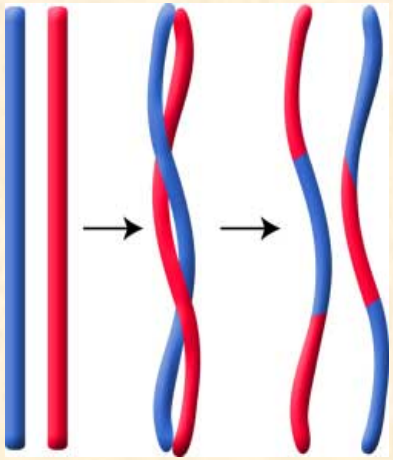
http://petang.cgu.edu.tw/bioinformatics/index.htm



Bioinformatics

Lecture 4 – Comparing Sequences & Multiple Sequence Alignment

AGTCCGCGAATACAGGCTCGGT AGTCCGCGAATACAGGCTCGGT

Petrus Tang, Ph.D. (鄧致剛**)**
Graduate Institute of Basic Medical Sciences
and
Bioinformatics Center, Chang Gung University.
petang@mail.cgu.edu.tw
EXT: 5136

助教：
蔡智宇(分機5690)

Comparing Sequences and Multiple Sequence Alignment

Comparison of your "query" DNA, RNA, or Amino acid sequence to a known sequence or database

Create an alignment of 2 or more sequences indicating matches

Comparing Sequences and Multiple Sequence Alignment

Pairwise Comparsion

```
137 AGACCAACCTGGCCAACATGGTGAAATCCCATCTCTAC.AAAAATACAAA 185
    |||||| |||||||||||||||||  |||||||||  |||||||||
  1 AGACCAGCCTGGCCAACATGGTGAAACTCCATCTCTACTGAAAATACAAA 50
```

Multiple Sequence Alignment

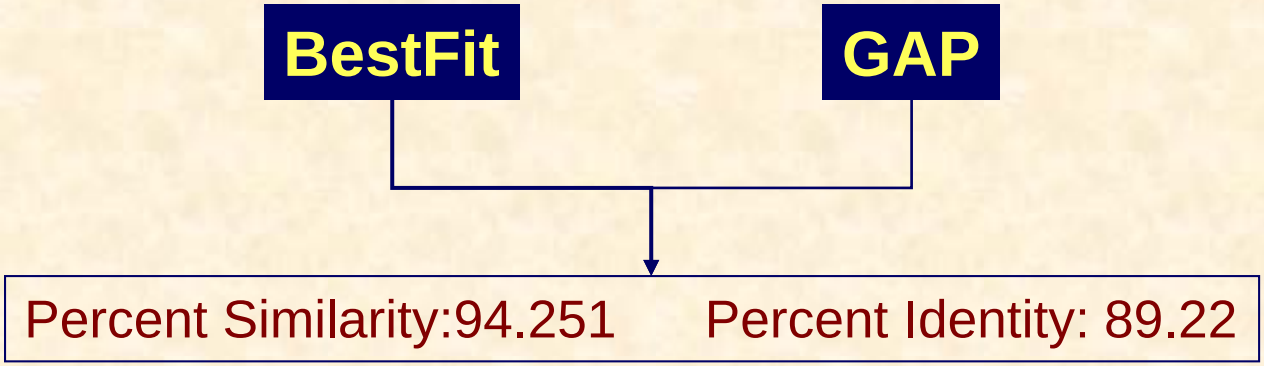
```
          1                                                  50
S11448    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MTFD GAIGIDLGTT YSCVGVWQNE
S06443    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MTFD GAIGIDLGTT YSCVGVWQNE
A25398    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MTYE GAIGIDLGTT YSCVGVWQNE
S06158    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MTYE GAIGIDLGTT YSCVGVWQNE
S42164    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~MS KAVGIDLGTT YSCVAHFAND
S20139    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~MS KAVGIDLGTT YSCVAHFSND
B36590    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~MS KAVGIDLGTT YSCVAHFAND
A25089    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~MAKSEG PAIGIDLGTT YSCVGLWQHD
S03250    ~~~~~~~~~~ ~~~~~~~~~~ ~~~MAGKGEG PAIGIDLGTT YSCVGVWQHD
A27077    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MSKG PAVGIDLGTT YSCVGVFQHG
S07197    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MSKG PAVGIDLGTT YSCVGVFQHG
```

Why Compare sequences?

