Sequence Based Function Annotation

Qi Sun

Bioinformatics Facility

Biotechnology Resource Center

Cornell University

Sequence Based Function Annotation

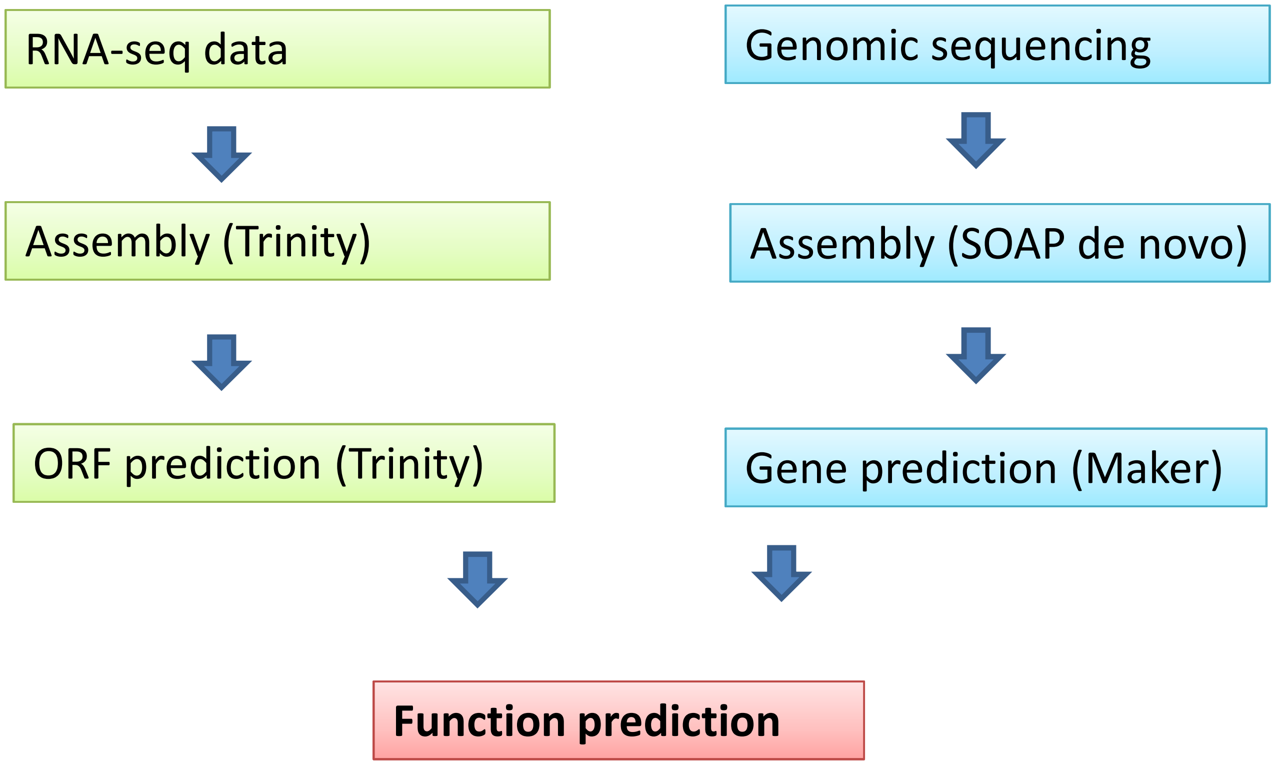
1. Given a sequence, how to predict its biological function?

2. How to describe the function of a gene?

3. How to works with 50,000 genes?

Usage scenarios for sequence based function annotation

Genomic scale function annotation for non-model organisms



RNA-seq data

Assembly (Trinity)

ORF prediction (Trinity)

Genomic sequencing

Assembly (SOAP de novo)

Gene prediction (Maker)

Function prediction

1. Given a protein sequence, how to predict its function?

```
>unknow_protein_1
MVHLTDAEKAAVSCLWGKVNSDEVGGEALGRLLVVYPWTQR
YFDSFGDLSSASAIMGNAKVKAHGKKVITAFNDGLNHLDSL
KGTFASLSELHCDKLHVDPENFRLLGNMIVIVLGHHLGKDF
TPAAQAAFQKVVAGVATALAHKYH
```

Common approaches

- **Identify the homologous gene in a different species with good function annotation; (BLAST, et al.)**

- **Identify conserved motif; (PFAM, InterProScan, et al.)**

Alternative approaches

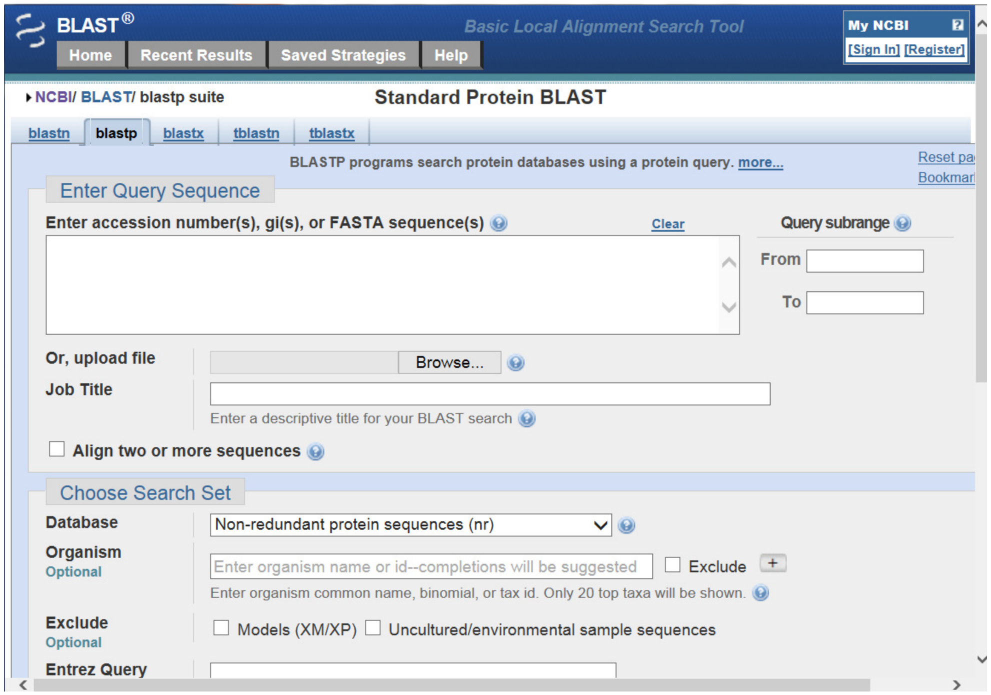
- **Protein 3D structure prediction (threading methods)**

- **Co-expression network modules; (Genevestigator)**

- **Linkage or association mapping;**
- **…**

NCBI BLAST

- How does BLAST work?
- BLAST and Psi-BLAST: Position independent and position specific scoring matrix.



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

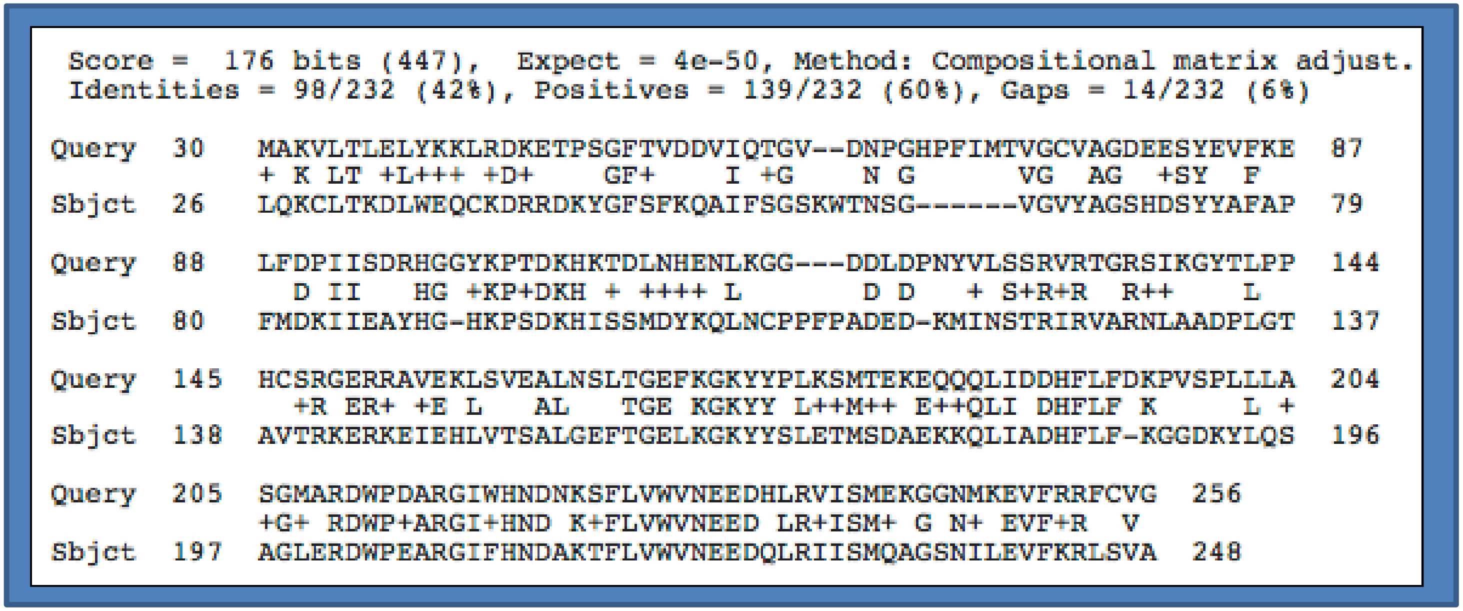
Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

