

# Concepts and tools for sequence alignment

Qi Sun

Bioinformatics Facility  
Cornell University

# How BLAST works?

## Query sequence

```
>unknown
MTAMEESQSDISLELPLSQETFSGLWKLPPEDILPSPHCMDDLLLHQDVEEFFEGPSEALRVSGAPAAQ
DPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSPPLNKLFQLAKTCPV
QLWVSATPPAGSRVRAMAIIKKSQHMTEVVRCPHHERCSGDGLAPPQHLIRVEGNLYPEYLEDRQTFR
HSVVPYEPPEAGSEYTTIHYKYMNCSSCMGGMNRPILTITLEDSSGNLLGRDSFEVRVCACPGRDRR
TEENFRKKEVLCPPEAKRALPTCTSAPPQKKPLDGEYFTLKIRGRKRKFEMFRELNEALELKDA
HATEESGDSRAHSSLQPRAFQALIKEESPNC
```

## NCBI BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

Clear      Query subrange [?](#)

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  Standard databases (nr etc.)  rRNA/ITS databases  Genomic + transcript databases  Nucleotide collection (nr/nt) [?](#)

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

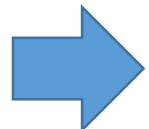
Limit to Optional  Sequences from type material

Entrez Query Optional  Enter an Entrez query to limit search [?](#) [YouTube](#) [Create custom database](#)

Program Selection

Optimize for  Highly similar sequences (megablast)  More dissimilar sequences (discontiguous megablast)  Somewhat similar sequences (blastn) Choose a BLAST algorithm [?](#)

BLAST  Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)  Show results in a new window



## BLAST Results

Descriptions      Graphic Summary      Alignments      Taxonomy

Download [?](#) Manage Columns [?](#) Show 100 [?](#)

Sequences producing significant alignments

GenPept      Graphics      Distance tree of results      Multiple alignment

	Description	Max Score	Total Score	Query Cover	E value	Per Ident	Accession
<input checked="" type="checkbox"/>	p53 [Mus musculus]	788	788	100%	0.0	100.00%	<a href="#">AAA39883.1</a>
<input checked="" type="checkbox"/>	transformation related protein p53 isoform CRA_a [Mus musculus]	785	785	100%	0.0	99.74%	<a href="#">EDL10505.1</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform b [Mus musculus]	784	784	100%	0.0	99.74%	<a href="#">NP_001120705.1</a>
<input checked="" type="checkbox"/>	transformation related protein p53 [Mus musculus]	778	778	99%	0.0	99.74%	<a href="#">CAI52015.1</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform X2 [Mus caroli]	760	760	100%	0.0	96.59%	<a href="#">XP_021032870.1</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform a [Mus musculus]	752	752	95%	0.0	99.73%	<a href="#">NP_035770.2</a>
<input checked="" type="checkbox"/>	transformation related protein p53 [Mus musculus]	752	752	95%	0.0	99.73%	<a href="#">AAK94783.1</a>
<input checked="" type="checkbox"/>	cellular tumour antigen p53 [Mus musculus]	751	751	95%	0.0	99.45%	<a href="#">CAA25420.1</a>
<input checked="" type="checkbox"/>	tumor suppressor p53 [Mus musculus]	750	750	95%	0.0	99.45%	<a href="#">AAC05704.1</a>
<input checked="" type="checkbox"/>	tumor suppressor p53 [Mus musculus]	749	749	95%	0.0	99.45%	<a href="#">AAD39535.1</a>
<input checked="" type="checkbox"/>	p53 [Mus musculus domesticus]	749	749	95%	0.0	99.45%	<a href="#">AAA39882.1</a>
<input checked="" type="checkbox"/>	antigen p53 tumor [Murid betaherpesvirus 1]	746	746	95%	0.0	98.90%	<a href="#">1001197A</a>
<input checked="" type="checkbox"/>	transformation related protein p53 [Mus musculus]	746	746	94%	0.0	99.72%	<a href="#">CAI52016.1</a>
<input checked="" type="checkbox"/>	unnamed protein product [Mus musculus]	738	738	95%	0.0	98.63%	<a href="#">CAA25323.1</a>
<input checked="" type="checkbox"/>	unnamed protein product [Mus musculus]	736	736	93%	0.0	99.72%	<a href="#">BAE28156.1</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform X1 [Mus caroli]	728	728	95%	0.0	96.43%	<a href="#">XP_021032869.1</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform X2 [Mus pahari]	709	709	100%	0.0	90.65%	<a href="#">XP_029402294.1</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform X1 [Mus pahari]	676	676	95%	0.0	90.22%	<a href="#">XP_021068233.2</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform X2 [Grammomys surdaster]	675	675	100%	0.0	88.08%	<a href="#">XP_028635883.1</a>
<input checked="" type="checkbox"/>	PREDICTED: cellular tumor antigen p53 isoform X2 [Rattus norvegicus]	661	661	99%	0.0	87.47%	<a href="#">XP_008765995.1</a>
<input checked="" type="checkbox"/>	unnamed protein product [Mus musculus]	648	648	82%	0.0	99.68%	<a href="#">BAC37729.2</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 [Arvicinalis niloticus]	647	647	99%	0.0	82.55%	<a href="#">XP_034361430.1</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform X3 [Mastomys coucha]	644	644	100%	0.0	85.64%	<a href="#">XP_031207310.1</a>
<input checked="" type="checkbox"/>	cellular tumor antigen p53 isoform X1 [Grammomys surdaster]	643	643	99%	0.0	84.94%	<a href="#">XP_028635882.1</a>

# BLAST

## Step1: Seeding

- Break down the query sequences into words;

### The BLAST Search Algorithm

query word ( $W = 3$ )

Query: TGSQSLAALLNKCKTP**PQG**QRLVNQWIKQPLMDKNRIEERLNLEAFV

# BLAST

- Identify candidate targets by matching to the “word”

