

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.

The screenshot shows the NCBI BLAST web interface. At the top, there are navigation tabs: Home, Recent Results, Saved Strategies, and Help. The main heading is 'Standard Protein BLAST'. Below this, there are tabs for different BLAST programs: blastn, blastp, blastx, tblastn, and tblastx. The 'blastp' tab is selected. The interface includes a text input field for 'Enter Query Sequence', a 'Clear' button, and a 'Query subrange' section with 'From' and 'To' input fields. There is also an option to 'Or, upload file' with a 'Browse...' button. Below that is a 'Job Title' field with a placeholder text 'Enter a descriptive title for your BLAST search'. A checkbox option is present for 'Align two or more sequences'. The 'Choose Search Set' section includes a 'Database' dropdown menu set to 'Non-redundant protein sequences (nr)', an 'Organism' field with a placeholder 'Enter organism name or id--completions will be suggested' and an 'Exclude' checkbox, and an 'Exclude' section with checkboxes for 'Models (XM/XP)' and 'Uncultured/environmental sample sequences'. At the bottom, there is an 'Entrez Query' field.

```
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--DNP GHPFIMTVGCVAGDEESYEVPFKE 87
      + K LT +L+++ +D+ GF+ I +G N G VG AG +SY F
Sbjct 26 LQKCLTKDLWEQCKDRRDKEYGFSFKQAIFSGSKWTNSG-----VGVYAGSHDSYYAFAP 79

Query 88 LFDPIISDRHGGYKPTDKHKTDLNHENLKG---DDLDPNYVLSSRVRTGRSIRKGYTLPP 144
      D II HG +KP+DKH + ++++ L D D + S+R+R R++ L
Sbjct 80 FMDKIIEAYHG-HKPSDKHISSMDYKQLNCPFPPEDED-KMINSTRIRVARNLAADPLGT 137

Query 145 HCSRGERRAVEKLSVEALNSLTGEFGKGYYP LKSMTEKEQQQLIDDHFLFDKPVSP LLLA 204
      +R ER+ +E L AL TGE KGKYY L++M++ E++QLI DHFLF K L +
Sbjct 138 AVTRKERKEIEHLVTSALGEFTGELKGKYY SLETMSDAEKKQLIADHFLF-KGGDKYLQS 196

Query 205 SGMARDWP DARGIWHNDNKSFLVWVNEEDHLRVISMEKGGNMKEVFRFRFCVG 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

The BLAST Search Algorithm

query word ($W = 3$)

Step1 Query: TGSQSLAALLNKCKT**PQG**QRLVNQWIKQPLMDKNRIEERLNLVEAFV

Step2 neighborhood words

PQG	18	
PEG	15	
PRC	14	
PKG	14	
PNG	13	
PDG	13	
PHG	13	neighborhood score threshold ($T = 13$)
PMG	13	
PSG	13	
<hr/>		
PQA	12	
PQN	12	
<i>etc...</i>		

Step3

Query: 325SLAALLNKCKT**PQG**QRLVNQWIKQPLMDKNRIEERLNLVEA365
+LA++L+ TPGR++W+P+D+ER+A

Subject: 290TLASVLDCTVT**PMG**SRMLKRWLHMPVRDTRVLLERQQTIGA330

High-scoring Segment Pair (HSP)

How does BLAST work

Step 2. scoring alignments

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

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

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

Query  1982  TTCTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTAATCAT  2041
      |||
Sbjct  33892  TTTGCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT  33952

Query  2042  TTTGCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT  2100
      |||
Sbjct  33893  TTTGCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT  33952

Query  2101  ACAGCACTGTTA-TGGTTCTGTGG  2159
      |||
Sbjct  33953  CTATGACTGTTATTGGTTCTATGA  34012

Query  2160  AAGTTCCAGTGTTCCTGTGGGCTA  2219
      |||
Sbjct  34013  AAATTCCACTATTCTCTCTTTCCCTATTTC AATGGAGGACTTCTAGTTCCTTCTGGATTA  34072

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

Match=+2

Mismatch=-3

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

How does BLAST work

Step 2. Score each alignment – protein alignment

Number of Chance Alignments = 4×10^{-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--DNPGHFPFIMTVGCVAGDEESYEVFKE  87
          + K LT +L+++ +D+    GF+    I +G    N G      VG AG +SY F
Sbjct  26  LQKCLTKDLWEQCKDRRDKYGFSPKQAI FSGSKWTNSG-----VGVYAGSHDSYYAFAP  79
  
