Concepts and tools for sequence alignment

Qi Sun

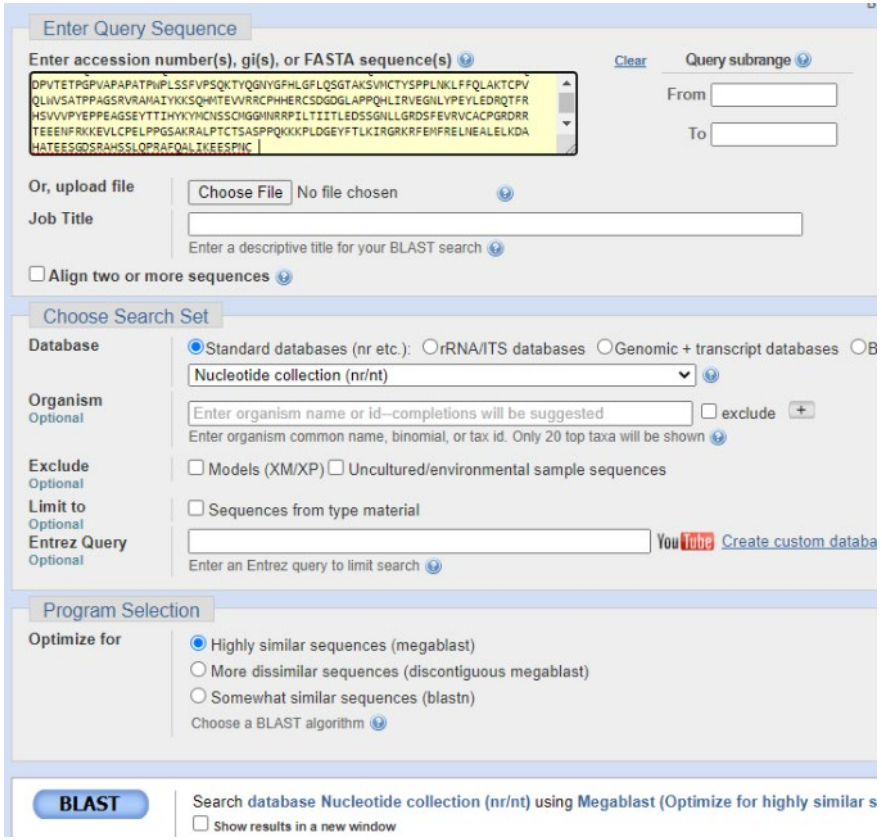
Bioinformatics Facility
Cornell University

How BLAST works?

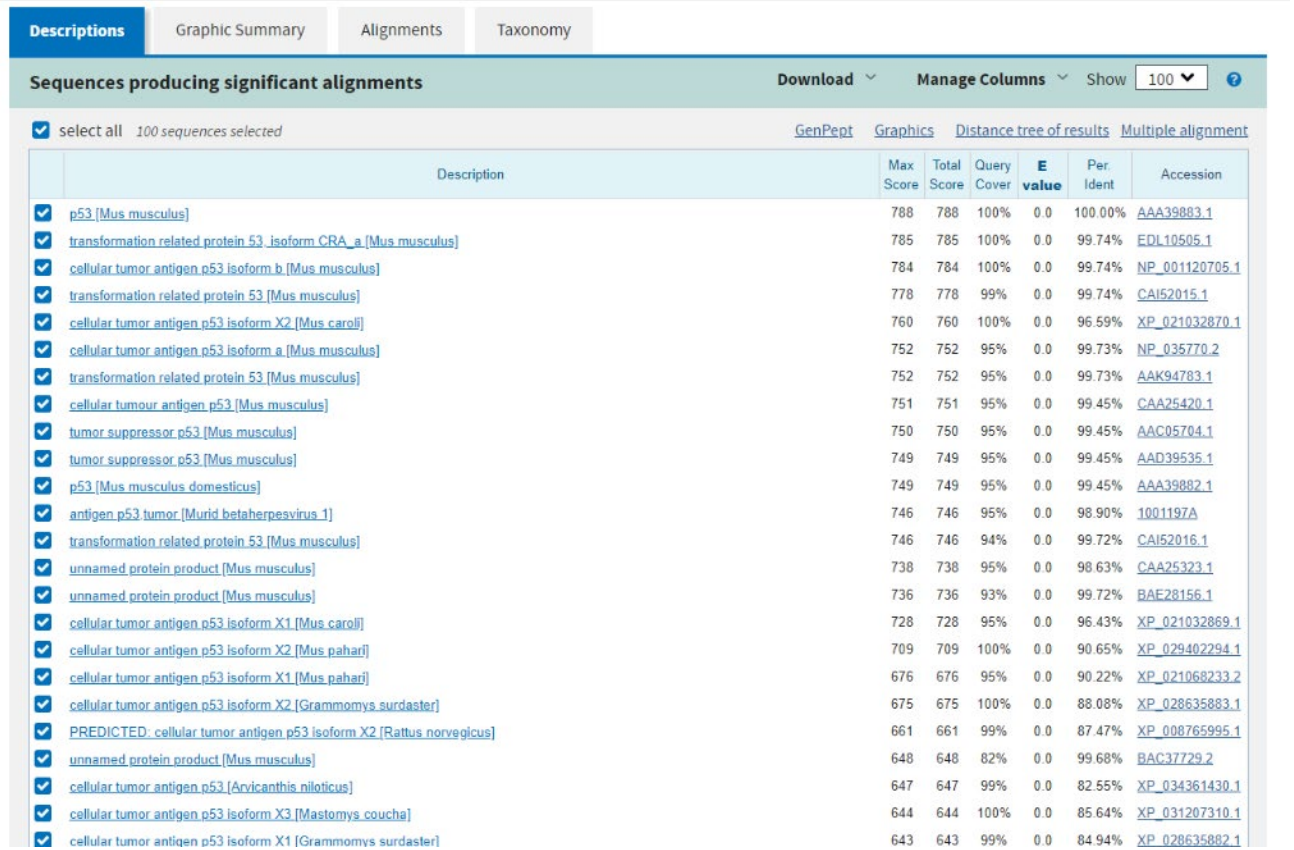
Query sequence

>unknown
MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEALRVSGAPAAQ
DPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSPPLNKLFFQLAKTCPV
QLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNLYPEYLEDRQTFR
HSVVVPYEPPEAGSEYTTIHYKYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRR
TEEENFRKKEVLCPELPPGSAKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDA
HATEESGDSRAHSSLQPRAFQALIKEESPNC

NCBI BLAST



BLAST Results



BLAST

Step1: Seeding

- Break down the query sequences into words;

BLAST

- Identify candidate targets by matching to the “word”