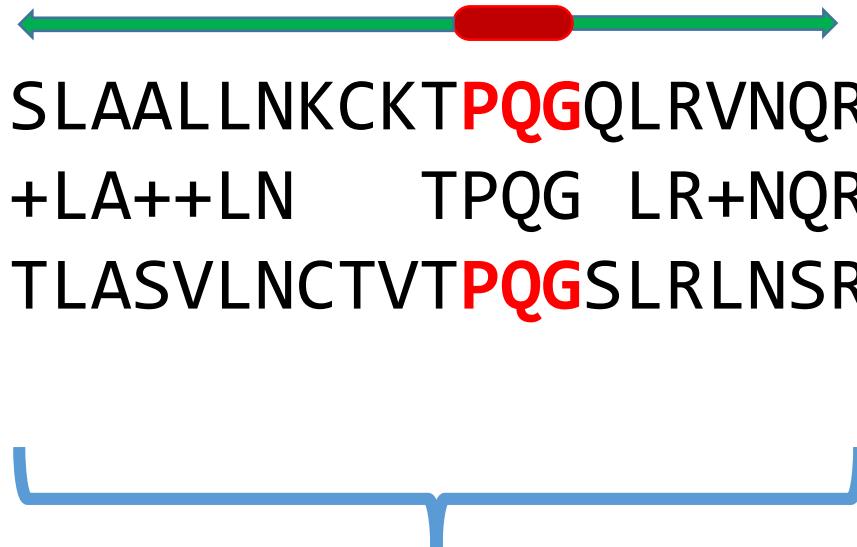
Seeding: **PQG**

MVNENTRMYIPEENHQGSNYGSPRPAHANMNANAAGLAPEHIPTPGAALSWQAIDAARQAKLMSAGN  
ATISTVSSTQRKRQYGPKKQGSTTATRPPRALCLTLKNPIRACISIVEWKPFEEIIILLTFANCVA  
LAIYIPFPEDDSNATNSNLERVEYLFLIIIFTVEAFLKVIAYGLLFHPNAYLRNGWNLLDFIIVVGLFSA  
ILEQATKADGANALGGKAGFDVKALRAFRVLRLGVPSLQVVLNSIIKAMVPLLHIALLVLFVII  
IYAIIGLELFMGKMHCTCYNQEGIADVPAAEDDPSPCALETGHGRQCQNGTVCKPGWDGPKHGITNFDNF  
FAMLTVCQICITMEGWTDVLYWVNDAVGRDWPIYFVTLLIIGSFVNLNLVGLVLSGEFSKEREKAKARGD  
FQKLREKQQLQLEDLKGYLDWITQAEDIDPENEDEGMDEEKPRNMSMPTSETESVNTENVAGGDIEGENCG  
ARLAHRISKSKFSRYWRWNRFCCRKCRAAVKSNVFYWLVLFLNTLTIASEHYNQPNWLTEVQDTAN  
KALLALFTAEMLLKMYSLGLQAYFVSLFNRFDCFVVCGGILETILVETKIMSPLGISVLCRVLLRIFKI  
TRYWNSLSNLVASLLNSVRSIASLLLLLFLFIIIFSLLGMQLFGGKFNFDEMOTRSTFDNFQSLLT  
QILTGEDWNSVMDGIPQGGSPFPGLMVCIFIIFLICGNYILLNVFLAIAVDNLADAESLTSQKEEEE  
EEKERKKLARTASPEKKQELVEKPAVGESKEEIKLSITADGESPATKINMDDLQPNENEDKSPYPNP  
ETTGEEDEEPEMPVGPRPRPLSELHLKEKA  
VPMPEASAFFIFSSNNRFRLQCHRIVNDTIFTNLILFFI  
LLSSISLAAEDPVQHTSFRNHILFYFDIVFTTIFTIEIALKILGNADYVFTSIFTLEIILKMTAYGAFLH  
KGSFCRNYFNILDLLVVSVSLSISFGIQSSAINVVKILRVLRLRAINRAKGLKHVVQCVFAIRTG  
NIVIVTLLQFMFACIGVQLFKGKLYTCSDSSKQTEAECKGNYITYKDGEVDHPIQPRSWENSKFDFDN  
VLAAMMALFTVSTFEGWP  
ELLYRSIDSHTEDKGP  
IYNYRVEISIFFIYIIIAFFMMNIFVGFVIVTFQ  
EQGEQEYKNC  
EELDKNQRC  
VEALKARPLRRYIPK  
NQHQYKVWV  
VNSTYFEYLMFV  
LILLNTICLAMQH  
YGSCLFKIAMNI  
LNMLFTGL  
VEMILK  
LIAFKPKGYF  
SDPWNV  
FDLIVIGSI  
IDVILSETN  
HYFCDA  
WNTFDAL  
LIVVGSIV  
DIAITE  
EVNPAE  
HTQC  
CPSMNA  
EENSRI  
SITFF  
RLFR  
MVLV  
KLLSR  
GEIGIR  
LLWT  
FIKS  
FQALPY  
VALLIV  
MLFFI  
YAVIG  
MQVFG  
KIALND  
TEINRN  
NNFQT  
FPQAV  
LLLFR  
CATGEA  
WQD  
LACMP  
GKKCAPE  
SEPSNST  
EGETPC  
GSSFAV  
FYFIS  
YMLCA  
FLIINLF  
VAVIMD  
NFDYLT  
RDWSI  
LGPH  
HLDEF  
KRIWA  
YDPEAK  
GRIKH  
LVDV  
TLLRRI  
QPPL  
FGKLC  
CPHR  
ACKRL  
VSMN  
PLNS  
SDGT  
VMF  
NATL  
FALV  
RTALRI  
KTEGN  
LEQANE  
ELRA  
IKKI  
WKR  
TSMK  
LDDQ  
VPPAG  
DEV  
TVG  
KFYAT  
FLI  
QEY  
FRKF  
KR  
KEQGL  
VGKPS  
QRNAL  
SQAGL  
RTHD  
DIGPE  
IRRA  
ISG  
DLTAE  
ELDK  
KAM  
KEAV  
SAASE  
DIFR  
RAGGL  
FGNH  
VSYY  
QSD  
GRSA  
FPQT  
TQR  
PLH  
INKAG  
SSQ  
GDTE  
PSHE  
KLV  
DST  
TPSY  
SST  
GSN  
ANIN  
NANN  
TAL  
GRL  
PRP  
AGYP  
PSTV  
STVE  
GHCP  
PLPAIR  
VQE  
VAK  
WL  
LSS  
SNR  
RHP  
MCED  
E  
L  
RRD  
SGS  
AQH  
CLL  
LRK  
ANPS  
RCHS  
RESQ  
AMAG  
QETS  
QDET  
YEV  
KMHD  
TEAC  
SEPS  
L  
STEM  
L  
SD  
NRQ  
L  
TLP  
EED  
DIR  
QSP  
KRG  
FLRS  
ASL  
GRR  
ASF  
H  
LECL  
K  
RQ  
DRGG  
DIS  
QKT  
VLP  
L  
HQA  
LAG  
V  
L  
R  
H  
SPAS  
MV  
NEN  
TRMY  
IPE  
ENHQ  
GSNY  
GS  
PR  
PA  
HAN  
MN  
ANA  
AG  
LA  
PE  
HI  
PT  
GA  
AL  
SW  
QA  
IDA  
AR  
QAK  
LMS  
AGN  
AT  
IST  
VS  
ST  
QR  
KRQ  
YGP  
KKQ  
GST  
TAT  
RPP  
RAL  
CL  
TL  
KN  
PI  
RAC  
IS  
IVE  
WK  
PF  
EE  
II  
LL  
TF  
ANC  
VA  
LA  
IY  
IP  
F  
PED  
D  
S  
N  
A  
T  
N  
S  
N  
L  
E  
R  
V  
E  
Y  
A  
F  
L  
K  
V  
I  
A  
Y  
G  
L  
L  
F  
H  
P  
N  
A  
Y  
L  
R  
N  
G  
W  
N  
L  
D  
F  
I  
I  
V  
V  
G  
L  
F  
S  
A  
T  
E  
S  
V  
N  
T  
E  
H  
Y  
N  
Q  
P  
N  
W  
L  
T  
E  
V  
Q  
D  
T  
A  
N  
T  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I  
Q  
S  
S  
A  
I  
N  
V  
V  
K  
I  
L  
R  
V  
L  
R  
P  
L  
R  
A  
I  
N  
R  
A  
K  
G  
L  
K  
H  
V  
V  
Q  
C  
V  
F  
A  
I  
R  
T  
G  
N  
Y  
I  
T  
Y  
K  
D  
G  
E  
V  
D  
H  
P  
I  
I  
Q  
P  
R  
S  
W  
E  
N  
S  
K  
F  
D  
F  
D  
N  
V  
L  
I  
L  
N  
T  
I  
C  
L  
A  
M  
Q  
H  
Y  
G  
Q  
S  
C  
L  
F  
K  
I  
A  
M  
N  
I  
L  
N  
M  
L  
F  
G  
L  
T  
V  
E  
M  
I  
L  
K  
L  
I  
A  
F  
K  
P  
K  
G  
Y  
F  
S  
D  
P  
W  
N  
V  
F  
D  
L  
I  
V  
I  
G  
S  
I  
I  
D  
V  
I  
L  
S  
E  
T  
N  
H  
Y  
F  
C  
D  
A  
W  
Q  
D  
I  
M  
T  
Y  
G  
A  
F  
L  
H  
K  
G  
S  
F  
C  
R  
N  
Y  
F  
N  
I  
L  
D  
L  
L  
V  
V  
S  
V  
S  
L  
I  
S  
F  
G  
I

# BLAST

## Step 2: Alignment

- align query and target at each candidate region



HSP (High-scoring segment pair)

# BLAST

## Step 3: Scoring

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

Nucleotide:

Score = 288 bits (318), Expect = 2e-73 Identities = 262/325 (81%), Gaps = 8/325 (2%) Strand=Plus/Plus			
Query 1923	TCAGCCTACCATGAGAATAAGAGAAAGA-AAATGAAGATCAAAAGCTTATTCTGT	1981	
Sbjct 33774	TCAGACTACCCCTGAGAATAAGAGAAAGAGAAATGAAGACCTAGA-CTTATCCATCT	33832	
Query 1982	TTCCTTTCTGGTGTAAAGCCAACACCCCTGTCTAAAAAACATAAATTCTTAATCAT	2041	
Sbjct Match=+2	TGACAAATTCTTTAAATAT	33892	
Query 2042	TTTGCCTCTTTCTGAGAATAGAGTGTT	2100	
Sbjct 33893	TTTGCCTCTTTCTGTGCTACAATTAAATAAAAAAGAATCTAATTAAATTGT	33952	
Query 2101	ACAGCACTGTTA-TGGTTCTGTGG	2159	
Sbjct 33953	CTATGACTGTTATTGGTTCTATGA	34012	
Query 2160	AAGTTCCAGTGTTC- Sbjct 34013 AAATTCCACTATTCTCTTCTCTTCTCTTCTCTTCTCTTCTGGATTAA	2219	
Query 2220	AT----TAAATAAATCATTAAACT 2240		
Sbjct 34073	ATTGCATAAAAAGAAACATTAAACT 34097		

