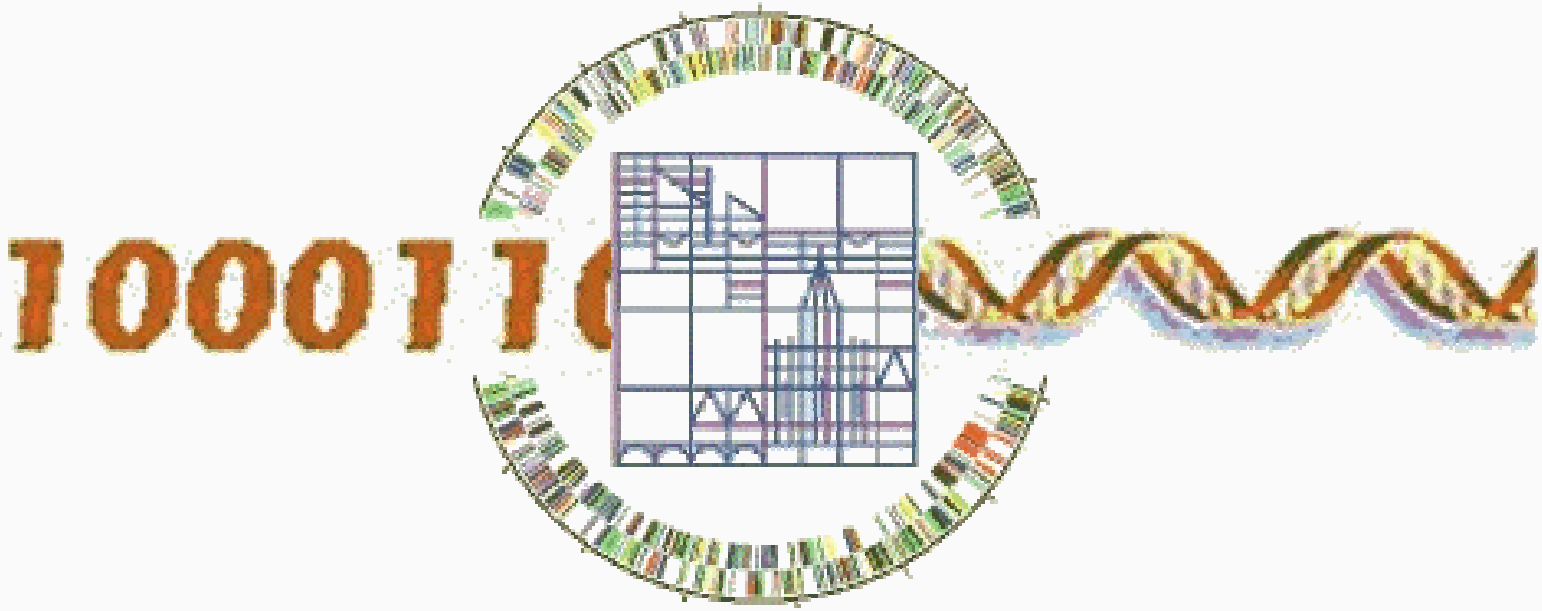
Bioinformatics and Drug Research

Klaus P. Schaefer



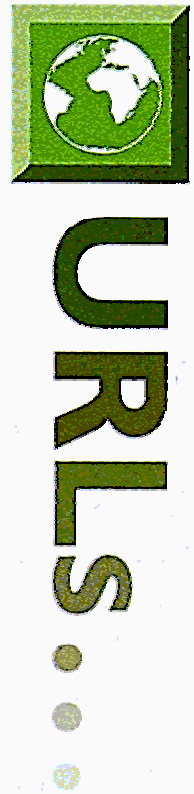
Sequence Searching and Analysis I

Summer Course - Lecture 03 - Version of May 26, 2003

http://www.uni-konstanz.de/FuF/Bio/Bioinformatik/

URLs for Sequence Analysis

- **BLAST WebSite** http://www.ncbi.nlm.nih.gov/BLAST/
- **BLAST FTP Site** ftp://ncbi.nlm.nih.gov/blast/
- **FASTA at EBI** http://www2.ebi.ac.uk/fasta3/
- **FASTA FTP site** ftp://ftp.virginia.edu/pub/fasta
- **Sequence Retrieval System (SRS) at EBI** http://srs.ebi.ac.uk/
- **Sequence Search Site** http://sss.berkeley.edu/



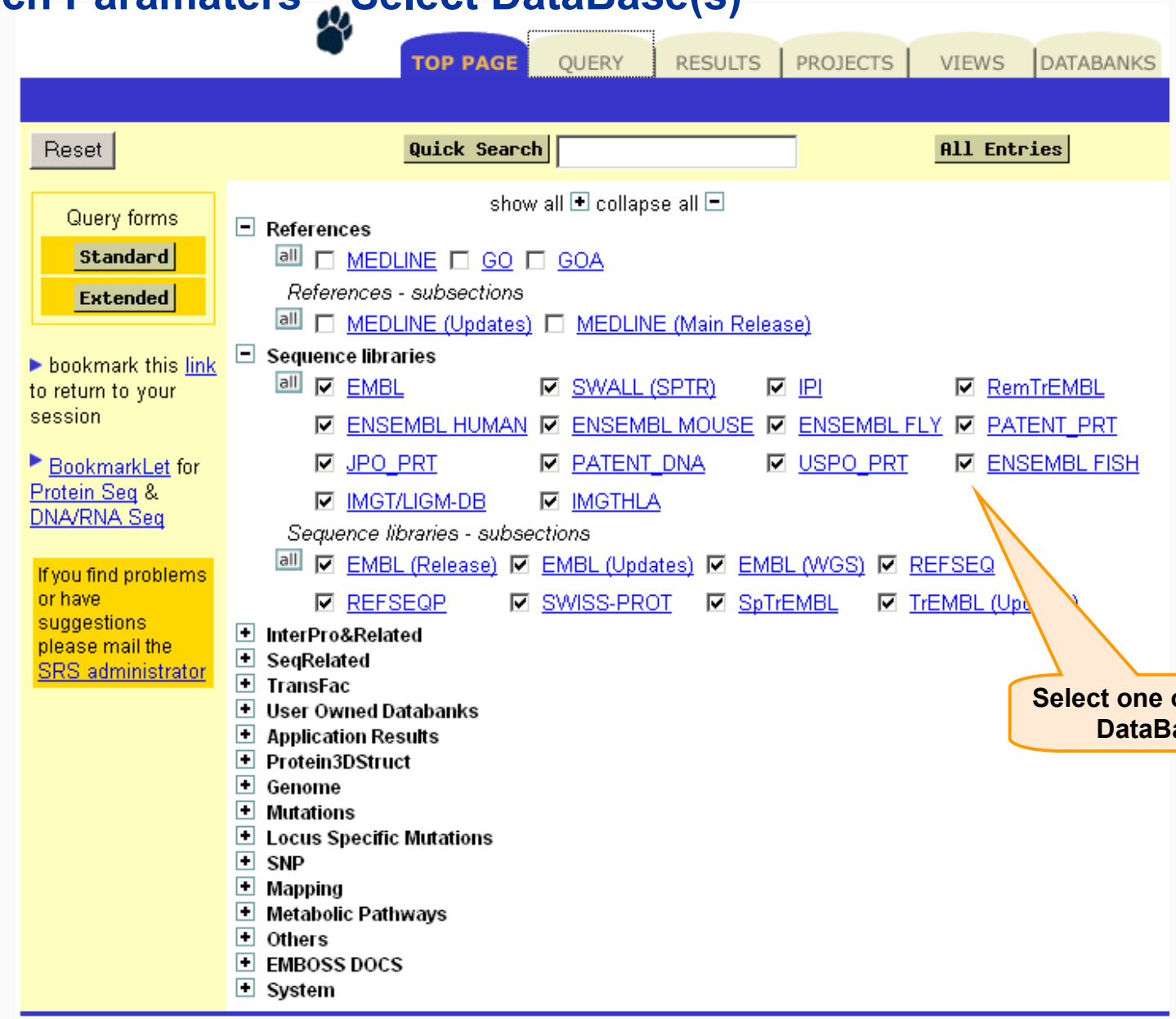
Searching for a Sequence

WebSite Infos - The Sequence Retrieval System –
Permanent Session: If you want to return to your session from time to time, if you think you may need to move work from one session to another, if you want a record of all your sessions in one place and the ability to recall any of your previous sessions or begin new sessions, permanent sessions are the way to go.
Temporary Session: Use temporary sessions for simple lookups, for example, looking up something once in a while or to run an occasional Blast search.

