

BIOC4004 - Industrial Biochemistry

Lecture 13 - Mon Feb 22, 04

Topics for the Day:

- Searching molecular databases
- BLAST
 - how it works
 - how to use
 - how to interpret
- Bioinformatics resources

Searching of Molecular Databases

- Why ?

To look for similarities between a sequence of interest and the sequences in the database

- could help elucidate the function of an unknown sequence
- could help find conserved motifs shared by different sequences

- How ?

Need to compare the sequence to every sequence in the DB

- need to align the query sequence to every sequence in the DB
- need to calculate some metric to assess the quality of the match
- need to report the “good” matches

- Early sequence alignment programs were meant to perform the best possible alignment of a pair of, at most, a few (dozen) sequences

- good results but computationally expensive
- no way could you use these to screen a DB of thousands of sequences

- Design of “heuristic” methods to perform the searches:

- an initial quick and dirty alignment between query and all the sequences
- assessment of potentially matching sequences worth “revisiting”
- a second, more accurate, alignment on the worthwhile sequences

- These heuristic methods are computationally economical:

- fast and good
- can handle lots of queries !!!
- Provide a statistical assessment of the match (more on this later)

- The program used for searching the GenBank DB is BLAST (Basic Local Alignment Tool)

[Some of the European DBs use a program called FASTA, however, GenBank cross-references entries in the European DBs anyway, so we'll mostly talk about GB]

BLAST (Basic Local Alignment Tool)

- find it at <http://www.ncbi.nlm.nih.gov/BLAST/>
also mirrored at the Canadian Bioinformatics Resource
<http://www.cbr.ca/blast/>
- uses an approach based on:
 - matching short sequence fragments (HSPs: High Scoring Segment Pairs)
 - HSPs can be perfect matches or highly homologous
 - looks for clusters of HSPs that are close to each other
 - a “cut-off alignment score” is calculated to determine the minimum score that should be met by an alignment in order to be significant
 - the best local alignments between the query sequence and the database sequence(s) are calculated and reported

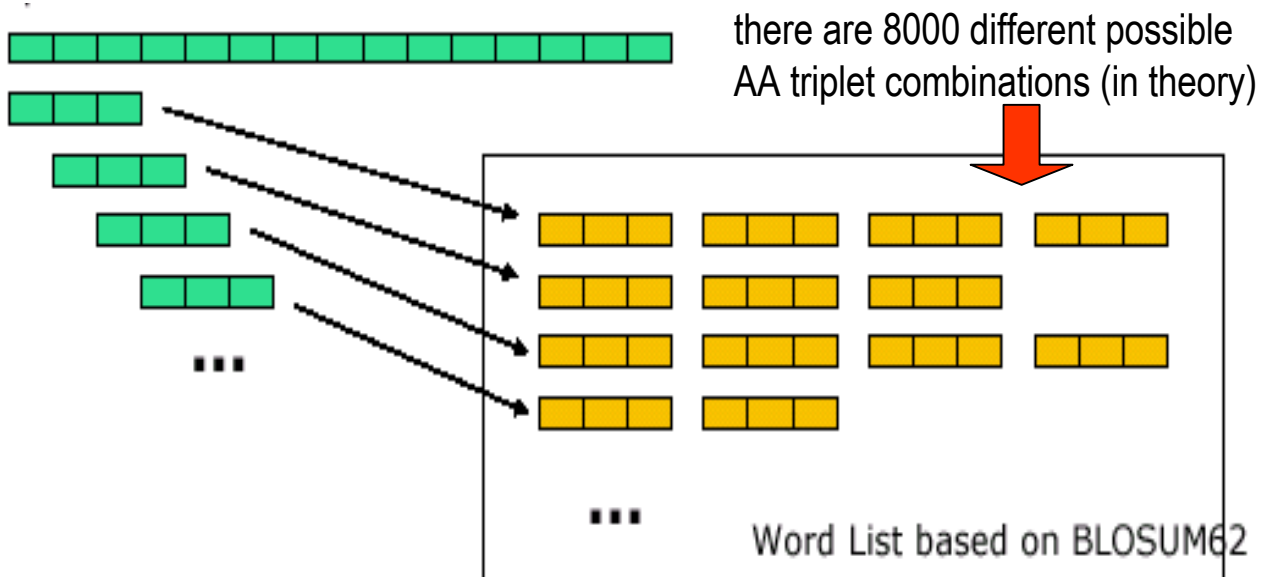
BLAST VARIANTS

PROGRAM	QUERY	DB	COMMENTS
BLASTP	protein	protein	compares amino acid query against protein sequences
BLASTN	DNA	DNA	compares nucleotide query against DNA sequences
BLASTX	DNA	protein	compares 6X translations of nucleotide query against protein sequences
TBLASTN	protein	DNA	compares protein query against 6X translations of DNA sequences
TBLASTX	DNA	DNA	compares 6X translations of nucleotide query against 6X translations of DNA sequences

