3000788 Intro to Comp Molec Biol

Lecture 5: Applications of sequence alignment

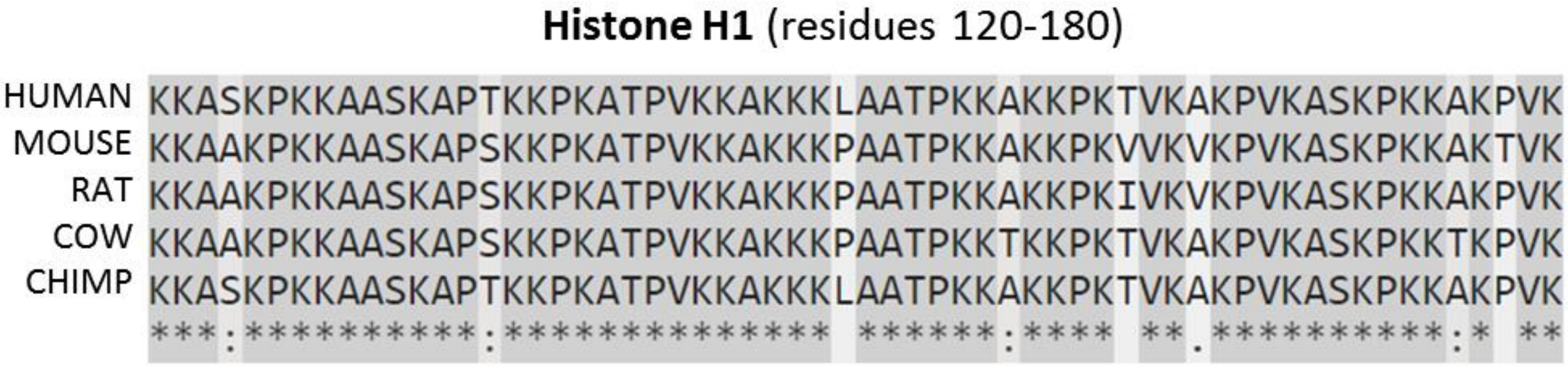
August 30, 2022

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Sequence homology

Evolution occurs at the sequence level



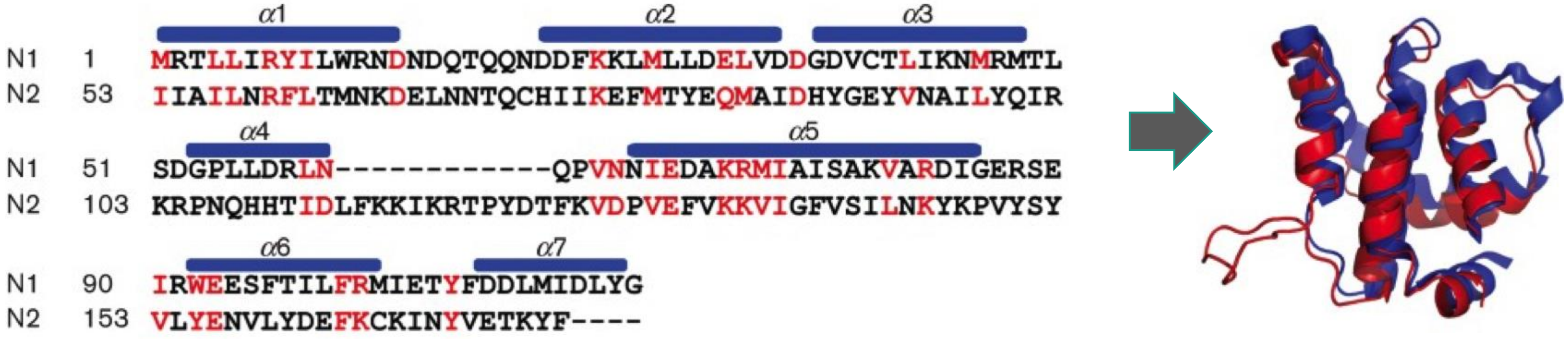
Histone H1 (residues 120-180)

```
HUMAN KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK
  RAT KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKAKPVK
  COW KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
CHIMP KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
      ***:**********:************** ******:**** **.**********:* **
```

https://en.wikipedia.org/wiki/Homology_(biology)

- Genes / proteins originating from the same ancestor will have similar sequence
- High sequence similarity → functional similarity, structural similarity, etc.

Sequence alignment enables inference

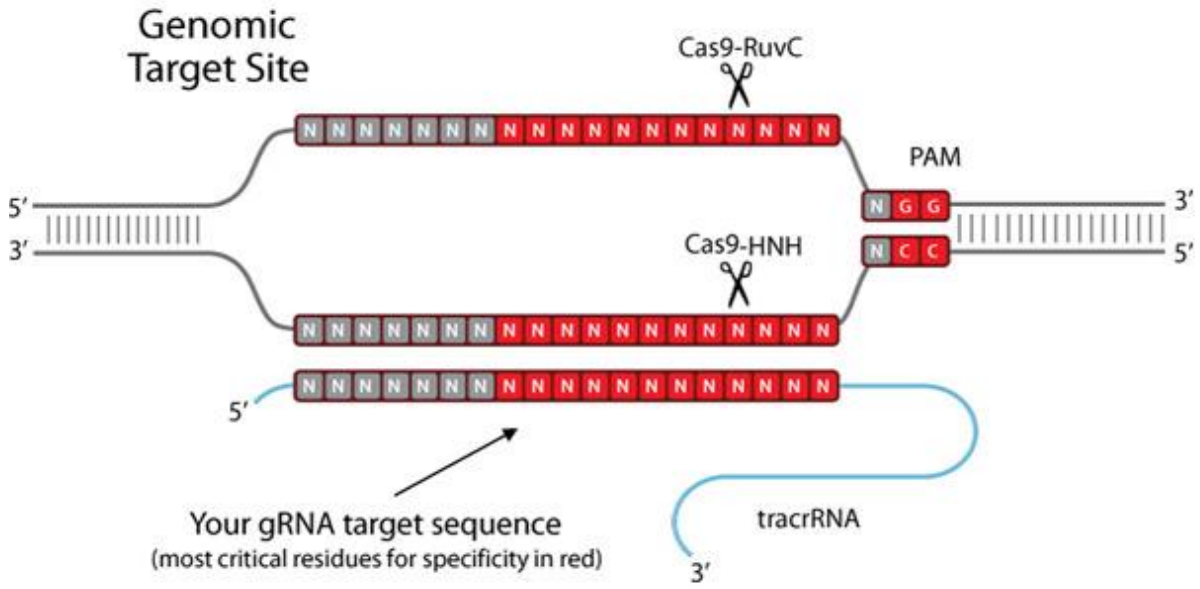


```
     α1                    α2              α3
N1   1    MRTLLIRYILWRNDNDQTQQNDDFKKLMLLDELVDDGDVCTLIKNMRMTL
N2   53   IIAILNRFLTMNKDELNNTQCHIIKEFMTYEQMAIDHYGEYVNAILYQIR

     α4                           α5
N1   51   SDGPLLDRLN-----------QPVNNIEDAKRMIAISAKVARDIGERSE
N2   103  KRPNQHHTIDLFKKIKRTPYDTFKVDPVEFVKKVIGFVSILNKYKPVYSY

     α6              α7
N1   90   IRWEESFTILFRMIETYFDDLMIDLYG
N2   153  VLYENVLYDEFKCKINYVETKYF----
```

Ferguson et al. J General Virology, 94: 2070-2081 (2013)

- Same amino acid residue positions are involved in similar secondary structure
- Properties of amino acid side chains are important

Molecular probe design


Genomic Target Site
Cas9-RuvC
N N N N N N N N N N N N N N N N N N N N
PAM
5′
3′
N G G
N C C
3′
5′
Cas9-HNH
N N N N N N N N N N N N N N N N N N N N
N N N N N N N N N N N N N N N N N N N N
5′
Your gRNA target sequence
(most critical residues for specificity in red)
tracrRNA
3′

http://www.sigmaaldrich.com/technical-documents/articles/biology/crispr-cas9-genome-editing.html

- Sequence alignment can check the specificity of your probes

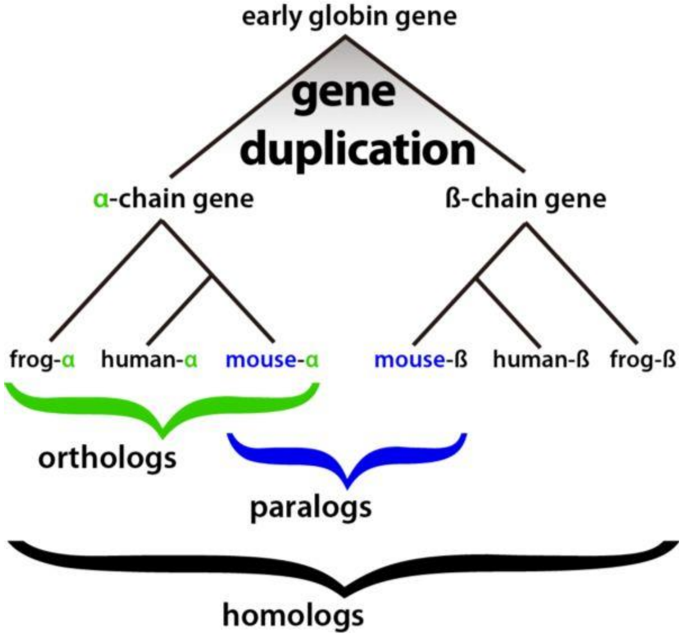
Broad applications of sequence homology

- Infer evolutionary relationship across species
  - Many-to-many alignment between gene lists

- Identify the species of origin for a sequence
  - One-to-many alignment against a reference database
  - Host vs pathogen

- Predict function and structure
  - Partial similarity is good enough
  - Locate conserved functional domain / motif

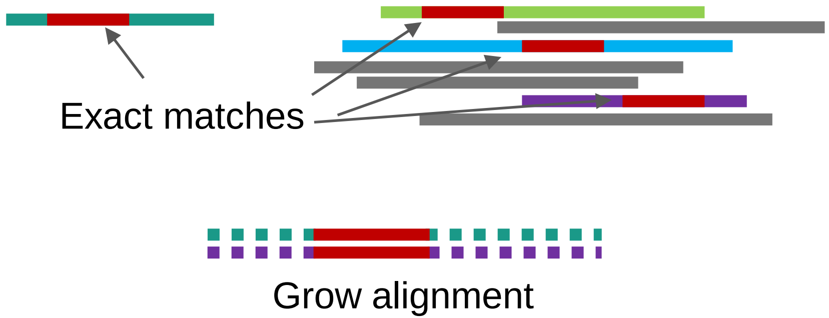
- Check the specificity of designed probes



https://sites.google.com/site/jkim339n/part2a

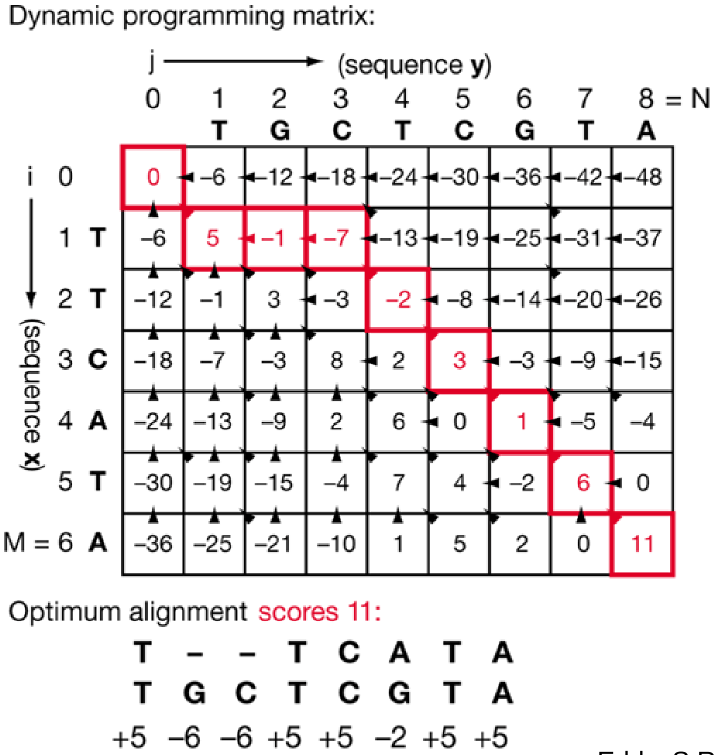
Components of sequence alignment

Starting from exact match (seed / word)