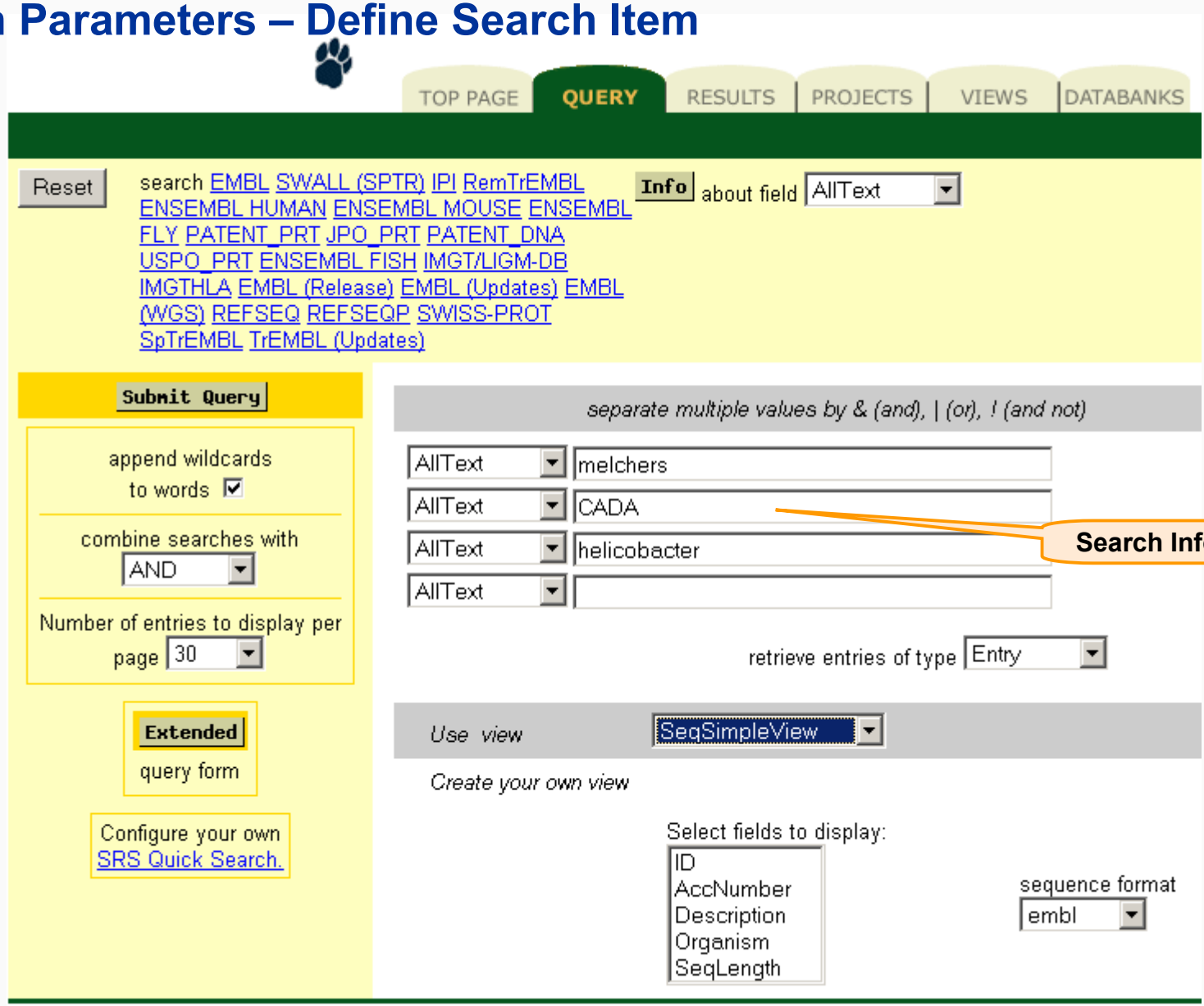
http://srs.ebi.ac.uk/



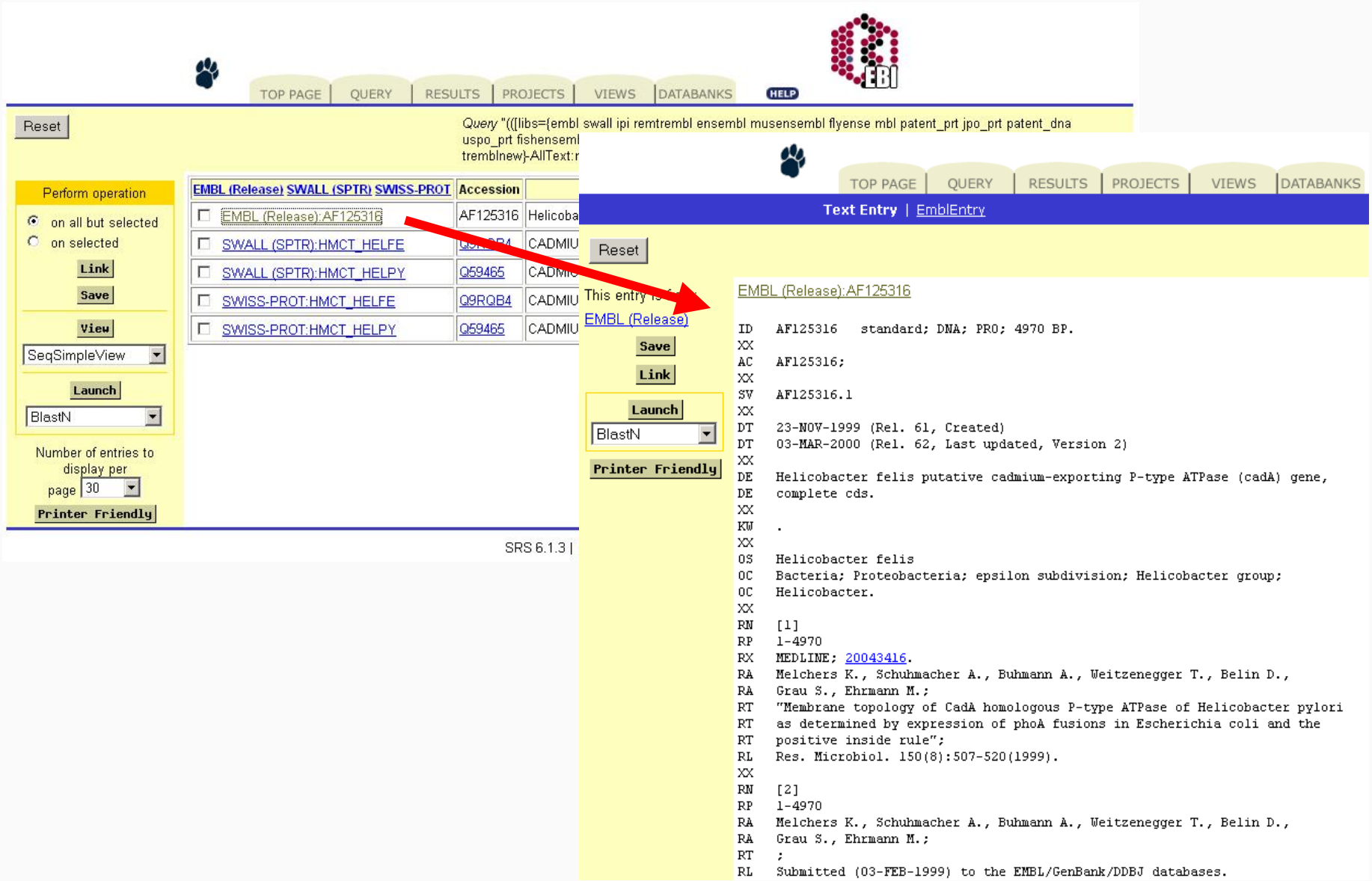
Search Paramaters – Select DataBase(s)