```

Query	80RH.....LM.....L.....
Sbjct	8E.....MI.....E.....
Query	1	K +5
Sbjct	138	AVTRKERKEIEHLVTSALGEFTGELKGKYYSLTMSD
Query	1	E +1
Sbjct	138	AVTRKERKEIEHLVTSALGEFTGELKGKYYSLTMSD
Query	1	F -3
Sbjct	138	AVTRKERKEIEHLVTSALGEFTGELKGKYYSLTMSD
Query	1	Gap
Sbjct	138	AVTRKERKEIEHLVTSALGEFTGELKGKYYSLTMSD
Query	1	- (11 + 6(1)) = -
Sbjct	138	AVTRKERKEIEHLVTSALGEFTGELKGKYYSLTMSD
Query	1	18
Sbjct	138	AVTRKERKEIEHLVTSALGEFTGELKGKYYSLTMSD

```

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

Scores from BLOSUM62, a position independent matrix

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

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--DNP GHPFIMTVGCVAGDEESYEVFKE 87
      + K LT +L+++ +D+ GF+ I +G N G VG AG +SY F
Sbjct 26 LQKCLTKDLWEQCKDRRDKYGFSPKQAI FSGSKWTNSG-----VGVYAGSHDSYYAFAP 79

Query 88 LFDPIISD H GYKPTD H TDLNHENLKGG-- D LDPNYVLSSRVRTGRS IKG YTLPP 144
      D II H +KP+D H + +++++ L D + S+R+R R++ L
Sbjct 80 FMDKIIEA H -HKPSD H SSM DYKQLNCPFF A ED-KMINSTRIRVARNLAADPLGT 137

Query 145 HCSRGERRAVEKLSVEALNSLTGEF K GKYYPLKSMTEKEQQQL D HFLFDKPV SPLLLA 204
      +R ER+ +E L AL TGE KGKYY L++M++ E++QL HFLF K L +
Sbjct 138 AVTRKERKEIEHLVTSALGEFTGELKGKYY SLETMSDAEKKQL A HFLF-KGGDKYLQS 196

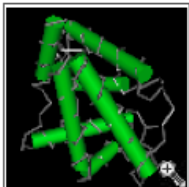
Query 205 SGMARDWPDARGI H DNKSFLVWVNEEDHLRVISMEKGGNMKEVFRFCVG 256
      +G+ RDWP+ARGI- H D K+FLVWVNEED LR+ISM+ G N+ EVF+R V
Sbjct 197 AGLERDWPEARGI H DAKTFLVWVNEEDQLRIISMQAGSNILEVFKRLSVA 248
```

Scores from BLOSUM62, a position independent matrix

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-dependent 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

Feature 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
lASH	1	ANKTREL	CMKSL	.	[12]	.	QDGI	EL	LD	R	H	A	R	R	K	Y	F	.	[16]	.	FAKQ	G	Q	K	I	L	L	A	C	H	V	L	C	A	.	[13]	.	EL	LD	R	H	A	R	99																																																								
query	2	TPAQIAL	VQQSF	.	[8]	.	QAAS	R	L	A	K	L	H	V	S	Q	P	L	F	.	[4]	.	IRDQ	G	K	K	L	M	G	T	L	A	V	V	G	.	[13]	.	R	L	A	K	L	H	V	84																																																						
gi 13810249	18	NILQRL	KVKNQW	.	[11]	.	SXGT	R	L	A	K	L	H	V	S	D	K	F	F	.	[12]	.	FQAH	I	Q	R	V	F	G	G	F	D	M	C	I	S	.	[10]	.	Q	L	A	H	L	H	A	Q	109																																																				
gi 20513982	3	SSHERS	LIRKTW	.	[7]	.	DVAE	Q	L	A	H	L	H	A	Q	Q	K	M	F	.	[16]	.	FLAQ	A	T	I	L	A	G	L	N	V	V	I	Q	.	[13]	.	A	L	G	G	A	H	Q	A	96																																																					
gi 22001638	14	GEEQEAL	VLKSW	.	[8]	.	NLGI	Q	L	A	H	L	H	A	Q	Q	E	Q	M	F	.	[15]	.	LKTH	A	M	S	V	F	V	M	T	C	E	A	A	.	[16]	.	R	L	G	A	T	H	L	R	110																																																				
gi 22960923	8	SPADIHR	VRTSF	.	[8]	.	EMAD	A	L	G	G	A	H	Q	A	R	T	L	F	.	[3]	.	MTRM	K	D	K	F	I	Q	T	L	A	V	L	V	G	.	[13]	.	K	L	A	V	D	H	V	R	89																																																				
gi 25495425	21	NEIKRL	KVKLQW	.	[11]	.	DFED	A	L	G	G	A	H	Q	A	E	K	F	F	.	[12]	.	FRAF	G	M	R	V	A	S	G	L	D	M	V	L	S	.	[13]	.	F	L	K	A	Q	H	A	P	115																																																				
gi 32417616	4	TYQQSK	LVRDTI	.	[8]	.	RITS	R	L	G	A	T	H	L	R	N	N	Y	F	.	[6]	.	NGRQ	P	R	A	L	T	A	V	I	L	G	F	A	S	.	[13]	.	R	M	C	N	K	H	C	S	88																																																				
gi 33300043	12	TQEEKN	DLEHSW	.	[8]	.	HIAC	R	L	G	A	T	H	L	R	R	R	L	F	.	[19]	.	QAMR	F	M	Q	V	I	E	G	A	V	K	A	L	D	.	[10]	.	N	L	G	R	R	H	G	K	106																																																				
gi 34447132	7	SIEDIRD	IQHDW	.	[13]	.	VFGQ	K	L	A	V	D	H	V	R	K	G	V	H	.	[8]	.	FKNH	V	L	R	V	L	N	G	L	D	N	L	I	N	.	[13]	.	H	L	S	Q	Q	H	K	E	102																																																				