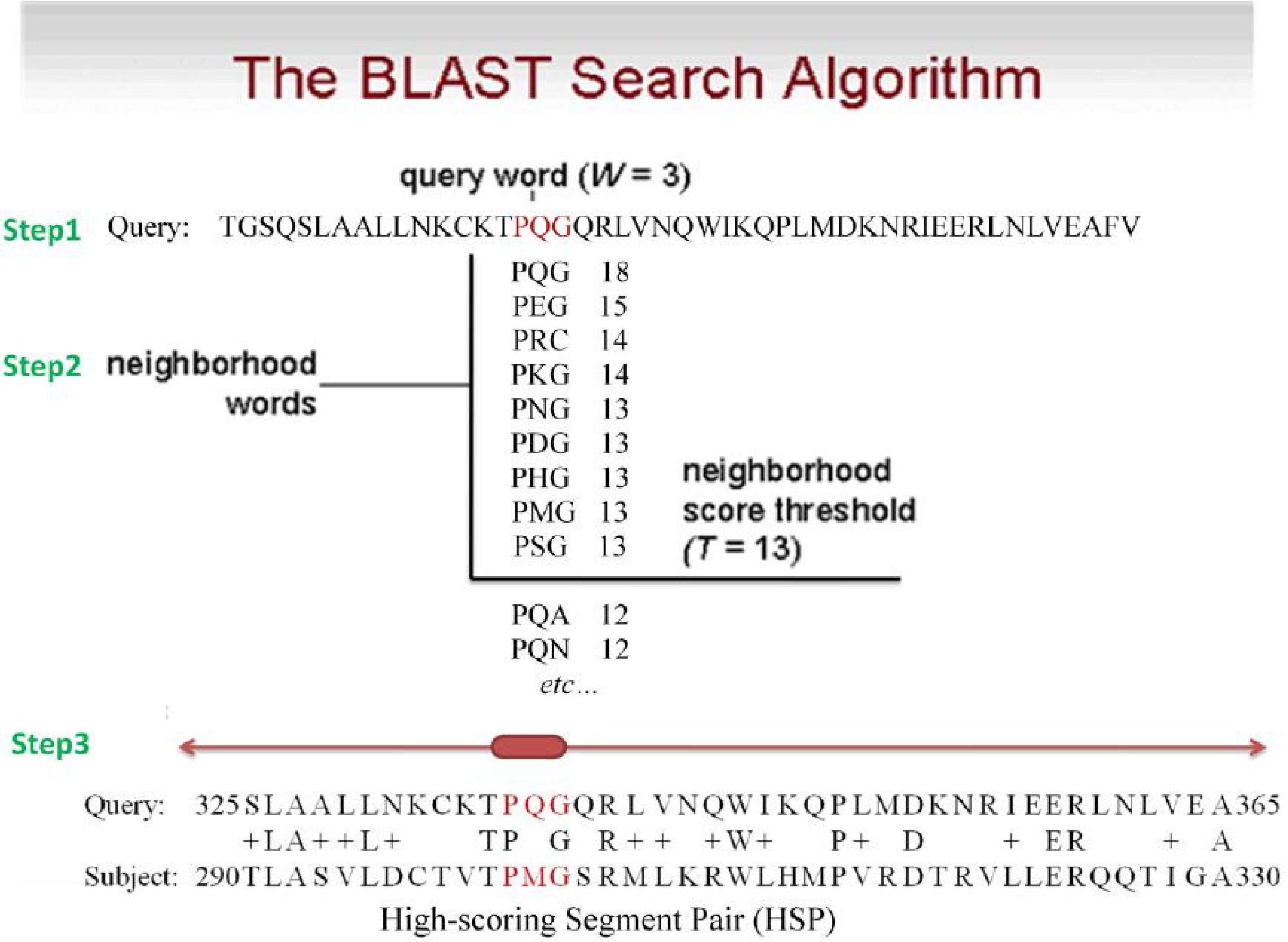
Entrez Query



```
 Score =  176 bits (447),  Expect = 4e-50, Method: Compositional matrix adjust.
 Identities = 98/232 (42%), Positives = 139/232 (60%), Gaps = 14/232 (6%)

Query  30    MAKVLTLELYKKLRDKETPSGFTVDDVIQTGV--DNPGHPFIMTVGCVAGDEESYEVFKE  87
             + K LT +L+++ +D+    GF+    I +G    N G      VG  AG  +SY  F
Sbjct  26    LQKCLTKDLWEQCKDRRDKYGFSFKQAIFSGSKWTNSG------VGVYAGSHDSYYAFAP  79

Query  88    LFDPIISDRHGGYKPTDKHKTDLNHENLKGG---DDLDPNYVLSSRVRTGRSIKGYTLPP  144
               D II   HG +KP+DKH + ++++ L       D D   + S+R+R  R++    L
Sbjct  80    FMDKIIEAYHG-HKPSDKHISSMDYKQLNCPPFPADED-KMINSTRIRVARNLAADPLGT  137

Query  145   HCSRGERRAVEKLSVEALNSLTGEFKGKYYPLKSMTEKEQQQLIDDHFLFDKPVSPLLLA  204
               +R ER+ +E L   AL   TGE KGKYY L++M++ E++QLI DHFLF K     L +
Sbjct  138   AVTRKERKEIEHLVTSALGEFTGELKGKYYSLETMSDAEKKQLIADHFLF-KGGDKYLQS  196

Query  205   SGMARDWPDARGIWHNDNKSFLVWVNEEDHLRVISMEKGGNMKEVFRRFCVG  256
             +G+ RDWP+ARGI+HND K+FLVWVNEED LR+ISM+ G N+ EVF+R  V
Sbjct  197   AGLERDWPEARGIFHNDAKTFLVWVNEEDQLRIISMQAGSNILEVFKRLSVA  248
```

How does BLAST work

Step 1. Create alignments between HSPs (High-scoring Segment Pair)



The BLAST Search Algorithm

query word (*W* = 3)

Step1 Query: TGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFV

Step2 neighborhood words

PQG	18
PEG	15
PRC	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12

etc…

neighborhood score threshold (*T* = 13)

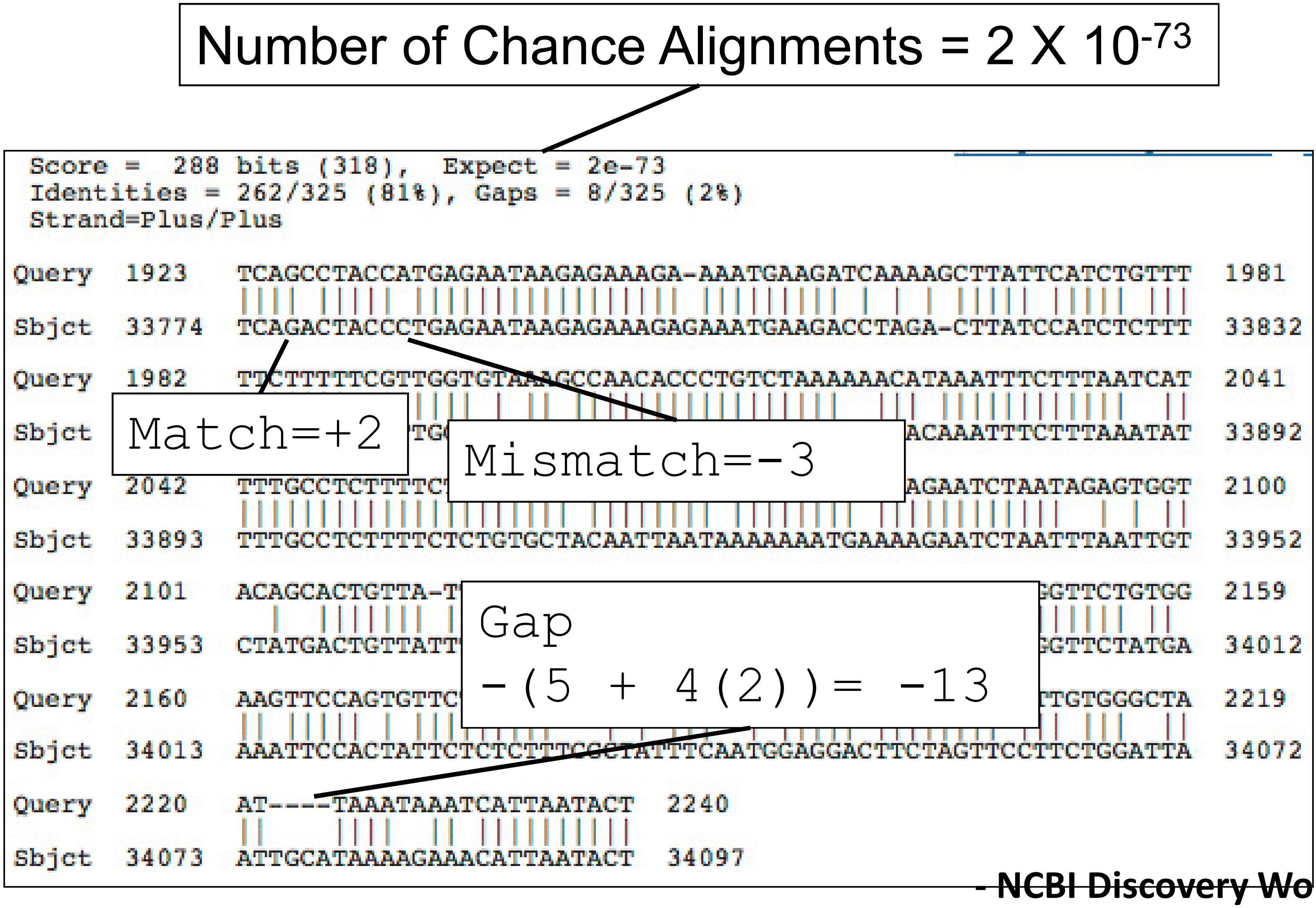
Step3

```
Query:   325SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Subject: 290TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA330
```

High-scoring Segment Pair (HSP)

How does BLAST work

Step 2. Score each alignment, and report the top alignments



Number of Chance Alignments = 2 X 10^{-73}

```
 Score =  288 bits (318),  Expect = 2e-73
 Identities = 262/325 (81%), Gaps = 8/325 (2%)
 Strand=Plus/Plus

Query  1923   TCAGCCTACCATGAGAATAAGAGAAAGA-AAATGAAGATCAAAAGCTTATTCATCTGTTT  1981
              |||| ||||| ||||||||||||||||| ||||||||| | | | ||||| ||||| |||
Sbjct  33774  TCAGACTACCCTGAGAATAAGAGAAAGAGAAATGAAGACCTAGA-CTTATCCATCTCTTT  33832

Query  1982   TTCTTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTAATCAT  2041
                                                                        
Sbjct             TG                        ACAAATTTCTTTAAATAT  33892

Query  2042   TTTGCCTCTTTTC                         AGAATCTAATAGAGTGGT  2100
              
Sbjct  33893  TTTGCCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT  33952

Query  2101   ACAGCACTGTTA-T                                  GGTTCTGTGG  2159
              
Sbjct  33953  CTATGACTGTTATT                                  GGTTCTATGA  34012

Query  2160   AAGTTCCAGTGTTC                                  TTGTGGGCTA  2219
              
Sbjct  34013  AAATTCCACTATTCTCTCTTTCGCTATTTCAATGGAGGACTTCTAGTTCCTTCTGGATTA  34072

Query  2220   AT----TAAATAAATCATTAATACT  2240
              ||    ||||  ||  ||||||||||
Sbjct  34073  ATTGCATAAAAGAAACATTAATACT  34097
```