- Input sequence length = 300
- Expected similarity between input and reference = 95% (genome re-sequencing)
- Expected 15 mismatches
- If mismatches are random, there should be a run of 285/16 ~ 18 positions with matches
  - MM…MEM…MEM………MEM…MM
  - NCBI's MEGABLAST searches for a run of 28 matches

Dynamic programming algorithm



Dynamic programming matrix:

j → (sequence **y**)

i ↓ (sequence **x**)

		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
i 0		0	–6	–12	–18	–24	–30	–36	–42	–48
1	T	–6	5	–1	–7	–13	–19	–25	–31	–37
2	T	–12	–1	3	–3	–2	–8	–14	–20	–26
3	C	–18	–7	–3	8	2	3	–3	–9	–15
4	A	–24	–13	–9	2	6	0	1	–5	–4
5	T	–30	–19	–15	–4	7	4	–2	6	0
M = 6	A	–36	–25	–21	–10	1	5	2	0	11

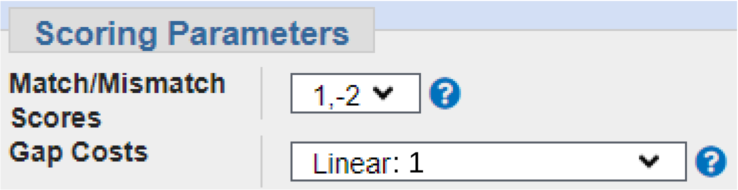
Optimum alignment scores 11:

T	–	–	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	–6	–6	+5	+5	–2	+5	+5

- The best alignment for TTCATA vs TGCTCGTA is either
  - T/T + best alignment for TCATA vs GCTCGTA
  - T/– + best alignment for TCATA vs TGCTCGTA
  - –/T + best alignment for TTCATA vs GCTCGTA
- Rely on the score function

Eddy, S.R. Nature Biotechnology 22:909-10 (2004)

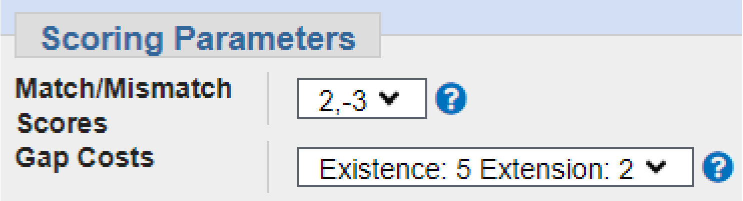
Alignment scores



Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear: 1



Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

```
  Ref: ACCGTATCG
       ||   ||||
Query: AC---ATCG
```

Score = +1+1-1-1-1+1+1+1+1
= +3

Score = +2+2-5-2-2-2+2+2+2+2
= +1

- Gap cost models
  - Constant = Same penalty regardless of length
  - Linear = Penalty x Length
  - Affine = Existence + (Extension x Length)

Alignment score interpretation

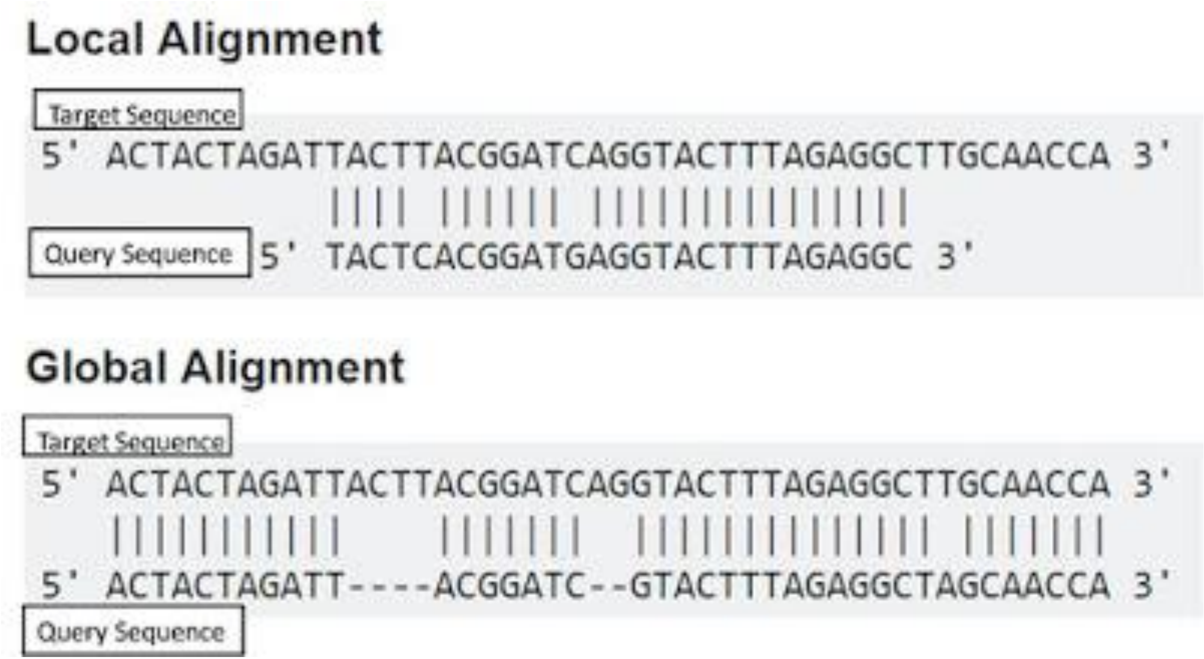
- **Match / Mismatch = +1 / –2**
  - To permit a mismatch, there must be >2 matches afterward to gain score
  - Want hits with high identity

- **Match / Mismatch = +2 / –3**
  - A mismatch followed by two matches = net +1 score
  - Want hits with intermediate identity

- **Gap cost**
  - Constant = An insertion/deletion can be of any length
  - Linear = Long indel is less likely than short indel
  - Affine = Existence + (Extension x Length)
    - Balance between constant and linear

Global and local alignment

Global vs local alignment



Local Alignment

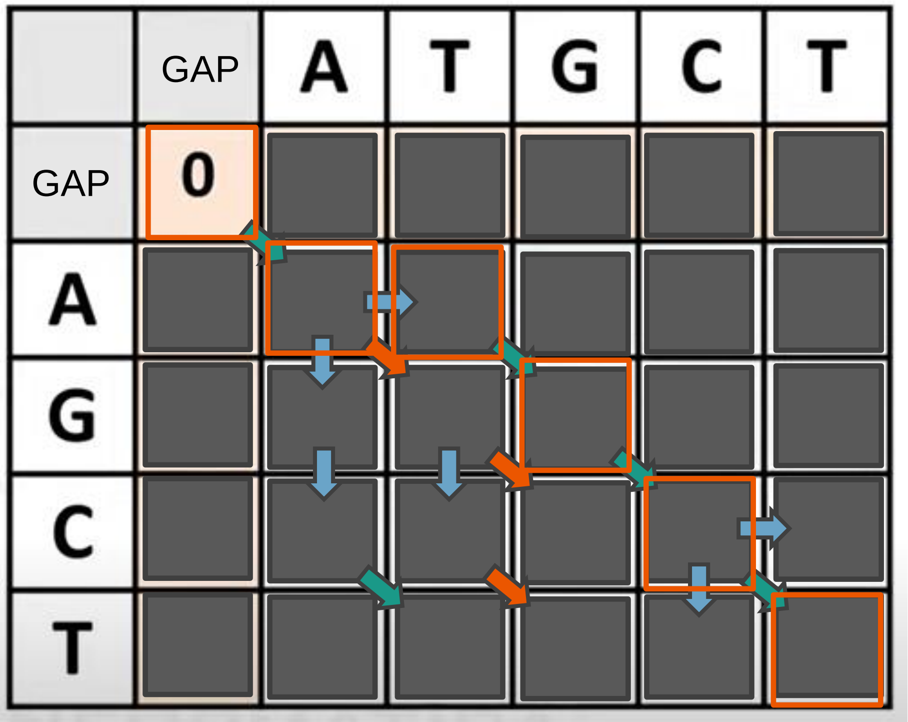
```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
Query Sequence
          5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
Query Sequence
```

Images generated from BABA http://baba.sourceforge.net/

Global alignment



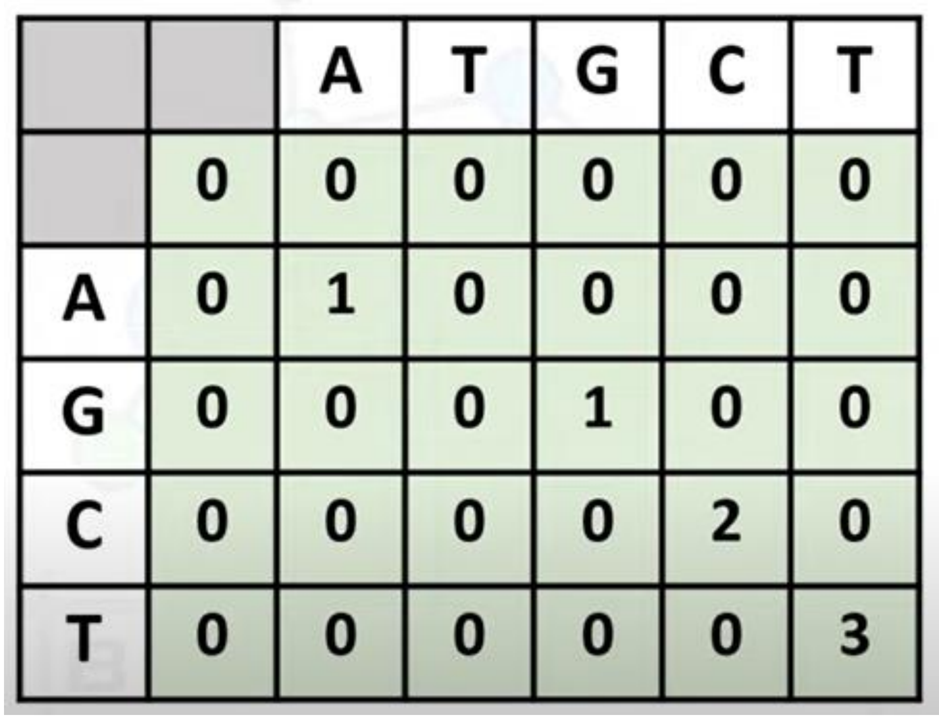
	GAP	A	T	G	C	T
GAP	0					
A						
G						
C						
T						

