3000788 Intro to Comp Molec Biol

Lecture 2: Sequence alignment

Fall 2025

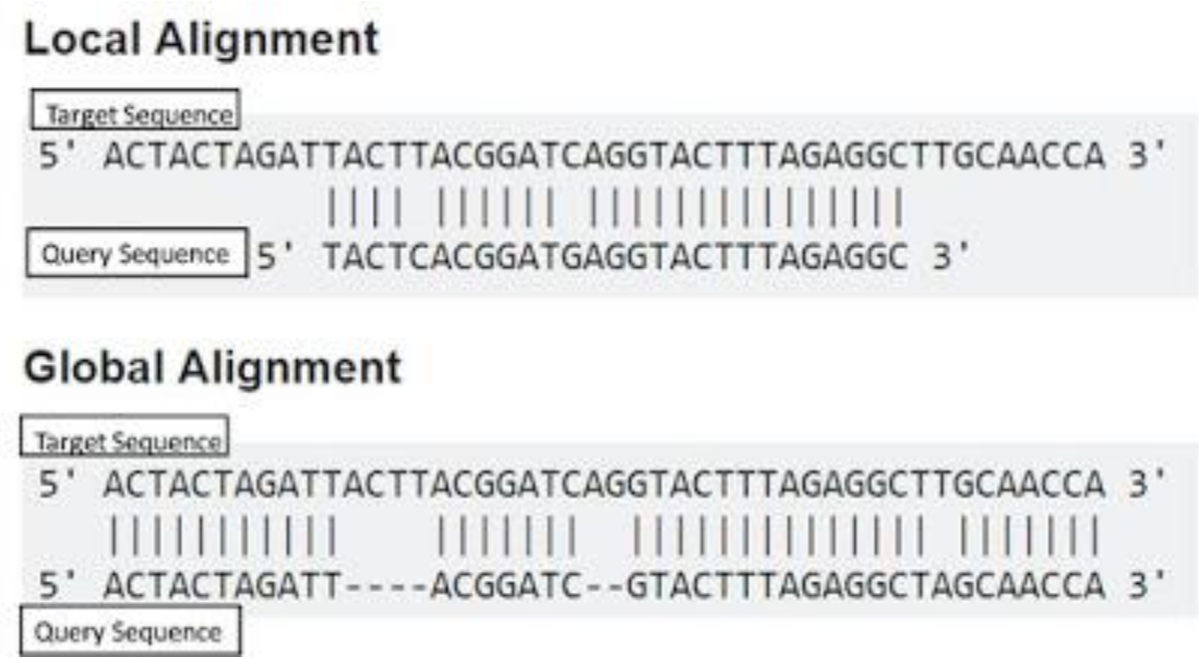
Sira Sriswasdi, PhD

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

Today’s agenda

- What is sequence alignment?
- Why do we perform sequence alignment?
- Strategies/algorithms for sequence alignment

