Sequence Database Searching

2007-01-14

Gao, Ge

Center for Bioinformatics, Peking University

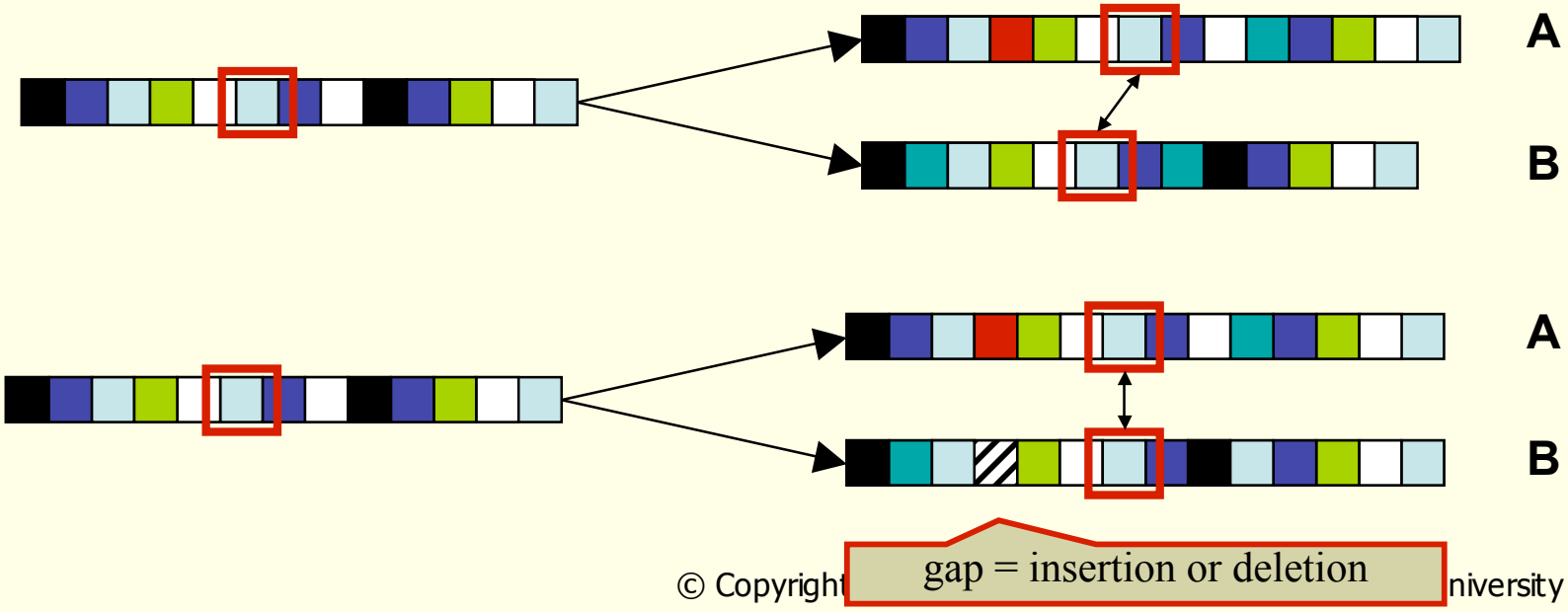
Sequence Alignment: A Quick Review

- Object: see how close two sequences are

- Usages:
  - To infer functions.
  - To determine the common area of some sequences
  - To infer protein structure, if the structure of one of the sequences is known.
  - To “guess” whether they are descended from a common ancestor (and construct evolutionary trees).

Sequence Alignment: in Biology

- The purpose of a sequence alignment is to line up all residues in the sequence that were derived from the same residue position in the ancestral gene or protein in any number of sequences



Sequence Alignment: Scoring

```
GAATC      GAAT-C     -GAAT-C
CATAC      C-ATAC     C-A-TAC

GAATC-     GAAT-C     GA-ATC
CA-TAC     CA-TAC     CATA-C
```

- Scoring function: measure the quality of a candidate alignment.
  - **scoring matrix**,
  - **gap penalty**

Sequence Alignment : Global vs. Local

- Global alignments: align residues in whole sequence,
  - are most useful when the sequences are similar and of roughly equal size.
  - Algorithm: Needleman-Wunsch
- Local alignments: align residues in regions
  - are more useful for diverged sequences
  - Algorithm: Smith-Wate
- With sufficiently similar s
  difference between loca

```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

Sequence Database Searching

- Rather than do the alignment pair-wise, it's necessary to run database searching in a high-throughput style.

- Identify similarities between
  - novel query sequences
    whose structures and functions are unknown and uncharacterized
  - sequences in (public) databases
    whose structures and functions have been elucidated.