Comparing two sequences of the same type
(e.g. genomic *vs*. genomic, mRNA *vs*. mRNA, protein *vs.* protein)
Shows you how similar sequences are.
Highlight regions of similarity or difference.
Find best region of similarity.
Look for overlaps.
Often more exacting alignments than database scanning programs.

Comparing genomic vs. EST or genomic vs. protein:
Reveal coding regions
Reinforce gene predictive methods
Many programs have been written to do pairwise comparisons, some of the major types are discussed below:

Pairwise Comparsion



Identity, Similarity and Homology

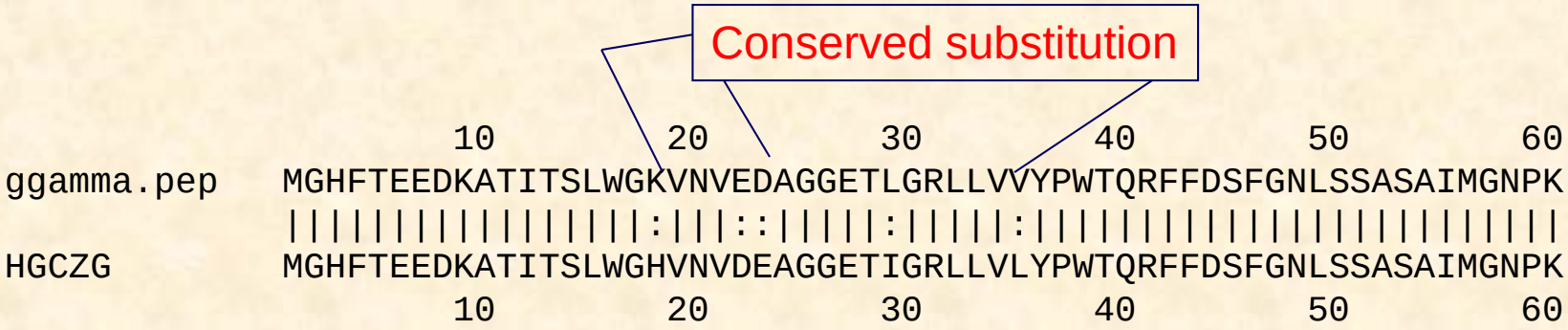
Identity and Similarity is a measurable property

Homology implies functional or evolutionary relatedness

Pairwise Comparsion

Nucleotide sequence alignments

Protein sequence alignments



Residues with shared chemical properties can substitute for each other
Size, charge, hydrophobicity, polarity
scored less than a match, but better than a mismatch
Conservative changes scored as better than non-conservative

Identity & Similarity

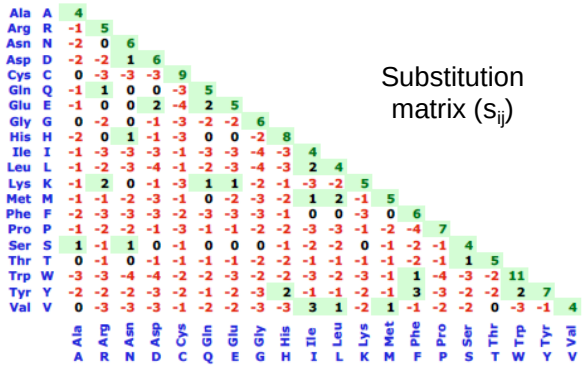
Score: A number used to assess the biological relevance of a finding.

