Genome Annotation

Qi Sun

Bioinformatics Facility

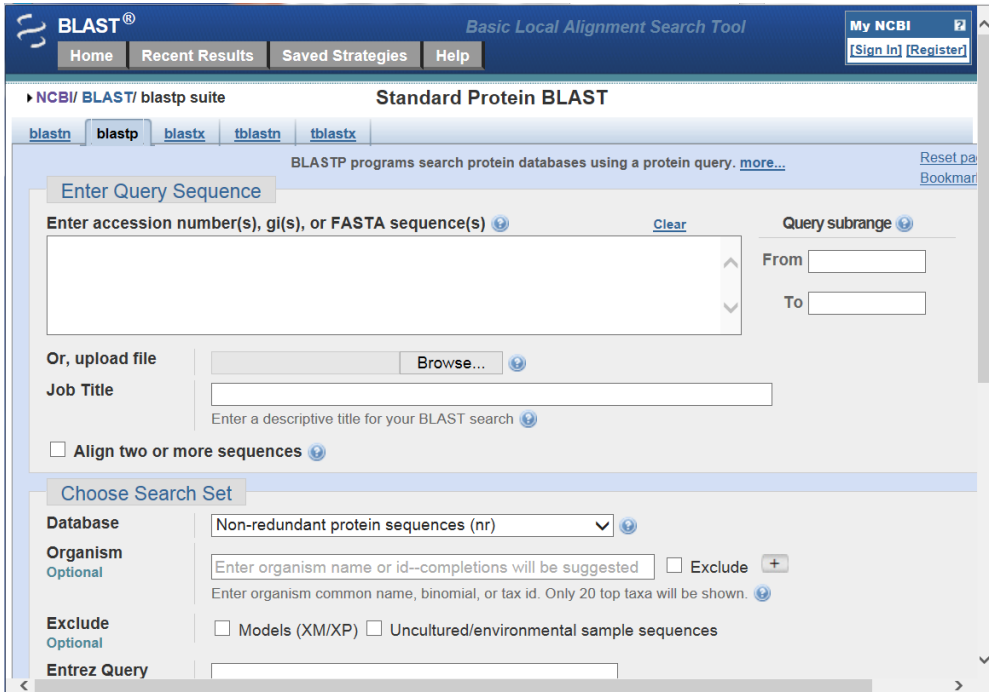
Cornell University

Some basic bioinformatics tools

- BLAST
- PSI-BLAST - Position-Specific Scoring Matrix
- HMM - Hidden Markov Model

NCBI BLAST

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



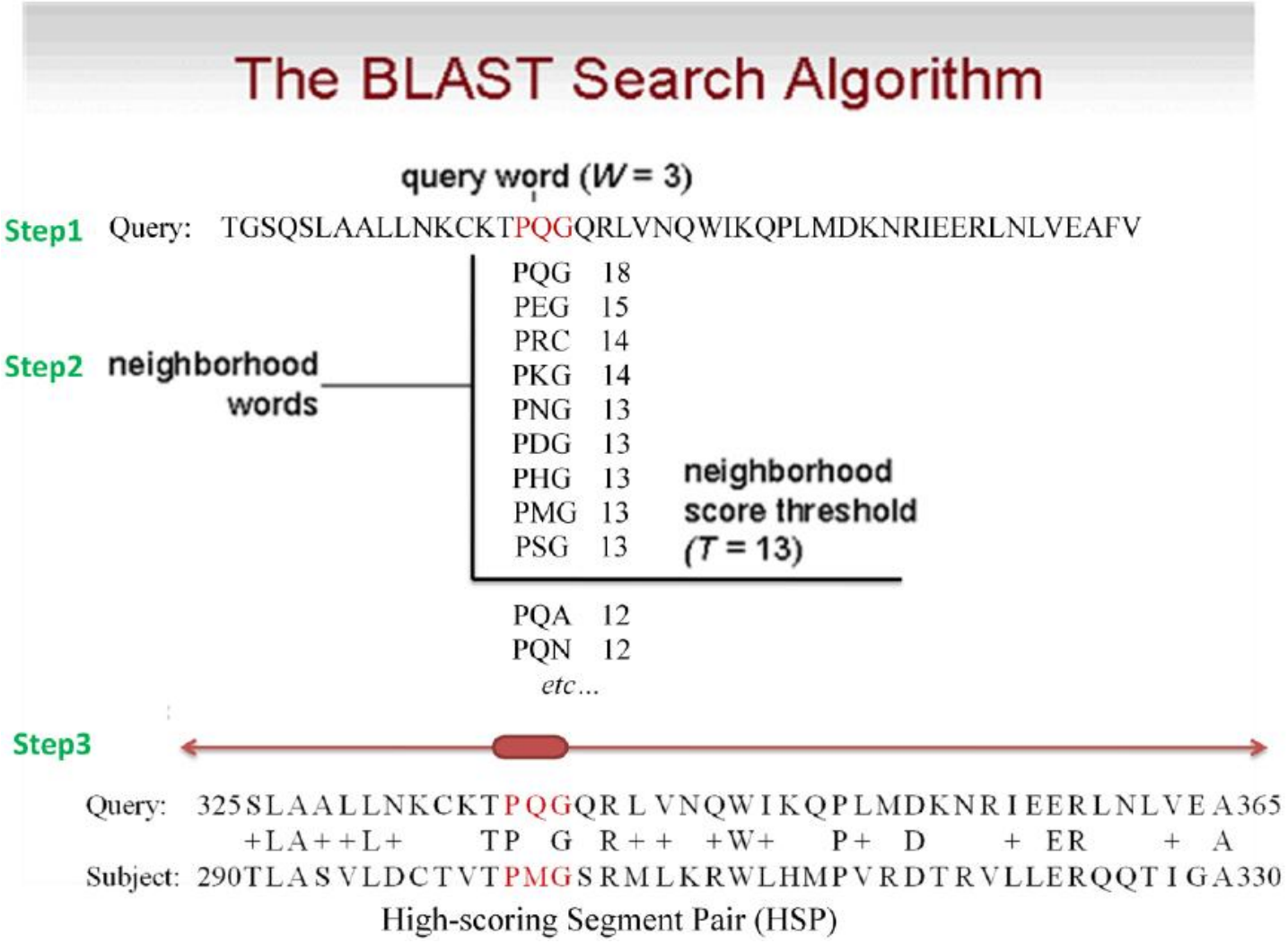

```
 Score =  176 bits (447),  Expect = 4e-50, Method: Compositional matrix adjust.
 Identities = 98/232 (42%), Positives = 139/232 (60%), Gaps = 14/232 (6%)

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
```