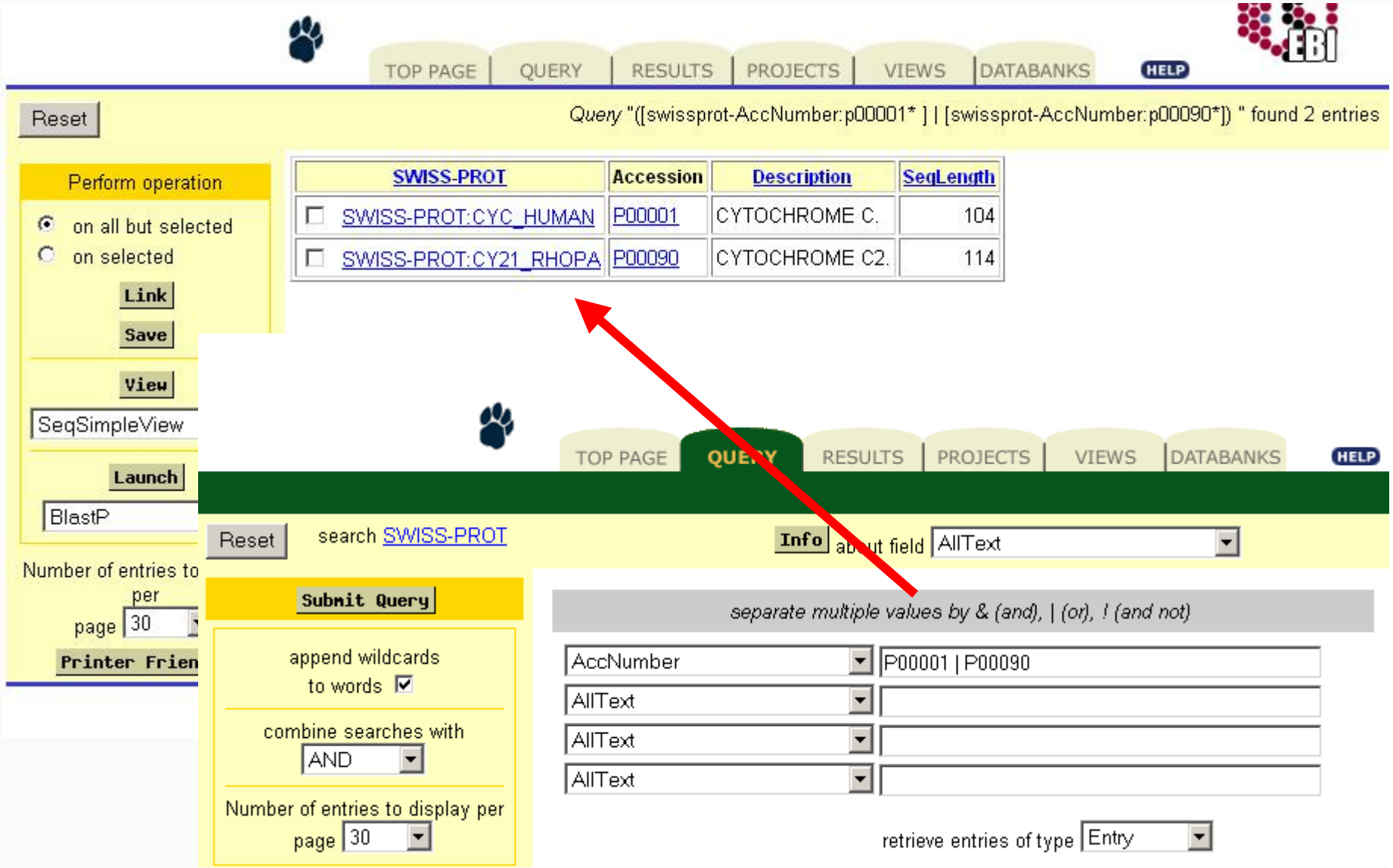
Search Parameters – Define Search Item



From the Hit it is only a Click to the DataBase File



Sequence Retrieval is easy with the Accession Number

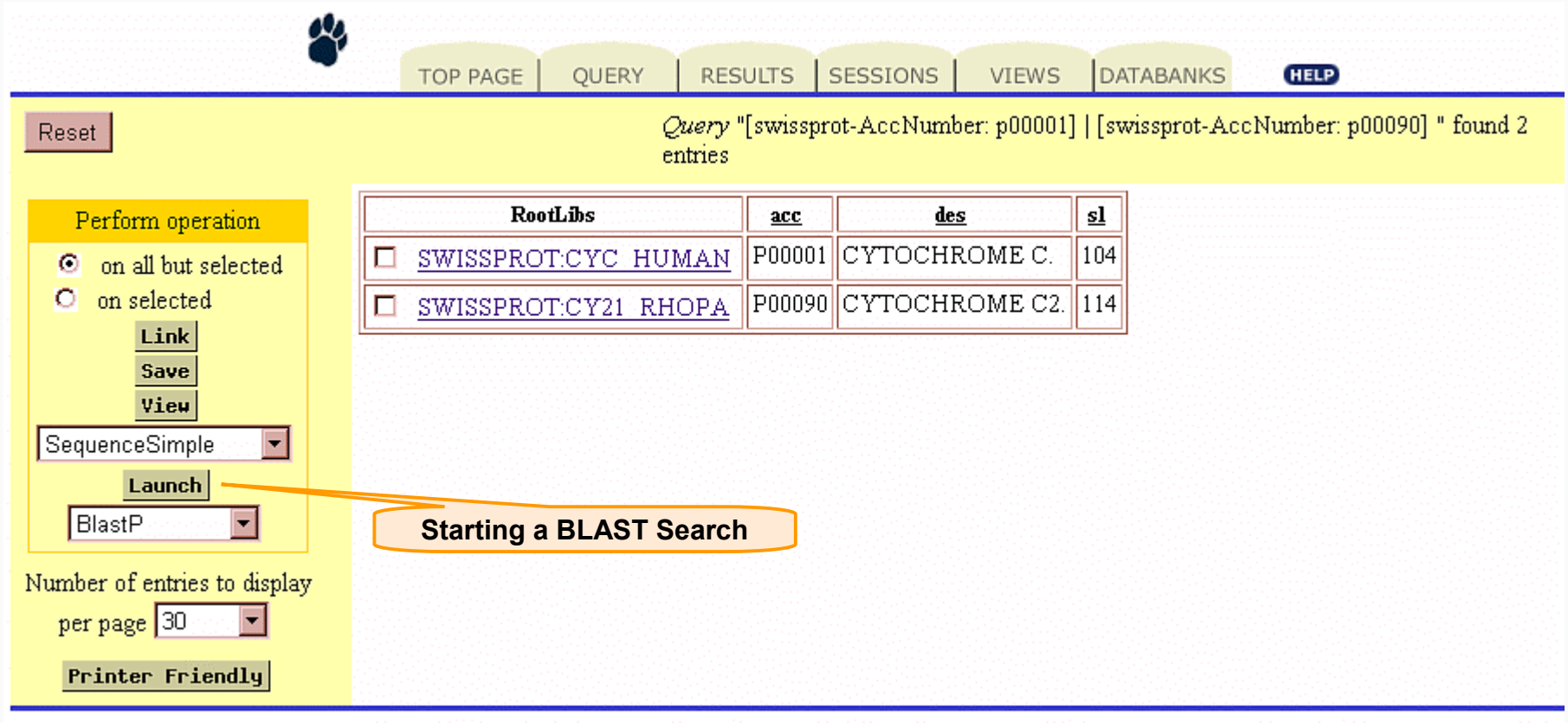




Blast

- **BLAST, or Basic Local Alignment Search Tool, uses the method of Altschul et al. (J. Mol. Biol. 215: 403-410 (1990)) to search for similarities between a query sequence and all the sequences in a database.** Reference
- **BLAST is a statistically driven search method that finds regions of similarity between your query and database sequences and produces gapped alignments of these regions. Within these aligned regions, the sum of the scoring matrix values of their constituent symbol pairs is higher than some level that you would expect to occur by chance alone.**

From here you can start a Blast...