Match=+2

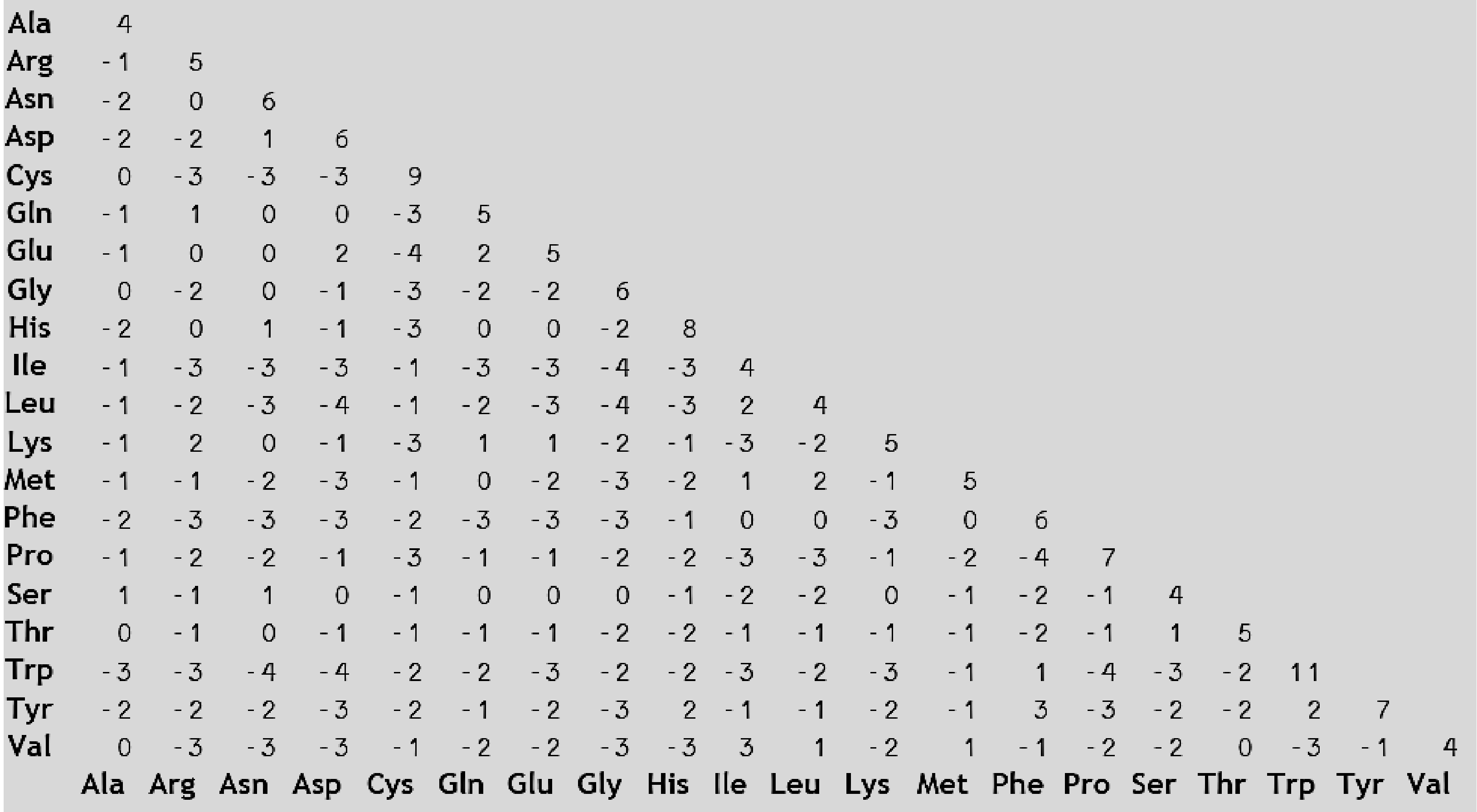
Mismatch=-3

Gap

-(5 + 4(2))= -13

- NCBI Discovery Workshops

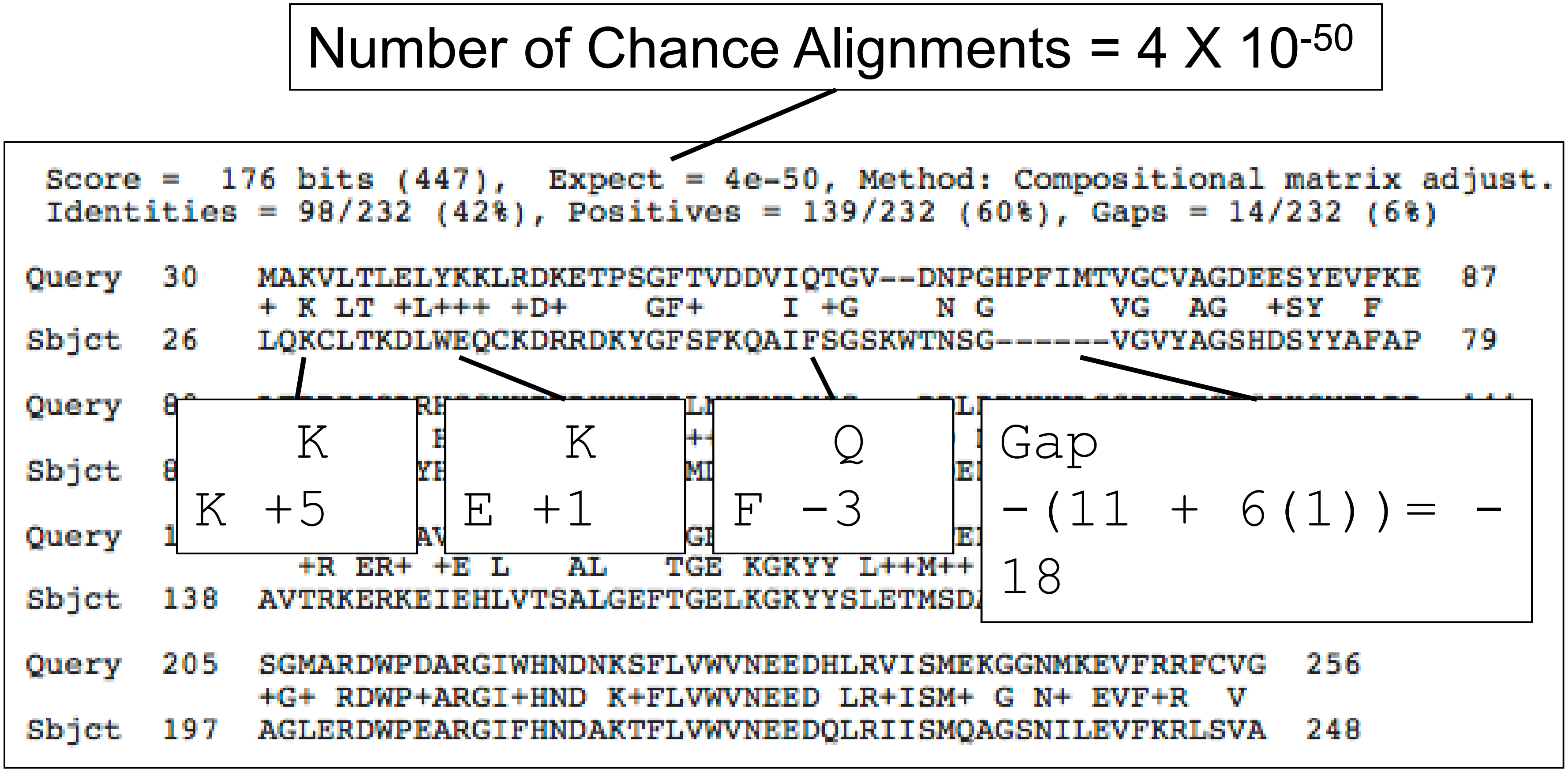
BLOSUM62, a position independent matrix



Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

How does BLAST work

Step 2. Score each alignment – protein alignment



Number of Chance Alignments = 4 X 10^{-50}

```
 Score =  176 bits (447),  Expect = 4e-50, Method: Compositional matrix adjust.
 Identities = 98/232 (42%), Positives = 139/232 (60%), Gaps = 14/232 (6%)

Query  30    MAKVLTLELYKKLRDKETPSGFTVDDVIQTGV--DNPGHPFIMTVGCVAGDEESYEVFKE  87
             + K LT +L+++ +D+     GF+     I +G     N G        VG   AG   +SY   F
Sbjct  26    LQKCLTKDLWEQCKDRRDKYGFSFKQAIFSGSKWTNSG------VGVYAGSHDSYYAFAP  79

Query  8                   RH               L                 L
                           E                ++                 
Sbjct  8                   YH              MI                 E

Query  1                   AV              GI                 E
                +R ER+ +E L    AL    TGE KGKYY L++M++
Sbjct  138   AVTRKERKEIEHLVTSALGEFTGELKGKYYSLETMSD

Query  205   SGMARDWPDARGIWHNDNKSFLVWVNEEDHLRVISMEKGGNMKEVFRRFCVG  256
             +G+ RDWP+ARGI+HND K+FLVWVNEED LR+ISM+ G N+ EVF+R   V
Sbjct  197   AGLERDWPEARGIFHNDAKTFLVWVNEEDQLRIISMQAGSNILEVFKRLSVA  248
```

