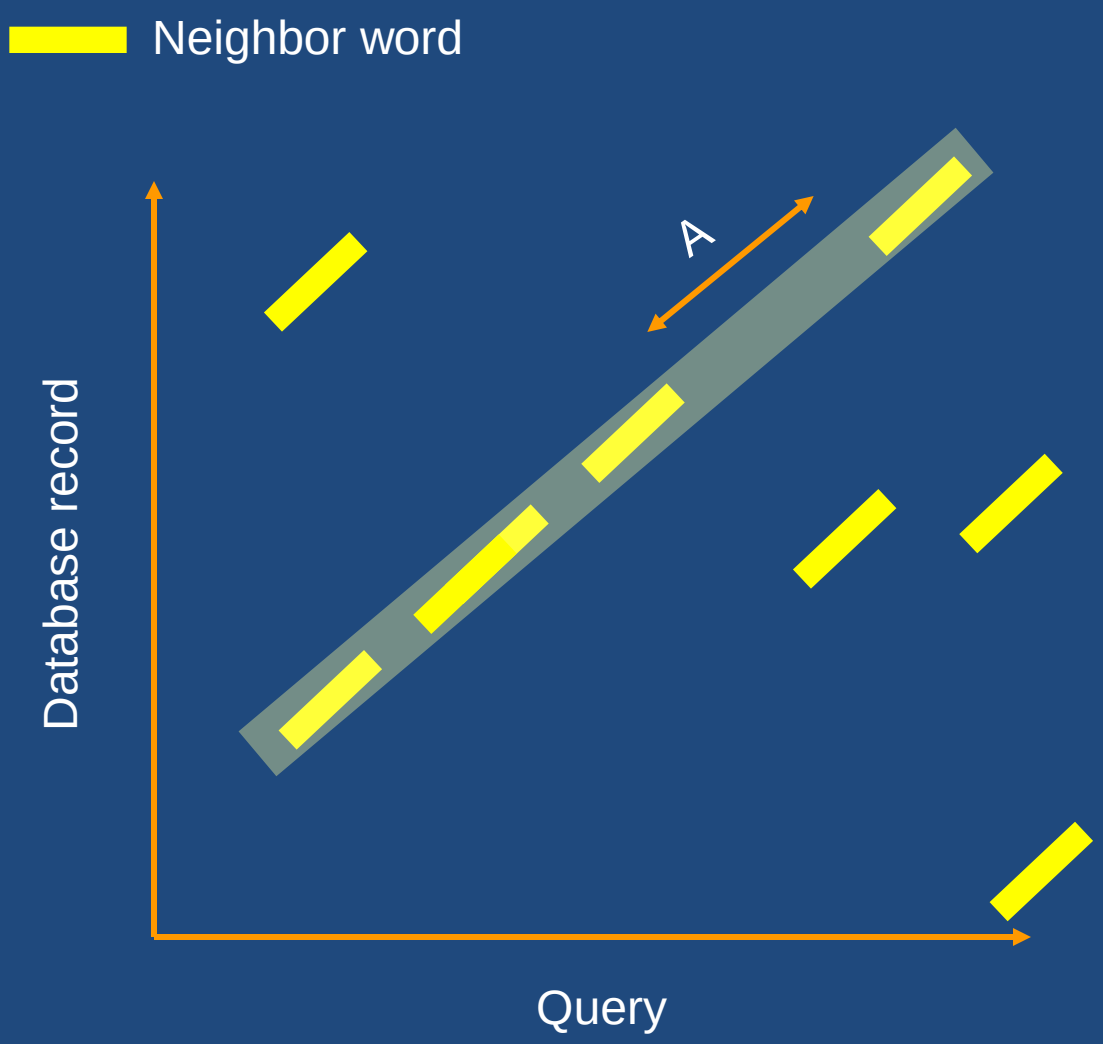
- When the user gives a query sequence, divide it also into words
- Search the **database** for consecutive neighbor words

Neighbour words

- neighbor words are defined according to a scoring matrix (e.g. BLOSUM62 for proteins) with a certain cutoff level



Search for consecutive words



Look for a seed: hits on the same diagonal which can be connected

At least 2 hits on the same diagonal with distance which is smaller than a predetermined cutoff

This is the filtering stage – many unrelated hits are filtered, saving lots of time!

The result – local alignment

- The result of BLAST will be a series of local alignments between the query and the different hits found

E-value

- The number of times we will theoretically find an alignment with a score ≥ **Y** of a random sequence vs. a random database

Theoretically, we could trust any result with an E-value ≤ **1**

In practice – BLAST uses estimations.
E-values of $\mathbf{10^{-4}}$ and lower indicate a significant homology.
E-values between $\mathbf{10^{-4}}$ and $\mathbf{10^{-2}}$ should be checked (similar domains, maybe non-homologous).
E-values between $\mathbf{10^{-2}}$ and **1** are suspicious…

Filtering low complexity

- Low complexity regions : e.g., Proline rich areas (in protein), Alu repeats (in DNA)
- Regions of low complexity generate high score of alignment BUT – this does not indicate homology

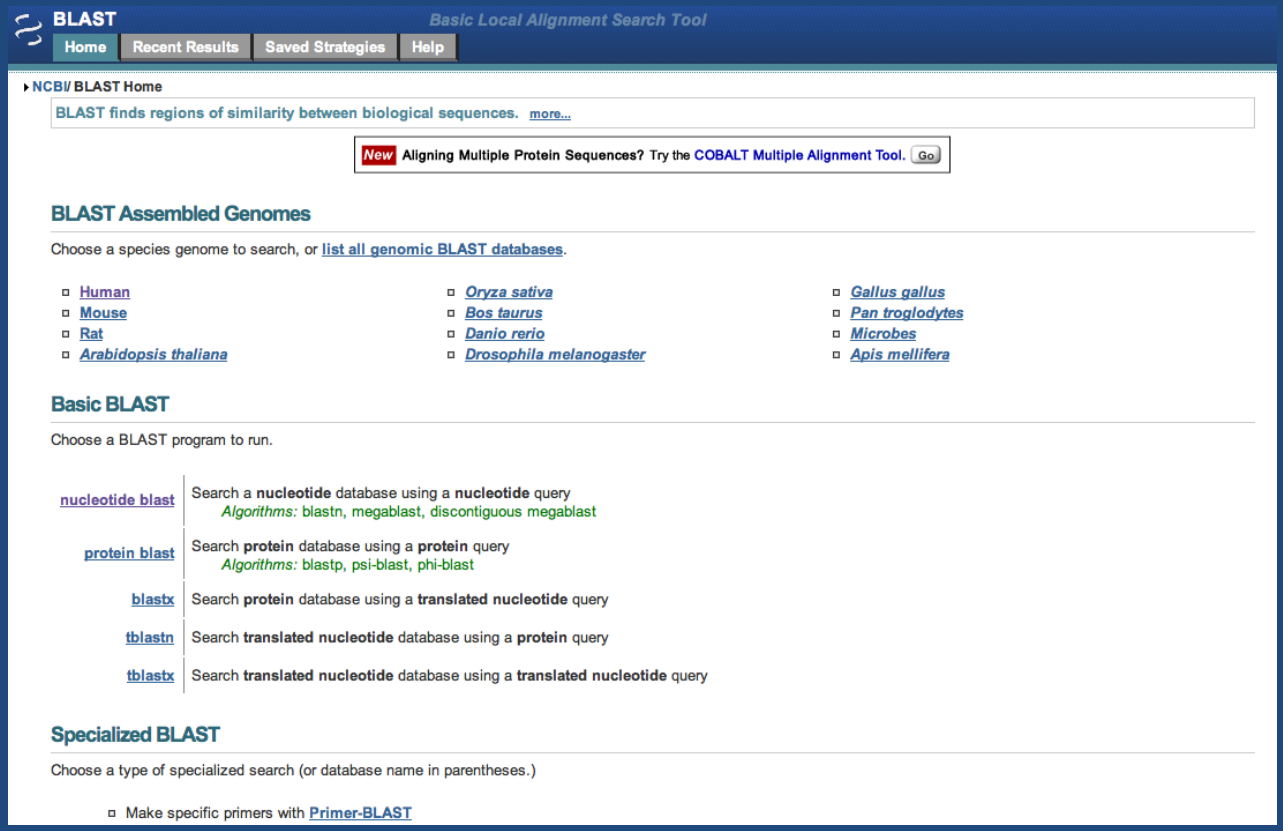
Solution

- In BLAST there is masking of low-complexity regions in the query sequence (such regions are represented as XXXXX in query)

