

---

# Sequence Database Searching

2007-01-14

Gao, Ge

Center for Bioinformatics, Peking University

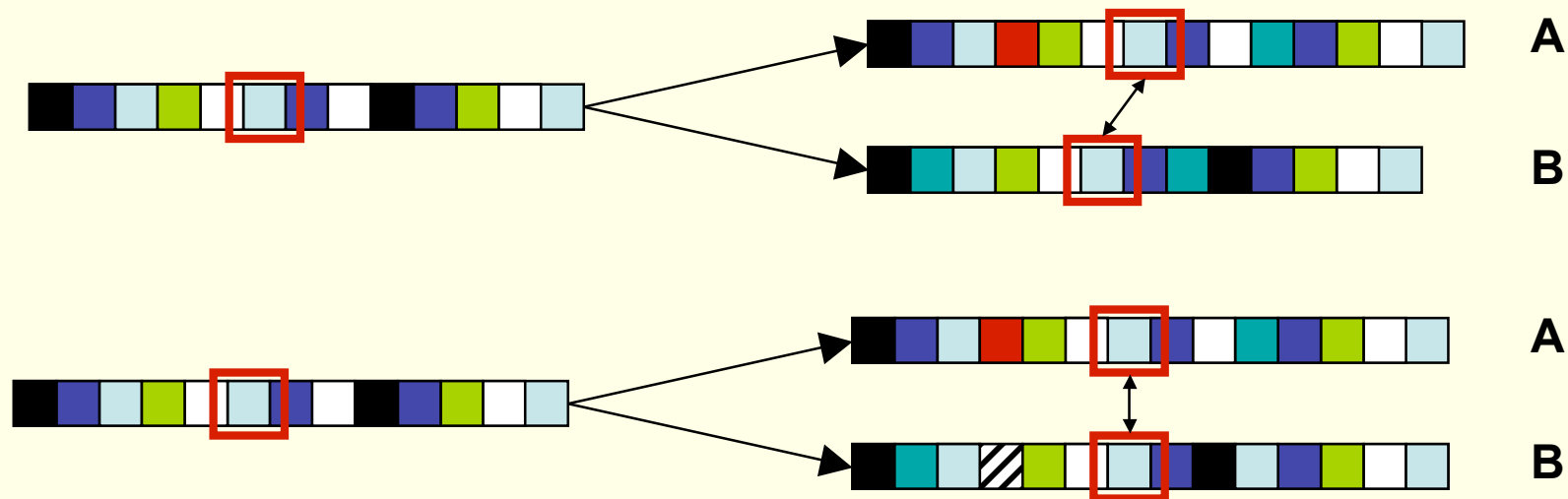
# Sequence Alignment: A Quick Review

---

- 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-Waterman**

- With sufficiently similar sequences, the difference between local and global alignments is small.

```
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

	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	