Global and local sequence alignments



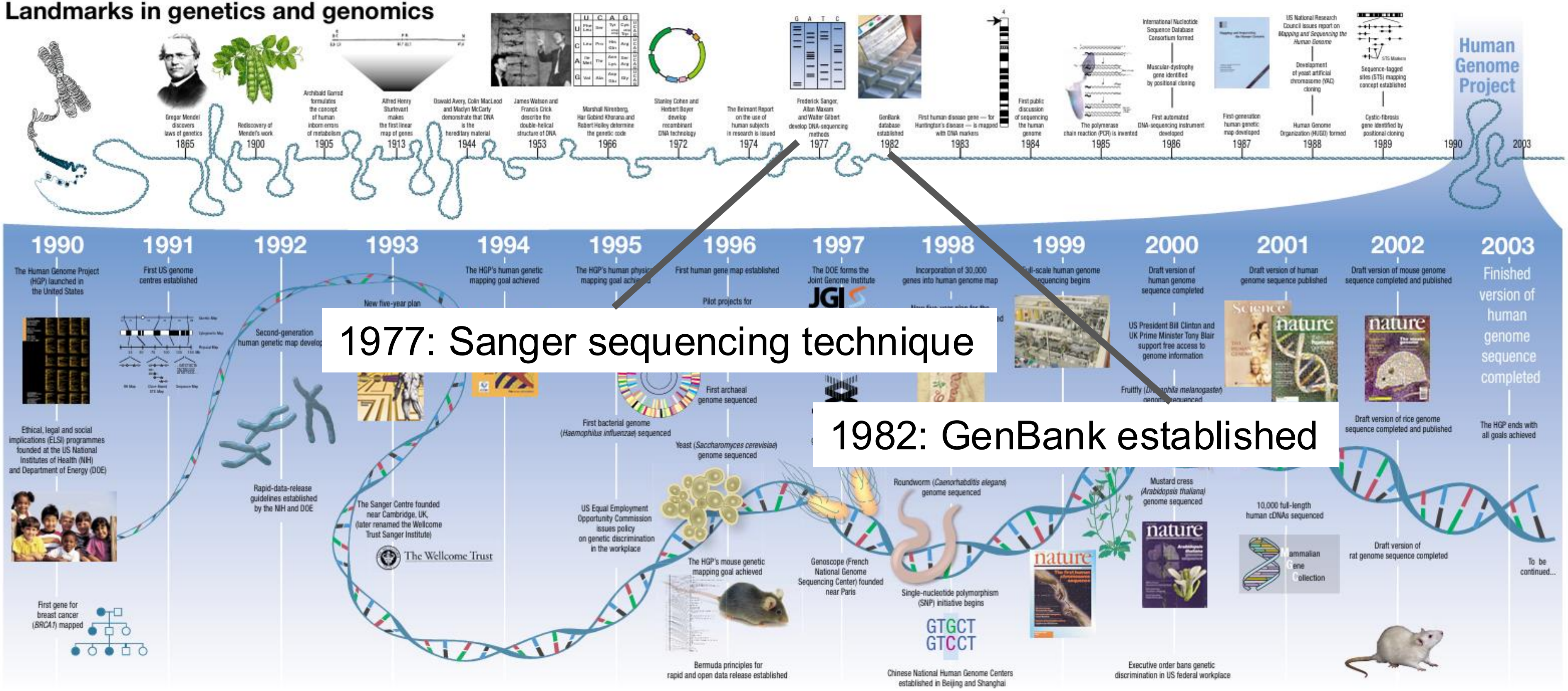
Local Alignment

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

Global Alignment

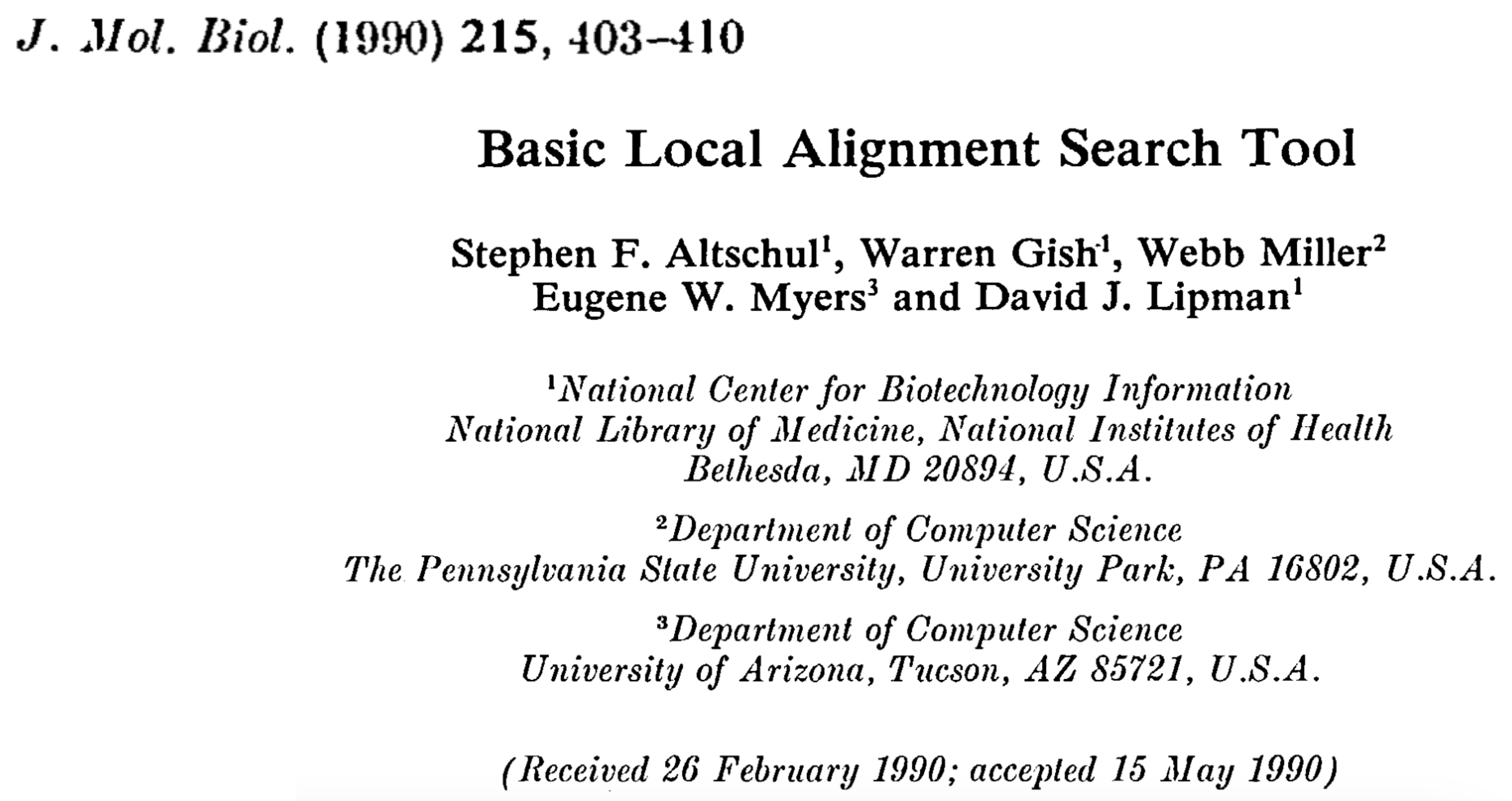
```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
Query Sequence
```

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


Landmarks in genetics and genomics
Gregor Mendel discovers laws of genetics
1865
Rediscovery of Mendel's work
1900
Archibald Garrod formulates the concept of human inborn errors of metabolism
1905
Alfred Henry Sturtevant makes the first linear map of genes
1913
Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrate that DNA is the hereditary material
1944
James Watson and Francis Crick describe the double-helical structure of DNA
1953
Marshall Nirenberg, Har Gobind Khorana and Robert Holley determine the genetic code
1966
Stanley Cohen and Herbert Boyer develop recombinant DNA technology
1972
The Belmont Report on the use of human subjects in research is issued
1974
G A T C
Frederick Sanger, Allan Maxam and Walter Gilbert develop DNA-sequencing methods
1977
GenBank database established
1982
First human disease gene — for Huntington's disease — is mapped with DNA markers
1983
First public discussion of sequencing the human genome
1984
The polymerase chain reaction (PCR) is invented
1985
International Nucleotide Sequence Database Consortium formed
Muscular-dystrophy gene identified by positional cloning
First automated DNA-sequencing instrument developed
1986
First-generation human genetic map developed
1987
US National Research Council issues report on *Mapping and Sequencing the Human Genome*
Development of yeast artificial chromosome (YAC) cloning
Human Genome Organization (HUGO) formed
1988
STS Markers
Sequence-tagged sites (STS) mapping concept established
Cystic-fibrosis gene identified by positional cloning
1989
1990
2003
Human Genome Project
1990
The Human Genome Project (HGP) launched in the United States
Ethical, legal and social implications (ELSI) programmes founded at the US National Institutes of Health (NIH) and Department of Energy (DOE)
First gene for breast cancer (*BRCA1*) mapped
1991
First US genome centres established
1992
Second-generation human genetic map develop
Rapid-data-release guidelines established by the NIH and DOE
1993
New five-year plan
The Sanger Centre founded near Cambridge, UK, (later renamed the Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
The HGP's human genetic mapping goal achieved
1995
The HGP's human physic mapping goal achi
First bacterial genome (*Haemophilus influenzae*) sequenced
US Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for
First archaeal genome sequenced
Yeast (*Saccharomyces cerevisiae*) genome sequenced
The HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
1997
The DOE forms the Joint Genome Institute
JGI
Genoscope (French National Genome Sequencing Center) founded near Paris
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the
Roundworm (*Caenorhabditis elegans*) genome sequenced
Single-nucleotide polymorphism (SNP) initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers established in Beijing and Shanghai
1999
ull-scale human genome equencing begins
nature
2000
Draft version of human genome sequence completed
US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information
Fruitfly (*Drosophila melanogaster*) genome sequenced
Mustard cress (*Arabidopsis thaliana*) genome sequenced
nature
Executive order bans genetic discrimination in US federal workplace
2001
Draft version of human genome sequence published
Science
nature
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
Draft version of rice genome sequence completed and published
Draft version of rat genome sequence completed
2003
Finished version of human genome sequence completed
The HGP ends with all goals achieved
To be continued...