BLAST Programs

- You can do it on line:

http://blast.ncbi.nlm.nih.gov/Blast.cgi



```
[shzhang@usa]$ ./formatdb –help
formatdb 2.2.15   arguments:
  -t  Title for database file [String]  Optional
  -i  Input file(s) for formatting [File In]  Optional
  -l  Logfile name: [File Out]  Optional
    default = formatdb.log
  -p  Type of file
        T - protein
        F - nucleotide [T/F]  Optional
    default = T
```

1. Check whether BLAST is in your path
   > which blastall

2. Target sequences should be formatted before it's searched against.
   a. Copy E.Coli protein sequences (NC_00913.faa)

   b. Now perform 'formatdb' in the BLAST directory
      >formatdb -i NC_000913.faa -n EColi -p T
   c. You will see these files created in the same directory.
      EColi.pin, EColi.psq, EColi.phr, formatdb.log

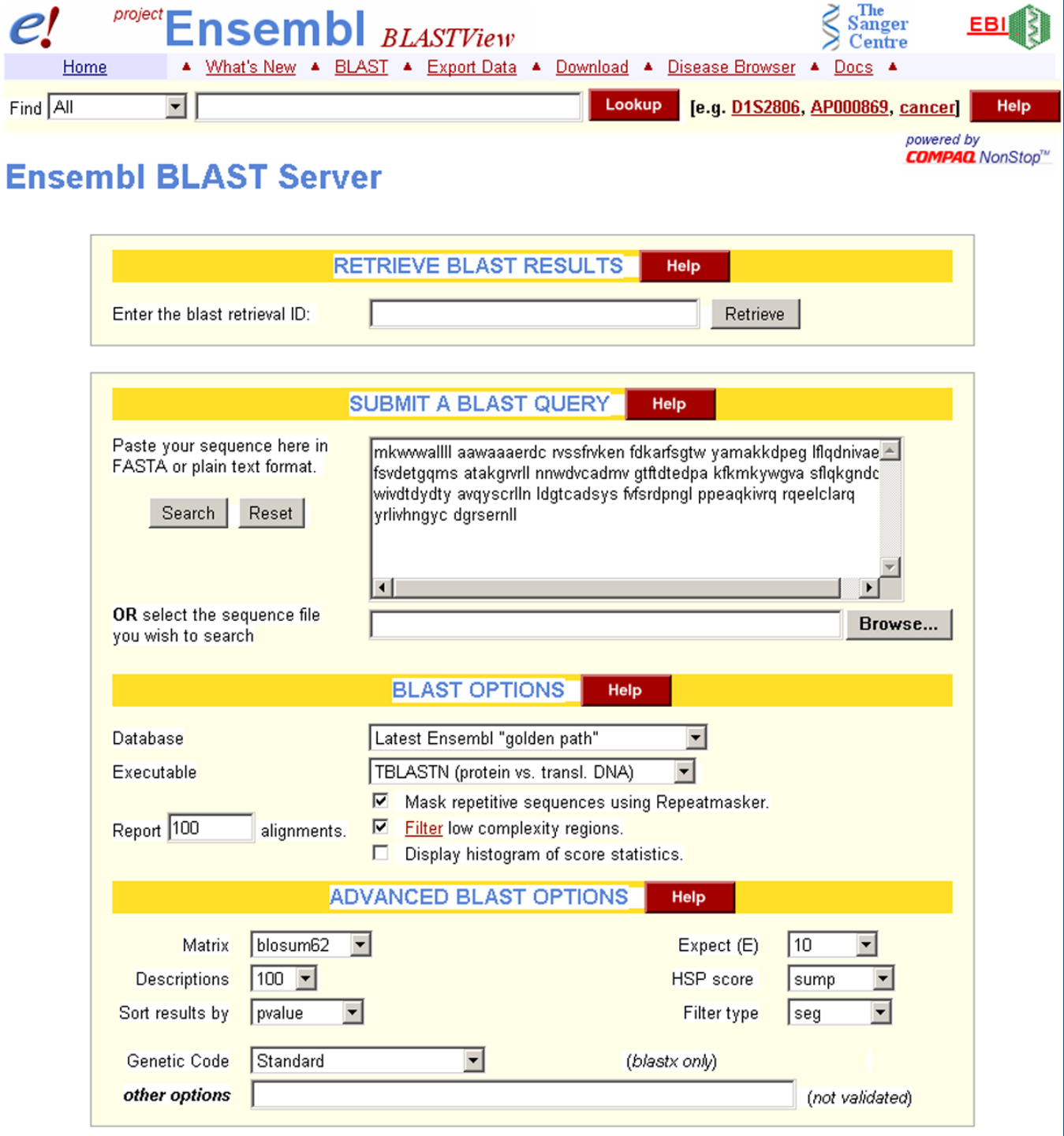
3. Let's perform a simple BLAST of "proteinSeq1.txt"
   a. Copy the "proteinSeq1.txt" into the BLAST directory.

   b. >blastall -p blastp -d EColi -i proteinSeq1.txt -o proteinSeq1.out
      blastall -p blastp -d EColi -i proteinSeq1.txt

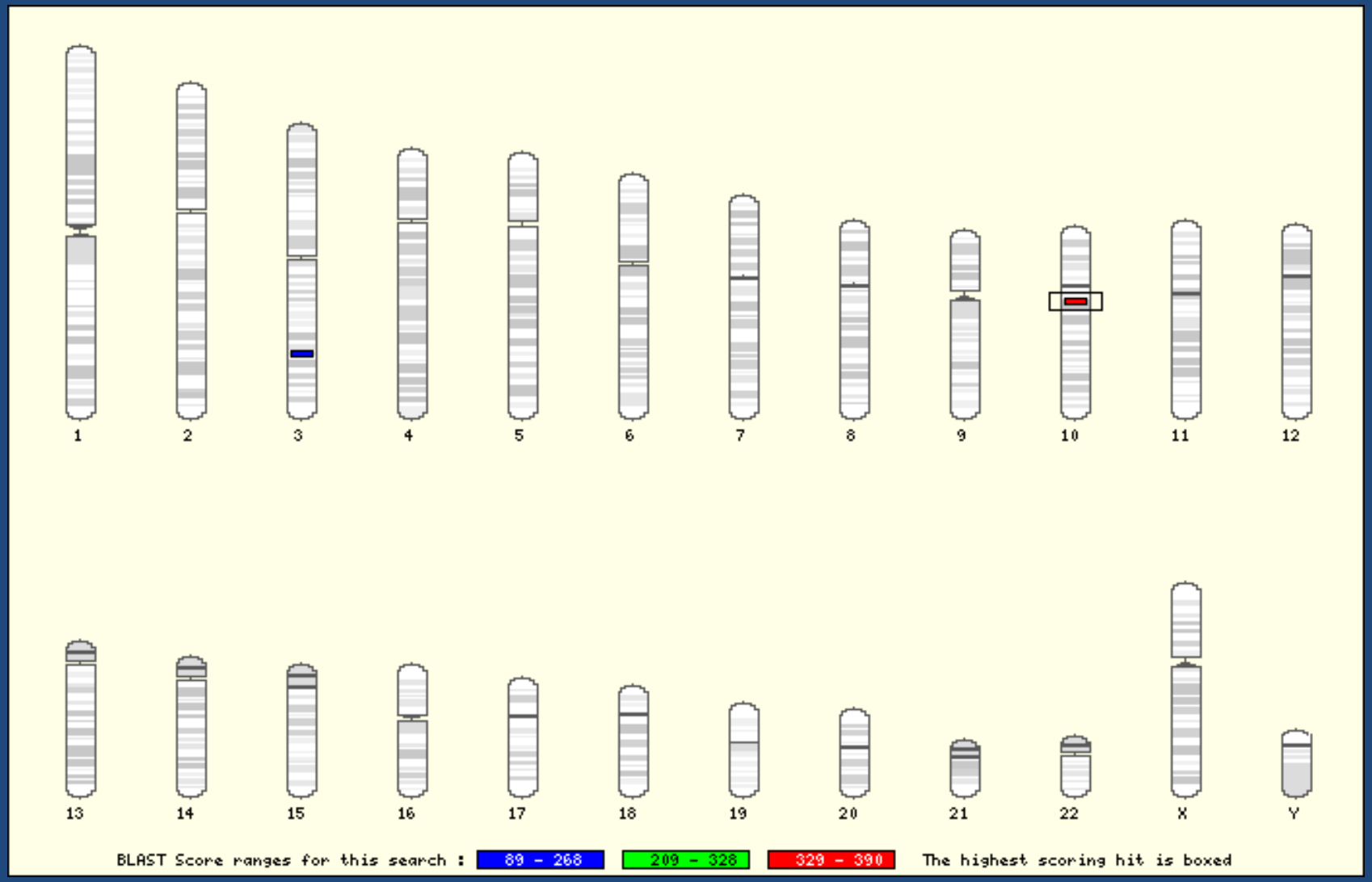
4. Change the following options
   A. -e : expectation value (Default: 10)
   B. -m : alignment view option (Default: 0)
   C. -b : Number of databse sequences to show alignments (Default: 250)
   D. -v : Number of database sequences to show one-line descriptor (Default: 500)
   E. -g : Perform gapped alignment (Default: T)
   F. -M : Scoring Matrix (Default: BLOSUM62)