BLAST programs

- **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

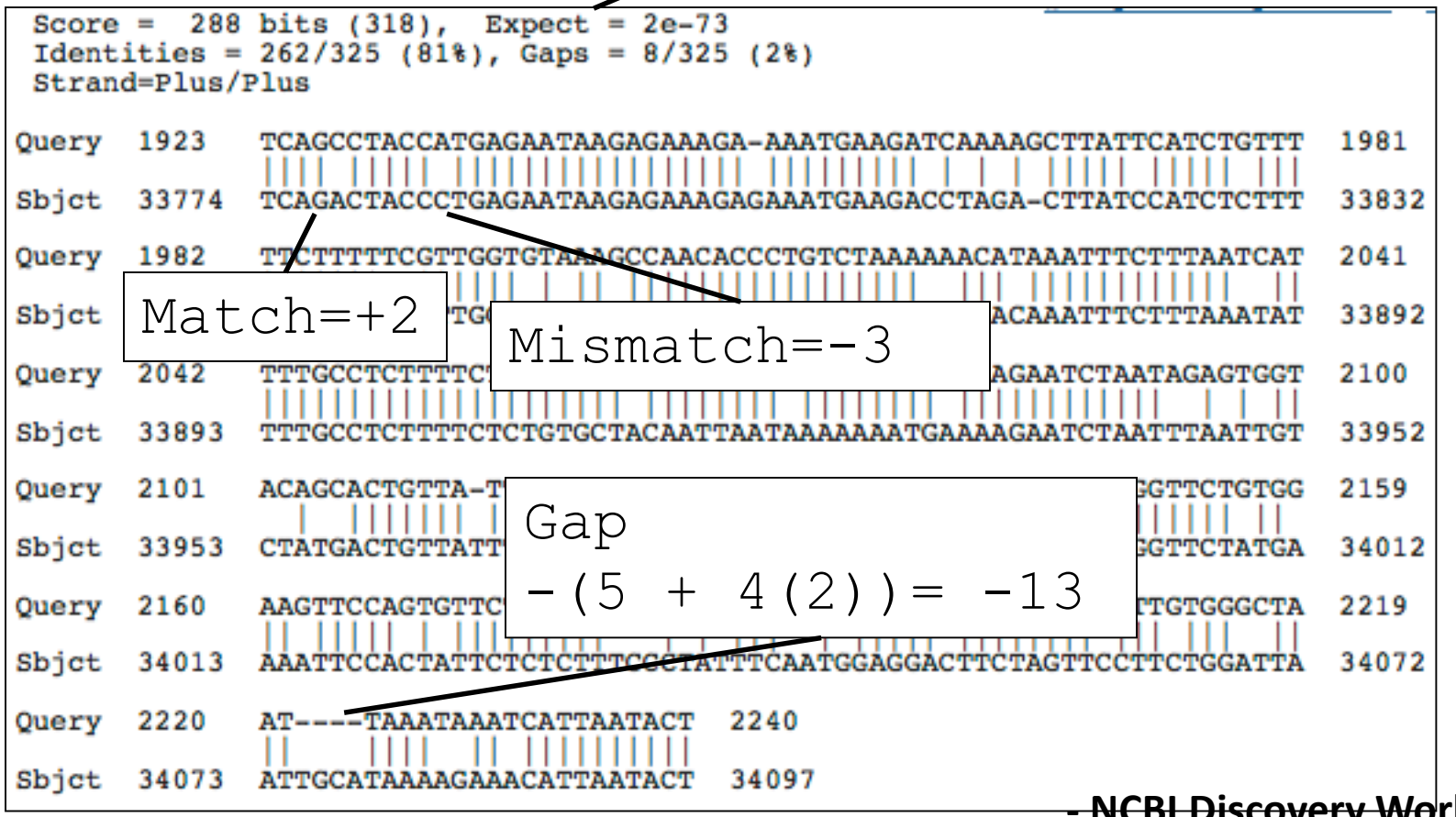
How does BLAST work

Step 1. find alignments



Step 2. scoring alignments

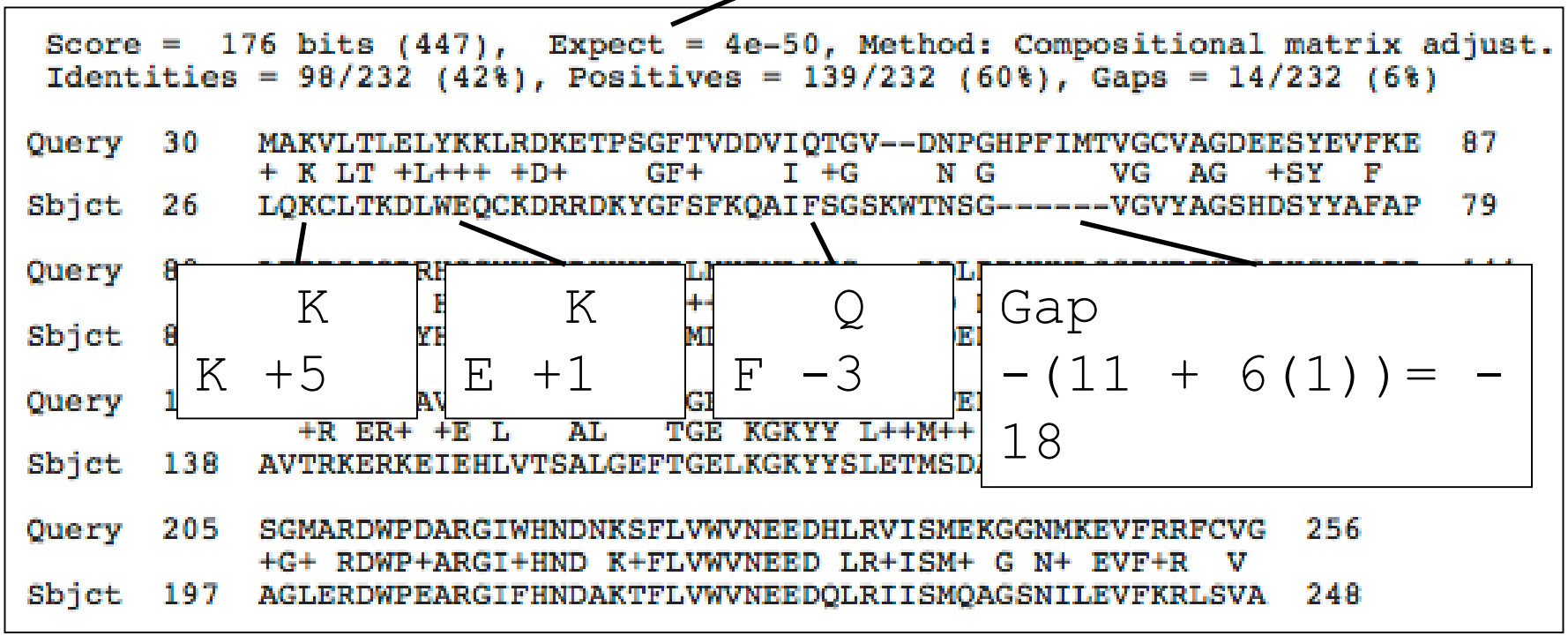
Number of Chance Alignments = 2 X 10^{-73}