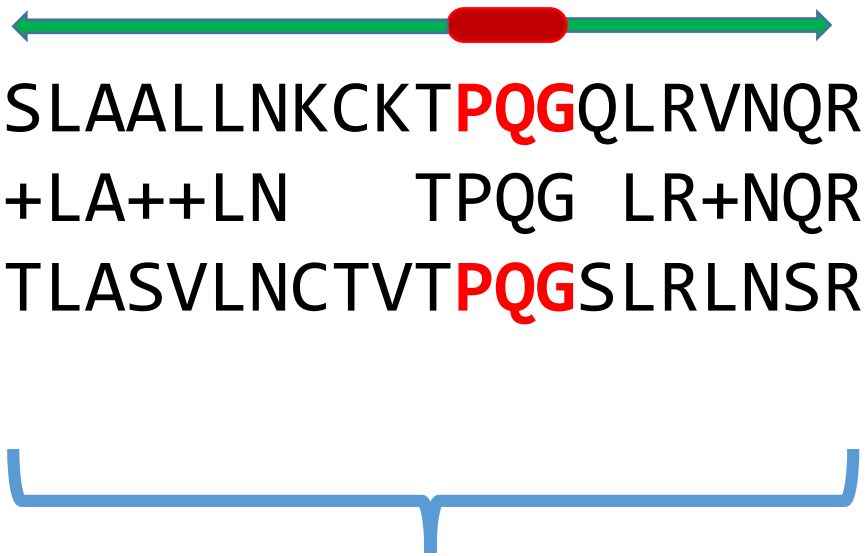
Seeding: **PQG**

MVNENTRMYIPEENHQGSNYGSPRPAHANMNANAAAGLAPEHIPTPGAALSWQAAIDAARQAKLMGSAGN
ATISTVSSTQRKRQQYGKPKKQGSTTATRPPRALLCLTLKNPIRRACISIVEWKPFEIIILLTIFANCVA
LAIYIPFPEDDSNATNSNLERVEYLFLIIFTVEAFLKVIAYGLLFHPNAYLRNGWNLLDFIIVVVGLFSA
ILEQATKADGANALGGKGAGFDVKALRAFRVLRPLRLVSGVPSLQVVLNSIIKAMVPLLHIALLVLFVII
IYAIIGLELFMGKMHKTCYNQEGIADVPAEDDPSPCALETGHGRQCQNGTVCKPGWDGPKHGITNFDNFA
FAMLTVFQCITMEGWTDVLYWVNDAVGRDWPWIYFVTLIIIGSFFVLNLVLGVLSGEFSKEREKAKARGD
FQKLREKQQLEEDLKGYLDWITQAEDIDPENEDEGMDEEKPRNMSMPTSETESVNTENVAGGDIEGENCG
ARLAHRISKSKFSRYWRRWNRFCRRKCRAAVKSNVFYWLVIFLVFLNTLTIASEHYNQPNWLTEVQDTAN
KALLALFTAEMLLKMYSLGLQAYFVSLFNRFDCFVVCGGILETILVETKIMSPLGISVLRCVRLLRIFKI
TRYWNSLSNLVASLLNSVRSIASLLLLLFLFIIIFSLLGMQLFGGKFNFDEMQTRRSTFDNFPQSLLTVF
QILTGEDWNSVMYDGI**PQG**GGPSFPGMLVCIYFIILFICGNYILLNVFLAIAVDNLADAESLTSAQKEEE
EEKERKKLARTASPEKKQELVEKPAVGESKEEKIELKSITADGESPPATKINMDDLQPNENEDKSPYPNP
ETTGEEDEEEPEMPVGPRPRPLSELHLKEKAVPMPEASAFFIFSSNNRFRLQCHRIVNDTIFTNLILFFI
LLSSISLAAEDPVQHTSFRNHILFYFDIVFTTIFTIEIALKILGNADYVFTSIFTLEIILKMTAYGAFLH
KGSFCRNYFNILDLLVVSVSLISFGIQSSAINVVKILRVLRVLRPLRAINRAKGLKHVVQCVFVAIRTIG
NIVIVTTLLQFMFACIGVQLFKGKLYTCSDSSKQTEAECKGNYITYKDGEVDHPIIQPRSWENSKFDFDN
VLAAMMALFTVSTFEGWPELLYRSIDSHTEDKGPIYNYRVEISIFFIIYIIIIAFFMMNIFVGFVIVTFQ
EQGEQEYKNCELDKNQRQCVEYALKARPLRRYIPKNQHQYKVWYVVNSTYFEYLMFVLILLNTICLAMQH
YGQSCLFKIAMNILNMLFTGLFTVEMILKLIAFKPKGYFSDPWNVFDFLIVIGSIIDVILSETNHYFCDA
WNTFDALIVVGSIVDIAITEVNPAEHTQCSPSMNAEENSRISITFFRLFRVMRLVKLLSRGEGIRTLLWT
FIKSFQALPYVALLIVMLFFIYAVIGMQVFGKIALNDTTEINRNNNFQTFPQAVLLLFRCATGEAWQDIM
LACMPGKKCAPESEPSNSTEGETPCGSSFAVFYFISFYMLCAFLIINLFVAVIMDNFDYLTRDWSILGPH
HLDEFKRIWAEYDPEAKGRIKHLDVVTLLRRIQPPLGFGKLCPHRVACKRLVSMNMPLNSDGTVMFNATL
FALVRTALRIKTEGNLEQANEELRAIIKKIWKRTSMKLLDQVVPPAGDDEVTVGKFYATFLIQEYFRKFK
KRKEQGLVGKPSQRNALSLQAGLRTLHDIGPEIRRAISGDLTAEEELDKAMKEAVSAASEDDIFRRAGGL
FGNHVSYYQSDGRSAFPQTFTTQRPLHINKAGSSQGDTESPSHEKLVDSTFTPSSYSSTGSNANINNANN
TALGRLPRPAGYPSTVSTVEGHGPPLSPAIRVQEVAWKLSSNRERHVPMCEDLELRRDSGSAGTQAHCLL
LRKANPSRCHSRESQAAMAGQEETSQDETYEVKMNHDTEACSEPSLLSTEMLSYQDDENRQLTLPEEDKR
DIRQSPKRGFLRSASLGRRASFHLECLKRQKDRGGDISQKTVLPLHLVHHQALAVAGLSPLLQRSHSPAS
MVNENTRMYIPEENHQGSNYGSPRPAHANMNANAAAGLAPEHIPTPGAALSWQAAIDAARQAKLMGSAGN
ATISTVSSTQRKRQQYGKPKKQGSTTATRPPRALLCLTLKNPIRRACISIVEWKPFEIIILLTIFANCVA
LAIYIPFPEDDSNATNSNLERVEYLFLIIFTVEAFLKVIAYGLLFHPNAYLRNGWNLLDFIIVVVGLFSA
ILEQATKADGANALGGKGAGFDVKALRAFRVLRPLRLVSGVPSLQVVLNSIIKAMVPLLHIALLVLFVII
IYAIIGLELFMGKMHKTCYNQEGIADVPAEDDPSPCALETGHGRQCQNGTVCKPGWDGPKHGITNFDNFA
FAMLTVFQCITMEGWTDVLYWVNDAVGRDWPWIYFVTLIIIGSFFVLNLVLGVLSGEFSKEREKAKARGD
FQKLREKQQLEEDLKGYLDWITQAEDIDPENEDEGMDEEKPRNMSMPTSETESVNTENVAGGDIEGENCG
ARLAHRISKSKFSRYWRRWNRFCRRKCRAAVKSNVFYWLVIFLVFLNTLTIASEHYNQPNWLTEVQDTAN
KALLALFTAEMLLKMYSLGLQAYFVSLFNRFDCFVVCGGILETILVETKIMSPLGISVLRCVRLLRIFKI
TRYWNSLSNLVASLLNSVRSIASLLLLLFLFIIIFSLLGMQLFGGKFNFDEMQTRRSTFDNFPQSLLTVF
QILTGEDWNSVMYDGIMAYGGPSFPGMLVCIYFIILFICGNYILLNVFLAIAVDNLADAESLTSAQKEEE
EEKERKKLARTASPEKKQELVEKPAVGESKEEKIELKSITADGESPPATKINMDDLQPNENEDKSPYPNP
ETTGEEDEEEPEMPVGPRPRPLSELHLKEKAVPMPEASAFFIFSSNNRFRLQCHRIVNDTIFTNLILFFI
LLSSISLAAEDPVQHTSFRNHILFYFDIVFTTIFTIEIALKILGNADYVFTSIFTLEI**PQG**MTAYGAFLH
KGSFCRNYFNILDLLVVSVSLISFGIQSSAINVVKILRVLRVLRPLRAINRAKGLKHVVQCVFVAIRTIG
NIVIVTTLLQFMFACIGVQLFKGKLYTCSDSSKQTEAECKGNYITYKDGEVDHPIIQPRSWENSKFDFDN
VLAAMMALFTVSTFEGWPELLYRSIDSHTEDKGPIYNYRVEISIFFIIYIIIIAFFMMNIFVGFVIVTFQ
EQGEQEYKNCELDKNQRQCVEYALKARPLRRYIPKNQHQYKVWYVVNSTYFEYLMFVLILLNTICLAMQH
YGQSCLFKIAMNILNMLFTGLFTVEMILKLIAFKPKGYFSDPWNVFDFLIVIGSIIDVILSETNHYFCDA
WNTFDALIVVGSIVDIAITEVNPAEHTQCSPSMNAEENSRISITFFRLFRVMRLVKLLSRGEGIRTLLWT
FIKSFQALPYVALLIVMLFFIYAVIGMQVFGKIALNDTTEINRNNNFQTFPQAVLLLFRCATGEAWQDIM
DIRQSPKRGFLRSASLGRRASFHLECLKRQKDRGGDISQKTVLPLHLVHHQALAVAGLSPLLQRSHSPAS

BLAST

Step 2: Alignment

- align query and target at each candidate region

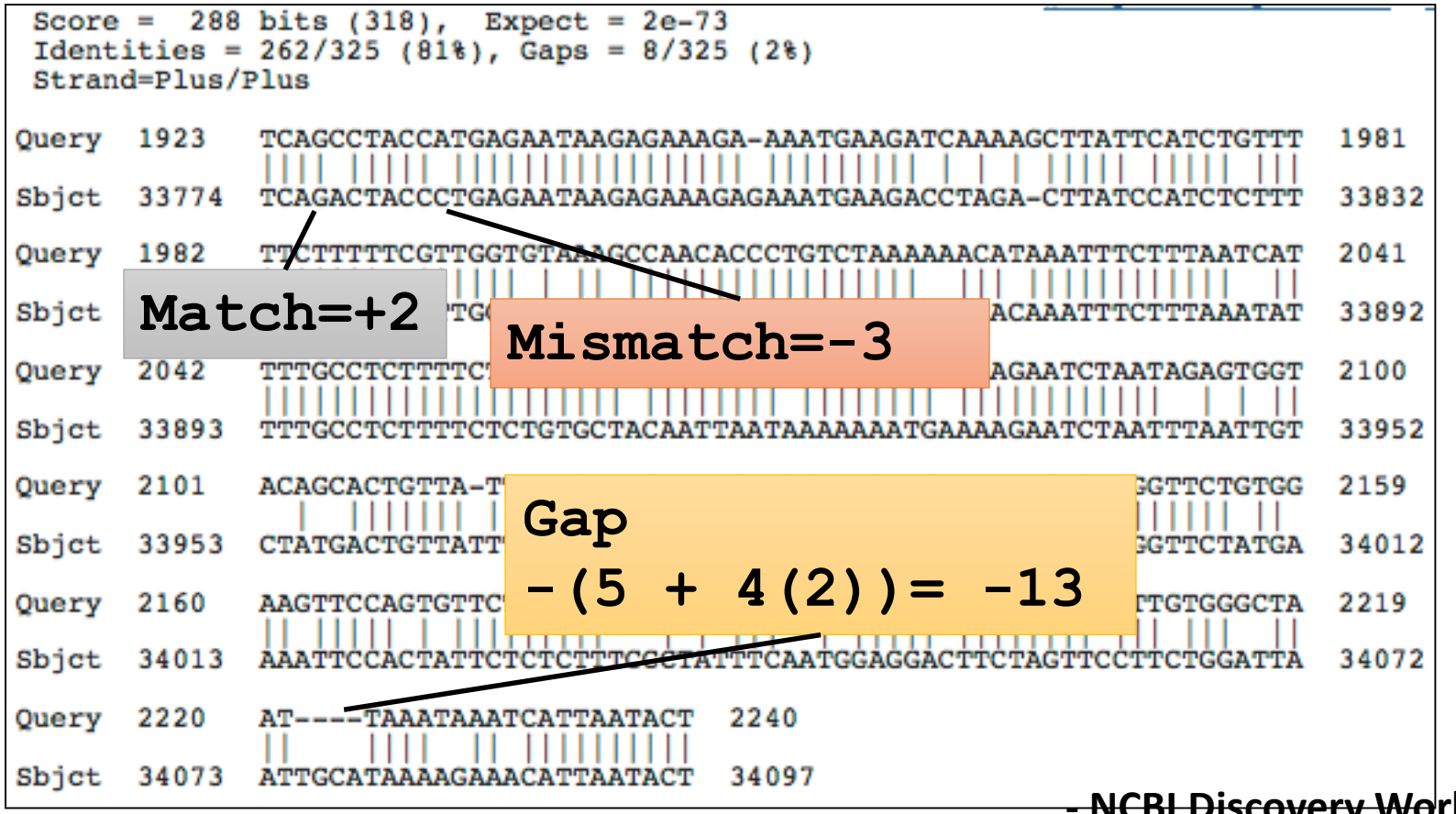


BLAST

Step 3: Scoring

- Give each HSP a score, report the targets ranked by the score

Nucleotide:

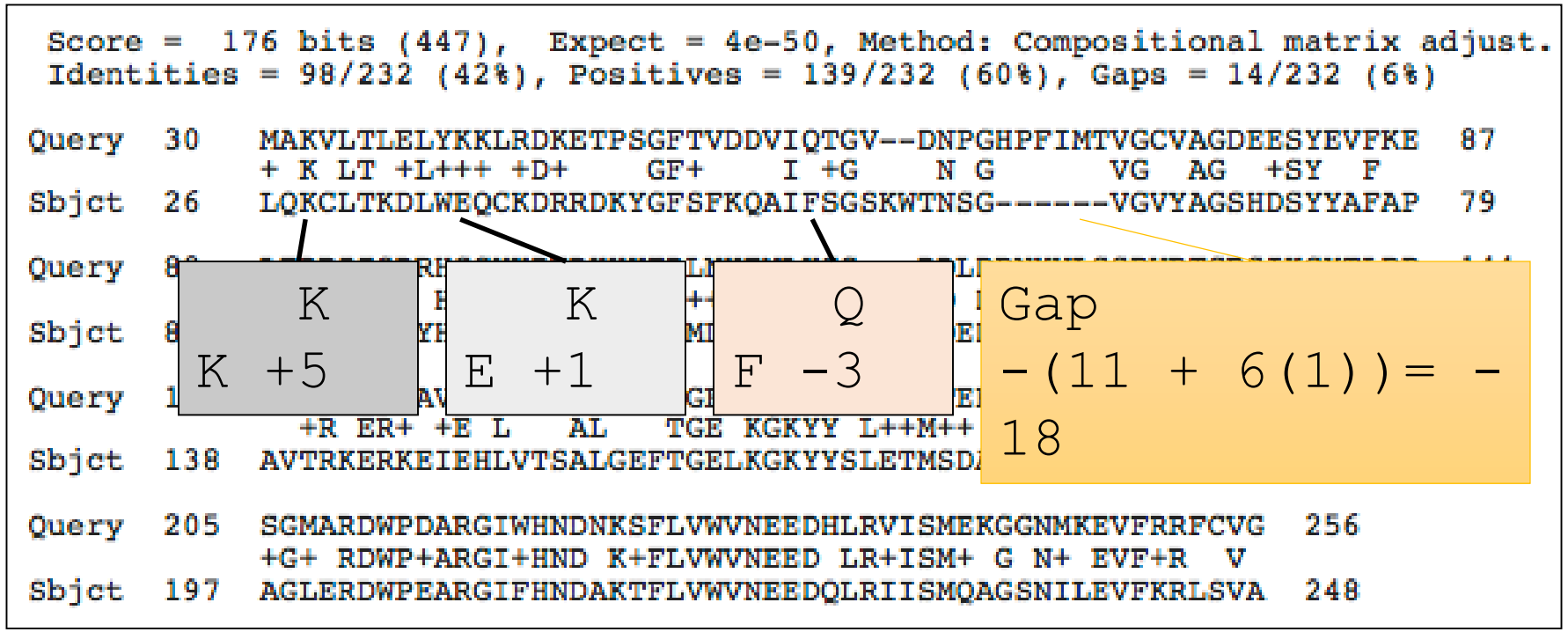


- NCBI Discovery Workshops

BLAST Step 3: Scoring

- Give each HSP a score

Protein



Scores from BLOSUM62, a position independent matrix

- NCBI Discovery Workshops

Scoring for protein alignment

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

BLAST statistics: from raw score to E-value

bit score: log transformed

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

E-value: p-value corrected for multiple-testing

$$Pval_S^{MSP} = Ke^{-\lambda S}$$
$$= Ke^{-\ln(2)S' + \ln(K}$$
$$= 2^{-S'}$$

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

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
```

* E-value 4e-50: Number of Chance Alignments = 4 X 10^{-50}

Local vs **Global** Alignment

BLAST: Basic **Local** Alignment Search Tool

Query: **ACGGTGAGGTGTCCGAGAGAGCT**

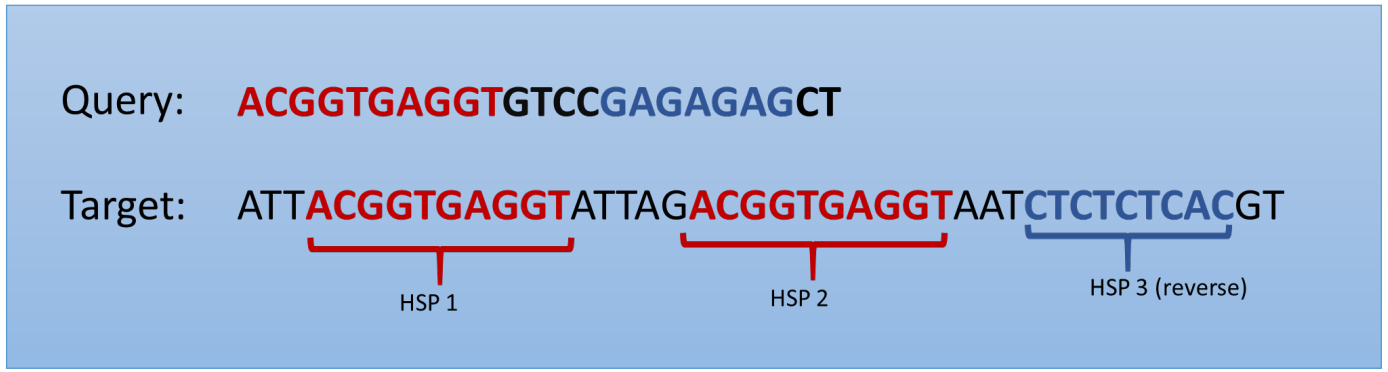
Target: ATT**ACGGTGAGGT**ATTAG**ACGGTGAGGT**AAT**CTCTCTCAC**GT

HSP 1

HSP 2

HSP 3 (reverse)

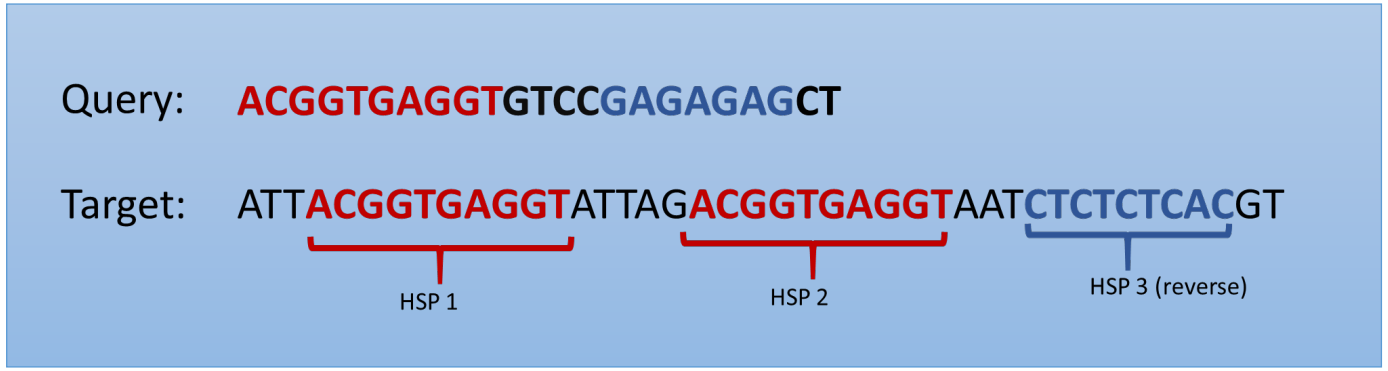
Local alignment results



3 HSPs in this target:

HSP1	ACGGTGAGGT |||||||||| ACGGTGAGGT	Forward
HSP2	ACGGTGAGGT |||||||||| ACGGTGAGGT	Forward
HSP3	GAGAGAG ||||||| GAGAGAG	reverse

Global alignment results



Global alignment:

```
---ACGGTGAGGT--------GT--------CCGAGAGAGCT
   ||||||||||          ||        |       | |
ATTACGGTGAGGTATTAGACGGTGAGGTAATCTCTCTCACGT
```

BLAST is a package including the following tools

command	Query	Hit database
blastn	nucleotide	nucleotide
blastp	protein	protein
blastx *	nucleotide	protein
tblastn *	protein	nucleotide
tblastx *	nucleotide	nucleotide

* Do 6-frame of the query, hit or both

Run BLAST on your local computer

(Windows, Mac, Linux)

For example, you just finish a genome assembly, and it is not available on NCBI web site yet.

```
# make a blast database from the genome sequence fasta file
makeblastdb -in myGenome.fasta -dbtype nucl


#run blast (do 6-frame translation of hits in the database), write results into a file
tblastn -query myProtein.fasta  -db myGenome.fasta -out result
```

Some useful parameters when running BLAST

https://www.ncbi.nlm.nih.gov/books/NBK279684/

-num_threads	Number of CPU threads to be used (e.g. 8)
-evalue	E-value cutoff (e.g. 1e-10)
-max_target_seqs	Maximum number of targets, e.g. 10
-max_hsps	Maximum number of HSPs per hit

-outfmt Output file format

-outfmt 5	xml format
-outfmt 6	tab-delimited (12 standard columns) qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore
-outfmt "6 std stitle staxids"	tab-delimited (12 standard columns + 2 extra columns) • Hit description • Hit taxonomy ID *

* Only if taxonomy is in the blast database. The blast database you download from NCBI contains taxonomy information

Short query sequences:

-task "blastn-short" Query <30 nucleotides

-task "blastp-short" Query <30 aa resides

Parallel Computing when running BLAST

Query: 100k protein sequences

DB: NCBI Genbank

```
>NP_001014992.2 inositol 1,4,5-triphosphate kinase [Apis mellifera]
MSRSINMDQEKKNNVENLKSGGSTTPASPTLSTPPTLNLMEQILLAKIEKQNLHESDDLHESDGRVGGKRRNILLRRTDS
MDSQNSASTYNSFLSSDSASSGNVYCKCDDCLLGIVDDYQRNPSVVGRKKSSGWRKLRNIVHWTPFFQTYKKQRYPWVQL
AGHQGNFRAGPTPGTILKKLCPQEEACFRLLMNDILRPYVPEFKGVLDVKDVEEGNVEETNSEETHQKDGSSDSVIKRTV
VSSYLQLQDLLGDFEHPCVMDCKVGVRTYLESELAKAKERPKLRKDMYEKMVQVDPTAPNAEERRVQGVTKPRYMVWRET
ISSTATLGFRVEGIKLAHGGSSKDFKTTRTREQVTEALRRFVEGYPHAVPKYIQRLKAIRATLKASPFFASHEVVGSSLL
FVHDTKNAGIWMIDFAKTLPLPQHLPRIHHDAEWKVGNHEDGYLIGVNNLIDIFQDIRNSEET
>NP_001014993.1 elongation factor 1-alpha [Apis mellifera]
MGKEKIHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAQEMGKGSFKYAWVLDKLKAERERGITIDIALWKF
ETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGTGEFEAGISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEP
PYSETRFEEIKKEVSSYIKKIGYNPAAVAFVPISGWHGDNMLEVSSKMPWFKGWTVERKEGKVEGKCLIEALDAILPPTR
PTDKALRLPLQDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPAGLTTEVKSVEM
>NP_001014994.1 glycerol-3-phosphate dehydrogenase [Apis mellifera]
MAEKLRICIVGSGNWGSTIAKIIGINAANFSNFEDRVTMYVYEEIINGKKLTEIINETHENVKYLPGHKLPPNIIAIPDV
VEAAKDADILTFVVPHQFIKRICSALFGKIKPTAIGLSLIKGFDKKQGGGIELISHIISKQLHIPVSVLMGANLASEVAN
EMFCETTIGCKDKNMAPILKDLMETSYFKVVVVEDVDSVECCGALKNIVACGAGFIDGLGLGDNTKAAVMRLGLMEIIKF
VNIFFPGGKKTTFFESCGVADLIATCYGGRNRKICEAFVKTGKKISELEKEMLNGQKLQGPFTAEEVNYMLKAKNMENRF
PLFTTVHRICIGETMPMELIENLRNHPEYIDETRNYQECKCSI
>NP_001019868.1 major royal jelly protein 9 precursor [Apis mellifera]
MSFNIWWLILYFSIVCQAKAHYSLRDFKANIFQVKYQWKYFDYNFGSDEKRQAAIQSGEYNYKNNVPIDVDRWNGKTFVT
ILRNDGVPSSLNVISNKIGNGGPLLEPYPNWSWAKNQNCSGITSVYRIAIDEWDRLWVLDNGISGETSVCPSQIVVFDLK
NSKLLKQVKIPHDIAINSTTGKRNVVTPIVQSFDYNNTWVYIADVEGYALIIYNNADDSFQRLTSSTFVYDPRYTKYTIN
DESFSLQDGILGMALSHKTQNLYYSAMSSHNLNYVNTKQFTQGKFQANDIQYQGASDILWTQASAKAISETGALFFGLVS
DTALGCWNENRPLKRRNIEIVAKNNDTLQFISGIKIIKQISSNIYERQNNEYIWIVSNKYQKIANGDLNFNEVNFRILNA
PVNQLIRYTRCENPKTNFFSIFL
>NP_001027532.1 follistatin-like 5 [Apis mellifera]
MRCMLEIAARSFLLLSIASTYVVSVAGYKHSRRHRDFTVAESYDASSSNSDSLSMTIPPSIDRSSIHEESYLAESSRSID
PCASKYCGIGKECELSPNSTIAVCVCMRKCPRRHRPVCASNGKIYANHCELHRAACHSGSSLTKSRLMRCLHHDIENAHI
RRTLHMNRTSLKTSKIVSYPKSRSRKKGGLKDNLIPDKNDPDSKECSNQEYEIMKDNLLLYNHARLMSQDNHSKEYLVSI
MFSHYDRNNNGNLEREELEQFAENEDLEELCRGCNLGHMISYDDTDGDGKLNVNEFYMAFSKLYSVSVVSLDKSLEVNHI
SARVGDNVEIKCDVTGTPPPPLVWRRNGADLETLNEPEIRVFNDGSLYLTKVQLIHAGNYTCHAVRNQDVVQTHVLTIHT
IPEVKVTPRFQAKRLKEEANIRCHVAGEPLPRVQWLKNDEALNHDQPDKYDLIGNGTKLIIKNVDYADTGAYMCQASSIG
GITRDISSLVVQEQPTPTTESEERRFFSFHQWGILVYEPSACRPRHEIRSTDVIPGTQEHVCGVKGIPCSWGRAINVANR
IGGLQHPGAVVWFTVSLH
>NP_001032395.1 putative tyramine receptor [Apis mellifera]
MANQTANYYGDVYQWNHTVSSGERDTRTEYYLPNWTDLVLAGLFTMLIIVTIVGNTLVIAAVITTRRLRSVTNCFVSSLA
AADLLVGLAVMPPAVLLQLTGGTWELGPMLCDSWVSLDILLCTASILSLCAISIDRYLAVTQPLIYSRRRRSKRLAGLMI
VAVWVLAGAITSPPLLGCFPRATNRDIKKCSYNMDSSYVIFSAMGSFFLPMLVMLYVYGRISCVIASRHRNLEATESENV
RPRRNVLIERAKSIRARRTECVTNSVTCDRPSDEAEPSSTSKKSGIVRSHQQSCINRVARETKTAGTLAVVVGGFVACWL
PFFILYLATPFVPVEPPDILMPALTWLGWINSAINPFIYAFYSADFRLAFWRLTCRKCFKSRTNLDPSNRKLPAPANWKK
DTTRT
```

• • •

Run 1 blast job with 64 threads