Match	: 1	🟩
Mismatch	: -1	🟥
GAP	: -2	🟥

```
Seq1:  ATGCT
       | |||
Seq2:  A-GCT
```

https://www.youtube.com/watch?v=ipp-pNRIp4g

Local alignment



		A	T	G	C	T
	0	0	0	0	0	0
A	0	1	0	0	0	0
G	0	0	0	1	0	0
C	0	0	0	0	2	0
T	0	0	0	0	0	3

https://www.youtube.com/watch?v=lu9ScxSejSE

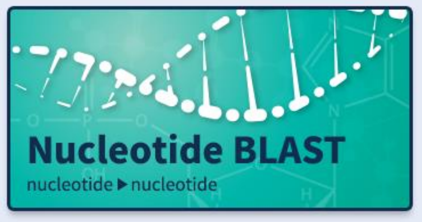
Match : 1 🟩
Mismatch : -1 🟥
GAP : -2 🟥

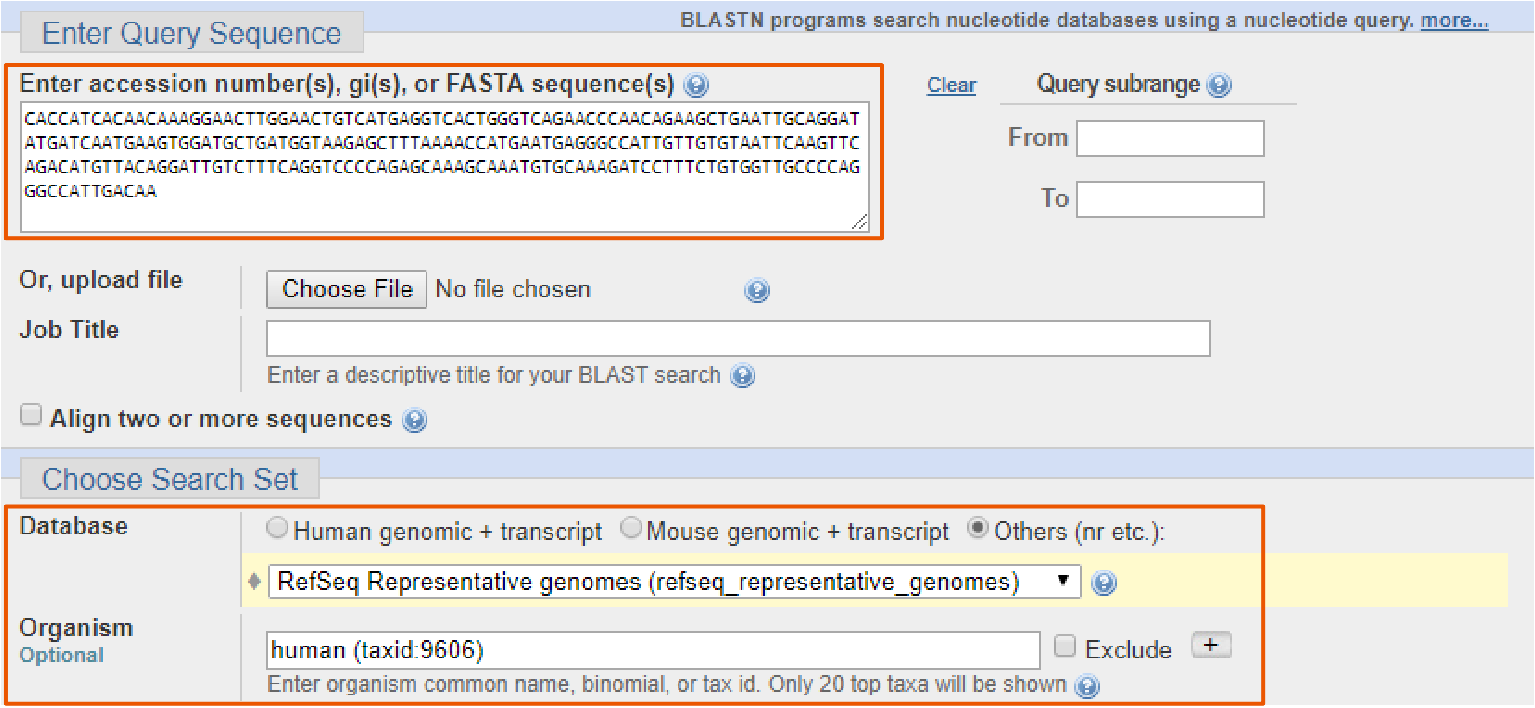
```
Seq1 :  ~~AT~~GCT
            | | |
Seq2 :   ~~A~~GCT
```

- Ignore possibilities with negative score
  - Start over is better

Basic Local Alignment Search Tool BLAST

NCBI’s nucleotide BLAST interface





Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. more...

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

```
CACCATCACAACAAAGGAACTTGGAACTGTCATGAGGTCACTGGGTCAGAACCCAACAGAAGCTGAATTGCAGGAT
ATGATCAATGAAGTGGATGCTGATGGTAAGAGCTTTAAAACCATGAATGAGGGCCATTGTTGTGTAATTCAAGTTC
AGACATGTTACAGGATTGTCTTTCAGGTCCCCAGAGCAAAGCAAATGTGCAAAGATCCTTTCTGTGGTTGCCCCAG
GGCCATTGACAA
```

Clear

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title:

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

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

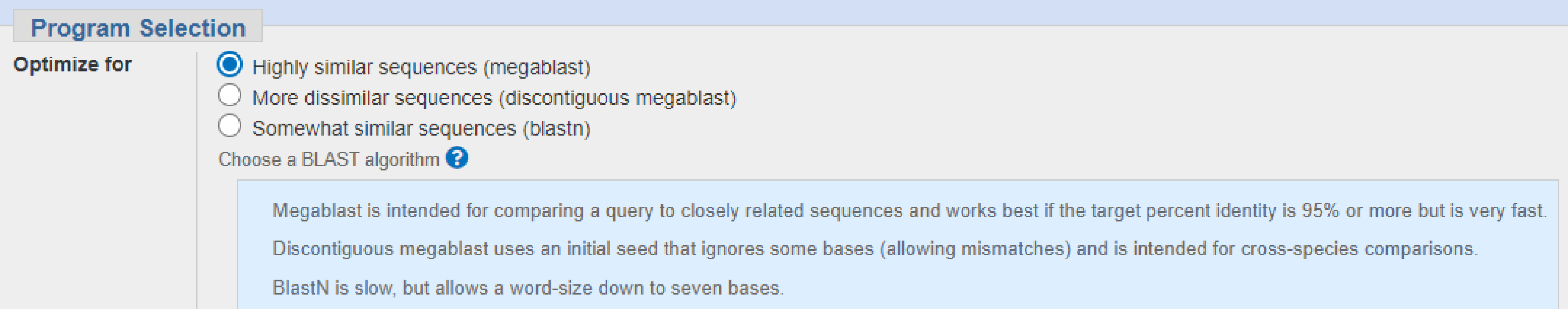
RefSeq Representative genomes (refseq_representative_genomes)

Organism
Optional

human (taxid:9606) □ Exclude +

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

Nucleotide BLAST algorithms



Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm ?

Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast.

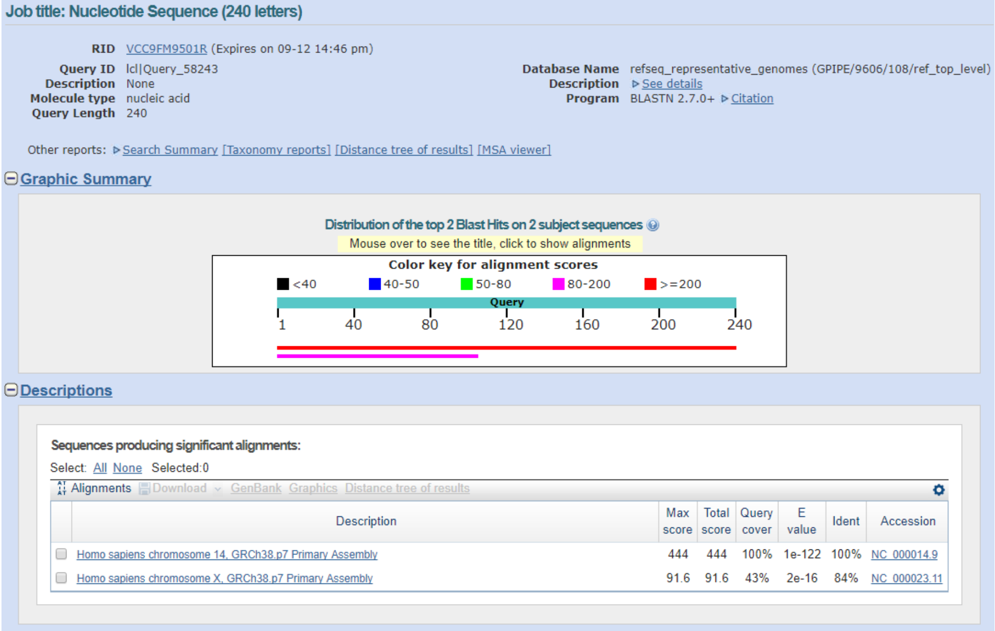
Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.

BlastN is slow, but allows a word-size down to seven bases.

- MEGABLAST: word size = 28, match/mismatch score = +1/−2, linear gap
- BLASTN: word size = 11, match/mismatch score = +2/−3, affine gap

MEGABLAST vs BLASTN

MEGABLAST = few, high-identity hits



Job title: Nucleotide Sequence (240 letters)

RID	VCC9FM9501R (Expires on 09-12 14:46 pm)		
Query ID	lcl|Query_58243	Database Name	refseq_representative_genomes (GPIPE/9606/108/ref_top_level)
Description	None	Description	▷ See details
Molecule type	nucleic acid	Program	BLASTN 2.7.0+ ▷ Citation
Query Length	240		

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [MSA viewer]

⊖ Graphic Summary

Distribution of the top 2 Blast Hits on 2 subject sequences

Mouse over to see the title, click to show alignments

⊖ Descriptions

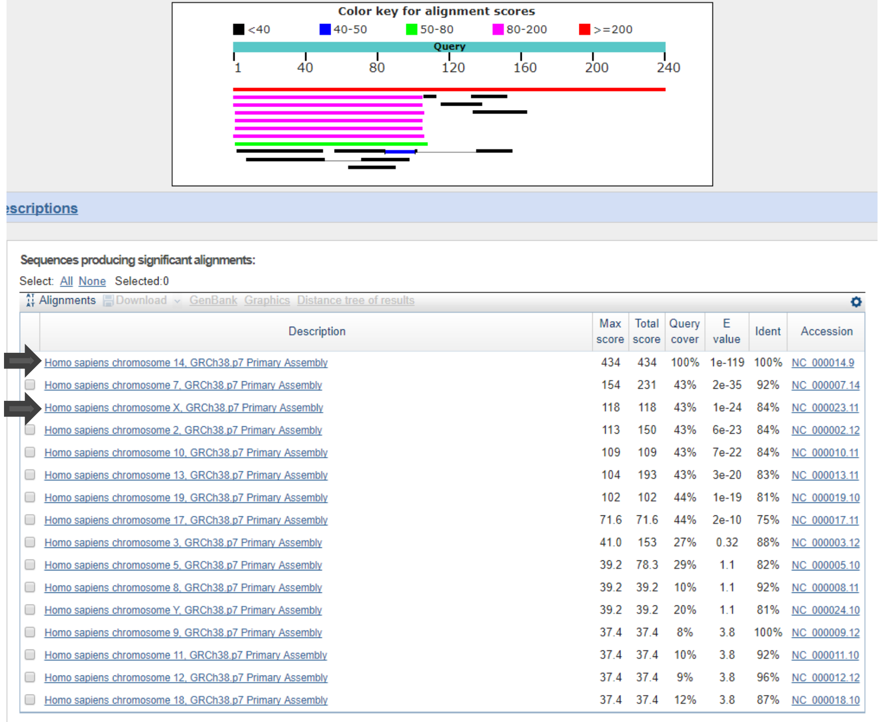
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly	444	444	100%	1e-122	100%	NC_000014.9
□	Homo sapiens chromosome X, GRCh38.p7 Primary Assembly	91.6	91.6	43%	2e-16	84%	NC_000023.11