How does BLAST work

Step 2. Score each alignment – protein alignment

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



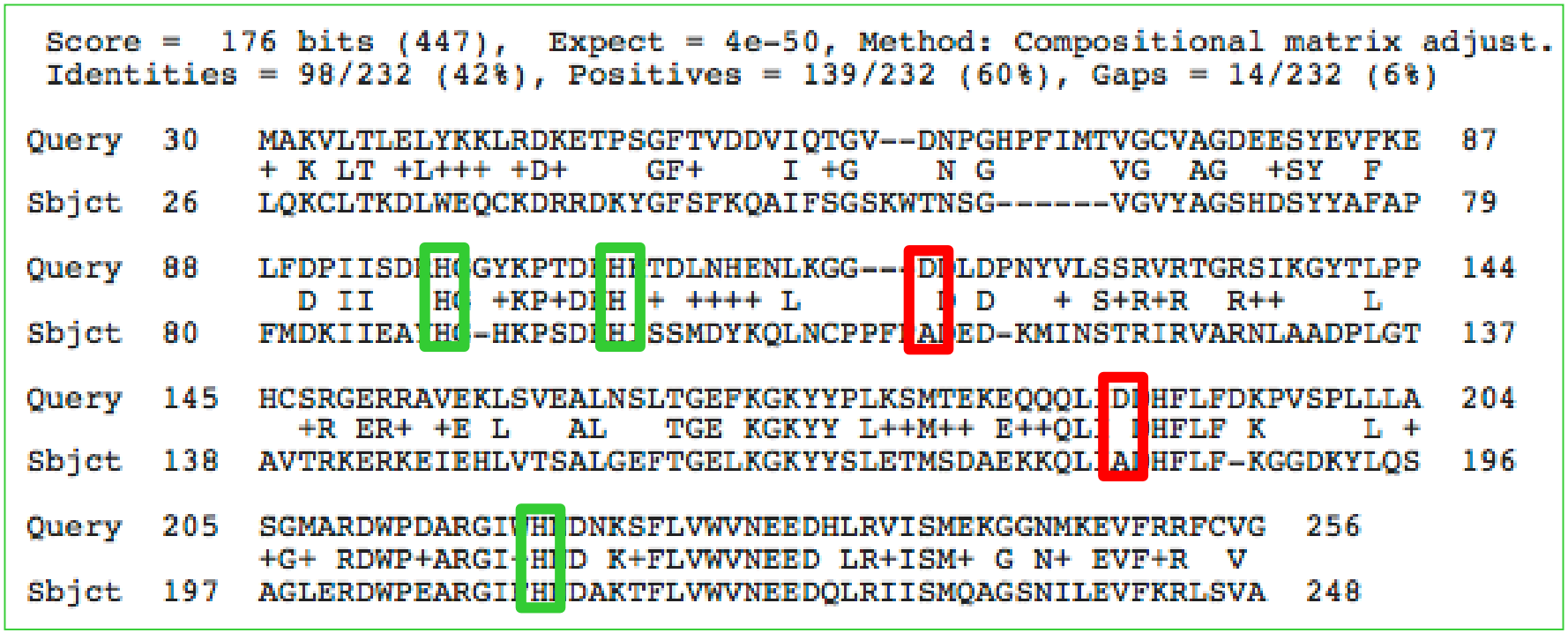
Scores from BLOSUM62, a position independent matrix

- NCBI Discovery Workshops

BLOSUM62, a position independent matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
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

BLOSUM62 substitution score is position independent

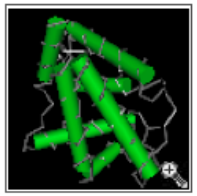


Scores from BLOSUM62, a position independent matrix

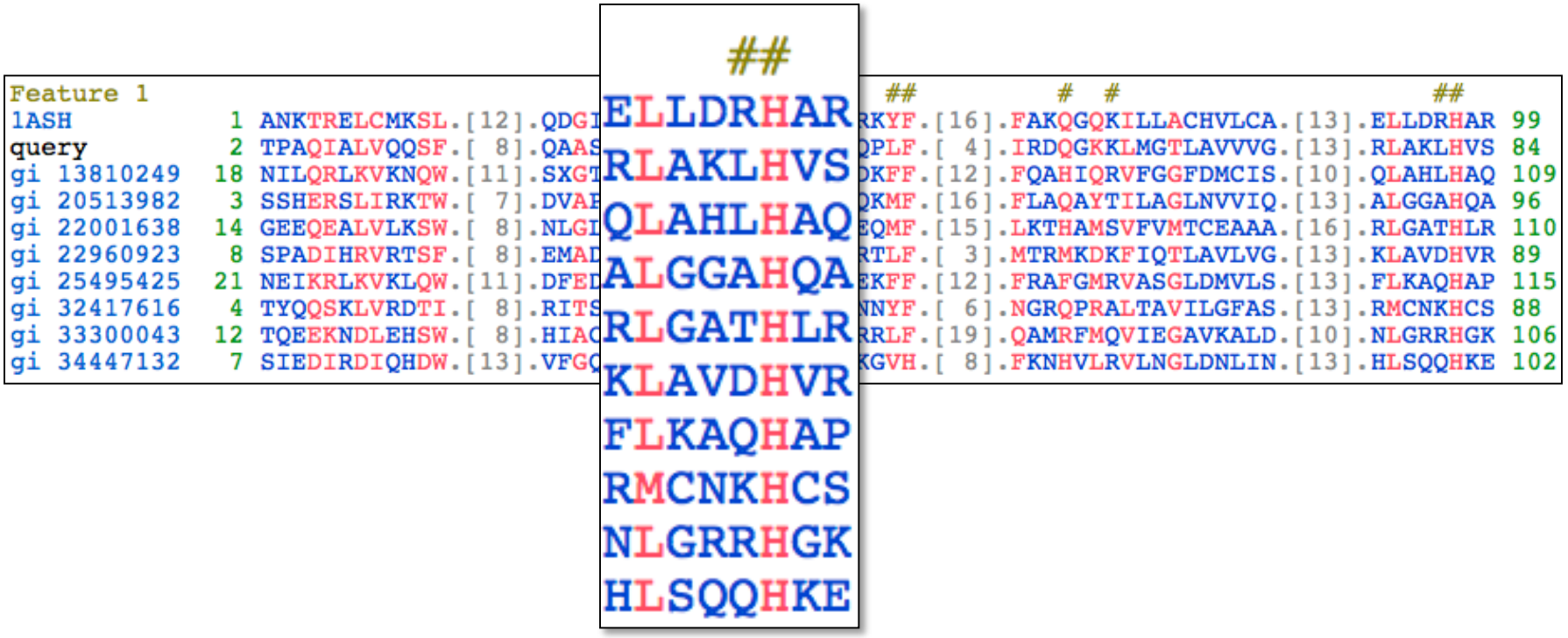
- NCBI Discovery Workshops

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.



Conserved Histidine

- NCBI Discovery Workshops

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

Conserved Histidine

bla

DELTA-BLAST

ILPAASRLAKLHVS-YGVQPTHYAPVG

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

- NCBI Discovery Workshops

Search PSSM with DELTA-BLAST

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

Enter Query Sequence

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

Reset page | Bookmark

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)