The Blast Suite of Homology Search Programs

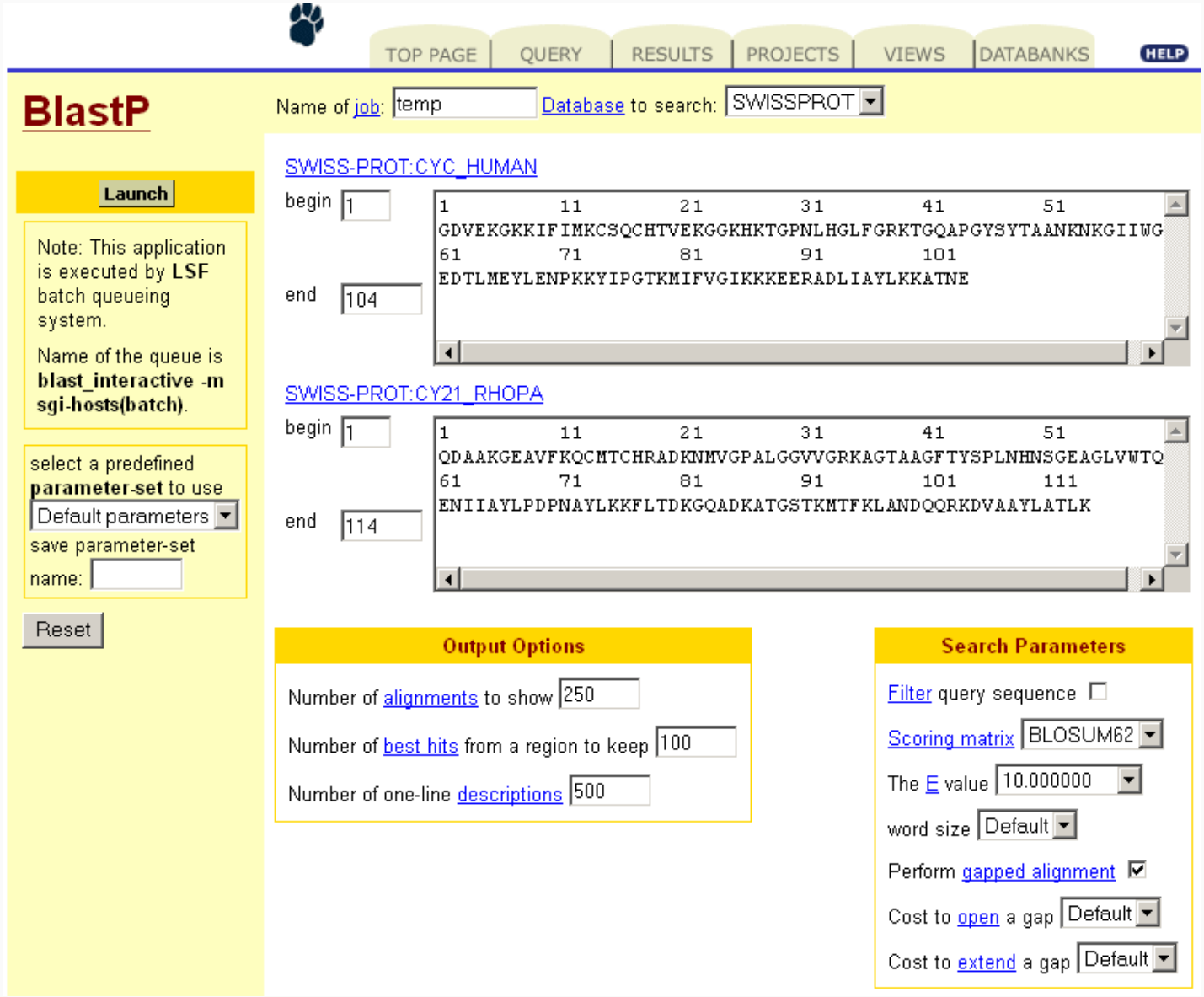
- The four programs perform the following tasks:
  - **blastp** compares an amino acid sequence against a protein database
  - **blastn** compares a nucleotide sequence against a nucleotide database
  - **blastx** compares the six-frame conceptual translation products of a nucleotide sequence (both strands) against a protein database
  - **tblastn** compares a protein sequence against a nucleotide database dynamically translated in all six reading frames (both strands)
  - **tblastx** compares the six-frame conceptual translation products of a nucleotide sequence (both strands) against a nucleotide database dynamically translated in all six reading frames (both strands)

Table I. BLAST variants for different searches[a]

Program	Query	Database	Comparison	Common use
blastn	DNA	DNA	DNA level	Seek identical DNA sequences and splicing patterns
blastp	Protein	Protein	Protein level	Find homologous proteins
blastx	DNA	Protein	Protein level	Analyze new DNA to find genes and seek homologous proteins
tblastn	Protein	DNA	Protein level	Search for genes in unannotated DNA
tblastx	DNA	DNA	Protein level	Discover gene structure

[a]Similar variant programs are available for FASTA. Protein-level searches of DNA sequences are performed by comparing translations of all six reading frames.

Where do you want to Blast today?



A BlastP Result



Reset

Query "[BlastP-JobName:temp_job1]" found 239 entries

Perform operation

- (•) on all but selected
- ( ) on selected

Link

Save

View

* Complete entries *

Launch

BlastP

Number of entries to display per page 30

Printer Friendly

☐ BLASTP:temp_job1_swissprot_1_CYC_HUMAN

```
>SW:CYC_HUMAN P00001 CYTOCHROME C.
          Length = 104

 Score =  220 bits (560), Expect = 4e-58
 Identities = 104/104 (100%), Positives = 104/104 (100%)

Query: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG 60
           GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG
Sbjct: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG 60

Query: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
           EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
Sbjct: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
```

☐ BLASTP:temp_job1_swissprot_2_CYC_MACMU

```
>SW:CYC_MACMU P00002 CYTOCHROME C.
          Length = 104

 Score =  218 bits (555), Expect = 1e-57
 Identities = 103/104 (99%), Positives = 103/104 (99%)

Query: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG 60
           GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGI WG
Sbjct: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGITWG 60

Query: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
           EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
Sbjct: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
```



Nucleotide Sequence Databases available for Blast I

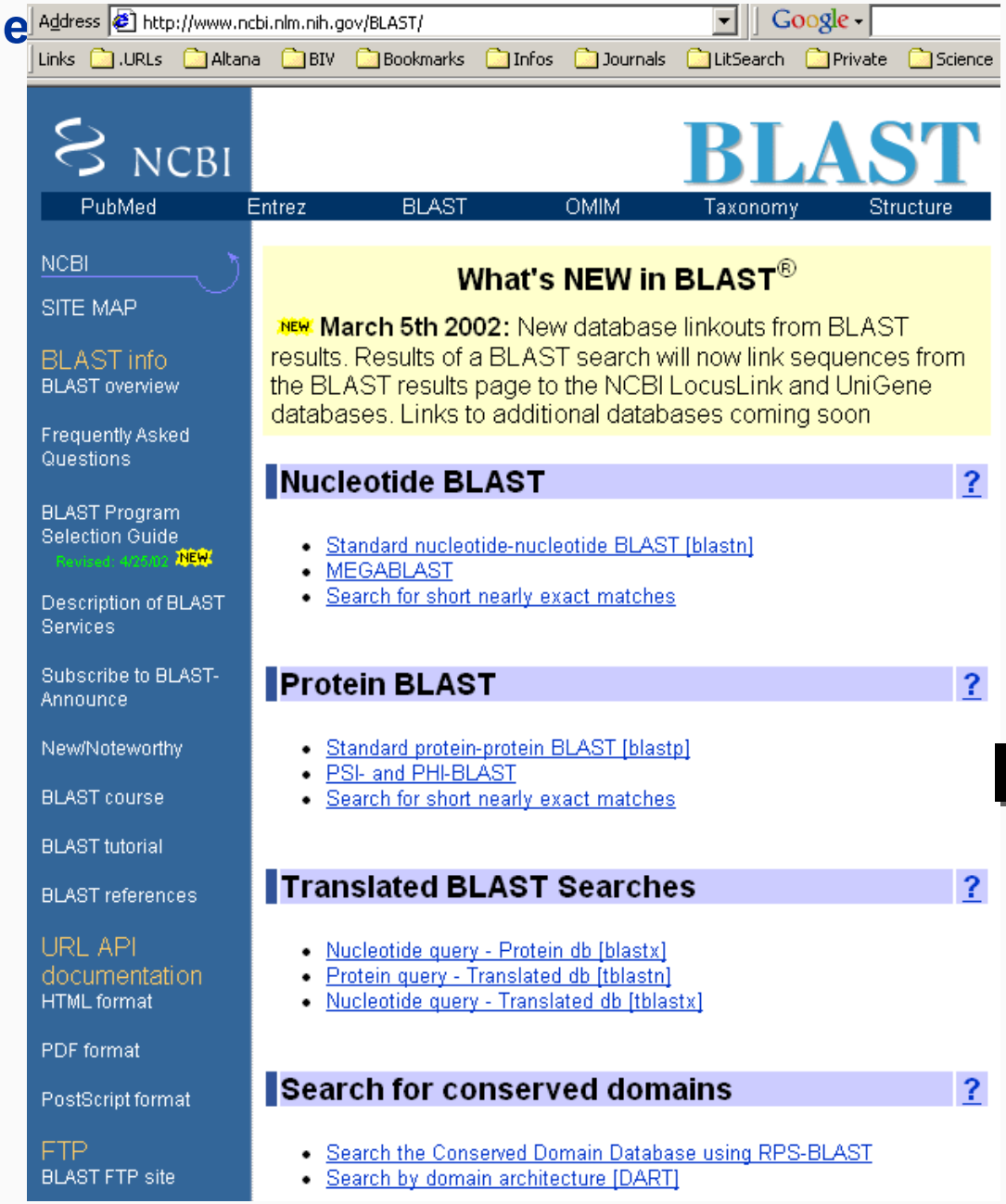
- nr
  - All Non-redundant GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences)
- month
  - All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
- dbest
  - Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions
- dbsts
  - Non-redundant Database of GenBank+EMBL+DDBJ STS Divisions
- htgs
  - htgs unfinished High Throughput Genomic Sequences: phases 0, 1 and 2 (finished, phase 3 HTG sequences are in nr)
- Drosophila genome
  - Drosophila genome provided by Celera and Berkeley Drosophila Genome Project (BDGP).
- yeast
  - Yeast (*Saccharomyces cerevisiae*) genomic nucleotide sequences
- *E. coli*
  - *E. coli* genomic nucleotide sequences



Nucleotide Sequence Databases available for Blast II

- pdb
  - ➔ Sequences derived from the 3-dimensional structure
- kabat [kabatnuc]
  - ➔ Kabat's database of sequences of immunological interest
- vector
  - ➔ Vector subset of GenBank(R), NCBI, in ftp://ncbi.nlm.nih.gov/blast/db/
- mito
  - ➔ Database of mitochondrial sequences"
- alu
  - ➔ Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. It is available by anonymous FTP from ncbi.nlm.nih.gov (under the /pub/jmc/alu directory). See "Alu alert" by Claverie and Makalowski, Nature 371:752 (1994).
- epd
  - ➔ Eukaryotic Promotor Database
- gss
  - ➔ Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.