1977: Sanger sequencing technique

1982: GenBank established



J. Mol. Biol. (1990) **215**, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]*National Center for Biotechnology Information*
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]*Department of Computer Science*
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]*Department of Computer Science*
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

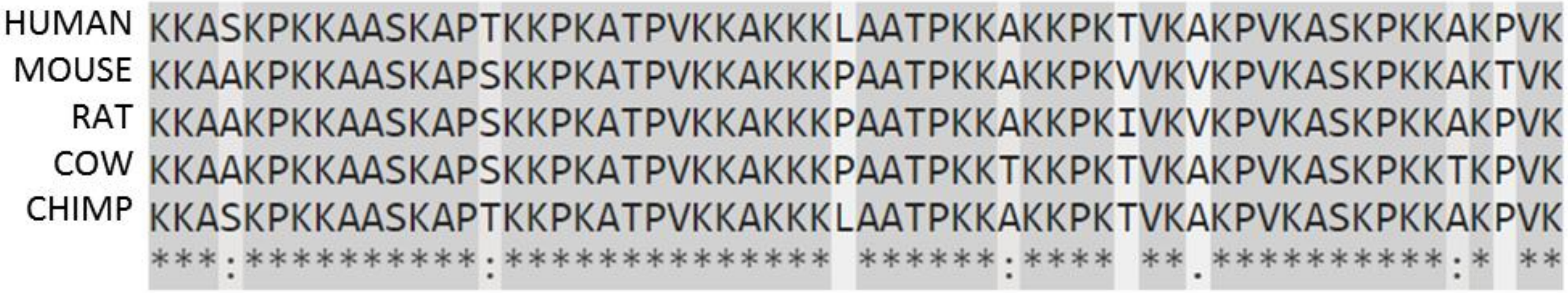
An early day of sequence alignment

- Found new cell extracts with interesting functions
- Cloned the genes, but which are responsible for the functions?
- Check the DNA sequences against known genes with validated functional experiments
  - Infer function
  - Infer taxonomy

Implications of sequence homology

Evolution occurs at the sequence level

Histone H1 (residues 120-180)

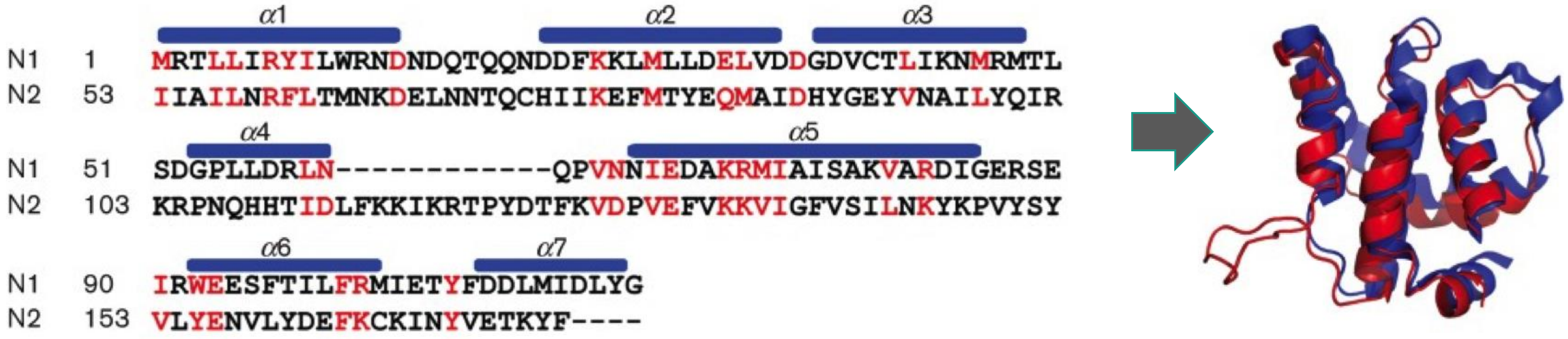


```
HUMAN KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK
  RAT KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKAKPVK
  COW KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
CHIMP KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK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

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

- Evolutionarily-related genes/proteins have similar sequences
- Evolutionarily-related genes/proteins tend to have similar functions
- Hence, genes/proteins with similar sequences may have similar functions, and may come from closely related organisms

Evolution tolerates conservation of function/structure

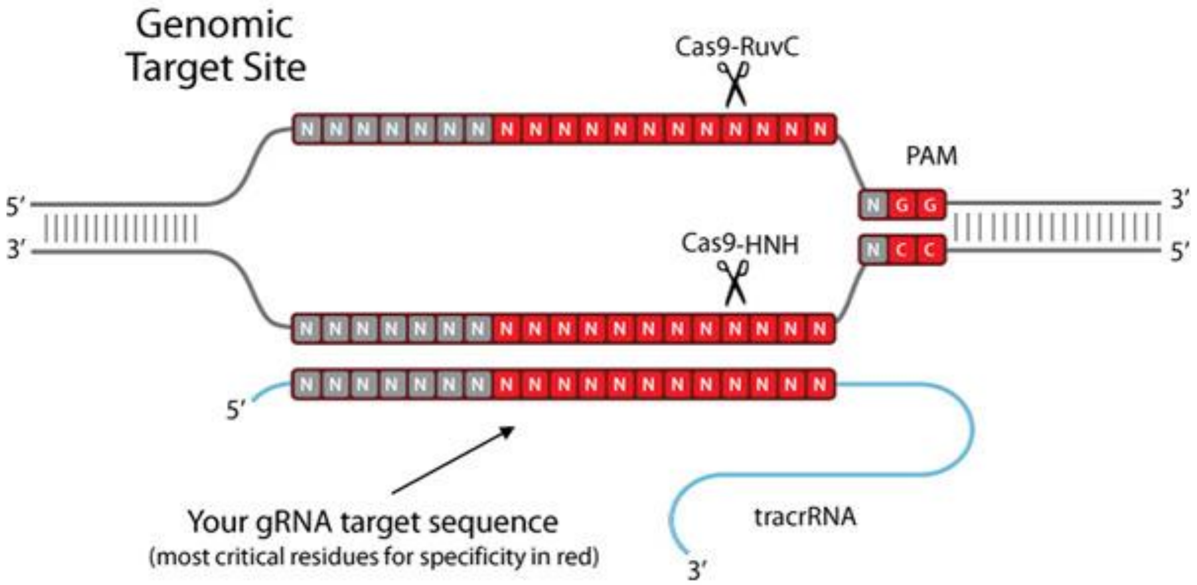