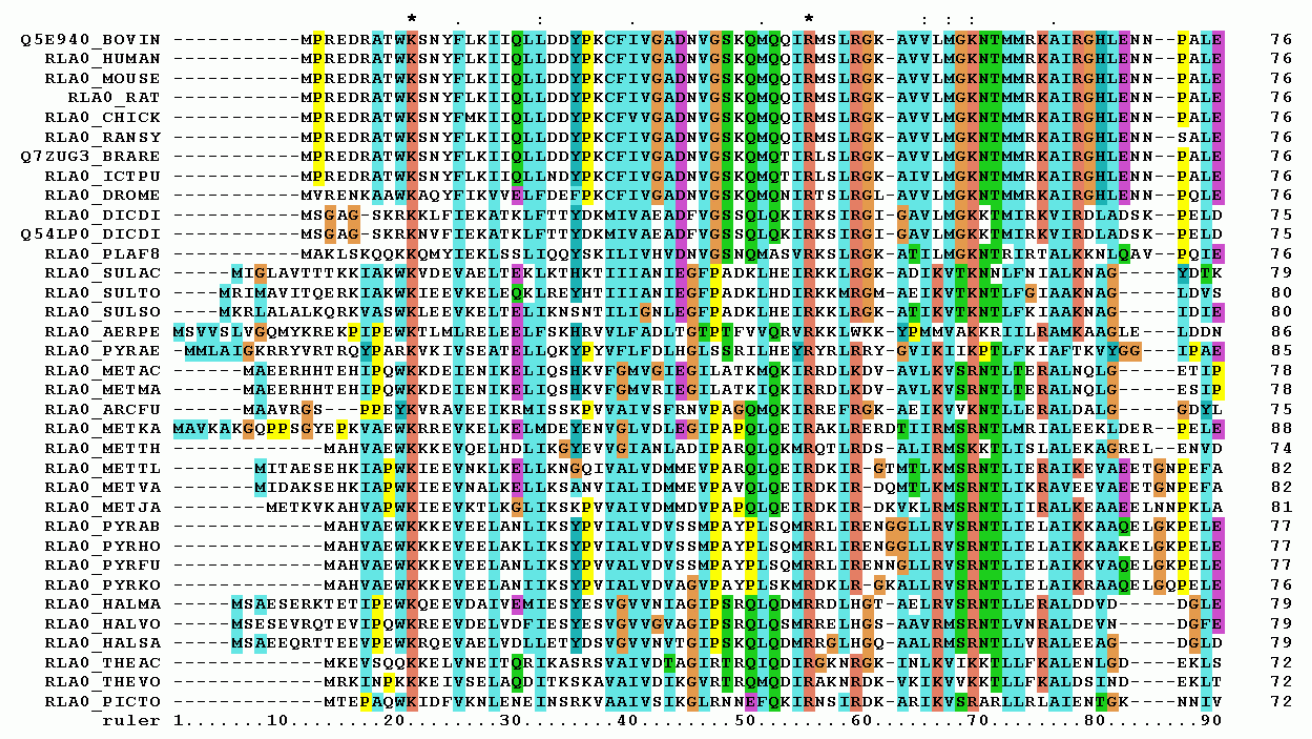
Show results in a new window

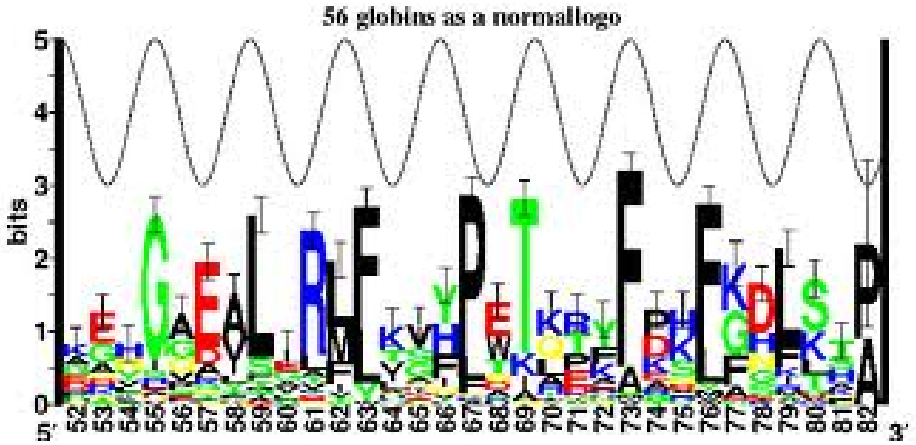
Algorithm parameters

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

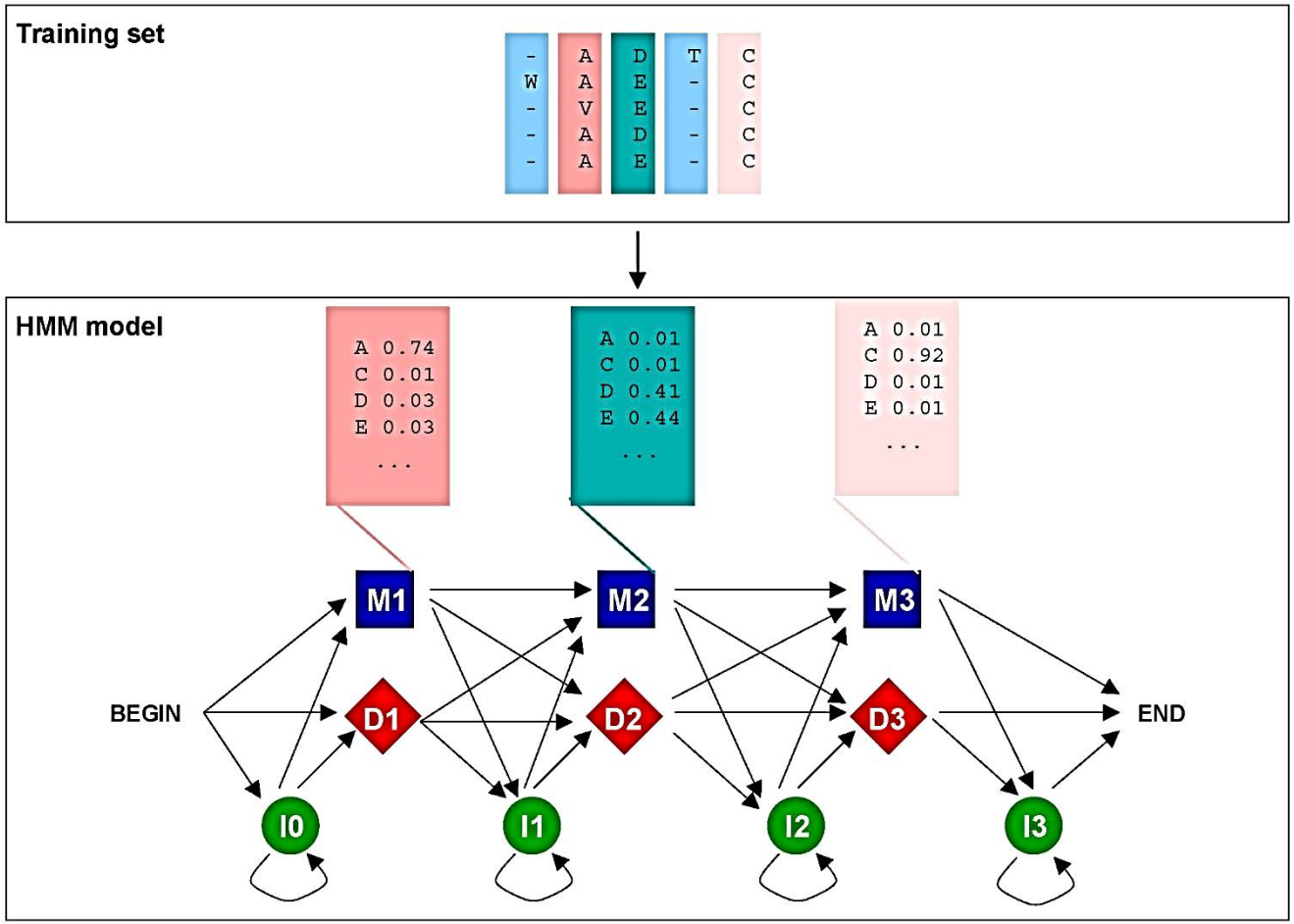
Hidden Markov Model

HMMs are trained from a multiple sequence alignment



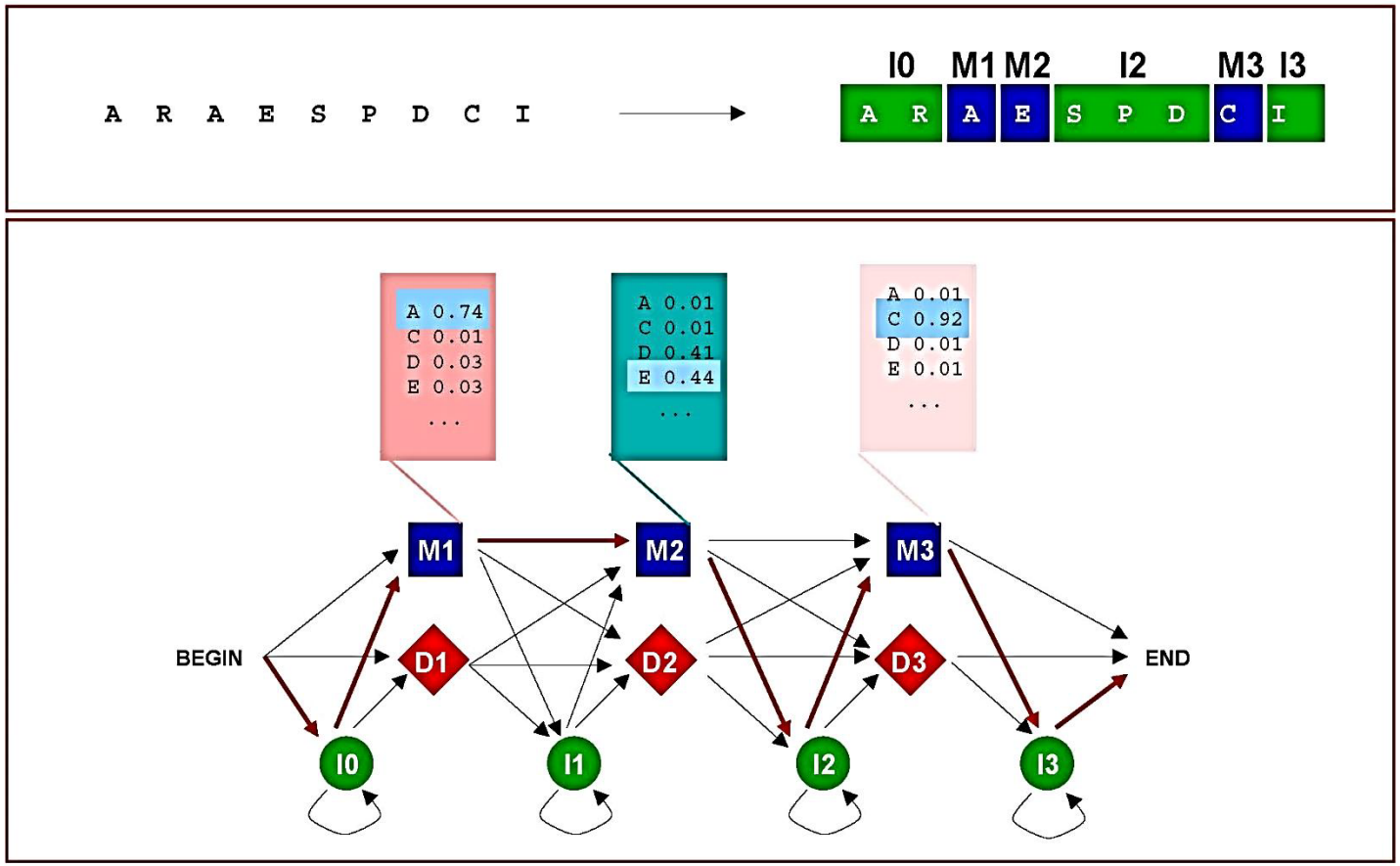


Hidden Markov Model (HMM) is more general than PSSM



Pagni, et al. EMBnet Course 2003

Match a sequence to a model
Application: Function Prediction



Pagni, et al. EMBnet Course 2003

```
>unknown_protein
MALLYRRMSMLLNIILAYIFLCAICVQGSVKQEWAEIGKNVSLECASENEAVAWKLGNQTINKNHTRYKI
RTEPLKSNDDGSENNDSQDFIKYKNVLALLDVNIKDSGNYTCTAQTGQNHSTEFQVRPYLPSKVLQSTPD
RIKRKIKQDVMLYCLIEMYPQNETTNRNLKWLKDGSQFEFLDTFSSISKLNDTHLNFTLEFTEVYKKENG
TYKCTVFDDTGLEITSKEITLFVMEVPQVSIDFAKAVGANKIYLNWTVNDGNDPIQKFFITLQEAGTPTF
TYHKDFINGSHTSYILDHFKPNTTYFLRIVGKNSIGNGQPTQYPQGITTLSYDPIFIPKVETTGSTASTI
TIGWNPPPPDLIDYIQYYELIVSESGEVPKVIEEAIYQQNSRNLPYMFDKLKTATDYEFRVRACSDLTKT