BLASTN = lots of intermediate-identity hits



escriptions

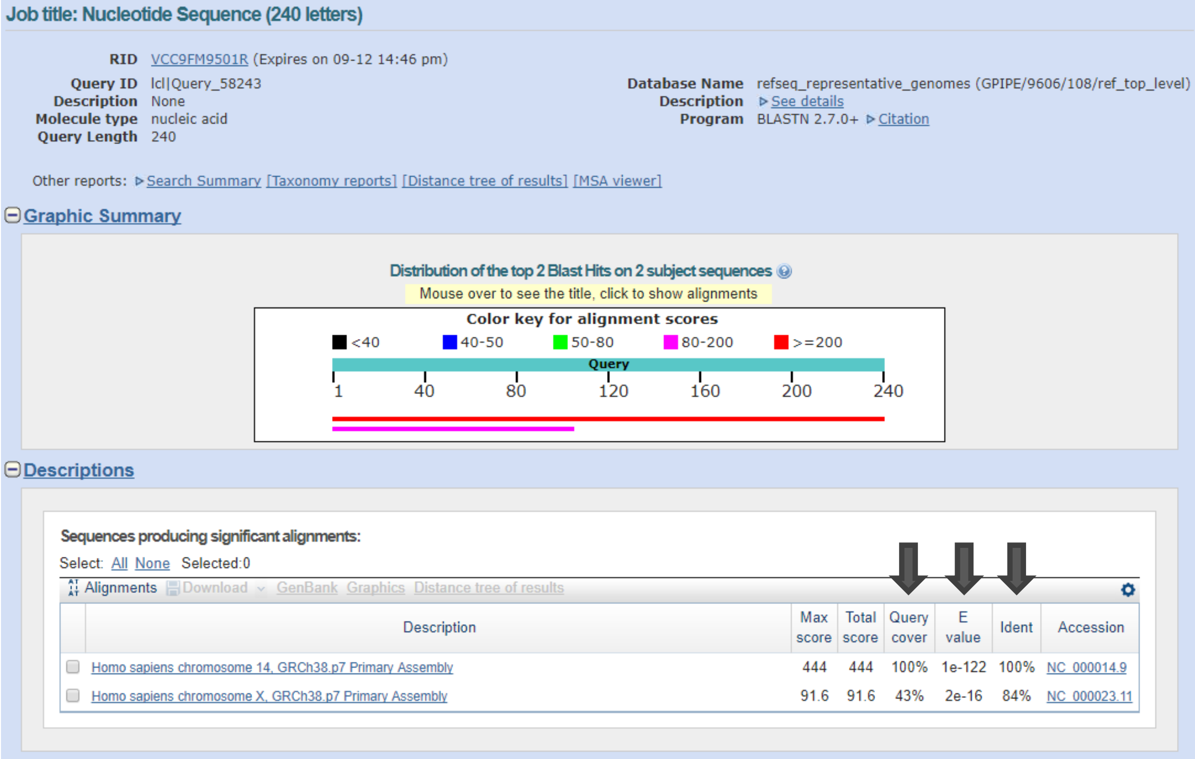
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
	Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly	434	434	100%	1e-119	100%	NC_000014.9
□	Homo sapiens chromosome 7, GRCh38.p7 Primary Assembly	154	231	43%	2e-35	92%	NC_000007.14
	Homo sapiens chromosome X, GRCh38.p7 Primary Assembly	118	118	43%	1e-24	84%	NC_000023.11
□	Homo sapiens chromosome 2, GRCh38.p7 Primary Assembly	113	150	43%	6e-23	84%	NC_000002.12
□	Homo sapiens chromosome 10, GRCh38.p7 Primary Assembly	109	109	43%	7e-22	84%	NC_000010.11
□	Homo sapiens chromosome 13, GRCh38.p7 Primary Assembly	104	193	43%	3e-20	83%	NC_000013.11
□	Homo sapiens chromosome 19, GRCh38.p7 Primary Assembly	102	102	44%	1e-19	81%	NC_000019.10
□	Homo sapiens chromosome 17, GRCh38.p7 Primary Assembly	71.6	71.6	44%	2e-10	75%	NC_000017.11
□	Homo sapiens chromosome 3, GRCh38.p7 Primary Assembly	41.0	153	27%	0.32	88%	NC_000003.12
□	Homo sapiens chromosome 5, GRCh38.p7 Primary Assembly	39.2	78.3	29%	1.1	82%	NC_000005.10
□	Homo sapiens chromosome 8, GRCh38.p7 Primary Assembly	39.2	39.2	10%	1.1	92%	NC_000008.11
□	Homo sapiens chromosome Y, GRCh38.p7 Primary Assembly	39.2	39.2	20%	1.1	81%	NC_000024.10
□	Homo sapiens chromosome 9, GRCh38.p7 Primary Assembly	37.4	37.4	8%	3.8	100%	NC_000009.12
□	Homo sapiens chromosome 11, GRCh38.p7 Primary Assembly	37.4	37.4	10%	3.8	92%	NC_000011.10
□	Homo sapiens chromosome 12, GRCh38.p7 Primary Assembly	37.4	37.4	9%	3.8	96%	NC_000012.12
□	Homo sapiens chromosome 18, GRCh38.p7 Primary Assembly	37.4	37.4	12%	3.8	87%	NC_000018.10

Interpreting BLAST result



Job title: Nucleotide Sequence (240 letters)

RID	VCC9FM9501R (Expires on 09-12 14:46 pm)		
Query ID	lcl|Query_58243	Database Name	refseq_representative_genomes (GPIPE/9606/108/ref_top_level)
Description	None	Description	▷ See details
Molecule type	nucleic acid	Program	BLASTN 2.7.0+ ▷ Citation
Query Length	240		

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [MSA viewer]

⊖ Graphic Summary

Distribution of the top 2 Blast Hits on 2 subject sequences

Mouse over to see the title, click to show alignments

⊖ Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly	444	444	100%	1e-122	100%	NC_000014.9
□	Homo sapiens chromosome X, GRCh38.p7 Primary Assembly	91.6	91.6	43%	2e-16	84%	NC_000023.11

Query coverage = % of input sequence used in the alignment

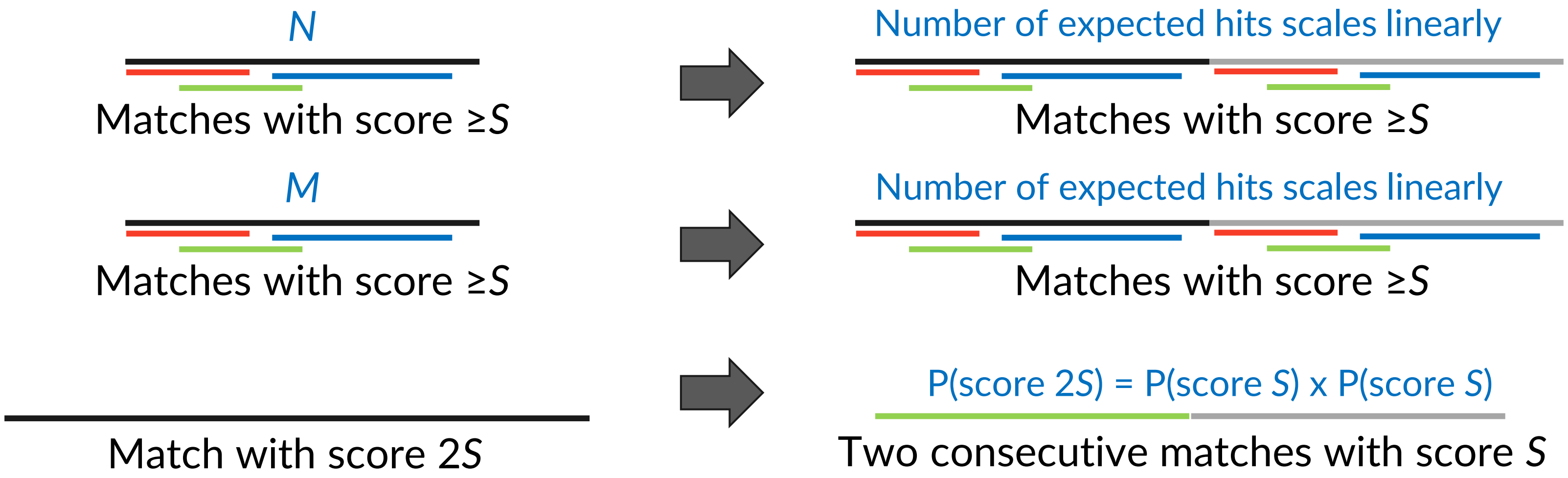
Identity = % of identity between input and matched sequences in the aligned region

E value = expected number of hits with the same or higher score by chance (given input length and database size)

Typical cutoff is 1e-5

Understanding E value

- Given an input sequence of length N and a reference sequence of length M
- E value for a hit with score S is proportional to N x M x $e^{-\lambda S}$



N

Matches with score $\geq S$

Number of expected hits scales linearly

Matches with score $\geq S$

M

Matches with score $\geq S$

Number of expected hits scales linearly

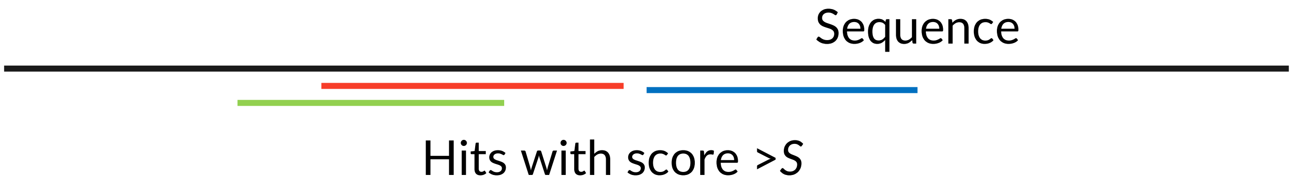
Matches with score $\geq S$

Match with score $2S$

P(score $2S$) = P(score S) x P(score S)

Two consecutive matches with score S

E value as Poisson distribution

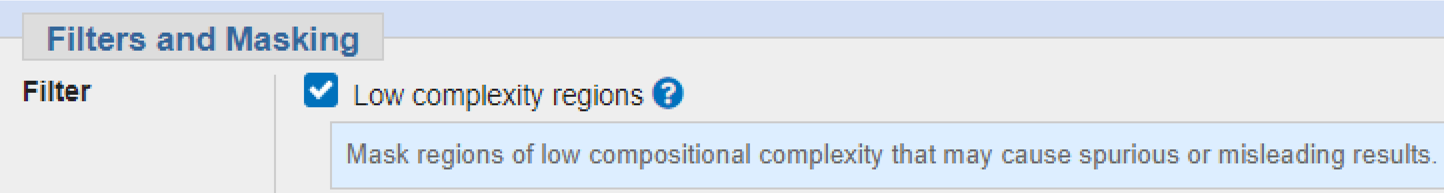


- Event of interest = hits with score >*S* occurs on the sequence of length *N*
- Expected value = E value
- Probability of observing *k* hits with score >*S* = $\frac{E^k e^{-E}}{k!}$