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

# BLAST

## Step 3: Scoring

- Give each HSP a score

### Protein

```
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  MAKVLTLLEYKKLRDKETPSGFTVDDVIQTGV--DNPGHFFIMTVGCVAGDEESEYEVFKE 87  
+ K LT +L+++ +D+ GF+ I +G N G VG AG +SY F  
Sbjct 26  LQKCLTKDLWEQCKDRRDKYGF SFKQAI FSGSKWTNSG-----VGVYAGSHDSYYAFAP 79  
  
Query 80  R P L Q L L L  
Sbjct 80  E Y H M I E E  
Query 1  K +5 E +1 F -3  
Sbjct 138 AVTRKERKEIEHLVTSALGEFTGELKGKYYSLETMSD  
  
Gap  
- (11 + 6(1)) = -  
18  
  
Query 205 SGMARDWPDARGIWHNDNKSFLVWWNEEDHLRVISMEKGGNMKEVFRRCVG 256  
+G+ RDWP+ARGI+HND K+FLVWWNEED LR+ISM+ G N+ EVF+R V  
Sbjct 197 AGLERDWPEARLGIFHNDAKTFLVWWNEEDQLRIISMQAGSMILEVFKRLSVA 248
```

Scores from BLOSUM62, a position independent matrix

# 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	-2	-2	-1	-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	

# 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

Score = 176 bits (447),	Expect = 4e-50,	Method: Compositional matr
Identities = 98/232 (42%),	Positives = 139/232 (60%),	Gaps = 14/232
Query 30	MAKVLTLELYKKLRDKEPSGFTVDDVIQTGV--DNPGHFFIMTVGCVAGDEESYEVFKE	87
	+ K LT +L+++ +D+ GF+ I +G N G VG AG +SY F	
Sbjct 26	LQKCLTKDLWEQCKDRRDKYGFSPKQAIIFSGSKWTNSG-----VGVYAGSHDSYYAFAP	79
Query 88	LFDPIISDRHGGYKPTDKHKTDLNHENLKGG---DDLDPNYVLSSRVRTGRSIKGYTLPP	144
	D II HG +KP+DKH + ++++ L D D + S+R+R R++ L	
Sbjct 80	FMDKIIIEAYHG-HKPSDKHISSMDYKQLNCPPPPADED-KMINSTRIRVARNLAADPLGT	137
Query 145	HCSRGERRAVEKLSVEALNSLTGEFKGKYYPLKSMTEKEQQQLIDDHFLFDKPVSPLLL	204
	+R ER+ +E L AL TGE KGKYY L++M++ E++QLI DHFLF K L +	
Sbjct 138	AVTRKERKEIEHLVTSALGEFTGELKGKYYSLTMSDAEKKQLIADHFLF-KGGDKYLQS	196
Query 205	SGMARDWPDARGIWHNDNKSFLVWWNEEDHLRVISMEKGGNMKEVFRRFCVG	256
	+G+ RDWP+ARGI+HND K+FLVWWNEED LR+ISM+ G N+ EVF+R V	
Sbjct 197	AGLERDWPEARGLFHNDAKTFLVWWNEEDQLRIISMQAGSNILEVKRLSVA	248

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

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

# Local vs Global Alignment

BLAST: Basic Local Alignment Search Tool

Query: ACGGTGAGGTGTCCGAGAGAGCT

Target: ATTACGGTGAGGTATTAGACGGTGAGGTAACTCTCTCACGT

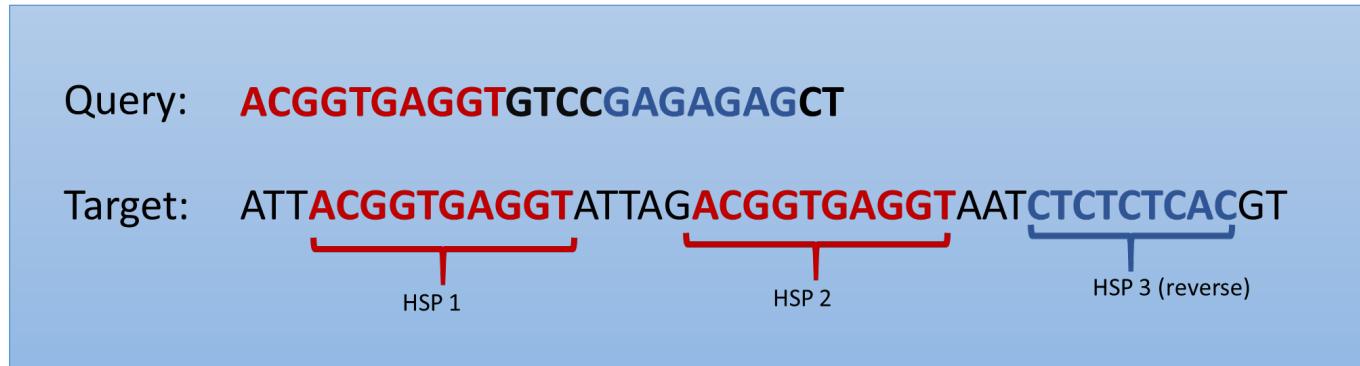
The diagram illustrates three local alignments (HSPs) between a query sequence and a target sequence. The query sequence is shown above in red: ACGGTGAGGTGTCCGAGAGAGCT. The target sequence is shown below in black: ATTACGGTGAGGTATTAGACGGTGAGGTAACTCTCTCACGT. Red brackets under the target sequence indicate two matches with the query (HSP 1 and HSP 2), while a blue bracket indicates a third match (HSP 3) that is described as being in reverse orientation.

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

Query: ACGGTGAGGTGTCCGAGAGAGCT

Target: ATTACGGTGAGGTATTAGACGGTGAGGTAACTCTCTCACGT

HSP 1

HSP 2

HSP 3 (reverse)

Global alignment:

---ACGGTGAGGT-----GT-----CCGAGAGAGCT  
| | | | | | | | | | | | | | | |  
ATTACGGTGAGGTATTAGACGGTGAGGTAACTCTCTCACGT

# **BLAST** is a package including the following tools

---

<u>command</u>	<u>Query</u>	<u>Hit database</u>
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)