In the context of sequence alignments, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used (substitution matrix, gap penalty).

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

Example:

```
R  L  A  S  V  -  E  T  D  M  W  T  P  L  T  L  R  Q  H
.  |  .  |  :     :  |  .  :        .  |  .  .  |
T  L  T  S  L  A  Q  T  T  L  -  -  K  A  H  L  G  T  H
-1 +4 +0 +4 +1 -4 +2 +5 -1 +2 -4 -1 -1 -1 -2 +4 -2 -1 +8 = 12
```



Substitution matrix (s_{ij})

gap penalty ($s_{i\text{-}}$)

gap opening	-4
gap extension	-1
end gap	0

Gap Penalty

Gap penalty	Alignment	Identity / Similarity	Gaps	Score
0	1 GTC-ATGCTA-GTCGT---GG---GTAGCATTTA-GCT-ATG-TGGG-GT 38 1 -TCGATGCT-GGTCG-CAAGGCAAGTAG---TTATG-TCATGCT---AG- 39	27/50 (54.0%)	23/50	S=135
5	1 GTC-ATGCTAGTCG--TGGGTAGCATTTA-GCT-ATG-TGGGGT 38 1 -TCGATGCTGGTCGCAAGGCAAGTAGTTATG-TCATGCTAG--- 39	26/44 (59.1%)	11/44	S=67
10	1 ----------GTCATGCTAGTCGTGGGTAGC 21 1 TCGATGCTGGTCGCAAGGCAAGTAGTTATGTCATGCTAG---------- 39 22 ATTTAGCTATGTGGGGT 38 39 ---------- 39	10/67 (14.9%)	57/67	S=50

Observations: If the gap penalty is too large, gaps are avoided and the sequences can not be properly aligned. If the gap penalty is too low, gaps are inserted everywhere to prevent mismatches. This does not produce any informative alignement. The "best" alignment is obtained for an intermediary gap penalty.

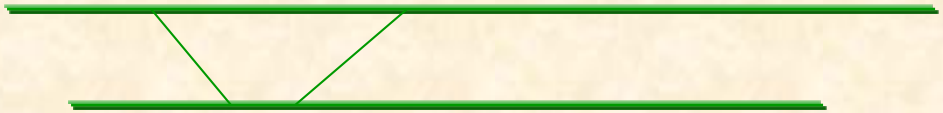
Remark: The scores of these different alignments can not be compared (neither used to select the best alignment) because their scale depends on the gap penalty.

Pairwise Comparsion

Local Alignment

BestFit

compares regions within two sequences and can return several matches



BLAST

VS

Global Alignment

GAP

compare entire sequences

FASTA

Pairwise Comparsion

1. **BestFit:**
   Make an optimal alignment of the best segment of similarity between two sequences by inserting gaps to maximize the number of matches using the *local homology* algorithm of Smith and Waterman.
2. **Compare:**
   Compare two protein or nucleic acid sequences
3. **DotPlot**:
   Make a dot-plot with the output file from Compare.
4. **Gap:**
   Alignment of two sequences which has maximum base matches and minimum gap by using the algorithm of Needleman and Wunsch.
5. **GapShow:**
   Graphic of alignment (use Gap or Bestfit first)
6. **FrameAlign:**
   Create an optimal alignment between a protein sequence and the codons in 3 reading frames on a nucleotide sequence
7. **ProfileGap:**
   Make an optimal alignment between a profile and one or more sequences

Pairwise Comparsion

There are three variations on the theme of sequence comparison.

The **BEST** region of similarity between two sequences,
The best **OVERALL** alignment of two sequences, or
ALL regions of similarity between them.

bestfit –
finds the best single region of similarity & displays it.

gap –
aligns two sequences over their entire length & displays it.

compare - finds all regions of potential homology & displays them.