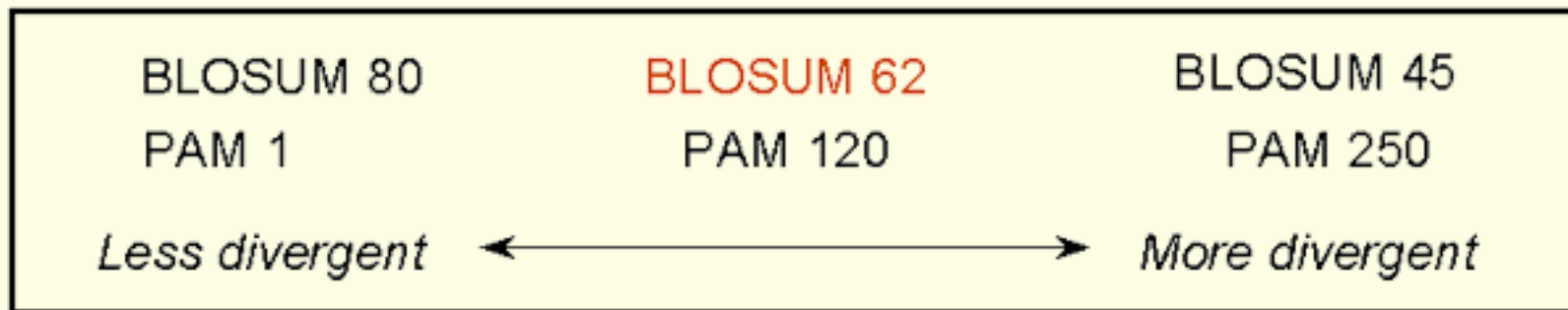
om

bed  
ple

# Which Matrix to use?

Close relationships (Low PAM, high BLOSUM)

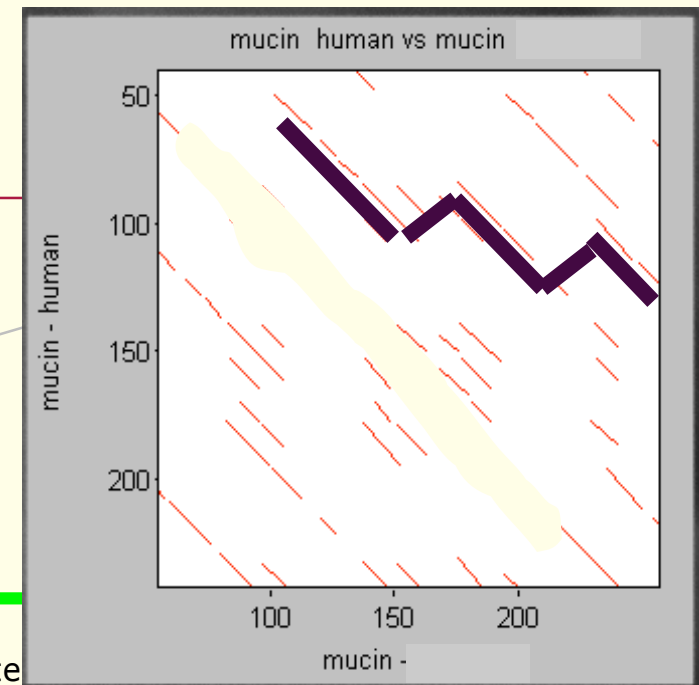
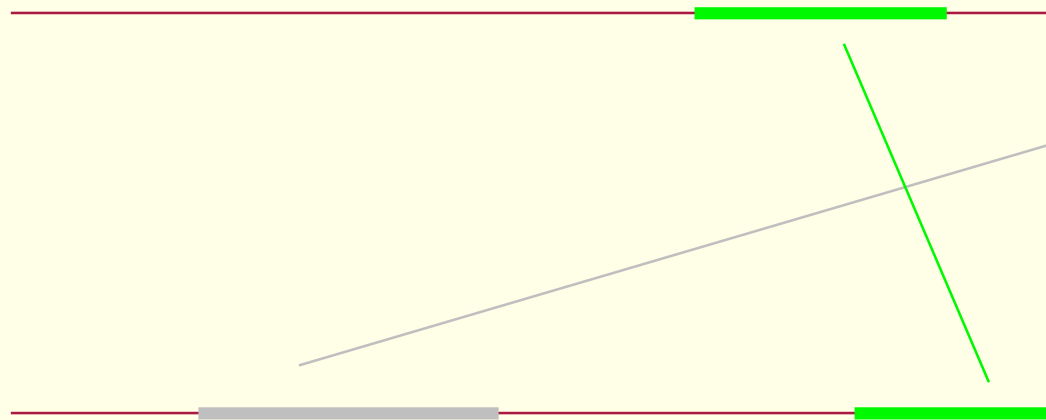
Distant relationships (High PAM, low BLOSUM)



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

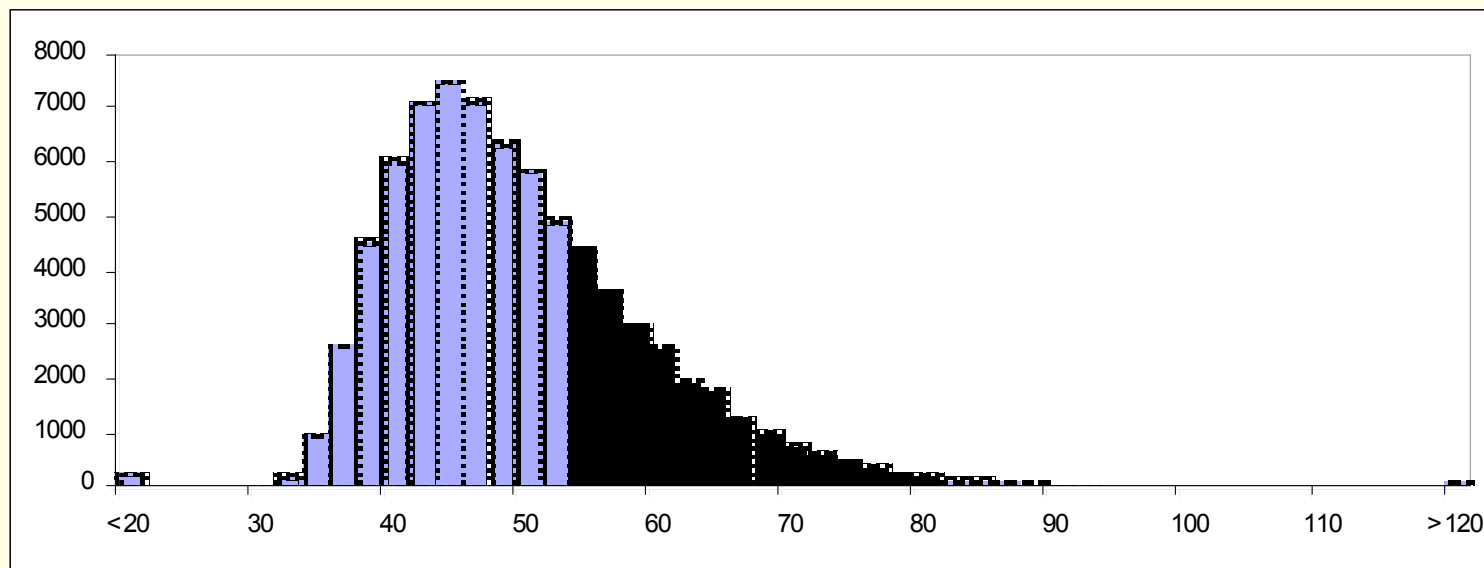
- The program tries to **extend** matching segments (**seeds**) out in both directions by adding pairs of residues.