- also :
 - PSI-BLAST: protein query and protein DB - different statistics used to detect weak similarities
 - PHI-BLAST: protein query and protein DB - used to look for protein patterns (small regions of conservation)
 - MEGA-BLAST: larges sets of long DNA sequences
 - CD Search - conserved domain detection
 - BLAST2: pairwise alignment of two sequences
 - Genome BLAST : alignment of DNA sequences to genome data
 - VecScreen: used to detect vector sequence within sequence data

How does BLAST work ? (part I)

- A list of words of length 3 in the query protein sequence is made.
- Using BLOSUM62, the query words are evaluated for an exact match with a word of any database sequence.(cf. $20 \times 20 \times 20 = 8000$)



- each triplet is searched against all of the triplets generated from the database
- use scoring matrix to get a score for each potential match

How does BLAST work ? (part II)

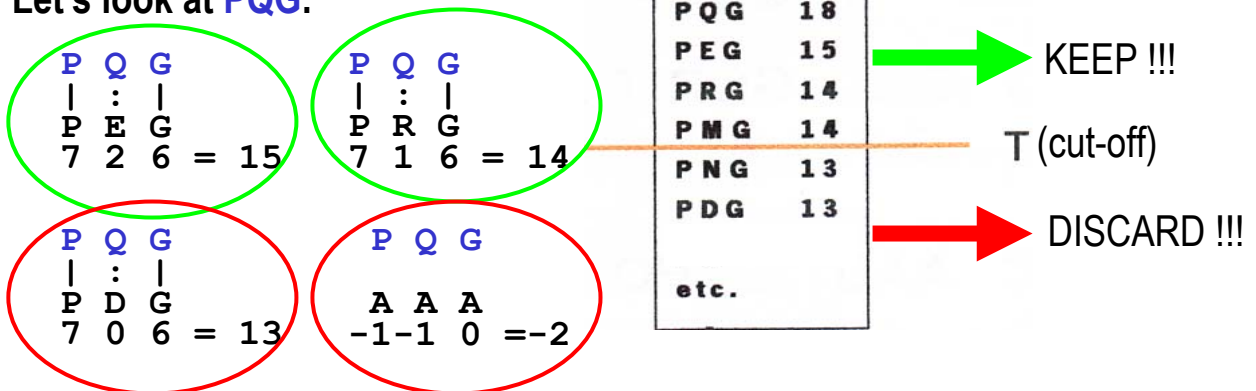
Let's look at one example:

Query: TPQGQRQGQ....

Query Wordlist	TPQ PQG QGQ	Dbase Wordlist	AAA AAC AAD ... PQG QGQ YYY
	GQR QRQ RQG		AGA AGC AAN ... PEG QGM ...
	QGQ etc etc		AAG CAC AAE ... PRG MGQ ...
	...		GAA ACC AAQ ... PMG QAQ ...

- A cutoff score called neighborhood word score threshold(T) is used to pull out low-score matches.

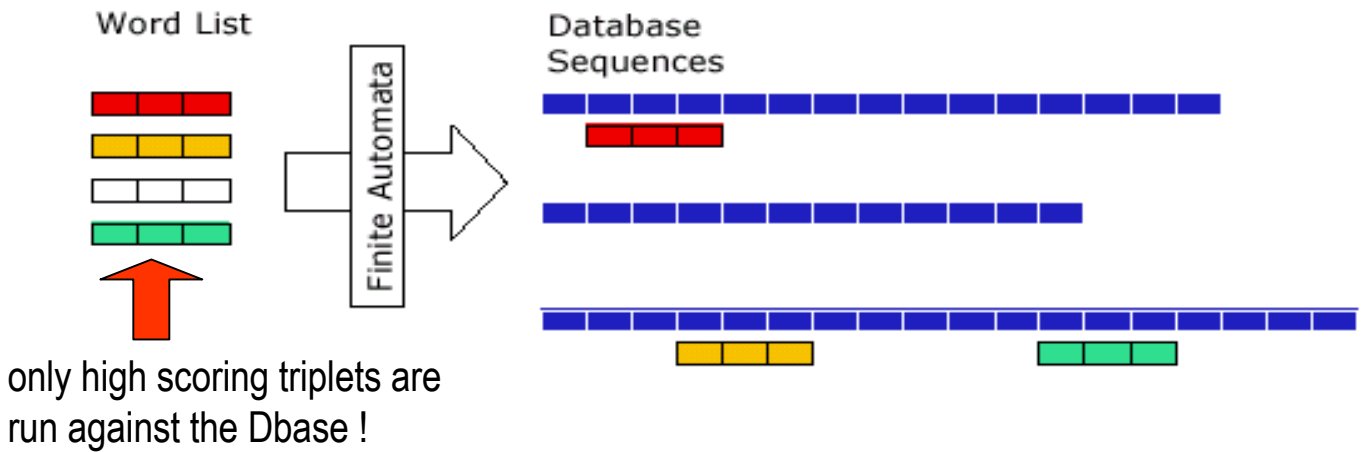
Let's look at **PQG**:




1. Select all Dbase triplets that score higher than cut-off (using BLOSUM62 score)
2. Discard low scoring Dbase triplets (ie.lower than cut-off)
3. Most triplets will be discarded:
 - less data to sift through in subsequent (computationally intensive) steps

How does BLAST work ? (part III)

- You now have a "restricted data-set" of all the high scoring triplets
- You scan every sequence in the database for an exact match to each of the high-scoring triplets (ie. do this with tens or hundreds of triplets, not 8000 !!!)
- Wherever there is a perfect match \Rightarrow SEED



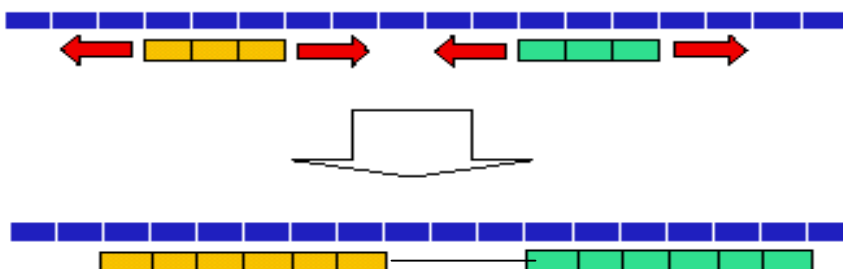
- Each SEED is extended in either direction to get high-scoring segment pair (HSP)



Query: 325 LNKCKT PQG QORQGQWIKQPLMDKN 350
 L TPQGQR++++W+ P+ D
 Sbjct: 290 LDCTVTPQGQREARWLHMPVRDTR 315

An HSP is obtained when the score can neither be improved by extending or trimming

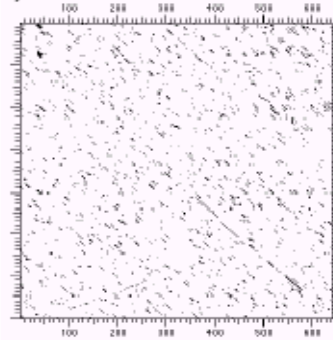
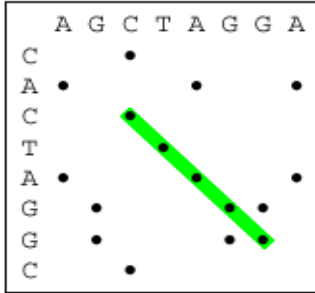
- HSPs that are in close physical proximity are merged into one long HSP
 - ie. a gapped alignment !



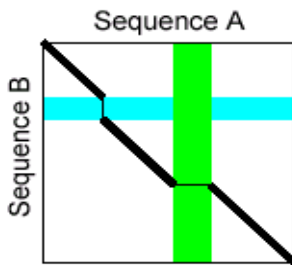
How does BLAST work ? (part IV)

Dot Matrix Analysis

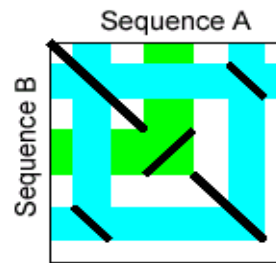
- A.J. Gibbs and G.A. McIntyre (1970)



