Sequence Analysis (part III)

BBSI 2006: Lecture #(χ+3)

Takis Benos (2006)

Outline

- Sequence variation
- Distance measures
- Scoring matrices
- Pairwise alignments (global, local)
- Database searches (BLAST, FastA)
- Multiple sequence alignments

Database Searches

Database search: why?

- Database searching is the first step in characterizing a newly discovered gene.
- It helps determining the function and the evolutionary relationships.
- It answers the question: "*Has anyone seen anything like that before?*"

Database search

- Database searching consists of **many** pairwise alignments combined in one search.
- Heuristic algorithms are used instead of DP. *Why?*
  - Size of SWISS-PROT + TrEMBL: 1M entries or 344M residues.
  - Exact algorithms are O(*nm*) fast.
- The goal of the heuristic methods is to look at a small fraction of the searching space that will include all (or most) of the high scoring pairs.

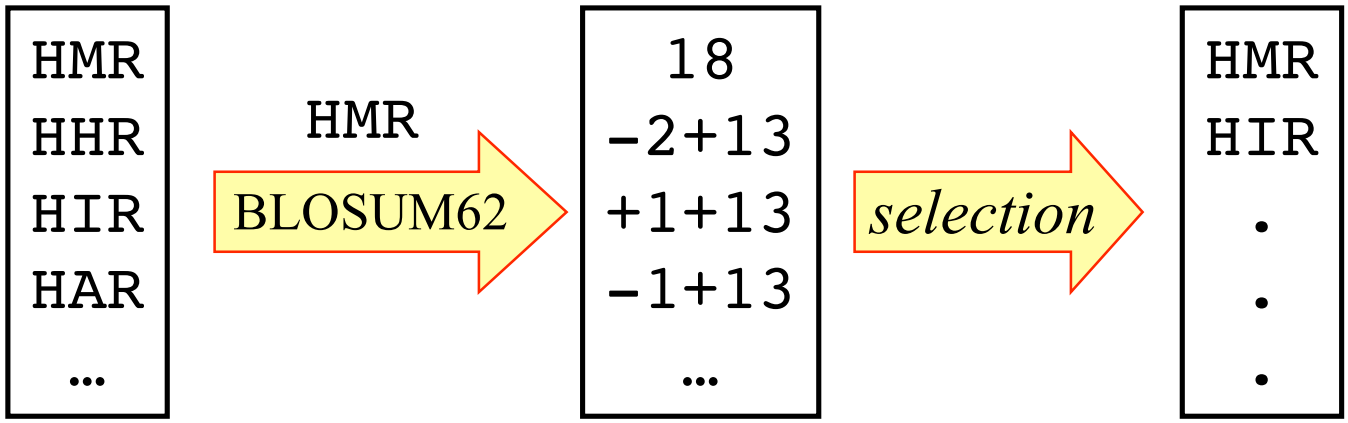
BLAST algorithm

- **B**asic **L**ocal **A**lignment **S**earch **T**ool - The method:
  - For each (fixed-length) “word” in the query sequence, make a list of all neighbouring “words” that score above some threshold.
  - Scan the database for these words.
  - Perform (ungapped) “hit extension”.
  - Stop at maximum scoring extension.

BLAST algorithm (cntd)

- An example:

```
Query:    CPICHRAFHRLEHQTRHMRIHTGEKPHAC
```



BLAST algorithm (cntd)

- An example:

```
Query:   CPICHRAFHRLEHQTRHMRIHTGEKPHAC
                           H+R
Sbjct:   CPLCDKAFHRLEHQTRHIRTHTGEKPHAC
```

BLAST algorithm (cntd)