```
                  α1                                  α2                    α3
N1   1    MRTLLIRYILWRNDNDQTQQNDDFKKLMLLDELVDDGDVCTLIKNMRMTL
N2   53   IIAILNRFLTMNKDELNNTQCHIIKEFMTYEQMAIDHYGEYVNAILYQIR
              α4                              α5
N1   51   SDGPLLDRLN------------QPVNNIEDAKRMIAISAKVARDIGERSE
N2   103  KRPNQHHTIDLFKKIKRTPYDTFKVDPVEFVKKVIGFVSILNKYKPVYSY
                α6              α7
N1   90   IRWEESFTILFRMIETYFDDLMIDLYG
N2   153  VLYENVLYDEFKCKINYVETKYF----
```

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

- Sequences can change if the function and structure remain intact
- **Hence, proteins with similar sequences may adopt similar 3D structure, down to the specific domains**

Molecular probe design

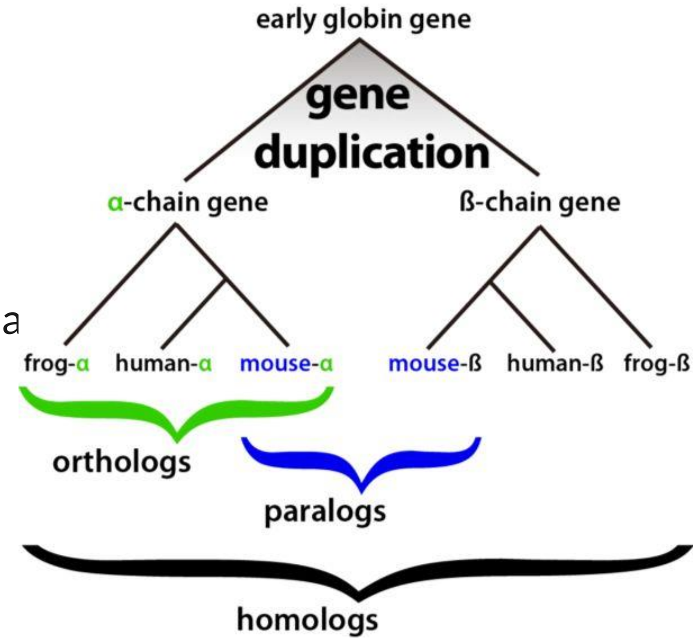

Genomic Target Site
Cas9-RuvC
N N N N N N N N N N N N N N N N N N N N
PAM
5′
3′
N G G
3′
N C C
5′
Cas9-HNH
N N N N N N N N N N N N N N N N N N N N
N N N N N N N N N N N N N N N N N N N N
5′
Your gRNA target sequence
(most critical residues for specificity in red)
tracrRNA
3′

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

- Sequence alignment confirm the **specificity** of your probes
  - Against targets and potential off-targets

Sequence alignment predicts many things

- 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 databa
  - 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

Sequence alignment strategies