- Insertions or deletions

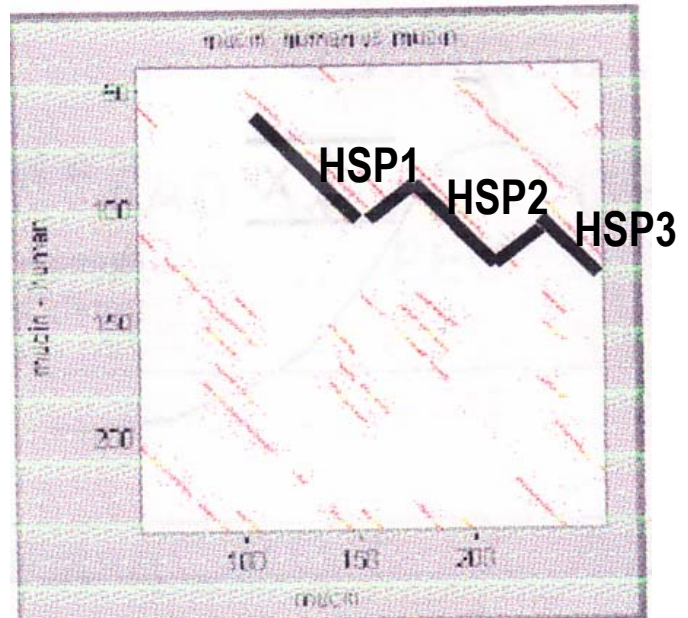


- Direct repeat and inverted region



- HSPs in the same diagonals(ungapped) or in near diagonals(gapped) are merged.

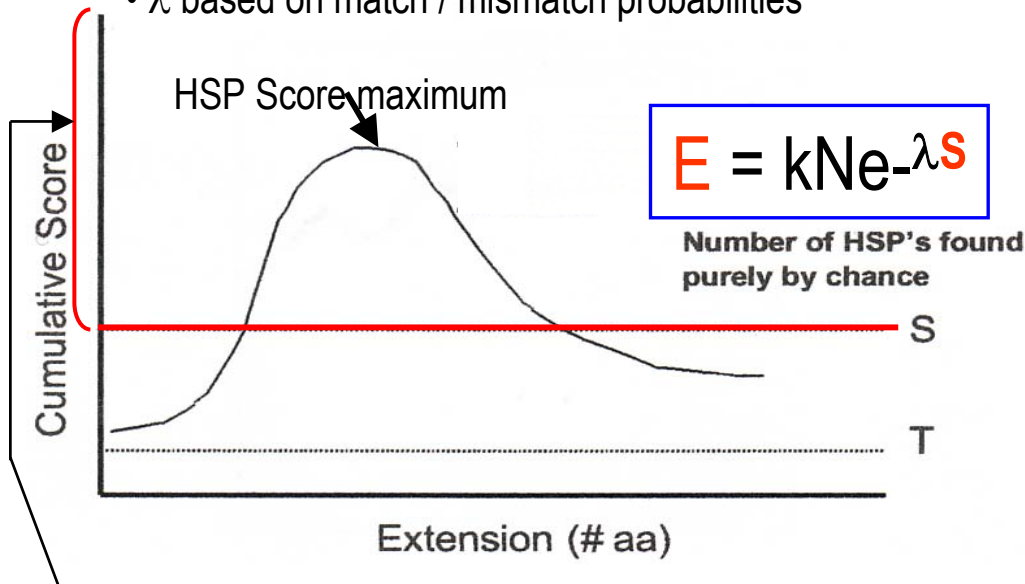
- close physical proximity
- gapped or ungapped



How does BLAST work ? (part V)

• E-value (or Expect value):

- Expected # of HSP alignments with a **Score** > **S** by chance alone
- S is computed for each query / dbase based on the other parameters
 - BLAST uses a default E = 10
 - k is based on the word length (default = 3)
 - N based on the query sequence length and the dbase size
 - λ based on match / mismatch probabilities



- All HSPs with score greater than cut-off score(S) are identified and returned as hits
- All others discarded

How do we distinguish between good HSPs and poor HSPs ?

- For every HSP an E-value statistic is calculated based on the HSPs score

$$P = 1 - e^{-E}$$

When $E < 0.01$, $P \sim E$ (ie. E approximates the probability of match occurring by chance alone as E gets smaller)

- **So the E value is a good measure of the significance of the alignment**
 - the lower the E-value (closer to 0), the more likely that the match is real
 - the higher the E-value (closer to 1), the more chance the match is bogus

Running a BLAST search at NCBI

- Paste in sequence (FASTA format or type in GI or accession number)

```
>Mysequence MT0895
KIQIYGTGCANCQMLEKNAREAVKELGIDAE
FEKIKEMDQILEAGLTALPGLAVDGELKIDS
```