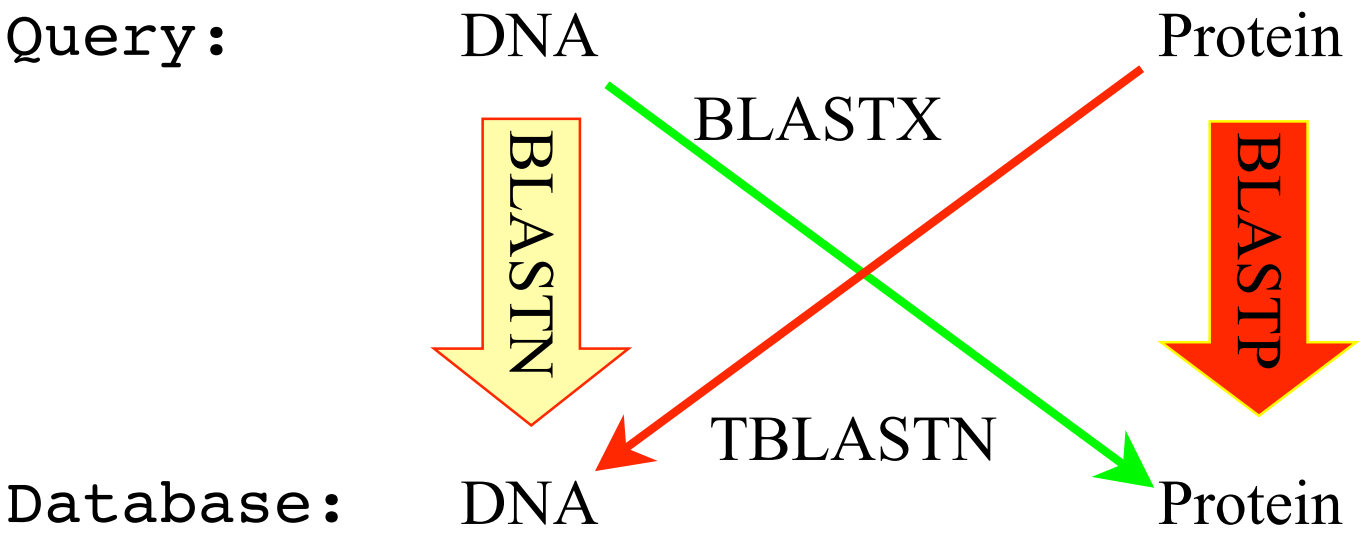
- An example:

```
Query:    CPICHRAFHRLEHQTRHMRIHTGEKPHAC
          CP+C +AFHRLEHQTRH+R HTGEKPHAC
Sbjct:    CPLCDKAFHRLEHQTRHIRTHTGEKPHAC
```

BLAST algorithm (cntd)

- The idea: a high scoring match alignment is very likely to contain a short stretch of very high scoring matches.
- Word length: 3 (proteins) and 11 (DNA).
- HSSP: multiple HSSPs can be reported for each database entry.
- Gapped alignments: more recently, BLAST versions perform gapped alignments.

BLAST flavours



TBLASTX: DNA Query to DNA Database *via* translation

Database searching programs

	Program	Query	Database	Examples of usage
1.	BLASTN	DNA	DNA	identical/closely related genes
2.	BLASTP	Prot	Prot	general use program *(recom.)*
3.	BLASTX	DNA(*)	Prot	find exons in your genomic seq.
4.	TBLASTN	Prot	DNA(*)	find the location/structure of your gene in the genome
5.	TBLASTX	DNA(*)	DNA(*)	*nothing really….*

(*) translated query/database

BLAST tutorial:
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

BLAST output

```
BLASTP 2.2.5 [Nov-16-2002]
Reference :Altschul, Stephen F., Thomas L. Madden,
Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb
Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein
database search
programs",  Nucleic Acids Res. 25:3389-3402.

RID: 1050176602-03205-14137

Query= gi|127091|sp|P27705|MIG1_YEAST Regulatory protein
MIG1 (Regulatory protein CAT4).
         (504 letters)

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF
          1,411,415 sequences; 454,141,287 total letters
```

BLAST output (cntd)