```
Query: 325 SLAALLNKCKTTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
      +LA++L+  TP G R++ +W+  P+ D  + ER  + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

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

NCBI → BLAST Latest news: 12 Dec 2006 : New search options

About

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.

Getting started

- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity

Nucleotide	Protein
<ul style="list-style-type: none"><li>• Quickly search for highly similar sequences (megablast)</li><li>• Quickly search for divergent sequences (discontiguous megablast)</li><li>• Nucleotide-nucleotide BLAST (blastn)</li><li>• Search for short, nearly exact matches</li><li>• Search trace archives with megablast or discontiguous megablast</li></ul>	<ul style="list-style-type: none"><li>• Protein-protein BLAST (blastp)</li><li>• Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)</li><li>• Search for short, nearly exact matches</li><li>• Search the conserved domain database (rpsblast)</li><li>• Protein homology by domain architecture (cdart)</li></ul>
<p>Downloads</p> <ul style="list-style-type: none"><li>• Developer info</li></ul>	<p>Special</p> <ul style="list-style-type: none"><li>• Protein query vs. translated database (tblastn)</li><li>• Translated query vs. translated database (tblastx)</li></ul>
<p>Other resources</p> <ul style="list-style-type: none"><li>• References</li><li>• NCBI Contributors</li><li>• Mailing list</li><li>• Contact us</li></ul>	<p>Meta</p> <ul style="list-style-type: none"><li>• Retrieve results</li></ul>