NB: Be careful when using these programes; it is possible to align one sequence with *any* other, if you really want to. False alignments, and the research you plan using them, may have no biological significance!

Pairwise Comparsion

FrameAlign creates an optimal alignment of the best segment of similarity (local alignment) between a protein sequence and the codons in all possible reading frames on a single strand of a nucleotide sequence. Optimal alignments may include reading frame shifts.

Query:Nucleotide sequence Against:Protein sequence

```
  3 GAAATCAAGAAGGCCATCAAGGAGGAATCTGAAGGCAAAATGAAGGGAAT 52
    ||||||||||||||||||||||||||||||||||||||:::|||||||||
261 GluIleLysLysAlaIleLysGluGluSerGluGlyLysLeuLysGlyIl 277
           .         .         .         .         .
 53 TTTGGGATACTCTGAGGATGATGTTGTGTCTACCGACTTTGTTGGTGACA 102
    ||||||||||...|||||||||||||||||||||||||||||||||||||
278 eLeuGlyTyrThrGluAspAspValValSerThrAspPheValGlyAspA 294
           .         .         .         .         .
103 ACAGGTCAAGCATTTTCGATGCCAAGGCTGGATTGCATTGCATTGAGCGA 152
    ||||||||||||||||||||||||||||    ||||||||||||||
295 snArgSerSerIlePheAspAlaLysAlaGly....IleAlaLeuSerAs 309
```

FrameAlign always finds an alignment for any protein and nucleotide sequences you compare, even if there is no significant similarity between them. You must evaluate the results critically to decide if the segment shown is not just a random region of relative similarity

EXERCISE 04-1

BestFit and GAP

FETCH the following sequences in GCG:
fetch k02938 (Xenopus 5S RNA gene transcription factor TFIIIA mRNA)
fetch x15785 (Xenopus TFIIIA gene 5' region)

Perform
(A)bestfit-call the output display file best.pair
(B)gap-call the output display file gap.pair
-->cat best.pair
-->cat gap.pair
-->Compare the results

ANSWER

Multiple Sequence Alignment

Compare three or more sequences to each other.

Uses

- Select appropriate primers for a gene family
- Identify conserved regions and motifs
- Identify gene families
- Generates a consensus sequence
- First step to the study of phylogenetic relationships

Programs trade sensitivity and alignment quality for computational speed

Use of more than one program is advised

Multiple Sequence Alignment

1. **MEME:**
   Find conserved motifs in a group of unaligned sequences similarity between two sequences.
2. **NoOverlap:**
   Identify the places where a group of nucleotide sequences do not share any common subsequences.
3. **OldDistances:**
   Make a table of the pairwise similarities within a group of aligned sequences.
4. **Overlap:**
   Compare two sets of DNA sequences to each other echo in both orientations.
5. **PileUp:**
   Create a multiple sequence alignment from a group of related sequences.
6. **PlotSimilarity:**
   Plot the running average of the similarity among multiple sequence alignment.
7. **Pretty:**
   Display multiple sequence alignments and calculates a consensus sequence.
8. **PrettyBox :**
   Display multiple sequence alignments in PostScript format.
9. **ProfileGap:**
   Make an optimal alignment between a profile and one or more sequences.
10. **ProfileMake:**
    Create a position-specific scoring table, called a profile.