Sequence Database Searching

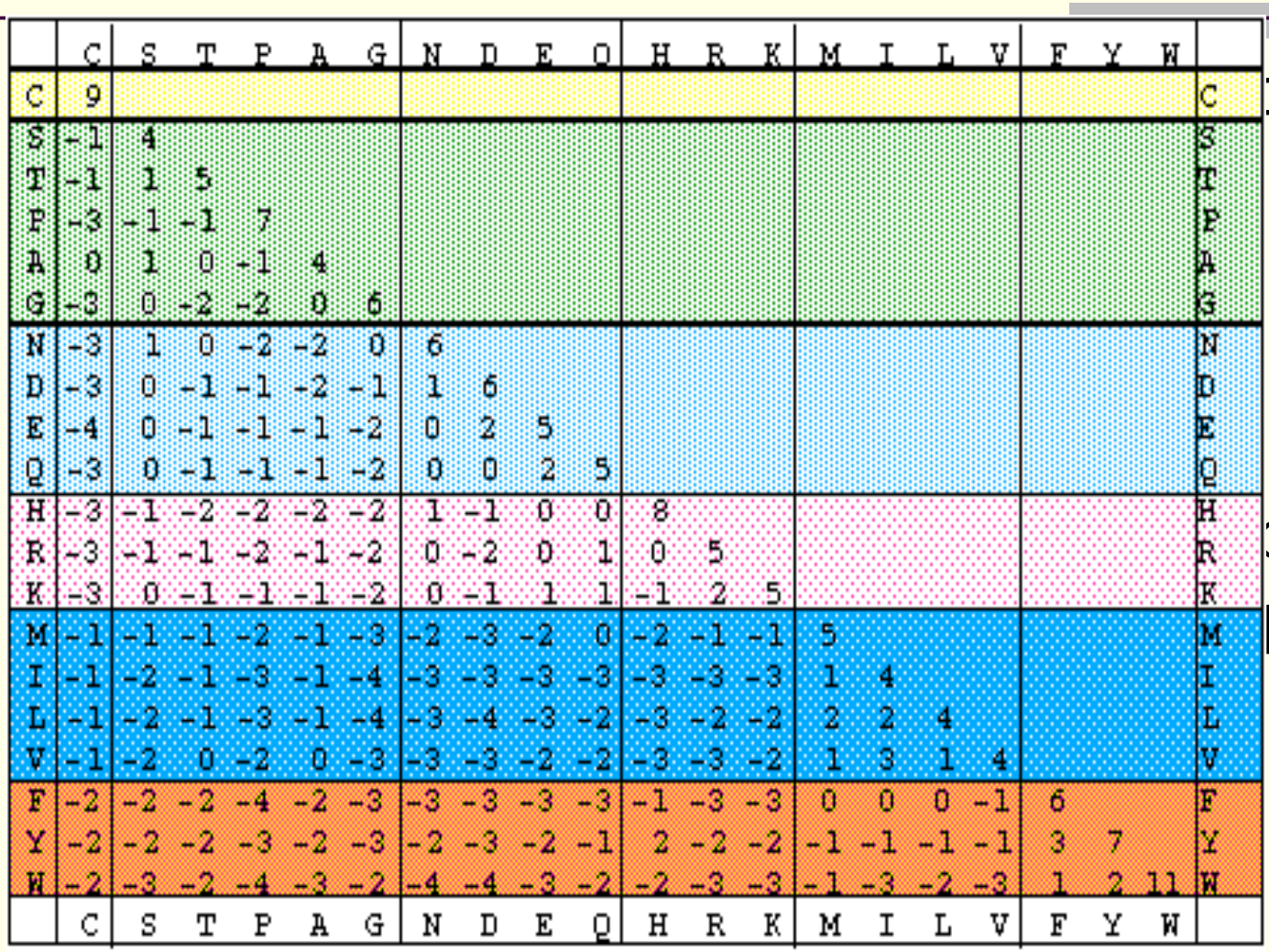
- The query sequence is compared/aligned with every sequence in the database
- High-scoring database sequences are assumed to be evolutionary related to the query sequence
  - Similar function
  - Similar structure
  - Closer evolutionary relationship

BLAST: Intro

- To make the alignment effectively, an algorithm BLAST (Basic Local Alignment Search Tool) is proposed by Altschul *et al* in 1990.
  - BLAST finds the highest scoring locally optimal alignments between a query sequence and a database.
  - Very fast algorithm
  - Can be used to search extremely large databases
  - Sufficiently sensitive and selective for most p
  - Robust – the default parameters can usually

BLOSUM62: revolution of PAM250

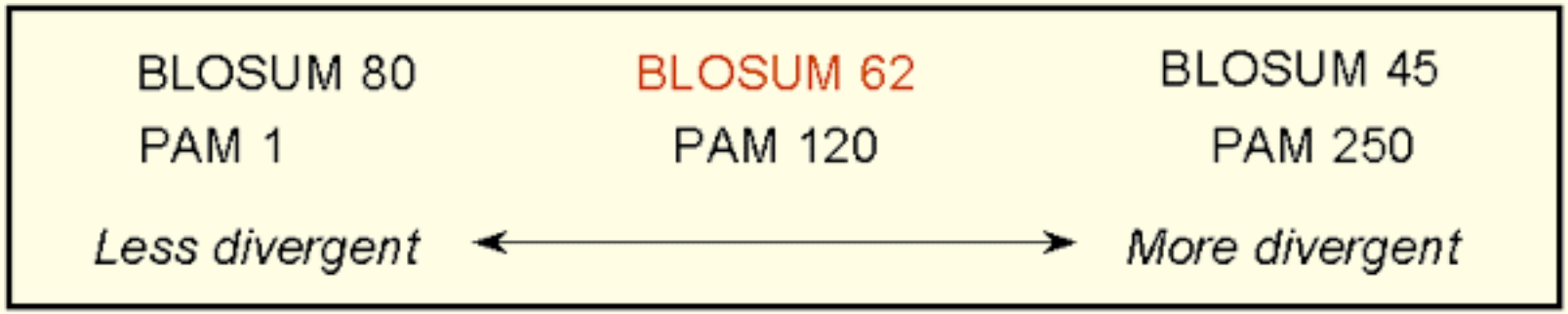
- om
- ped
- ple