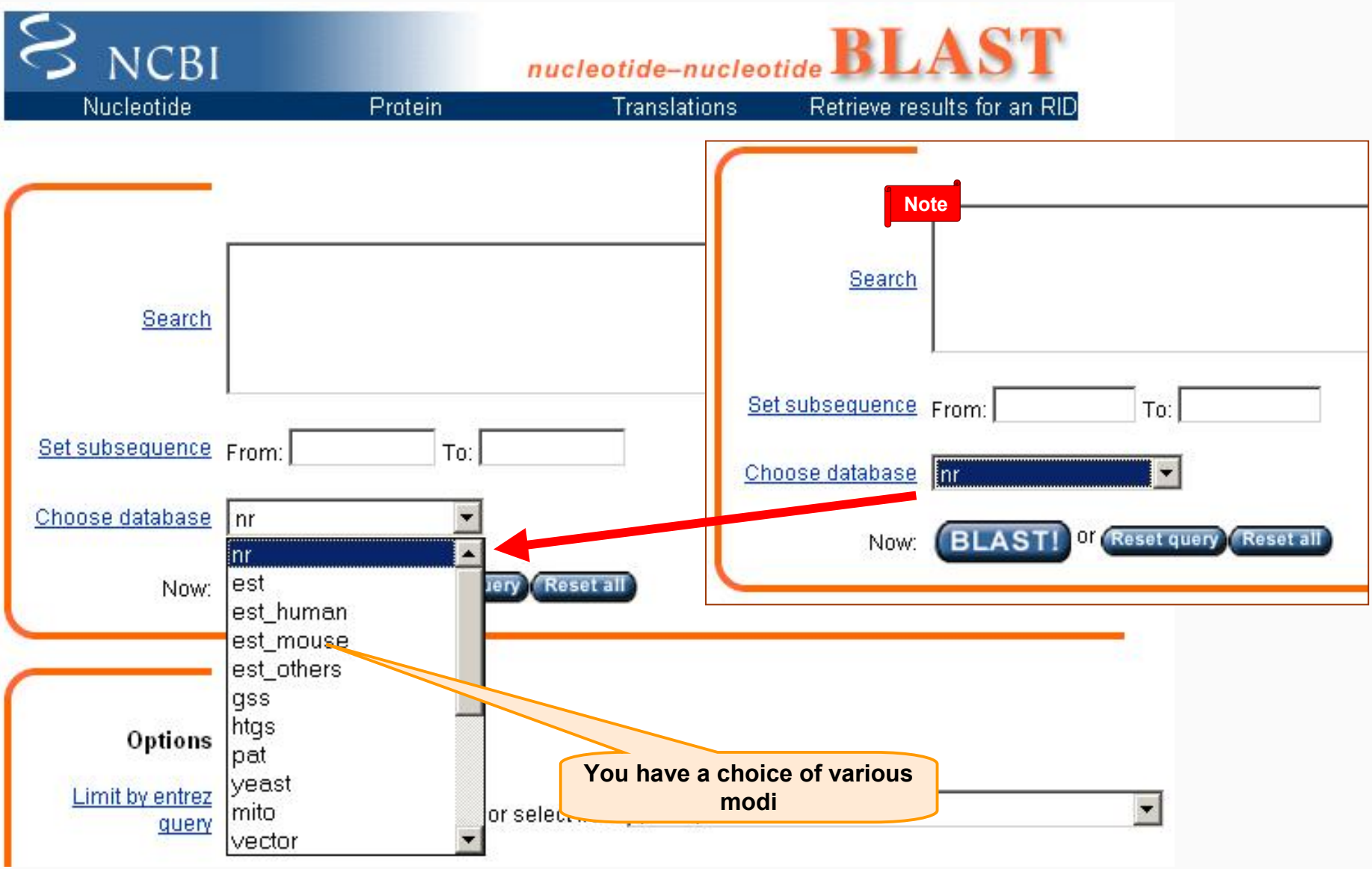
The



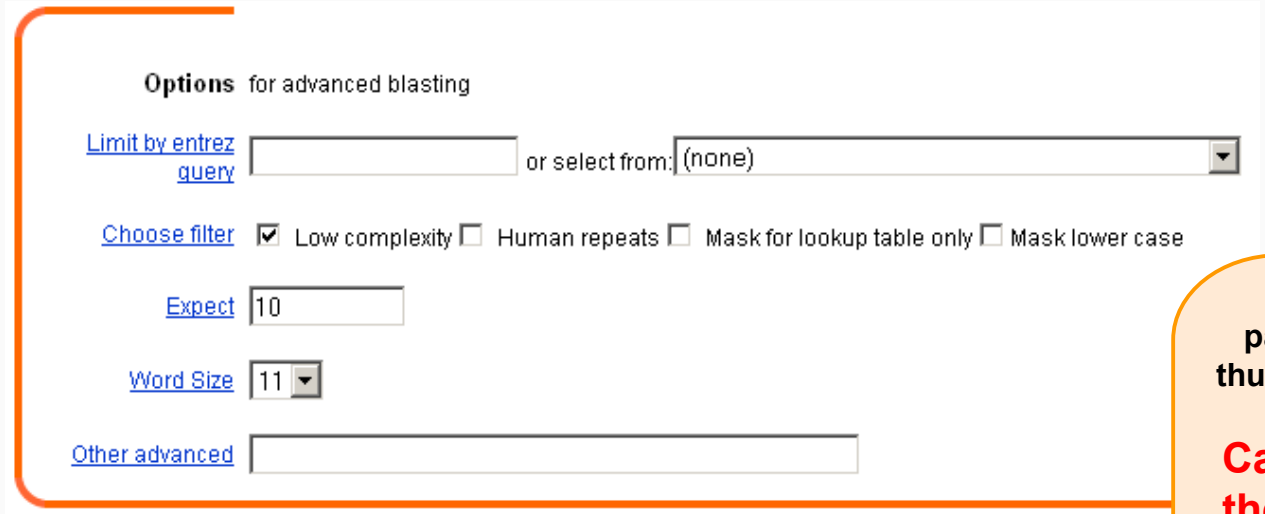
WebSite Infos - What is BLAST?
BLAST® (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships.