5. There are many options you can adjust. Simply run blastall without any option.

6. Try to make BLAST print out result in html (with -T T)
   >blastall -p blastp -d EColi -i proteinSeq1.txt -o //index.html -T T


project Ensembl BLASTView
The Sanger Centre
EBI
Home
What's New
BLAST
Export Data
Download
Disease Browser
Docs
Find All
Lookup
[e.g. D1S2806, AP000869, cancer]
Help
powered by COMPAQ NonStop™
Ensembl BLAST Server
RETRIEVE BLAST RESULTS
Help
Enter the blast retrieval ID:
Retrieve
SUBMIT A BLAST QUERY
Help
Paste your sequence here in FASTA or plain text format.
mkwwwallll aawaaaerdc rvssfrvken fdkarfsgtw yamakkdpeg lflqdnivae
fsvdetgqms atakgrvrll nnwdvcadmv gtftdtedpa kfkmkywgva sflqkgndc
wivdtdydty avqyscrlln ldgtcadsys fvfsrdpngl ppeaqkivrq rqeelclarq
yrlivhngyc dgrsernll
Search
Reset
OR select the sequence file you wish to search
Browse...
BLAST OPTIONS
Help
Database
Latest Ensembl "golden path"
Executable
TBLASTN (protein vs. transl. DNA)
Report 100 alignments.
Mask repetitive sequences using Repeatmasker.
Filter low complexity regions.
Display histogram of score statistics.
ADVANCED BLAST OPTIONS
Help
Matrix
blosum62
Expect (E)
10
Descriptions
100
HSP score
sump
Sort results by
pvalue
Filter type
seg
Genetic Code
Standard
(blastx only)
other options
(not validated)

Ensembl BLAST output includes an ideogram



Program: tblastn

Database:
- Porcine (Sus scrofa)
- African clawed frog (Xenopus laevis)
- Zebrafish (Danio rerio)
- ----------
- Arabidopsis thaliana
- Barley (Hordeum vulgare)
- Ice Plant (Mesembryanthemum crystallinum)
- Maize (Zea mays)
- Medicago truncatula
- Potato (Solanum tuberosum)

Alignments: 20

Upload a file containing a sequence OR paste it into the textbox:
(Note: If both are entered, the file will be ignored.)

Enter the name of the file containing a sequence in FASTA or raw format:

Browse...

Enter your sequence in FASTA or raw format:

```
MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMS
ATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLN
LDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
```

Sequence identifier (fo raw sequence entered):

blosum62
- blosum62
- blosum100
- blosum30
- blosum35
- blosum40
- blosum45
- blosum50
- blosum55
- blosum60
- blosum65
- blosum70
- blosum75
- blosum80
- blosum85
- blosum90
- blosumn
- dayhoff
- gonnet
- identity
- match

Options:

Matrix: blosum62

Filter: default

Expect: 10

Cutoff: default

Strand: default

Descriptions: 20

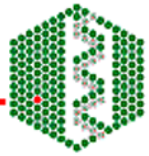
Wordlength (for blastn only): default

Echofilter ☐

Graphical Overview ☑

Ignore Hypotheticals ☐

EMBL
European Bioinformatics Institute



WU-Blast2 Help | Tools | EBI Home | RUN BLAST | RESET FORM

YOUR EMAIL	SEARCH TITLE	RESULTS	DATABASE	PROGRAM
	Sequence	interactive	swall	WU-blastp
MATRIX	**DNA STRAND**	**EXP.THR**	**FILTER**	**VIEW FILTER**
blosum62	none	default	none	no
HISTOGRAM	**STATS**	**SORT**	**SCORES**	**ALIGNMENTS**
no	sump	pvalue	default	default

Enter or Paste a PROTEIN Sequence in any format:

Upload a file: Browse...

RUN BLAST | RESET FORM

This document was last modified on : Thursday, July 05, 2001 10:52:03
Comments or suggestions support@ebi.ac.uk

If you plan to use these services during a course please contact us using the email above.

BLAST-related tools for genomic DNA

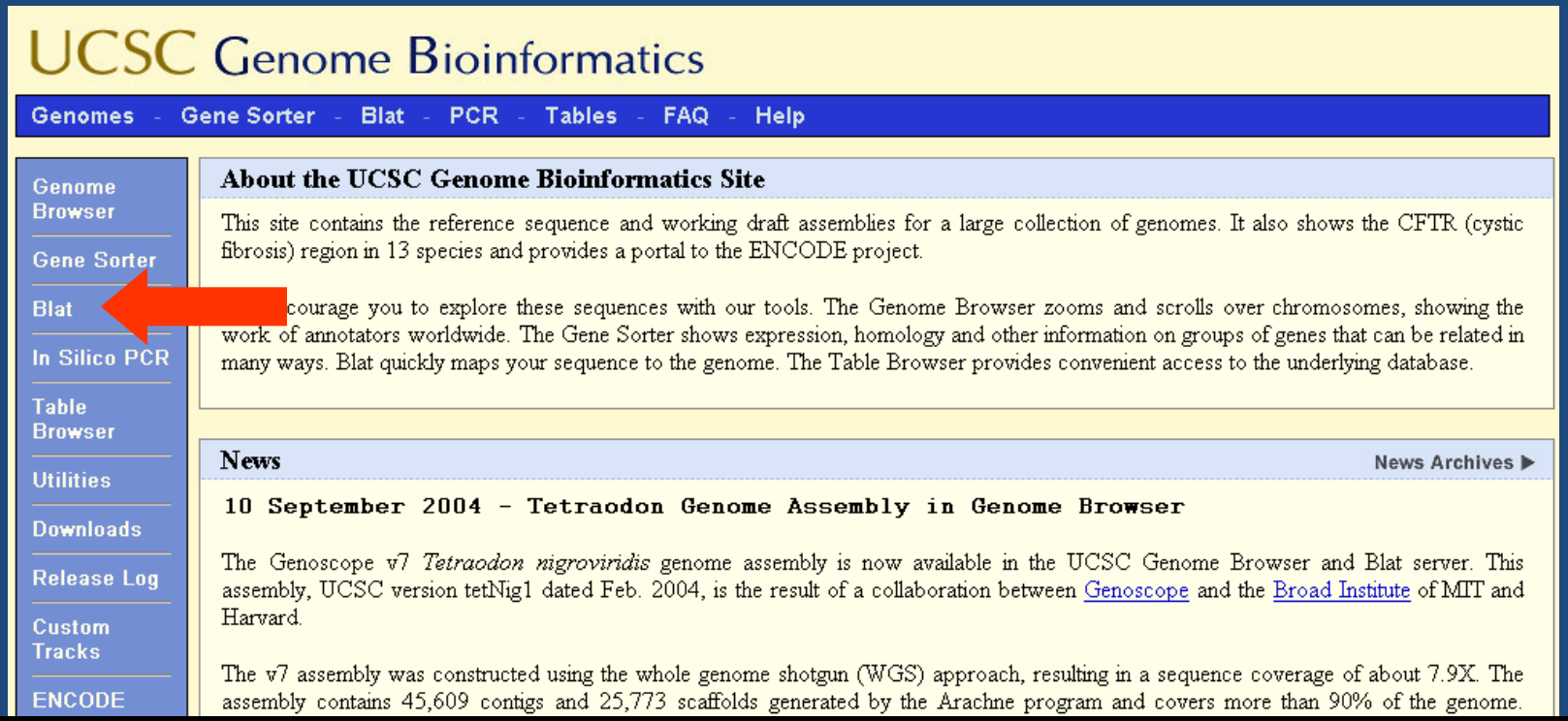
Recently developed tools include:

- MegaBLAST at NCBI.