```
blastp -num_threads 64 -query input.fasta -db swissprot
```

Run 8 blast jobs in **parallel**, 8 threads per job

```
cat input.fasta | \
parallel -j 8  \
--blocks 10k \
--recstart '>' \
--pipe blastp -num_threads 8 -outfmt 6 -db swissprot -query - \
> combined_results.txt
```

* Parallel allows you parallelize the job through multiple computer servers

The BLAST algorithm was published in 1990*, Sequence alignment has developed since then.

Step 1: Seeding: Identify candidates for alignment

Step 2: Alignment: Do sequence alignment;

Step 3: Scoring: HSP scores

* Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410.

Step 1: Seeding

How to reduce number of candidate matches?

1. Increase word size, e.g. megablast

   Blastn: 11 bp

   Megablast: 28 bp

2. Using multiple spaced seeds, e.g. DIAMOND

Seed: *** QPG *** ALD ***

DIAMOND 100x - 20,000x speed of BLAST

http://www.diamondsearch.org/

```
diamond makedb --in nr.faa -d nr

diamond blastx -d nr -q reads.fna -o matches.m8
```

- blastp and blastx only;
- Run on AVX2 machines

Sensitivity of Diamond

--mid-sensitive (default): >70% identity

--sensitive: >40%

--more-sensitive:

--very-sensitive: <40%

--ultra-sensitive

Is 70% good enough?

- DIAMOND is so fast, you can use a super large database;
- If the database is big enough to include all species, you would more likely find a hit that >70%

*On biohpc, there is a pre-indexed file for UniRef90 /shared_data/genome_db/uniref90.dmnd)

Step 2 & 3: alignment and scoring

Improving accuracy

- Smith-Waterman (slow but more accurate alignment algorithm)
- Position weighted alignment matrix

Why weighted by position?

```
 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    MAKVLTLELYKKLRDE                QTGV--DNPGHPFIMTVGCVAGDEESYEVFKE  87
             + K LT +L+++ +D+                +G    N G        VG  AG  +GY  F
Sbjct  26    LQKCLTKDLWEQCKDRRDKIGPSFKQAIFSGSKWTNSG---          YYAFAP  79

Query  88    LFDPIISDIHCGYKPTDIHITDLNHENLKGG---DILDPNYVLSSRVRTGRSIKGYTLPP  144
               D II   HC +KP+DIH + ++++ L       I D   + S+R+R  R++    L
Sbjct  80    FMDKIIEAIHC-HKPSDIHISSMDYKQLNCPPFIAIED-KMINSTRIRVARNLAADPLGT  137

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

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

Critial H

Critial D

Blossom62 matrix:

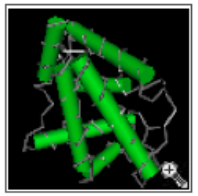
H->H +8

D->A -2

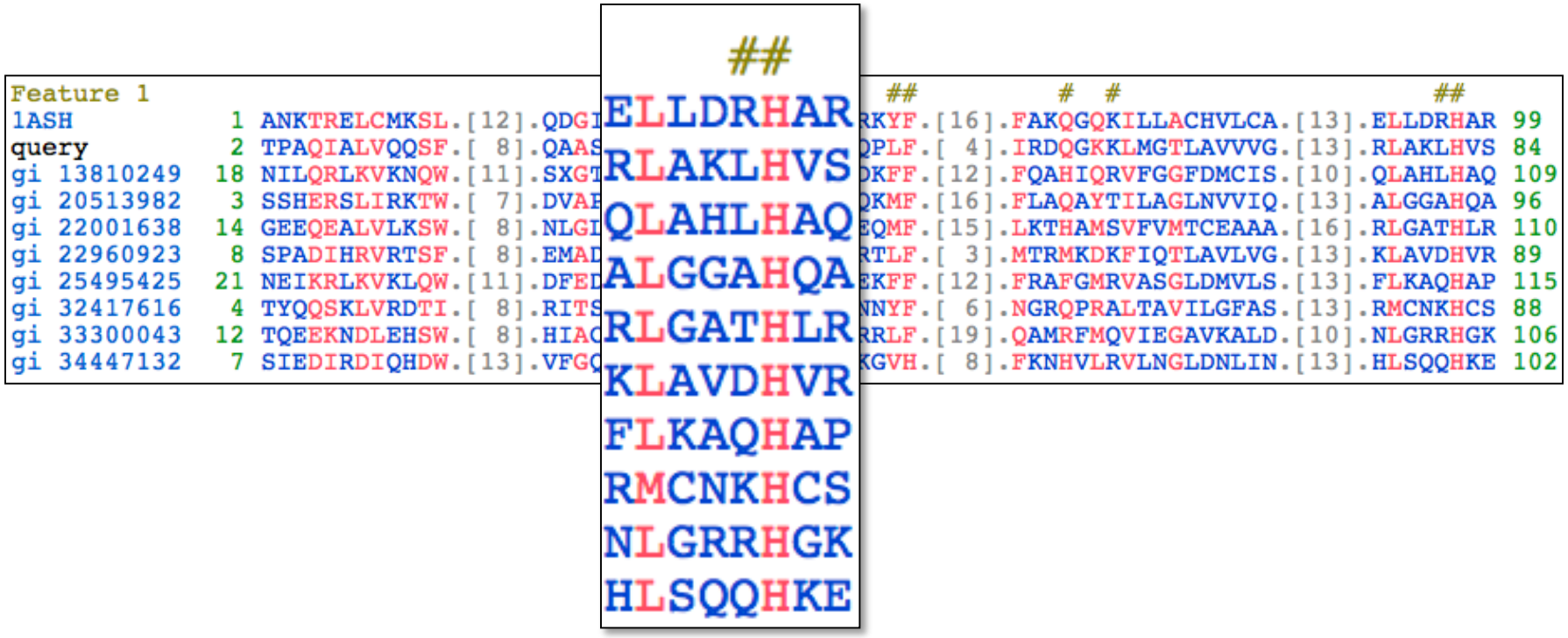
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1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

PSSM (Position-Specific Scoring Matrix):

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

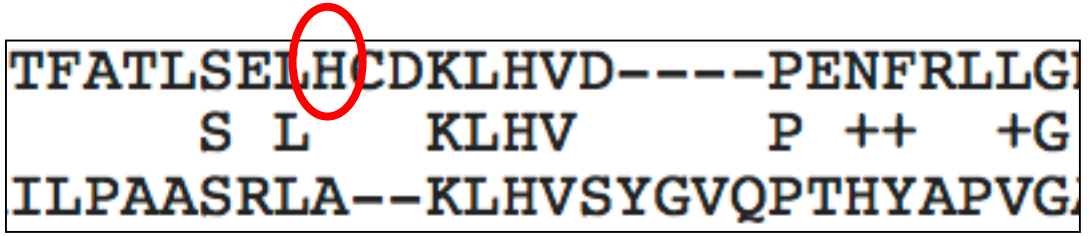
Using PSSM improve the BLAST accuracy

Critial **H**

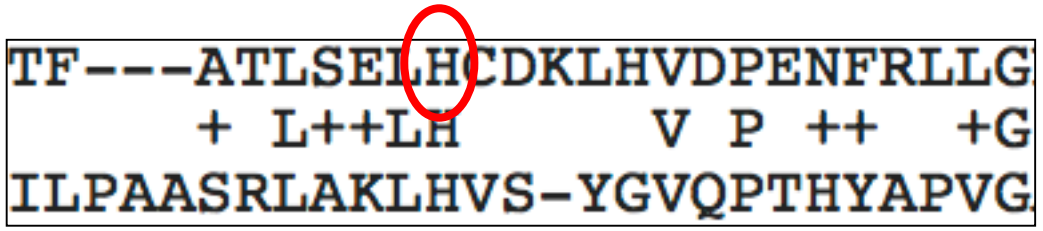
Query: TFATLSELHCDKLHVDPENFRLLG

DB: genome of a distant species

Best BLAST hit



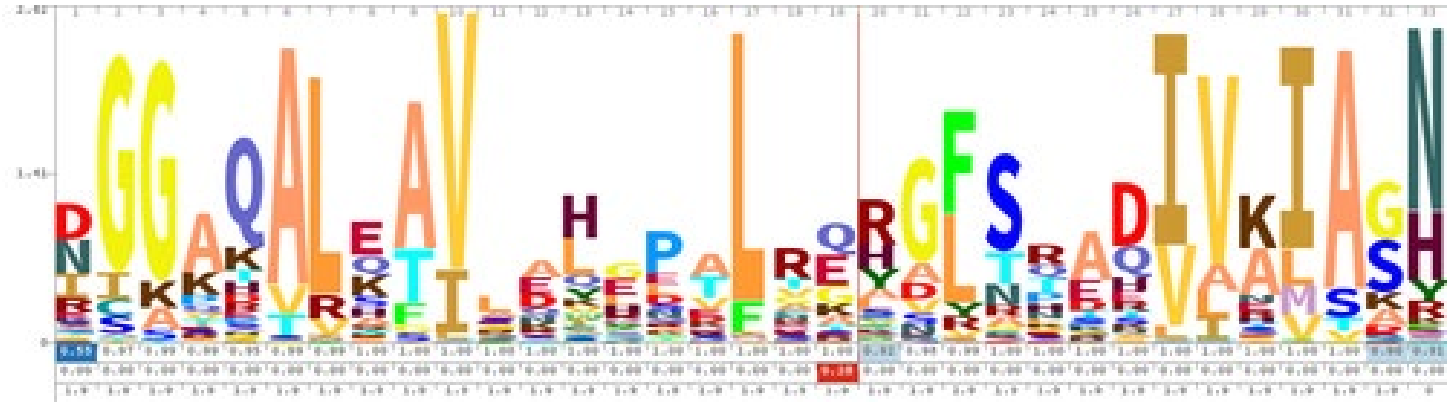
Best DELTA-BLAST hit (using PSSM)