CGPWSENVNGTTMDGVATKPTNLSIQCHHDNVTRGNSIAINWDVPKTPNGKVVSYLIHLLGNPMSTVDRE
MWGPKIRRIDEPHHKTLYESVSPNTNYTVTVSAITRHKKNGEPATGSCLMPVSTPDAIGRTMWSKVNLDS
KYVLKLYLPKISERNGPICCYRLYLVRINNDNKELPDPEKLNIATYQEVHSDNVTRSSAYIAEMISSKYF
RPEIFLGDEKRFSENNDIIRDNDEICRKCLEGTPFLRKPEIIHIPPQGSLSNSDSELPILSEKDNLIKGA
NLTEHALKILESKLRDKRNAVTSDENPILSAVNPNVPLHDSSRDVFDGEIDINSNYTGFLEIIVRDRNNA
LMAYSKYFDIITPATEAEPIQSLNNMDYYLSIGVKAGAVLLGVILVFIVLWVFHHKKTKNELQGEDTLTL
RDSLRALFGRRNHNHGHEITSGNHKGEDAGDIHRLDLENAYKNPHKDTDYGFLDEYEMLDNPESDRTTKN
SDLKENACKNRYP
EQHLEIIVMLTNL
RRQITQYHYLTWK
SVSIYNTVCDLRH
EKLLATADEISKS
QDPLENTIGDFWR
TNCKIDDTLKVTQ
VAMCILVQHLRLE
```

PFAM

a pre-constructed HMM model database

for protein function domain prediction

wellcome trust sanger institute

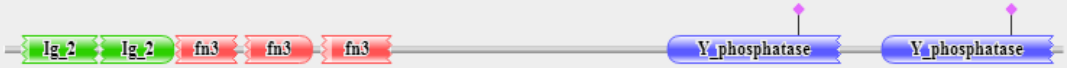
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Sequence search results

Show the detailed description of this results page.

We found **7** Pfam-A matches to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.



Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Ig_2	Immunoglobulin domain	Domain	CL0011	24	127	35	126	**11**	**78**	80	27.0	3.5e-06	n/a	Show
Ig_2	Immunoglobulin domain	Domain	CL0011	132	233	135	233	**4**	80	80	19.8	0.00063	n/a	Show
fn3	Fibronectin type III domain	Domain	CL0159	237	321	244	320	**8**	**84**	85	39.3	5.2e-10	n/a	Show
fn3	Fibronectin type III domain	Domain	CL0159	333	425	340	425	**6**	85	85	40.9	1.6e-10	n/a	Show
fn3	Fibronectin type III domain	Domain	CL0159	439	534	452	532	**11**	**83**	85	27.3	2.8e-06	n/a	Show
Y_phosphatase	Protein-tyrosine phosphatase	Domain	CL0031	916	1154	916	1153	1	**234**	235	283.6	9.6e-85	1096,1096	Show
Y_phosphatase	Protein-tyrosine phosphatase	Domain	CL0031	1212	1448	1212	1447	1	**234**	235	211.8	8.5e-63	1390,1390	Show

Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**. Our **cookie policy**.

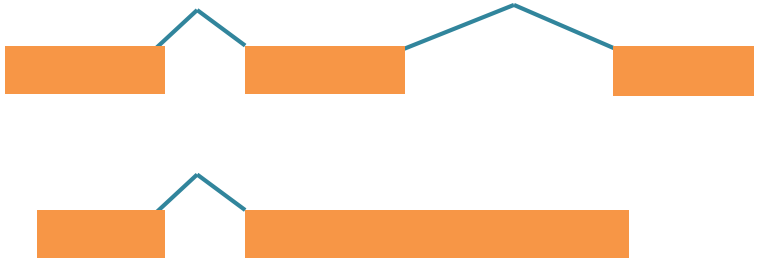
The Wellcome Trust

http://pfam.sanger.ac.uk/

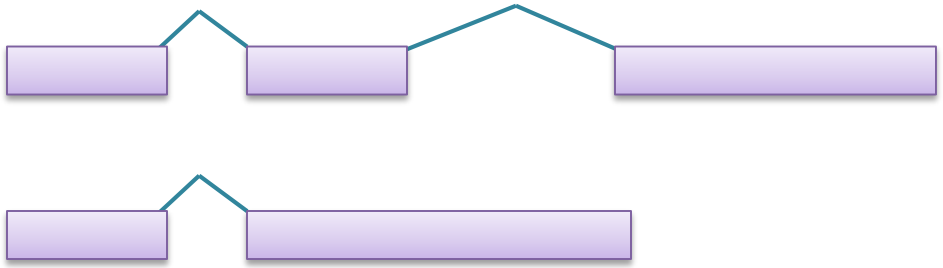
Genome Annotation

Genome Assembly

Evidence based



***Ab initio* prediction**



Genome Annotation Tools

Procaryotic genomes

Online Services

1. **RAST**
2. **NCBI**

Eucaryotic genomes

MAKER

Mark Yandell Lab
University of Utah

To run MAKER, you need the following files:

Genome sequence FASTA file

Transcript sequences:

- **Assembled from RNA-seq**
- **Transcriptome from related species**

Protein sequencies:

- **From related species**
- **Uniprot/Swissprot**

Where to run MAKER:

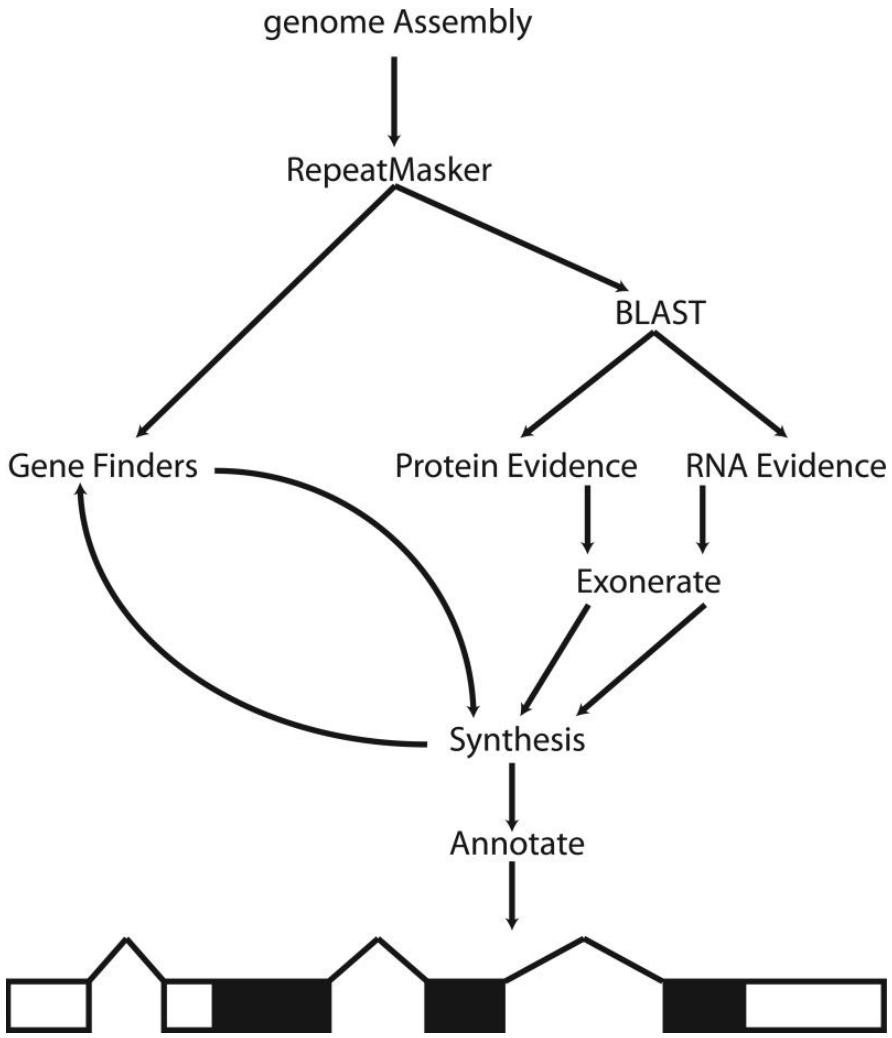
- **Use CyVerse/XSEDE**

  https://wiki.cyverse.org/wiki/display/TUT/MAKER+2.31.9+with+CCTOOLS+Jetstream+Tutoria

- **Use BioHPC**

  https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=65#c

The MAKER Pipeline:



Campbell et al. Curr Protoc Bioinformatics. 2014; 48: 4.11.1–4.11.39.