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Which Matrix to use?

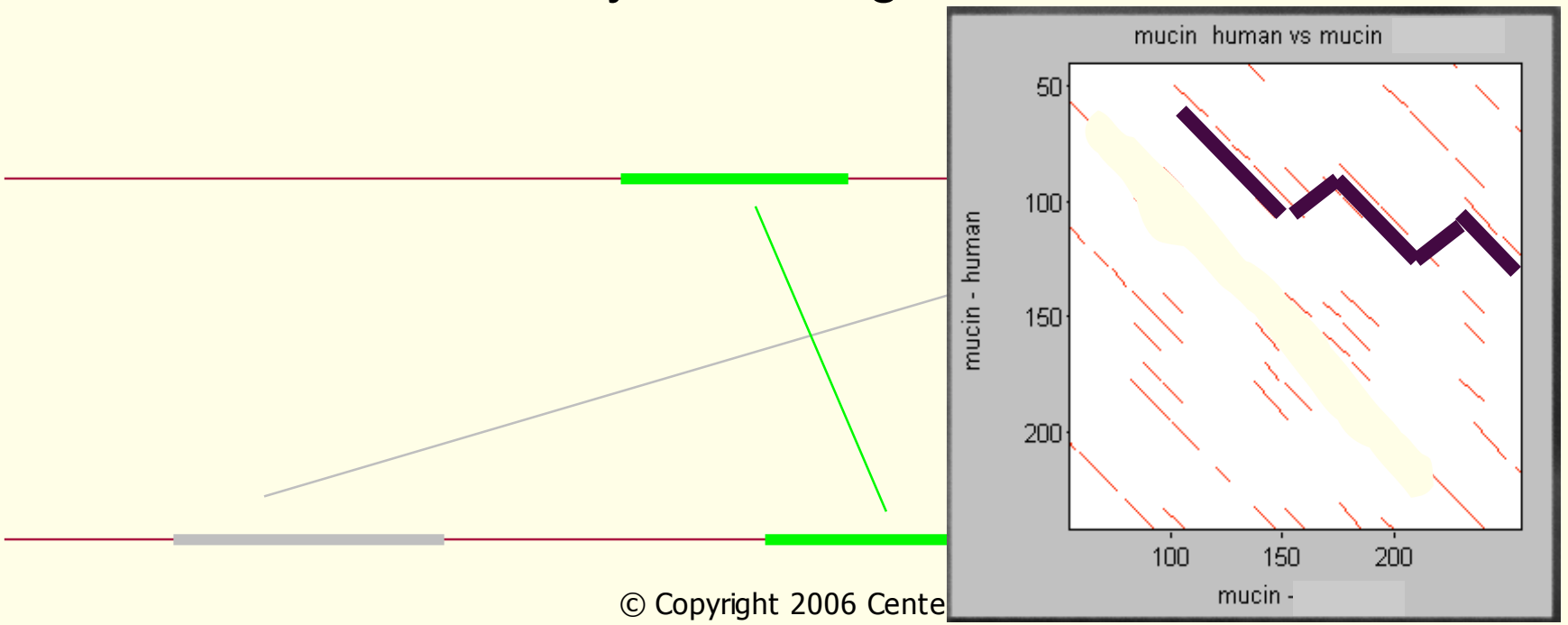
Close relationships (Low PAM, high BLOSUM)
Distant relationships (High PAM, low BLOSUM)



Reasonable defaults: PAM250, BLOSUM62

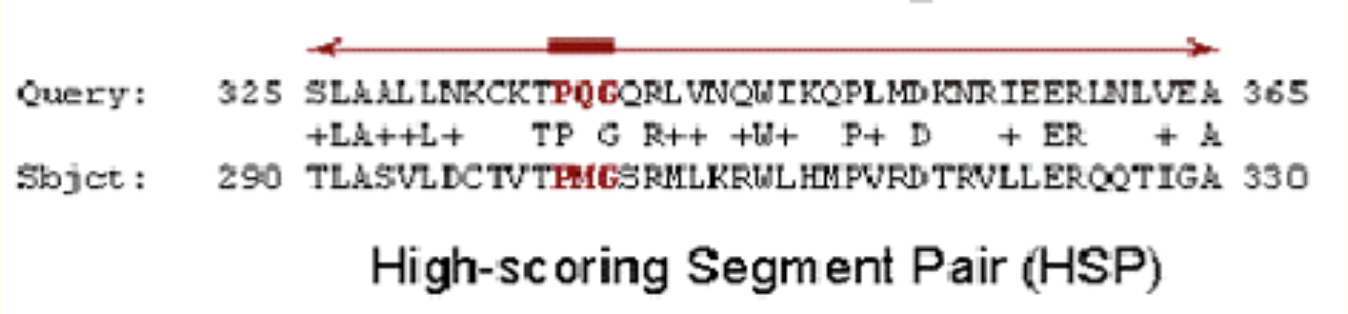
BLAST Ideas: Seeding-and-extending

- Find matches (seed) between the query and subject
- Extend seed into High Scoring Segment Pairs (HSPs)
  - Stop extension when total score doesn't increase
- Assess the reliability of the alignment.



BLAST HSP

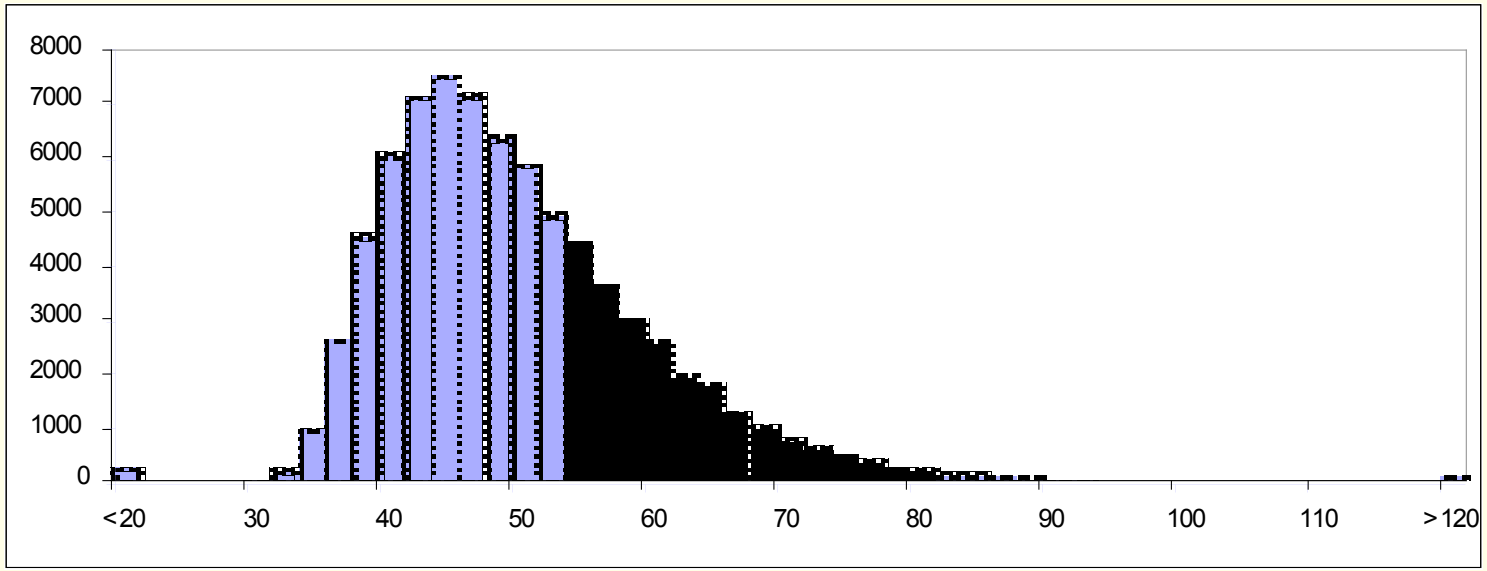
- The program tries to extend matching segments (seeds) out in both directions by adding pairs of residues.



High-scoring Segment Pair (HSP)

BLAST Algorithm: E-Value

- Given the large data volume, it's critical to provide some measures for accessing the significance of a given hit.