BLAST is not reliable between distantly related species

NCBI tools that implement PSSM

PSI-BLAST
&
CD-Search

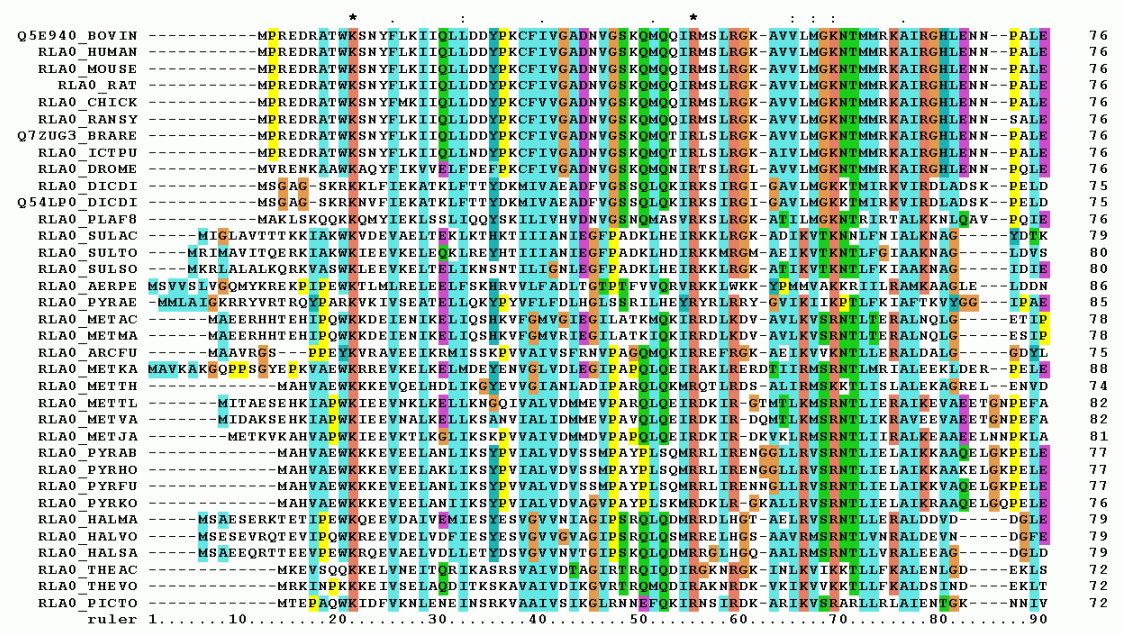


PSI-BLAST: Custom-made PSSM **(Removed from NCBI web site now)**

CD-Search: Pre-constructed PSSM (NCBI **Conserved Domain Database, CDD)**

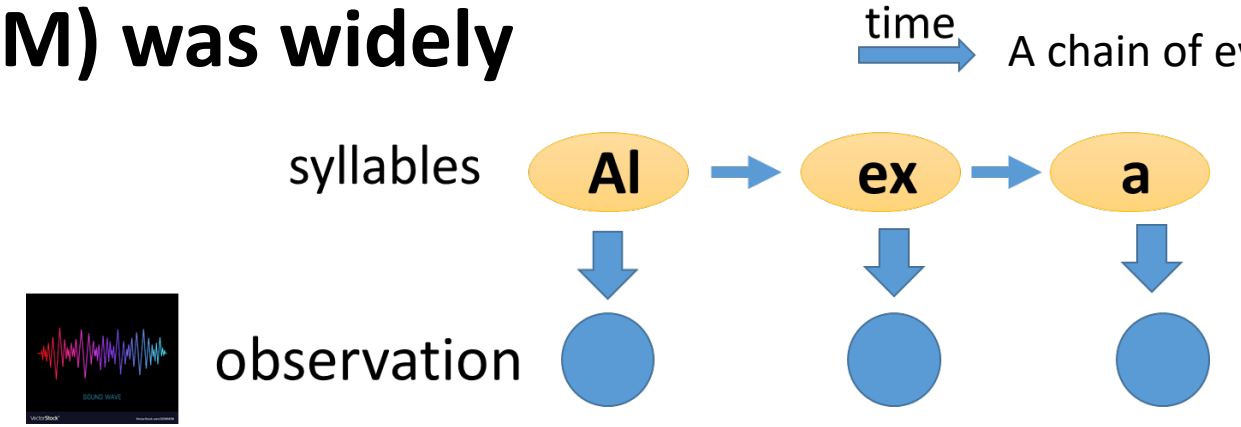
Hidden Markov Model

HMMs are trained from a multiple sequence alignment



Pre-constructed HMM database: **PFAM**

Hidden Markov Model (HMM) was widely used for voice recognition

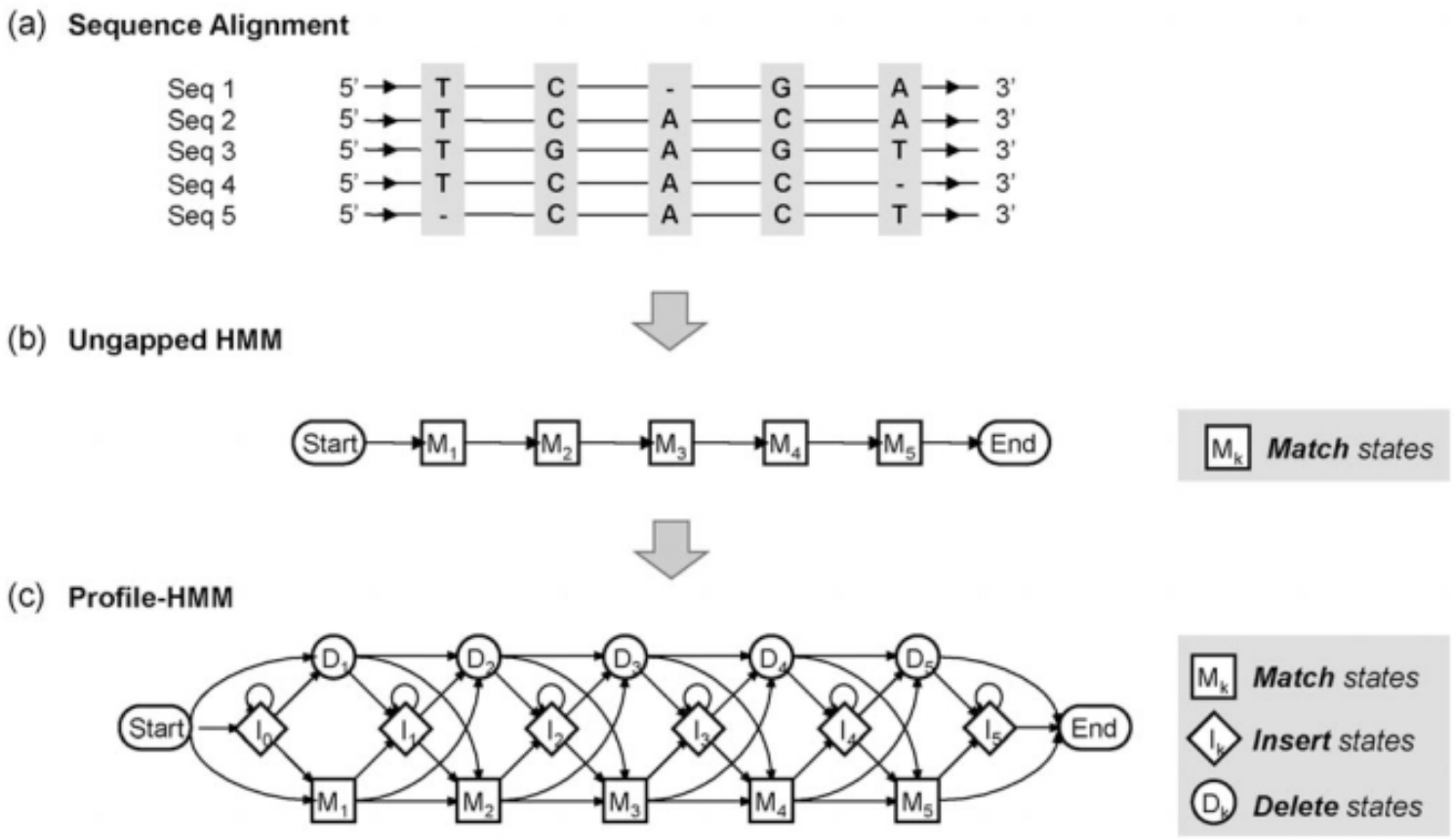


You can train an HMM model with real world data.
These are the elements in the model:

- hidden **states**: syllables in a given language
- **Transition probability** between two states
- Observed **symbols**: wave patterns
- **Emission probability**: probability of symbols of each state

Viberbi algorithm: match a new series with the model

HMM is good for model sequence profile of a domain



Three hidden states:

M: match

I: insertion

D: deletion

Observations symbols

A C G T

Transition and emission probability: trained from sequence alignment

Applications of HMM

- Given a new protein, identify protein domains;
- Find all genes that contains a domain;
- Multiple sequence alignment

Should I use **BLAST** or **HMM**?