Sequences producing significant alignments:	Score (bits)	E Value
gi|6321403|ref|NP_011480.1| Transcription factor involved i...	549	e-15
gi|3437|emb|CAA39084.1| MIG1 [Saccharomyces cerevisiae]	441	e-12
gi|1709031|sp|P52288|MIG1_KLUMA Regulatory protein MIG1 >gi...	106	1e-2
gi|1709030|sp|P50898|MIG1_KLULA Regulatory protein MIG1 >gi...	104	3e-2
gi|416840|sp|Q01981|CREA_EMENI DNA-binding protein creA (Ca...	99	1e-19
gi|101802|pir||A41694 regulatory protein creA - Emericella ...	99	1e-19
gi|544095|sp|Q05620|CREA_ASPNG DNA-binding protein creA (Ca...	99	2e-19
gi|2293072|emb|CAA04425.1| carbon catabolite repressor CRES...	99	2e-19
gi|12229763|sp|Q9P889|CREA_ASPOR DNA-binding protein creA (...	99	2e-19
gi|12229746|sp|O94130|CREA_BOTCI DNA-binding protein creA (...	99	2e-19

BLAST output (cntd)

```
>gi|3437|emb|CAA39084.1| MIG1 [Saccharomyces cerevisiae]
Length = 386 Score =  441 bits (1133), Expect = e-122 Identities = 262/380
(68%), Positives = 262/380 (68)%

Query: 1   MQSPYPMTQVSNVDDGXXXXXXXXXXXXXXXXXPRPHACPICHRAFHRLEHQTRHMRIH 60
           MQSPYPMTQVSNVDDG                 PRPHACPICHRAFHRLEHQTRHMRIH
Sbjct: 1   MQSPYPMTQVSNVDDGSLLKESKSKSKVAAKSEAPRPHACPICHRAFHRLEHQTRHMRIH 60

Query: 61  TGEKPHACDFPGCVKRFSRSDELTRHRRIHTNSHPXXXXXXXXXXXXXXXXXXXXXXXX 120
           TGEKPHACDFPGCVKRFSRSDELTRHRRIHTNSHP
Sbjct: 61  TGEKPHACDFPGCVKRFSRSDELTRHRRIHTNSHPRGKRGRKKKVVGSPINSASSSATSI 120

Query: 121 XDLNTANFXXXXXXXXXXXXXXXAIAPKENXXXXXXXXXXXXFEIGESGGNDPYMVSSP 180
            DLNTANF               AIAPKEN             FEIGESGGNDPYMVSSP
Sbjct: 121 PDLNTANFSPPLPQQHLSPLIPIAIAPKENSSRSSTRKGRKTKFEIGESGGNDPYMVSSP 180
```

```
>gi|8926704|emb|CAB96530.1| MIG repressor [Pichia jadinii]
Length = 345 Score = 85.9 bits (211), Expect = 2e-15 Identities = 42/58
(72%), Positives = 49/58 (84%), Gaps = 2/58 (3%)

Query: 36 RPHACPICHRAFHRLEHQTRHMRIHTGEKPHACDFPGCVKRFSRSDELTRHRRIHTNS 93
          RP+ C +C++AFHRLEHQTRHMRIHTGEKP  C F  C K+FSRSDELTRH RIH+N+
Sbjct: 17 RPYVCTVCNKAFHRLEHQTRHMRIHTGEKPFQCTF--CSKKFSRSDELTRHTRIHSNT 72
```

Other database searching programs

- MEGABLAST.

  It can be used for comparing two large sets of sequences with each other.
- PSI-BLAST (*Position-Specific Iterated*).

  It performs iterative searches; the sequences found in one searching round are used to build models for searching in the next round. To be used when seeking increased sensitivity.

FASTA algorithm

- The method:
  - For each pair of sequences (query, subject), identify all identical "word" matches of (fixed) length.
  - Look for diagonals with many mutually supporting "word" matches.
  - The best diagonals are used to extend the word matches to find the maximal scoring (ungapped) regions.