Conserved Histidine

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
```

Heme Binding Site

Conserved Histidine

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

bla
DELTA-BLAST

ILPAASRLAKLHVS-YGVQPTHYAPVG

Search 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) [?](#) [Clear](#) [Query subrange](#) [?](#)

NP_001265090
 From
 To

Or, upload file No file chosen [?](#)

Job Title
 Enter a descriptive title for your BLAST search [?](#)

[Align two or more sequences](#) [?](#)

Choose Search Set

Database [?](#)

Organism
 Optional 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 [?](#)

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

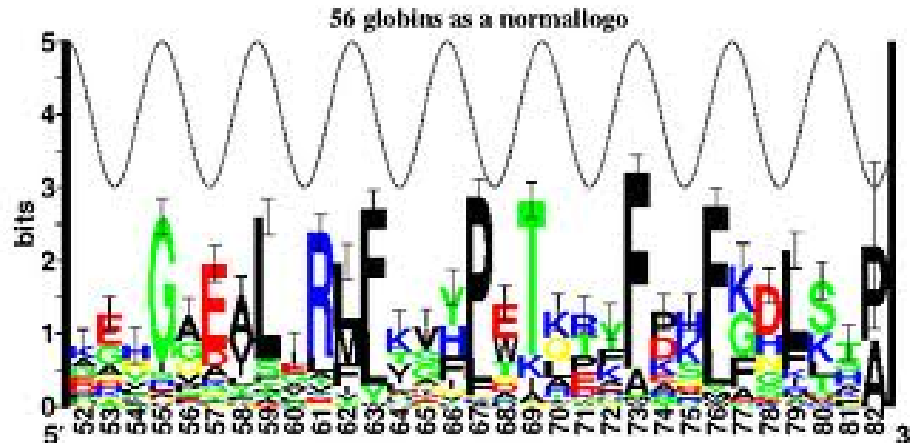
Show results in a new window

[+ Algorithm parameters](#) **Note: Parameter values that differ from the default are highlighted in yellow and marked with * sign**

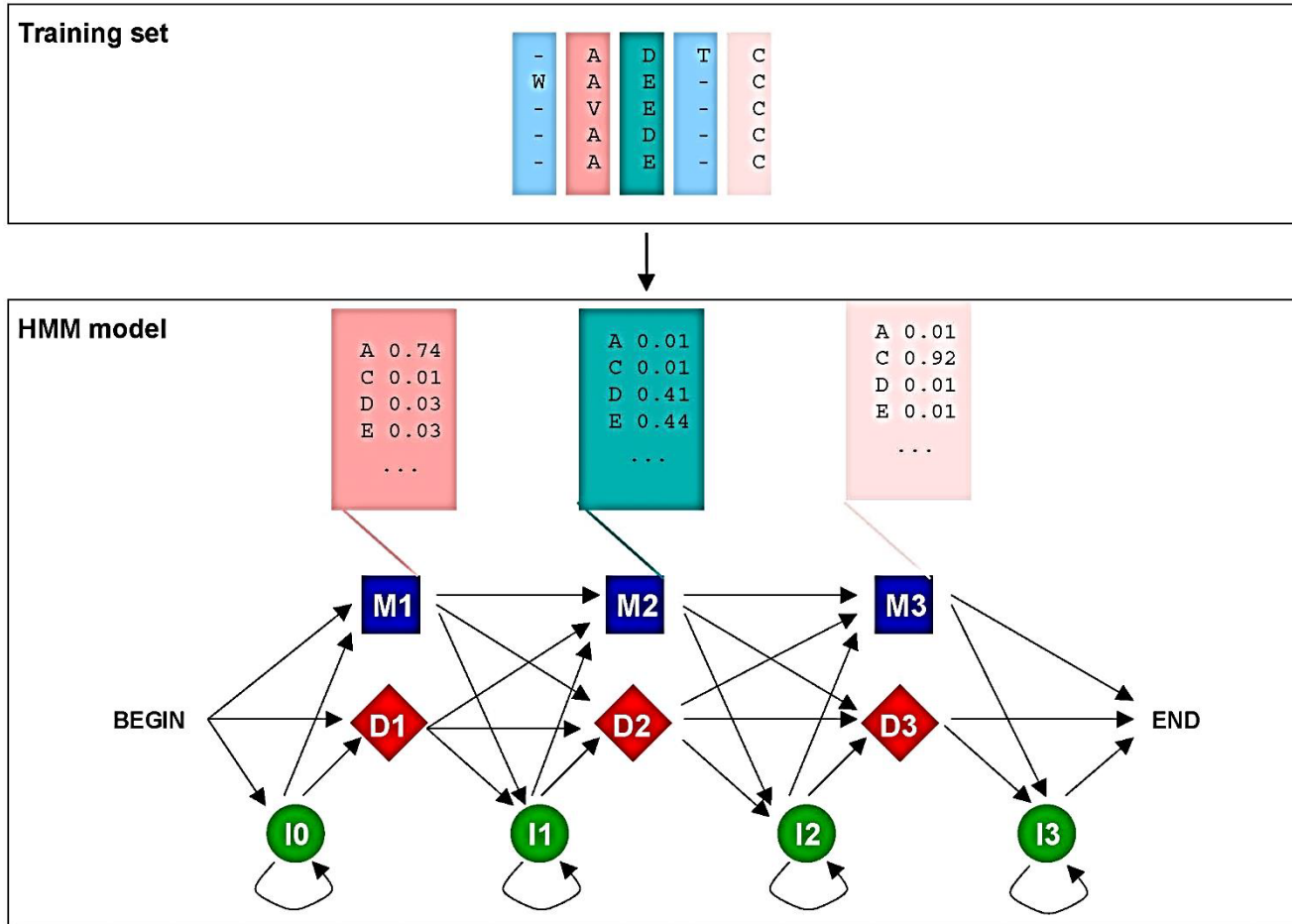
Hidden Markov Model

HMMs are trained from a multiple sequence alignment

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMOQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMOQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMOQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMOQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLDDYPKCFVVGADNVGSKQMOQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMOQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMOQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_ICTPU	