- BLAT (BLAST-like alignment tool). BLAT parses an entire genomic DNA database into words (11mers), then searches them against a query. Thus it is a mirror image of the BLAST strategy. See http://genome.ucsc.edu

- SSAHA at Ensembl uses a similar strategy as BLAT. See http://www.ensembl.org

To access BLAT, visit http://genome.ucsc.edu



"BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 20 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates." --BLAT website

Home - Genomes - Gene Sorter - Blat - Tables - FAQ - Help

Human BLAT Search

BLAT Search Genome

Genome: Human | Assembly: July 2003 | Query type: BLAT's guess | Sort output: query,score | Output type: hyperlink | Submit

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched at once if separated by a line starting with > followed by the sequence name. Reset

```
>gi|8400727|ref|NM_006744.2| Homo sapiens retinol binding protein 4, plasma
(RBP4), mRNA
CGCTCGCCTCCCTCGCTCCACGCGCGCCCGGACGCGGCGGCCAGGCTTGCGCGTGGTTCCCCTCCCGGTG
GGCGGATTCCTGGGCAAGATGAAGTGGGTGTGGGCGCTCTTGCTGTTGGCGGCGTGGGCAGCGGCCGAGC
GCGACTGCCGAGTGAGCAGCTTCCGAGTCAAGGAGAACTTCGACAAGGCTCGCTTCTCTGGGACC
CGCCATGGCCAAGAAGGACCCCGAGGGCCTCTTTCTGCAGGACAACATCGTCGCGGAGTTCTCC
GAGACCGGCCAGATGAGCGCCACAGCCAAGGGCCGAGTCCGTCTTTTGAATAACTGGGACGTGTGCGCAG
ACATGGTGGGCACCTTCACAGACACCGAGGACCCTGCCAAGTTCAAGATGAAGTACTGGGGCGTAGCCTC
CTTTCTGCAGAAAGGAAATGATGACCACTGGATCGTCGACACAGACTACGACACGTATGCCGTACAGTAC
TCCTGCCGCCTCCTGAACCTCGATGGCACCTGTGCTGACAGCTACTCCTTCGTGTTTTCCCGGGACCCCA
ACGGCCTGCCCCCAGAAGCGCAGAAGATTGTAAGGCAGCGGCAGGAGGAGCTGTGCCTGGCCAGGCAGTA
CAGGCTGATCGTCCACAACGGTTACTGCGATGGCAGATCAGAAAGAAACCTTTTGTAGCAATATCAAGAA
TCTAGTTTCATCTGAGAACTTCTGATTAGCTCTCAGTCTTCAGCTCTATTTATCTTAGGAGTTTAATTTG
CCCTTCTCTCCCCATCTTCCCTCAGTTCCCATAAAACCTTCATTACACATAAAGATACACGTGGGGGTCA
```

Paste DNA or protein sequence here in the FASTA format

Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

Upload sequence: Browse... Submit File

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 5000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 12,500 letters.

About BLAT

BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 22 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates.

BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers except for those heavily involved in repeats. The index takes up a bit less than a gigabyte of RAM. The genome itself is not kept in memory, allowing BLAT to deliver high performance on a reasonably priced Linux box. The index is used to find areas of probable homology, which are then loaded into memory for a detailed alignment.

BLAT output includes browser and other formats

Home - Genomes - Gene Sorter - Blat - Tables - FAQ - Help

Human BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	NM_006744.2	902	1	919	919	99.5%	10	-	95016188	95025584	9397
browser details	NM_006744.2	21	887	909	919	86.4%	9	-	77698017	77698038	22

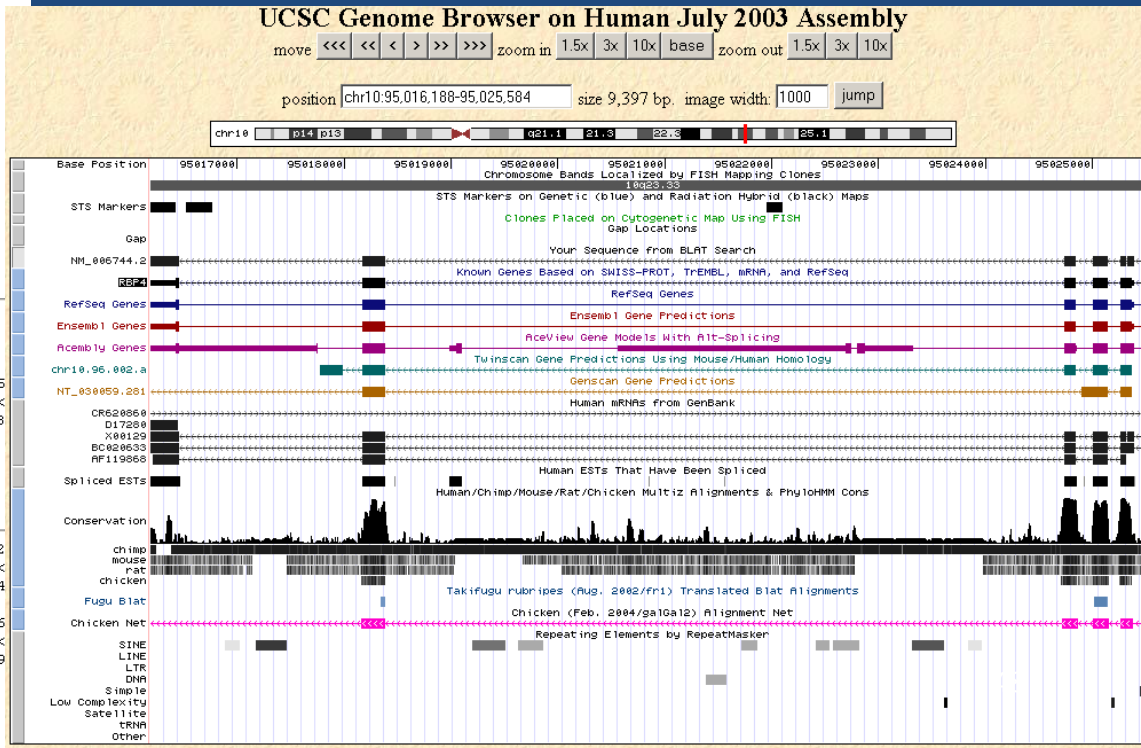
Alignment of NM_006744.2

NM_006744.2
Human.chr10
block1
block2
block3
block4
block5
block6
together

```
ttggtcacga actgacccac tacaccaaac agatgggaac ttagaccaaa  95017085
atcaaattga ttttgaagct tcacccttca aaattaaatg tagggccagc  95017035
aaggattcca ccacagcagg gcaccgtagg aagcagatta tgcacattgt  95016985
acagatgaga aaacagaggc tgacaatgag gaagaaagtt gccctaaatc  95016935
tttaataggt agctgagttg gatgcaaacc tagagttttg tgactctgaa  95016885
gtcccttgtg ccacaccatc tggtgacaca tgacatgaga catagaagca  95016835
ctttataaaa ccatcaccag tgtgtcccag gtcaggctgc cttggctccc  95016785
acttcctgaa agctgagggt ggcctccgac actttgaaat gcaataaggc  95016735
agacttaaga gaaatcaaag gaagccctgt tcatccagca agtcatattc  95016685
tctccacccc cattacgtcc agagaaaatt cagtgggttt cagaaacagc  95016635
cttgaaggtg tttatgaatt acagccacct gtatccgtaa gggttgcaat  95016585
ctgctgtgat gcatttgaat taagaagcat ttgaatgaag ccagctctaa  95016535
ggccatcatg atgttcatta tctacataaa aactgggtcc actttctgtc  95016485
gttaatggct ttgttttgta ttttccagGT TACTGCGATG GCAGATCAGA  95016435
AAGAAACCTT TTGTAGCAAT ATCAAGAATC TAGTTTCATC TGAGAACTTC  95016385
TGATTAGCTC TCAGTCTTCA GCTCTATTTA TCTTAGGAGT TTAATTTGCC  95016335
CTTCTCTCCC CATCTTCCCT CAGTTCCCAT AAAACCTTCA TTACACATAA  95016285
AGATACACGT GGGGGTCAGT GAATCTGCTT GCCTTTCCTG AAAGTTTCTG  95016235
GGGCTTAAGA TTCCAGACTC TGATTCATTA AACTATAGTC ACCCGTGtcc  95016185
tgtgatttta gttttcattt gtgtttatgt ctgtgctgca gacggatggg  95016135
tggggtgcgc ttctttatac caggagcacg tggctctttc tgacctt
```