OR

```
>
KIQIYGTGCANCQMLEKNAREAVKELGIDAE
FEKIKEMDQILEAGLTALPGLAVDGELKIDS
```

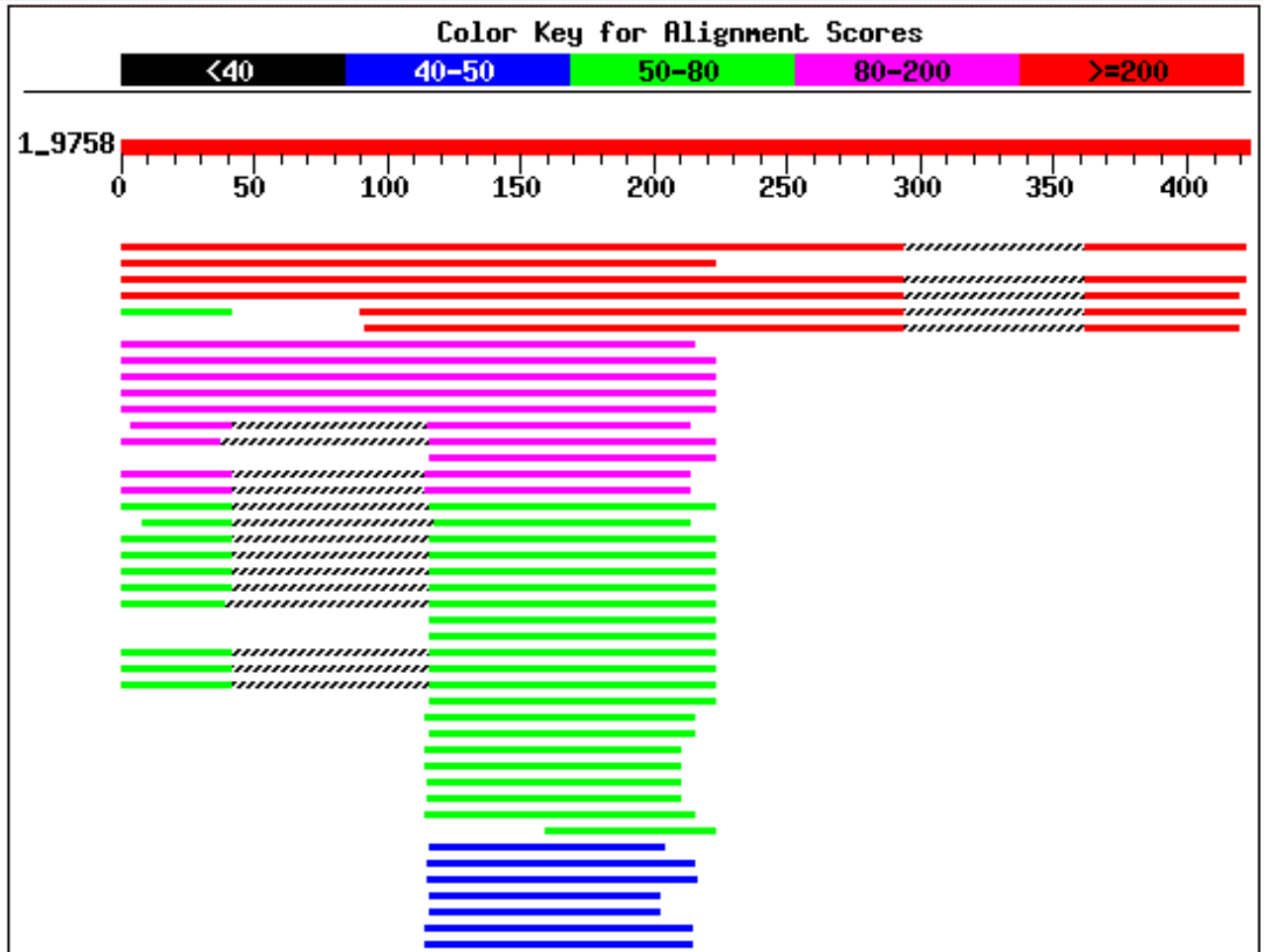
- Choose a range of interest in the sequence “set subsequences” (not usually used)
- Select the database from pull-down menu (usually choose nr = non-redundant)
- Keep CD Search “check box” on ———> Conserved Domain
- Leave “Options” unchanged (use defaults)
- Go to “Format” menu and adjust Number of descriptions and alignments as desired

C'est tout !!!

An Example of a Typical BLAST search (part I)

Distribution of 176 Blast Hits on the Query Sequence

Mouse-over to show define and scores. Click to show alignments



- The first thing you get is the distribution of Blast hits on the query
 - evaluate if the match occurs along the whole length of the query
- vs
- matches on restricted are of query
- How good is the match ? The higher the score !!!!