Starting from exact match (seed / word)

..TTACGATAAGC**ATTTTCATAATA**CGACGTCA..

..AACCGACGT**ATTTTCATAATA**GCATAGCAT..

- **ATTTTCATAATA** pattern appears in both sequences
- Do you think the best alignment **must** have this at the center? Why?

.....ATTTTCATAATA....

.....ATTTTCATAATA....

How long of a match can be expected?

- Input sequence length = 300
- Expected 95% similarity (genome re-sequencing)

- How many matches and mismatches to expect? **285 and 15**
  - We can expect to see at least (285/16) **18 matches in a row**

MM...MEM...MEM........MEM...MM

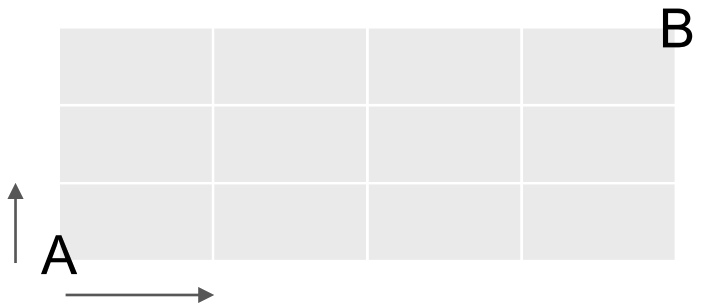
  - NCBI's MEGABLAST searches for a run of 28 matches

Dynamic programming

No. of rocks	1	2	3	4	5	6	7	8	9	10
Winner	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
No. of rocks	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Winner	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	?

- There is a pile of 20 rocks.
- Two players take turns by removing 1 or 2 rocks from the pile. Whoever removes the last rock(s) win.
- Who is the winner?

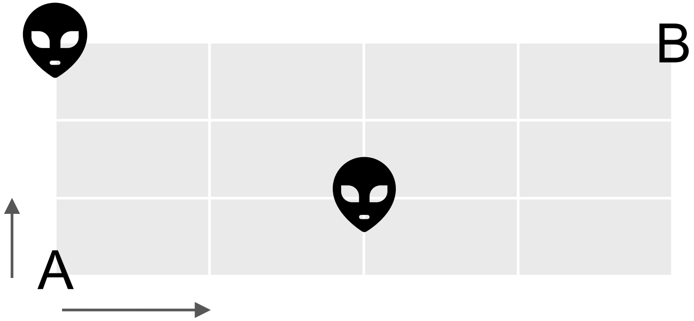
Dynamic programming



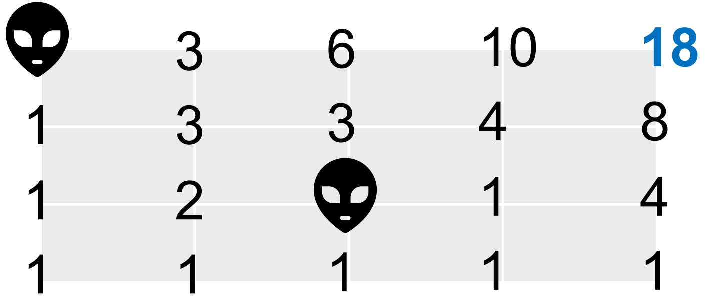
- How many ways to travel from A to B using only **right** and **up**?
  - (7 choose 3) = 35 ways

1	4	10	20	**35**
1	3	6	10	15
1	2	3	4	5
1	1	1	1	1

Dynamic programming

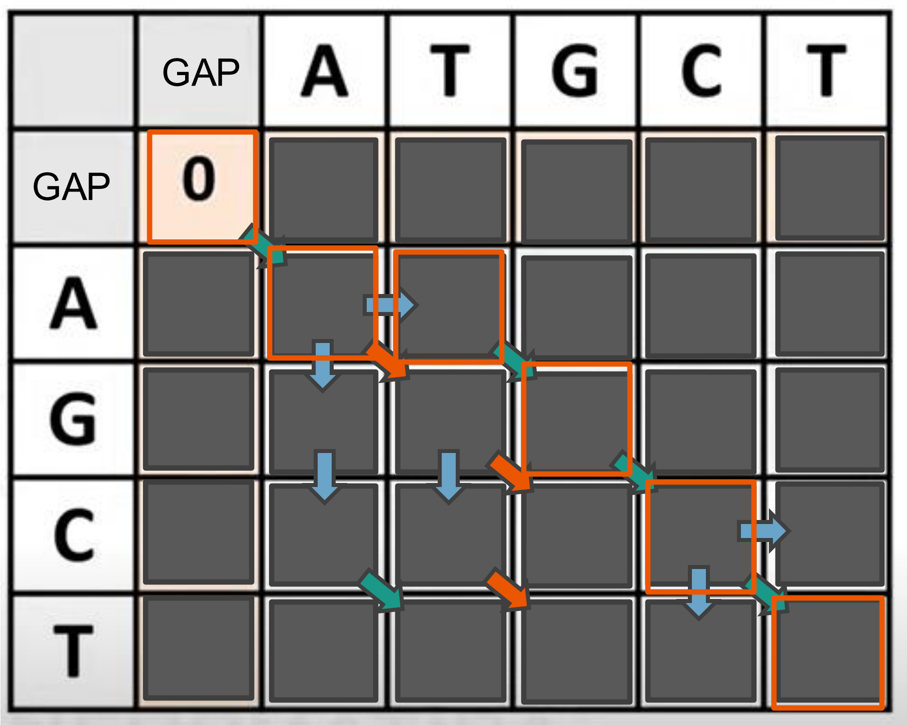