Chicken, puffer fish, zebrafish

- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes

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

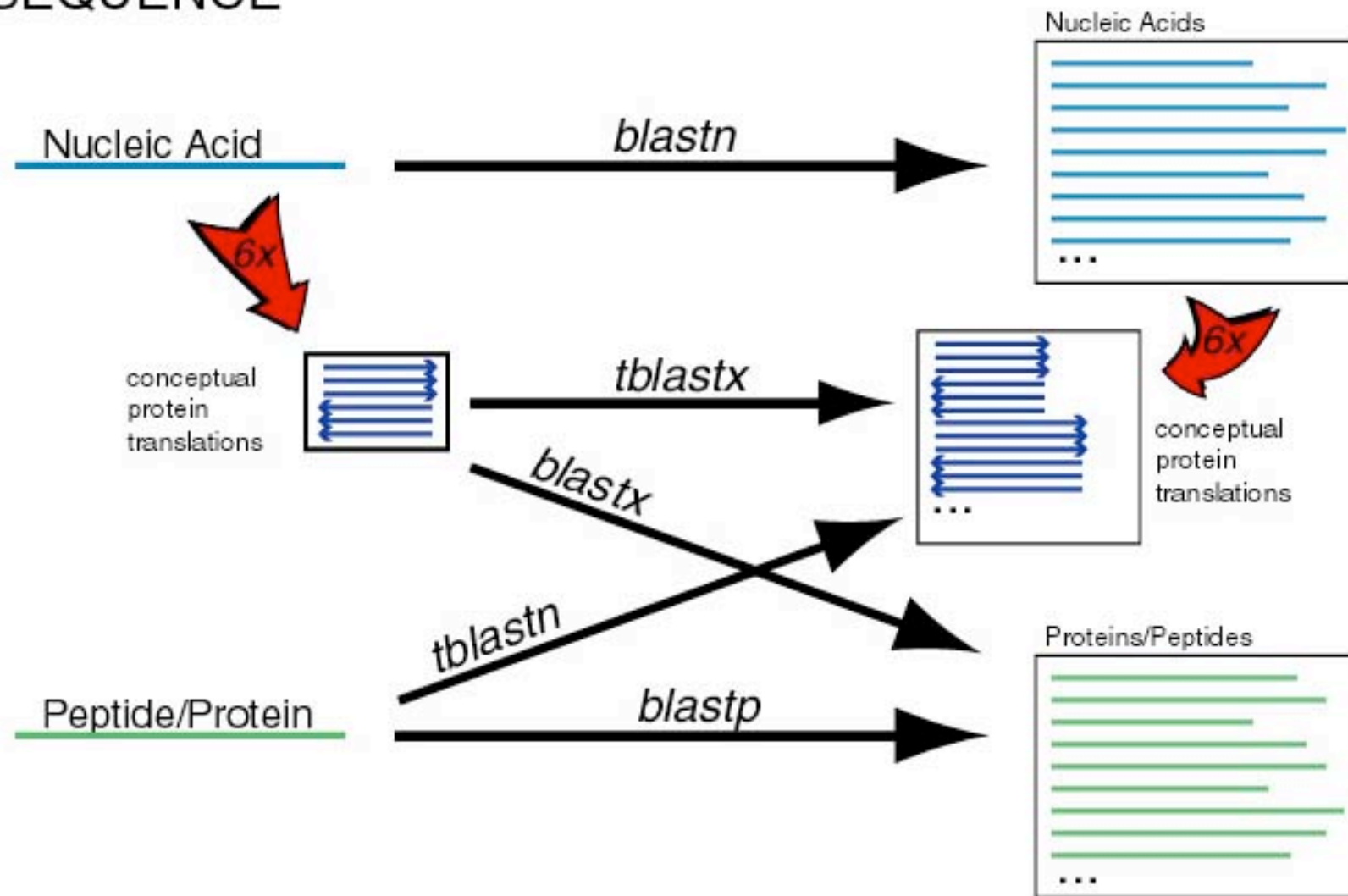
# BLAST family

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



# 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  
CGGSLISEDWVVTAAHCGVRTSEILIAGEFDQGSDEDNIQVLRIAKVKQPKYSILTVNND  
ITLLKLASPARYSQTISAVCLPSVDDDAGSLCATTGWGRTKYNANKSPDKLERAALPLLT  
NAECKRSWGRRLTDVMICGAASGVSSCMGDSGGPLVCQKDGAYTLVAIVSWASDTCSSASS  
GGVYAKVTKIIPWVQKILSSN
```

# Query Sequence and Database

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

Search

Set subsequence From:

Choose database nr

Do CD-Search

Now: **BLAST!** or **Reset query** **Reset all**

Click here to **Run the BLAST**

- **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.

# Choose Alignment Arguments

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

Search your query again **ONLY** given specie.

Options for advanced blasting

[Limit by entrez query](#)  or select from:

[Composition-based statistics](#)

[Choose filter](#)

[Expect](#)

[Word Size](#)

[Matrix](#)

[PSSM](#)

[Other advanced](#)

[PHI pattern](#)

## Example Entrez Queries

nucleotide all[Filter] NOT mammalia[Organism]

decapoda[Organism]

biomol mrna[Properties]

biomol genomic[Properties]

## Other Advanced

-v 2000 descriptions

-b 2000 alignments

# Format the Result

The screenshot shows the NCBI BLAST formatting interface. At the top, the NCBI logo is on the left, and the text "formatting BLAST" is in the center. Below the logo are links for "Nucleotide", "Protein", "Translations", and "Retrieve results for an RID". A message states: "Your request has been successfully submitted and put into the Blast Queue." Below this, it says "Query = (261 letters)". A horizontal bar represents the query sequence, with a red segment labeled "Tryp\_SPc" indicating a conserved domain. A callout box points to this segment with the text: "The Possible conserved domains in the query sequence detected by the NCBI CDD". Below the bar, a button labeled "CD search result summary" is visible. A red box highlights the "The request ID is" field, which contains the value "1098198699-4436-205850111949.BLASTQ1". Below this field are "Format!" and "Reset all" buttons. A callout box points to the "Format!" button with the text: "Click here to do the real work". At the bottom, a message says: "The results are expected to be ready in 1 minutes 10 seconds but may be done sooner. Please press the 'Format!' button to see the results. You may change the formatting options for your result via the form below and press 'FORMAT!' again. You may also use a valid request ID to see other recent jobs."