Side by Side Alignment

```
00000001 cgctcgcctcccctcgctccacgcgcgcccggacgcggcggccaggcttgc 0000005
>>>>>>>> ||||||||||||||||||||||||||||||||||||| ||||||||||||| <<<<<<<
95025584 cgctcgcctcccctcgctccacgcgcgcccggactcggcggccaggcttgc 9502553

00000051 gcgtggttccccctcccggtg 00000070
>>>>>>>> ||| |||||||||||||||| <<<<<<<<
95025534 gcgcggttccccctcccggtg 95025515

00000071 ggcggattcctgggcaagatgaagtgggtgtgggcgctcttgctgttggc 0000012
>>>>>>>> |||||||||||||||||||||||||||||||||||||||||||||||||| <<<<<<<
95025390 ggcggattcctgggcaagatgaagtgggtgtgggcgctcttgctgttggc 9502534

00000121 ggcg.tgggcagcggcc.....gagcgcgactgccgagtgagcagcttcc 0000016
>>>>>>>> |||| |||||||||||||     |||||||||||||||||||||||||| <<<<<<<
95025340 ggcgctgggcagcggccgcgcggagcgcgactgccgagtgagcagcttcc 9502529

00000165 gagtcaaggagaacttcgacaaggctcgc 00000193
>>>>>>>> ||||||||||||||||||||||||||||| <<<<<<<<
95025290 gagtcaaggagaacttcgacaaggctcgc 95025262
```



