Sequence Analysis and Alignment

Definition of Sequence Alignment

Computational procedure (“algorithm”) for comparing two/many sequences

- identify series of identical residues or patterns of identical residues that appear in the same order in the sequences
- visualized by writing sequences as follows:

```
MLGPSSKQTGKGS-SRIWDN*
||     |   |||    |   |
MLN-ITKSAGKGAIMRLGDA*
```

Pairwise Global Alignment
(over whole length of sequences)

```
    GKG
    |||
    GKG
```

Pairwise Local Alignment
(similar parts of sequences)

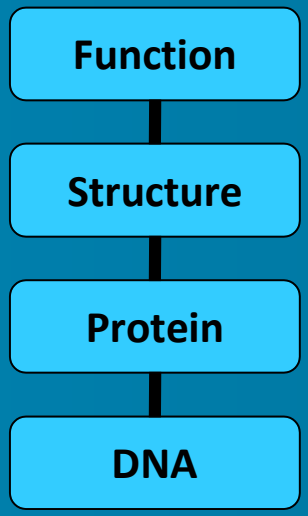
- **sequence alignment is an optimiztion problem**
  bringing as many identical residues as possible into corresponding positions

Algorithms for Local Sequence Aignments

- **Sequence Similarity and Homology**
  - Origins of homology
  - Sequence alignment
  - Global Alignment
  - Local Alignment
- **Content of Sequence DBs**
  - GenBank, SwissProt, RefSeq
  - Size of sequence DB requires special search tools
- **Algorithms for searching Sequence Databases**
  - Basics of sequence DB searches
  - Efficient detection of identical k-mers
  - BLAST2 improvements
  - Statistical significance of hits

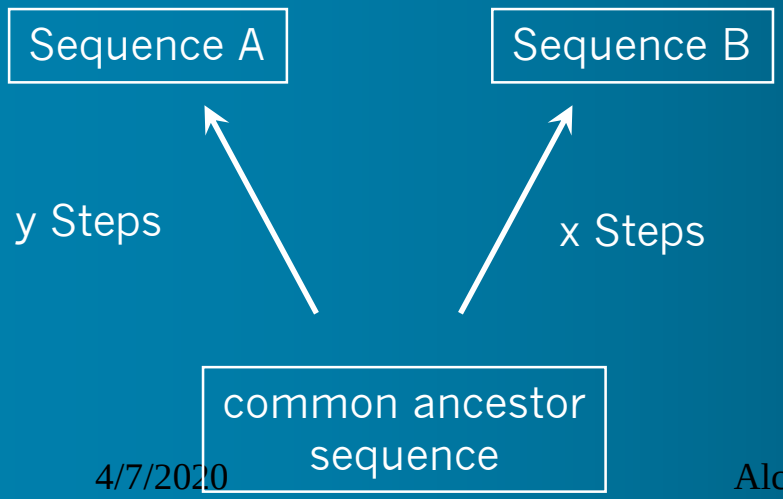
Outline follows: David W. Mount, "Bioionformatics - Sequence and Genome Analysis“ Cold Spring Harbour Laboratory Press, 2001. Online: http://www.bioinformaticsonline.org

Rational for Sequence Analysis, Origins of Sequence Similarity



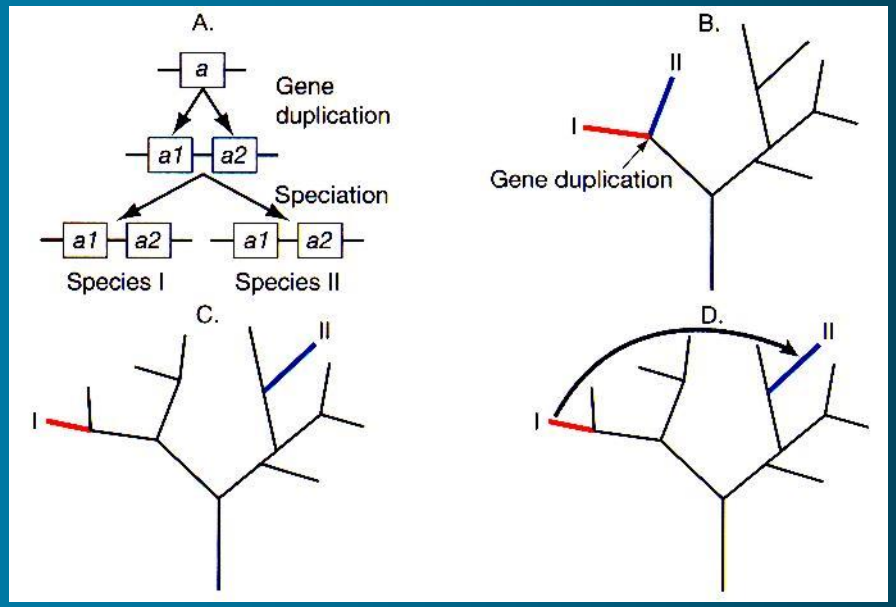
Similar sequence leads to similar function