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMOQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENKAANKAQYFIKVVPLFDEPKCFIVGADNVGSKQMONIRTSIRGL-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIEKATKLFETTVDKMIVAEADFGSSQLOKIRKSIIRGI-GAVLMGKNTMIRKVIIRDLDADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIEKATKLFETTVDKMIVAEADFGSSQLOKIRKSIIRGI-GAVLMGKNTMIRKVIIRDLDADSK--PELD	75
RLA0_PLAF8	-----MAKLSKQKKQMYIEKLSLTDYQSKILLVHVDDNVGSKQMASVRKSLRGK-ATLLMGKNTIRRTALKKNLQAV--DQIE	76
RLA0_SULAC	-----MIGLAVTTTKIAKWKVDEVAELTKLTKHTLIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLNFNIALKNAG----YDIE	79
RLA0_SULTO	-----MRIMAVITQERKIATKWKIEEVKLEOKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG----LDVS	80
RLA0_SULSO	-----MKRLALALKQRKVASWKLIEEVKLETELTKNSNTLLIGNEGFPADKLHEIRKKLRGK-ADIKVTKNTLFGIAAKNAG----IDIE	80
RLA0_AERPE	MSVVSIVGQMYKREKIPENKTLMLRELEELFSKRVVLFADLTGIPTFVYQVVRKKLWKK-YPMVVAKRRIILRAMKAGLE--LDDN	86
RLA0_PYRAE	MMLATGKRRYVRTROYPAARKVKIVSEATELLQKYPYVFLFDLHGLSSRIIHEVRYRLRRY-GVIKIIKPTLFGIAFTKVVYGG--TPAE	85
RLA0_METAC	MAEERHHEHTPQWKDEIENIKELIQSHKVFQMGVTEGLATKMKIRRDLDKV-AVIVKSRNTLTERALNQLG----ETIP	78
RLA0_METMA	MAEERHHEHTPQWKDEIENIKELIQSHKVFQMGVTEGLATKMKIRRDLDKV-AVIVKSRNTLTERALNQLG----ESIP	78
RLA0_ARCFU	MAAVRGS--PPEYKVRAVEEIKRMISSKVVAIVSFRNYPAGQMOKIRREFRGI-AEIKVVKNTLLEALDADG--GDYL	75
RLA0_METKA	MAVKAAGQPPSGYEPKVAENKRREVEKLELMDENVGLVDLEGIPAPQLOEIRAKLRERDTIIRMSRNTLMRIALEKLDER--PELE	88
RLA0_METTH	MAHVAEWKKEVEEQLHDLIKGEYVVGIANLADIPAROLOKMRQTLRDS-ALIRMSKKTLLISLAKKAGREL--ENVV	74