- How many ways to travel from A to B using only **right** and **up**?
  - Without running into aliens



👽	3	6	10	**18**
1	3	3	4	8
1	2	👽	1	4
1	1	1	1	1

Global sequence alignment

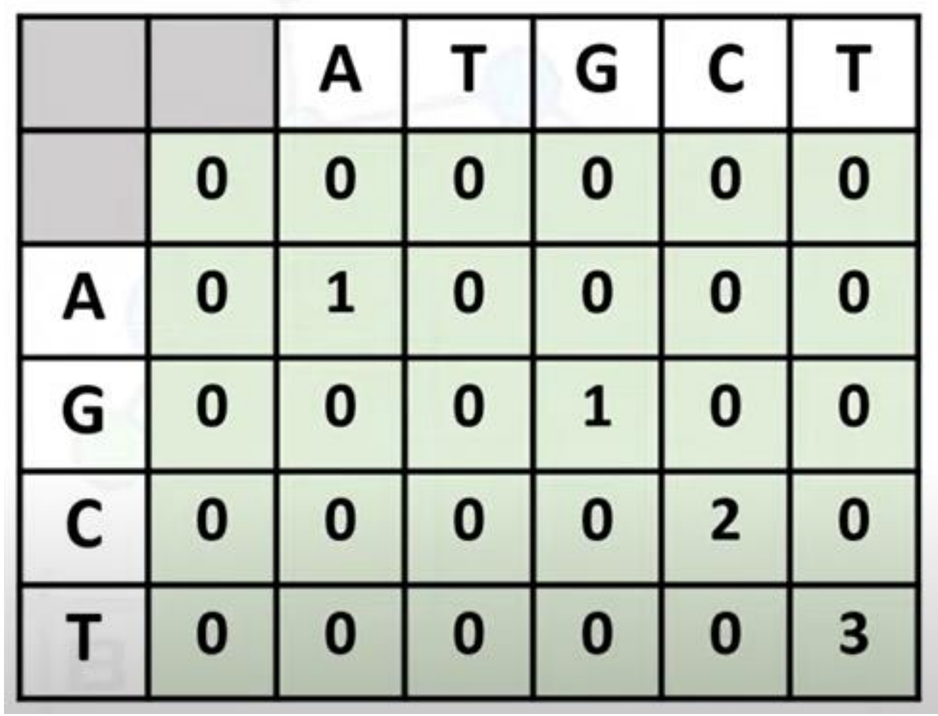


https://www.youtube.com/watch?v=ipp-pNRIp4g

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

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

Local sequence alignment = reset when score <0



		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

- Starting over is better than negative score

Dynamic programming for sequence alignment

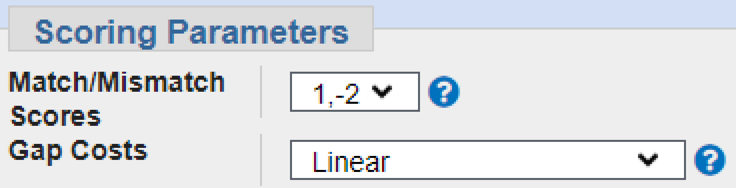
- The best alignment for TTCATA vs TGCTCGTA
  - T/T with the best alignment for TCATA vs GCTCGTA
  - T/– with the best alignment for TCATA vs TGCTCGTA
  - –/T with the best alignment for TTCATA vs GCTCGTA

- Different score for each possibility
  - Match
  - Mismatch
  - Gap

Scoring of sequence alignment

Nucleotide alignment scores

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

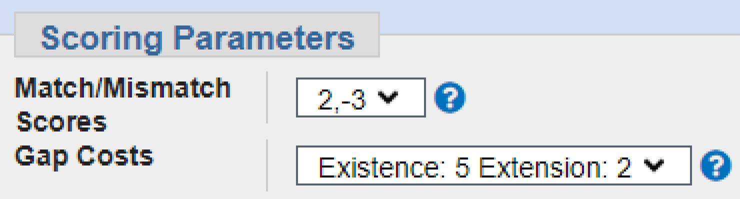


Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear

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



Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

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)

Interpretation of match/mismatch scores

- **Match / Mismatch = +1 / −2**
  - A mismatch followed by two matches = no score gain