BLAST:

Between closely related species;

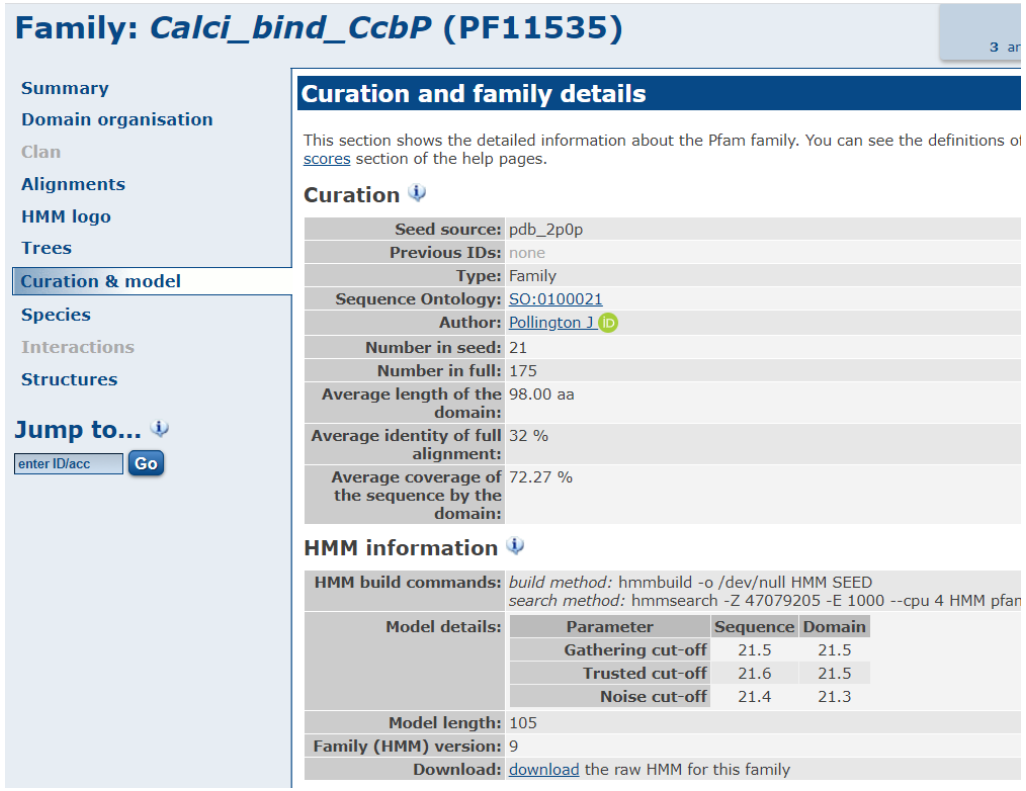
HMM:

Between distantly related species;

Pre-built HMM domains can be downloaded from PFAM web site

https://pfam.xfam.org/

Download a domain from the web site



Family: *Calci_bind_CcbP* (PF11535)

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Curation and family details

This section shows the detailed information about the Pfam family. You can see the definitions of scores section of the help pages.

Curation

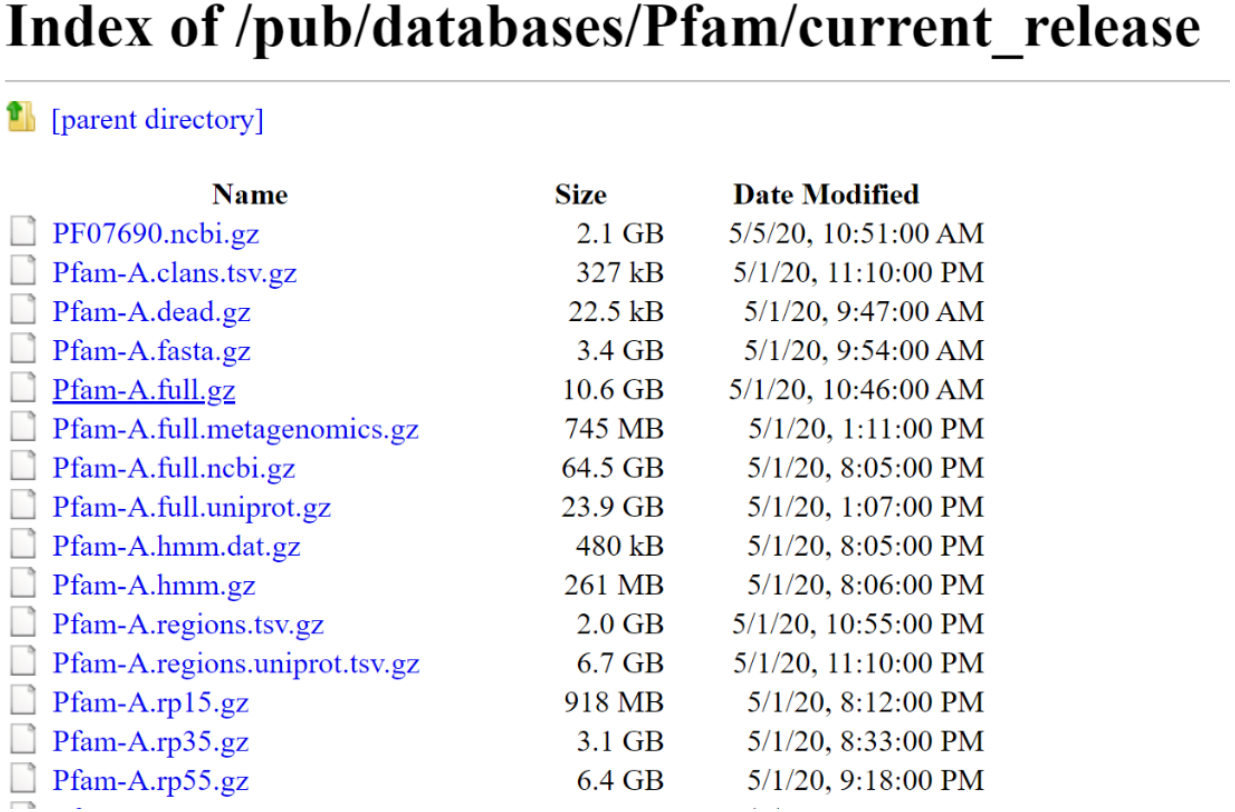
Seed source:	pdb_2p0p
Previous IDs:	none
Type:	Family
Sequence Ontology:	SO:0100021
Author:	Pollington J
Number in seed:	21
Number in full:	175
Average length of the domain:	98.00 aa
Average identity of full alignment:	32 %
Average coverage of the sequence by the domain:	72.27 %

HMM information

HMM build commands:	*build method:* hmmbuild -o /dev/null HMM SEED *search method:* hmmsearch -Z 47079205 -E 1000 --cpu 4 HMM pfan		
Model details:	Parameter	Sequence	Domain
	Gathering cut-off	21.5	21.5
	Trusted cut-off	21.6	21.5
	Noise cut-off	21.4	21.3
Model length:	105		
Family (HMM) version:	9		
Download:	download the raw HMM for this family		

Download all 18259 entries from the FTP site

Pfam_A_full.gz



Index of /pub/databases/Pfam/current_release

[parent directory]

Name	Size	Date Modified
PF07690.ncbi.gz	2.1 GB	5/5/20, 10:51:00 AM
Pfam-A.clans.tsv.gz	327 kB	5/1/20, 11:10:00 PM
Pfam-A.dead.gz	22.5 kB	5/1/20, 9:47:00 AM
Pfam-A.fasta.gz	3.4 GB	5/1/20, 9:54:00 AM
Pfam-A.full.gz	10.6 GB	5/1/20, 10:46:00 AM
Pfam-A.full.metagenomics.gz	745 MB	5/1/20, 1:11:00 PM
Pfam-A.full.ncbi.gz	64.5 GB	5/1/20, 8:05:00 PM
Pfam-A.full.uniprot.gz	23.9 GB	5/1/20, 1:07:00 PM
Pfam-A.hmm.dat.gz	480 kB	5/1/20, 8:05:00 PM
Pfam-A.hmm.gz	261 MB	5/1/20, 8:06:00 PM
Pfam-A.regions.tsv.gz	2.0 GB	5/1/20, 10:55:00 PM
Pfam-A.regions.uniprot.tsv.gz	6.7 GB	5/1/20, 11:10:00 PM
Pfam-A.rp15.gz	918 MB	5/1/20, 8:12:00 PM
Pfam-A.rp35.gz	3.1 GB	5/1/20, 8:33:00 PM
Pfam-A.rp55.gz	6.4 GB	5/1/20, 9:18:00 PM

Two major software packages for HMM

HMMER

Protein-HMM alignment

https://pfam.xfam.org/

HH-Suite:

Protein-HMM alignment or HMM-HMM alignments.

https://toolkit.tuebingen.mpg.de/tools/hhblits

Identify domains from an unknown protein.

Use the web sites if you only have small number of proteins.

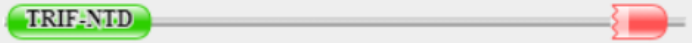
PFAM Search

https://pfam.xfam.org/

Sequence search results

Show the detailed description of this results page.

We found **3** Pfam-A matches to your search sequence (**2** significant and **1** insignificant)



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					
TRIF-NTD	TRIF N-terminal domain	Domain	n/a	4	157	5	156	**2**	**156**	157	148.0	2.2e-43	n/a	Show
RHIM	RIP homotypic interaction motif	Domain	n/a	659	717	677	716	**13**	**51**	52	43.1	4.5e-11	n/a	Show

Insignificant 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					
TIR_2	TIR domain	Domain	CL0173	442	567	443	540	**2**	**96**	118	26.7	7.0e-06	n/a	Show

Command line tools (HMMER):

hmmbuild

```
hmmbuild Pfam-A.hmm mySeqs.msa
```

Build an HMM profile from a multiple sequence alignment file

hmmsearch

Search an HMM profile against a sequence database

hmmscan

Search a sequence file against an HMM database

Comparing **hmmsearch** and **hmmscan**

	hmmsearch	hmmscan
Query	HMM file	Sequence file
Target	Sequence file	HMM file

Commands

```
hmmsearch my.hmm myProteins.fasta
```

```
hmmscan my.hmm myProteins.fasta
```

HMMscan: for small number of proteins;

HMMsearch: for large number of protines;

HH-Suite another HMM package

HH stands for “HMM” to “HMM”. While HMMER does protein to HMM matches, HH-Suite does HMM to HMM matches

---

Two search tools in HH-Suite:

HHsearch

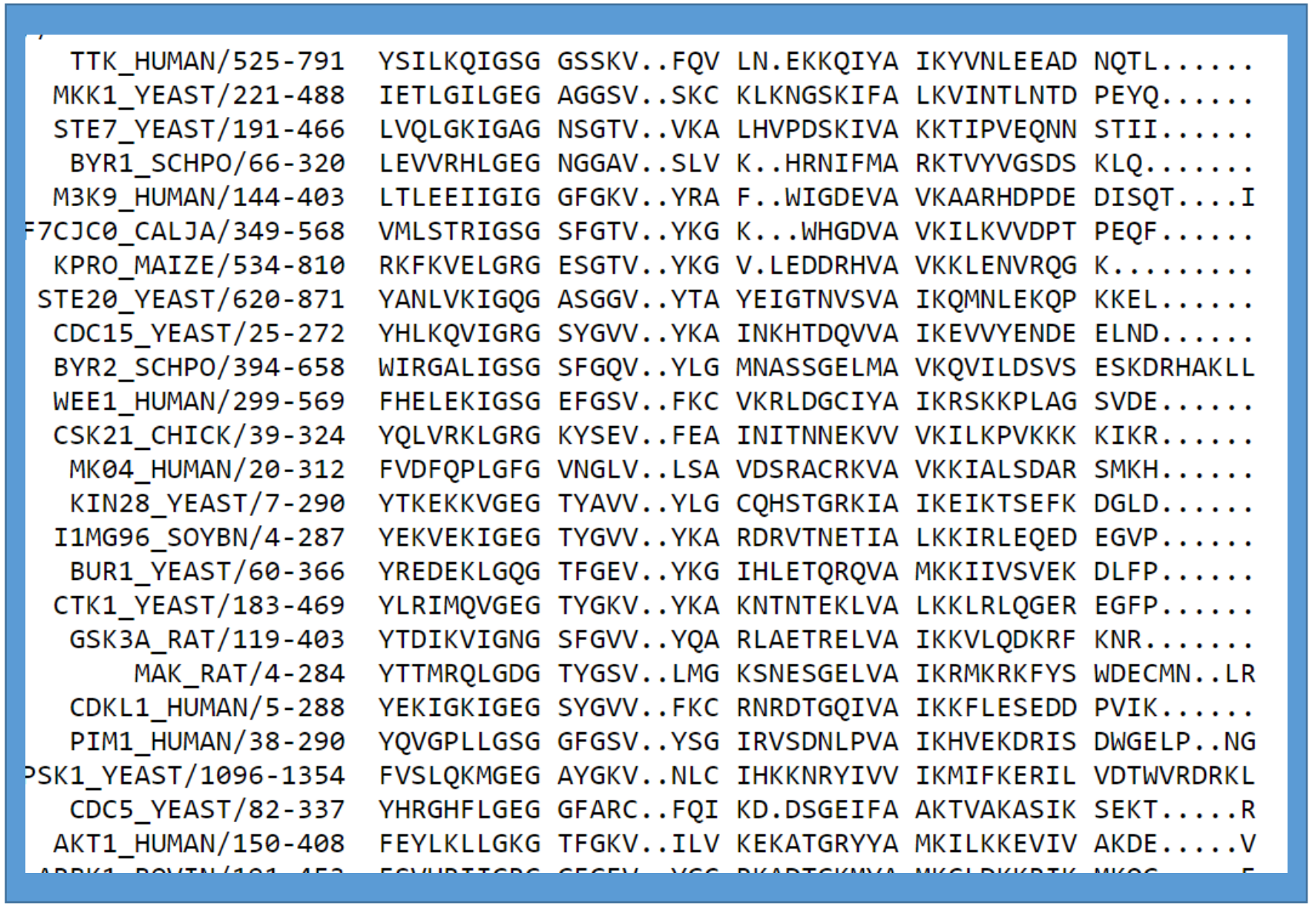
```
hhsearch -i query.a3m -d scop70_1.75_hhm_db
```

HHblits

```
hhblits -i query.a3m -d scop70_1.75_hhm_db
```

* HHblits has an extra prefitering step to reduce the number of alignments.

Multiple Sequence Alignment (MSA)



```
  TTK_HUMAN/525-791  YSILKQIGSG GSSKV..FQV LN.EKKQIYA IKYVNLEEAD NQTL......
 MKK1_YEAST/221-488  IETLGILGEG AGGSV..SKC KLKNGSKIFA LKVINTLNTD PEYQ......
 STE7_YEAST/191-466  LVQLGKIGAG NSGTV..VKA LHVPDSKIVA KKTIPVEQNN STII......
  BYR1_SCHPO/66-320  LEVVRHLGEG NGGAV..SLV K..HRNIFMA RKTVYVGSDS KLQ.......
 M3K9_HUMAN/144-403  LTLEEIIGIG GFGKV..YRA F..WIGDEVA VKAARHDPDE DISQT....I
F7CJC0_CALJA/349-568 VMLSTRIGSG SFGTV..YKG K...WHGDVA VKILKVVDPT PEQF......
 KPRO_MAIZE/534-810  RKFKVELGRG ESGTV..YKG V.LEDDRHVA VKKLENVRQG K.........
STE20_YEAST/620-871  YANLVKIGQG ASGGV..YTA YEIGTNVSVA IKQMNLEKQP KKEL......
 CDC15_YEAST/25-272  YHLKQVIGRG SYGVV..YKA INKHTDQVVA IKEVVYENDE ELND......
 BYR2_SCHPO/394-658  WIRGALIGSG SFGQV..YLG MNASSGELMA VKQVILDSVS ESKDRHAKLL
 WEE1_HUMAN/299-569  FHELEKIGSG EFGSV..FKC VKRLDGCIYA IKRSKKPLAG SVDE......
 CSK21_CHICK/39-324  YQLVRKLGRG KYSEV..FEA INITNNEKVV VKILKPVKKK KIKR......
  MK04_HUMAN/20-312  FVDFQPLGFG VNGLV..LSA VDSRACRKVA VKKIALSDAR SMKH......
  KIN28_YEAST/7-290  YTKEKKVGEG TYAVV..YLG CQHSTGRKIA IKEIKTSEFK DGLD......
 I1MG96_SOYBN/4-287  YEKVEKIGEG TYGVV..YKA RDRVTNETIA LKKIRLEQED EGVP......
  BUR1_YEAST/60-366  YREDEKLGQG TFGEV..YKG IHLETQRQVA MKKIIVSVEK DLFP......
 CTK1_YEAST/183-469  YLRIMQVGEG TYGKV..YKA KNTNTEKLVA LKKLRLQGER EGFP......
  GSK3A_RAT/119-403  YTDIKVIGNG SFGVV..YQA RLAETRELVA IKKVLQDKRF KNR.......
      MAK_RAT/4-284  YTTMRQLGDG TYGSV..LMG KSNESGELVA IKRMKRKFYS WDECMN..LR
  CDKL1_HUMAN/5-288  YEKIGKIGEG SYGVV..FKC RNRDTGQIVA IKKFLESEDD PVIK......
  PIM1_HUMAN/38-290  YQVGPLLGSG GFGSV..YSG IRVSDNLPVA IKHVEKDRIS DWGELP..NG
PSK1_YEAST/1096-1354 FVSLQKMGEG AYGKV..NLC IHKKNRYIVV IKMIFKERIL VDTWVRDRKL
  CDC5_YEAST/82-337  YHRGHFLGEG GFARC..FQI KD.DSGEIFA AKTVAKASIK SEKT.....R
 AKT1_HUMAN/150-408  FEYLKLLGKG TFGKV..ILV KEKATGRYYA MKILKKEVIV AKDE.....V
```

Global alignment is forced for MSA

Seq1: **ACGGTGAGGTGTCCGAGAGAGCT**

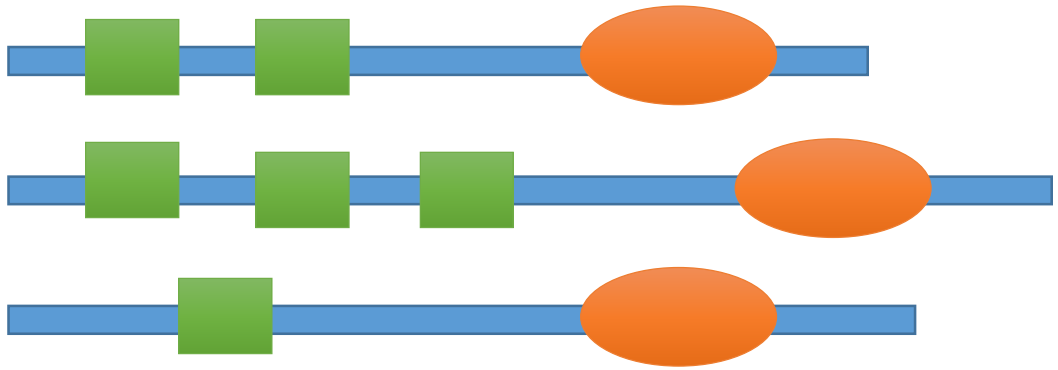
Seq2: ATT**ACGGTGAGGT**ATTAG**ACGGTGAGGT**AAT**CTCTCTCAC**GT

Global alignment:

```
---ACGGTGAGGT---------GT--------CCGAGAGAGCT
   ||||||||||          ||        |       | |
ATTACGGTGAGGTATTAGACGGTGAGGTAATCTCTCTCACGT
```