Sequence Analysis as the basic tool to discover functional, structural, evolutionary information in biological sequences



Evolutionary relationship between two similar sequences and a possible common ancestor. The number of steps to convert one sequence into the other is the "evolutionary" distance between the sequences (x + y). Usually, the ancestor sequence is not available, only (x + y) can be computed.

Origins of Homology → Significance of Sequence Alignments



Possible Origins of Sequence Homology:

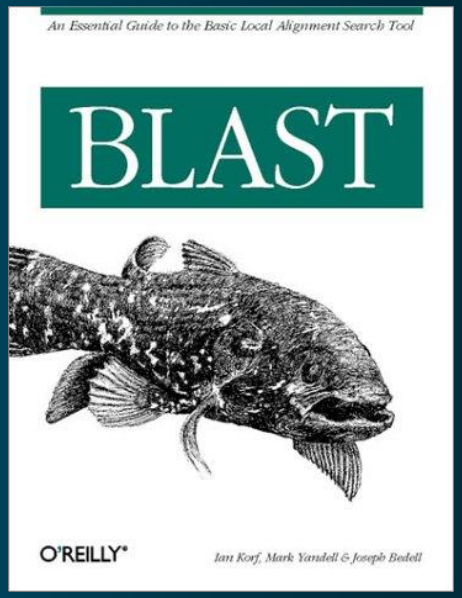
- **orthologs** (panel A and B) a1 in species I and a1 in species II (same ancestor!)
- **paralogs** (panel A and B) a1 and a2 (arose from gene duplication event)
- **analogs** (panel C): different genes converge to same function by different evolutionary paths
- **transfer of genetic material** (panel D) between different species

Homology vs. Similarity

- Similarity can be **computed** (by sequence alignments)
- Homology is deduced (e.g. from similarity, but also from other evidence!)

Basic Local Alignment Search Tool (BLAST)

- 3rd most cited paper in MEDLINE
- Most widely used program to find similar sequences within large databases
- Search flexibility enables many different kinds of match possibilities