**-evaluate**

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]
MSRSINMDQEKKNNVENLKGSGTTPASPTLSTPPTLNLMEQILLAKIEKQNLHESDDLHESDGRVGGKRRNILLRRTDS
MDSQNSASTYNFLSSDSASSGVYCKCDDCLLGIVDDYQRNPSVGRKKSSGWRKLRLRNIVWTPFFQTYYKKQRYPWQL
AGHQGNFRAGPTPTGTLKKLCPQEEACFRLLMNDILRPVPEFKGVLDVKDVEEGNVEETNSEETHQKDGSVDVIKRTV
VSSYLQLQDLLGFHPCVMDCVKVGVRTYLESELAKERPLRKDMYEKMVQDVPTAPNAERRVQGVTKPRYMVWRET
ISSTATLGFRVEGIKLAHGSSKDFKTTTRREQVTEALRFFVEGYPHAVPKYIQLKAIRATLKASPFASHEVVGSSL
FVHDTKNAGIWMIDFAKTLPLPQHLPRIHDAEWKVGNNHEDGYLIGVNNLIDIFQDIRNSEE
>NP_001014993.1 elongation factor 1-alpha [Apis mellifera]
MGKEKIHNIVVIGHVDGKSTTGHILYKCGGIDKRTIEKEAQMKGSKYAWVLDKLKAERERGITDIALWKF
ETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGTGEFEAGISKNGQTREHALFTLGVKQLIVGVNKMDSTEP
PYSETRFEEIKKEVSSYIKKIGYNPAAVAFVPISGWHGDNMLEVSSKMPWFKGWTVERKEGVGKCLIEALDAILPPTR
PTDKALRLPLQDVYKIGGIGTVPGVGRVETGVLPKGMVTFAPAGLTTEVKSLEM
>NP_001014994.1 glycerol-3-phosphate dehydrogenase [Apis mellifera]
MAEKLRICIVGSGNWGSTIAKIGINAANFSNFEDRVTMYVYEEINGKKLTEINETHENVKYLPGHKLPPNIIAIPDV
VEAAKDADILTFVPHQFIKRICSALFGKIKPTAIGLSLIKQGFDKKQGGIELISHIISKQLHIPVSVMGANLASEEVAN
EMFCETTIGCKDKNMAPILKDLMETSYFKVVVVEDVSVCECGALKNIACGAGFIDGLGLGDNTKAAVMRLGLMEITKF
VNIFPPGGKKTFFESCGVADLIATCYGGRNRKICEAFVTKGKKISELEKEMLNQQLQGPFTAEEVNMLAKNMENRF
PTTTVHRICGETMPMELIENLRNHPYIDETrNYQECKCSI
>NP_001019868.1 major royal jelly protein 9 precursor [Apis mellifera]
MSFNIWWLILYFSIVCQAKAHYSLRDKKANIFQVKYQWKYFDYNFGSDEKRQAAIQSGEYNYKNNVPIDVRWNGKTFVT
ILRNDGPVSSLNVSNKIGNGGPLLEPYPNWSWAKNCNSCGITSVYRIAIDEWDRLWLVDNGISGETSVCPSPQIVVFDLK
NSKLLKQVKIPHDIAINSTTGKRNVTPIVQSFDDYNTTWYIADVEGYALIINYNNADDSFQLRTSSTFVDPRTKYTN
DEFSLQLDGILGMALSHKTQNLYYSAMSSHNLNYVNTKQFTQGKFQANDIQYQGASDILWTQASAKAISETGALFFGLVS
DTALGCWNENRPLKRRNIEIVAKNNDTLQFISGIKIKQISSNIYERQNNEYIWVSNKYQKIANGDLNFNEVNFRILNA
PVNQLIRYTRCENPKTNFFSIFL
>NP_001027532.1 follistatin-like 5 [Apis mellifera]
MRCMLEIAARSFLLSIASTYVSVAGYKHSRRHRDTVAEYDASSNSDSLMTIPPSIDRSSIHEEYSLAEASSRSID
PCASKYCGIGKECELSPTIAVCVMRKPCRHPVCASNGKIYANHCELHRAACHSGSSLTKSRLMRCLHHDIENAH
RRTLHMNRTSLKTSKIVSYPKSRSRKKGLDNLPDKNDPDSKCESNQYEYIMKDNLLYNHARLMSQDNHSKEYLVSI
MFSHYDRNNNGNLEREELEQFAEEDLEELCRGNCNLGHMISYDDTDGDKLNVNEFYMAFSKLYSVSVSLDKSLEVNHI
SARVGDNVEIKCDVTGTPPPPLVWRNRNGADLETNEPEIRVFNDGSLYLTKVQLIHAGNYTCHAVRNQDVQTHVLTHT
IPEVKVTPRFQAKRLKEEANIRCHVAGEPLPQWLKNDEALNHQDPDKYDLIGNGTKLIKNVDYADTGAYMCQASSIG
GITRDISSLVQEQPTPTTESEERRFFSFHQWQGILVYEPSACRPRHEIRSTDVIPGTQEHVCGVKGIPCSWGRAINVANR
IGGLQHGPAGVWFTVSLH
>NP_001032395.1 putative tyramine receptor [Apis mellifera]
MANQTANYYGDVYQWNHTVSSGERDTRTEYLYLPNWTDLVLAGLFTMLIIVTIVGNTLVIAAVITRRLRSVTNCFVSSLA
AADLLVGLAVMPPAVLLQLTGGTWEGLPMLCDSWLSDILLCTASILCAISIDRYLAVTQPLIYSRRRSKRLAGLMI
VAVWVLAGAITSPPLLGCFCRATNRDIKKCSYNMDSYYVIFSAMGSFFLPMVLYVYGRISCVIASRHRNLEATESENV
RPRRNVLIERAKSIRARRTECVTSVTCDRPSDEAEPSSTSKSGIVRSHQQSCINRARETKTAGTLAVVGGFVACWL
PFFILYLATPFVPVEPPDILMPALTWLGWINSAINPFIYAFYSAFRLAFWRLTCRCKCFKSRTNLDPNSRNKLPAWANW
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 \
--restart '>' \
--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.dmdn )

## 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	MAKVLTLELYKKLRDI + K LT +L+++ +D-	QTGV--DNPGHPFIMTVGCVAGDEESYEVFKE	87
Sbjct 26	LQKCLTKDLWEQDKRDKRIGRGRKQAI	FSGSKWTNSG--- Critical D	79
Query 88	LFDPIISDIH D II	H GYKPTDIH TDLNHENLKGG-- H +KP+D+H + +++++ L	144
Sbjct 80	FMDKIIIEA -HKPSDIH	SSMDYKQLNCPPFI AED-KMINSTRIRVARNLAAADPLGT	137
Query 145	HCSRGERRAVEKLSVEALNSLTGEFKGKYYPLKSMTEKEQQQL	D HFLFDKPVSPILL +R ER+ +E L AL TGE KGKYY L++M++ E++QL	204
Sbjct 138	AVTRKERKEIEHLVTSALGEFTGELKGKYYSL ETMSDAEKKQL	A HFLF-KGGDKYI LQS	196
Query 205	SGMARDWPDARGI +G+ RDWP+ARGI	H DNKSFLVVVNEEDHLRVI H D K+FLVVNEED LR+ISM+ G N+ EVF+R V	256
Sbjct 197	AGLERDWPEARGI HI DAKTFLVVVNEEDQLRI	I ISMQAGSNILEVFKRLSVA	248

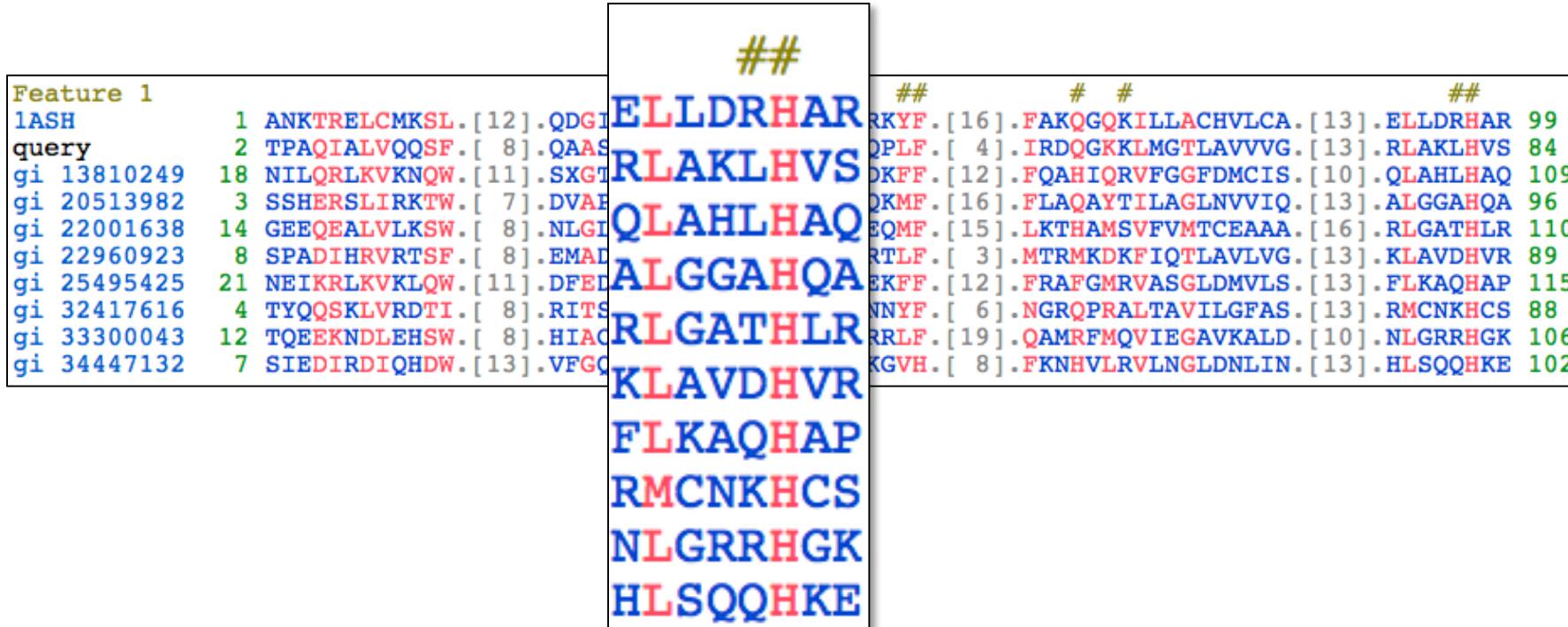
Ala	4	Blossom62 matrix:																
Arg	-1	5																
Asn	-2	0	6															
Asp	-2	-2	1	6														
Cys	0	-3	-3	-3	9													
Gln	-1	1	0	0	-3	5												
Glu	-1	0	0	2	-4	2	5											
Gly	0	-2	0	-1	-3	-2	6											
His	-2	0	1	-1	-3	0	0	-2	8									
Ile	-1	-3	-3	-3	-1	-3	-3	-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	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-1	-1	-1	-2	-1	1	5	
Trp	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	-2	-1	-1	-2	-1	3	-3	-2	-2	7
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3