An Example of a Typical BLAST search (part II)

Query= Pbpp58b (423 letters)

Database: nr (493,611 sequences; 154,780,071 total letters)

Probability of match appearing by chance

Sequences producing significant alignments:	Score (bits)	E Value
sp Q08168 HRP_PLABE 58 KD PHOSPHOPROTEIN (HEAT SHOCK-RELATED PRO...	334	1e-90
gb AAC37300.1 (L21710) 58 kDa phosphoprotein [Plasmodium berghei]	329	3e-89
pir T10455 heat shock related protein - Plasmodium berghei >gi ...	250	2e-65
sp P50503 HIP_RAT HSC70-INTERACTING PROTEIN >gi 4379408 emb CAA5...	106	5e-22
sp P50502 HIP_HUMAN HSC70-INTERACTING PROTEIN (PROGESTERONE RECE...	87	3e-16
gb AAF45894.1 (AE003429) CG2947 gene product [Drosophila melano...	87	4e-16
pir T24865 hypothetical protein T12D8.8 - Caenorhabditis elegan...	86	5e-16
pir T04562 hypothetical protein T12H17.60 - Arabidopsis thalian...	81	2e-14

Good!

Entries removed for clarity

Alignment Score

gb AAC60555.2 (S59774) STI1 stress-inducible protein homolog [S...	48	2e-04
gb AAD33401.1 AF129086_1 (AF129086) carboxy terminus of Hsp70-in...	48	2e-04
pir A56534 P58 protein - bovine >gi 468012 gb AAA17795.1 (U046...	48	2e-04
gb AAB49720.1 (U89984) transformation-sensitive protein homolog...	47	4e-04
pir T16689 hypothetical protein R05F9.10 - Caenorhabditis eleg...	47	4e-04
gb AAD33400.1 AF129085_1 (AF129085) carboxy terminus of Hsp70-in...	46	6e-04

CRAP

Identifier Line:
database | accession # | name or locus

Entries removed for clarity

emb CAA61595.1 (X89416) protein phosphatase 5 [Homo sapiens]	43	0.007
pdb 1A17 Tetratricopeptide Repeats Of Protein Phosphatase 5	43	0.007
ref NP_006238.1 protein phosphatase 5, catalytic subunit >gi 1...	43	0.007
pir S52570 phosphoprotein phosphatase (EC 3.1.3.16) 5, catalyti...	43	0.007
gb AAB18614.1 (U12203) phosphoprotein phosphatase [Rattus norve...	43	0.007
sp P53042 PPP5_RAT SERINE/THREONINE PROTEIN PHOSPHATASE 5 (PP5) ...	43	0.007
gb AAB60384.1 (U25174) serine-threonine phosphatase [Homo sapiens]	43	0.007
sp Q60676 PPP5_MOUSE SERINE/THREONINE PROTEIN PHOSPHATASE 5 (PP5...	42	0.009
gb AAB70573.1 (AF018262) protein phosphatase 5; PP5 [Mus musculus]	42	0.009

CRAP

- Note that the first few hits have very low E-values (1e-90)
 - very close to 0; very good match
- I wouldn't trust anything with an E-value higher than ~1e-10 (if that...)
 - as E gets closer to 1, you get closer to a 100% probability that the match could have occurred by chance alone

An Example of a Typical BLAST search (part III)

```
>sp|P50503|HIP_RAT_HSC70-INTERACTING_PROTEIN >qi|4379408|emb|CAA57546.1| (X82021)
Hsc70-interacting protein [Rattus norvegicus] (Length = 368)
```

```
Score = 106 bits (261), Expect = 5e-22
Identities = 60/224 (26%), Positives = 97/224 (42%)
```

Filtering—Glu rich

```
Query: 1 MDIRKIEDLKKFVASCEBNPSILLKPELSFFKDFIRSPGGKIKKDKMGYXXXXXXXXXX 60
MD K+ +L+ FV C ++PS+L E+ P +++E+ GGK+
Sbjct: 1 MDPRKVSELRAFVKMCRQDPSVLHTEEMRFLREHWVESMGQKVPDPATHKAKREENTKEEKR 60

(SDREEREDEEEEEERESDDDDPEKLE)
Query: 61 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX 120
+ P + K A
Sbjct: 61 DKTTEENIKTEREPSSRESLDLEIDNEGVIHADTDAPQEMGDENARITRANMDEANKEKGGAA 120

Query: 121 VDLVENKKYERALEKYNKIISPFGNPSAMIYTKRASILLNLKRPKACIRDCTEALNLNVDS 180
+D + + + ++A++ + I A++Y KRAS+ + L++P A IRDC A+ +N DS
Sbjct: 121 IDALNDGELQKAIDLFTDAIKLNRLAIIYAKRASVFKLQKPNAAIRDCDRAIRINPDS 180

Query: 181 ANAYKIRAKAYRYLGKWEFAHADMEQGQKIDYDENLWDMQKLIQ 224
A YK R KA+R LG WE A D+ K+DYDE+ M + +Q
Sbjct: 181 AQPYKWRGKAHRLIGHWEBAARDLALACKLDYDEDASAMLRVQ 224
```