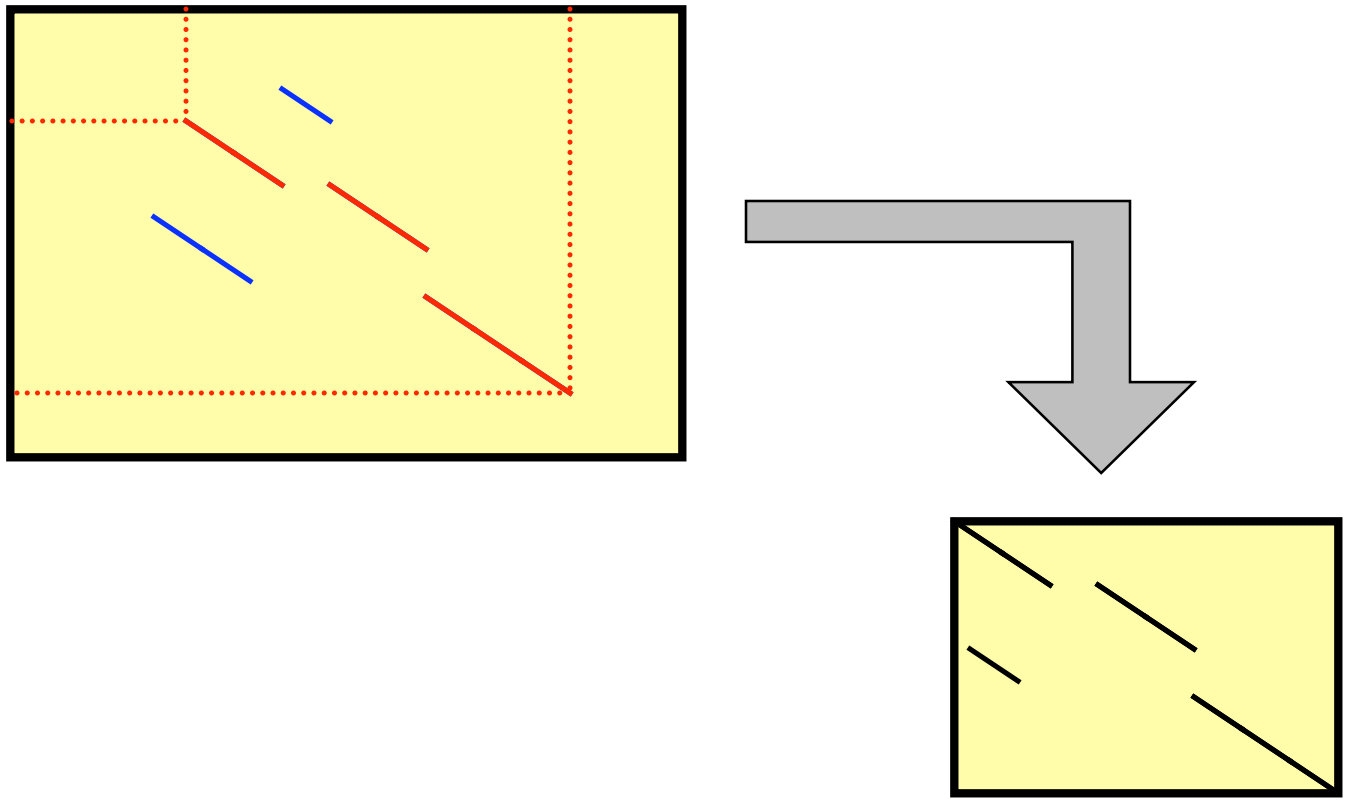
FASTA algorithm (cntd)

- The method:
  - Join ungapped regions, using gap costs.
  - Align the two (sub)regions using full dynamic programming techniques.