Low complexity region

CG island

CCCGCGCGCCCCGGCGCCCGATGCAACTAGC



Filters and Masking

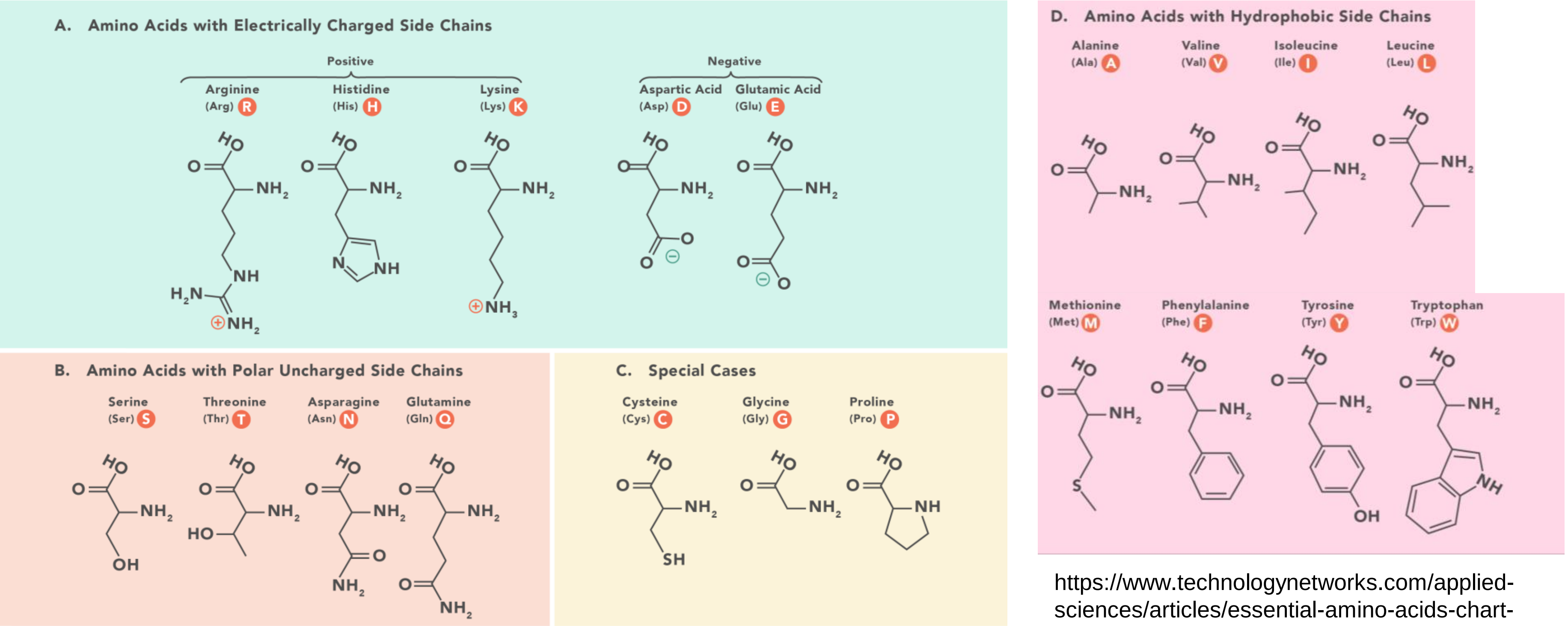
Filter ☑ Low complexity regions ?

Mask regions of low compositional complexity that may cause spurious or misleading results.

- Probability of getting a hit with score >*S* will be high if both sequences contain only C’s and G’s
- BLAST withholds these regions from score calculation

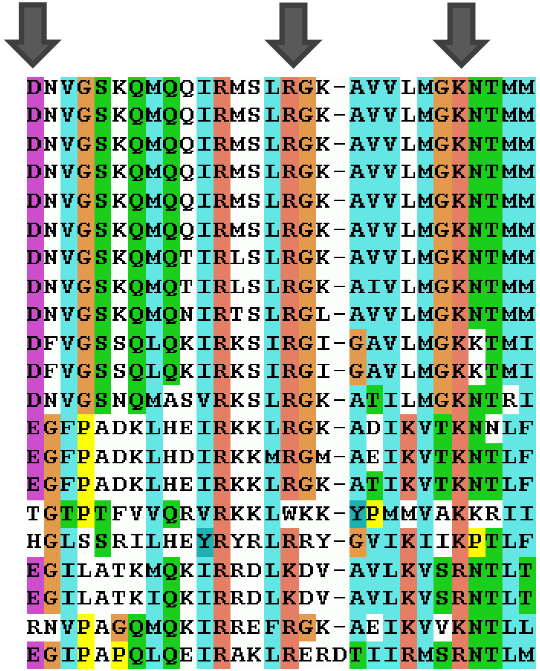
Protein sequence alignment

Amino acid side chains



https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357

Integrated Genomics Viewer (IGV)



```
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQTIRLSLRGK-AVVLMGKNTMM
DNVGSKQMQTIRLSLRGK-AIVLMGKNTMM
DNVGSKQMQNIRTSLRGL-AVVLMGKNTMM
DFVGSSQLQKIRKSIRGI-GAVLMGKKTMI
DFVGSSQLQKIRKSIRGI-GAVLMGKKTMI
DNVGSNQMASVRKSLRGK-ATILMGKNTRI
EGFPADKLHEIRKKLRGK-ADIKVTKNNLF
EGFPADKLHDIRKKMRGM-AEIKVTKNTLF
EGFPADKLHEIRKKLRGK-ATIKVTKNTLF
TGTPTFVVQRVRKKLWKK-YPMMVAKKRII
HGLSSRILHEYRYRLRRY-GVIKIIKPTLF
EGILATKMQKIRRDLKDV-AVLKVSRNTLT
EGILATKIQKIRRDLKDV-AVLKVSRNTLT
RNVPAGQMQKIRREFRGK-AEIKVVKNTLL
EGIPAPQLQEIRAKLRERDTIIRMSRNTLM
```

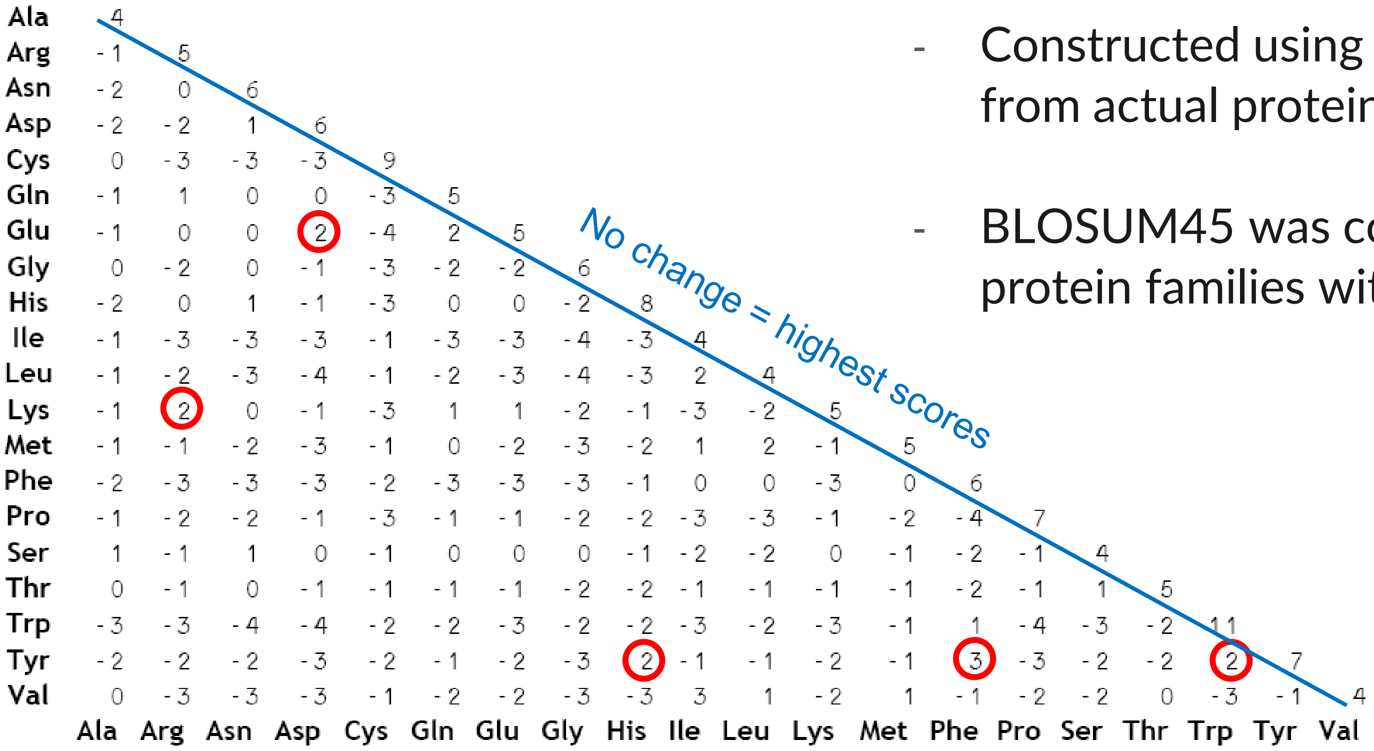
wikipedia.com

- Amino acids with similar properties can replace each other with minimal impact on protein function

- D, E have –COOH groups
- K, R have positively charged –NH_2 groups
- A, V, I, L have small hydrocarbon side chains
- F, Y, W have benzene rings

- Alignment score must reflect these!

Block Substitution Matrix (BLOSUM)