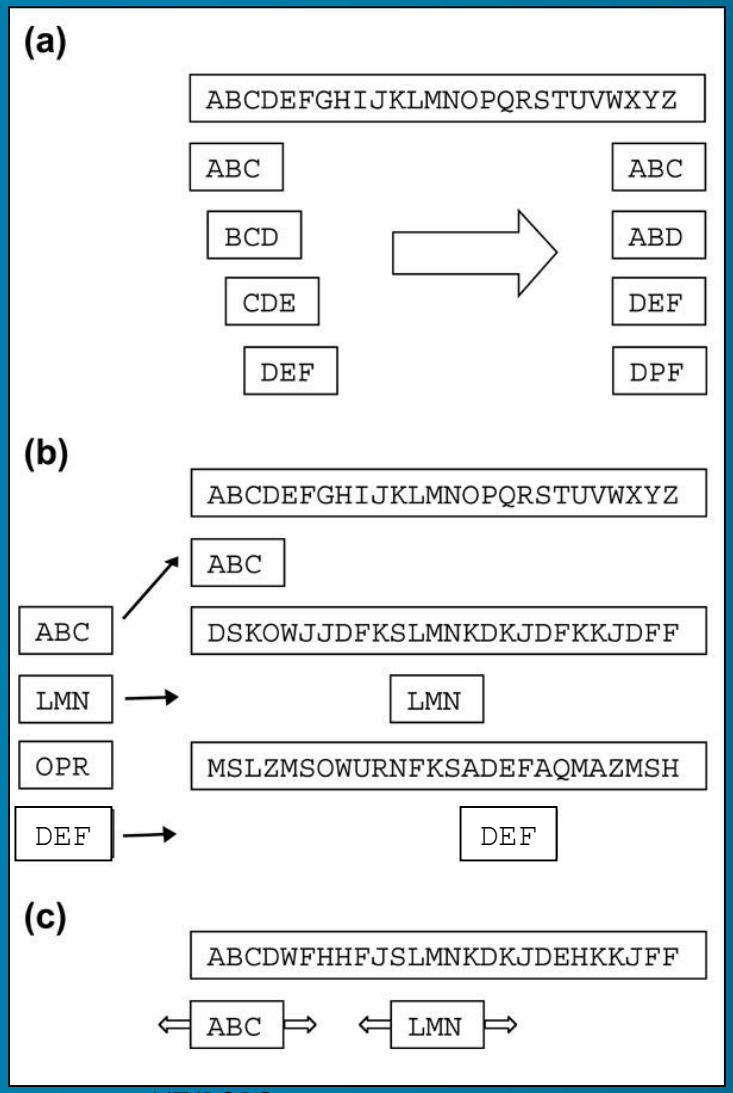
J Mol Biol 1990 Oct 5;215(3):403-10

Basic local alignment search tool.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.

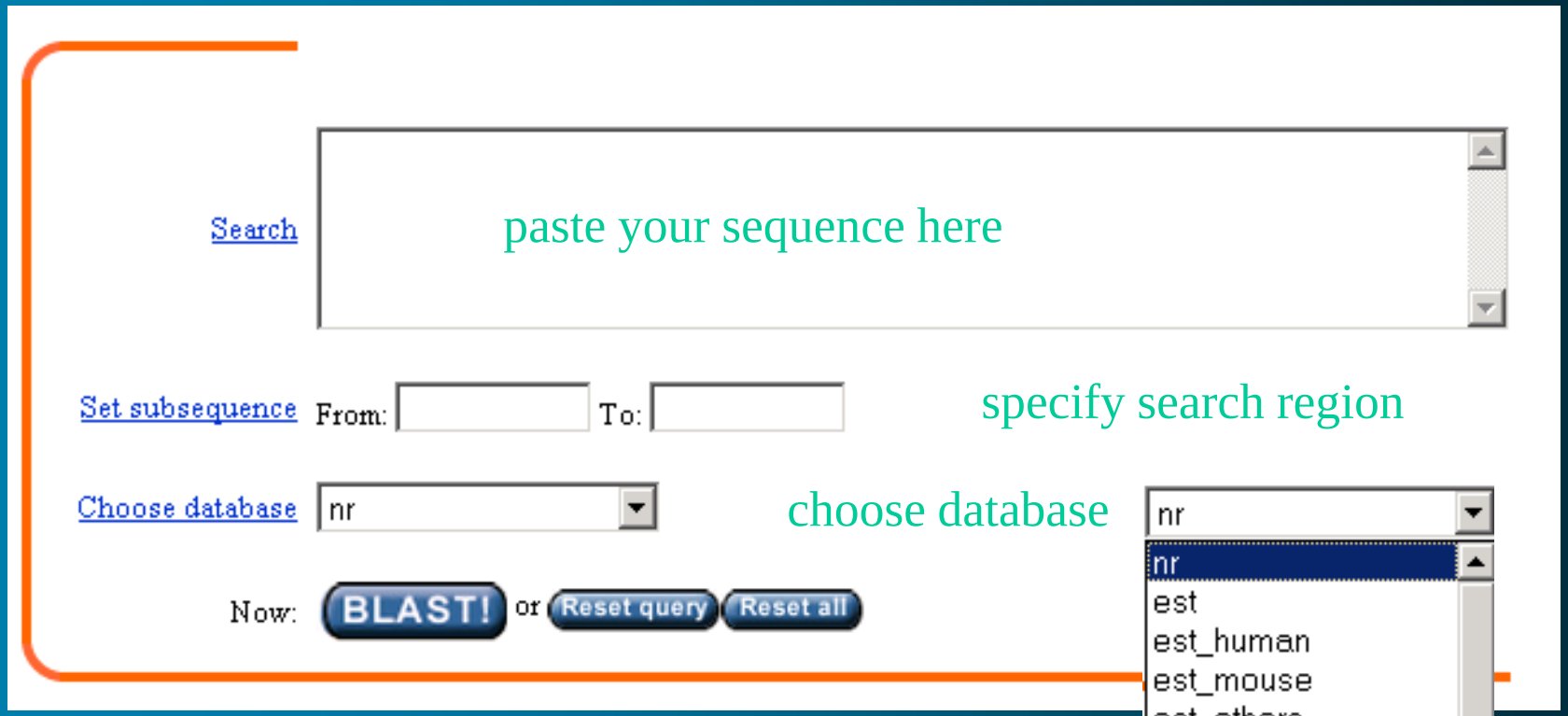
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