  - Get hits with **high similarity**
  - Re-sequencing, closely related species

- **Match / Mismatch = +2 / −3**
  - A mismatch followed by two matches +1 score
  - Get hits with **intermediate similarity**
  - More distance species

Interpretation of gap scores

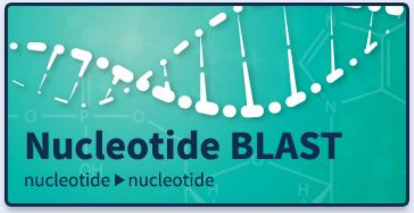
- Constant = An insertion/deletion can be of any length

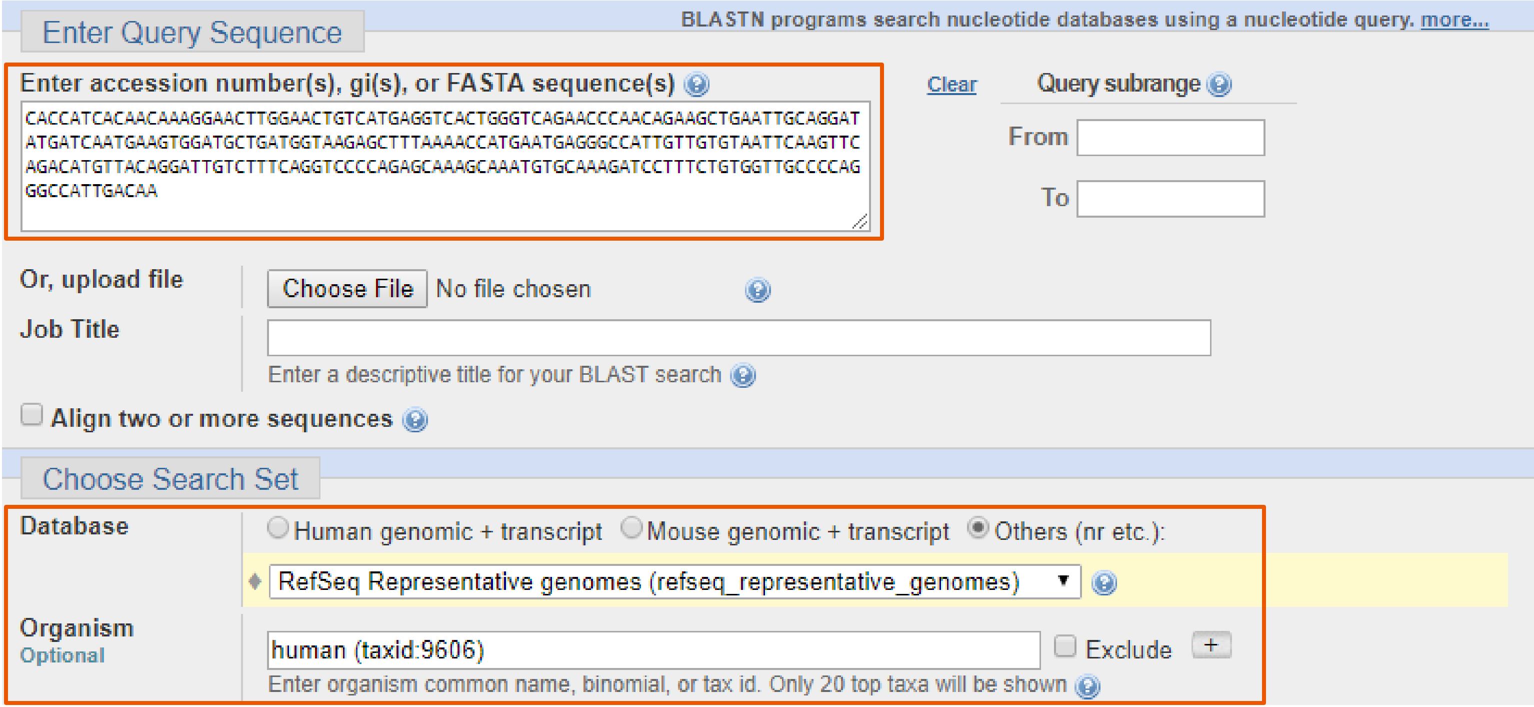
- Linear = Long indel is less likely than short indel

- Affine = Existence + (Extension x Length)
    - Combination of constant and linear

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.):

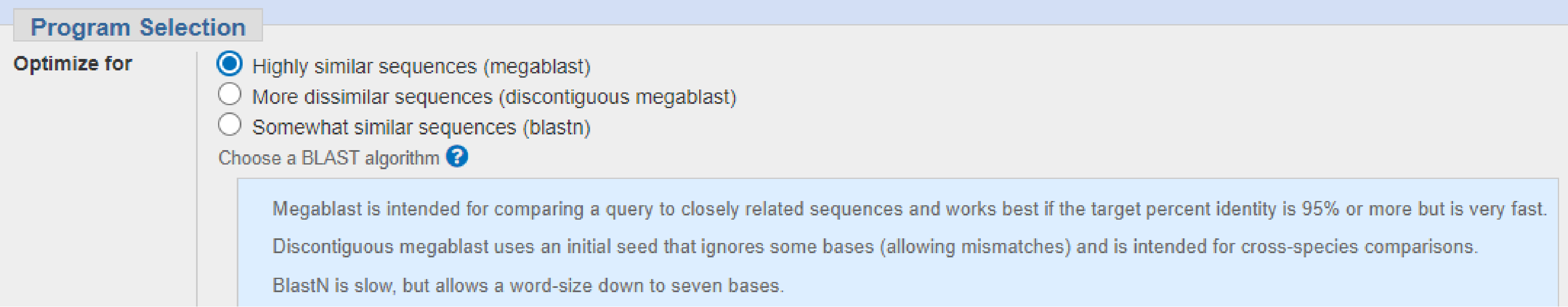
RefSeq Representative genomes (refseq_representative_genomes) ▾

Organism
Optional

human (taxid:9606) □ Exclude +

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

Nucleotide BLAST algorithm



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.

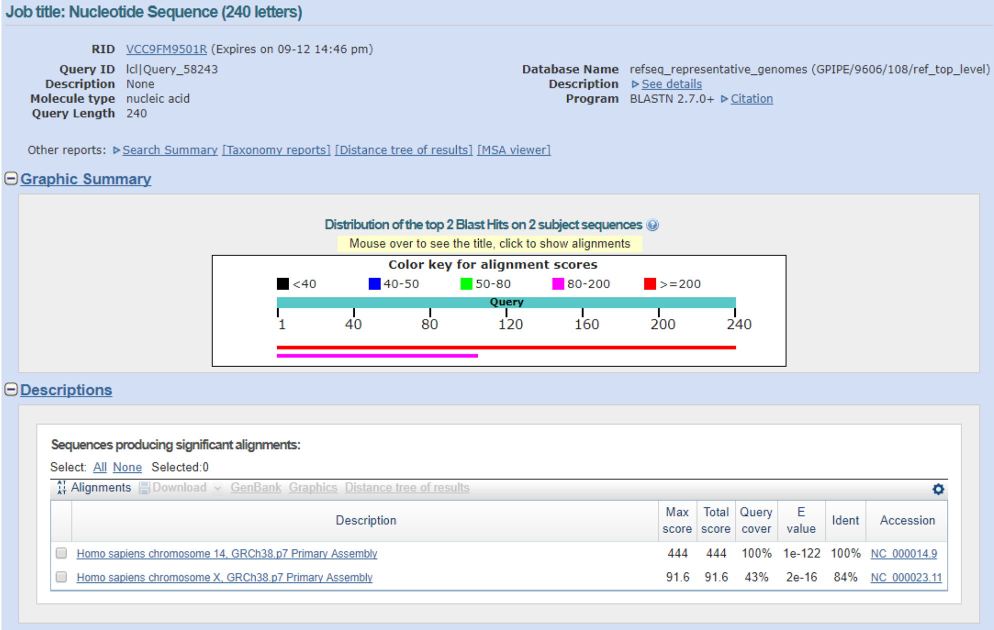
Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.

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

- MEGABLAST: word size = 28, match/mismatch = +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

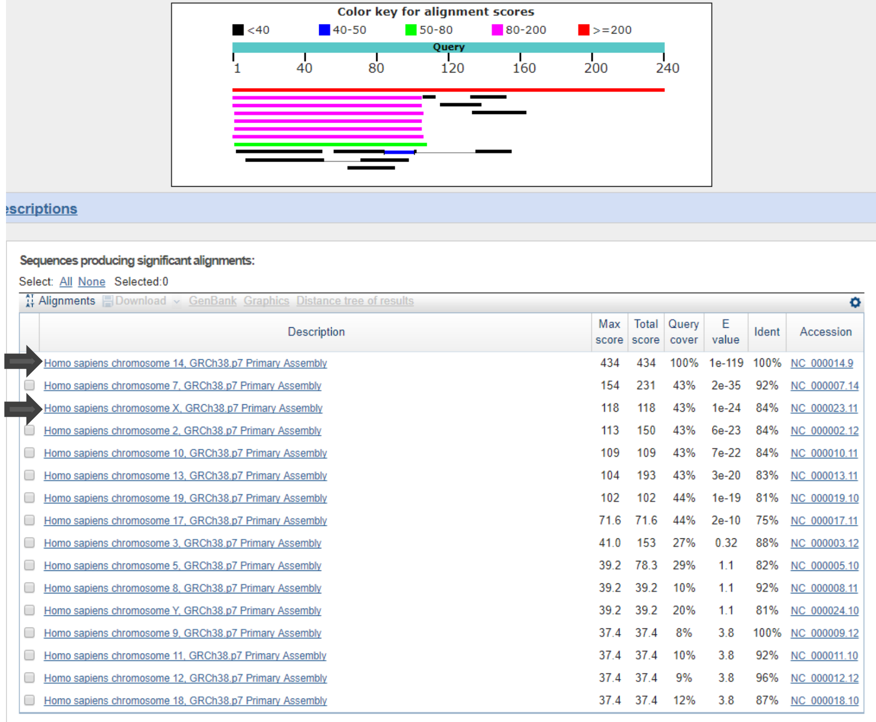
Sequences producing significant alignments:

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

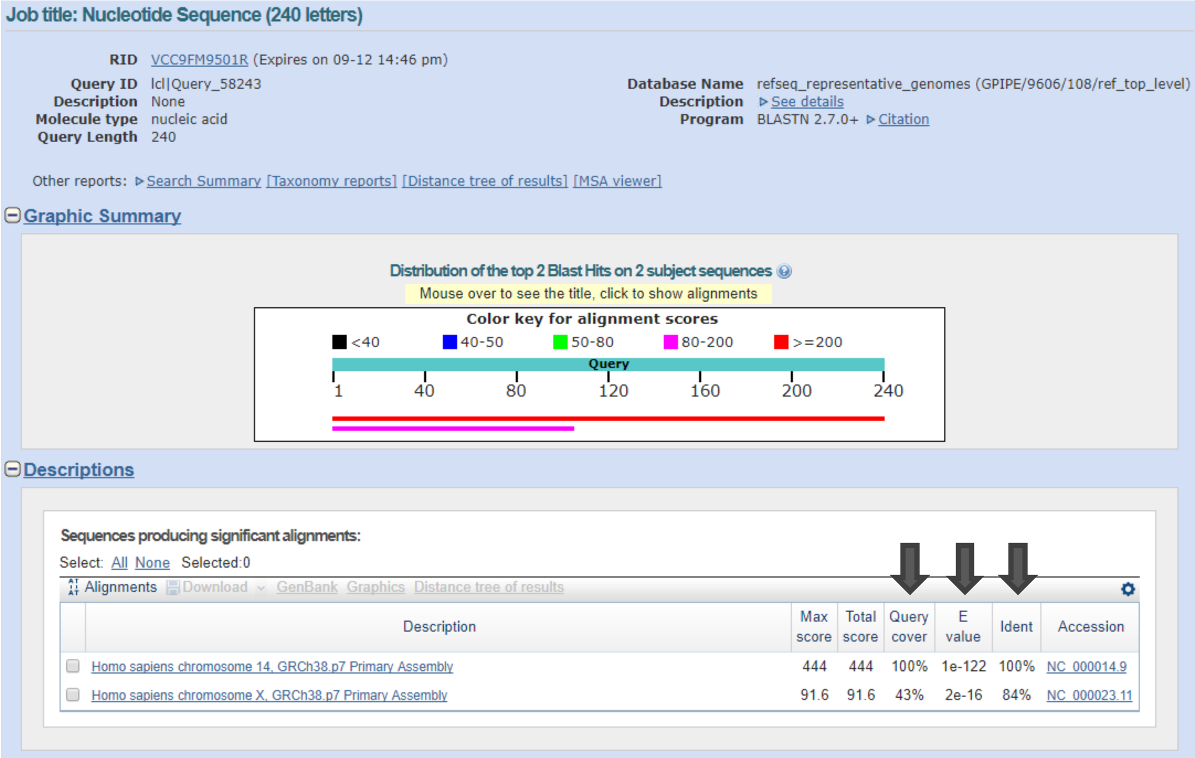
Sequences producing significant alignments:

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