	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

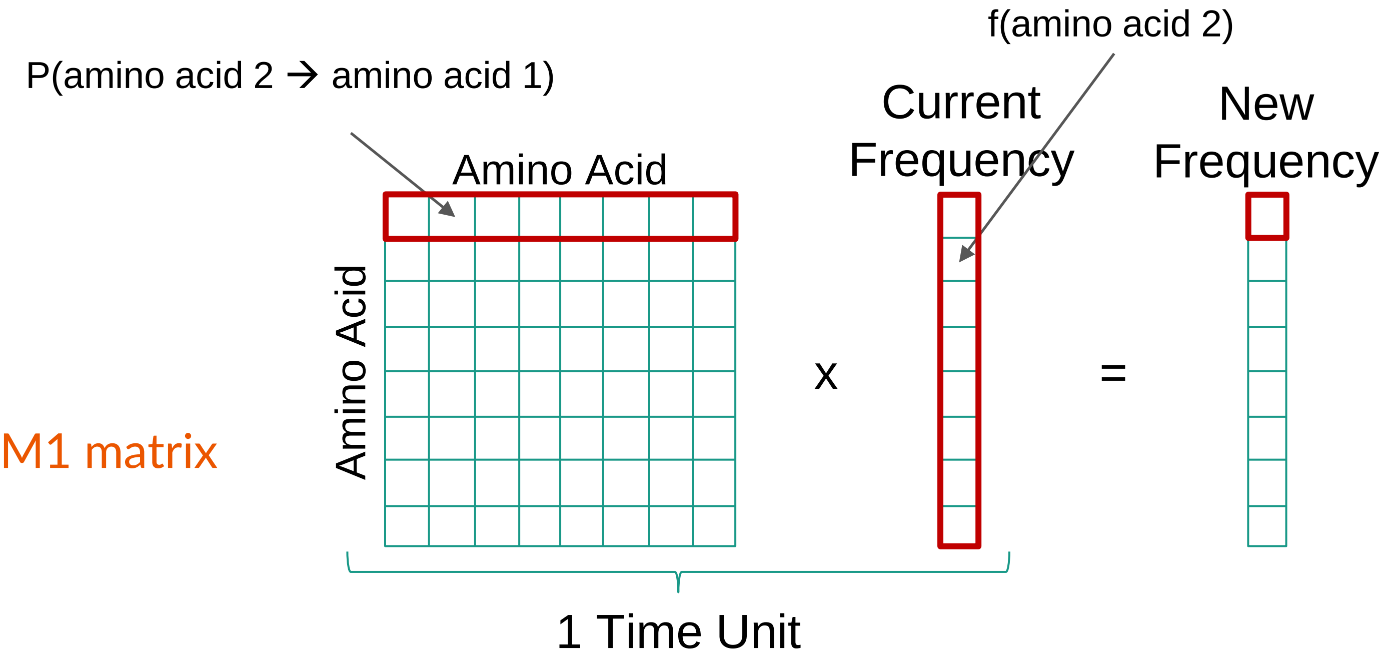
No change = highest scores

https://en.wikipedia.org/wiki/BLOSUM

- Constructed using substitution rates from actual protein families

- BLOSUM45 was constructed using protein families with >45% conservation

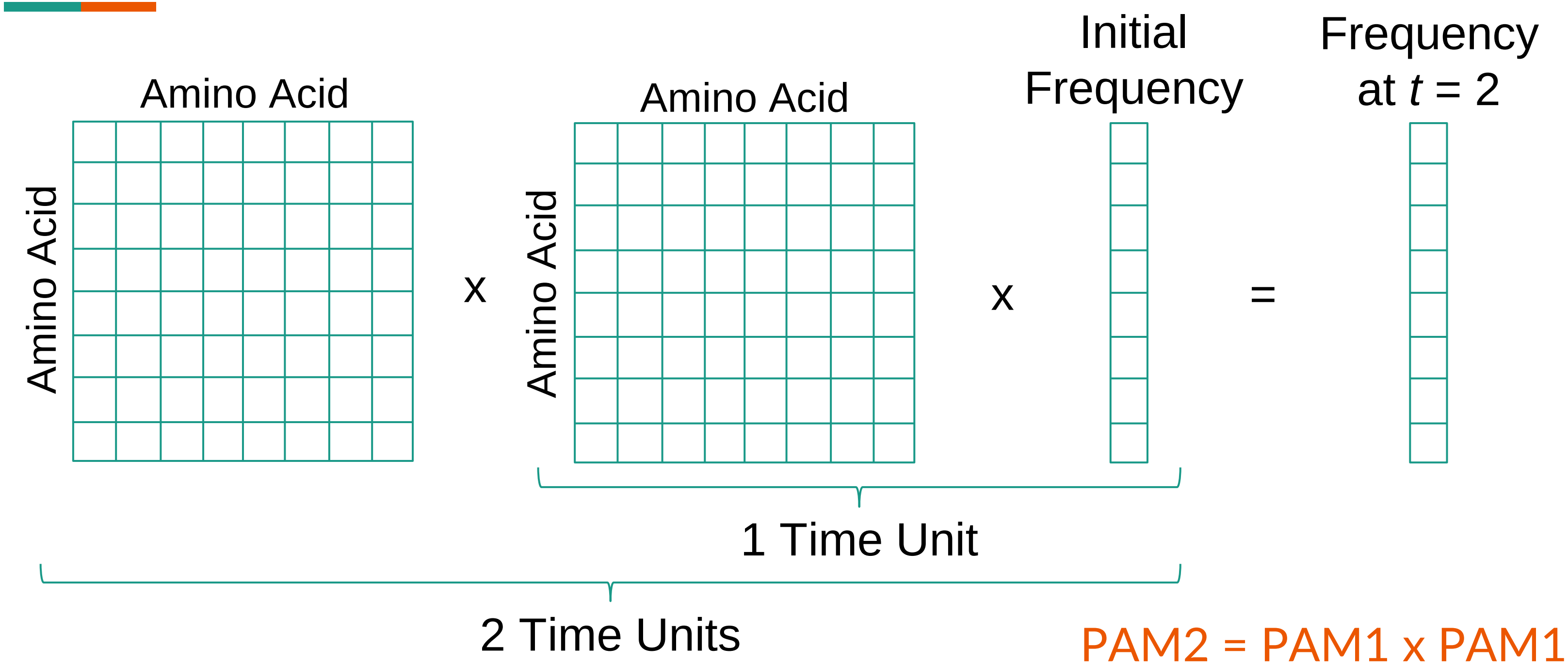
Point Accepted Mutation (PAM)



PAM1 matrix

- Estimate amino acid substitution rate between highly similar proteins (>85%)

Point Accepted Mutation (PAM)



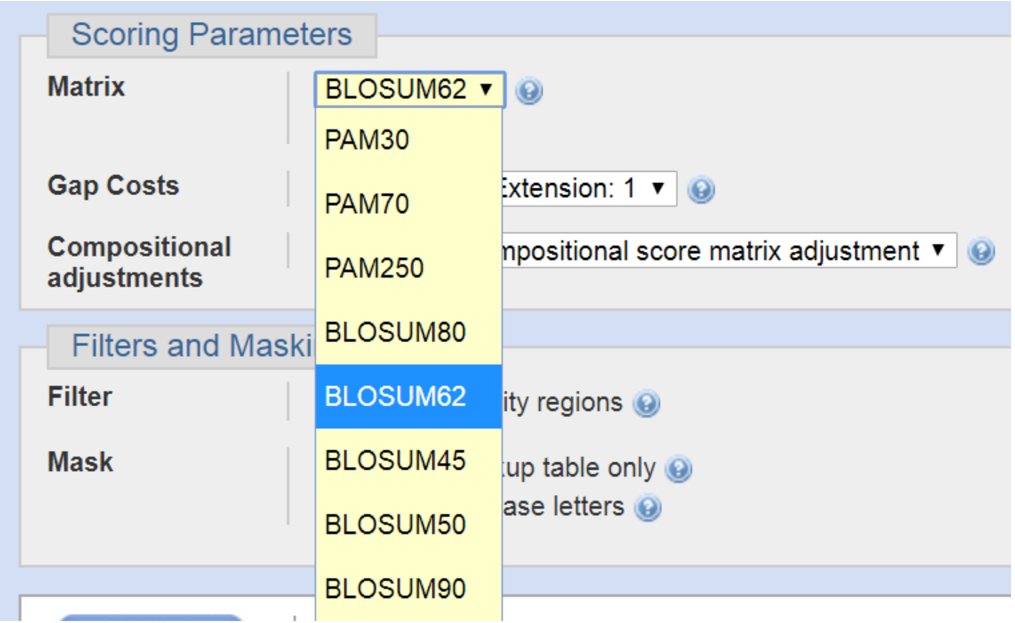
PAM2 = PAM1 x PAM1

- Extrapolate substitution rates for more distant proteins

PAM vs BLOSUM

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

Data from https://en.wikipedia.org/wiki/BLOSUM



Scoring Parameters

Matrix: BLOSUM62 ▼

Gap Costs: xtension: 1 ▼

Compositional adjustments: mpositional score matrix adjustment ▼

Filters and Maski

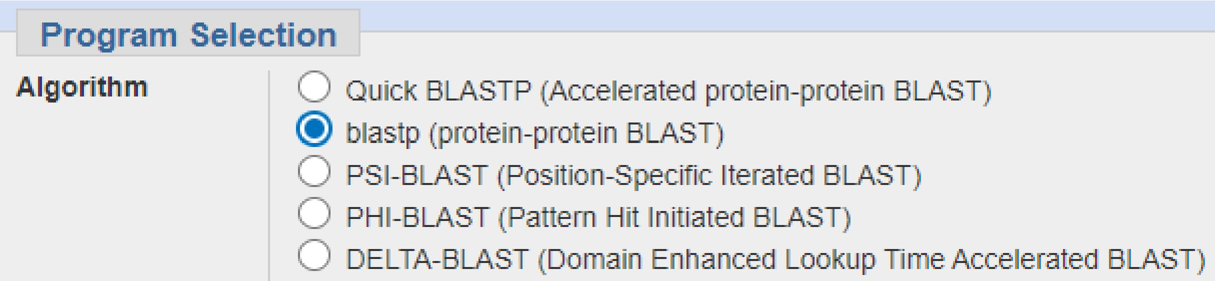
Filter: ity regions

Mask: up table only

ase letters

- BLOSUM for low identity, PAM for high identity

Protein BLAST algorithms



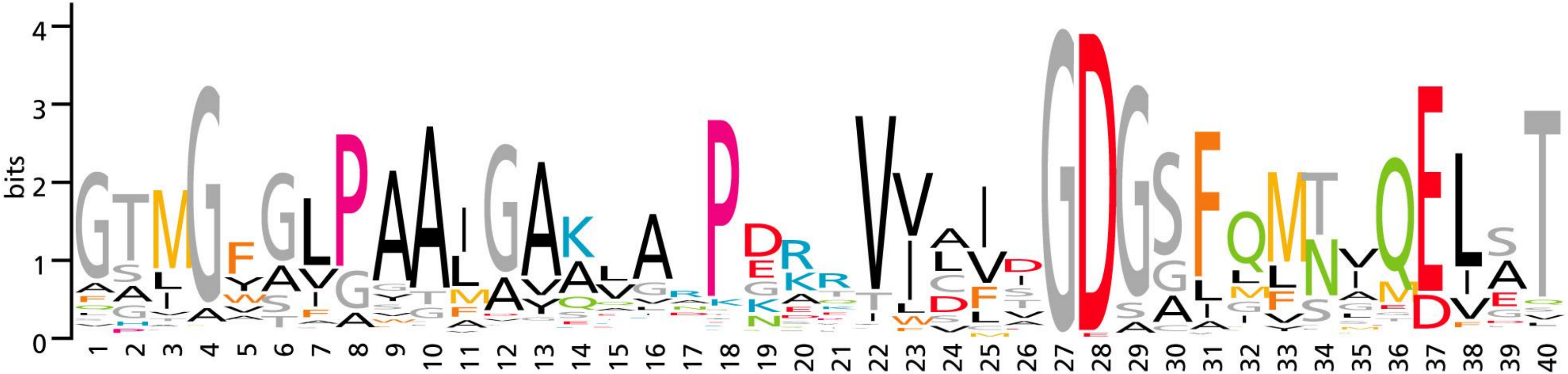
Program Selection

Algorithm

○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

- Standard BLASTP assumes that all amino acid residue positions are the same
- But there are protein domains & motifs with specific patterns

Position-specific scoring matrix (PSSM)



www.nemates.org/uky/520/Lecture/Lect6/BIO520_2010_Lect6.pp

weblogo.berkeley.edu

- Different scoring matrix for each position in the motif
- But how do we know the position-specific amino acid profile?

Pattern hit initiated (PHI-BLAST)

x = any amino acid

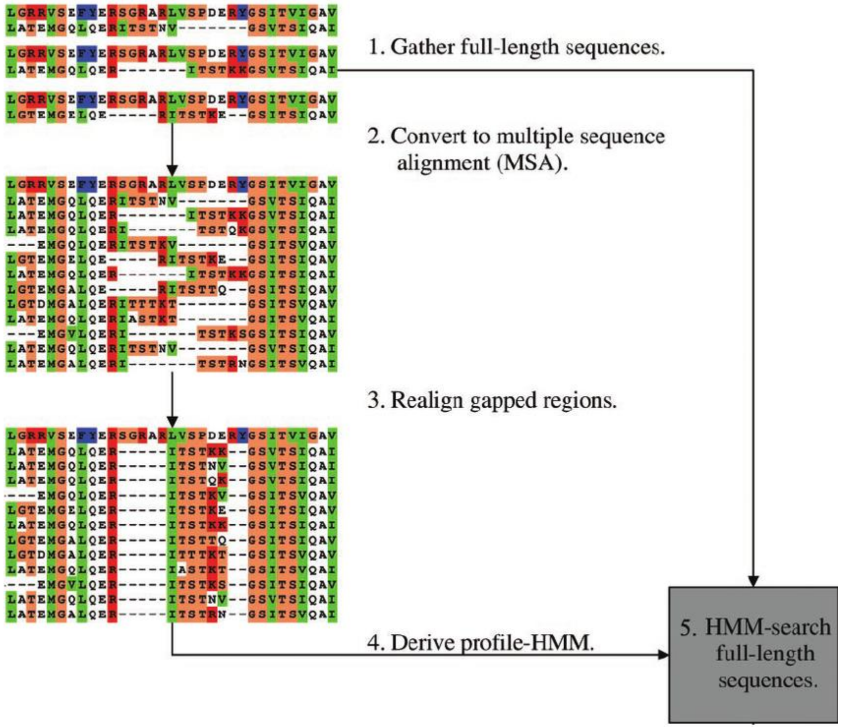
```
[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]
```

L, I, V, M, or F

any sequences of 5-11 amino acids

- Combine regular BLASTP with user-specified pattern
- Hits must be similar to the input sequence AND match the pattern
- Search for known protein domain