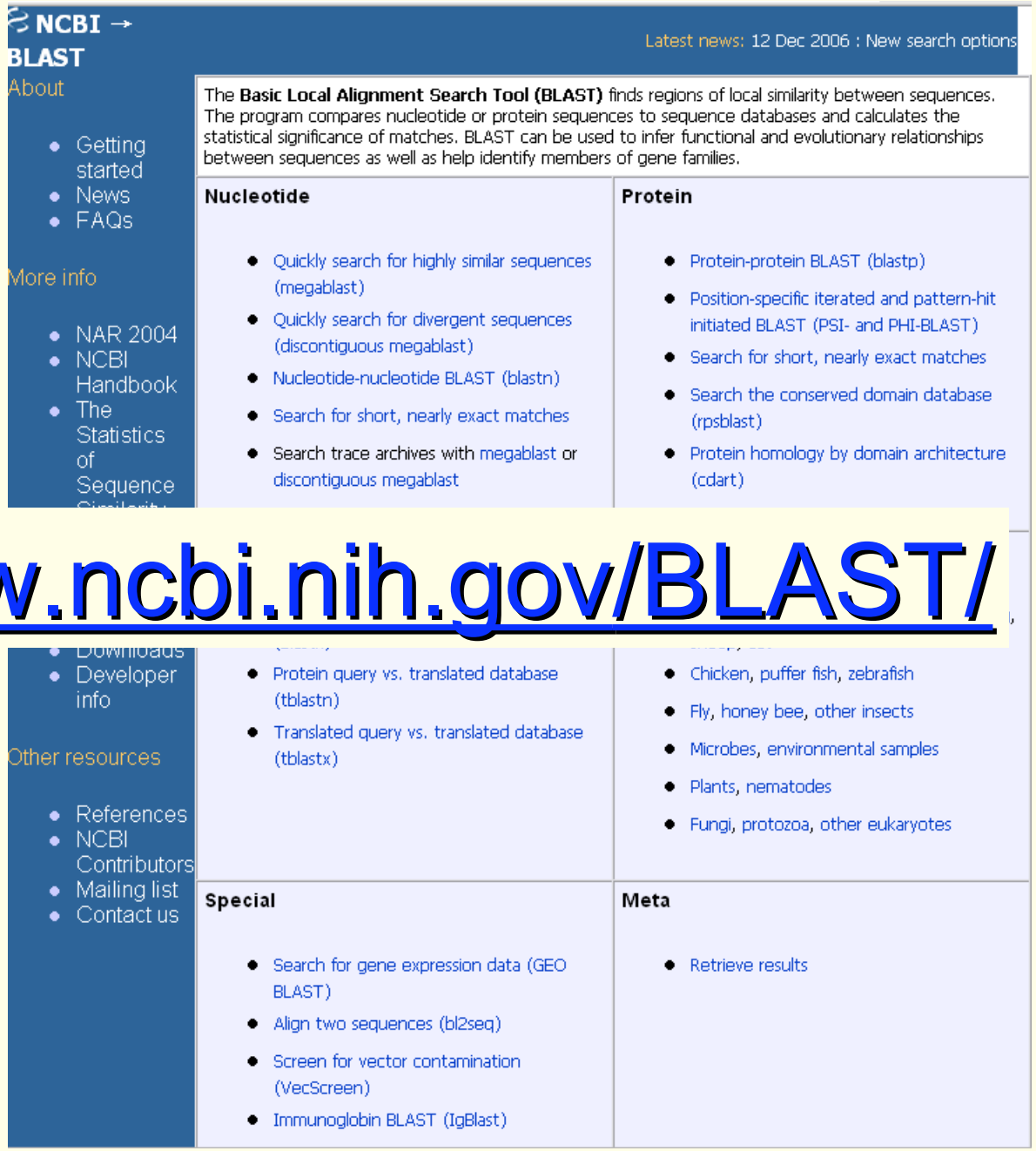


BLAST Algorithm: E-Value

- E-value: expect value
  - the number of alignments with a given score that would be expected to occur at random in the database that has been searched
  - e.g. if E=10, 10 matches with scores this high are expected to be found by chance

$$E = kmne^{-\lambda s}$$

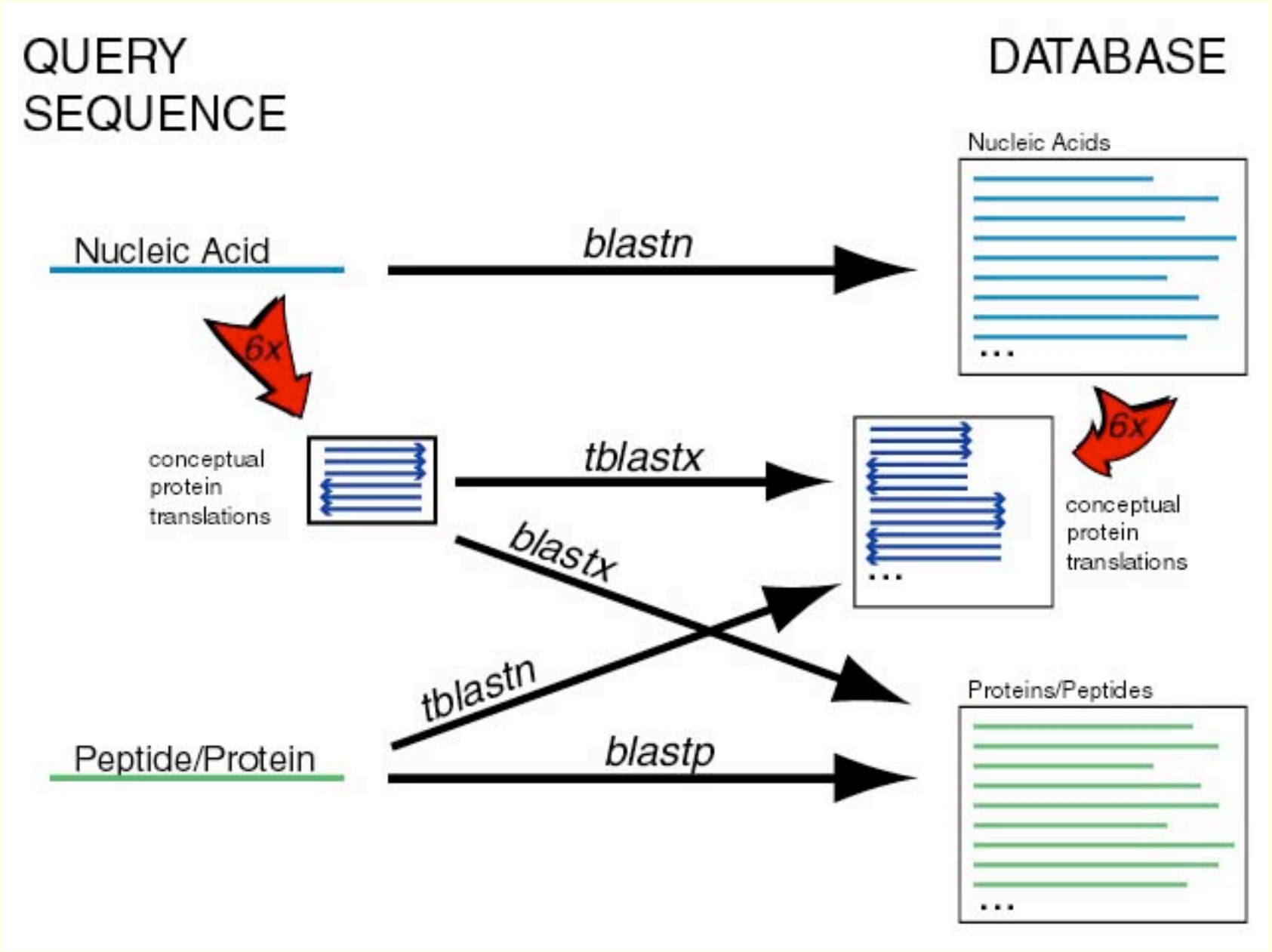
NCBI BLAST



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

BLAST family

Program	Query	Database	Typical Uses
BLASTP	Protein	Protein	Identifying common regions between proteins. Collecting related proteins for phylogenetic analysis.