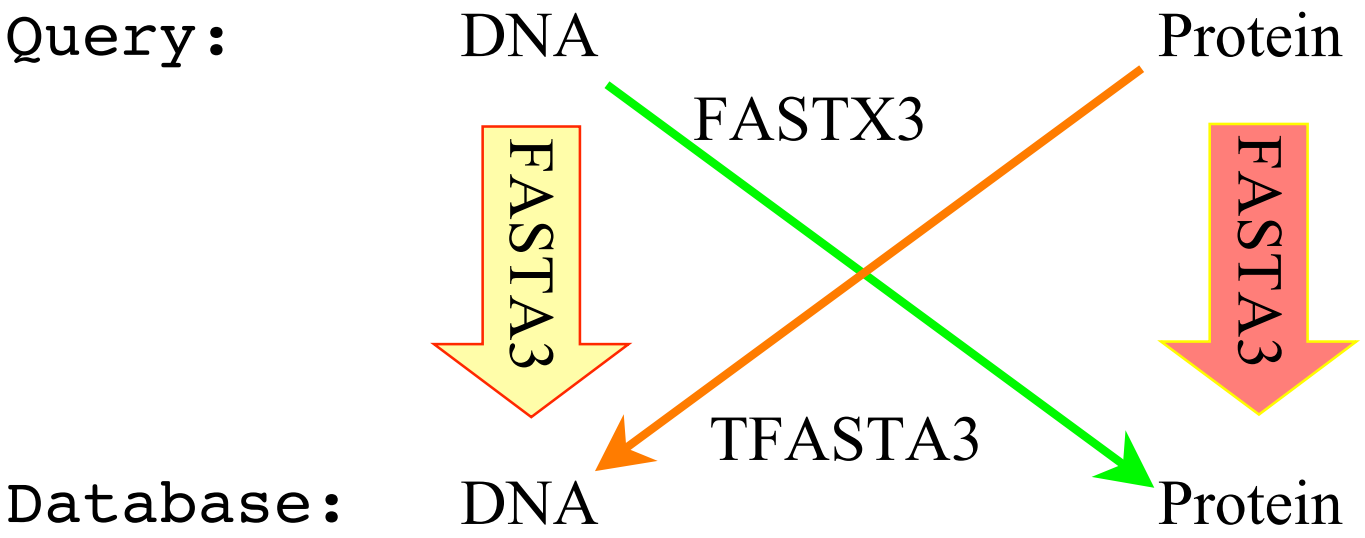
FASTA algorithm (cntd)

- The idea: a high scoring match alignment is very likely to contain a short stretch of identities.
- Word length: 2 (proteins) and 4-6 (DNA).
- HSSP: usually one (extended) gapped alignment is presented.

FASTA algorithm (cntd)



FASTA flavours



DNA or protein?

- It depends…

  Question: What is more conserved?

```
ATG aat cgt ctt att gaa
 M   N   R   L   I   E
ATG aag agg ttg ata gag
```

DNA or protein?

- It depends…

  Question: What is more conserved?

```
ATG aat cgt ctt att gaa
||| ||   |   |  ||  ||
ATG aag agg ttg ata gag
```

DNA or protein? (cntd)

- Some facts:
  - DNA sequences generally change quicker than the protein sequences.
  - DNA databases are larger than the protein ones (e.g. human genome: 2.9 billion bases; SWISS-PROT+TrEMBL: 1 million a.a.)
  - DNA: 4 symbols; a.a.: 20 symbols

DNA or protein? (cntd)

- So…
  - DNA searches have lower signal to noise ratio.
- However…
  - …they can still be useful in searching for closely related genes and establishing evolutionary relationships.
  - More sensitive in EST hunting.