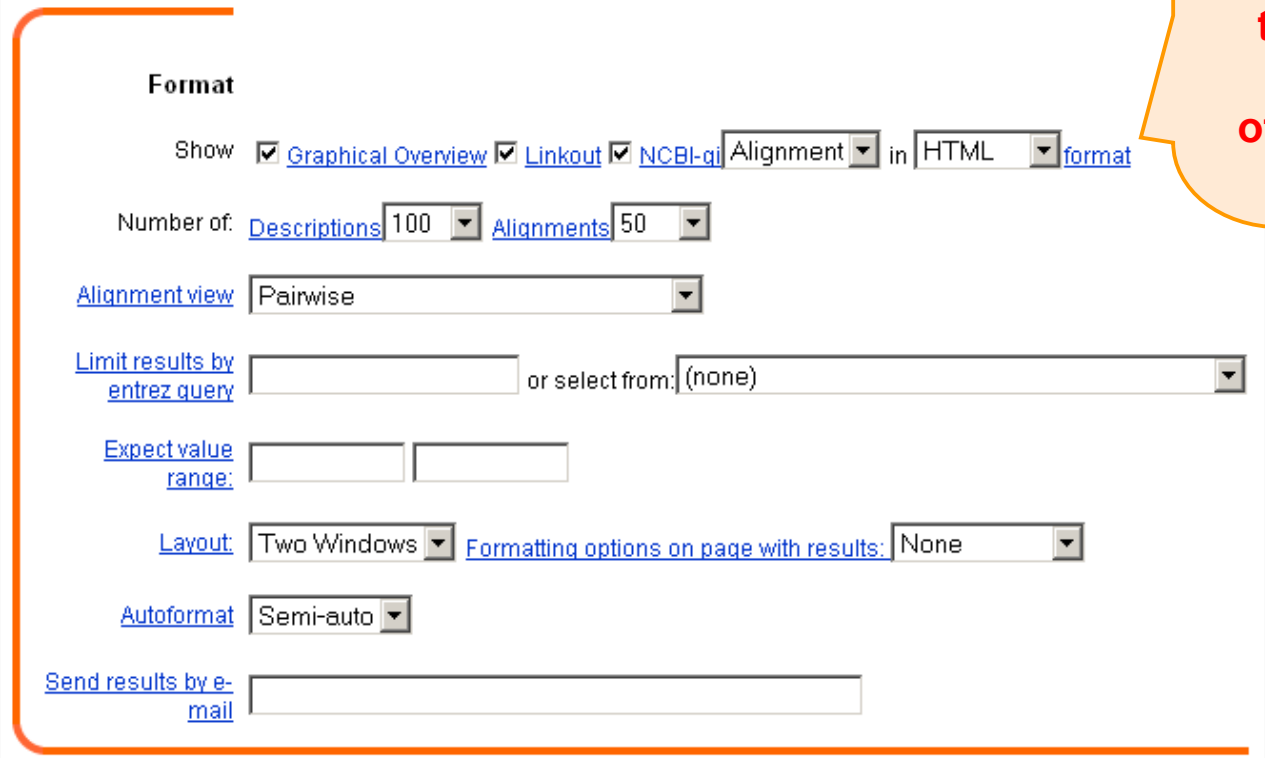
http://www.ncbi.nlm.nih.gov/BLAST/

Decide first on the Blast Mode





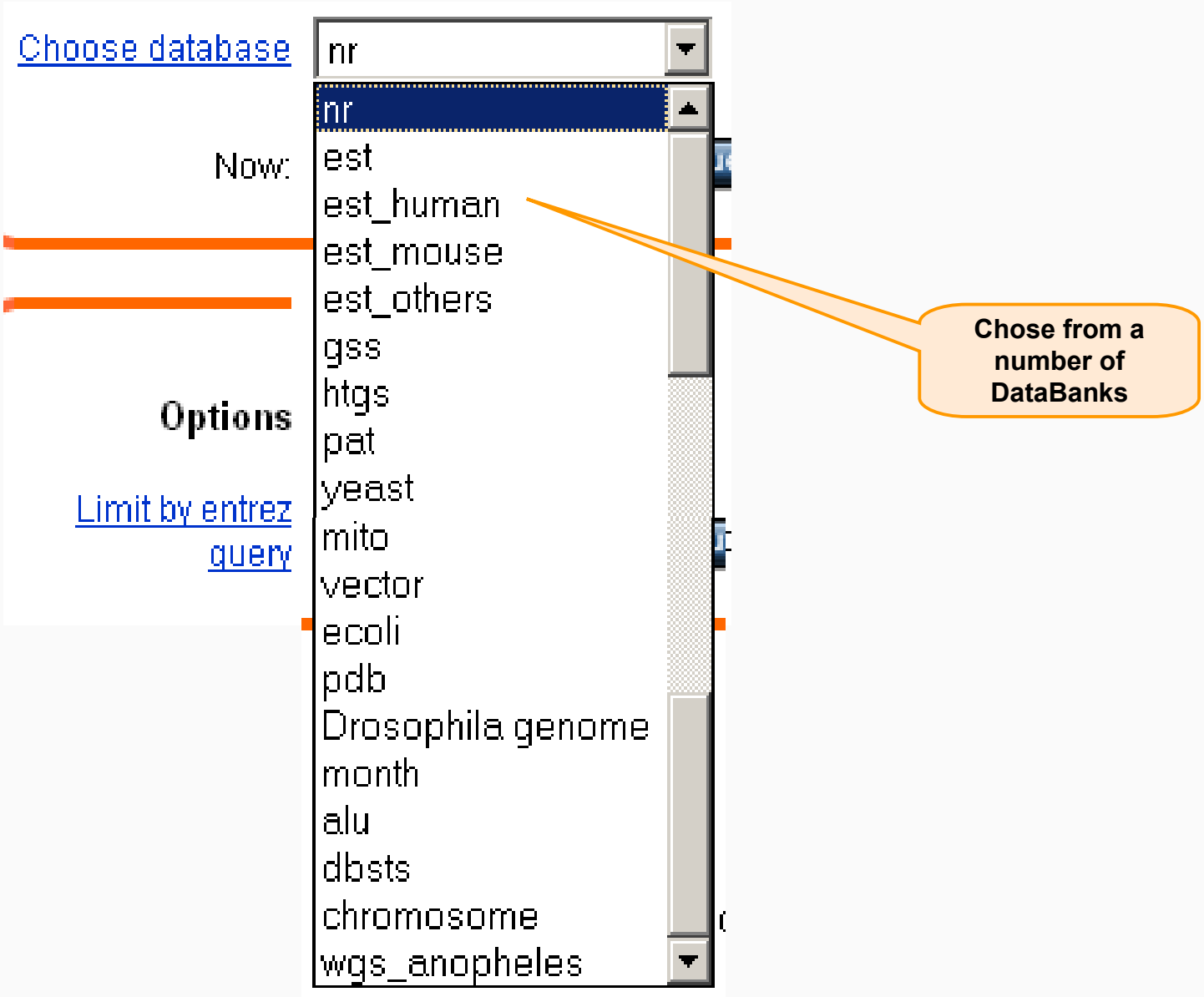




Here you can set even more parameters and some variables - thus controlling your output results.

Caveat: inform yourself on the meaning and effects of the various parameters before using them, otherwise results may be misleading!

In which DataBank will you search?



Inputting your Sequence

Search

```
gacgtatagcatcgattacgatcgtatatgacgcgtacgtgtactatcgatgctatccga
tatcgtacttatagctagctatatgcatcgagctcggcgatgcattatattcgatagtat
cgtagcatgactacgtagcttagcgtatagcatcgtagcatatttacgactgtagcggca
tgctagcattagctagctagctgtatatcgatgctagctagtgcatgc
```

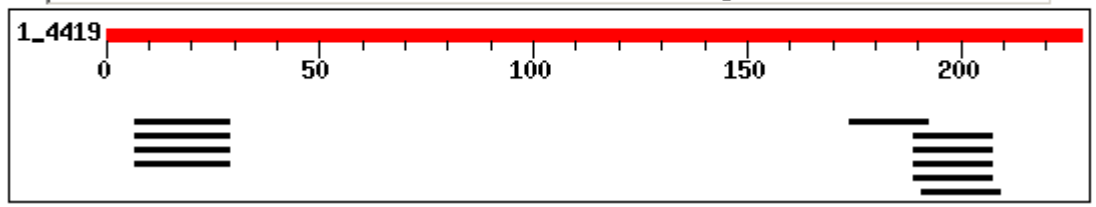
Set subsequence From: To:

Choose database nr

Now: BLAST! or Reset query

Distribution of 10 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Alignments

>gi|17570973|ref|NG_000621.1| Drosophila melanogaster (Doa) on chromosome 1
Length = 44642

Score = 38.2 bits (19), Expect = 3.7
Identities = 22/23 (95%)
Strand = Plus / Minus

```
Query: 8     agcatcgattacgatcgtatatg 30
              ||||||||||| ||||||||||||
Sbjct: 13031 agcatcgattgcgatcgtatatg 13009
```

Sequences producing significant alignments:	Score (bits)	E Value
gi|17570973|ref|NG_000621.1| Drosophila melanogaster (Doa)...	38	3.7
gi|17571060|ref|NG_000552.1| Drosophila melanogaster (Cdk4...	38	3.7
gi|13811822|gb|AC091501.1|AC091501 Drosophila melanogaster,...	38	3.7
gi|13129407|gb|AC006170.11|AC006170 Drosophila melanogaster...	38	3.7
gi|13027522|gb|AC007854.4|AC007854 Drosophila melanogaster,...	38	3.7
gi|10727503|gb|AE003806.2|AE003806 Drosophila melanogaster ...	38	3.7
gi|7301701|gb|AE003767.1|AE003767 Drosophila melanogaster g...	38	3.7
gi|16030170|emb|AL583887.9|AL583887 Mouse DNA sequence from...	38	3.7
gi|4309695|gb|AC005647.3|AC005647 Drosophila melanogaster, ...	38	3.7
gi|14587202|dbj|AP003199.3| Oryza sativa (japonica cultivar...	38	3.7

Note



FASTA Format Description

- A sequence in FASTA format begins with a single-line description, followed by lines of sequence data.
- The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column.
- It is recommended that all lines of text be shorter than 80 characters in length.
- An example sequence in FASTA format is:

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNN
SFNVATLPAEKMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEK
RRVKVYLPQMKIEEKYNLTSVLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELS
EDGIEMAGSTGVIEDIKHSPESEQFRADHPFLFLIKHNPTNTIVYFGRYWSP
```



IUPAC Nucleotide Codes

- **Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters.**

  Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue).

- **A → Adenosine**
- **B → G T C**
- **C → Cytidine**
- **D → G A T**
- **G → Guanine**
- **H → A C T**
- **K → G T (keto)**
- **M → A C (amino)**
- **N → A G C T (any)**
- **R → G A (purine)**
- **S → G C (strong)**
- **T → Thymidine**
- **U → Uridine**
- **V → G C A**
- **W → A T (weak)**
- **Y → T C (pyrimidine)**
- **- gap of indeterminate length**



IUPAC Aamino Acid Codes

- A alanine
- B aspartate or asparagine
- C cystine
- D aspartate
- E glutamate
- F phenylalanine
- G glycine
- H histidine
- I isoleucine
- K lysine
- L leucine
- M methionine
- N asparagine
- P proline
- Q glutamine
- R arginine
- S serine
- T threonine
- U selenocysteine
- V valine
- W tryptophan
- X any
- Y tyrosine
- Z glutamate or glutamine
- * translation stop
- - gap of indeterminate length

A Blast Result

NCBI results of BLAST

BLASTP 2.2.3 [Apr-24-2002]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejand
Jinghui Zhang, Zheng Zhang, Webb Miller, and De
"Gapped BLAST and PSI-BLAST: a new generation o
programs", Nucleic Acids Res. 25:3389-3402.

RID: 1021293204-020389-23844

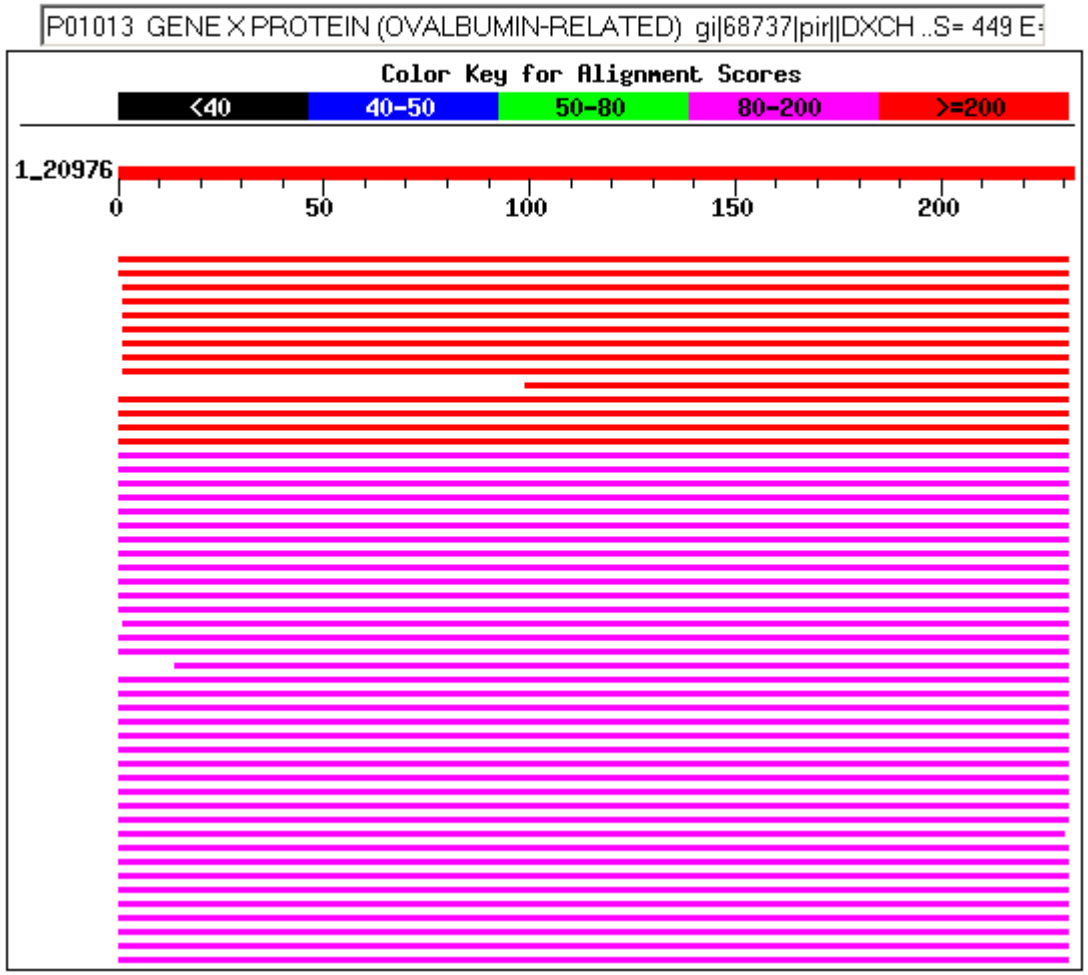
Query= gi|129295|sp|P01013|OVAX_CHICK GENE X PR
(232 letters)

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF
922,422 sequences; 289,623,403 total

If you have any problems or questions with the
please refer to the **BLAST FAQs**

Taxonomy reports

Distribution of 100 Blast Hits on the Query Sequence





You may want to use the Advanced Options Page

Sequences producing significant alignments:	Score (bits)	E Value
gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RE...	449	e-126
gi|129296|sp|P01014|OVAY_CHICK GENE Y PROTEIN (OVALBUMIN-RE...	367	e-101
gi|223299|prf||0705172A ovalbumin [Gallus gallus]	293	1e-78
gi|230201|pdb|1OVA|A Chain A, Ovalbumin (Egg Albumin) >gi|2...	293	1e-78
gi|129293|sp|P01012|OVAL_CHICK Ovalbumin (Plakalbumin) (All...	292	2e-78
gi|15826578|pdb|1JTI|A Chain A, Loop-Inserted Structure Of ...	291	4e-78
gi|808969|emb|CAA23682.1| (V00383) reading frame [Gallus ga...	290	7e-78
gi|3136338|gb|AAC16664.1| (AF064546) ovalbumin [Meleagris g...	285	2e-76
gi|129294|sp|P19104|OVAL_COTJA OVALBUMIN >gi|104501|pir||S1...	281	3e-75
gi|763101|emb|CAA23685.1| (V00387) seventh exon [Gallus gal...	265	3e-70
gi|4758906|ref|NP_004146.1| (NM_004155) serine (or cysteine...	205	4e-52
gi|481621|pir||S38962 serpin - pig	202	2e-51
gi|4826902|ref|NP_005015.1| (NM_005024) serine (or cysteine...	202	2e-51
gi|6678097|ref|NP_033280.1| (NM_009254) serine protease inh...	201	5e-51
gi|6018510|emb|CAA04937.1| (AJ001698) hurpin [Homo sapiens]	198	4e-50
gi|8393956|ref|NP_036529.1| (NM_012397) serine (or cysteine...	197	5e-50
gi|7522623|pir||JC7118 headpin serine proteinase inhibitor ...	197	5e-50
gi|15214956|gb|AAH12609.1|AAH12609 (BC012609) Similar to se...	197	7e-50
gi|15988197|pdb|1JRR|A Chain A, Human Plasminogen Activator...	197	8e-50
gi|4505595|ref|NP_002566.1| (NM_002575) serine (or cysteine...	197	8e-50
gi|189545|gb|AAA36413.1| (J02685) plasminogen activator inh...	195	3e-49
gi|15826846|ref|NP_035586.1| (NM_011456) serine protease in...	194	4e-49
gi|6678101|ref|NP_033282.1| (NM_009256) serine protease inh...	194	6e-49
gi|12833752|dbj|BAB22650.1| (AK003220) data source:MGD, sou...	193	1e-48
gi|417185|sp|P80229|ILEU_PIG Leukocyte elastase inhibitor (...	192	3e-48
gi|15826842|ref|NP_035582.1| (NM_011452) serine protease in...	191	4e-48
gi|20345217|ref|XP_111407.1| (XM_111407) similar to serine ...	189	3e-47
gi|20345377|ref|XP_109319.1| (XM_109319) similar to EIA [Mu...	187	7e-47
gi|15826848|ref|NP_035589.1| (NM_011459) serine protease in...	187	9e-47
gi|20345219|ref|XP_110035.1| (XM_110035) similar to SPI3C [...	187	1e-46
gi|18251228|gb|AAL65909.1|AF425083_1 (AF425083) NK21B [Mus ...	186	1e-46

The Blast Output Parameters I

Reference

- **Bit Score**
  - **Each aligned segment pair has a normalized score expressed in bits that lets you estimate the magnitude of the search space you would have to look through before you would expect to find an HSP score as good as or better than this one by chance. If the bit score is 30, you would have to score, on average, about 1 billion independent segment pairs (2(30)) to find a score this good by chance. Each additional bit doubles the size of the search space. This bit score represents a probability; one over two raised to this power is the probability of finding such a segment by chance. Bit scores represent a probability level for sequence comparisons that is independent of the size of the search.**
  - **The size of the search space is proportional to the product of the query sequence length times the sum of the lengths of the sequences in the database. This product, referred to as N in Altschul's publications, is multiplied by a coefficient K to get the size of the search space. When searching protein databases with protein queries, K is about 0.13. BLAST uses estimates of K produced before it runs by random simulation**

Note

(Altschul & Gish, Methods in Enzymology 266:460-480 (1996)).



The Blast Output Parameters II

- **E Value**
  - **There is a probability, for instance 3.3e-160 in the example, associated with each pairwise comparison in the list and with each segment pair alignment. In the list, this number (which means 3.3 x 10(-160)) is the probability that you would observe a score or group of scores as high as the observed score purely by chance when you do a search of this size.**
  - **An ideal search would find hits that go from extremely unlikely to ones whose best scores should have occurred by chance alone (that is, with probabilities approaching 1.0).**



The Blast Output Parameters III

- **N**
  - **If you specify ungapped alignments to BLAST, a third column of data will appear in your output under the heading N. The number in that column indicates how many HSPs were involved in computing the statistics for the sequence. If the number is greater than 1, the scores of multiple HSPs were combined to produce the result. See the ALGORITHM topic for more information.**



The Top Hits on the Blast List

BLASTP:temp_job1_SW_CYC_HUMAN

