

Computational Biology or Bioinformatics

- **ability to rapidly sequence DNA has led to large databases**
 - **development of new algorithms**
 - **data analysis and interpretation**
- **similar concepts also applied to epidemiological databases**
 - **genetic epidemiology**
 - **evolutionary genetics**
- **align related sequences and search databases**

Sequence Alignments

- **easy to obtain DNA sequence data**
- **difficult to predict protein structure and function**
- **structure/function can be inferred from sequence similarities**
- **similarities identified by aligning DNA or protein sequences**
- **alignments can be 'global' or 'local'**

- **homolog (common ancestor)**
 - **ortholog (between species)**
 - **paralog (within species)**
- **analog (no common ancestor)**

Pair-wise Sequence Alignments

- ‘scoring matrix’ calculates an alignment score
 - eg, match = 0.9 and mismatch = -0.1 for DNA
 - amino acids have different weights (abundance and chemical or structural similarities)
 - BLOSUM and PAM + variants

GCGCCTC

||| ||

GCGGGTC

$$(5 \times 0.9) + (2 \times -0.1) = 4.3$$

Amino Acid Similarities

Chemical	Physical
A, G	C, S
D, E	D, L, N
F, Y	E, Q
K, R	F, H, W, Y
I, L, M, V	I, T, V
Q, N	K, M, R
S, T	

Pair-wise Sequence Alignments

- gap penalties (opening and extending)
 - optimal penalties depend on relatedness of sequences
 - 'trial and error' approach
- alignment with maximum score is returned for prescribed gap penalties and scoring matrix
 - not necessarily most biological significant

Multiple sequence alignments

- gives 1st approximation of best score
- human eye + biological insight better at refining the alignments

Databases

- **two types:**
 - 1° (original biological data)
 - 2° (value added)
- **three 1° DNA databases**
 - GenBank
 - EMBL
 - DDBJ
- **subdivisions (taxonomic groups, genome projects, ESTs, etc)**
- **annotated to include ancillary information (author, publications, etc.)**

Searching Databases

- **text-based (annotations)**
 - gene name, authors, species, etc.
- **information retrieval systems**
 - Entrez can access all databases + medline
- **sequence comparisons**
 - submit 'query' sequence
 - compare to all sequences in database(s)
- **pairwise is too time consuming**
 - heuristic programs (eg, FASTA and BLAST)
 - match short sequence fragments
 - alignments of sequence regions showing promise
 - scores and statistics

Basic Local Alignment Search Tool

BLAST PROGRAMS

PROGRAMS	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

Doing a BLAST Search

- <http://www.ncbi.nlm.nih.gov/BLAST/>
- choose BLAST program
- paste in query sequence or acc. no. → BLAST!
- change default options:
 - database (nr = non-redundant)
 - scoring matrix and gap penalties
 - filtering
 - E-value cutoff (ie, Expect)
 - limit subset of database (organism, keyword, etc.)
 - display options (eg, # of descriptions, alignments, etc.)

Blast Search Results

Query= Pbpp58b (423 letters)

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

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...	<u>86</u>	5e-16
pir T04562 hypothetical protein T12H17.60 - Arabidopsis thalian...	81	2e-14

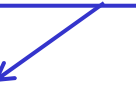
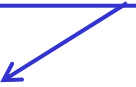
.
.
.
.
.

database | accession # | entry name or locus

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

Probability

Score (bits) E Value



Example of Blast Alignment

>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 EAVDLVENKKYEEALEKYNKIISFGNPSAMIYTKRASILLNLKRPKACIRDCTEALNLNV 178
      +A +  N  ++ AL  +  I      SAM++ KRA++LL LKRP A I DC +A+++N
Sbjct: 121 KAQEAFSNGDFDTALHTFTAAIEANPGSAMLHAKRANVLLKLKRPVAAIADCDKAISINP 180
```

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

Matches



Gap



Score = 41.4 bits (95), Expect = 0.016

Identities = 16/34 (47%), Positives = 23/34 (67%)

```
Query: 9  LKKFVASCEENPSILLKPELSFFKDFIESFGGKI 42
      LK+FV  C+ NP++L  PE  FFKD++ S G  +
Sbjct: 7  LKQFVGMQCQANPAVLHAPEFGFFKDYLVS LGATL 40
```

A 2nd high scoring segment



Blast Search Results

Query= Pbpp58b (423 letters)

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

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...	<u>106</u>	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 elegans...	86	5e-16
pir T04562 hypothetical protein T12H17.60 - Arabidopsis thalian...	81	2e-14

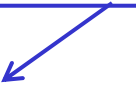
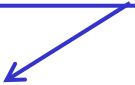
.
.
.
.
.

database | accession # | entry name or locus

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

Probability

Score
(bits) E
Value



Example of Blast Alignment

```
>sp|P50503|HIP_RAT HSC70-INTERACTING PROTEIN >gi|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

```
Query: 1 MDIEKIEDLKKFVASCEENPSILLKPELSFFKDFIESFGGKIKKDKMGYXXXXXXXXXXXXX 60  
MD K+ +L+ FV C ++PS+L E+ F ++++ES GGK+  
Sbjct: 1 MDPKRVSELRAFVKMCRQDPSVLHTEEMRFLREWVESMGGKVPPATHKAKSEENTKEEKR 60  
  
(SDEEEEEDEEEEEEEEEEDDDPEKLE)  
Query: 61 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX 120  
+ P + K A  
Sbjct: 61 DKTTEDNIKTEEPSSEESDLEIDNEGVIEADTDAPQEMGDENAEITEAMMDEANEEKGAA 120  
  
Query: 121 VDLVENKKYEEALEKYNKIISFGNPSAMIYTKRASILLNLKRPKACIRDCTEALNLNVDS 180  
+D + + + ++A++ + I A++Y KRAS+ + L++P A IRDC A+ +N DS  
Sbjct: 121 IDALNDGELQKAIDLFTDAIKLNPRLAILYAKRASVFKLQKPNAAIRDCDRAIEINPDS 180  
  
Query: 181 ANAYKIRAKAYRYLGKWEFAHADMEQGQKIDYDENLWDMQKLIQ 224  
A YK R KA+R LG WE A D+ K+DYDE+ M + +Q  
Sbjct: 181 AQPYKWRGKAHRLLGHWEEAARDLALACKLDYDEDASAMLRVQ 224
```

LOCUS RNHSRP 1694 bp mRNA ROD 14-JAN-1996

DEFINITION R.norvegicus mRNA for heat shock related protein.

ACCESSION X82021

REFERENCE 2 (bases 1 to 1694)

AUTHORS Hohfeld, J., Minami, Y. and Hartl, F.U.

JOURNAL Cell 83 (4), 589-598 (1995)

MEDLINE [96069860](#)

FEATURES Location/Qualifiers

source

1..1694

/organism="Rattus norvegicus"

gene

67..1173

/gene="hip"

CDS

67..1173

/product="Hsc70-interacting protein"

/protein_id="[CAA57546.1](#)"

/db_xref="[SWISS-PROT:P50503](#)"

/translation="MDPRKVSELRAFVKMCRQDPSVLHTEEMRFLREWVESMGK

.....

QDVAQNPSNMSKYQNNPKVMNLIKLSAKFGGHS"

BASE COUNT

542 a 342 c 423 g 387 t

1 gcgtcgacgg gcttggcatc gggcctccgc agccgcccac cgccagaagc ttccagcctc

.....

1681 aaaaaaaaaa aaaa

//