Some links…

- BLAST programs on the web (NCBI), includes PSI-BLAST and PHI-BLAST:
  http://www.ncbi.nlm.nih.gov/BLAST/
- BLAST parameters:
  http://www.ncbi.nlm.nih.gov/BLAST/newoptions.html
  http://blast.wustl.edu/blast/parameters.html
- BLAST help:
  http://www.ncbi.nlm.nih.gov/blast/html/BLASThomehelp.html

Some links…(cntd)

- BLAST program selection guide:
  http://www.ncbi.nlm.nih.gov/BLAST/producttable.html
- BLAST tutorials:
  http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

Some links…(cntd)

- FASTA programs on the web (EMBL-EBI and DDBJ):

  http://www.ebi.ac.uk/fasta33/

  http://gib.genes.nig.ac.jp/single/fasta3/main.php

- FASTA parameters/help:

  http://fasta.genome.ad.jp/dbget-bin/show_man?fasta3

Multiple sequence alignments - sequence evolution

Background

- Proteins are related to each other through evolution.
- There is a unique true underlying evolutionary tree...
- …but we do not know it!
- There is no objective way to define the "correct" alignment, for the interesting cases (i.e. ~30% average a.a. identity pairwise).

Background (cntd)

- Different parts of the proteins have different evolutionary constraints.
- Multiple alignment methods, in principle, can identify the better conserved regions.
- Ideally, the amino acids in a multiple alignment column occupy similar three-dimensional positions in the folded protein.

Multiple alignment scoring

- Complete probabilistic model:

  Impractical (very complex; not enough data).
- Simplifying assumptions:
  1. Individual columns are statistically independent.
  2. Residues *within* the column are considered independent (i.e. information on phylogeny is ignored).

Multiple alignment scoring (cntd)

Method-1: minimum entropy

$$Sc(alignment) = Sc(gaps) + \sum_i Sc(col_i)$$

$$Sc(col_i) = - \sum_a c_a(i) \log p_a(i)$$

$$p_a(i) = c_a(i) / N(i)$$

Multiple alignment scoring(cntd)

Method-2: sum of pairs (SP)

$$SP(i) = Sc(col_i) = \sum_l \sum_{k<l} Sc_i(k,l)$$

BLOSUM

Problem: Compare SP scores

- N sequences with *Arg* at position *i*.
- N-1 sequences with *Arg* and one with *Lys*.

Multiple alignment scoring(cntd)

N sequences with *Arg* at position *i*.

- BLOSUM62: Sc(*Arg*, *Arg*) = 5
- SP1 = 5 x N (N - 1) / 2

N-1 sequences with *Arg*, one with *Lys*.

- BLOSUM62: Sc(*Arg*, *Lys*) = 2
- SP2 = SP1 - 3 x (N - 1)
- (SP1-SP2)/SP1 = 6/5N !!

Methods

- A naïve approach:

  Use dynamic programming to calculate all possible alignments of the N sequences of length L and choose the best.

- Problem:
  - Memory complexity $O(L^N)$, time complexity $O(2^N L^N)$.