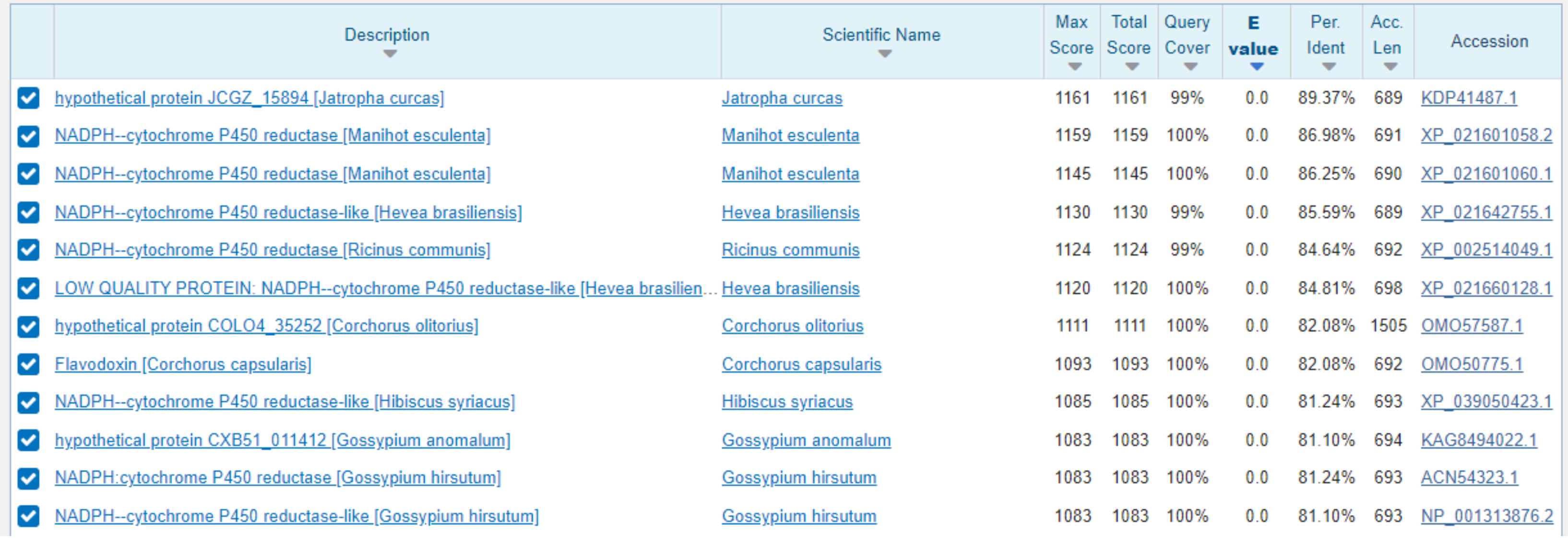
Position-specific iteratred (PSI-BLAST)


LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LATEMGQLQERITSTNV-------GSVTSIQAI
LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LATEMGQLQER-------ITSTKKGSVTSIQAI
LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LGTEMGELQE-----RITSTKE--GSITSIQAV
1. Gather full-length sequences.
2. Convert to multiple sequence alignment (MSA).
LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LATEMGQLQERITSTNV-------GSVTSIQAI
LATEMGQLQER-------ITSTKKGSVTSIQAI
LATEMGQLQERI-------TSTQKGSVTSIQAI
---EMGQLQERITSTKV-------GSITSVQAV
LGTEMGELQE-----RITSTKE--GSITSIQAV
LATEMGQLQER-------ITSTKKGSITSIQAI
LGTEMGALQE-----RITSTTQ--GSITSIQAV
LGTDMGALQERITTTKT-------GSITSVQAV
LATEMGQLQERIASTKT-------GSITSVQAI
---EMGVLQERI-------TSTKSGSITSIQAV
LATEMGQLQERITSTNV-------GSVTSIQAI
LATEMGALQERI-------TSTRNGSITSVQAI
3. Realign gapped regions.
LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LATEMGQLQER-----ITSTKK--GSVTSIQAI
LATEMGQLQER-----ITSTNV--GSVTSIQAI
LATEMGQLQER-----ITSTQK--GSVTSIQAI
---EMGQLQER-----ITSTKV--GSITSVQAV
LGTEMGELQER-----ITSTKE--GSITSIQAV
LATEMGQLQER-----ITSTKK--GSITSIQAI
LGTEMGALQER-----ITSTTQ--GSITSIQAV
LGTDMGALQER-----ITTTKT--GSITSVQAV
LATEMGQLQER-----IASTKT--GSITSVQAI
---EMGVLQER-----ITSTKS--GSITSIQAV
LATEMGQLQER-----ITSTNV--GSVTSIQAI
LATEMGALQER-----ITSTRN--GSITSVQAI
4. Derive profile-HMM.
5. HMM-search full-length sequences.

Frickey, T. and Lupas, A. NAR 32:5231-8 (2004)

- Start from user inputs
- First round of BLASTP
- Construct PSSM from hits
- Re-search using the PSSM
- Repeat

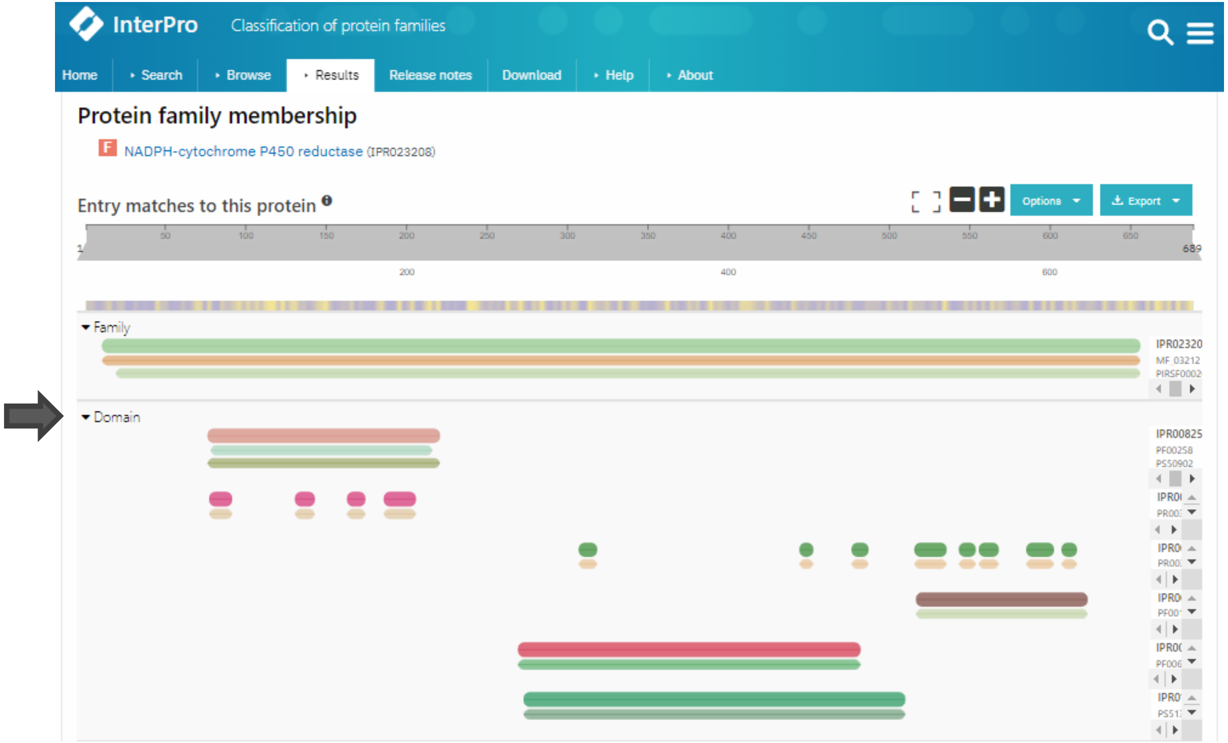
Using BLASTP to annotate protein function



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	hypothetical protein JCGZ_15894 [Jatropha curcas]	Jatropha curcas	1161	1161	99%	0.0	89.37%	689	KDP41487.1
☑	NADPH--cytochrome P450 reductase [Manihot esculenta]	Manihot esculenta	1159	1159	100%	0.0	86.98%	691	XP_021601058.2
☑	NADPH--cytochrome P450 reductase [Manihot esculenta]	Manihot esculenta	1145	1145	100%	0.0	86.25%	690	XP_021601060.1
☑	NADPH--cytochrome P450 reductase-like [Hevea brasiliensis]	Hevea brasiliensis	1130	1130	99%	0.0	85.59%	689	XP_021642755.1
☑	NADPH--cytochrome P450 reductase [Ricinus communis]	Ricinus communis	1124	1124	99%	0.0	84.64%	692	XP_002514049.1
☑	LOW QUALITY PROTEIN: NADPH--cytochrome P450 reductase-like [Hevea brasilien...	Hevea brasiliensis	1120	1120	100%	0.0	84.81%	698	XP_021660128.1
☑	hypothetical protein COLO4_35252 [Corchorus olitorius]	Corchorus olitorius	1111	1111	100%	0.0	82.08%	1505	OMO57587.1
☑	Flavodoxin [Corchorus capsularis]	Corchorus capsularis	1093	1093	100%	0.0	82.08%	692	OMO50775.1
☑	NADPH--cytochrome P450 reductase-like [Hibiscus syriacus]	Hibiscus syriacus	1085	1085	100%	0.0	81.24%	693	XP_039050423.1
☑	hypothetical protein CXB51_011412 [Gossypium anomalum]	Gossypium anomalum	1083	1083	100%	0.0	81.10%	694	KAG8494022.1
☑	NADPH:cytochrome P450 reductase [Gossypium hirsutum]	Gossypium hirsutum	1083	1083	100%	0.0	81.24%	693	ACN54323.1
☑	NADPH--cytochrome P450 reductase-like [Gossypium hirsutum]	Gossypium hirsutum	1083	1083	100%	0.0	81.10%	693	NP_001313876.2

- Suspected novel CYP reductase from an indigenous plant
- BLASTP against plant sequences
- >80% similarity to known and predicted CYP reductase class I

InterPro: Protein domain search



InterPro Classification of protein families

Home · Search · Browse · Results · Release notes · Download · Help · About

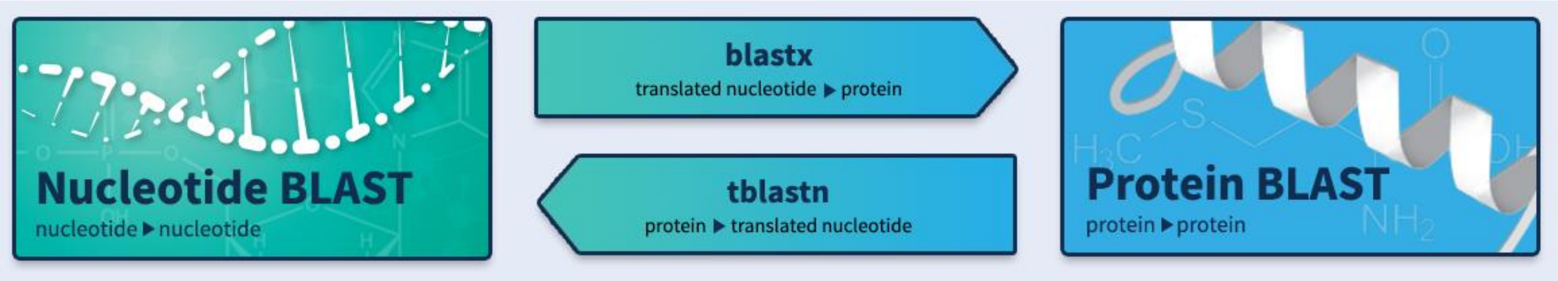
Protein family membership

F NADPH-cytochrome P450 reductase (IPR023208)