How BLAST works



1) Low complexity regions in query sequence are filtered
2) List of all k-tuples (words) that make up your query sequence are generated
3) Scoring matrix is used to determine all word matches above a specific threshold (about 50 matches per word)
4) Database is searched for sequences with exact matches to the word list generated (b)
5) Matches are used to seed possible alignments between the query sequence and the database (c)
6) Alignment is extended as long as score continues to increase and is retained if score is greater than empirically determined cutoff
7) The statistical significance of the score is calculated

Blastn queries



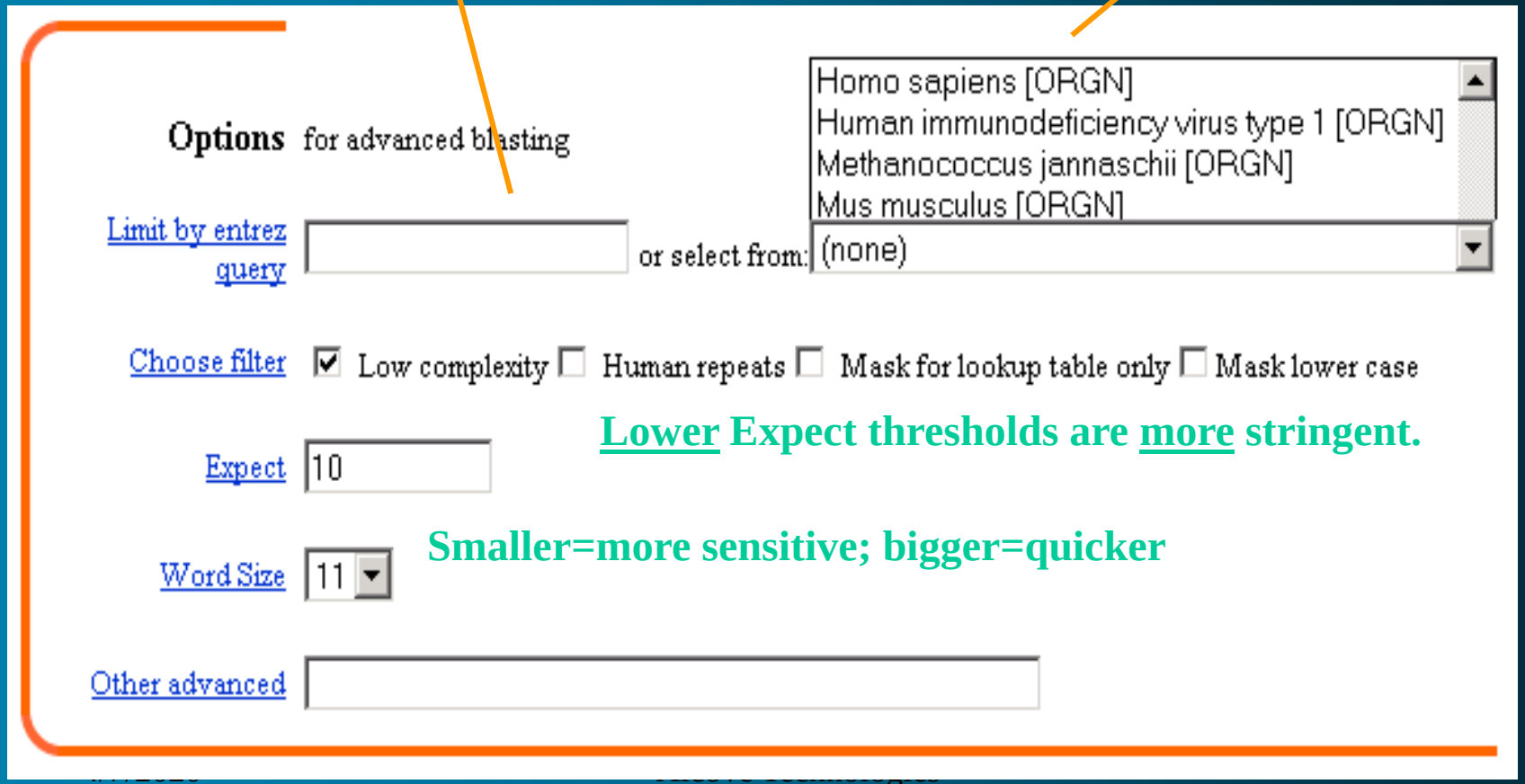
nr = **n**on-**r**edundant database

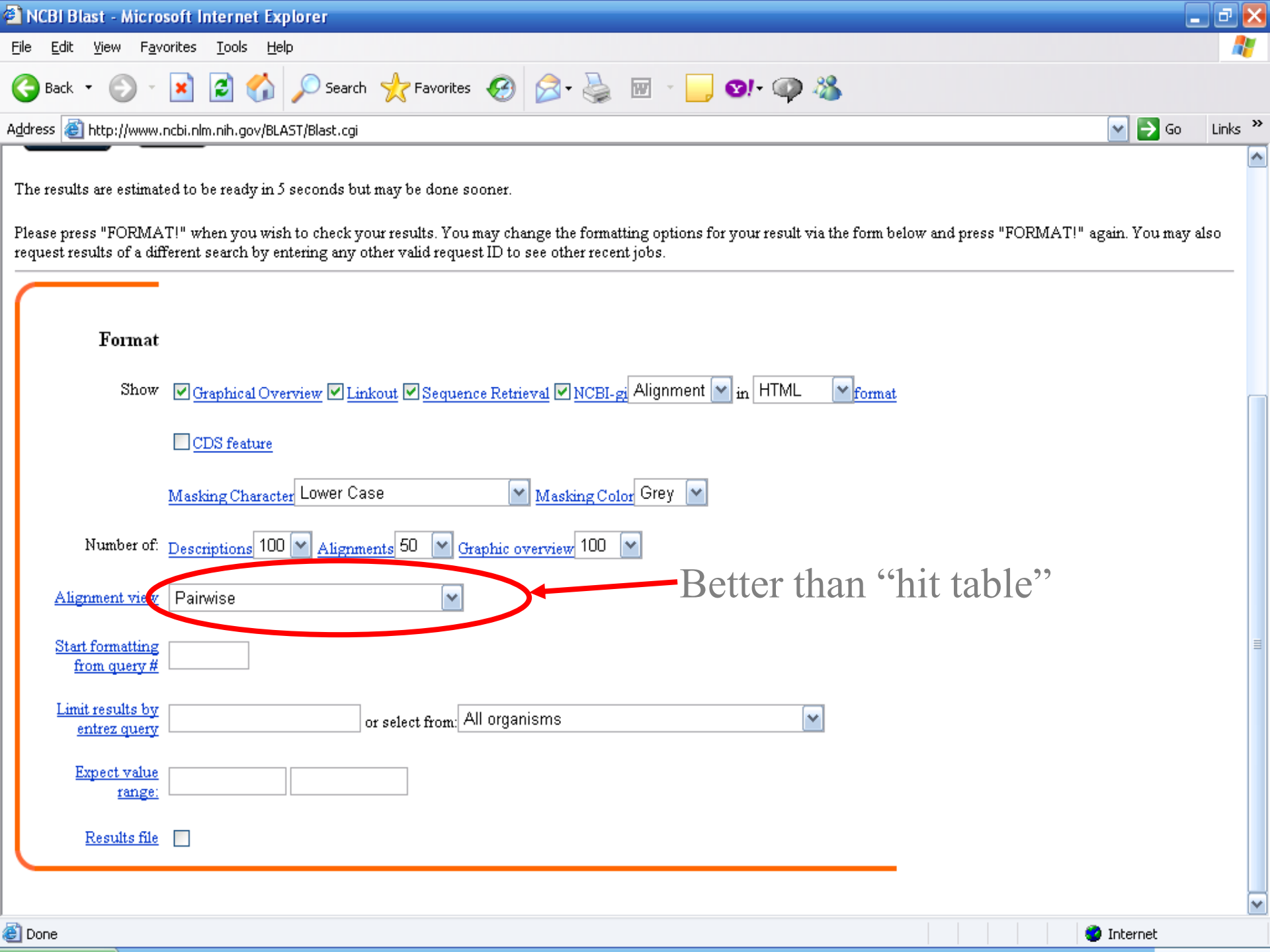