# 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

Critical H

Query: TFATLSEL**H**CDKLHVDPENFRLLG

DB: genome of a distant species

Best BLAST hit

TFATLSEL <b>H</b> CDKLHVD-----PENFRLLG
S L      KLHV            P ++    +G
ILPAASRLA--KLHVSYGVQPTHYAPVG

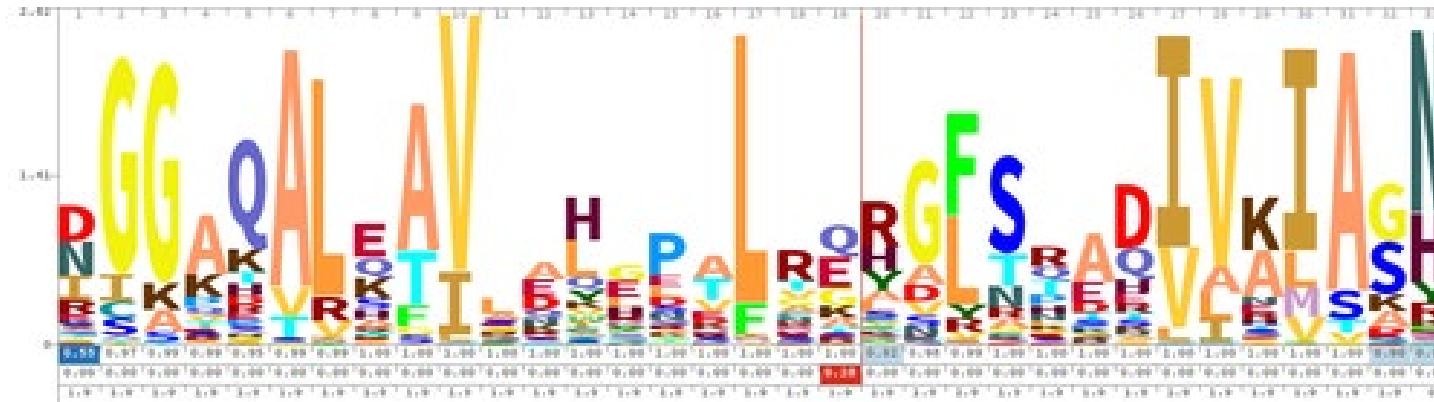
Best DELTA-BLAST hit (using PSSM)

TF--- <b>A</b> TLSEL <b>H</b> CDKLHVDPENFRLLG
+ L++LH            V P ++    +G
ILPAASRLAKLHVS-YGVQPTHYAPVG

BLAST is not reliable between distantly related species

# NCBI tools that implement PSSM

PSI-BLAST  
&  
CD-Search



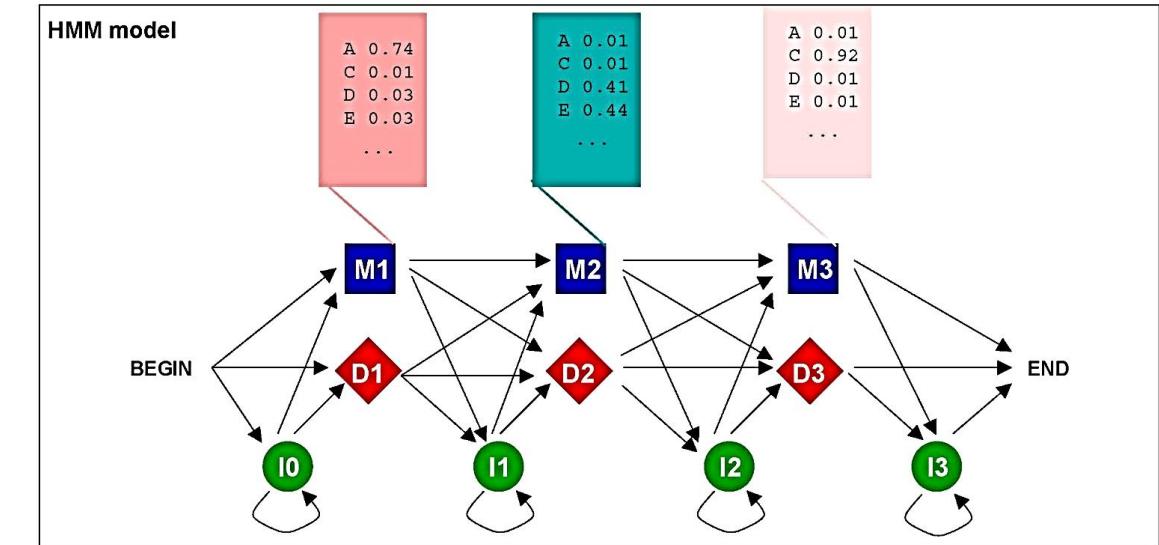
**PSI-BLAST:** Custom-made PSSM (Removed from NCBI web site now)