Difficulty in MSA

Protein families with multiple domains



Deletions

GGAC AA T AA TT

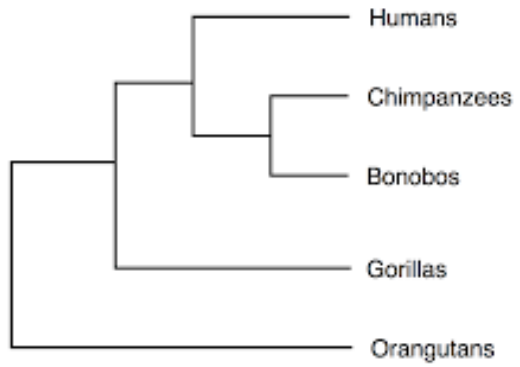
GGAC AA G AA TT

GGAC AA TT

MSA on individual domains

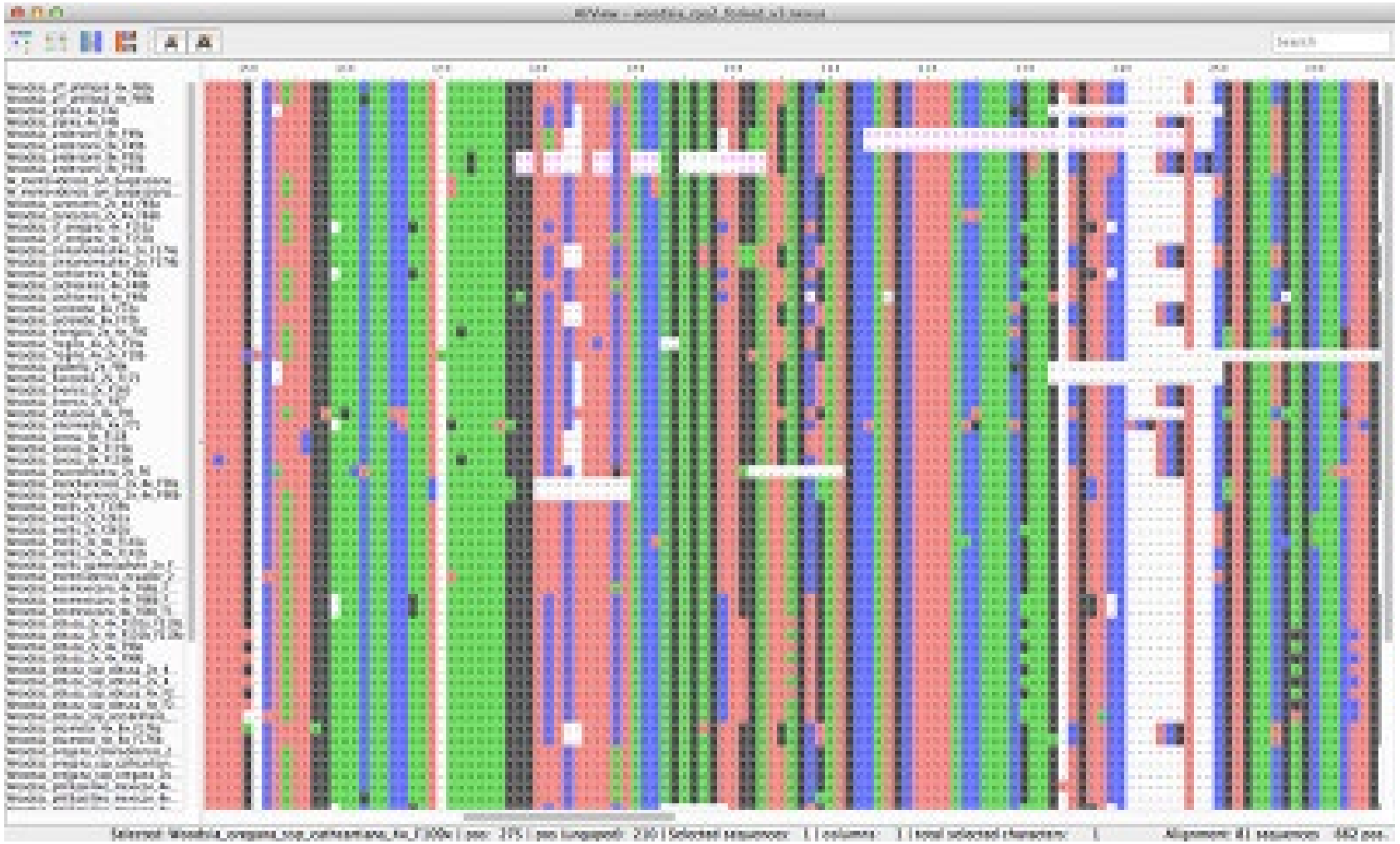
Trim gaps in MSA

Popular MSA construction software

Progressive /iterative (Guided by a neighbor-joining tree)	Clustal Omega (replace ClustalW) MAFFT MUSCLE T-Coffee	
Codon alignment	PRANK	More accurate placement of insertions and deletions
HMM guided	HMMAlign	align sequences to an HMM profile

Alignment viewers and editors

AliView (https://ormbunkar.se/aliview/)

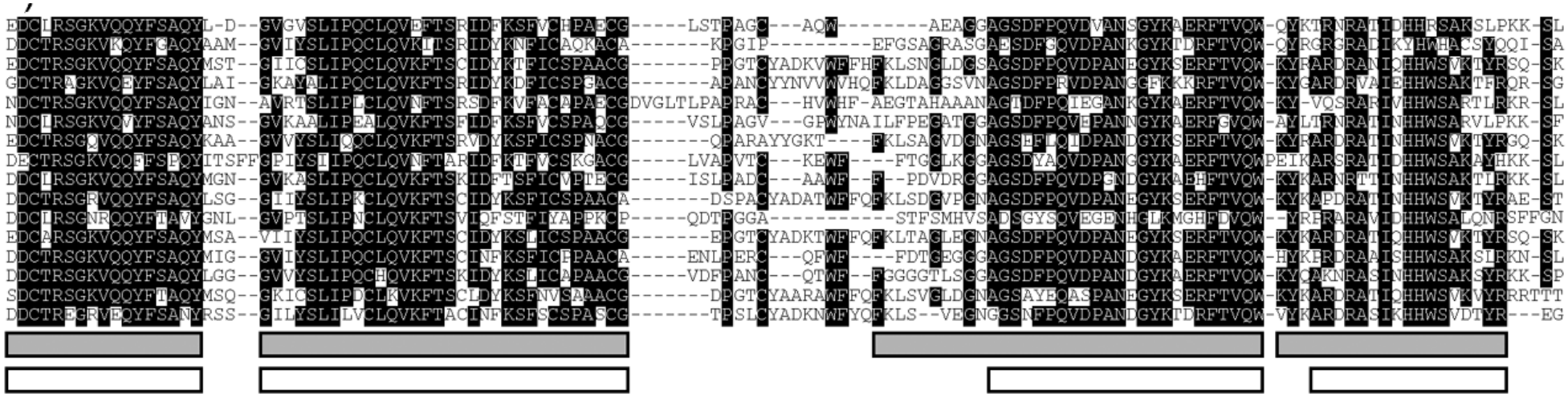


MSA Trimming software

Trim un-reliable regions of alignment:

1. Genome assembly/annotation errors. Commonly at 5’ and 3’ end of sequences;

2. Regions with too much variations, especially with insertion/deletions;



Gblocks (the grey and white box represent regions kept by Gblocks)

MSA File formats

clustal fasta nexus paup phylip selex a2m a3m

```
CLUSTAL W(1.83) multiple sequence alignment


IXI_234         TSPASIRPPAGPSSRPAMVSSRRTRPSPPGPRRPTGRPCCSAAPRRPQAT
IXI_235         TSPASIRPPAGPSSR---------RPSPPGPRRPTGRPCCSAAPRRPQAT
IXI_236         TSPASIRPPAGPSSRPAMVSSR--RPSPPPPRRPPGRPCCSAAPPRPQAT
IXI_237         TSPASLRPPAGPSSRPAMVSSRR-RPSPPGPRRPT----CSAAPRRPQAT


IXI_234         GGWKTCSGTCTTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSRSAG
IXI_235         GGWKTCSGTCTTSTSTRHRGRSGW----------RASRKSMRAACSRSAG
IXI_236         GGWKTCSGTCTTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSR--G
IXI_237         GGYKTCSGTCTTSTSTRHRGRSGYSARTTTAACLRASRKSMRAACSR--G


IXI_234         SRPNRFAPTLMSSCITSTTGPPAWAGDRSHE
IXI_235         SRPNRFAPTLMSSCITSTTGPPAWAGDRSHE
IXI_236         SRPPRFAPPLMSSCITSTTGPPPAGDRSHE
IXI_237         SRPNRFAPTLMSSCLTSTTGPPAYAGDRSHE
```

clustal

```
>1ymg_A  THE CHANNEL ARCHITE
SASFWRAICAEFFASLFYVFFGLGASLRW-----AG------P---------lHVLQVAL
AFGLALATLVQAVGHISGAHVNPAVTFAFLVGSQMSLLRAICYMVAQLLGAVAGAAVLYS
VT--PPAvRGNlALNTLHPGVSVGQATIVEIFLTLQFVLCIFATYDERRNGRLGSVALAV
GFSLTLGHLFGMYYTGAGMNPARSFAPAILTR------NFTNHWVYWVGPVIGAGLGSLL
YDFLLFPRLKSVSERLSILKG
>2d57_A  DOUBLE LAYERED 2D C
TQAFWKAVTAEFLAMLIFVLLSVGSTINW-----GG-SENPLP---------VDMVLISL
CFGLSIATMVQCFGHISGGHINPAVTVAMVCTRKISIAKSVFYITAQCLGAIIGAGILYL
VT--PPSVVGGLGVTTVHGNLTAGHGLLVELIITFQLVFTIFASCDSKRTDVTGSVALAI
GFSVAIGHLFAINYTGASMNPARSFGPAVIMG------NWENHWIYwVGPIIGAVLAGAL
YEYVF--------------CP
>2f2b_A  CRYSTAL STRUCTURE O
MVSLTKRCIAEFIGTFILVFFGAGSAAVTLMIASGGTSPNPFNIGIGLLGGLGDWVAIGL
AFGFAIAASIYALGNISGCHINPAVTIGLWSVKKFPGREVVPYIIAQLLGAAFGSFIFLQ
CAGIGAATVGGLGATAPFPGISYWQAMLAEVVGTFLLMITIMGIAvDERAP-KGFAGIII
GLTVAGIITTLGNISGSSLNPARTFGPYLNDMifagtDlWNYYSIYvIGPIVGAVLAALT
YQYL---------------TS
```

fasta

```
>d1a1x__ b.63.1.1 (-) p13-MTCP1 {Human (Homo sapiens)}
PPDHLWVHQEGIYRDEYQRTWVAVVEE..E..T..SF.........LR..........ARVQQIQVPLG.......DAARPSHLLTS.....QL
>gi|6678257|ref|NP_033363.1|:(7-103) T-cell lymphoma breakpoint 1 [Mus musculus]
HPNRLWIWEKHVYLDEFRRSWLPVVIK..S..N..EK.........FQ..........VILRQEDVTLG.......EAMSPSQLVPY.....EL
>gi|7305557|ref|NP_038800.1|:(8-103) T-cell leukemia/lymphoma 1B, 3 [Mus musculus]
PPRFLVCTRDDIYEDENGRQWVVAKVE..T..S..RSpygsrietcIT..........VHLQHMTTIPQ.......EPTPQQPINNN.....SL
>gi|11415028|ref|NP_068801.1|:(2-106) T-cell lymphoma-1; T-cell lymphoma-1A [Homo sapiens]
HPDRLWAWEKFVYLDEKQHAWLPLTIEikD..R..LQ.........LR..........VLLRREDVVLG.......RPMTPTQIGPS.....LL
>gi|7305561|ref|NP_038804.1|:(7-103) T-cell leukemia/lymphoma 1B, 5 [Mus musculus]
----------GIYEDEHHRVWIAVNVE..T..S..HS.........SHgnrietcvt.VHLQHMTTLPQ.......EPTPQQPINNN.....SL
>gi|7305553|ref|NP_038801.1|:(5-103) T-cell leukemia/lymphoma 1B, 1 [Mus musculus]
LPVYLVSVRLGIYEDEHHRVWIVANVE..TshS..SH.........GN..........RRRTHVTVHLW.......KLIPQQVIPFNplnydFL
>gi|27668591|ref|XP_234504.1|:(7-103) similar to Chain A, Crystal Structure Of Murine Tcl1
-PDRLWLWEKHVYLDEFRRSWLPIVIK..S..N..GK.........FQ..........VIMRQKDVILG.......DSMTPSQLVPY.....EL
>gi|27668589|ref|XP_234503.1|:(9-91) similar to T-cell leukemia/lymphoma 1B, 5;
-PHILTLRTHGIYEDEHHRLWVVLDLQ..A..ShlSF.........SN..........RLLIYLTVYLQqgvafplESTPPSPMNLN.....GL
>gi|7305559|ref|NP_038802.1|:(8-102) T-cell leukemia/lymphoma 1B, 4 [Mus musculus]
PPCFLVCTRDDIYEDEHGRQWVAAKVE..T..S..SH.........SPycskietcvtVHLWQMTTLFQ.......EPSPDSLKTFN.....FL
>gi|7305555|ref|NP_038803.1|:(9-102) T-cell leukemia/lymphoma 1B, 2 [Mus musculus]
---------PGFYEDEHHRLWMVAKLE..T..C..SH.........SPycnkietcvtVHLWQMTRYPQ.......EPAPYNPMNYN.....FL
```

A2m (hhsuite)

Most likely to work: Fasta formats

File format converting:

Online tool:

mview https://www.ebi.ac.uk/Tools/msa/mview/

Emboss/Seqret https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/

Phylogeny.fr http://phylogeny.lirmm.fr/phylo_cgi/data_converter.cgi

Command line tool:

HH-Suite: reformat.pl

EMBOSS/Seqret

BioPERL

BioPython

Any questions?

Email:
brc-bioinformatics@cornell.edu

Office hours
https://biohpc.cornell.edu/lab/office1.aspx