K
K +5

K
E +1

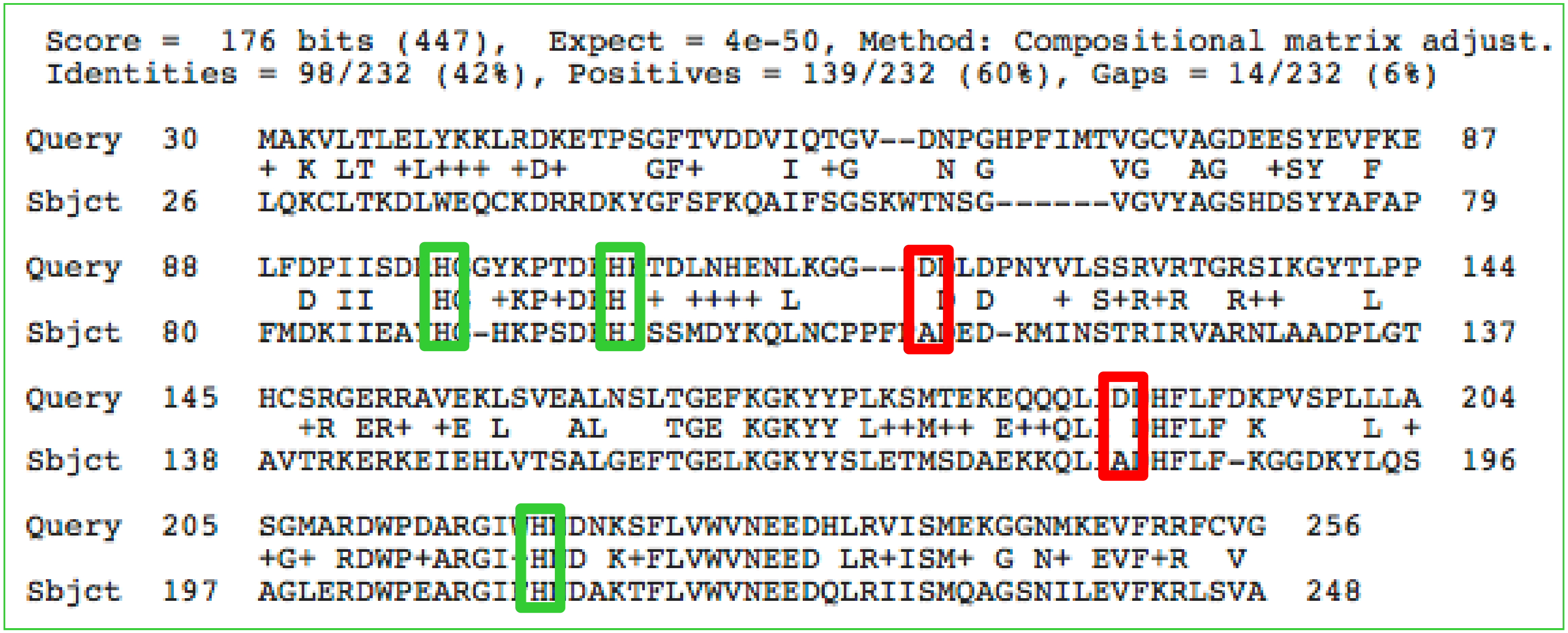
Q
F -3

Gap
-(11 + 6(1))= -18

Scores from BLOSUM62, a position independent matrix

- NCBI Discovery Workshops

BLOSUM62 substitution score is position independent

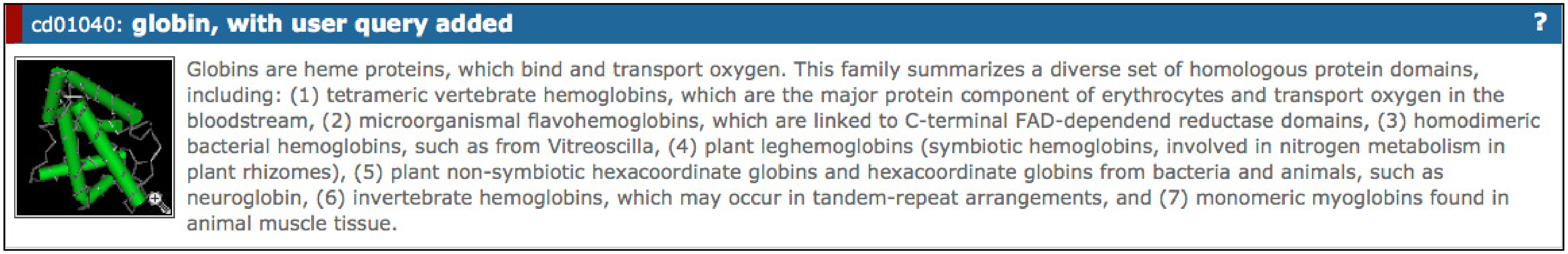


```
 Score =  176 bits (447),  Expect = 4e-50, Method: Compositional matrix adjust.
 Identities = 98/232 (42%), Positives = 139/232 (60%), Gaps = 14/232 (6%)

Query  30    MAKVLTLELYKKLRDKETPSGFTVDDVIQTGV--DNPGHPFIMTVGCVAGDEESYEVFKE  87
             + K LT +L+++ +D+    GF+    I +G    N G      VG  AG  +SY  F
Sbjct  26    LQKCLTKDLWEQCKDRRDKYGFSFKQAIFSGSKWTNSG------VGVYAGSHDSYYAFAP  79

Query  88    LFDPIISDIHGGYKPTDIHITDLNHENLKGG---DILDPNYVLSSRVRTGRSIKGYTLPP  144
               D II   HG +KP+DIH + ++++ L       I D   + S+R+R  R++    L
Sbjct  80    FMDKIIEAIHG-HKPSDIHISSMDYKQLNCPPFIAIED-KMINSTRIRVARNLAADPLGT  137

Query  145   HCSRGERRAVEKLSVEALNSLTGEFKGKYYPLKSMTEKEQQQLIDIHFLFDKPVSPLLLA  204
               +R ER+ +E L   AL   TGE KGKYY L++M++ E++QLI IHFLF K     L +
Sbjct  138   AVTRKERKEIEHLVTSALGEFTGELKGKYYSLETMSDAEKKQLIAIHFLF-KGGDKYLQS  196

Query  205   SGMARDWPDARGIVHIDNKSFLVWVNEEDHLRVISMEKGGNMKEVFRRFCVG  256
             +G+ RDWP+ARGI+HID K+FLVWVNEED LR+ISM+ G N+ EVF+R  V
Sbjct  197   AGLERDWPEARGIIHIDAKTFLVWVNEEDQLRIISMQAGSNILEVFKRLSVA  248
```

Scores from BLOSUM62, a position independent matrix

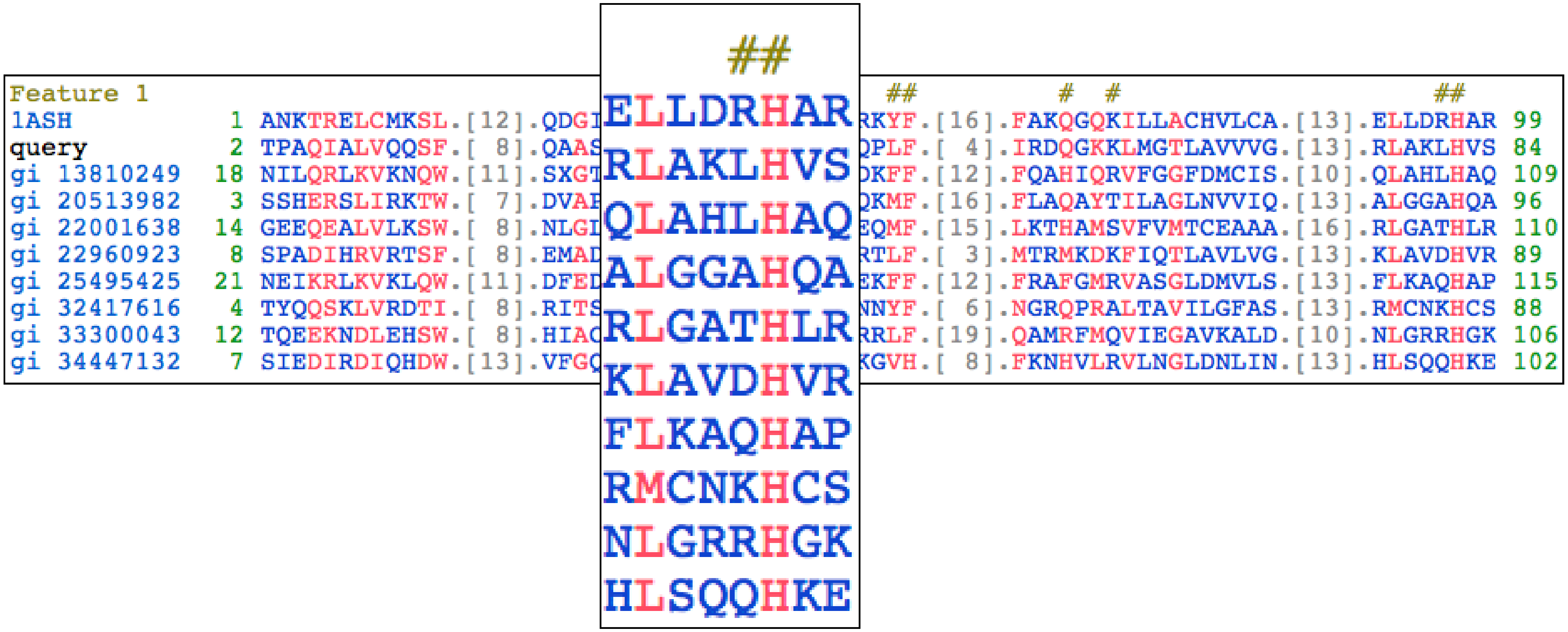
- NCBI Discovery Workshops

PSSM Alignment: Globins



cd01040: **globin, with user query added** ?

Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependend reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue.



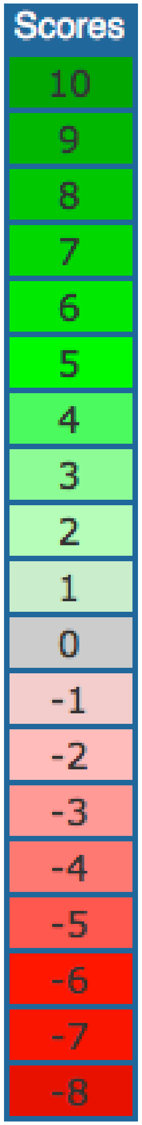
```
Feature 1                                                    ##          #   #                          ##
1ASH          1 ANKTRELCMKSL.[12].QDGI      RKYF.[16].FAKQGQKILLACHVLCA.[13].ELLDRHAR  99
query         2 TPAQIALVQQSF.[ 8].QAAS      QPLF.[ 4].IRDQGKKLMGTLAVVVG.[13].RLAKLHVS  84
gi 13810249  18 NILQRLKVKNQW.[11].SXGT      DKFF.[12].FQAHIQRVFGGFDMCIS.[10].QLAHLHAQ 109
gi 20513982   3 SSHERSLIRKTW.[ 7].DVAF      QKMF.[16].FLAQAYTILAGLNVVIQ.[13].ALGGAHQA  96
gi 22001638  14 GEEQEALVLKSW.[ 8].NLGI      EQMF.[15].LKTHAMSVFVMTCEAAA.[16].RLGATHLR 110
gi 22960923   8 SPADIHRVRTSF.[ 8].EMAD      RTLF.[ 3].MTRMKDKFIQTLAVLVG.[13].KLAVDHVR  89
gi 25495425  21 NEIKRLKVKLQW.[11].DFED      EKFF.[12].FRAFGMRVASGLDMVLS.[13].FLKAQHAP 115
gi 32417616   4 TYQQSKLVRDTI.[ 8].RITS      NNYF.[ 6].NGRQPRALTAVILGFAS.[13].RMCNKHCS  88
gi 33300043  12 TQEEKNDLEHSW.[ 8].HIAC      RRLF.[19].QAMRFMQVIEGAVKALD.[10].NLGRRHGK 106
gi 34447132   7 SIEDIRDIQHDW.[13].VFGQ      KGVH.[ 8].FKNHVLRVLNGLDNLIN.[13].HLSQQHKE 102
```

```
   ##
ELLDRHAR
RLAKLHVS
QLAHLHAQ
ALGGAHQA
RLGATHLR
KLAVDHVR
FLKAQHAP
RMCNKHCS
NLGRRHGK
HLSQQHKE
```