```
>SW:CYC_HUMAN P00001 homo sapiens (human), and pan troglodytes (chimpanzee).
           cytochrome c. 12/98
           Length = 104

 Score =  222 bits (560), Expect = 1e-58
 Identities = 104/104 (100%), Positives = 104/104 (100%)

Query: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG 60
           GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG
Sbjct: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG 60

Query: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
           EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
Sbjct: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
```

The top hit is the search sequence because it was included in the DataBank

BLASTP:temp_job1_SW_CYC_MACMU

```
>SW:CYC_MACMU P00002 macaca mulatta (rhesus macaque). cytochrome c. 11/97
           Length = 104

 Score =  220 bits (555), Expect = 6e-58
 Identities = 103/104 (99%), Positives = 103/104 (99%)

Query: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG 60
           GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGI WG
Sbjct: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGITWG 60

Query: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
           EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
Sbjct: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
```

The second hit is a sequence from a closely related species - a monkey

An Example further down the list ...



BLASTP:temp job1 SW CYC CANAL

```
>SW:CYC CANAL P53698 candida albicans (yeast). cytochrome c. 7/98
           Length = 109

 Score =  153 bits (382), Expect = 1e-37
 Identities = 68/103 (66%), Positives = 81/103 (78%)

Query: 1   GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG 60
           G  +KG  +F  +C QCHTVEKGG HK GPNLHG+FGRK+G A GYSYT ANK KG+ W
Sbjct: 7   GSEKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGVFGRKSGLAEGYSYTDANKKKGVEWT 66

Query: 61  EDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATN 103
           E T+ +YLENPKKYIPGTKM F G+KK ++R DL+ YLKKAT+
Sbjct: 67  EQTMSDYLENPKKYIPGTKMAFGGLKKPKDRNDLVTYLKKATS 109
```

Similiarity decreases - gaps appear



This is not a <homologous> Sequence

BLASTP:temp_job1_SW_C550_CYAPA

```
>SW:C550_CYAPA P48263 cyanophora paradoxa. cytochrome c-550 precursor (c550).
           11/97
           Length = 162

 Score = 33.2 bits (74), Expect = 0.15
 Identities = 30/102 (29%), Positives = 46/102 (44%), Gaps = 17/102 (16%)

Query: 3   VEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGED 62
           V++GK++F    C   CH     GG   KT PN+ GL         A      T     N         +
Sbjct: 51  VKRGKRLFNSTCGICHV---GGITKTNPNV-GLDSEALALA-----TPPRNN------IE 95

Query: 63  TLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE 104
           +L++Y++NP   Y   G++ I+   I       AD     ++   T E
Sbjct: 96  SLVDYMKNPTSY-DGSEEIY-DIHPSIRSADAFPKMRNLTEE 135
```



The Blast Search Summary Table

BLASTP:temp_job1_Statistics

```
  Database: swissprot
    Posted date:  Jul 27, 1999 12:38 PM
  Number of letters in database: 29,085,965
  Number of sequences in database:  80,000

Lambda     K      H
   0.314    0.137    0.407

Gapped
Lambda     K      H
   0.270   0.0470    0.230


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 6714533
Number of Sequences: 80000
Number of extensions: 279396
Number of successful extensions: 579
Number of sequences better than 10.0: 139
Number of HSP's better than 10.0 without gapping: 130
Number of HSP's successfully gapped in prelim test: 9
Number of HSP's that attempted gapping in prelim test: 421
Number of HSP's gapped (non-prelim): 140
length of query: 104
length of database: 29085965
effective HSP length: 46
effective length of query: 58
effective length of database: 25405965
effective search space: 1473545970
effective search space used: 1473545970
T: 11
A: 40
X1: 16 ( 7.3 bits)
X2: 38 (14.8 bits)
X3: 64 (24.9 bits)
S1: 42 (21.9 bits)
```

All variables used during the BLAST search are listed in the Summary Table

You can use a BLAST Site in Japan



WebSite Infos - IDEAS -
Integrated Database and Extended Analysis System for Nucleic Acids and Proteins
DBGET/LinkDB: Integrated Database Retrieval System

http://blast.genome.ad.jp/



Homology Search Example

Fig. 1. Two protein alignments. **(a)** An optimal global alignment of human cytochrome *c* (105 residues; SWISS-PROT accession number P00001) and *Rhodopseudomonas palustris* cytochrome c_2 (114 residues; SWISS-PROT accession number P00090). **(b)** An optimal local alignment of the human cystic fibrosis transmembrane conductance regulator (1480 residues; SWISS-PROT accession number P13569) and *Escherichia coli* nickel transport ATP-binding protein NIKD (253 residues; SWISS-PROT accession number P33593). Scores for both alignments are calculated using the BLOSUM62 amino acid substitution matrix[2], and affine gap costs[7] that assign the score $-(11 + k)$ to a gap of length k. The global alignment, with score 131, is required to include the whole of the two input sequences and is constructed using the Needleman–Wunsch algorithm[5]. The local alignment, with score 89, involves only those segments of the two input sequences that optimize the score and is constructed using the Smith–Waterman algorithm[6]. On the central line of each alignment, characters indicate identical amino acids and '+' signs indicate similar amino acids (i.e. those whose alignment receives a positive score).

An Example from the TRENDS Booklet
Results on next Page

Homology Search



(a)

```
P00001     1  MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNK---GI   58
              D  KG+ +F     QC T  +  K+  GP L G+ GRK G A G++Y+  N N    G+
P00090     1  Q-DAAKGEAVF----KQCMTCHRADKNMVGPALGGVVGRKAGTAAGFTYSPLNHNSGEAGL   56

P00001    59  IWGEDTLMEYLENPKKYIP--------------GTKMIFVGIKKKEERADLIAYLKKATNE  105
              +W ++ ++ YL +P  Y+                TKM F  +   ++R D+ AYL  AT +
P00090    57  VWTQENIIAYLPDPNAYLKKFLTDKGQADKATGSTKMTF-KLANDQQRKDVAAYL--ATLK  114
```

(b)

```
P13569  1221    EGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLL-----NTEGEIQIDGVS 1273
                +    ++  +S ++  G+ + L+G +GSGKS   +A L +L      T GEI  DG
P33593    13    QAAQPLVHGVSLTLQRGRVLALVGGSGSGKSLTCAATLGILPAGVRQTAGEILADGKP   70

P13569  1274    WDSITL---------QQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEV 1322
                    L         Q  R AF  +              +  + + +   + K AD+
P33593    71    VSPCALRGIKIATIMQNPRSAFNPL----------HTMHTHARETCLALGKPADDA   116

P13569  1323    GLRSVIEQFP-GKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPV 1379
                 L + IE         VL      +S G  Q M +A +VL ++  ++ DEP+  LD V
P33593   117    TLTAAIEAVGLENAARVLKLYPFEMSGGMLQRMMIAMAVLCESPFIIADEPTTDLDVV   174
```