Multiple sequence alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWWSNG--
```

Like pairwise alignment BUT compare ***n*** sequences instead of ***2***

Rows represent individual sequences
Columns represent ‘same’ position

May be gaps in some sequences

MSA & Evolution

MSA can give you a picture of the forces that shape evolution!

- Important amino acids or nucleotides are not "allowed" to mutate
- Less important positions change more easily

Conserved positions

- Columns where all the sequences contain the same amino acids or nucleotides
- Important for the function or structure

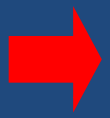
```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGSSSNIGS--ITVNWYQQLPG
LRLSCTGSGFIFSS--YAMYWYQQAPG
LSLTCTGSGTSFDD-QYYSTWYQQPPG
```

Consensus Sequence

- The consensus sequence holds the most frequent character of the alignment at each column

Profile

A	T	C	T	T	G	T
A	A	C	T	T	G	T
A	A	C	T	T	C	T



	1	2	3	4	5	6
A	1	0.67	0	0	.	.
T	0	0.33	1	1	.	.
C	0	0	0	0	.	.
G	0	0	0	0	.	.

Profile =

PSSM – Position Specific Score Matrix

Alignment methods

- Progressive alignment (Clustal)
- Iterative alignment (mafft, muscle)
- All methods today are an approximation strategy (**heuristic algorithm**), yield a possible alignment, but not necessarily the best one

Progressive alignment

First step:

Compute the pairwise alignments for all against all (6 pairwise alignments) the similarities are stored in a table

	A	B	C	D
A				
B	11			
C	3	1		
D	2	2	10	

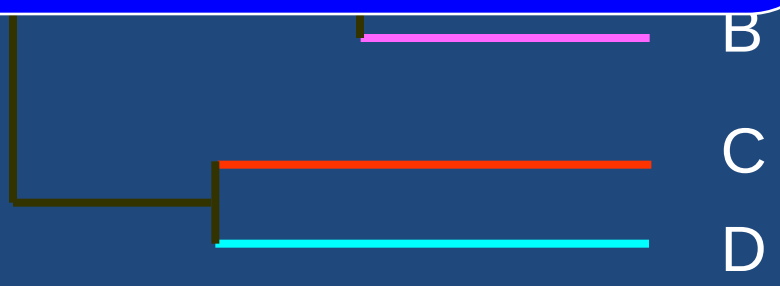
Second step:

	A	B	C	D
A				
B	11			
C	3	1		

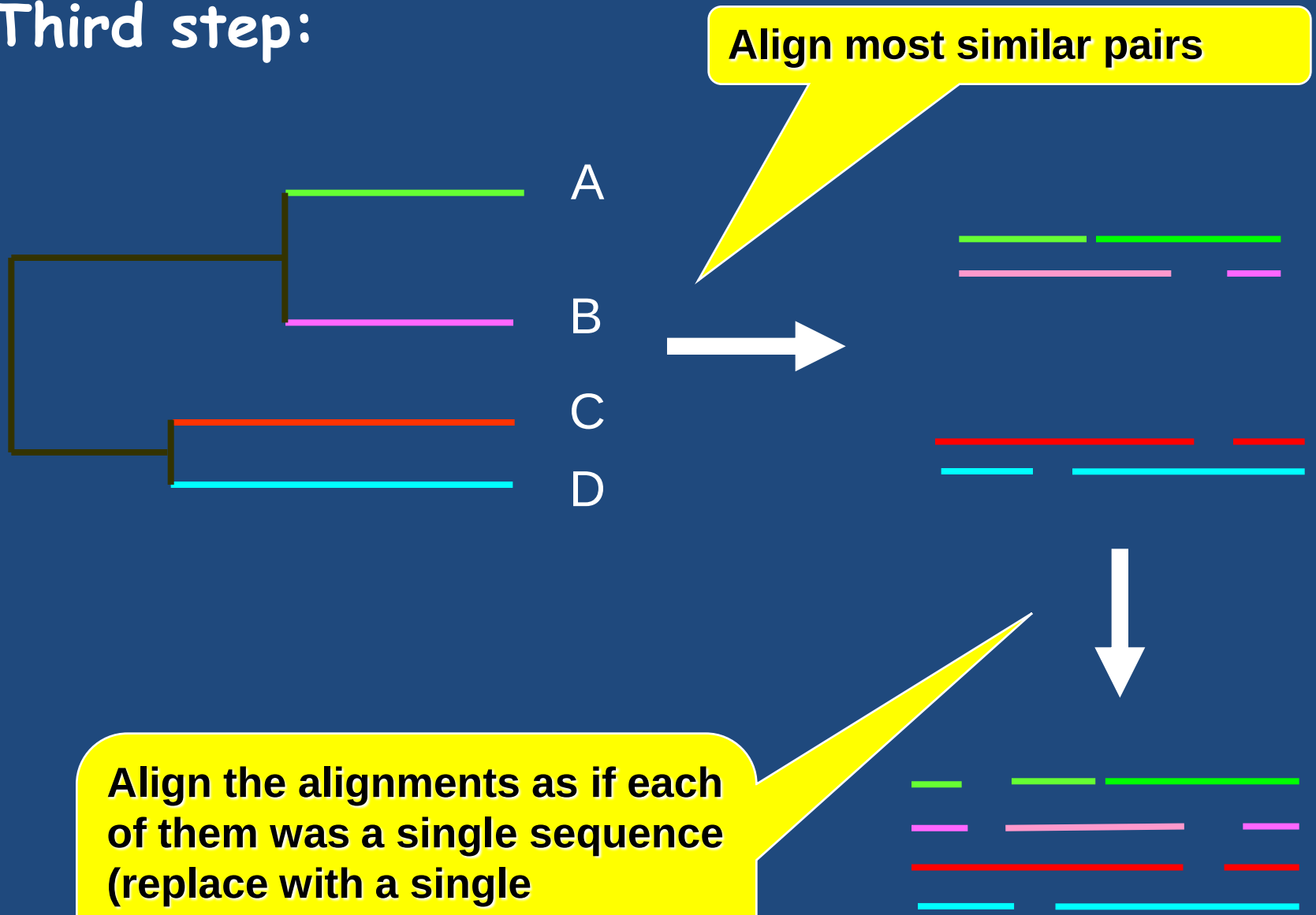
cluster the sequences to create a tree (guide tree):

- Represents the ord
  sequences are to be
- similar sequences a
  tree
- distant sequences a
  other in the tree

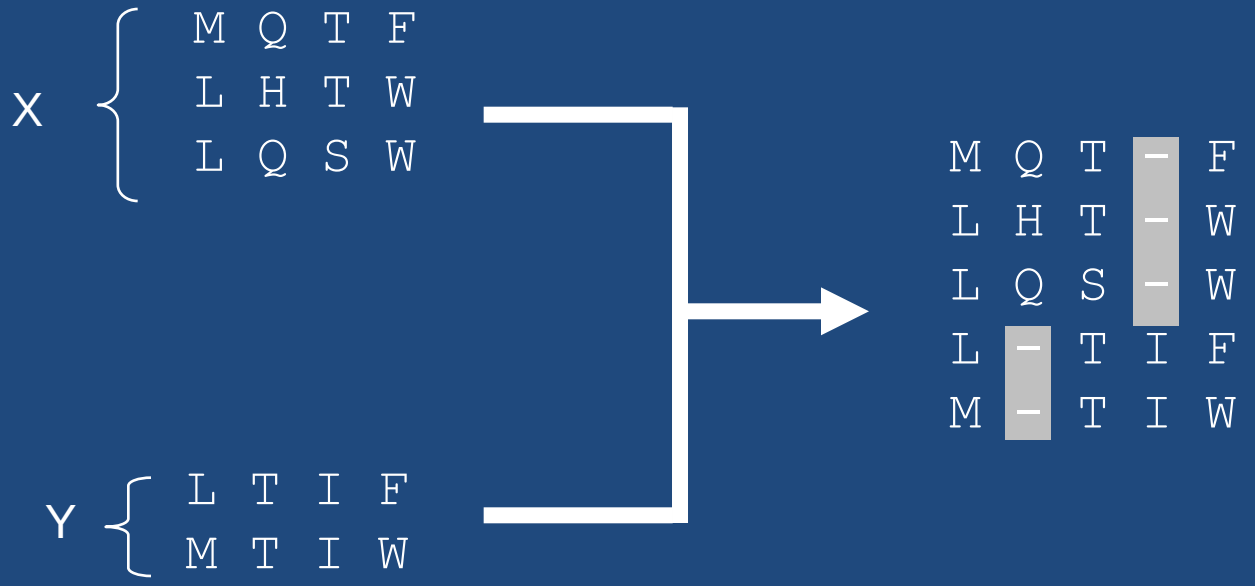
The guide tree is imprecise and is NOT the tree which truly describes the relationship between the sequences!



Third step:


Align most similar pairs
A
B
C
D
Align the alignments as if each of them was a single sequence (replace with a single consensus sequence or use a profile)

Alignment of alignments



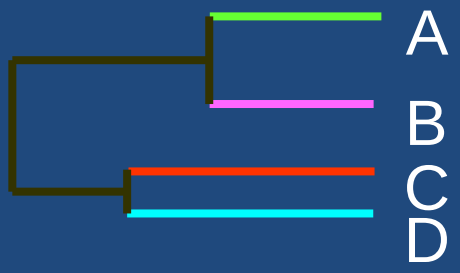
Iterative alignment

Pairwise distance table

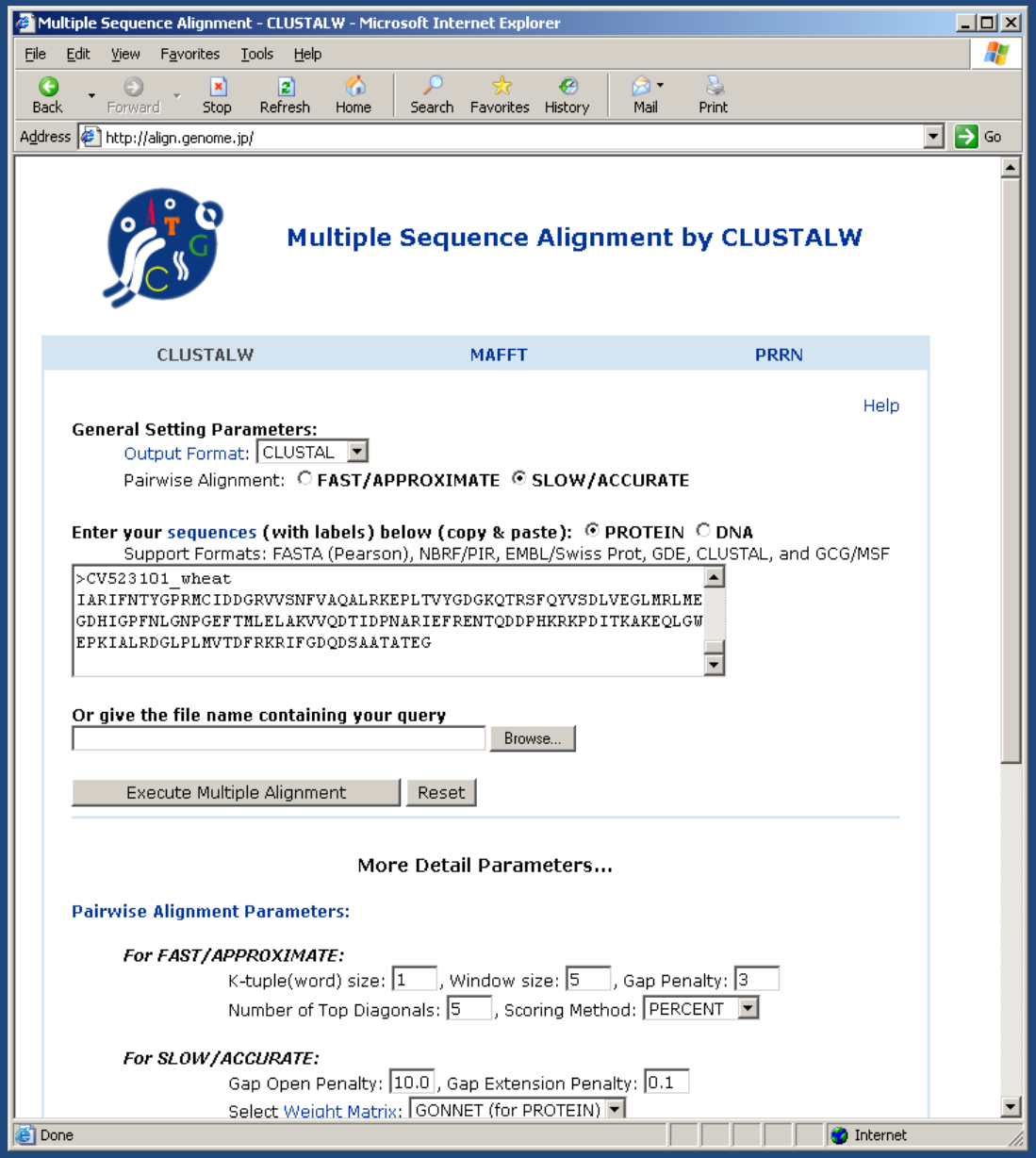
	A	B	C	D
A				
B	11			
C	3	1		
D	2	2	10	

Iterate until the MSA doesn't change

Guide tree



MSA



Online version

- >gi|115023|sp|P10425| MKKNTLLKVGLCVSLLGTTQFVSTISSVQAS QKVEQIVIKNETGTISISQLNKNVW VHTELGYFNGEAVPSNGLVLNTSKGL VLVDSSWDNKLTKELIEMVEKKFQKRVTD VIITHAHADRIGGITALKERGIK AHSTALTAELAKKSGYEEPLGDLQTVTNLKFGNTK VETFYPGKGHTEDNIV VWLPQYQILAGGCLVKSAEAKNLGNVADAYVNEWSTSIE NMLKRYRNINL VVPGHGKVGDKGLLLHTLDLLK >gi|115030|sp|P25910| MKTVFILIS MLFPVAVMAQKSVKISDDISITQLSDKVYTYVSLAEIEGWGMVPSNGM IVI NNHQAALLDTPINDAQTEMLVNWVTDSLHAKVTTFIPNHWHGDCIGGLG YLQR KGVQSYANQMTIDLAKEKGLPVPEHGFTDSLTVSLDGMPLQCYYLG GGHATDNIV VWLPTENILFGGCMLKDNQATSIGNISDADVTAWPKTLDK VKAKFPSARYVVPGH GDYGGTELIEHTKQIVNQYIESTSKP >gi|282554| pir||S25844
  MTVEVREVAEGVYAYEQAPGGWCVSNAGIVVGGDGALVVDTLSTIPRAR RLAEWV DKLAAGPGRTVVNTH
FHGDHAFGNQVFAPGTRIIAHEDMRSAMVTTGLALTGLWP RVDWGEIEL RPPNVTFRDRLTLHVGERQVE
LICVGPAHTDHDVVVWLPEERVLFAGD VVMSGVTPFALFGSVAGTLAALD RLAELEPEVVVGGHGPVAGP
EVIDANRDYLRWV QRLAADAVDRRLTPLQAARRADLGAFAGLLDAERLVA NLHRAHEELLGGHVRDAM EI FAELVAYNGGQLPTCLA

An output from ClustalW

sequences have significant similarity