Conserved Histidine

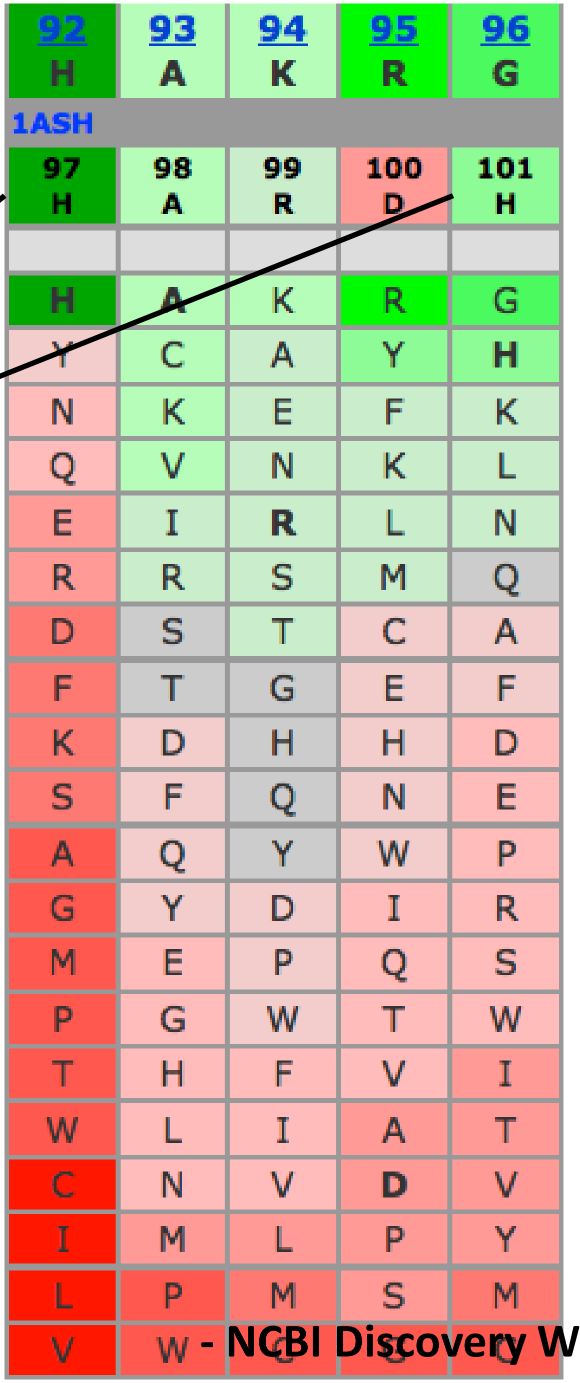
- NCBI Discovery Workshops

PSSM Viewer



Scores
10
9
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8

Histidine scored differently at two positions



92 H	93 A	94 K	95 R	96 G
1ASH				
97 H	98 A	99 R	100 D	101 H
H	A	K	R	G
Y	C	A	Y	H
N	K	E	F	K
Q	V	N	K	L
E	I	R	L	N
R	R	S	M	Q
D	S	T	C	A
F	T	G	E	F
K	D	H	H	D
S	F	Q	N	E
A	Q	Y	W	P
G	Y	D	I	R
M	E	P	Q	S
P	G	W	T	W
T	H	F	V	I
W	L	I	A	T
C	N	V	D	V
I	M	L	P	Y
L	P	M	S	M
V	W	C	G	C

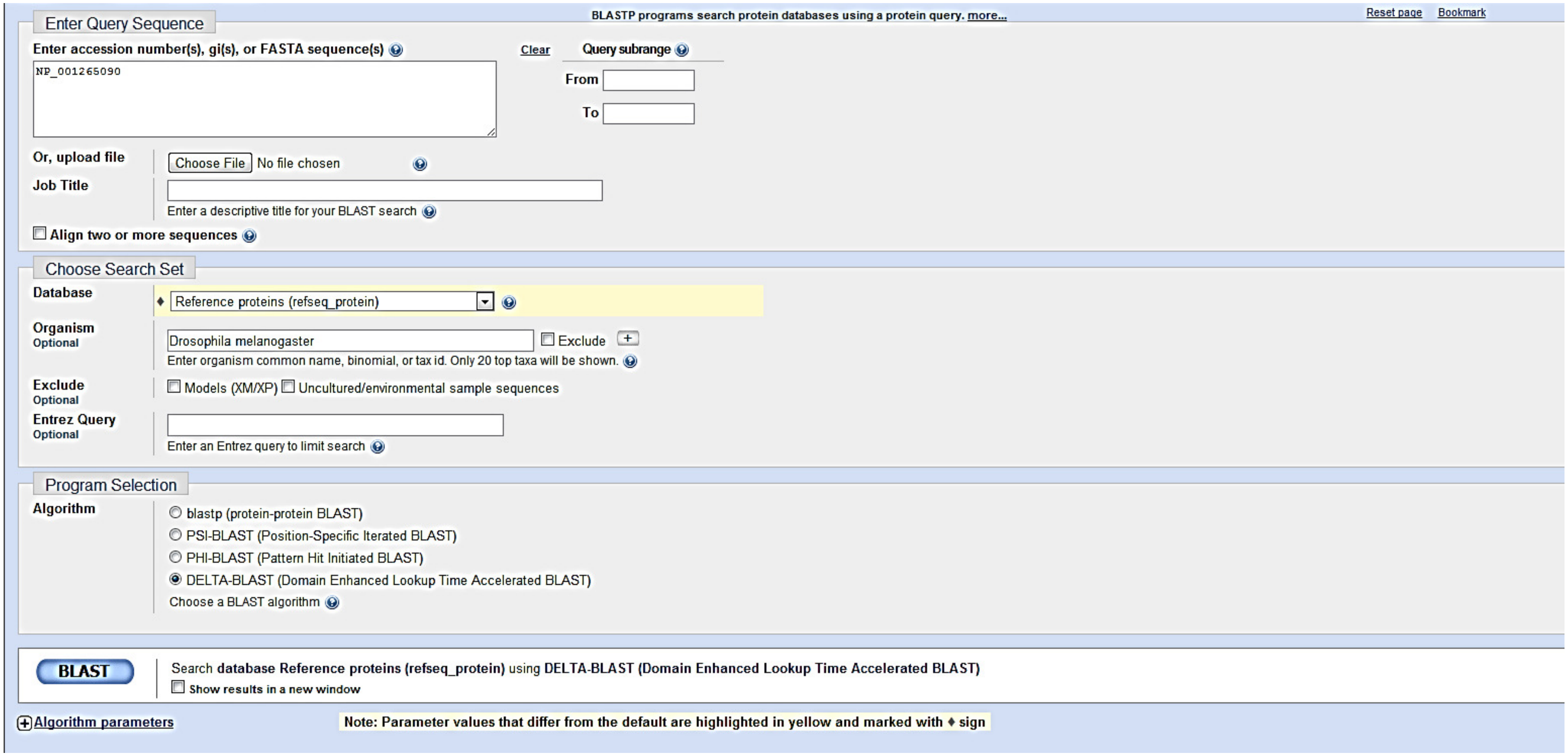
Build PSSM with PSI-BLAST

- PSI-BLAST
  1. Iteration 1: Regular BLASTP (BLOSSOM62) to identify a list of closely related proteins. Build PSSM from these proteins.
  2. Iteration 2: Use the PSSM built from Iteration 1 to score alignment in this Iteration.
  3. Repeat multiple iterations.

- NCBI Discovery Workshops

Build PSSM with DELTA-BLAST

DELTA-BLAST employs a subset of NCBI's Conserved Domain Database (CDD) to construct PSSM



BLASTP programs search protein databases using a protein query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): NP_001265090

Clear

Query subrange
From:
To:

Or, upload file: Choose File No file chosen

Job Title:
Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ♦ Reference proteins (refseq_protein)

Organism (Optional): Drosophila melanogaster □ Exclude +
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional):
Enter an Entrez query to limit search

Program Selection

Algorithm:
○ blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
● DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

BLAST

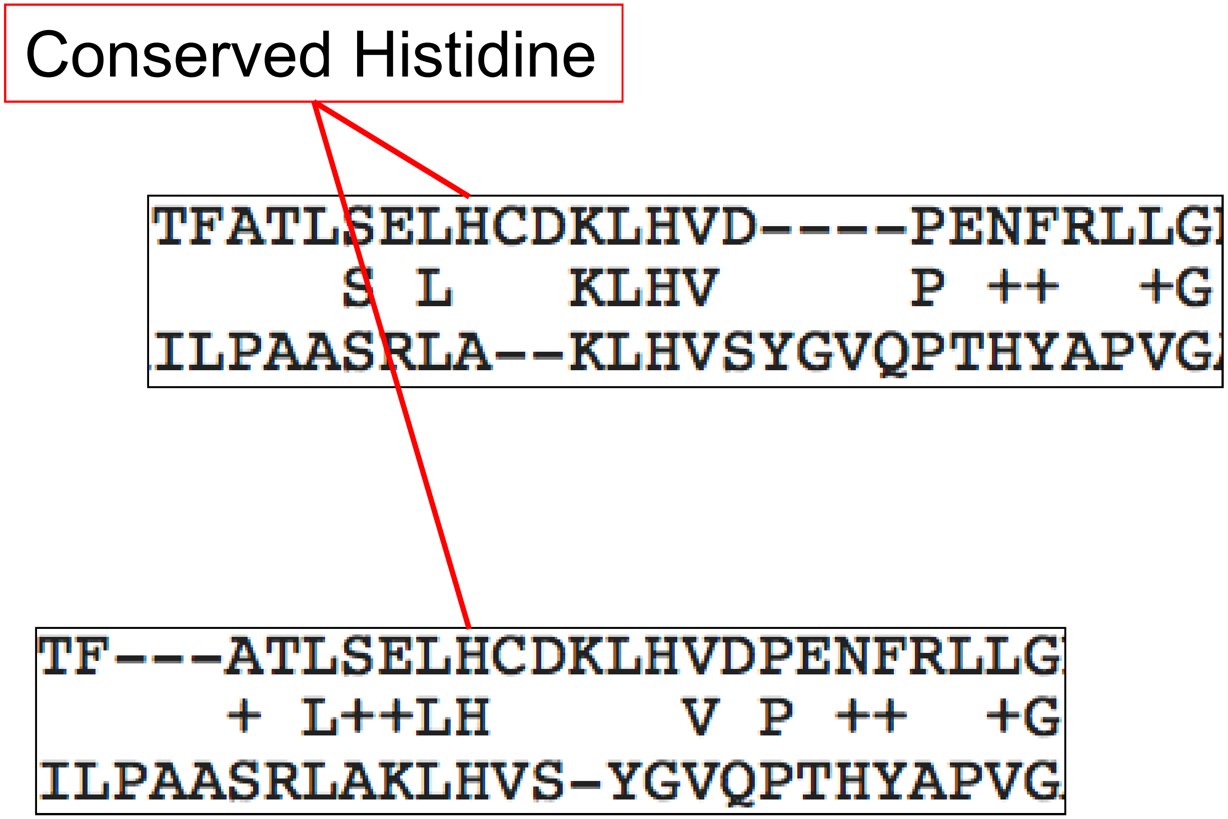
Search database Reference proteins (refseq_protein) using DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