PILEUP

PileUp creates a multiple sequence alignment from a group of related sequences by using a simplification of the progressive alignment method of Feng and Doolittle.

```
           1                                                  50
S11448     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MTFD GAIGIDLGTT YSCVGVWQNE
S06443     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MTFD GAIGIDLGTT YSCVGVWQNE
A25398     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MTYE GAIGIDLGTT YSCVGVWQNE
S06158     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MTYE GAIGIDLGTT YSCVGVWQNE
S42164     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~MS KAVGIDLGTT YSCVAHFAND
S20139     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~MS KAVGIDLGTT YSCVAHFSND
B36590     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~MS KAVGIDLGTT YSCVAHFAND
A25089     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~MAKSEG PAIGIDLGTT YSCVGLWQHD
S03250     ~~~~~~~~~~ ~~~~~~~~~~ ~~~MAGKGEG PAIGIDLGTT YSCVGVWQHD
A27077     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MSKG PAVGIDLGTT YSCVGVFQHG
S07197     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MSKG PAVGIDLGTT YSCVGVFQHG
A25646     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~MSGKG PAIGIDLGTT YSCVGVFQHG
S10859     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~MSARG PAIGIDLGTT YSCVGVFQHG
A29160     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MAKA AAVGIDLGTT YSCVGVFQHG
JH0095     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MAKN TAIGIDLGTT YSCVGVFQHG
A03310     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~MATKG VAVGIDLGTT YSCVGVFQHG
JT0285     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~MSKH NAVGIDLGTT YSCVGVFMHG
```

Sequence Files for PILEUP

gcg 1% pileup
gcg 2% Pileup of what sequences ?

(1) Use wild cards
Ex:mouse.psq, rat.psq, human.psq, chicken.psq
→ *.psq
Ex:pkc.mouse, pkc.rat, pkc.human, pkc.chicken
→ pkc.*

Preparing an Alignment as a Figure

SeqWEB

Save as html format

Done by hand with a word processor

Transfer *.pair or *.msf files to PC
Set font to Courier or other fixed spacing font
Use shaded boxes to highlight important domains
Use color sparingly, red for the most important feature

GenDoc

A free msf file viewer and editor.

Download pileup.msf

Download and decompress gd322602.exe

EXERCISE 04-2

PileUP

"fetch" the following sequences:

capzb_chick

capzb_mouse

capzb_human

capzb_caeel

-->Perform pileup capzb_*.*

-->call the output display file fetch.msf *ANSWER*

-→cat fetch.msf

Download fetch.msf to PC & open with Gendoc

EXERCISE 04-3

Pretty and Prettybox

(A)Use "Pretty" to display *.msf files
-->pretty fetch.msf{*}
-->call the output display file fetch.pretty
-->cat fetch.pretty

(B)Use "Prettybox" to display pretty result
-->prettybox fetch.msf{*}
-->call the output display file fetch.ps
-->use FTP to transfer file to you PC

ANSWER

(C) Msf file viewers
1. MS-Word
2. GenDoc

DNA vs Protein Sequence

Why do people suggest that translated sequences be used to search for relatives in databanks?

DNA is composed of only four kinds of units -A, G, C and T- and even if gaps were not allowed, it would be anticipated that, on the average, 25% of the residues of any two aligned sequences would be identical. In fact, there would be a dispersion around the mean expectation, and a predictable fraction of random cases would be as much as 35% identical. Once we decide to allow gaps in the sequences, then the range of chance similarities between two unrelated sequences can exceed 50%, thereby obscuring any genuine relationships that may exist.

Nucleotide sequence alignment

Why Protein Sequence

Why do people suggest that translated sequences be used to search for relatives in databanks?

Protein sequences are composed of a 20 aa alphabet determined by 61 degenerate codons. When the DNA sequences are translated into 21 different types of codons (20 aa and a terminator), the information is sharpened up considerably. The 'wrong-frame' information is discarded, and third-base degeneracies are consolidated. All in all, the signal-to-noise ratio is greatly improved for the specific purpose of identifying protein relatives. It is accepted that convergence phenomena in aa sequences are very rare and thus aa similarity almost always means homology. Furthermore, aa sequences may still show a similarity derived from common folding patterns and function of the proteins, even while their coding DNA sequences might have strongly diverged due to other selective pressures existent at the genome level (e.g., G+C pressure, preferential usage of synonymous codons, etc.). Protein evolution is governed by the constraint of maintaining a characteristic fold which enables some function. Thus, it is possible to infer relationships between proteins that last shared a common ancestor 1-2.5 billion years ago by conducting protein searches, doubling the lookback time obtained performing DNA database searches.