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

# Hidden Markov Model

**HMMs are trained from a multiple sequence alignment**

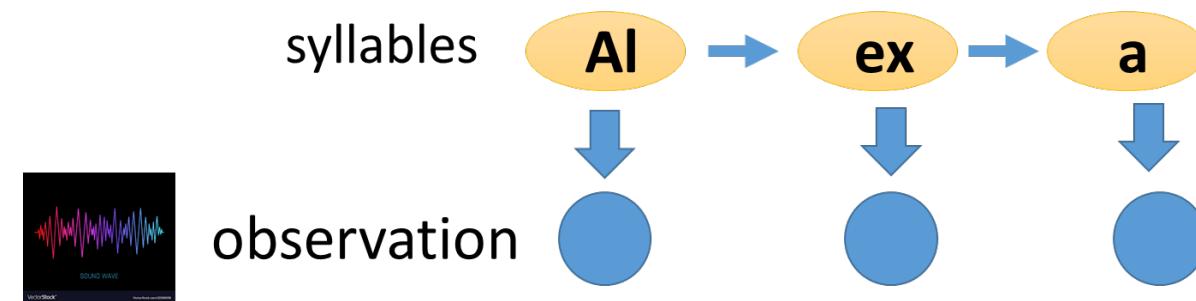
Q5E940_BOV_IN		*	*
RLA0_HUMAN	-MPREDRATWKSNSYFLKIIQLLDDYPKCFIVGADNVGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	7	7
RLA0_MOUSE	-MPREDRATWKSNSYFLKIIQLLDDYPKCFIVGADNVGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	7	7
RLA0_RAT	-MPREDRATWKSNSYFLKIIQLLDDYPKCFIVGADNVGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	7	7
RLA0_CHICK	-MPREDRATWKSNSYFLMIIQLLDDYPKCFIVGADNVGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	7	7
RLA0_RANSY	-MPREDRATWKSNSYFLKIIQLLDDYPKCFIVGADNVGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE	7	7
Q7ZUG3_BRARE	-MPREDRATWKSNSYFLKIIQLLDDYPKCFIVGADNVGSKOMOTIRSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	7	7
RLA0_ICTPU	-MPREDRATWKSNSYFLKIIQLLDDYPKCFIVGADNVGSKOMOTIRSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	7	7
RLA0_DROME	-MVRENKAQWAQYFIRKVBFEDPFPKCFIVGADNVGSKOMONIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--POLE	7	7
RLA0_DICDI	-MSGAC-SKRKKLFIEIKAFTKLFTYDMDKIVMAAEADFVGSSOLKIRKSIRGI-CAVLMGKTTMRKIVRLDLASD-K-PEDL	7	7
Q5ALP0_DICDI	-MSGAC-SKRKKNFIEIKAFTKLFTYDMDKIVMAAEADFVGSSOLKIRKSIRGI-CAVLMGKTTMRKIVRLDLASD-K-PEDL	7	7
RLA0_PLAF8	-MAKLSKQKKQYMMIEYLKKLLIQNSKILLVHVWDNGINOMASVRSKRS-LGK-AVVLMGKNTTRTALKNNLDAV--DQIE	7	7
RLA0_SULAC	-MIGLAVTTTKIIAKRVDEVAELTSSLKIKHETKTIIIANIEGFPAKDLHESERKLRKGK-ADLVTKVNTNLNIALKNAQG--YDVK	7	7
RLA0_SULTO	-MRIMAVITQERKIAKWWIEVEKELELRLKREHTIILIANIEGFPAKDLHEDRVTKVNLTNLNIALKNAQG--LDVS	8	8
RLA0_SULSO	-MKRKLALAKQYRKVASWLEEVEKTELIELKNSNTILLNQLEGPFPADKHLIEERKLRKGK-ATVTKVNTNLNIALKNAQG--IDIE	8	8
RLA0_AERPE	-MSVVSILQVGQMKYRKEPKIDPEWTLMLREREEFLSQRKVRVFLWADTFPTFVFRVQKRLWKW-YPPMVKARQRLILRAMKAGCLE--LDON	8	8
RLA0_PYRAE	-MMLAIGKRRYVHTRFQDPAKVKIVSEFATELQQWYFYELFDLHGS-BRILHEYYRRLRY-GVIVLIIKPTLEKIAFKTWYGCG--IPAE	8	8
RLA0_MET_AC	-MAEERHTEHITPQWKDELENKELIQSOKVFGMVGICELLATTMKRIVRDLKD-VDV-ALVKVSRNLTERALRNQLG--ETIP	7	7
RLA0_MET_MTA	-MAEERHTEHITPQWKDELENKELIQSOKVFGMVGICELLATTMKRIVRDLKD-VDV-ALVKVSRNLTERALRNQLG--ESIP	7	7
RLA0_ARCFU	-MAA VRGS---PPEYVVAKEEVIKRMISSSPKVVAIISFRNVPAQOMCQKIREFRGK-AEKIVVKVNTLNRALDALG---GDYL	7	7
RLA0_METKA	-MAVKAKQPPSSGYEPKVAEWKREVRKELKELMDENYVGNLVDLCEIPAPOLEIRAKIRLDRDILIRNSRNTLMLRIALEKELDR-PELE	7	7
RLA0_METTH	-MAHVAEWKKKEVQELHDILKGVVYGGIANLADIPAROLQKMRQTLRDS-ALIEMSKKTTLISLALEKAGREL-ENVD	7	7
RLA0_METTL	-MITAESHKIAPWKEIEVNKLKELKNQIVAIYALDMMEVPVAPOLOEIRDKR-CTMTLKMRSRNTLILRKAIVEEFTGNEPEFA	8	8
RLA0_METVA	-MIDAKSEHKAPWKEIEVNKLKELKNQIVAIYALDMMEVPVAPOLOEIRDKR-DQMTLKMRSRNTLILRKAIVEEFTGNEPEFA	8	8
RLA0_METJA	-METKVKAHAPWKEIEVTKLGLIKSKPSPKVVAIDMDVMPVAPOLOEIRDKR-DRVKLIRMSRNTLILRKAEEALENNPKLA	8	8
RLA0_PYRAB	-MAHVAEWKKKEVEELANIJKS-PVITALWDVSSMPAYLPSQMRRLIRENGLLRSVSRNTLILLELAIKKAAELELGKPELE	7	7
RLA0_PYRHO	-MAHVAEWKKKEVEELANIJKS-PVITALWDVSSMPAYLPSQMRRLIRENGLLRSVSRNTLILLELAIKKAAELELGKPELE	7	7
RLA0_PYRFU	-MAHVAEWKKKEVEELANIJKS-PVITALWDVSSMPAYLPSQMRRLIRENGLLRSVSRNTLILLELAIKKAAELELGKPELE	7	7
RLA0_PYRKO	-MAHVAEWKKKEVEELANIJKS-PVITALWDVSSMPAYLPSQMRRLIRENGLLRSVSRNTLILLELAIKKAAELELGKPELE	7	7
RLA0_HALMA	-MSAEASERKETIDPEWQEEVDAVEMIESYESVCVNVNIAGPVRDOLQMDRRLHGCT-EALRVSRTNLLERALRDWD---DGLE	7	7
RLA0_HALVO	-MSESEFRVQTETIPQWKREVEELDVEIESYESCVVGVGAIPR-ROLOSMRSTNLRLVRALEDEVN---DGFE	7	7
RLA0_HALSA	-MSAEQPRTEEVPEWQREOVAEVLDDILETDSVGVWNVTGKFLKOLOMDRGLHQG-AALRMSRNTLILRVALEEAG---DGLD	7	7
RLA0_THE_AC	-MKEYSQOKRELVNETRPIKASRSVVAIETEAGTR-ROIQDIEGKNGKQ-INLIVKIKTLLFKCALENLG-EKLS	7	7
RLA0_THE_VO	-MRKINPKKIEIVSELADOTIKSKAWAIVDICKVRLQDMRQDMLRNRDK-VKIKVVKKLLFKALDSDIND-EKLT	7	7
RLA0_PICTO	-MTEPQWVIFDFVKVNLNEINSRKVVAIISIKGLRNNEFCKIRNSIRDK-ARIKVKSRARLLRLAATENICK--NNIV	7	7
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90		



# Pre-constructed HMM database: PFAM

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

time → A chain of events



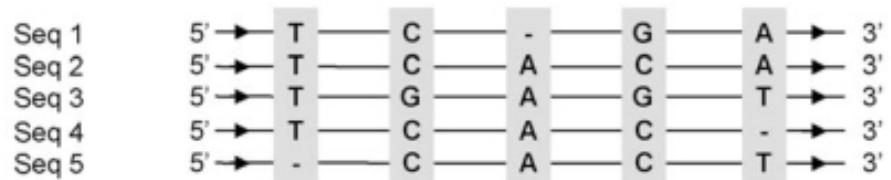
You can train an HMM model with real world data.  
These are the elements in the model:

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

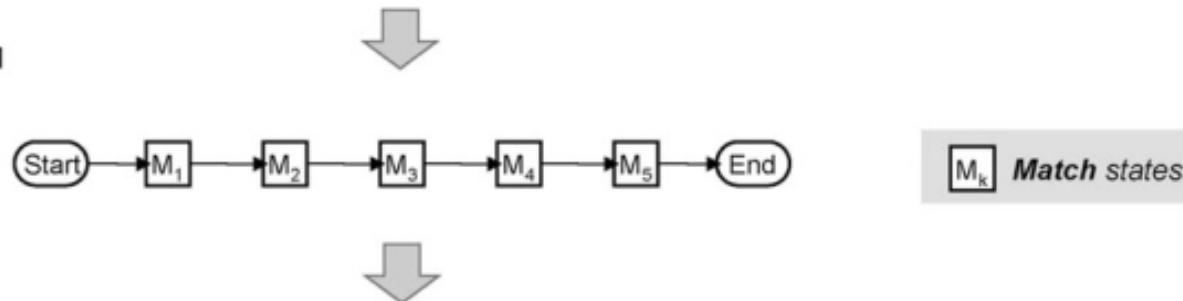
**Viterbi algorithm**: match a new series with the model

HMM is good for model sequence profile of a domain

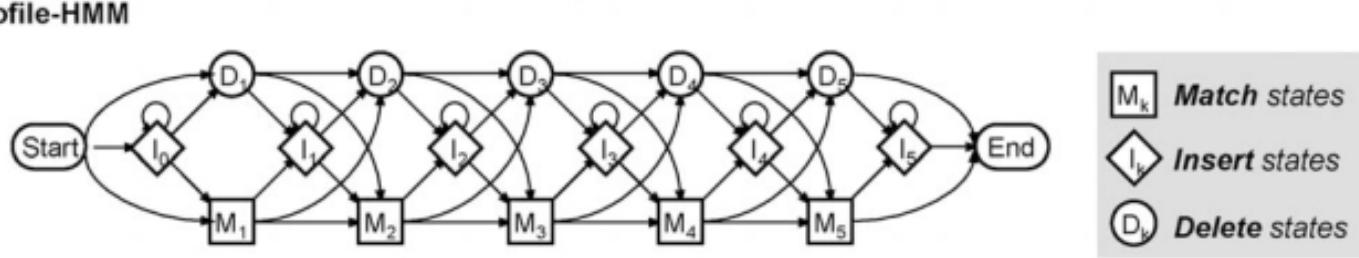
(a) Sequence Alignment



(b) Ungapped HMM



(c) Profile-HMM



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:	<a href="#">Pollington J</a> ⓘ
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 pfam
```

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
<a href="#">PF07690.ncbi.gz</a>	2.1 GB	5/5/20, 10:51:00 AM	
<a href="#">Pfam-A.clans.tsv.gz</a>	327 kB	5/1/20, 11:10:00 PM	
<a href="#">Pfam-A.dead.gz</a>	22.5 kB	5/1/20, 9:47:00 AM	
<a href="#">Pfam-A.fasta.gz</a>	3.4 GB	5/1/20, 9:54:00 AM	
<a href="#">Pfam-A.full.gz</a>	10.6 GB	5/1/20, 10:46:00 AM	
<a href="#">Pfam-A.full.metagenomics.gz</a>	745 MB	5/1/20, 1:11:00 PM	
<a href="#">Pfam-A.full.ncbi.gz</a>	64.5 GB	5/1/20, 8:05:00 PM	
<a href="#">Pfam-A.full.uniprot.gz</a>	23.9 GB	5/1/20, 1:07:00 PM	
<a href="#">Pfam-A.hmm.dat.gz</a>	480 kB	5/1/20, 8:05:00 PM	
<a href="#">Pfam-A.hmm.gz</a>	261 MB	5/1/20, 8:06:00 PM	
<a href="#">Pfam-A.regions.tsv.gz</a>	2.0 GB	5/1/20, 10:55:00 PM	
<a href="#">Pfam-A.regions.uniprot.tsv.gz</a>	6.7 GB	5/1/20, 11:10:00 PM	
<a href="#">Pfam-A.rp15.gz</a>	918 MB	5/1/20, 8:12:00 PM	
<a href="#">Pfam-A.rp35.gz</a>	3.1 GB	5/1/20, 8:33:00 PM	
<a href="#">Pfam-A.rp55.gz</a>	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	<a href="#">Show</a>
RHIM	RIP homotypic interaction motif	Domain	n/a	659	717	677	716	13	51	52	43.1	4.5e-11	n/a	<a href="#">Show</a>

### 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	<a href="#">Show</a>

# 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 `hmmpress` and `hmmscan`

`hmmpress`

`hmmscan`

Query

HMM file

Sequence file

Target

Sequence file

HMM file

Commands