```
CLUSTAL W (1.82) multiple sequence alignment
gi|42542791|gb|AAH66228.1|     MSTAGKVIKCKAAVLWELKKPFSIEEVEVAPPKAHEVRIKMVAAGICRS- 49
gi|825623|emb|CAA39813.1|      MGTKGKVIKCKAAIAWEAGKPLCIEEVEVAPPKAHEVRIQIIATSLCHT- 49
gi|42738724|gb|AAS42652.1|     --MQNFVFRNPTKLIFGKGQ---LEQLKTEIPQFGKKVLLVYGGGSIKRN 45
                               . *:: : : :  :  :*:::. *: : : : . .  : .

gi|42542791|gb|AAH66228.1|     ---DEHVVSGNLV-TPLPVILGHEAAGIVESVGEGVTTVKPG--DKVIPL 93
gi|825623|emb|CAA39813.1|      ---DASVIDSKFEGLAFPVIVGHEAAGIVESIGPGVTNVKPG--DKVIPL 94
gi|42738724|gb|AAS42652.1|     GIYDNVISILKDINAEVFELTGVEPNPRVSTVKKGIQICKDNGVEFILAV 95
                               . * :  : . . : * *.  *.:: *:  * .. : ::. :

gi|42542791|gb|AAH66228.1|     FTPQCGKCRICKNPESNYCLKN-DLGNPRG----------T 123
gi|825623|emb|CAA39813.1|      YAPLCRKCKFCLSPLTNLCGKISNLKSPASDQ----------QL 128
gi|42738724|gb|AAS42652.1|     GGGSVIDCTKAIAAGSKYDGDVWDIVTKKAFASEALPFGTVLTLAATGSE 145
                               .*  .  . ::   .  :: .  .  :.: . .:  : :::..

gi|42542791|gb|AAH66228.1|     LQDGTRRFTCSGKPIHHFVGVSTFSQYTVVDENAVAKIDAASPLEKVCLI 173
gi|825623|emb|CAA39813.1|      MEDKTSRFTCKGKPVYHFFGTSTFSQYTVVSDINLAKIDDDANLERVCLL 178
gi|42738724|gb|AAS42652.1|     MNAGSVITNWETNEKYGWGSPVTFPQFSILDPVHTASVPRDQTIYGMVDI 195
                               :: :   . . : : : . **.*:::.   *.:    : : :
```

alcohol dehydrogenase, iron-containing [Bacillus cereus

Class I alcohol dehydrogenase, gamma subunit [Homo sapiens]

Different form of alcohol dehydrogenase [Homo sapiens]

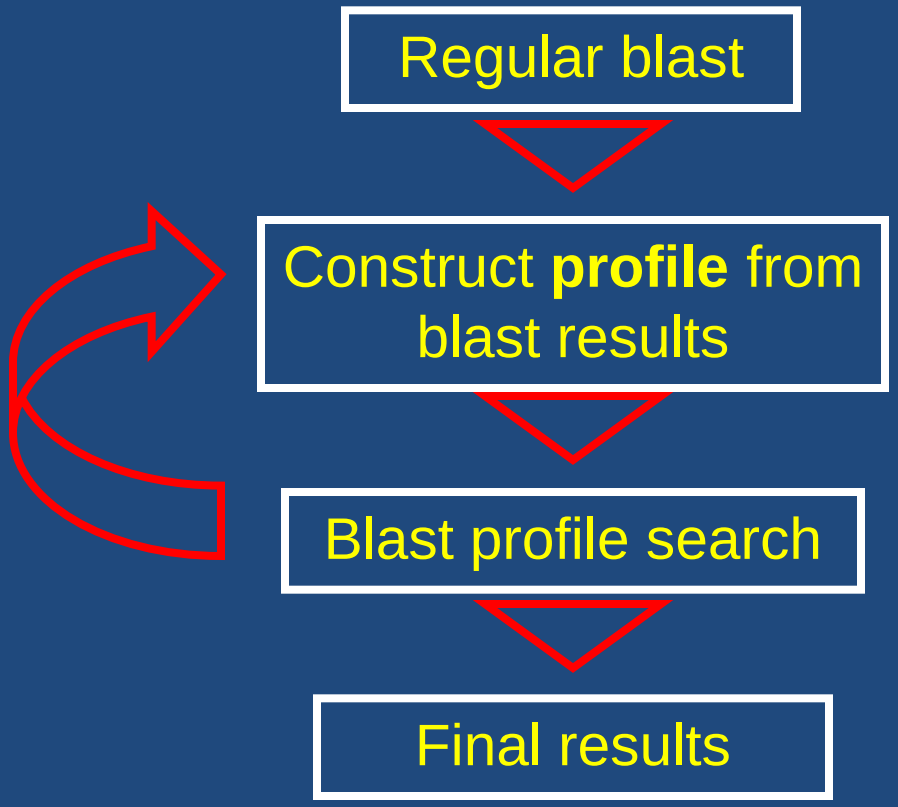
Searching for remote homologs

- Sometimes BLAST isn't enough.
- Large protein family, and BLAST only gives close members. We want more distant members

- PSI-BLAST

PSI-BLAST

- **P**osition **S**pecific **I**terated BLAST



Position specific iterated BLAST: PSI-BLAST

The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query.

PSI-BLAST is performed in five steps

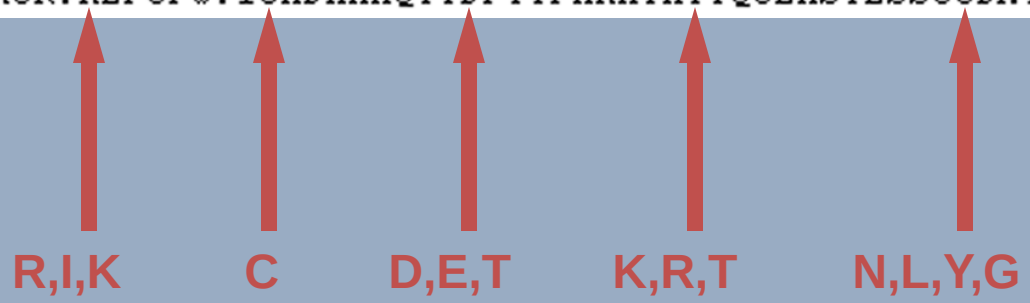
[1] Select a query and search it against a protein database

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