RLA0_METTL	MITAESEHKIAPWKIEEVNKLKELLKNGQIVAVLDVDMVPPAROLOEIRDKIR-ETMLKMSRNTLIERAIKEVAEETGNPEFA	82
RLA0_METVA	MIDAKSEHKIAPWKIEEVNALKLELLKNSANVIALIDMMEVPVAVQLOEIRDKIR-DQMLKMSRNTLIERAIVEVAEETGNPEFA	82
RLA0_METJA	METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMVPPAPQLOEIRDKIR-DKVKLRMSRNTLIERALKEAEELENNPKLA	81
RLA0_PYRAB	MAHVAEWKKEVEEELANLIKSPVIALVDVSSMPAYPLSQMRRILIRENGLLRVSNTLIEELAIKKAAGELGKPELE	77
RLA0_PYRHO	MAHVAEWKKEVEEELAKLIKSPVIALVDVSSMPAYPLSQMRRILIRENGLLRVSNTLIEELAIKKAAGELGKPELE	77
RLA0_PYRFU	MAHVAEWKKEVEEELANLIKSPVIALVDVSSMPAYPLSQMRRILIRENNGLRVSNTLIEELAIKKAAGELGKPELE	77
RLA0_PYRKO	MAHVAEWKKEVEEELANLIKSPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSNTLIEELAIKKAAGELGQPELE	76
RLA0_HALMA	MSAESEKRTETIPENKQEEVDIVEMIESYEVSGVNVITAGIPSRLODMRRDLHGT-AELRVSNTLIERALDDVD--DGLV	79
RLA0_HALVO	MSESEVRQTEVIPQWKREVEDELVDLLETDSYVGVVGVAGIPSRLODMRRDLHGS-AAVRMSRNTLVNRRALDEVN--DGFV	79
RLA0_HALSA	MSAEEQRTTEVIPENKQEVAVLVDLLETDSYVGVVNVITAGIPSRLODMRRDLHGS-AAVRMSRNTLVNRRALDEVN--DGLV	79
RLA0_THEAC	MKEVSSQKKELVNETORIKASRSVAIVDTAGIRTRQIDIRGNRQK-INLKVIKKTLFLFKALENLGD--EKLS	72
RLA0_THEVO	MRKINPKKKEIVSELAQDITKSKAVAVDITKGYRTIRMODIRAKNRDK-VKIKVYKTLFLFKALDLSND--EKLT	72
RLA0_PICTO	MTEPAQWKIDFVKNLENEINSRKVAIVDSIKGLRNNFQKIRNSIRDK-ARIKVSARLLRLAIENTGK--NNIV	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	

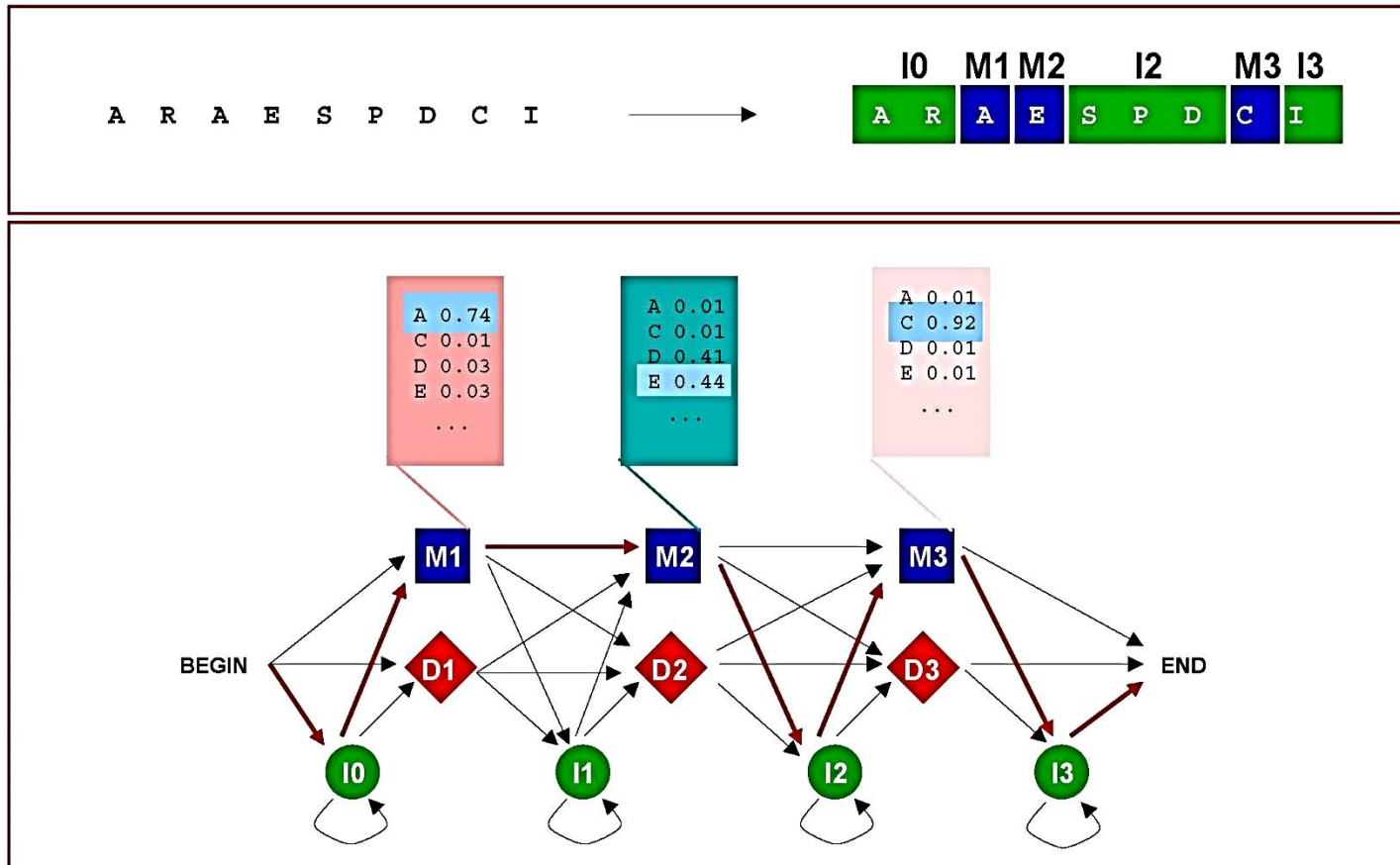


Hidden Markov Model (HMM) is more general than PSSM



Match a sequence to a model

Application: Function Prediction