```
hmmpress my.hmm myProteins.fasta
```

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

HMMScan: for small number of proteins;

HMMsearch: for large number of proteins;

# 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 KLKNNGSKIFA 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 VKILKVVDPTE 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 IKEVYYENDE ELND.....
BYR2_SCHPO/394-658	WIRGALIGSG SFGQV..YLG MNASSGELMA VKQVILDVS 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 RDRVNETIA LKKIRLEQED EGVP.....
BUR1_YEAST/60-366	YREDEKLGQG TFGEV..YKG IHLETQRQVA MKKIIIVSVEK DLFP.....
CTK1_YEAST/183-469	YLRIMQVGEQ TYGKV..YKA KNTNTEKLVA LKKLRLQGER EGFP.....
GSK3A_RAT/119-403	YTDIKVIGNG SFGVV..YQA RLAETRELVA IKKVLQDKRF KNR.....
MAK_RAT/4-284	YTTMRQLGDG TYGSV..LMG KSNESGELVA IKRMKRKFYS WDECMMN..LR
CDKL1_HUMAN/5-288	YEKIGKIGEG SYGVV..FKC RNRDTGQIVA IKKFLESEDD PVIK.....
PIM1_HUMAN/38-290	YQVGPLLGSQ GFGSV..YSG IRVSDNLPVA IKHVEKDRIS DWGELP..NG
PSK1_YEAST/1096-1354	FVSLQKMGEQ 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
APB1_P0V7N/101-452	ECVYRPTTCRQG CECGV..YGC RKAETGKIVVA MKCLPKKRTK MKCG.....

# Global alignment is forced for MSA

Seq1: ACGGTGAGGTGTCCGAGAGAGCT

Seq2: ATT**ACGGTGAGGT**ATTAG**ACGGTGAGGT**AAT**CTCTCACGT**

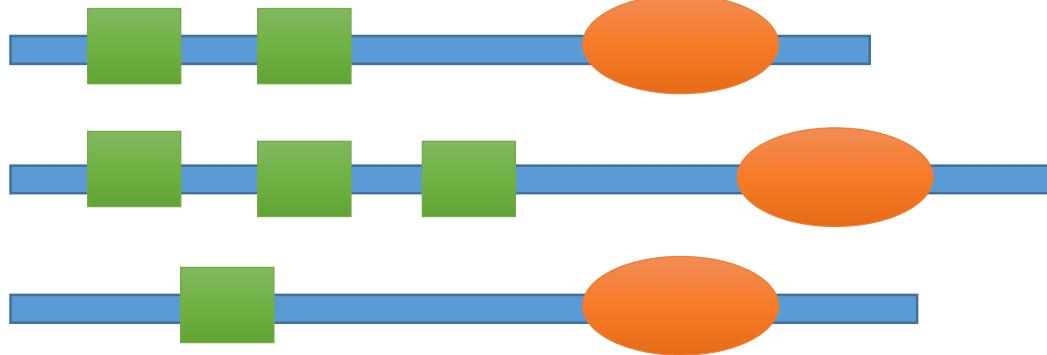
## Global alignment:

**---ACGGTGAGGT-----GT-----CCGAGAGAGCT**

ATTACGGTGAGGTATTAGACGGTGAGGTAACTCTCTCACGT

# Difficulty in MSA

Protein families with  
multiple domains



## Deletions

GGAC	AA	T	AA	TT
GGAC	AA	G	AA	TT
GGAC	AA			TT

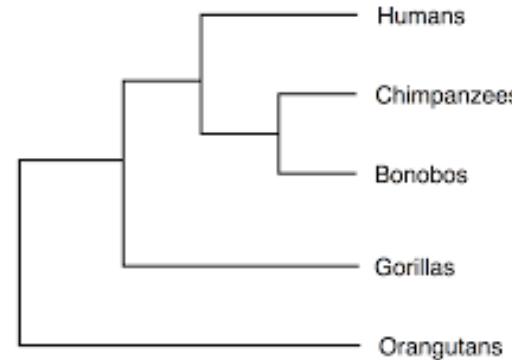
MSA on individual  
domains

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

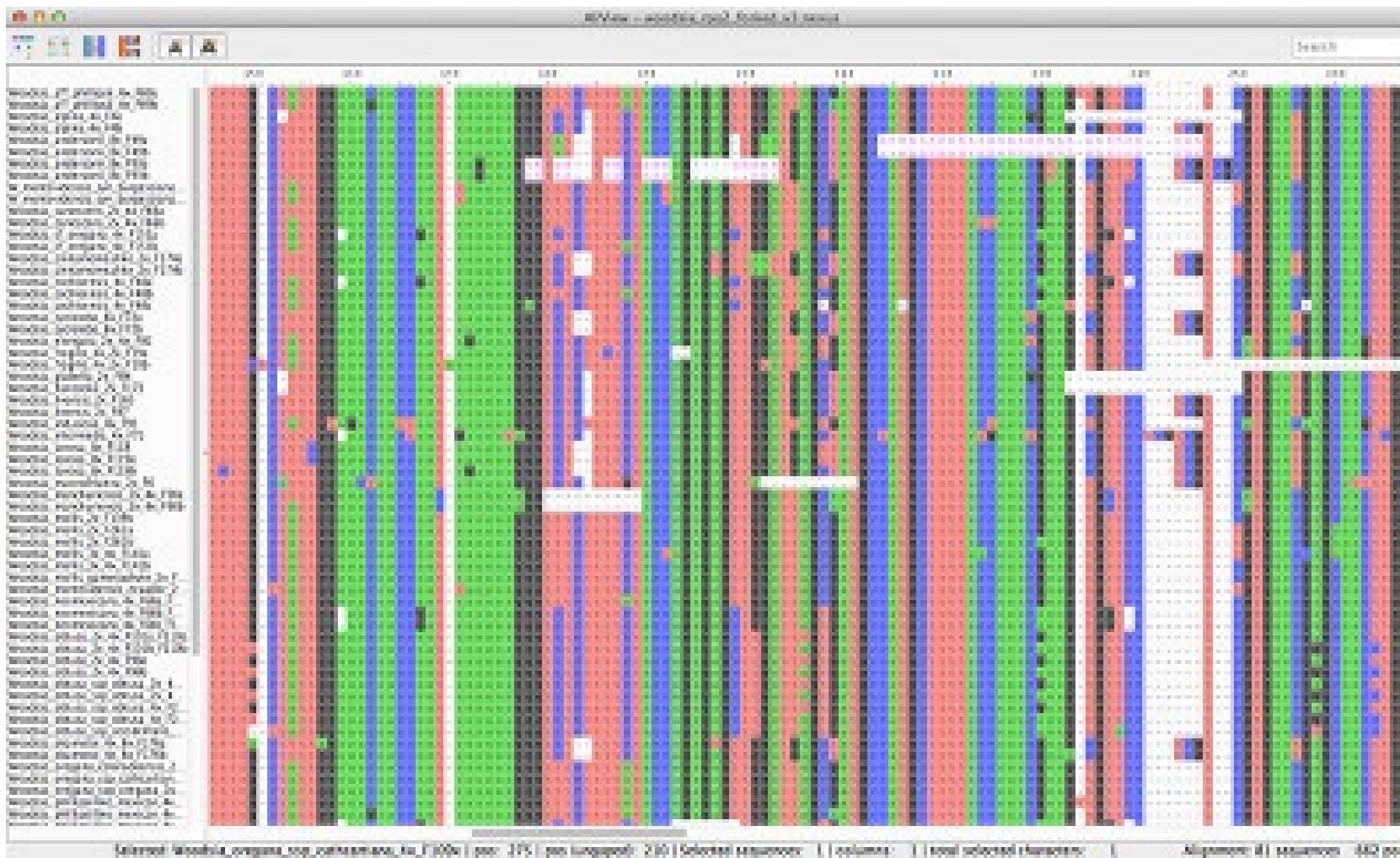
**HMM guided**

HMMAign

align sequences to an HMM profile

# Alignment viewers and editors

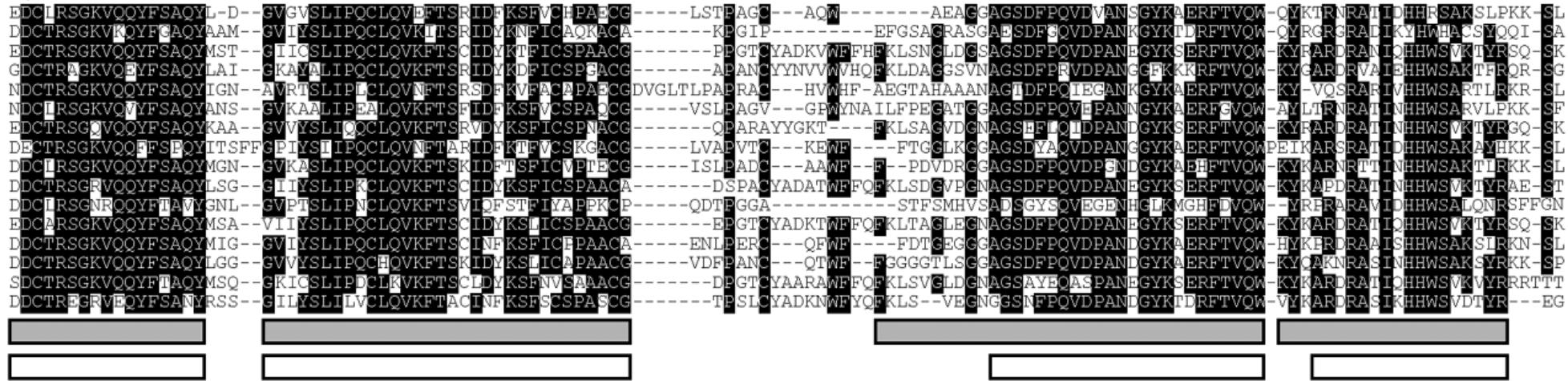
AliView (<https://ormbunkar.se/aliview/>)



# 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 TSPASIRPPAGPSSRPAMVSSRRTRPSPPGPRRPTGRPCCSAAPRRQAT  
IXI\_235 TSPASIRPPAGPSSR-----RPSPPGPRRPTGRPCCSAAPRRQAT  
IXI\_236 TSPASIRPPAGPSSRPAMVSSR--RSPPPPRPPGRPCCSAAPRRQAT  
IXI\_237 TSPASLRPPAGPSSRPAMVSSR-RPSPPGPRRPT---CSAAPRRQAT

IXI\_234 GGIKTCGTCCTTSTRHRGRSGWSARTTAACLRASRKSMRAACSRSA  
IXI\_235 GGIKTCGTCCTTSTRHRGRSGW-----RASRKSMRAACSRSA  
IXI\_236 GGIKTCGTCCTTSTRHRGRSGWSARTTAACLRASRKSMRAACSR-G  
IXI\_237 GGYKTCGTCCTTSTRHRGRSGYSARTTAACLRASRKSMRAACCSR-G

IXI\_234 SRPNRFAPTLMSSCITSTTGPPAWAGDRSHE  
IXI\_235 SRPNRFAPTLMSSCITSTTGPPAWAGDRSHE  
IXI\_236 SRPNRFAPPLMSSCITSTGPPPAGDRSHE  
IXI\_237 SRPNRFAPTLMSCLSTTGPPAYAGDRSHE

clustal

>1ymg\_A THE CHANNEL ARCHITE  
SASFWRRAICAEFFASLFYVFFGLGASLRW----AG-----P-----1HVLQVAL  
AFGLALATLVQAVGHISGAHVNPATF AFLVGSMQSLRAICYMVAQLLGAVAGAAVLYS  