□ Show results in a new window

+ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Heme Binding Site



Conserved Histidine

blastp

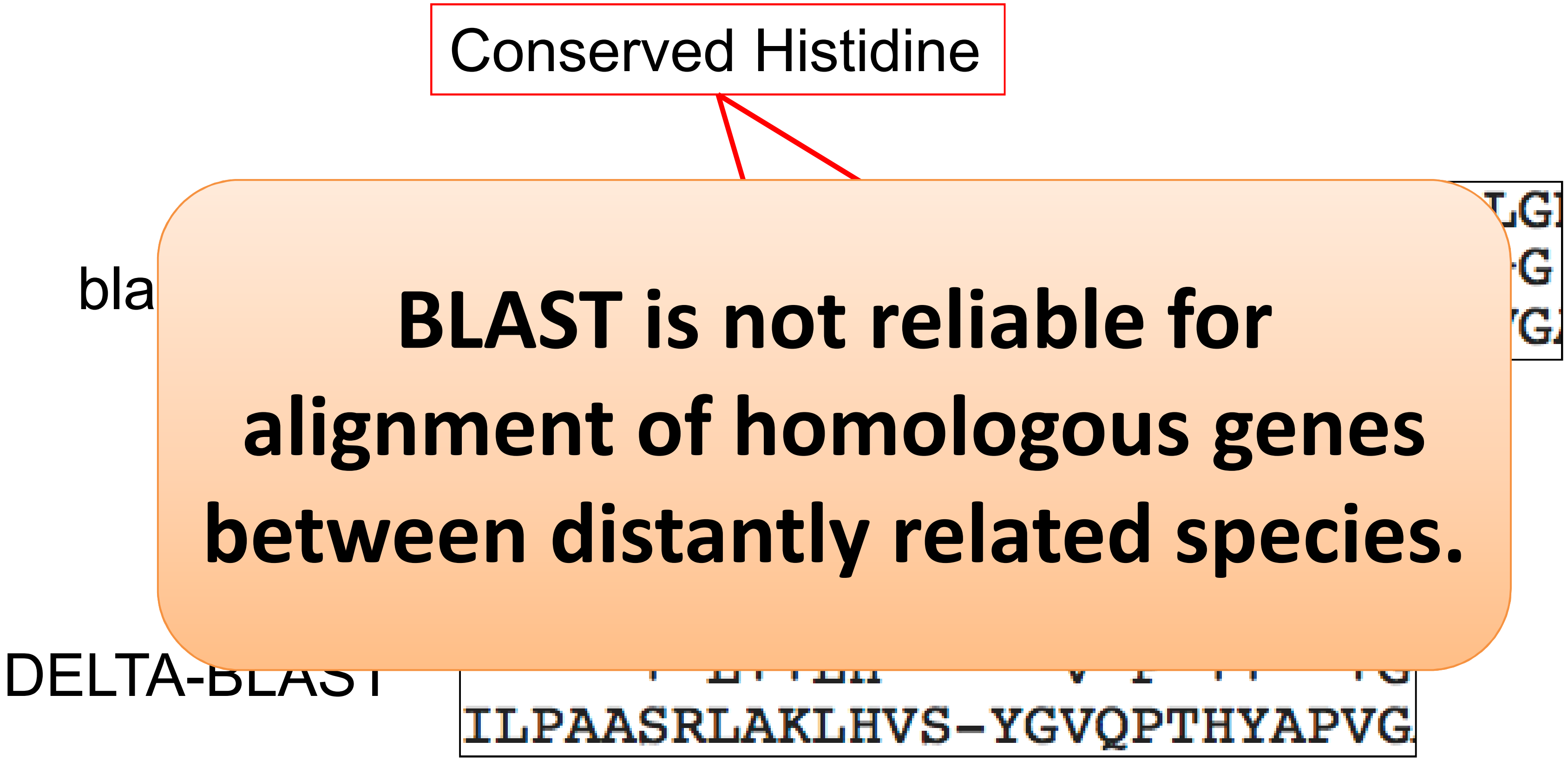
```
TFATLSELHCDKLHVD----PENFRLLG
     S L   KLHV     P ++  +G
ILPAASRLA--KLHVSYGVQPTHYAPVG
```

DELTA-BLAST

```
TF---ATLSELHCDKLHVDPENFRLLG
     + L++LH     V P ++  +G
ILPAASRLAKLHVS-YGVQPTHYAPVG
```

- NCBI Discovery Workshops

Heme Binding Site



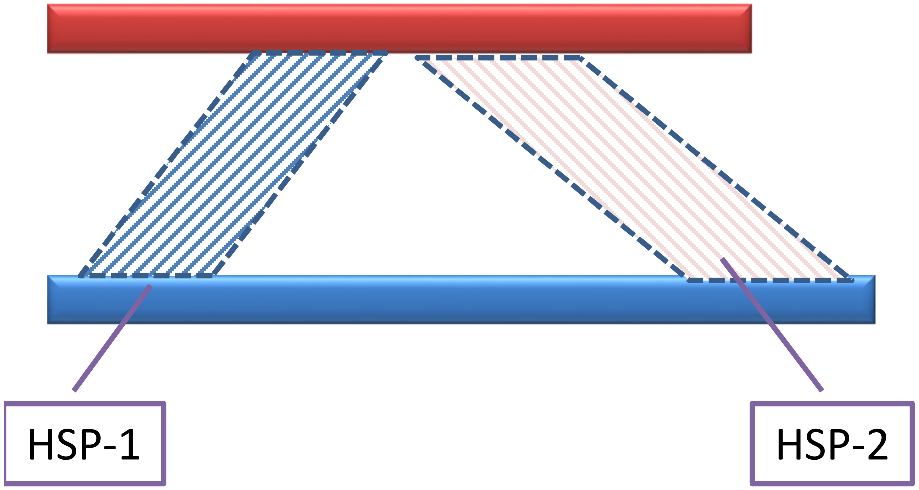
```
ILPAASRLAKLHVS-YGVQPTHYAPVG
```

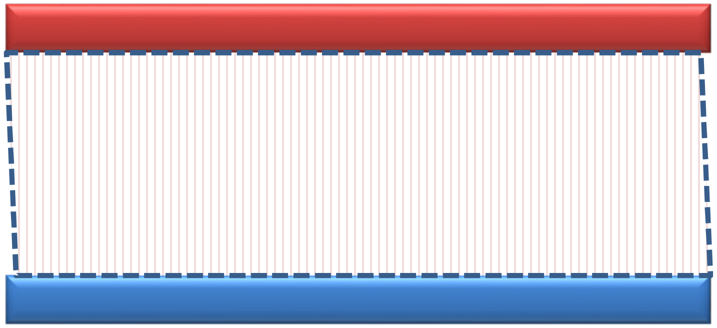
BLAST is not reliable for alignment of homologous genes between distantly related species.

- NCBI Discovery Workshops

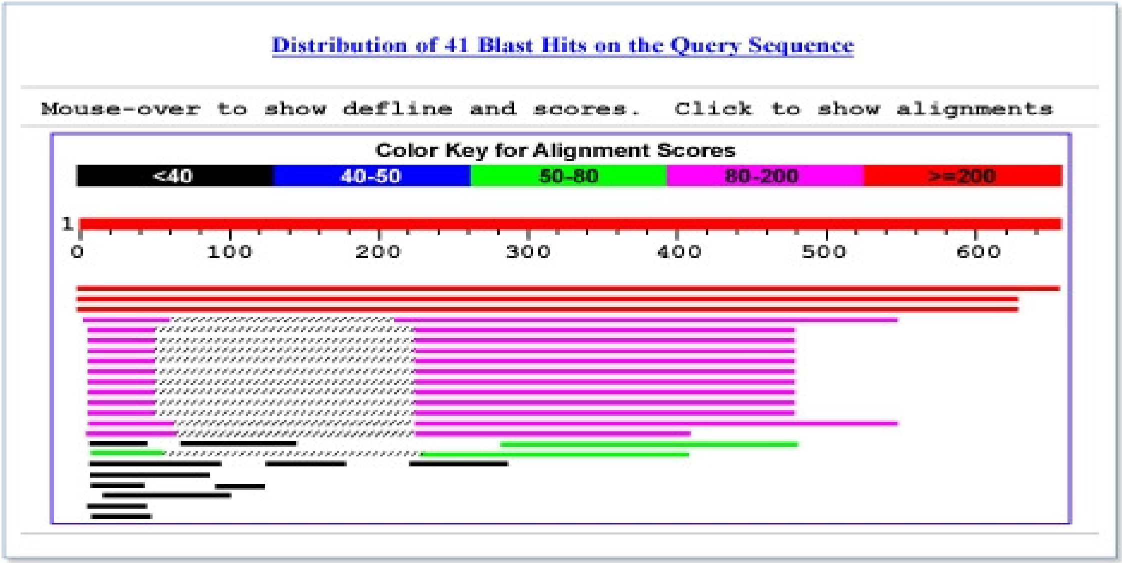
BLAST does Local Alignment
(Basic Local Alignment Search Tool)