```
>unknown_protein
MALLYRRMSMLLNIIILAYIFLCAICVQGSVKQEWAEIGKNVSLCASENEAVAWKLGNTQINKNHTRYKI
RTEPLKSNDDGSENNDSQDFIKYKNVLALLLDVNIKDSGNYTCTAQTGQNHSTEFQVRPYLPSKVLQSTPD
RIKRKIKQDVMLYCLIEMPQNETTNRNLKWLKDGSGQFEFLDTFSSISKLNTHLNFLEFTEVYKKENG
TYKCTVFDDTGLEITSKEITL FVMEVPQVSIDFAKAVGANKIYLNWTVNDGNDPIQKFFITLQEAGTPTF
TYHKDFINGSHTSYILDHFKNPTTYFLRIVGKNSIGNGQPTQYPQGITTLSYDPIFIPKVETTGSTASTI
TIGWNPPLDIDYIQYYELIVSESGEVPKVEEAIYQQNSRNL PYMFDLKLKATDYEFVRVACSDLTKT
CGPWSENVNGTMDGVATKPTNLSIQCHHDNVTGRNSIAINWDVPKTPNGKVVSYLIHLLGNPMSTVDRE
MWGPKIRRIDEPHHKTLYESVSPNTNYTVTVSAITRHKKNGEPATGSCMLPVPSTDAIGRTMWSKVNLD
KYVLKLYLPKISERNGPICCYRLYLVRINNDNKELPDPEKLNIAIYQEVHSDNVTSSAYIAEMISSKYF
RPEIFLGDEKRFSENNDIIRDNDICRCKLEGTPFLRKPEIIHIPPQGSLSNSDSELPILSEKDNLIKGA
NLTEHALKILESCLRDRNAVTSDENPILSAVNPVPLHDSSRDVFDGEIDINSNYTGFLIIVRDRNNA
LMAYSKYFDIITPATEAEPIQSLNNDYYLSIGVKAGAVLLGVILVFLVWVFFHKKTKNELQGEDTLTL
RDSLRLALFGRRNHNSHETTCENKUCFACRTURLDLFNAYKRNKQKQDYKCELEVEVMDNDFECPDTTAN
SDLKENACKNRYPI
EQHLEIIVMLTNL
RRQITQYHYLTWK
SVSIYNTVCDLRH
EKLLATADEISKS
QDPLENTIGDFWR
TNCKIDDTLKVTQ