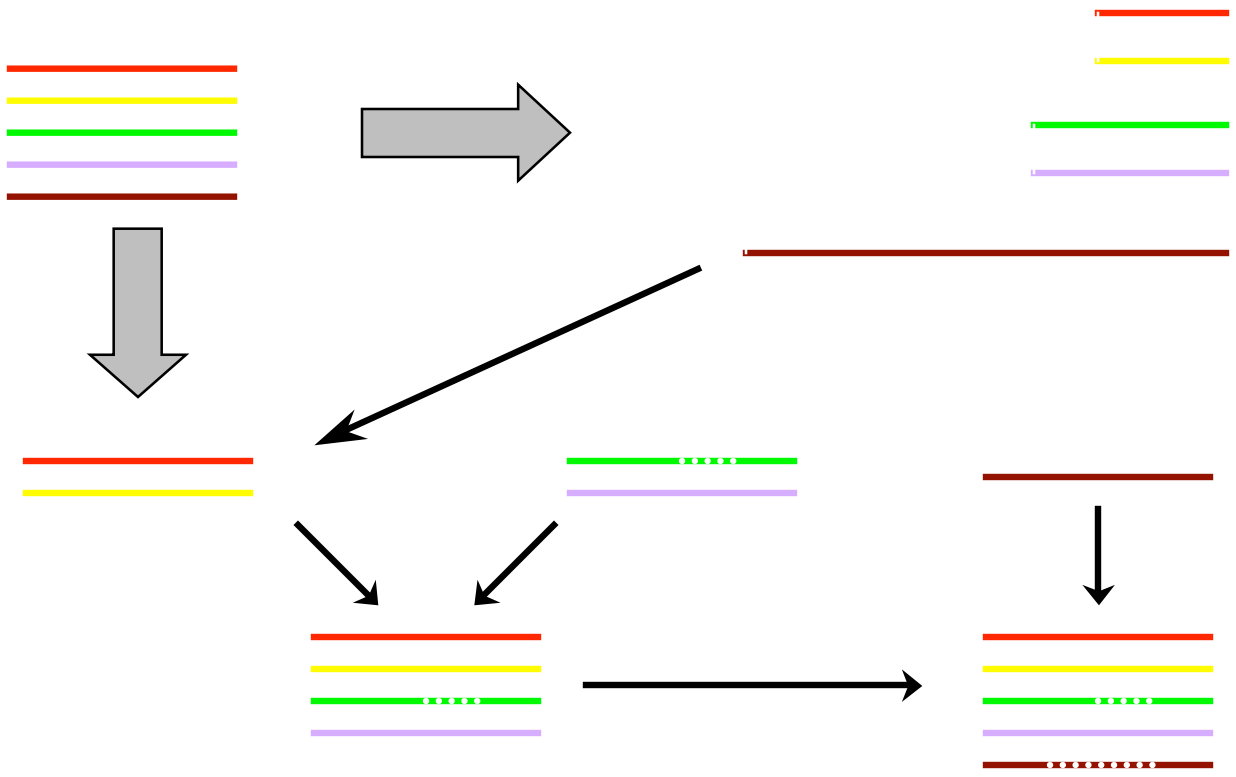
Methods (cntd)

- Example:
  - Aligning N sequences requires $(2L)^{N-2}$ pairwise comparisons.
  - You have 15 sequences, 50 a.a. long.
  - Your computer needs 1 sec for each pairwise comparison.
  - How many sequences you'll align until the end of our sun? (i.e. approx. 5 billion years)

Progressive algorithms

- General idea:
  - Calculate all pairwise alignments.
  - Cluster the sequences according to some scoring scheme.
  - Align the two closest sequences; fix their alignment.
  - Continue with next sequence and/or alignment, until all sequences are aligned.

Progressive algorithms (cntd)



Feng-Doolittle

- [Feng & Doolittle, 1987]:
  - Calculate a “distance matrix”, using all pairwise scores.
  - Construct a *guide tree* from this distance matrix.
  - Starting from the first node added to the guide tree, align the child nodes.

Feng-Doolittle (cntd)

- [Feng & Doolittle, 1987]:
  - Repeat for other nodes in the order they were added to the tree.
  - Each new sequence is added after compared to *every* sequence in the current alignment.
  - When an alignment is added, there is an all-to-all comparison.

CLUSTALW

- [Thompson, Higgins & Gibson, 1994]:
  - Similar to Feng-Doolittle.
  - Uses Kimura's model for the evolutionary distance and NJ algorithm to construct the tree.
  - Builds profiles and aligns the profiles.
  - Sequences are weighted to compensate for biased representation.

CLUSTALW (cntd)

- [Thompson, Higgins & Gibson, 1994]:
  - Uses a variety of scoring substitution matrices, depending on the expected similarity.
  - Penalties for gaps and mismatches are varying, depending on the position of the alignment that they occur.
  - The guide tree can be re-adjusted on the fly, if the score of the alignment becomes very low.

Barton-Sternberg

- [Barton & Sternberg, 1987]:
  - Find the two sequences with the highest pairwise score; build a profile.
  - Find the sequence that is closest to this profile; align it to it.
  - Repeat until all sequences have been aligned to a single profile.