Middle row = matches and similar residues (+)

```
>pir||T24865 hypothetical protein T12D8.8 -Caenorhabditis elegans (Length = 422)
```

```
Score = 86.2 bits (210), Expect = 5e-16
Identities = 44/101 (43%), Positives = 60/101 (58%), Gaps = 2/101 (1%)
```

```
Query: 119 EAVDLVENKKYERALEKYNKIISPFGNPSAMIYTKRASILLNLKRPKACIRDCTEALNLNV 178
+A + N ++ AL + I SAM++ KRA++LL LKRP A I DC +A+++N
Sbjct: 121 KAQBAPSNQDFDFTALHTPTAAIRANPGSAMLHAKRANVLLKLRPVAAIADCDKAIINP 180
```

```
Query: 179 DSAANAYKIRAKAYRYLGKWEFAHADMEQGQKIDYDE--NLW 217
DSA YK R +A R LGKW A D+ K+DYDE W
Sbjct: 181 DSAQGYKFRGRANRLLGKWVEAKTDLATAACKLDYDEANEN 221
```

Gaps to maximize alignment

```
Score = 41.4 bits (95), Expect = 0.016
Identities = 16/34 (47%), Positives = 23/34 (67%)
```

```
Query: 9 LKKFVASCEBNPSILLKPELSFFKDFIRSPGGKI 42
LK+FV C+ NP++L PE FFKD++ S G +
Sbjct: 7 LKQFVGMCCQANPAVLHAFEPFGFFKDYLVSLGATL 40
```

A second high-scoring segment

- note that low complexity regions have been filtered out
- can generate bogus matches

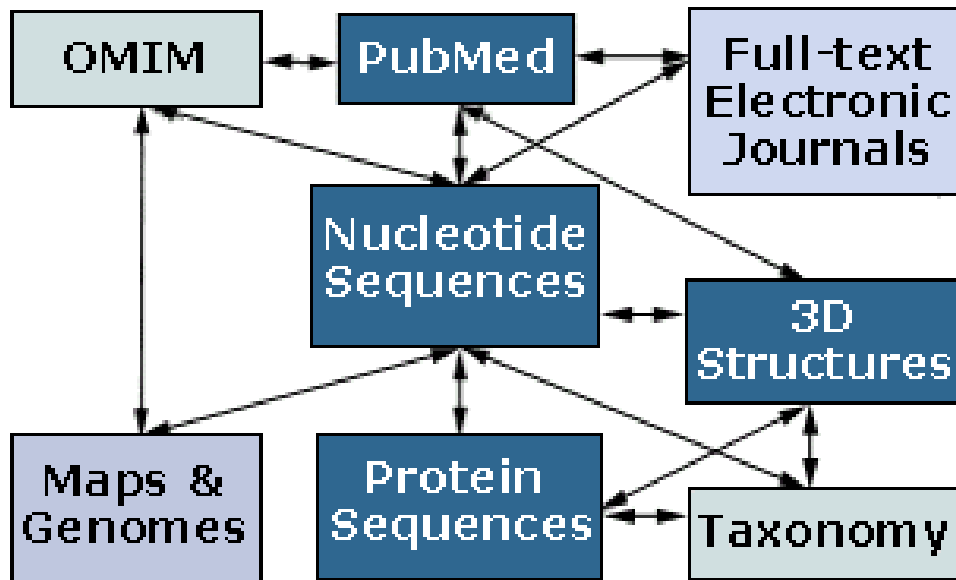
Bioinformatics Analysis (revisited)

- Sequence Alignments
 - Sequence Assembly
 - Database Searching
 - Function Prediction
 - Domain Identification
 - Molecular phylogenetics
- Motif-Finding
 - DNA:
 - restriction sites
 - gene structure
 - Transcription/Translation signals
 - Intron/Exon boundaries
 - gene prediction
 - Proteins
 - PTM prediction (phosphorylation, glycosylation, ...)
 - structural motifs (zinc-finger, ATP-binding cassette...)
 - transmembrane domains
 - export sequences
- Protein Structure Analysis
 - physical characteristics (hydrophobicity, α -helix/ β -sheet potential...)
 - protein structure comparison
 - protein structure prediction (protein-fold prediction or threading)
 - X-ray crystallography and NMR data analysis
- Gene Expression Analysis
 - microarray data analysis (image analysis, data clustering)
 - proteomics data analysis (2-D gel analysis, Mass Spec data analysis)
 - metabolomics data analysis (Mass Spec data analysis)
- Pathway Analysis
 - metabolic pathway reconstruction
 - elucidation of regulatory networks

The original bioinformatics resource.....