# BLAST Output

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

Sequences producing significant alignments:

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

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

About

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.

- 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

**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)
- Immunoglobulin 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 cyclase signalling in retraction

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

Max Planck Institut für Züchtungsbiologie, D-50829 Köln, Germany

Cyclic AMP is an important second messenger in prokaryotes and eukaryotes<sup>1</sup>, but its role in plant cells has generally been doubted<sup>2</sup> because of its low concentration and barely detectable amounts in plant cells. We used T-DNA tagging to create a mutant in the absence of the phytochrome A gene. The sequence tagged in this mutant was complementary DNA encoding a protein in a higher plant. Sequence analysis shows that this cyclase is probably soluble in the cytosol, has several repeats, and bears similarity to the cyclase of *Schizosaccharomyces pombe*. The *Escherichia coli* results showing that the *cry1* mutation levels, and in yeast its expression is repressed by a *cry1* mutation. Tobacco adenylyl cyclase activates cell division. This finding, together with the finding that adenylyl cyclase inhibitor dideoxyadenosine inhibits cell division in the presence of auxin, suggests that cAMP is involved in auxin-triggered cell division in higher plants.

## 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 incorrect. We apologize for the error caused.

*Note from the Editor:* Our apologies for the inaccuracy of parts of this Letter. The retraction and awaits the publication of the corrected version.

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

ing organism, the  
it were likely to  
rched for similar  
rched contained  
theses about the

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

umiliation was a

# 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!