```
730496     66   FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDH 125
200679     63   FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 122
206589     34   FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 93
2136812    2            MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 53
132408     65   FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124
267584     44   FSVDESGKVTATAHGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAASYLQTGNDDH 103
267585     44   FSVDGSGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 103
8777608    63   FTIHEDGAMTATAKGRVIILNNWEMCADMMATFETTPDPAKFRMRYWGAASYLQTGNDDH 122
6687453    60   FKVEEDGTMTATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMKYWGAAAYLQTGYDDH 119
10697027   81   FKVQEDGTMTATATGRVIILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTGYDDH 140
13645517   1                                MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 32
13925316   38   FSVDGSGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 97
131649     65   YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY 126
```



PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)

	☑	gi|6978523|ref|NP_036909.1|	apolipoprotein D [Rattus norvegicus]...	147	4e-35
	☑	gi|1542847|dbj|BAA13453.1|	(D87752) alpha1-microglobulin/bikunin...	144	6e-34
	☑	gi|619383|gb|AAB32200.1|	apolipoprotein D, apoD [human, plasma, ...	143	8e-34
	☑	gi|5419892|emb|CAB46489.1|	(X02824) RBP (aa 101-172) [Homo sapiens]	139	1e-32
	☑	gi|4502163|ref|NP_001638.1|	apolipoprotein D precursor [Homo sap...	138	4e-32
	☑	gi|584763|sp|P37153|APD_RABIT	APOLIPOPROTEIN D PRECURSOR >gi|482...	134	4e-31
	☑	gi|1703341|sp|P51909|APD_CAVPO	APOLIPOPROTEIN D PRECURSOR >gi|11...	133	7e-31
	☑	gi|2895204|gb|AAC02945.1|	(AF025334) mutant retinol binding prot...	80	9e-15
	☑	gi|1246096|gb|AAB35919.1|	(S80440) apolipoprotein D, apoD (C-ter...	77	8e-14
	☑	gi|2895206|gb|AAC02946.1|	(AF025335) mutant retinol binding prot...	67	8e-11
NEW	☑	gi|1346419|sp|P49291|LAZA_SCHAM	LAZARILLO PROTEIN PRECURSOR >gi|...	63	1e-09
NEW	☑	gi|2506821|sp|P00978|AMBP_BOVIN	AMBP PROTEIN PRECURSOR [CONTAINS...	63	2e-09
NEW	☑	gi|2497696|sp|Q07456|AMBP_MOUSE	AMBP PROTEIN PRECURSOR [CONTAINS...	63	2e-09
NEW	☑	gi|6680684|ref|NP_031469.1|	alpha 1 microglobulin/bikunin [Mus m...	62	2e-09
NEW	☑	gi|12836446|dbj|BAB23659.1|	(AK004907) putative [Mus musculus]	62	3e-09
NEW	☑	gi|6978497|ref|NP_037033.1|	alpha-1 microglobulin/bikunin [Rattu...	62	3e-09
NEW	☑	gi|2507586|sp|P04366|AMBP_PIG	AMBP PROTEIN PRECURSOR [CONTAINS: ...	61	8e-09
NEW	☑	gi|1085207|pir||JC2556	alpha-1-microglobulin/inter-alpha-trypsin...	60	1e-08
NEW	☑	gi|2988354|dbj|BAA25305.1|	(AB006444) alpha-1-microglobulin/biku...	59	2e-08
NEW	☑	gi|108233|pir||S13493	alpha-1-microglobulin - pig	59	2e-08
NEW	☑	gi|1882|emb|CAA36306.1|	(X52087) precursor codes for two protein...	59	2e-08
NEW	☑	gi|9181923|gb|AAF85707.1|AF276505_1	(AF276505) neural Lazarillo ...	59	3e-08
NEW	☑	gi|7296083|gb|AAF51378.1|	(AE003586) NLaz gene product [Drosophi...	58	3e-08
NEW	☑	gi|117330|sp|P80007|CRA2_HOMGA	CRUSTACYANIN A2 SUBUNIT >gi|10275...	57	8e-08
NEW	☑	gi|2497695|sp|Q60559|AMBP_MESAU	AMBP PROTEIN PRECURSOR [CONTAINS...	57	1e-07
NEW	☑	gi|102968|pir||S22400	insecticyanin A - tobacco hornworm >gi|971...	56	1e-07
NEW	☑	gi|4502067|ref|NP_001624.1|	alpha-1-microglobulin/bikunin precur...	56	2e-07
NEW	☑	gi|1146408|gb|AAA85089.1|	(L41641) gallerin [Galleria mellonella]	56	2e-07
NEW	☑	gi|2497694|sp|Q62577|AMBP_MERUN	AMBP PROTEIN PRECURSOR [CONTAINS...	55	3e-07
NEW	☑	gi|1213589|dbj|BAA12075.1|	(D83712) Prostaglandin D Synthase [Xe...	54	5e-07
	☑	gi|539717|pir||A61233	retinol-binding protein - cat (fragment)	54	8e-07
NEW	☑	gi|266472|sp|Q01584|LIPO_BUFMA	LIPOCALIN PRECURSOR >gi|104284|pi...	53	1e-06
	☑	gi|265042|gb|AAB25283.1|	retinol-binding protein, RBP (N-termina...	52	3e-06
NEW	☑	gi|1079295|pir||S52354	gene cpl-1 protein - African clawed frog ...	52	3e-06
NEW	☑	gi|732003|sp|P39281|BLC_ECOLI	OUTER MEMBRANE LIPOPROTEIN BLC PRE...	51	9e-06

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)

[5] Repeat steps [3] and [4] iteratively, typically 5 times. At each new search, a new profile is used as the query.

Results of a PSI-BLAST search

Iteration	# hits	# hits > threshold
1	104	49
2	173	96
3	236	178
4	301	240
5	344	283
6	342	298
7	378	310
8	382	320

PSI-BLAST alignment of RBP and β-lactoglobulin: iteration 1

```
Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
           V+ENFD  ++ G WY + +K P      + I A +S+ E G +    K          ++
Sbjct: 33  VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIEVLNK---------ELS 82

Query: 87  ADMVGTF---------TDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCR 137
            D  GT          ++  +PAK +++++ +          +WI+ TDY+ YA+ YSC
Sbjct: 83  PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
               L ++D      + ++  R+P  LPPE
Sbjct: 136 TFFWLFHVD------FFWILGRNPY-LPPE 158
```

PSI-BLAST alignment of RBP and β-lactoglobulin: iteration 2

```
Score =  140 bits (353), Expect = 1e-32
Identities = 45/176 (25%), Positives = 78/176 (43%), Gaps = 33/176 (18%)

Query: 4   VWALLLLAAWAAAERDCRVSSF--------RVKENFDKARFSGTWYAMAKKDPEGLFLQD 55
           V  L+ LA  A      +  +F          V+ENFD  ++ G WY + +K P      +
Sbjct: 2   VTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEI-EKIPASFEKGN 60

Query: 56  NIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMV---GTFTDTEDPAKFKMKYWGVASF 112
            I A +S+ E G +    K       + D   + V      ++  +PAK +++++ +
Sbjct: 61  CIQANYSLMENGNIEVLNKEL-----SPDGTMNQVKGEAKQSNVSEPAKLEVQFFPL--- 112

Query: 113 LQKGNDDHWIVDTDYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA 164
                +WI+ TDY+ YA+ YSC     L ++D       + ++  R+P  LPPE
Sbjct: 113 --MPPAPYWILATDYENYALVYSCTTFFWLFHVD------FFWILGRNPY-LPPET 159
```

PSI-BLAST alignment of RBP and β-lactoglobulin: iteration 3

```
Score =  159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3   WVWALLLLAAWAAAERD--------CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
            V  L+ LA  A               +  S  V+ENFD  ++ G WY + K
Sbjct: 1   MVTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIPASFE-KG 59

Query: 55  DNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQ 114
            + I A +S+ E G +    K           V  +     ++  +PAK +++++ +
Sbjct: 60  NCIQANYSLMENGNIEVLNKELSPDGTMNQVKGE--AKQSNVSEPAKLEVQFFPL----- 112

Query: 115 KGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEA 164
                +WI+ TDY+ YA+ YSC             + ++  R+P  LPPE
Sbjct: 113 MPPAPYWILATDYENYALVYSCTTFFWL--FHVDFFWILGRNPY-LPPET 159
```

1

```
Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
           V+ENFD  ++ G WY + +K P      + I A +S+ E G +    K          ++
Sbjct: 33  VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIEVLNK---------ELS 82

Query: 87  ADMVGTF---------TDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCR 137
            D  GT          ++  +PAK +++++ +          +WI+ TDY+ YA+ YSC
Sbjct: 83  PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
               L ++D      + ++  R+P  LPPE
Sbjct: 136 TFFWLFHVD------FFWILGRNPY-LPPE 158
```

3

```
Score =  159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3   WVWALLLLAAWAAAERD--------CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
            V  L+ LA  A               +  S  V+ENFD  ++ G WY + K
Sbjct: 1   MVTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIPASFE-KG 59

Query: 55  DNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQ 114
           + I A +S+ E G +    K          V  +      ++  +PAK +++++ +
Sbjct: 60  NCIQANYSLMENGNIEVLNKELSPDGTMNQVKGE--AKQSNVSEPAKLEVQFFPL----- 112

Query: 115 KGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEA 164
               +WI+ TDY+ YA+ YSC             + ++  R+P  LPPE
Sbjct: 113 MPPAPYWILATDYENYALVYSCTTFFWL--FHVDFFWILGRNPY-LPPET 159
```