BLAST vs FASTA

FASTA - a sensitive search engine

The early personal computers had insufficient memory and were too slow to carry out a database scan using a rigorous searching method (dynamic programming). Accordingly, Wilbur and Lipman [(1983) Proc. Nat. Acad. Sci. 80, 726-730] developed a fast procedure for DNA scans that in concept searches for the most significant diagonals in a dot plot. FASTA only shows the top scoring region, it does not locate all high scoring alignments between two sequences. As a consequence, FASTA may not directly identify repeats or multiple domains that are shared between two proteins

BLAST - a faster alternative

BLAST (Basic Local Alignment Search Tool) is a heuristic method to find the highest scoring locally optimal alignments between a query sequence and a database. Previous versions of BLAST did not allow gapped alignments, but BLAST2 (from the HGMP-RC telnet and www menus) does. A gapped BLAST search allows gaps (deletions and insertions) to be introduced into the alignments that are returned. Allowing gaps means that similar regions are not broken into several segments. The scoring of these gapped alignments tends to reflect biological relationships more closely.

The BLAST Family

Program	QUERY	Database
blastp	amino acid sequence	protein sequence database.
blastn	nucleotide sequence	nucleotide sequence database.
blastx	nucleotide sequence translated in all reading frames	protein sequence database (use this option to find potential translation products of an unknown nucleotide sequence)
tblastn	amino acid sequence	nucleotide sequence database translated in all reading frames
tblastx	six-frame translations of a nucleotide sequence	six-frame translations of a nucleotide sequence database. (tblastx program cannot be used with the nr database on the BLAST Web page because it is computationally intensive)

The FASTA Family of Programs

FastA : uses the method of Pearson and Lipman (Proc. Natl. Acad. Sci. USA 85; 2444-2448 (1988)) to search for similarities between one sequence (the *query*) and any group of sequences of the same type (nucleic acid or protein) as the query sequence.

TFastA : treats each of the six reading frames of a query nucleotide sequence as a separate sequence, resulting in three separate alignments for each strand.

TFastX : compares the protein query sequence to only one translated protein per strand of the nucleotide sequence, resulting in one alignment per strand.

SEARCHING in SeqWEB/GCG

Reference Searching

1. LookUp - Identifies sequences in sequence database (name, accession number, author, et al..)
2. Names - Identifies sequences entries by name.
3. StringSearch - Identifies sequences by character patterns.

Sequence Searching

1. BLAST - Finds sequences in a database that are similar to a query sequence (ver.2.0)
2. FastA - Search for similarity sequences of the same type
3. FastX - Search for similarity sequences between a nucleotide sequence and protein database, taking frameshifts into account.
4. FindPatterns - Identifies sequences with short sequence pattern
5. FrameSearch - Search protein sequences for similarity to nucleotide query sequences, or nucleotide sequences for similarity to protein query sequences.
6. Motifs - Search through proteins for the patterns defined in the PROSITE.
7. MotifSearch - Use a set of profiles search a database for new sequences.
8. NetBLAST - Search database maintained at NCBI
9. ProfileSegments - Make optimal alignments found by ProfileSearch.
10. ProfileSearch - Use a profile to search the database for new sequence.
11. Segments - Aligns and displays the segments found by WordSearch.
12. Ssearch - Does a rigorous Smith-Waterman search for similarity
13. TFastA - Search for similarity sequences between a protein sequence and nucleotide database
14. TFastX - Search for similarity sequences between a protein sequence and nucleotide database, taking frameshifts into account.
15. WordSearch - Identifies sequences in the database that share large numbers of common words