Others are subsets of nr

Blastn advanced options

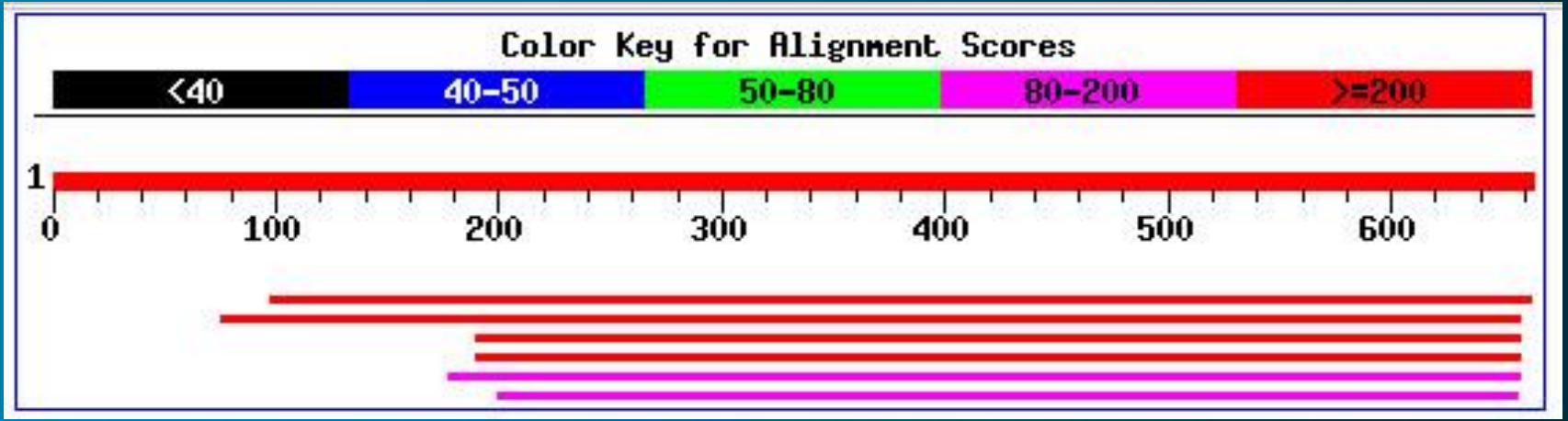
Example: protease NOT hiv1[Organism]

Restrict analysis to sequences only from a certain organism




NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Go Links
The results are estimated to be ready in 5 seconds but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
Format
Show Graphical Overview Linkout Sequence Retrieval NCBI-gi Alignment in HTML format
CDS feature
Masking Character Lower Case Masking Color Grey
Number of: Descriptions 100 Alignments 50 Graphic overview 100
Alignment view Pairwise
Better than "hit table"
Start formatting from query #
Limit results by entrez query
or select from: All organisms
Expect value range:
Results file
Done
Internet

BLAST output



Sequences producing significant alignments:		Score (bits)	E Value
gi|13879340|gb|AAH06645.1|AAH06645	(BC006645) Similar to al...	918	0.0
gi|6678788|ref|NP_032574.1|	(NM_008548) mannosidase 1, alph...	330	2e-90
gi|6754620|ref|NP_034893.1|	(NM_010763) mannosidase 1, beta...	322	5e-88
gi|1083217|pir||A54407	alpha-mannosidase (EC 3.2.1.24) - mo...	320	2e-87
gi|14164377|dbj|BAB55676.1|	(AB042828) Type II membrane pro...	177	2e-44
gi|14198417|gb|AAH08268.1|AAH08268	(BC008268) Similar to hy...	175	1e-43

S’ E

BLAST Scoring System

Raw score (S): Sum of scores for each aligned position and scores for gaps

$S = \Sigma(\text{matches}) - \Sigma(\text{mismatches}) - \Sigma(\text{gap penalties})$

note: this score varies with the scoring matrix used and thus may not be meaningfully compared for different searches