Local Alignment vs Global Alignment





(Bowtie, BWA, ClustalW, et al)



Distribution of 41 Blast Hits on the Query Sequence

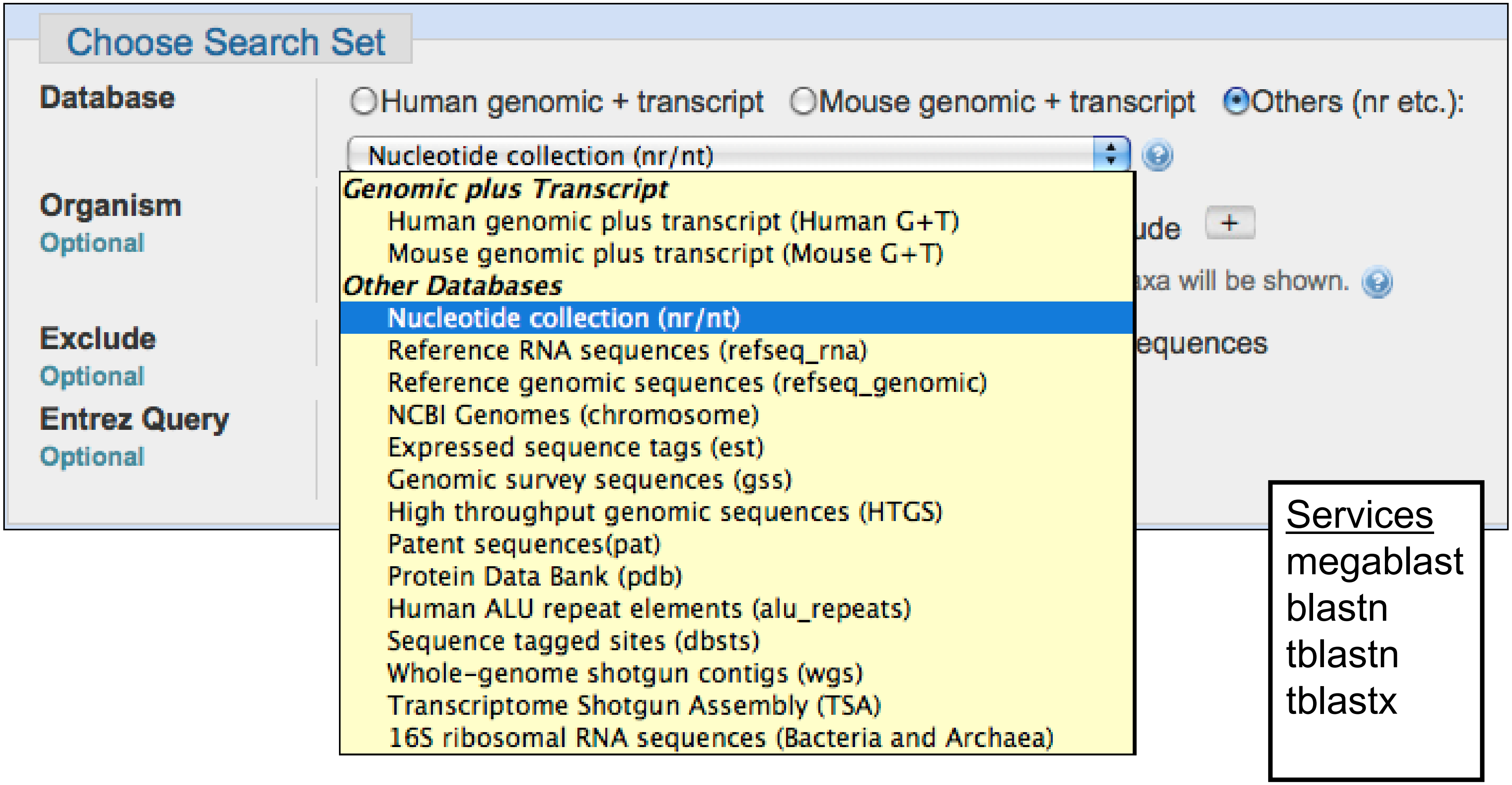
Mouse-over to show defline and scores. Click to show alignments

BLAST and BLAST-like programs

- **Traditional BLAST (formerly blastall) nucleotide, protein, translations**
  - **blastn nucleotide query vs. nucleotide database**
  - **blastp protein query vs. protein database**
  - **blastx nucleotide query vs. protein database**
  - **tblastn protein query vs. translated nucleotide database**
  - **tblastx translated query vs. translated database**
- **Megablast nucleotide only**
  - **Contiguous megablast**
    - **Nearly identical sequences**
  - **Discontiguous megablast**
    - **Cross-species comparison**

- NCBI Discovery Workshops

Nucleotide Databases: List



Choose Search Set

Database

○Human genomic + transcript ○Mouse genomic + transcript ●Others (nr etc.):

Nucleotide collection (nr/nt)

Genomic plus Transcript

- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases

- Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun contigs (wgs)
- Transcriptome Shotgun Assembly (TSA)
- 16S ribosomal RNA sequences (Bacteria and Archaea)

Organism
Optional

ude +

axa will be shown.

Exclude
Optional

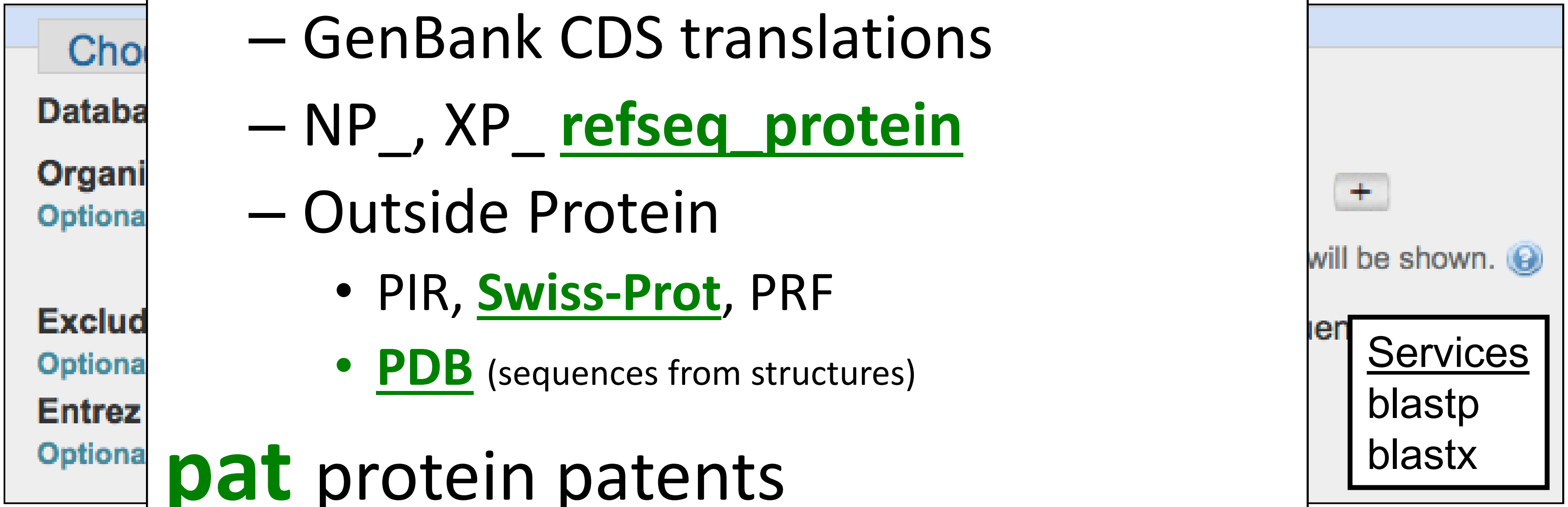
equences

Entrez Query
Optional

Services

megablast
blastn
tblastn
tblastx

Non-redundant protein



Services
blastp
blastx

nr (non-redundant protein sequences)

- GenBank CDS translations
- NP_, XP_ **refseq_protein**
- Outside Protein
  - PIR, **Swiss-Prot**, PRF
  - **PDB** (sequences from structures)

pat protein patents

env_nr metagenomes
(environmental samples)

Reference Sequence Databases

Archive Databases

Genbank & Genpept
(NCBI nt and nr)

Genome Based Reference

NCBI Refseq

UCSC Genomes (Animals)

Ensembl (Animals)

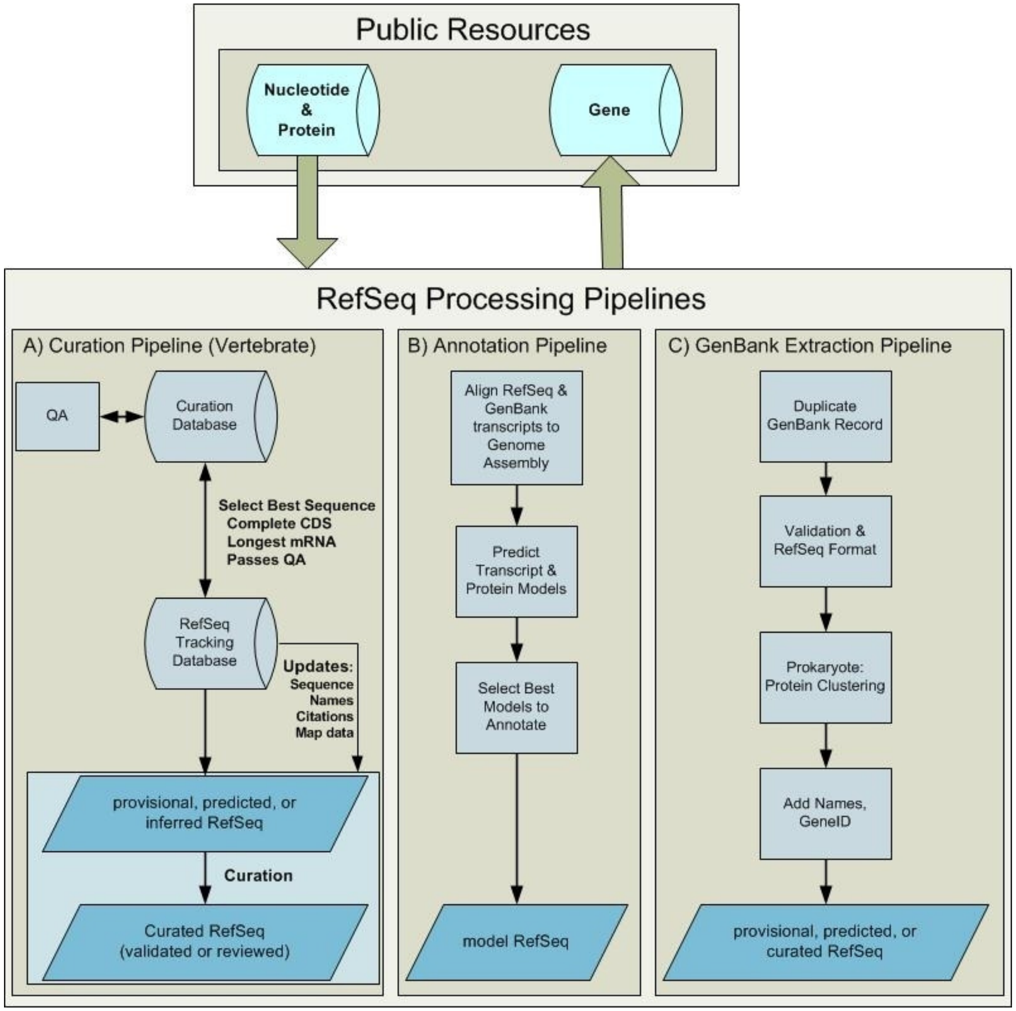
Phytozome (Plants)

Ensembl Plants (Plants)

Species Specific Databases

Flybase
Wormbase
TAIR
et al.