BLASTN	DNA	DNA	Mapping oligonucleotides, amplimers, ESTs, and repeats to a genome. Identifying related transcripts.
BLASTX	Translated DNA	protein	Finding protein-coding genes in genomic DNA.
TBLASTN	protein	Translated DNA	Identifying transcripts similar to a known protein (finding proteins not yet in GenBank). Mapping a protein to genomic DNA.
TBLASTX	Translated DNA	Translated DNA	Cross-species gene prediction. Searching for genes missed by traditional methods.


QUERY SEQUENCE
DATABASE
Nucleic Acids
Nucleic Acid
blastn
6x
conceptual protein translations
tblastx
6x
conceptual protein translations
blastx
tblastn
Proteins/Peptides
Peptide/Protein
blastp

(Joel H.

BLAST Input

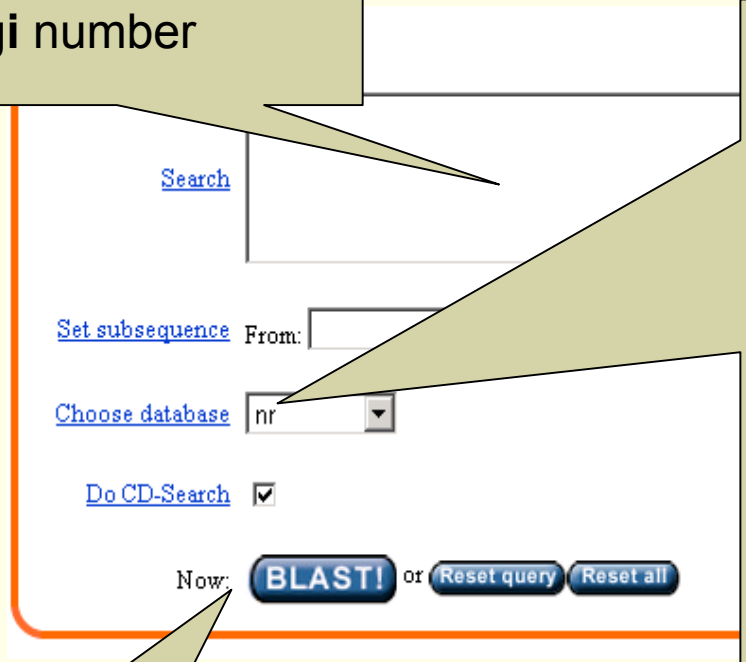
Steps in running BLAST:

- Entering your query sequence (cut-and-paste)
- Select the database(s) you want to search
- Choose alignment parameters (e.g. scoring matrix, filters,….)
- Choose output parameters
- Example query=

MAFIWLLSCYALLGTTFGCGVNAIHPVLTGLSKIVNGEEAVPGTWPWQVTLQDRSGFHF
CGGSLISEDWVVTAAHCGVRTSEILIAGEFDQGSDEDNIQVLRIAKVFKQPKYSILTVNND
ITLLKLASPARYSQTISAVCLPSVDDDAGSLCATTGWGRTKYNANKSPDKLERAALPLLT