Bit score (S'): Version of the raw score that is normalized by the scale of the scoring matrix (λ) and the scale of the search space size (K)

$S' = (\lambda S - \ln(K)) / \ln(2)$

note: because it is normalized the bit score can be meaningfully compared across searches

E value: Number of alignments with score S' or better that one would expect to find by chance in a search of a database of the same size

$E = mn2^{-S'}$

m = effective length of database

n = effective length of query sequence

note: E values may change if databases of different sizes are searched

BLAST output (cont.)

```
>gi|14164377|dbj|BAB55676.1| (AB042828) Type II membrane protein of ER~mouse gene similar to
          alpha-mannosidase [Mus musculus]
          Length = 652

 Score =  177 bits (449), Expect = 2e-44
 Identities = 173/546 (31%), Positives = 250/546 (45%), Gaps = 96/546 (17%)

Query: 179 PEQGTELPPRRAEVPTKPPLPPARTQGTPVHLNY--------------RQKGVI------ 218
           P +GTE    R E P +P   P      P H  Y              R  G
Sbjct: 67  PRRGTE---GRLETPPEPGPTPGPGVCGPAHWGYALGGGGCGPDEYERRYSGAFPPQLRA 123
```

S’ S E

```
  Database: All non-redundant GenBank CDS
  translations+PDB+SwissProt+PIR+PRF
    Posted date:  Jan 31, 2002 12:59 AM
  Number of letters in database: 270,667,070
  Number of sequences in database:  0

Lambda     K      H
   0.319    0.137    0.420

Gapped
Lambda     K      H
   0.267   0.0410    0.140
```

λ K

```
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 26,665,471
Number of Sequences: 0
Number of extensions: 1297690
Number of successful extensions: 5012
Number of sequences better than 1.0e-05: 6
Number of HSP's better than  0.0 without gapping: 6
Number of HSP's successfully gapped in prelim test: 0
Number of HSP's that attempted gapping in prelim test: 4968
Number of HSP's gapped (non-prelim): 10
length of query: 663
length of database: 18,444,546
effective HSP length: 108
effective length of query: 555
effective length of database: 12,037,230
effective search space: 6680662650
effective search space used: 6680662650
```

n

m

Types of BLAST

BLASTn

```
ACTACGAT
|  |||  ||
A-TACCAT
```

BLASTp

```
GWREIVN
||||   |
GWREVAN
```

Types of BLAST

- Nucleotide to nucleotide
  - Mega BLAST – looking for identical match
  - Discontinuous Mega BLAST – look for nearly identical match
  - BLASTn – Similarity unknown
- BLASTx – Only if you think your sequence is coding

```
CCTCATAT      CCTCATAT      CCTCATAT
 ↓  ↓          ↓  ↓          ↓  ↓
 P  H          L  I          S  Y
```

Frame 1 Frame 2 Frame 3

Plus the reverse strand too…

Types of BLAST

- BLASTp – Protein to protein

Position-Specific Iterated BLAST (PSI-BLAST):PSI-BLAST searches with iterations against protein database until no new significant alignments are found.

Pattern-Hits Integrated BLAST(PHI-BLAST):It searches against protein database based on protein conserved patterns .

- BLASTx – Translate all six possible frames and then compare to protein database
- tBLASTn – Compare protein versus a six-frame translated nucleotide database
- tBLASTx - Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