Use BioHPC

Create and edit control file

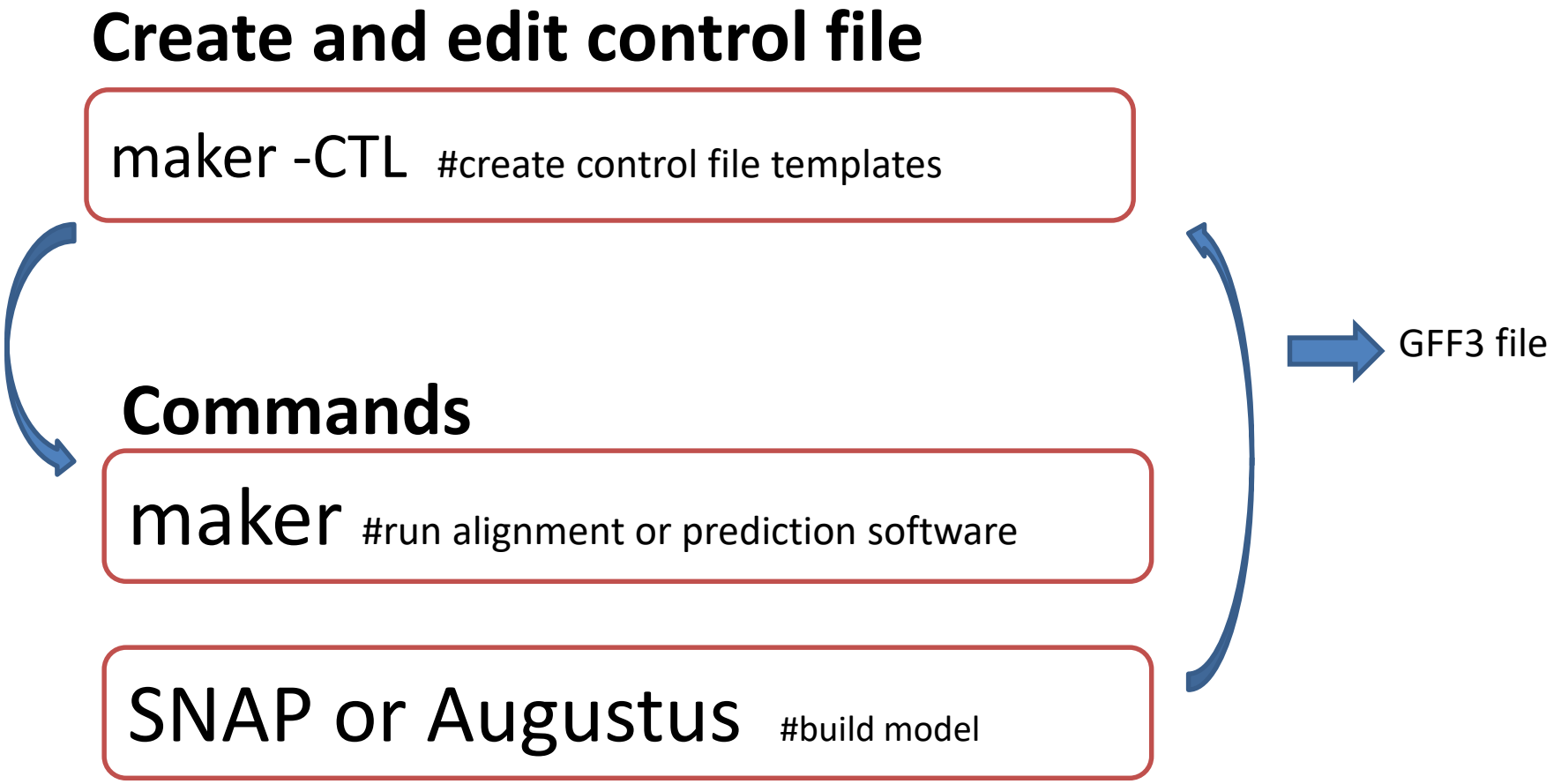
maker -CTL #create control file templates

Commands

maker #run alignment or prediction software

SNAP or Augustus #build model

Use BioHPC


Create and edit control file
maker -CTL #create control file templates
Commands
maker #run alignment or prediction software
SNAP or Augustus #build model
GFF3 file

Step 1. Repeat Masking

- Simple repeats: e.g. “AAAAA…AAA”
- Prebuilt DB: Repbase
- Custom DB: build with RepeatModeler

Control file: **maker_opts.ctl** *

```
genome=dpp_contig.fasta #genome sequence
…
model_org=all #select a model organism for RepBase masking in RepeatMasker
rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker
repeat_protein= #provide a fasta file of transposable element proteins for RepeatRunner
rm_gff= #pre-identified repeat elements from an external GFF3 file
prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no
softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and dust filtering)

…
cpus=1
```

- Keep this file in the directory where you execute the maker command.

Step 2. Train a gene prediction model

Training data set:

- Assembled RNA-seq
- Transcripts from related species
- Proteins from related species

Procedures:

- Alignment with BLAST
- Refine exon-intron junctions with EXONERATE
- Build HMM model with SNAP or AUGUSTUS

MAKER control file for step 2:

```
genome=dpp_contig.fasta #genome sequence
…
est=transcriptome.fasta
altest=otherspecesgene.fasta
protein=protein.fasta
est2genome=1
protein2genome=1
```

Build model with SNAP

Step 3. *Ab initio* Prediction with SNAP or AUGUSTUS

genome=dpp_contig.fasta #genome sequence

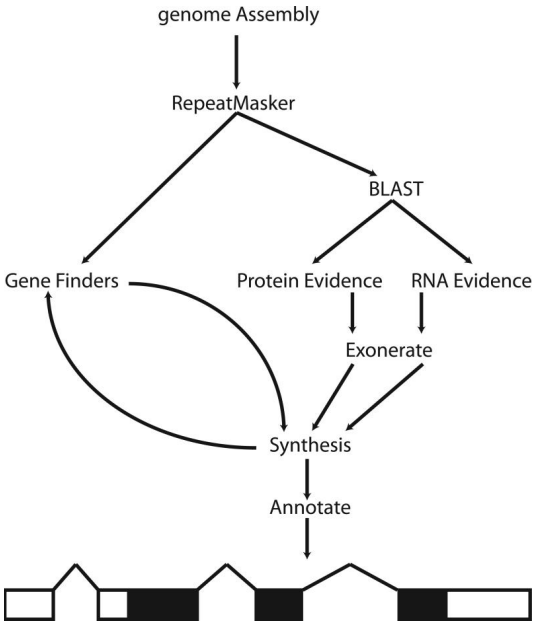
…

snaphmm=pyu1.hmm

est2genome=0

protein2genome=0

GFF file with predicted genes



Two or more iterations

genome=dpp_contig.fasta #genome sequence

…

snaphmm=pyu2.hmm

est2genome=0

protein2genome=0

Control file in first round of MAKER

```
genome=contig.fasta

est=EST1.fa,EST2.fa
altest=myAltEST.fa
protein=myprotein.fa

model_org=simple
rmlib=myRepeat.fa
repeat_protein=rp.fa

est2genome=1
protein2genome=1
```

Following rounds of MAKER

```
genome=contig.fasta

est=
altest=
protein=

model_org=
rmlib=
repeat_protein=

est2genome=0
protein2genome=0
```

Also included in following rounds of MAKER

maker_gff=pyu_rnd1.all.gff

est_pass=1 #pass on EST alignment in GFF
protein_pass=1 #pass on protein alignment in GFF
rm_pass=1 #pass on repeat alignment in GFF

pred_stats=1 #report AED stats

This way, MAKER would not run BLAST again, instead it will use the alignment from GFF file

A few other notes

1. Use MPI to parallelize the run:

   In control file: **cpus=1**

   mpiexec **-n 40** maker -base output_rnd1

2. Use both SNAP and AUGUSTUS for prediction.

3. There is a tool to create custom gene, transcript and protein names.

Related to running MAKER on BioHPC

1. **Set tmp directory to /workdir/xxxxx:**

   Create directory:

   ```
   mkdir /workdir/xxxxx/tmp
   ```

   In control file:

   ```
   TMP=/workdir/xxxxx/tmp
   ```

2. **Run mpiexec**

   ```
   /usr/local/mpich/bin/mpiexec -n 40 ....
   ```

   Use machines with >=40 cores on BioHPC, and use all cores

3. **Copy maker and repeatMasker to /workdir/xxxxx**

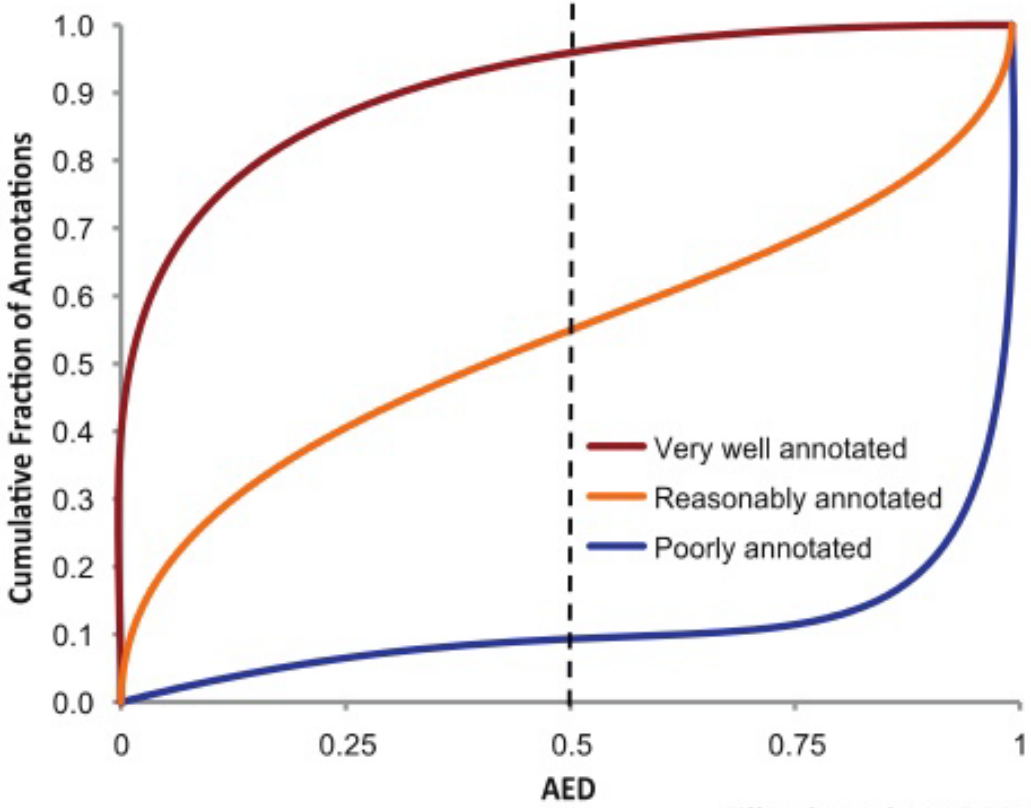
These two directories contain large data files, better to keep them on /workdir

Avoid under-fitting and over-fitting:
Evaluate results with AED Score

(Annotation Edit Distance)

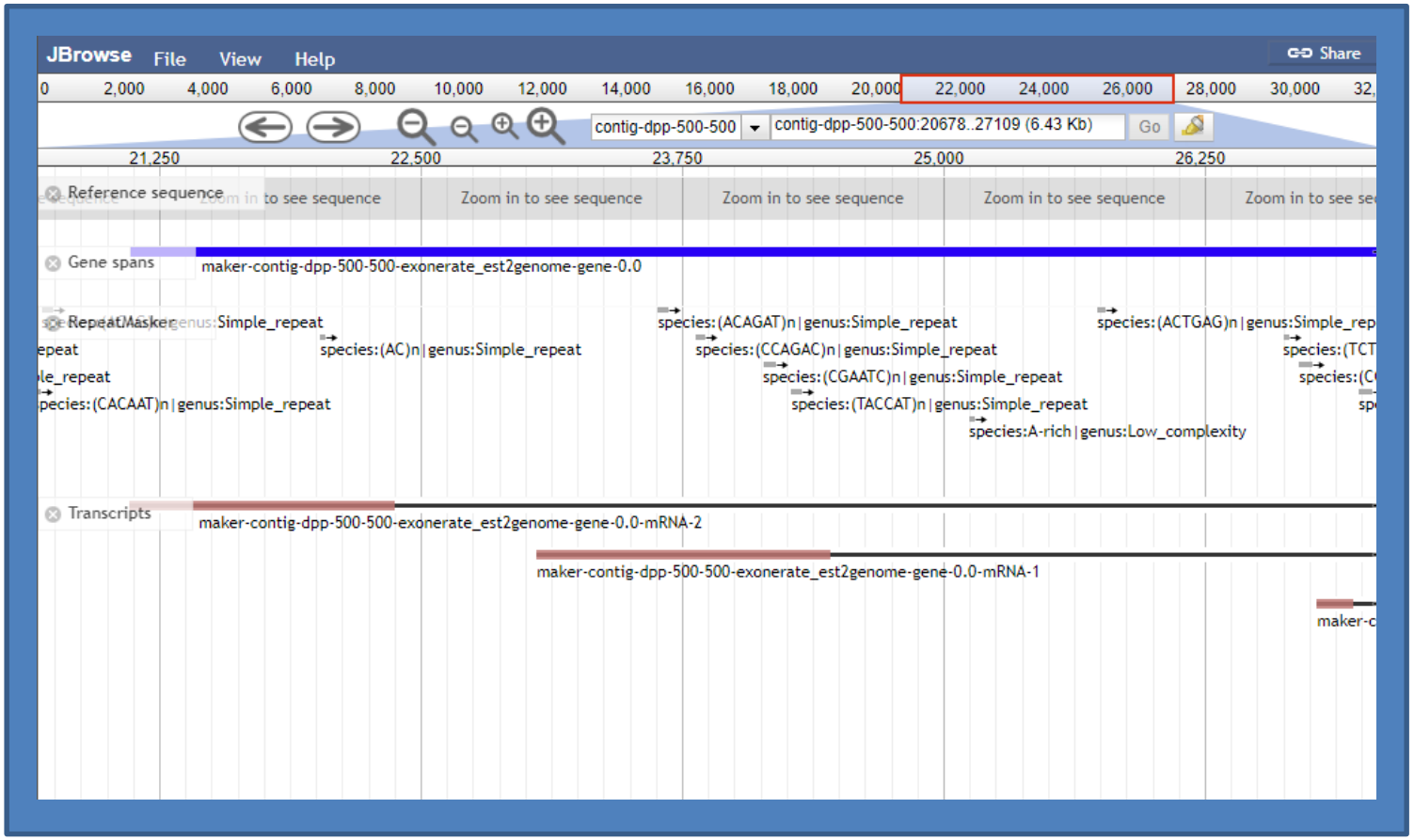
AED=0: Genes models of perfect concordance with the evidence;
AED=1: Genes models with no evidence support

Cumulative Distribution of AED



Eilbeck et al BMC Bioinformatics 2009

Visualization - JBrowse or IGV



JBrowse: https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=357#c

IGV: http://software.broadinstitute.org/software/igv/UserGuide

Can I trust MAKER annotation?

If a gene of interest is missed from annotation:

Run TBLASTN with a closely related protein:

```
makeblastdb -in myGenome.fa -parse_seqids -dbtype nucl

tblastn -query myProtein.fa -db myGenome.fa -out output_file
```

Run PFAM on all ORF (slow, and exons only)

```
getorf -minsize 100 -sequence myGenome.fa -outseq myorf.fa

pfam_scan.pl -fasta myorf.fa –pfamB mydomain.hmm
```