NCBI (<http://ncbi.nlm.nih.gov/>): BLAST, Medline, GenBank

The screenshot shows the NCBI website interface. At the top, the NCBI logo is on the left, and the text "National Center for Biotechnology Information" is centered, with "National Library of Medicine" and "National Institutes of Health" below it. A navigation bar contains tabs for PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Structure. Below this is a search bar with a dropdown menu set to "Nucleotide" and a "Go" button. On the left side, there are links for "SITE MAP", "About NCBI", "GenBank", and "Molecular databases". The main content area features a "What does NCBI do?" section with a paragraph about its history and a "Hot Spots" section with a list of projects. A "Mouse Genome" banner is also present, featuring a mouse icon and links to "Map Viewer", "Sequencing Progress", and "Human-Mouse Homology".



- as it's matured: more user friendly
- plenty of FAQs and tutorials !!!

For a general sequence analysis tools

BCM Search launcher (<http://searchlauncher.bcm.tmc.edu/>)

BCM Search Launcher
Baylor College of Medicine HGSC

HGSC | SL Home | MBCR | Search Tools | Help | FAQ | What's New | Contact Us | Comments | User Survey | Site Search

Start Search Launcher: Choose a type of search from the pull-down menu

Start Search Launcher:

- Nucleic Acid Search
- General Protein Search
- Species Protein Search
- Multiple Sequence Alignment
- Pairwise Sequence Alignment
- Gene Features Searches
- Sequence Utilities
- Protein Structure Prediction
- Other Services
- BCM Human Transcript Database

with client. (ver 2.8) for your operating system.

The launcher is an on-going project to organize molecular biology-... on the WWW by function by providing a single point-of-entry... launching protein sequence searches using standard

For a myriad of protein analysis tools

ExpASy (<http://www.expasy.ch>)

ExpASy Molecular Biology Server - Netscape

Site Menu Search ExpASy Contact Us

Hosted by SIB Switzerland Mirror sites: Australia Canada China Korea Taiwan

ExpASy Molecular Biology Server Expert Protein Analysis System

This is the ExpASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB). This server is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer).

Announcements | About | Mirror Sites

Databases	Tools and Software Packages
<ul style="list-style-type: none">• SWISS-PROT and TrEMBL - Protein knowledgebase• PROSITE - Protein families and domains• SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis• SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules• SWISS-MODEL Repository - Automatically generated protein models• CD40L base - CD40 ligand defects• ENZYME - Enzyme nomenclature• SeeAnalRef - Sequence analysis bibliographic references• Links to many other molecular biology databases	<ul style="list-style-type: none">• Proteomics tools<ul style="list-style-type: none">◦ Identification and characterization◦ DNA -> Protein◦ Similarity Searches◦ Pattern and profile searches◦ Posttranslational modification prediction◦ Primary structure analysis◦ Secondary structure prediction◦ Tertiary structure◦ Transmembrane regions detection◦ Signalment• Melanie 3 - Software for 2-D PAGE analysis• SWISS-MODEL - Automated knowledge-based protein modelling server• Swiss-PathViewer - Software for structure display and analysis• Boehringer Mannheim's Biochemical Pathways

Education and services	Documentation
<ul style="list-style-type: none">• The ExpASy FTP server• SWISS-SIBdb - automatically obtain (by email) new sequence entries relevant to your field(s) of interest• Masters Degree in Bioinformatics• 2-D PAGE training - attend a one-week course in Geneva• SWISS-2DSEARCH - get your 2-D gels performed according to Swiss standards	<ul style="list-style-type: none">• What's New on ExpASy• SWISS-FLASH electronic bulletins• SWISS-PROT documents• How to create HTML links to ExpASy• Complete table of available documents

Another excellent Site

CBR (<http://www.cbr.nrc.ca/>): lotsa good stuff !!!!!

- A lot more sequence analysis tools than offered at NCBI



National Research
Council Canada

Conseil national
de recherches Canada



English

Français

Canada

- Mirror for other bioinformatics sites:
 - BLAST
 - ExPASy
 - CMS Molecular Biology resource (tons of stuff !!!)
a full suite of protein and DNA analysis
- Also offers other analysis programs:
 - ClustalW: multiple sequence alignments
 - DNA fold: DNA and RNA folding
 - EMBOSS: a full suite of protein and DNA analysis
 - GeneMatcher: highly sensitive database and pattern searching
 - MAGPIE: genome data analysis
 - ReadSeq: sequence format conversion tool
 - PRIMER3: PCR primer design
 - WebPhylip: phylogenetic analysis