VT--PPAvRGN1ALNTLHPGVSGQATIVEIFLTQFLCIFATYDERRNGRLGSVALAV  
GFSLTLGHLFGMYYTGAGMNPARSFAPAILTR----NFTNHWWVWGPVIGAGLGSLL  
YDFLFLFPRLKSUSERLSILKG  
>2d57\_A DOUBLE LAYERED 2D C  
TQAFWKAVTAELFLAMIIVLVLGVGSTINW----GG-SENPLP-----VDMVLISL  
CFGLLSIATMVQCFGHISGGHINPAVTAVMCTRKISIAKSVFYITAQCLGAIIGAGILYL  
VT--PPSVVGGLVTTVHGNTAGHGLLVELIITFQLVFTIFASCDSKRTDVTGSVALAI  
GFSVAIGHFAINYTGASMNPARSFGPAVIMG----NWENHWIYwVGPIIGAVLAGAL  
YEYVF-----CP  
>2f2b\_A CRYSTAL STRUCTURE O  
MVSLLTKRCIAEFIGTFILVFFGAGSAAVTLMIASGGTSPNPFNIGIGLLGGLDWVAIGL  
AFGFAIAASIYALGNISGCHINPAVTIGLWSVKKFPGREVVVPYIIAQLLGAAFGSFIFLQ  
CAGIGAACATVGGLATAFPFGISYWQAMLAEVVGTFLLMITIMGIAvDERAP-KGFAGIII  
GLTVAGIITLGNISGSSLNPARTFGPYLNDMifagtDlWNYYSIYvIGPIVGAVLAALT  
YQYL-----TS

fasta

>d1a1x\_b.63.1.1 (-) p13-MTCP1 {Human (Homo sapiens)}  
PPDHLMWHQEIGIYRDEYQRTWAVVEE..E..T..SF.....LR.....ARVQQIQVPLG.....DAARPShLLTS....QL  
>gi|6678257|ref|NP\_033363.1|:(7-103) T-cell lymphoma breakpoint 1 [Mus musculus]  
HPNRLWIKWKEKHVYLDEFRSWLPVVIK..S..N..EK.....FQ.....VILRQEDVTLG.....EAMSPSQLVPY....EL  
>gi|7305557|ref|NP\_038800.1|:(8-103) T-cell leukemia/lymphoma 1B, 3 [Mus musculus]  
PPRFLVCTRDDIYEDENGRQWVAKVE..T..S..RSpygarietcIT.....VHLQHMTTIPQ.....EPTPQQPINNN....SL  
>gi|11415028|ref|NP\_068801.1|:(2-106) T-cell lymphoma-1; T-cell lymphoma-1a [Homo sapiens]  
HPDLRWAWEKFVYLDKQHAWPLTIEkD..R..LQ.....LR.....VLLRREDVVLG.....RPMPTPQICPS....LL  
>gi|7305561|ref|NP\_038804.1|:(7-103) T-cell leukemia/lymphoma 1B, 5 [Mus musculus]  
-----GIYEDEHHRWIAVNVE..T..S..HS.....SHgnrietcvt.VHLQHMTTLPQ.....EPTPQQPINNN....SL  
>gi|7305553|ref|NP\_038801.1|:(5-103) T-cell leukemia/lymphoma 1B, 1 [Mus musculus]  
LPVYLVSVRIGIYEDEHHRWIVANVE..TshS..SH.....GN.....RRRTHTVHLW.....KLIPQQVIPNplnydFL  
>gi|27668591|ref|XP\_234504.1|:(7-103) similar to Chain A, Crystal Structure Of Murine Tc1  
-PDRWLWEKHVYLDEFRSWLPVVIK..S..N..GK.....FQ.....VIMRQKDVLG.....DSMTPSQLVPY....EL  
>gi|27668589|ref|XP\_234503.1|:(9-91) similar to T-cell leukemia/lymphoma 1B, 5;  
-PHILTLRTHGIGIYEDEHHRWLVLDLQ..A..Sh1SF.....SN.....RLLIYTIVLQggvafp1ESTPPSPMNLN....GL  
>gi|7305559|ref|NP\_038802.1|:(8-102) T-cell leukemia/lymphoma 1B, 4 [Mus musculus]  
PPCFLVCTRDDIYEDENGRQWVAKVE..T..S..SH.....SPycskietcvtVHLWQMTTLFQ.....EPSPDSLKTFN....FL  
>gi|7305555|ref|NP\_038803.1|:(9-102) T-cell leukemia/lymphoma 1B, 2 [Mus musculus]  
-----PGFYDEHHRLWMVAKLE..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/](https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/)

Phylogeny.fr [http://phylogeny.lirmm.fr/phylo\\_cgi/data\\_converter.cgi](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](mailto:brc-bioinformatics@cornell.edu)

**Office hours**

<https://biohpc.cornell.edu/lab/office1.aspx>