NAECKRSWGRRLTDVMICGAASGVSSCMGDSGGPLVCQKDGAYTLVAIVSWASDTCSASS
GGVYAKVTKIIPWVQKILSSN

Query Sequence and Database

Input the query sequence as **FASTA** format, **bare** sequence or NCBI **gi** number



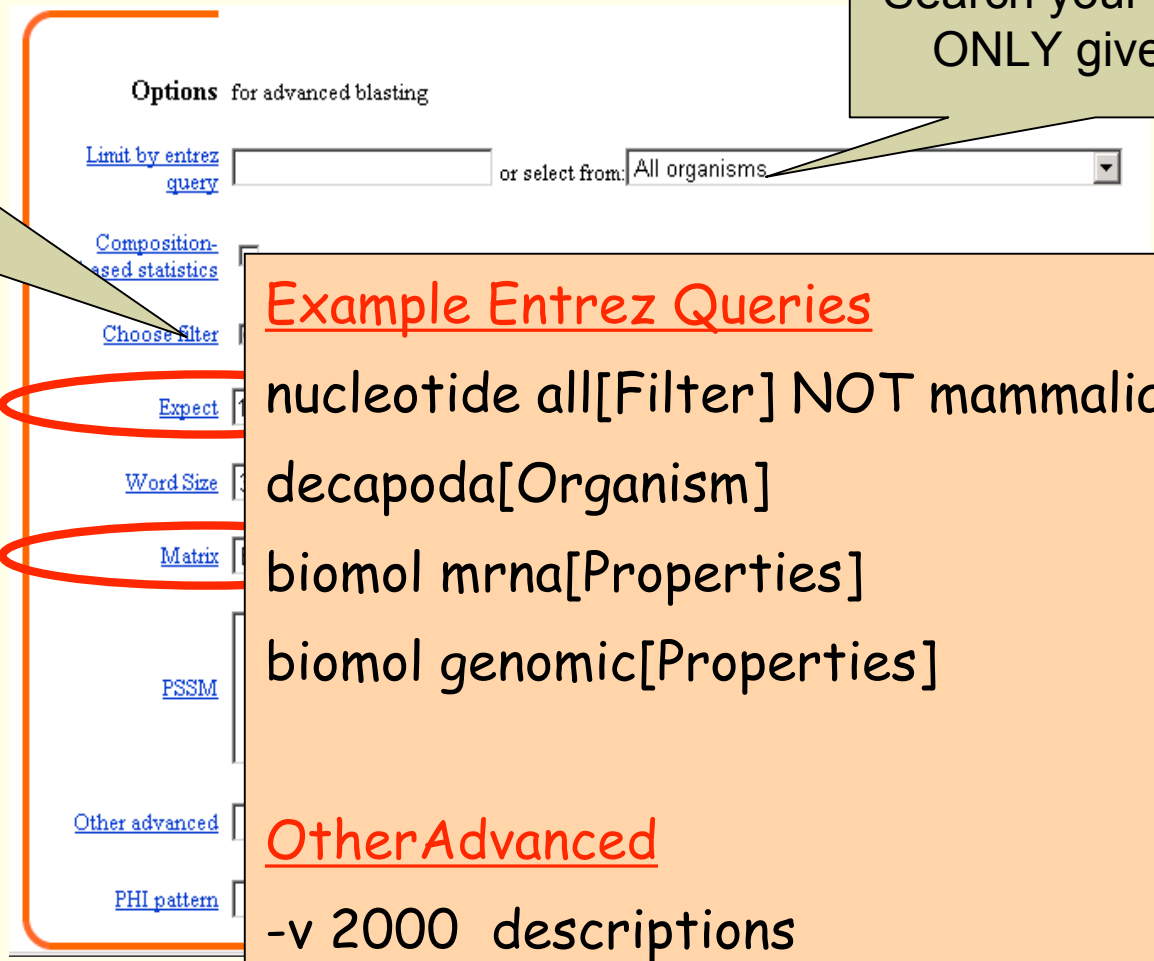
- **nr:** All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF
- **swissprot**: Last major release of the SWISS-PROT protein sequence database (no updates)
- **month**: All new or revised GenBank CDS translation + PDB + SwissProt + PIR + PRF released in the last 30 days.

Click here to **Run** the BLAST

Choose Alignment Arguments

The **filter** provides a way to prevent *false positive hit.*

Search your query again ONLY given specie.



Example Entrez Queries

nucleotide all[Filter] NOT mammalia[Organism]

decapoda[Organism]

biomol mrna[Properties]

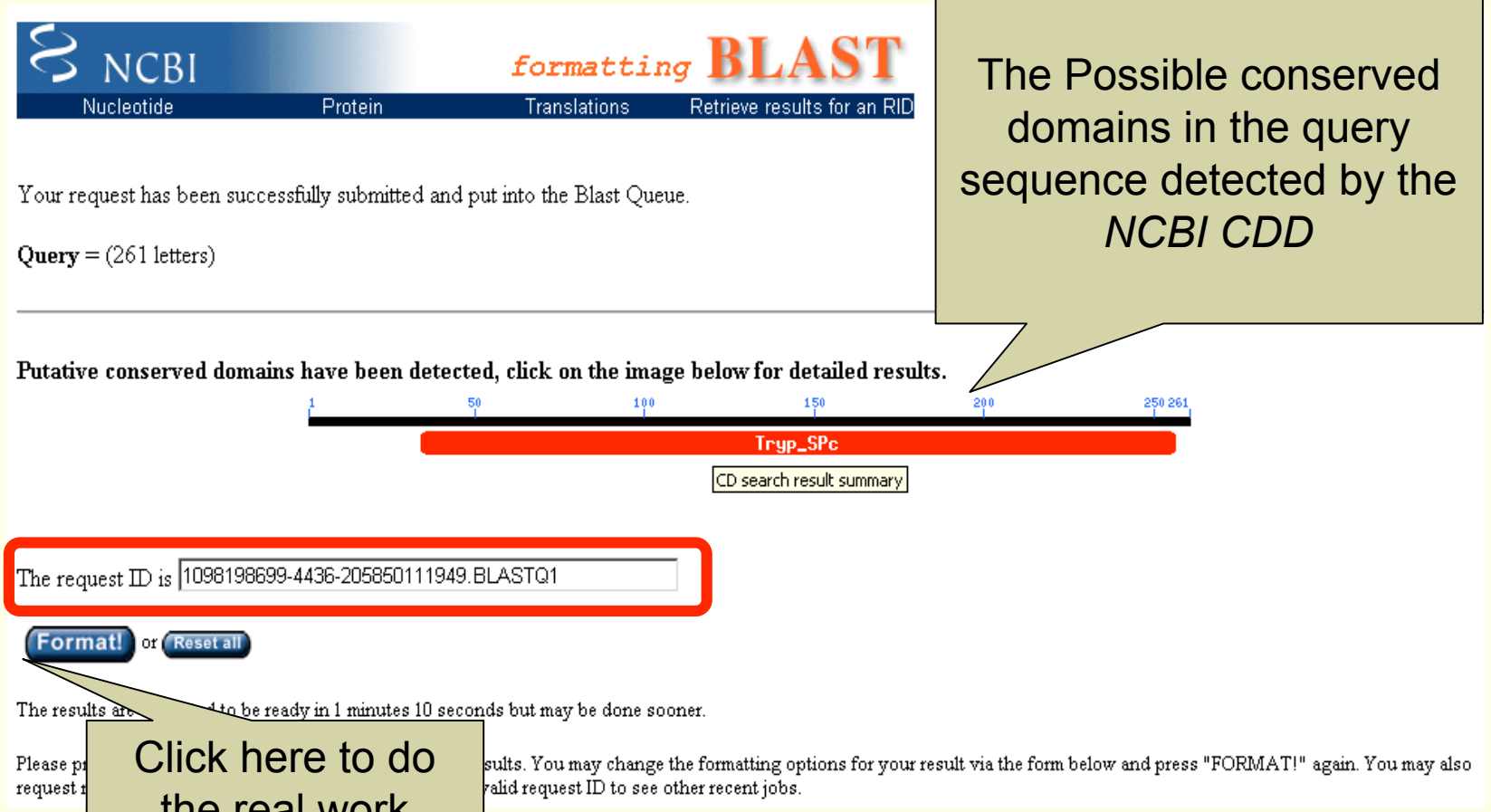
biomol genomic[Properties]

OtherAdvanced

-v 2000 descriptions

-b 2000 alignments

Format the Result

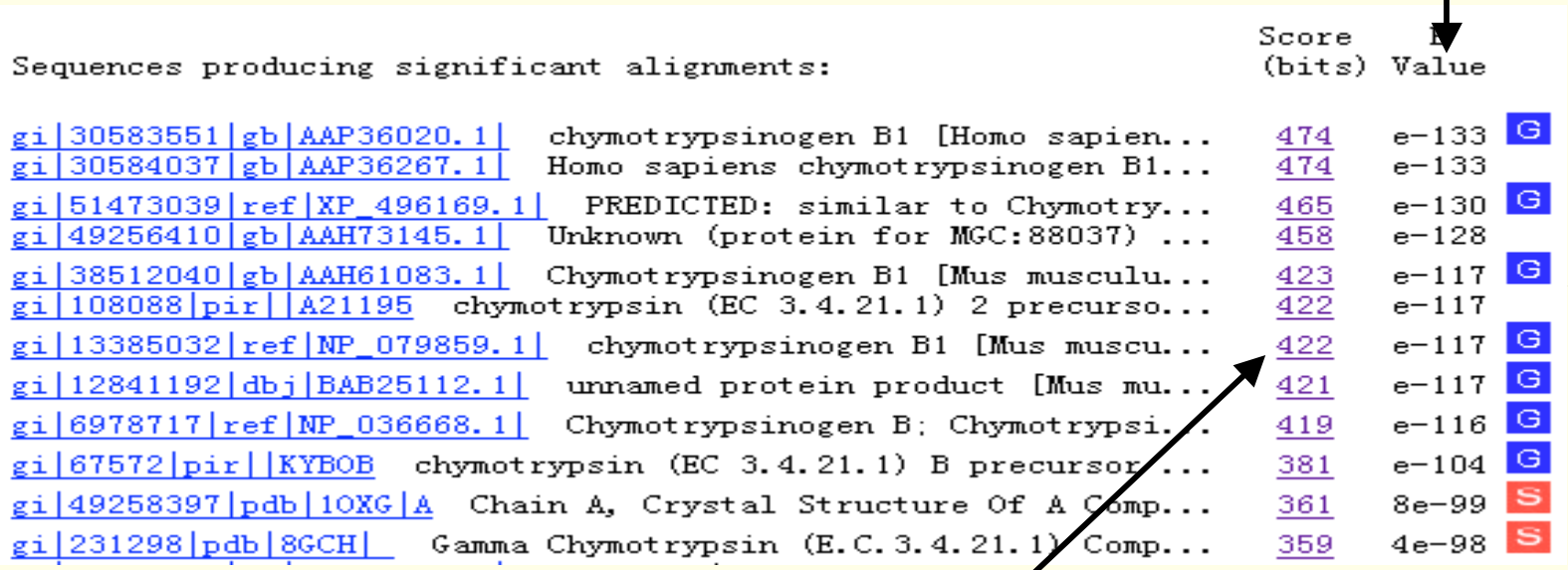


The Possible conserved domains in the query sequence detected by the *NCBI CDD*

Click here to do the real work

BLAST Output

A low "***Expectation value"*** indicates that a match is unlikely to arise by chance