NCBI Blast vs GCG Blast

Download CDK2 amino acid sequence

Copy & Paste

Upload to GCG
Reformat
GCG> blast -BAT

NCBI Blast

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

WWW system
Larger database
Interlinked Data

Slow
Single search only

GCG Blast

Unix system
Smaller database
Data not interlinked

Built your own database
Fast
Support multiple search
Output file easier to parse

Exercise 04-4

GCG Blast

```
v8803: petang [users/petang]>blast -BAT

BLAST searches one or more nucleic acid or protein databases
for sequences similar to one or more query sequences of any
type. BLAST can produce gapped alignments for the matches it
finds.

 BLAST with what query sequence(s) ? x15785.gb_ov

                Begin (* 1 *) ?
              End (*   515 *) ?

 Search for query in what sequence database:

   1) uniprot   p uniprot
   2) genbank   n GenBank
   3) genpept   p GenPept (Translated GenBank)
   4) htg       n High Throughput Genomes (HTG from GenBank and EMBL)
   5) rs_rna    n Refseq RNA
   6) rs_prot   p Refseq Prot
   7) est_human n Human Expressed Sequence Tags (GenBank and EMBL)
   8) est_mouse n Mouse Expressed Sequence Tags (GenBank and EMBL)
   9) est_other n All Other Expressed Sequence Tags (GenBank and EMBL)
  10) gss       n Genome Survey Sequences (GSS from GenBank and EMBL)
  11) htc       n HTC

Please choose one (* 1 *):  2
```

```
 Ignore hits expected to occur by chance more than (* 10.0 *) times?

  Limit the number of sequences in my output to (* 500 *) ? 10

  What should I call the output file (* x15785.blastn *) ?

  ** blast will run as a batch or at job.

  ** blast was submitted using the command:
     "  at now  "

v8803: petang [users/petang]> ls
blast_21760_1         blast_21760_1.init     x15785.gb_ov

v8803: petang [users/petang]>ls
x15785.blastn     x15785.gb_ov

v8803: petang [users/petang]>more x15785.blastn
```

Exercise 04-5

(1) What is cdk2?

-search UNIGENE, OMIM…..

(2) How many cdk2 proteins already discovered in different organisms?

-try ENTREZ protein,
-start search protein for “cdk2”, then “cyclin dependent kinase 2”
-search again with the same keywords but limit to “protein name”.

(3) Display & Save the sequences in NCBI

-DISPLAY the “cdk2” sequences (limit to protein name) in fasta format (xx sequences)
-SAVE ALL THE SEQUENCES in FASTA with the file name cdk2-psq.fasta
-Upload cdk2.txt and cdk2-psq.fasta to GCG
-Change to GCG format
cdk2.txt and
cdk2-psq.fasta (ALL SEQUENCES IN THE FILE WILL BE REFORMATED)

Build Your Own Database

Search for human cdk2 proteins in NCBI
Transfer to GCG and change to GCG format

formatdb+ combines any set of GCG sequences into a database that you can search with BLAST.

formatdb+ of what input sequence(s) ? *.pep
What should I call the database ? cdk2psq

Change xp132341 to gcg format

blast -BAT -IN2=cdk2psq

BLAST searches one or more nucleic acid or protein databases for sequences similar to one or more query sequences of any type. BLAST can produce gapped alignments for the matches it finds.
Blast with what query sequence(s) ? **"just pick one of the cdks sequences that you uploaded"**

ASSIGNMENT 01

Use the database searching techniques you learned today to retrieve the **amino acid sequences** of

Human (Homo sapiens) Vacuolar ATP synthase

Question:

(1) How many human V-ATP synthase deposited in NCBI

(2) Built a V-ATP synthase database in GCG

→ download this sequence [vatpase.txt]

→ TELL ME WHICH SEQUENCE IN YOUR DATABASE MATCHES BEST

E-mail the ANSWER as attached files to

--petang@mail.cgu.edu.tw. before 12:00 19 Oct 2017

****郵件主旨： ASS01 bioinfo – (學號)