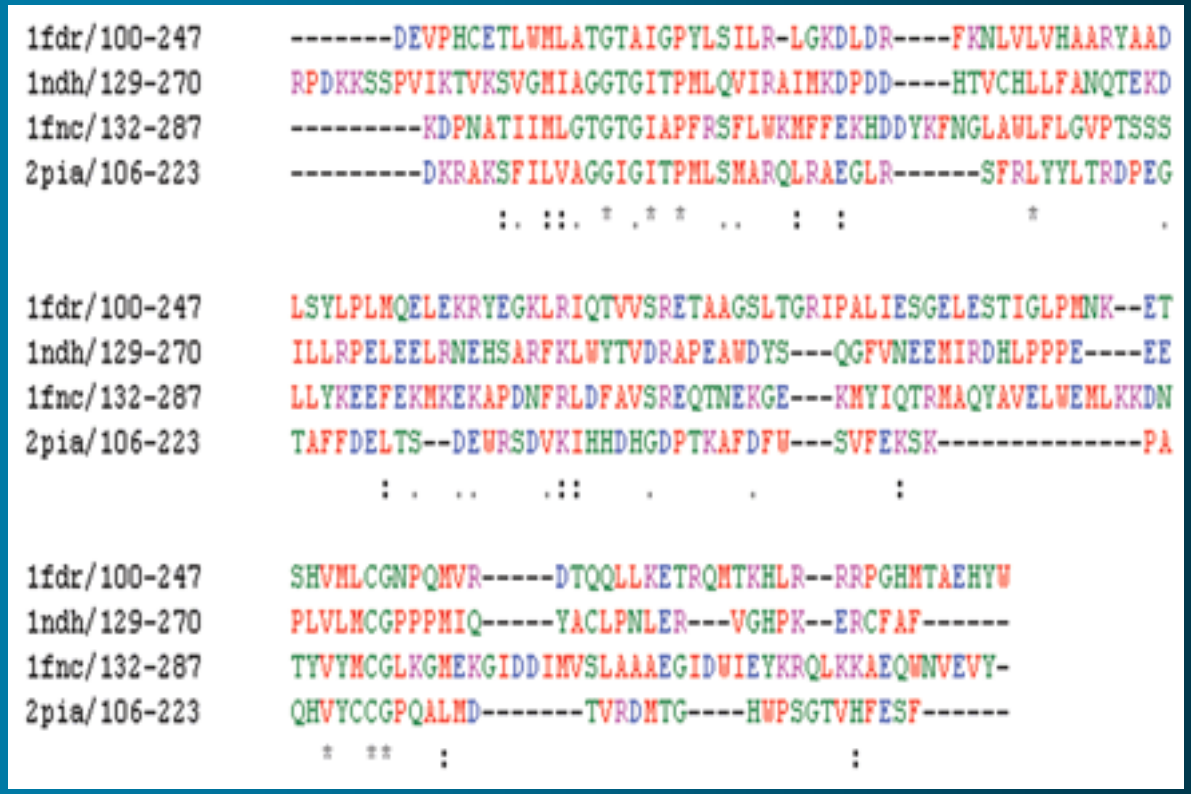
Multiple Sequence Alignment

Multiple Sequence Alignment

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

- Goal: Bring the greatest number of similar characters into the same column of the alignment
- Similar to alignment of two sequences.

CLUSTALW MSA



MSA of four oxidoreductase NAD binding domain protein sequences. **Red**: AVFPMILW. **Blue**: DE. **Magenta**: RHK. **Green**: STYHCNGQ. **Grey**: all others. Residue ranges are shown after sequence names.

Chenna et al. *Nucleic Acids Research*, 2003, Vol. 31, No. 13 **3497-3500**

Multiple Sequence Alignment: Motivation

- Correspondence. Find out which parts "do the same thing"
  - Similar genes are conserved across widely divergent species, often performing similar functions
- Structure prediction
  - Use knowledge of structure of one or more members of a protein MSA to predict structure of other members
  - Structure is more conserved than sequence
- Create "profiles" for protein families
  - Allow us to search for other members of the family
- Genome assembly: Automated reconstruction of "contig" maps of genomic fragments such as ESTs
- MSA is the starting point for phylogenetic analysis

Multiple Sequence Alignment: Approaches

- **Optimal Global Alignments** -Dynamic programming
  - Generalization of Needleman-Wunsch
  - Find alignment that maximizes a score function
  - Computationally expensive: Time grows as product of sequence lengths
- **Global Progressive Alignments** - Match closely-related sequences first using a guide tree
- **Global Iterative Alignments** - Multiple re-building attempts to find best alignment
- **Local alignments**
  - Profiles, Blocks, Patterns

Clustal W

- W stands for Weighted
- Different weights are given to sequences and parameters in different parts of the alignment.
- Position Specific Gap Penalties
- The goal is to insert gaps only in “loop” regions
- Higher penalties in the middle of helices and strands

Large penalty for closely related sequences

Small penalty for divergent sequences

Practical Considerations

- When to use Clustal

Can be used to align any group of protein or nucleic acid sequences that are related to each other over their entire lengths.

- Clustal is optimized to align sets of sequences that are entirely colinear, i.e. sequences that have the same protein domains, in the same order.

Thank you