RefSeq Curation Pipeline



Public Resources

Nucleotide & Protein

Gene

RefSeq Processing Pipelines

A) Curation Pipeline (Vertebrate)

QA

Curation Database

Select Best Sequence
Complete CDS
Longest mRNA
Passes QA

RefSeq Tracking Database

Updates:
Sequence
Names
Citations
Map data

provisional, predicted, or inferred RefSeq

Curation

Curated RefSeq (validated or reviewed)

B) Annotation Pipeline

Align RefSeq & GenBank transcripts to Genome Assembly

Predict Transcript & Protein Models

Select Best Models to Annotate

model RefSeq

C) GenBank Extraction Pipeline

Duplicate GenBank Record

Validation & RefSeq Format

Prokaryote: Protein Clustering

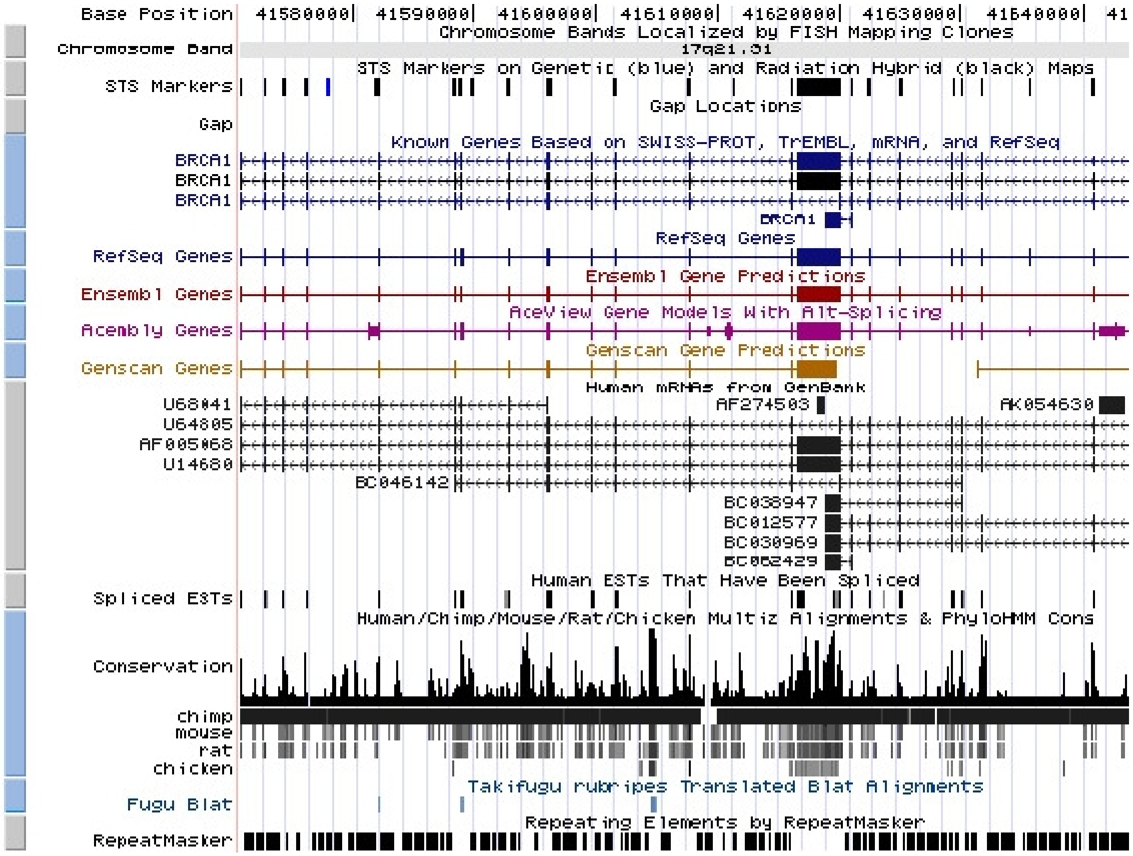
Add Names, GeneID

provisional, predicted, or curated RefSeq

Accessions:

NM_000000	mRNA
NP_000000	Protein
NC_000000	genome assembly
XM_000000	predicted mRNA
XP_000000	predicted protein

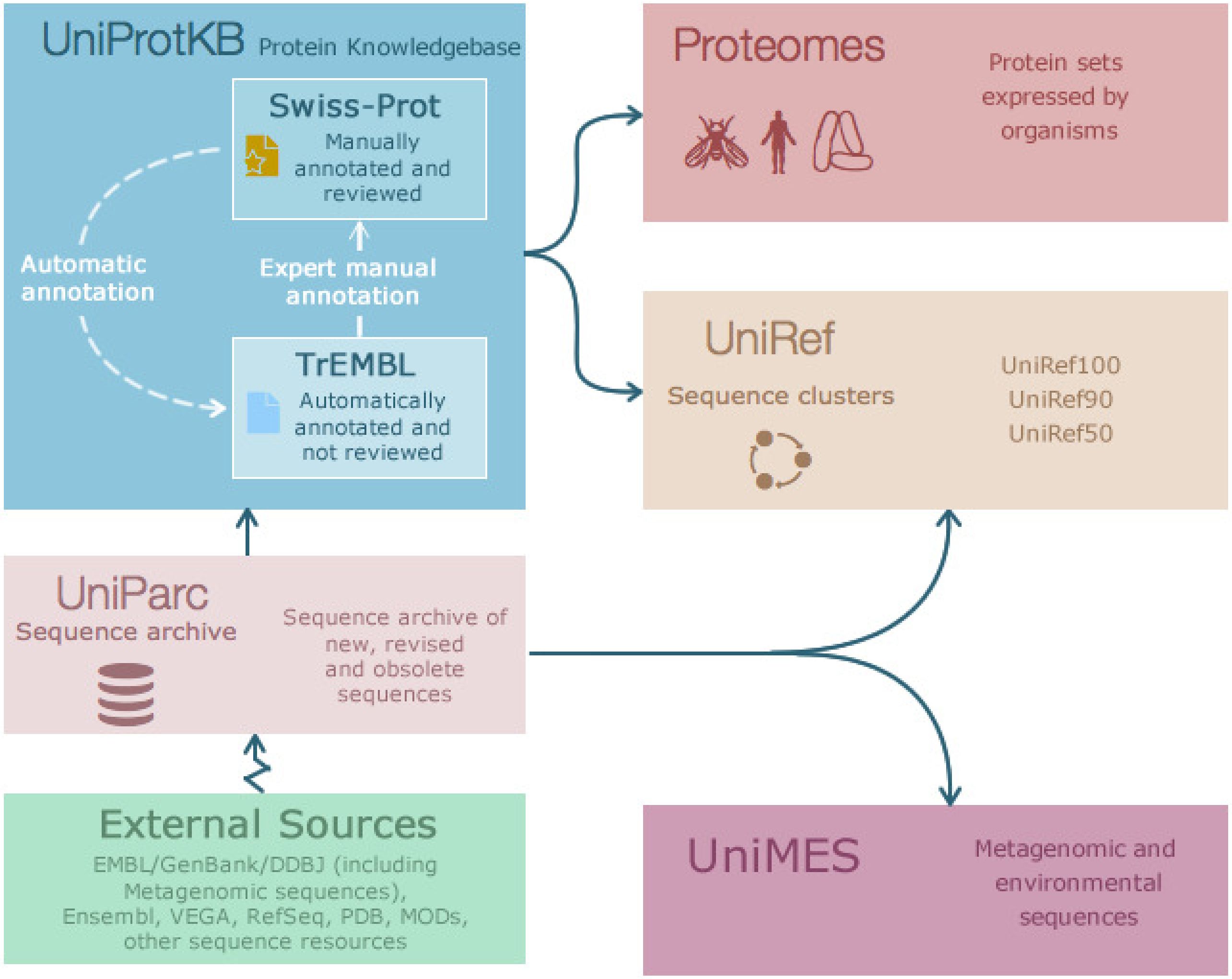
UCSC Genome Database



RefSeq records are mapped to the genome by UCSC.

UniProtKB/Swiss-Prot

UniRef



UniProtKB Protein Knowledgebase

Swiss-Prot
Manually annotated and reviewed

Automatic annotation

Expert manual annotation

TrEMBL
Automatically annotated and not reviewed

Proteomes
Protein sets expressed by organisms

UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50

UniParc
Sequence archive
Sequence archive of new, revised and obsolete sequences

External Sources
EMBL/GenBank/DDBJ (including Metagenomic sequences), Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources

UniMES
Metagenomic and environmental sequences

Versions of NCBI, UCSC and Ensembl

Versions of genome assembly:
NCBI: GRCh38 (December 2013)
UCSC: HG38
Ensembl: GRCh38

Versions of gene annotation:
NCBI: release 69 (January, 2015)
UCSC: Daily incremental updates as NCBI
Ensembl: release 79 (March 2015)

Database usage examples

1. **Identify species for a list of sequences.**

   Tool on BioHPC: fastq_species_detector (using NCBI nt database)
   https://cbsu.tc.cornell.edu/lab/userguide.aspx?a=software&i=149#c

2. **Function annotation**

- BLAST to a closely related species. (Ensembl, Flybase, et al)
- BLAST to swiss-prot (used by BLAST2GO) or Uni-ref (used by Trinotate).