VAMCILVQHLRLE
```

PFAM

a pre-constructed HMM model database
for protein function domain prediction

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](#).

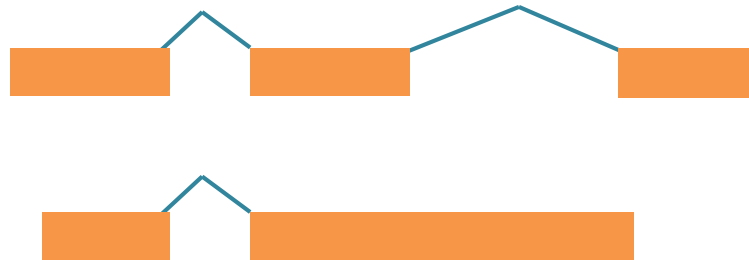
The Wellcome Trust

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

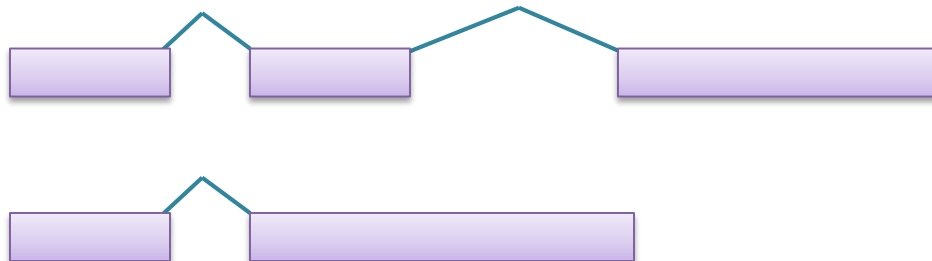
Genome Annotation

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 sequences:

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

Where to run MAKER:

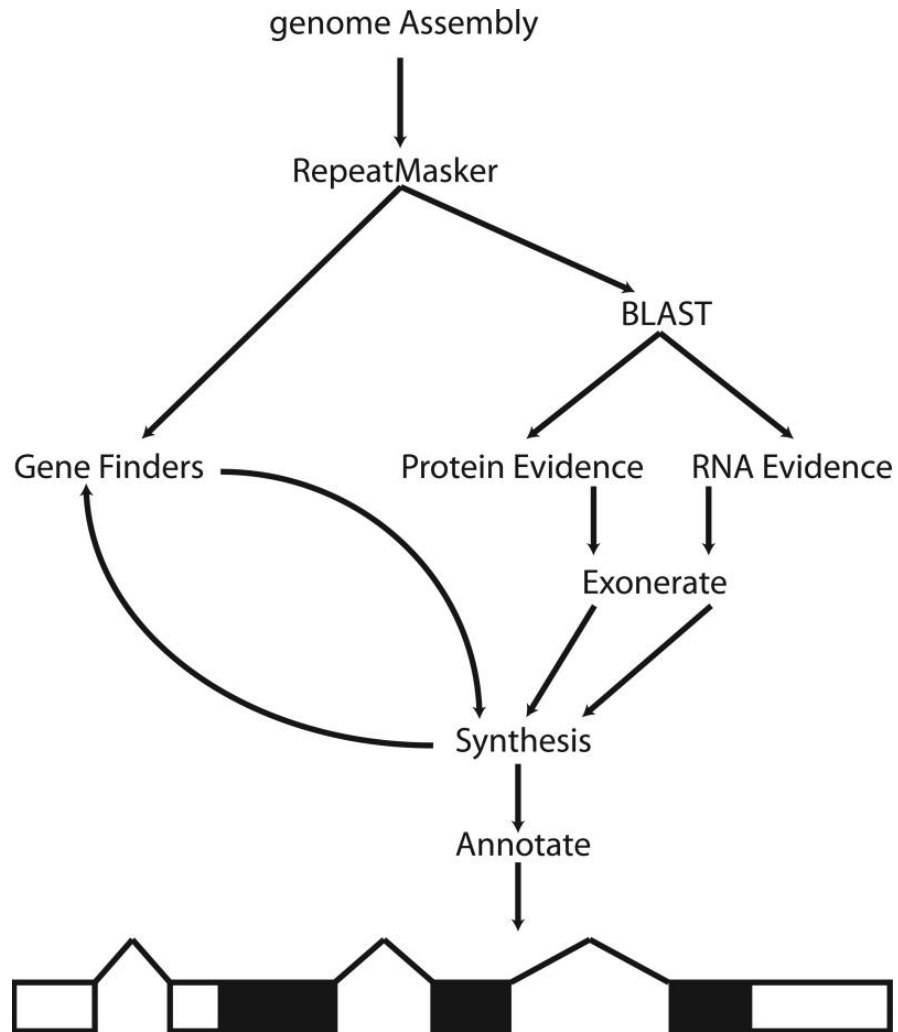
- **Use CyVerse/XSEDE**

<https://wiki.cyverse.org/wiki/display/TUT/MAKER+2.31.9+with+CCTOOLS+Jetstream+Tutorial>

- **Use BioHPC**

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

The MAKER Pipeline:



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

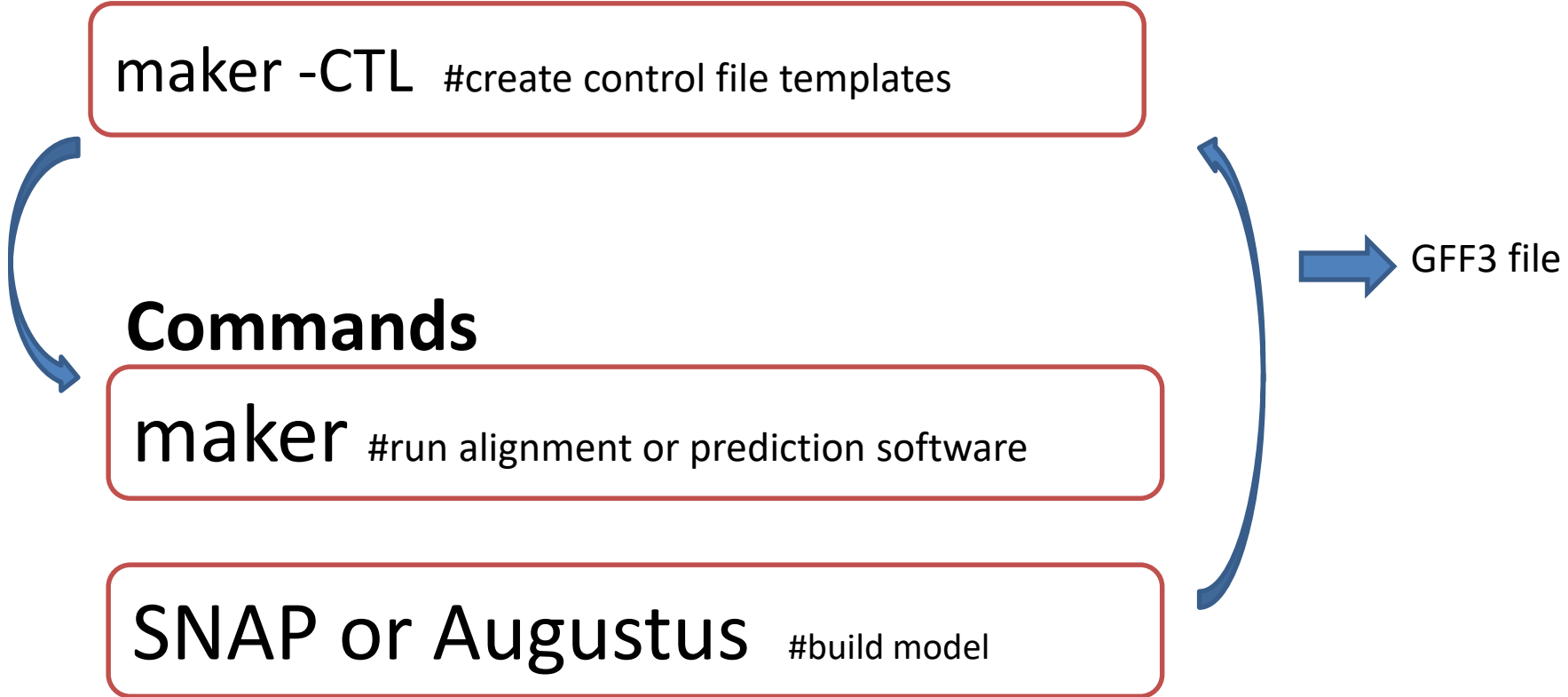
```
maker -CTL #create control file templates
```

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=otherspeciesgene.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
```

```
...
```