Entry matches to this protein

Mixing protein-nucleotide alignment

BLASTX and TBLASTN



- For alignment of coding DNA sequence
  - Codon structure = not all nucleotide positions evolve in the same manner
  - Similarity in protein is more informative than similarity in DNA
- Align translated DNA to protein database
- Align protein to translated DNA database

Example use cases

- BLASTX = align translated DNA to protein database
  - You perform RNA-seq
  - Unsure which open reading frame is correct
  - Check whether this RNA translated to known protein or function

- TBLASTN = align protein to translated DNA database
  - You identified novel protein
  - No evidence in protein database
  - But there might be transcriptomics studies that identified the RNA of related proteins

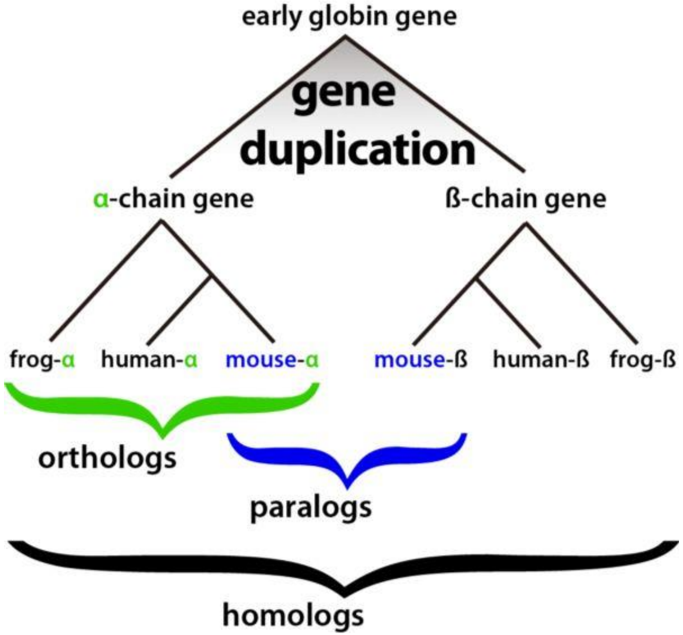
Beyond one-vs-all BLAST

All-vs-all BLAST

- Compare genes between related species to identify genes originated from a common ancestor
    - {Mouse-a, Human-a}, {Mouse-b, Human-b}

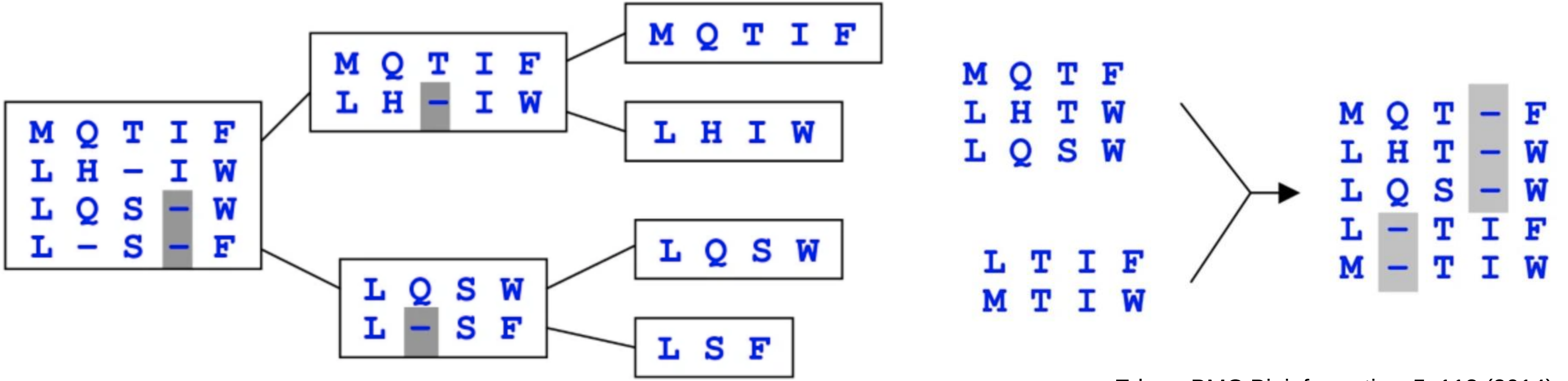
- BLAST mouse to human
- BLAST human to mouse

- Reciprocal best hit:
    - Human-a should be the best hit for Mouse-a
    - Mouse-a should be the best hit for Human-a



https://sites.google.com/site/jkim339n/part2a

Multiple sequence alignment (MSA)

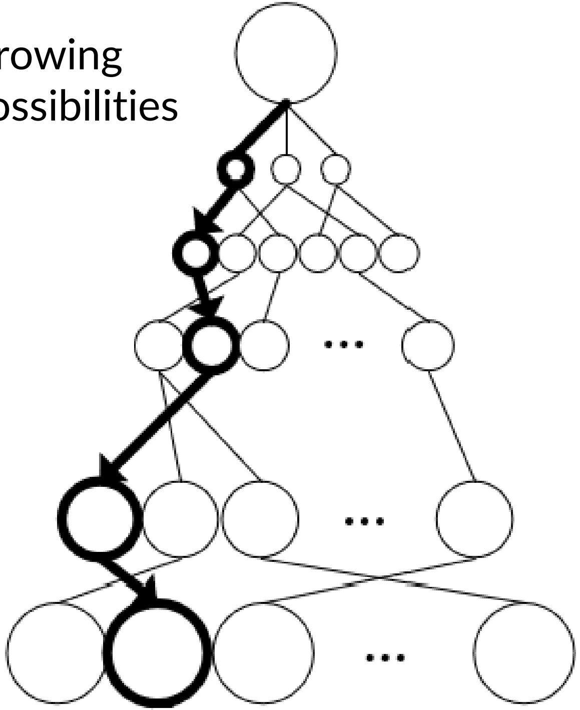


Edgar, BMC Bioinformatics, 5, 113 (2014)

- Dynamic programming is not feasible because of too many possibilities for grouping sequences
- Rely on heuristic algorithm

When the space of possible solutions is too large

Growing possibilities



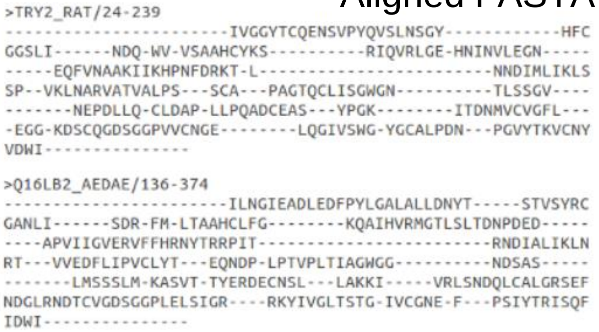
- **Heuristic** algorithm makes a decision by estimating the cost of all future steps

- **Greedy** algorithm makes a decision by optimizing the cost of only the next step

- **Randomized** algorithm makes a lot of random decisions and keeps the best one found

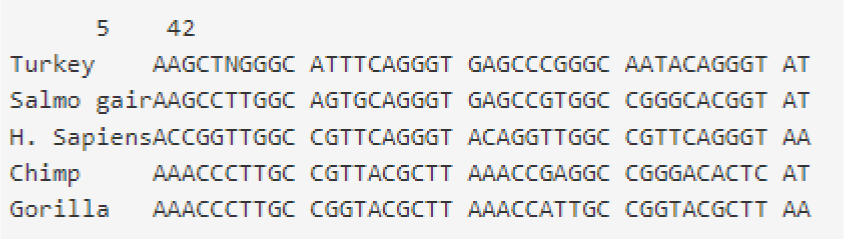
Alignment output format

Aligned FASTA



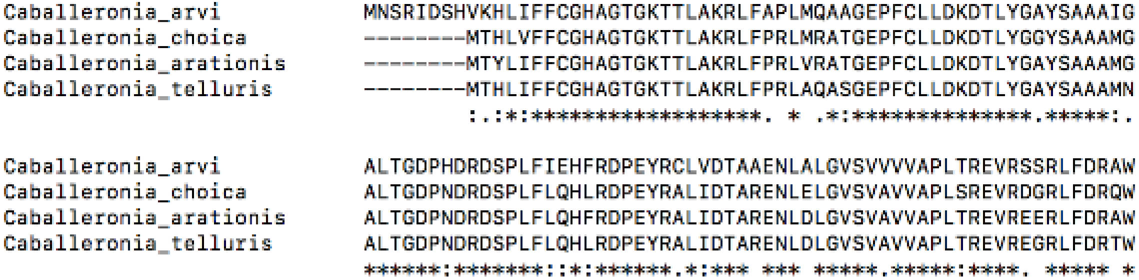
```
>TRY2_RAT/24-239
-----------------------IVGGYTCQENSVPYQVSLNSGY------------HFC
GGSLI------NDQ-WV-VSAAHCYKS----------RIQVRLGE-HNINVLEGN-----
-----EQFVNAAKIIKHPNFDRKT-L------------------------NNDIMLIKLS
SP--VKLNARVATVALPS---SCA---PAGTQCLISGWGN----------TLSSGV----
-------NEPDLLQ-CLDAP-LLPQADCEAS---YPGK--------ITDNMVCVGFL---
-EGG-KDSCQGDSGGPVVCNGE--------LQGIVSWG-YGCALPDN---PGVYTKVCNY
VDWI---------------

>Q16LB2_AEDAE/136-374
-----------------------ILNGIEADLEDFPYLGALALLDNYT-----STVSYRC
GANLI------SDR-FM-LTAAHCLFG--------KQAIHVRMGTLSLTDNPDED-----
----APVIIGVERVFFHRNYTRRPIT------------------------RNDIALIKLN
RT---VVEDFLIPVCLYT---EQNDP-LPTVPLTIAGWGG----------NDSAS-----
-------LMSSSLM-KASVT-TYERDECNSL---LAKKI-----VRLSNDQLCALGRSEF
NDGLRNDTCVGDSGGPLELSIGR----RKYIVGLTSTG-IVCGNE-F---PSIYTRISQF
IDWI---------------
```

PHYLIP



```
      5    42
Turkey    AAGCTNGGGC ATTTCAGGGT GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT ACAGGTTGGC CGTTCAGGGT AA
Chimp     AAACCCTTGC CGTTACGCTT AAACCGAGGC CGGGACACTC AT
Gorilla   AAACCCTTGC CGGTACGCTT AAACCATTGC CGGTACGCTT AA
```

ClustalW



```
Caballeronia_arvi           MNSRIDSHVKHLIFFCGHAGTGKTTLAKRLFAPLMQAAGEPFCLLDKDTLYGAYSAAAIG
Caballeronia_choica         --------MTHLVFFCGHAGTGKTTLAKRLFPRLMRATGEPFCLLDKDTLYGGYSAAAMG
Caballeronia_arationis      --------MTYLIFFCGHAGTGKTTLAKRLFPRLVRATGEPFCLLDKDTLYGAYSAAAMG
Caballeronia_telluris       --------MTHLIFFCGHAGTGKTTLAKRLFPRLAQASGEPFCLLDKDTLYGAYSAAAMN
                                    :.:*:******************. * .*:**************.*****:.

Caballeronia_arvi           ALTGDPHDRDSPLFIEHFRDPEYRCLVDTAAENLALGVSVVVVAPLTREVRSSRLFDRAW
Caballeronia_choica         ALTGDPNDRDSPLFLQHLRDPEYRALIDTARENLELGVSVAVVAPLSREVRDGRLFDRQW
Caballeronia_arationis      ALTGDPNDRDSPLFLQHFRDPEYRALIDTARENLDLGVSVAVVAPLTREVREERLFDRAW
Caballeronia_telluris       ALTGDPNDRDSPLFLQHLRDPEYRALIDTARENLDLGVSVAVVAPLTREVREGRLFDRTW
                            ******:*******::*:******.*:*** *** *****.*****:****. ***** *
```

Any question?

- See you on September 2nd 1-2:30pm