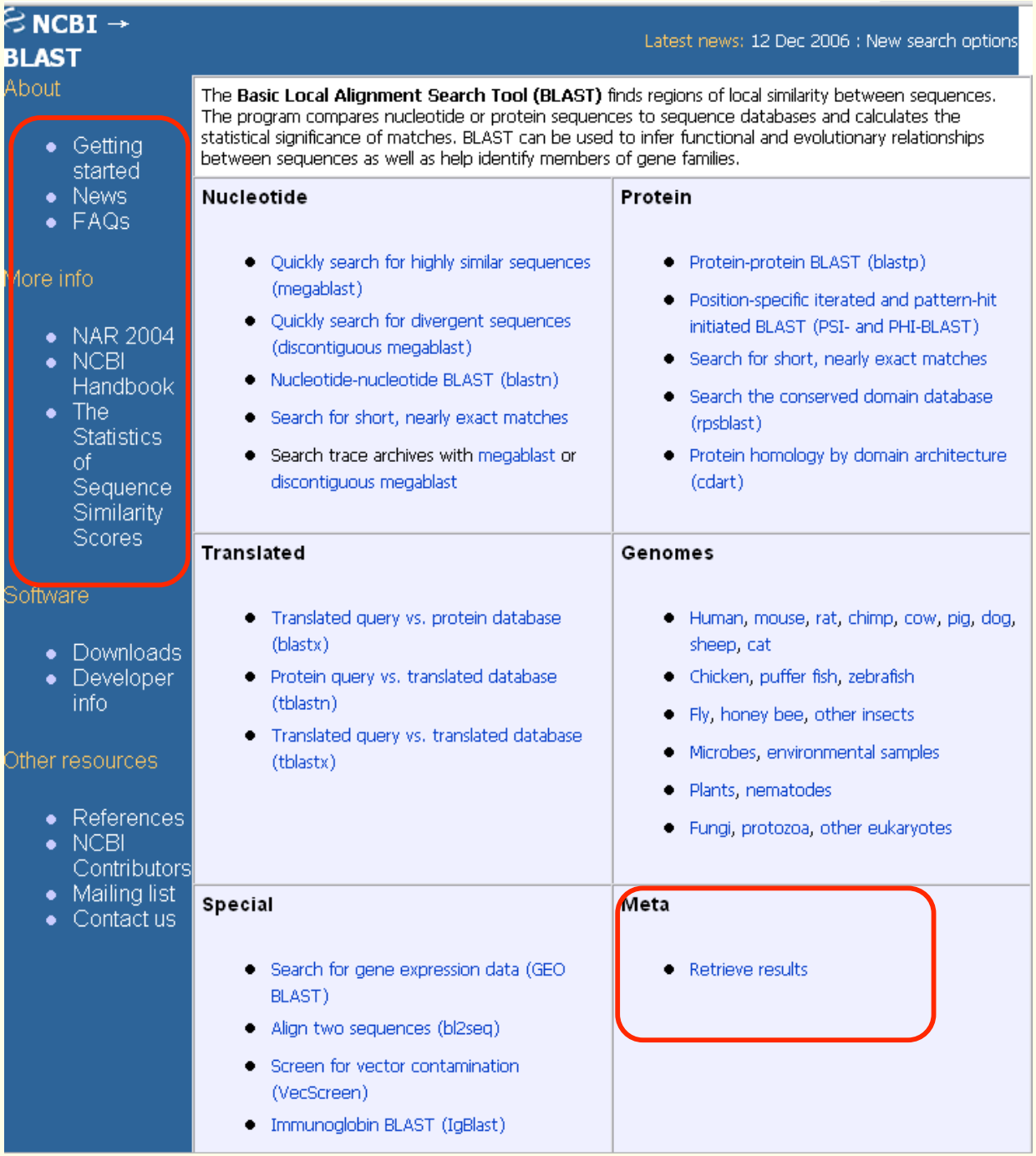
Sequences producing significant alignments:	Score (bits)	E Value	
gi|30583551|gb|AAP36020.1| chymotrypsinogen B1 [Homo sapien...	474	e-133	G
gi|30584037|gb|AAP36267.1| Homo sapiens chymotrypsinogen B1...	474	e-133	
gi|51473039|ref|XP_496169.1| PREDICTED: similar to Chymotry...	465	e-130	G
gi|49256410|gb|AAH73145.1| Unknown (protein for MGC:88037) ...	458	e-128	
gi|38512040|gb|AAH61083.1| Chymotrypsinogen B1 [Mus musculu...	423	e-117	G
gi|108088|pir||A21195 chymotrypsin (EC 3.4.21.1) 2 precurso...	422	e-117	
gi|13385032|ref|NP_079859.1| chymotrypsinogen B1 [Mus muscu...	422	e-117	G
gi|12841192|dbj|BAB25112.1| unnamed protein product [Mus mu...	421	e-117	G
gi|6978717|ref|NP_036668.1| Chymotrypsinogen B; Chymotrypsi...	419	e-116	G
gi|67572|pir||KYBOB chymotrypsin (EC 3.4.21.1) B precursor...	381	e-104	G
gi|49258397|pdb|1OXG|A Chain A, Crystal Structure Of A Comp...	361	8e-99	S
gi|231298|pdb|8GCH| Gamma Chymotrypsin (E.C.3.4.21.1) Comp...	359	4e-98	S

A high score, or preferably, clusters of high scores, indicates a likely relationship

Direct links to related information.