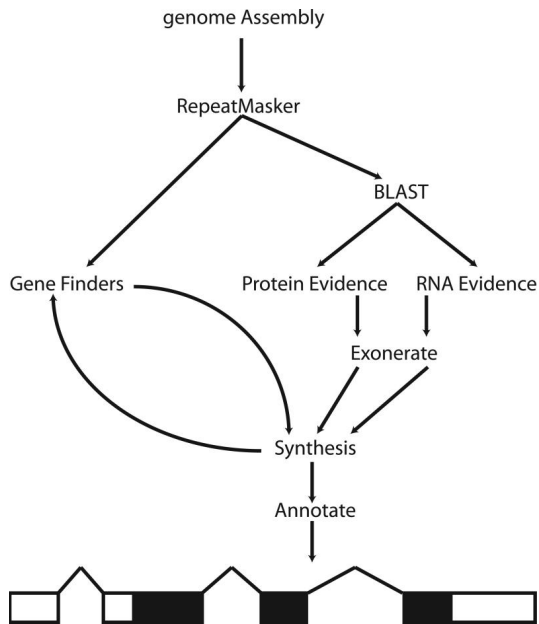
```
snaphmm=pyu1.hmm
```

```
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
```

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

2. Run mpiexec

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

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

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

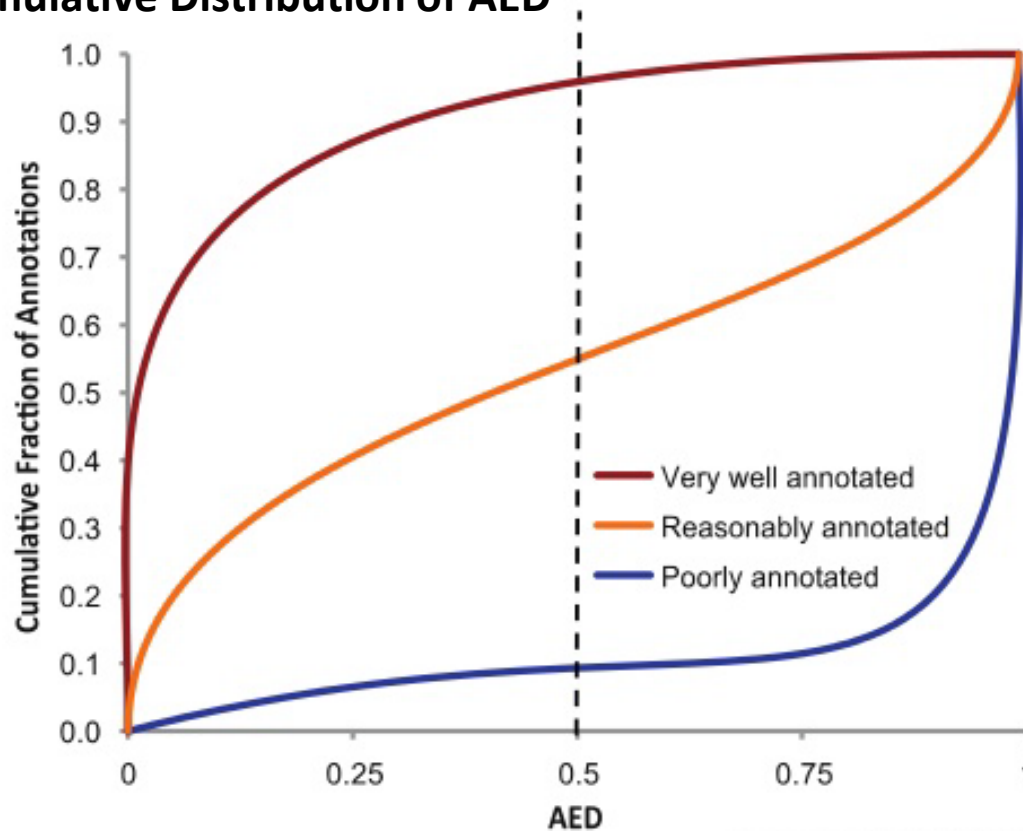
Avoid under-fitting and over-fitting: Evaluate results with AED Score

(Annotation Edit Distance)

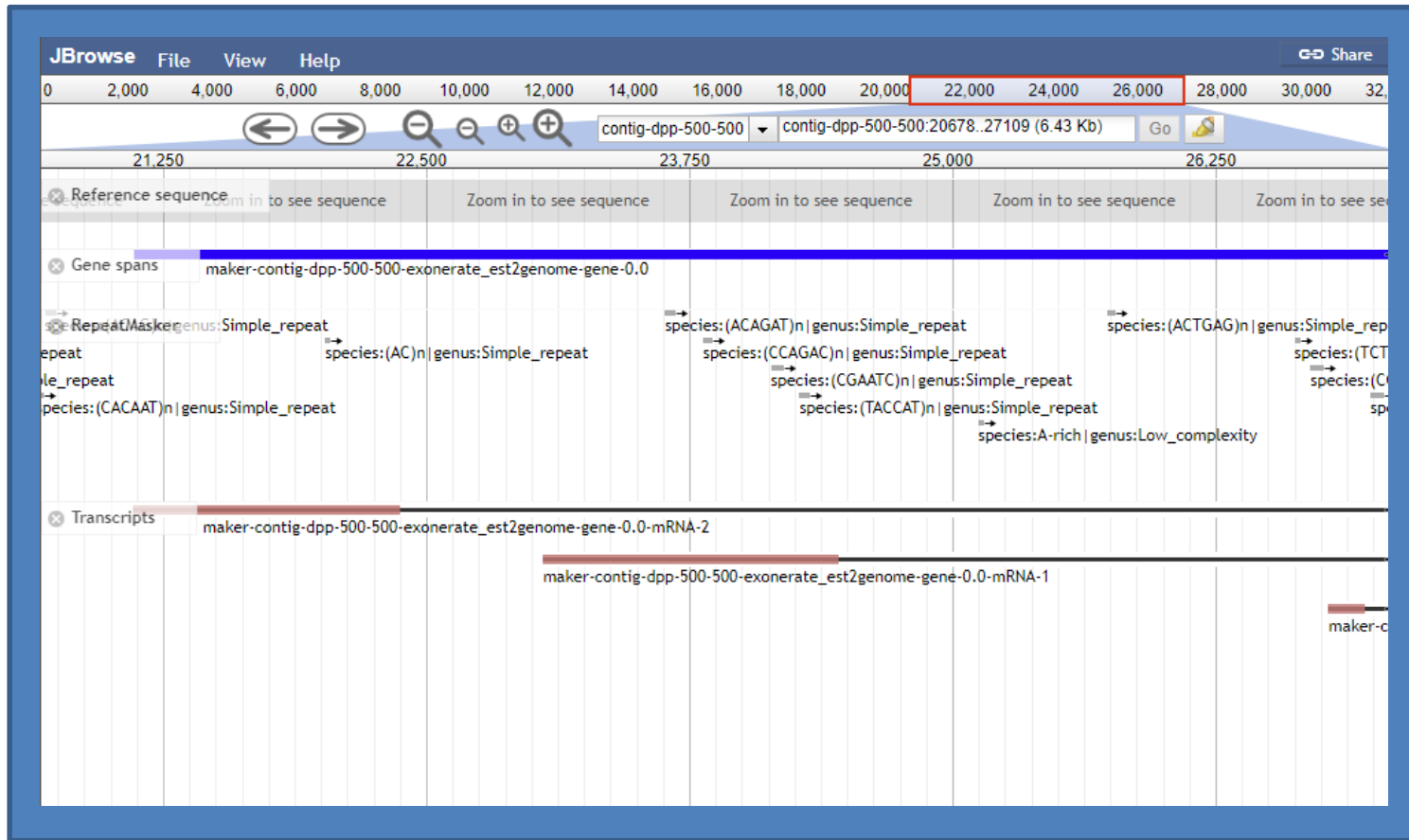
AED=0: Genes models of perfect concordance with the evidence;

AED=1: Genes models with no evidence support

Cumulative Distribution of AED



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
```