Barton-Sternberg (cntd)

- [Barton & Sternberg, 1987]:
  - Remove sequence-1 and re-align it to the profile; calculate the new score.
  - Repeat with sequence-2, etc.
  - Repeat the procedure a fixed number of times, or until convergence occurs (i.e. score doesn't change).

Comments

- Unlike pairwise alignments, multiple alignment methods **are not** guaranteed to find the optimal alignments.

Multiple alignments: general (cntd)

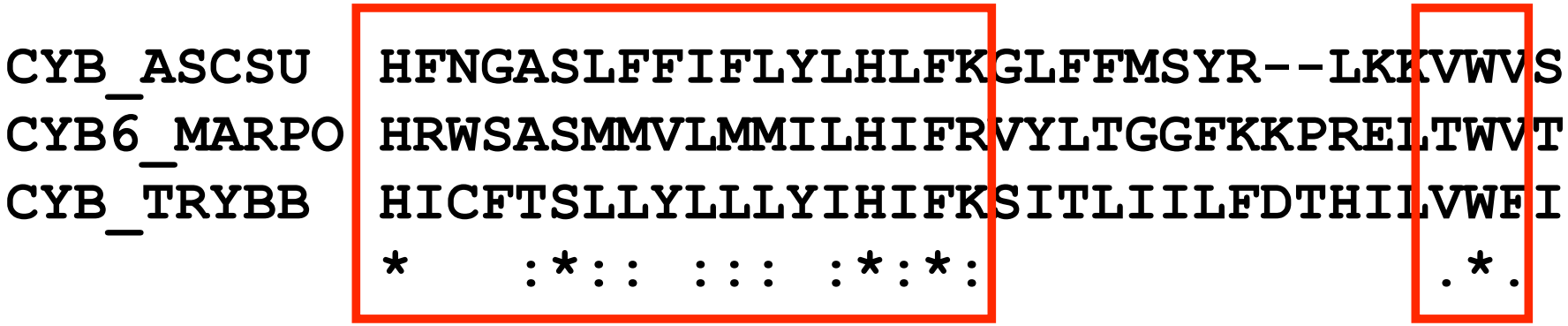
```
CYB_ASCSU   HFNGASLFFIFLYLHLFK
CYB6_MARPO  HRWSASMMVLMMILHIFR
CYB_TRYBB   HICFTSLLYLLLYIHIFK
            *    :*:: ::: :*:*:

CYB_ASCSU   GLF....FMSY..RLKK..VWVS
CYB6_MARPO  VYL....TGGFKKPREL..TWVT
CYB_TRYBB   SITLIILFDTH..IL....VWFI
                                .*.
```

Manually curated (Pfam):

http://pfam.wustl.edu/cgi-bin/getdesc?name=cytochrome_b_N

Multiple alignments: general (cntd)



Automatically aligned: CLUSTALW

Comments

- Unlike pairwise alignments, multiple alignment methods are not guaranteed to find the optimal alignments.
- Multiple alignments are used to calculate profiles characteristic for protein families.
- These profiles can be used to identify new (distant) members of these families.

Comments (cntd)

- E.g., if your BLAST searches yield many poor(ish) results, profile searches might hint the function of your newly sequenced gene.
- Also, you can align all the top hits of your BLAST search, to create a profile and check if your sequence belongs to this profile.

Resources

- CLUSTALW servers:
  - EBI: http://www.ebi.ac.uk/clustalw/
  - Baylor College of Medicine: http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html

Resources (cntd)

- Multiple alignments are the primary source of information for the *motif* databases:
  - PROSITE: http://us.expasy.org/prosite/
  - Pfam: http://pfam.wustl.edu/
  - PRINTS: http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/

Additional Reference

- Durbin, Eddy, Krogh & Mitchison, "*Biological Sequence Analysis*", 1998, Cambridge University Press