Advanced BLAST

- Megablast
  - nucleotide only
  - optimized for large batch searches
- PSI-BLAST
  - constructs PSSMs automatically
  - searches protein database with PSSMs
- RPS BLAST
  - searches a database of PSSMs
  - basis of conserved domain database


NCBI →
BLAST
Latest news: 12 Dec 2006 : New search options
About
Getting started
News
FAQs
More info
NAR 2004
NCBI Handbook
The Statistics of Sequence Similarity Scores
Software
Downloads
Developer info
Other resources
References
NCBI Contributors
Mailing list
Contact us
The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
Nucleotide
Quickly search for highly similar sequences (megablast)
Quickly search for divergent sequences (discontiguous megablast)
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or discontiguous megablast
Protein
Protein-protein BLAST (blastp)
Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Protein homology by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
Chicken, puffer fish, zebrafish
Fly, honey bee, other insects
Microbes, environmental samples
Plants, nematodes
Fungi, protozoa, other eukaryotes
Special
Search for gene expression data (GEO BLAST)
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
Meta
Retrieve results

Lies with BLAST

- With low gap penalties, you can make alignments between just about anything.
  - For *BLASTN*, NCBI-BLAST always uses ungapped statistics, so you don't have to do much work to lie. Just hope that nobody notices all the gaps.
- Another way to trick the unobservant is to remove complexity filters.
  - This works especially well when claiming that some anonymous low-complexity region or transcript is a cool gene.
- You can almost always find a small ORF that has a poor match to something with an interesting definition line.

Identification and role of adenylyl cyc signalling in

Takanari Ichikawa, Yos
Carla Schommer, Ange
& Richard Walden

Max Planck Institut für Züchtu
D-50829 Köln, Germany

Cyclic AMP is an impor
and eukaryotes[1], but its
generally doubted[2] becau
and barely detectable am
T-DNA tagging to create
the absence of the phyto
The sequence tagged in
complementary DNA en
higher plant. Sequence a
cyclase is probably solu
repeats, and bears simila
Schizosaccharomyces po
Escherichia coli results
levels, and in yeast its ex
cry1 mutation. Tobacco
adenylyl cyclase activato
divide. This finding, to
adenylyl cyclase inhibitor dideoxyadenosine inhi
eration in the presence of auxin, suggests that cAM
auxin-triggered cell division in higher plants.

retraction

Identification and role of adenylyl cyclase in auxin signalling in higher plants

Takanari Ichikawa, Yoshihito Suzuki, Inge Czaja, Carla Schommer, Angela Leßnick, Jeff Schell & Richard Walden

Nature **390**, 698–701 (1997)

Some of the results reported in this Letter cannot be reproduced. We know that the data on protoplast division described for Figs 1a, c, 2a, b and 4, and the corresponding experimental procedures described in the Methods section and the text, are wrong. In fact, the data showing that cAMP can stimulate protoplast division in the absence of auxins are r
paper. We apologize for
caused.
Note from the Editor: Or
accuracy of parts of thi
retraction and awaits th

ng organism, the
t were likely to
rched for similar
rched contained
heses about the

finding because
of biochemistry
e similarity, and
e was aligned to
e characteristics

umiliation was a

```
  1  MQRVLKARQL VRVLRKSSSP ILLNSVSRIQ SHCTYEATES CLNSSSRRGY
 51  FTSGTAICGN YMQTKHNIQR NVCQCVKCST MLKASFSTEA GTVESSAATV
101  SVKELYDKML KSVVEQRSAP PNAWLWSLIQ SCANREDVNL LHDILQRLRI
151  FRLSNLRIHE NFNCALCQDI TKACVRVGAI DLGKKVLWKH NVYGLTPNIG
201  SAHHLLLFAK QHNDVKLLVE IMKLVKKNDL NLQPGTAEIV FSICFQTDNW
251  DLMCKYGKRV VKAGVKLRKT SLDTWMEFAS KIGDVDALWK IEKIRSESMK
301  EHTLASGLSC AKAFLIDHKP GDAAAIIQSL NQTIPDSRRQ NFMIELQKLV
351  ADWPLEVIKR QKDEKRKELA ATLQHDIPAM LSALPNRGLN LDINLEDLTR
401  KEGVLS
```

Some ways to be good

- Check data, whenever possible.
  - Carefulness is good.
- Check the output, carefully
  - Extraordinary claims require extraordinary evidence
- Use your mind.
  - Tool is good, understanding tool is better.

Exercises

- Try to search NR database with following protein sequence, you may re-annotation them by analyzing the BLAST result:
  - Q57997
  - NP_149073
  - XP_372459

Thank you for your attentions!