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

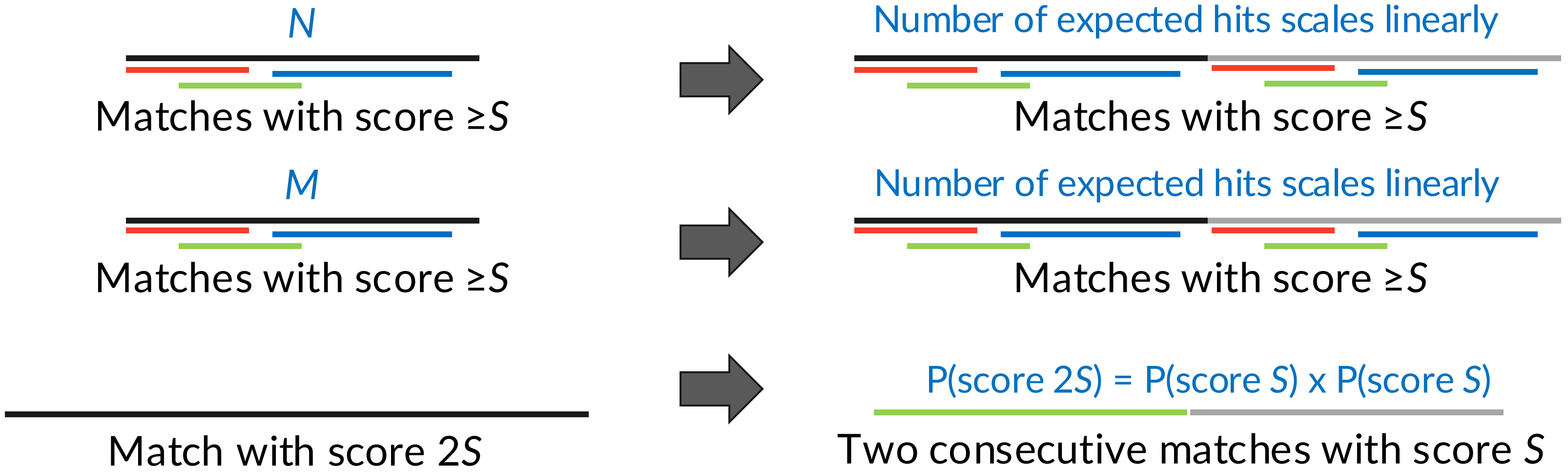
Identity = % of identity in the aligned region

E value = expected number of hits with the same or higher score by chance

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

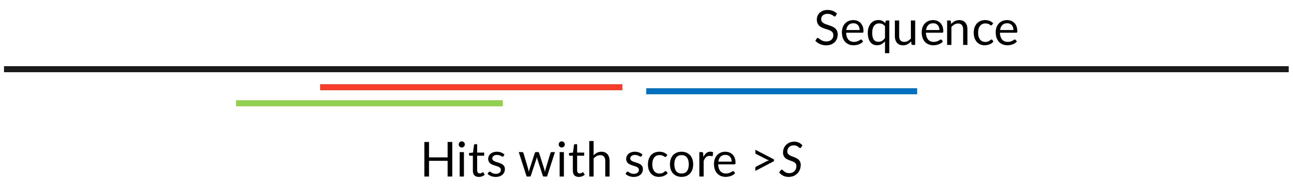
Matches with score $\geq S$

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

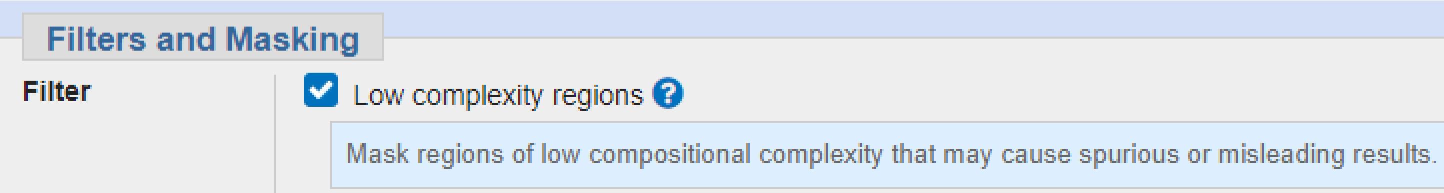


- Event of interest = hits with score >*S* on the sequence of length *N*
- Expected number of events = E value
- Probability of observing *k* hits with score >*S* = $\frac{E^k e^{-E}}{k!}$

Low complexity region

CG island

CCCGCGCGCCCCGGCGCCCGATGCAACTAGC



Filters and Masking

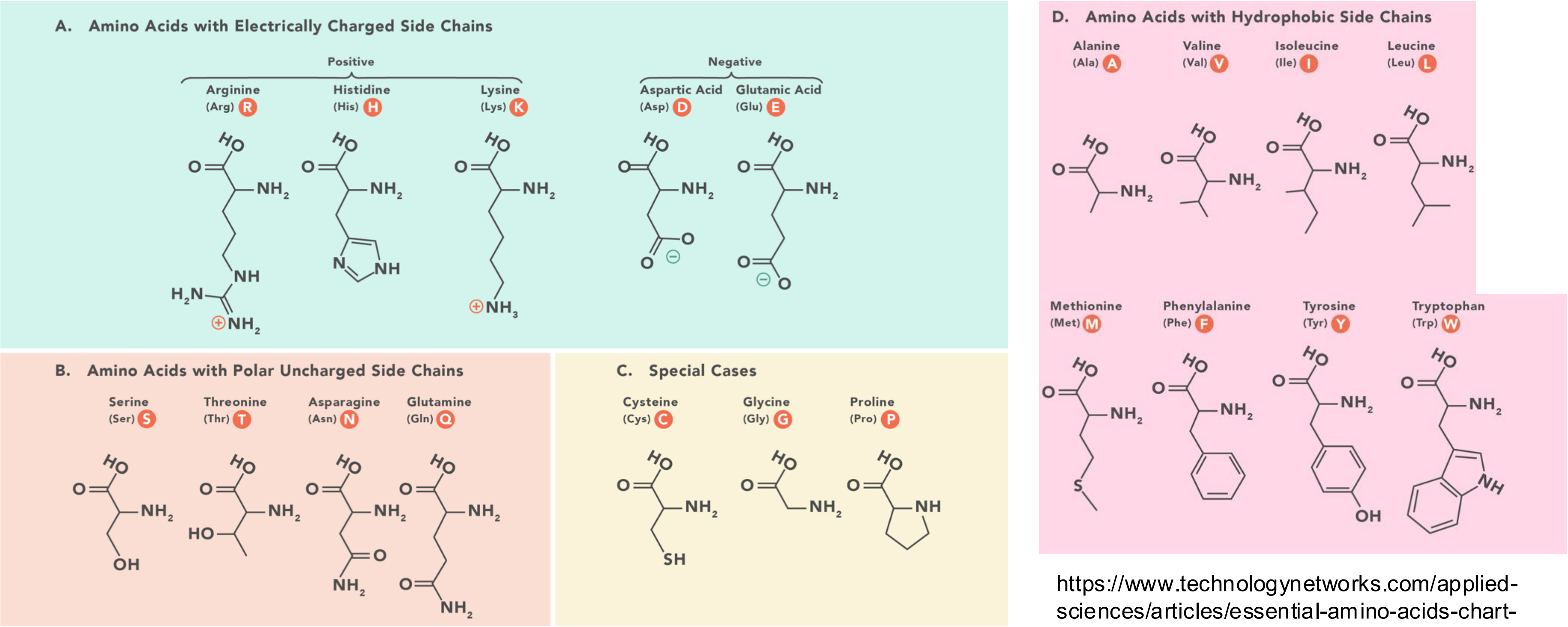
Filter ☑ Low complexity regions

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

- Probability of getting a hit with score >*S* depends on the nucleotide composition (easier with only C and G)
- BLAST withholds these regions from score calculation

Protein sequence alignment

Amino acid side chains



A. Amino Acids with Electrically Charged Side Chains

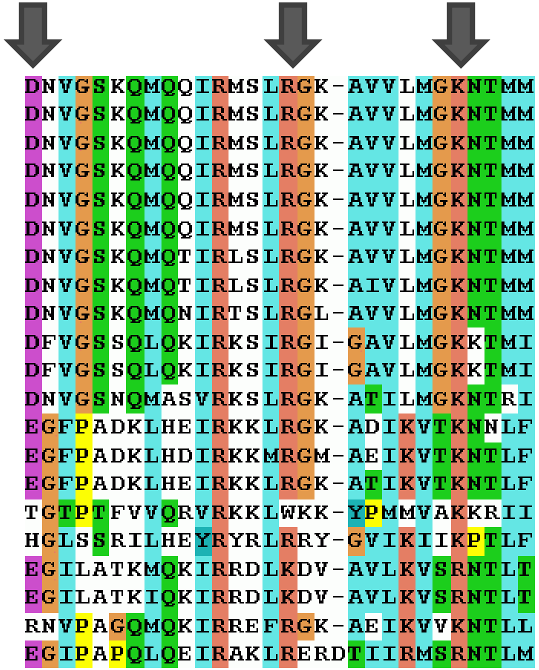
B. Amino Acids with Polar Uncharged Side Chains

C. Special Cases

D. Amino Acids with Hydrophobic Side Chains

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

Similar side chains are tolerated by evolution



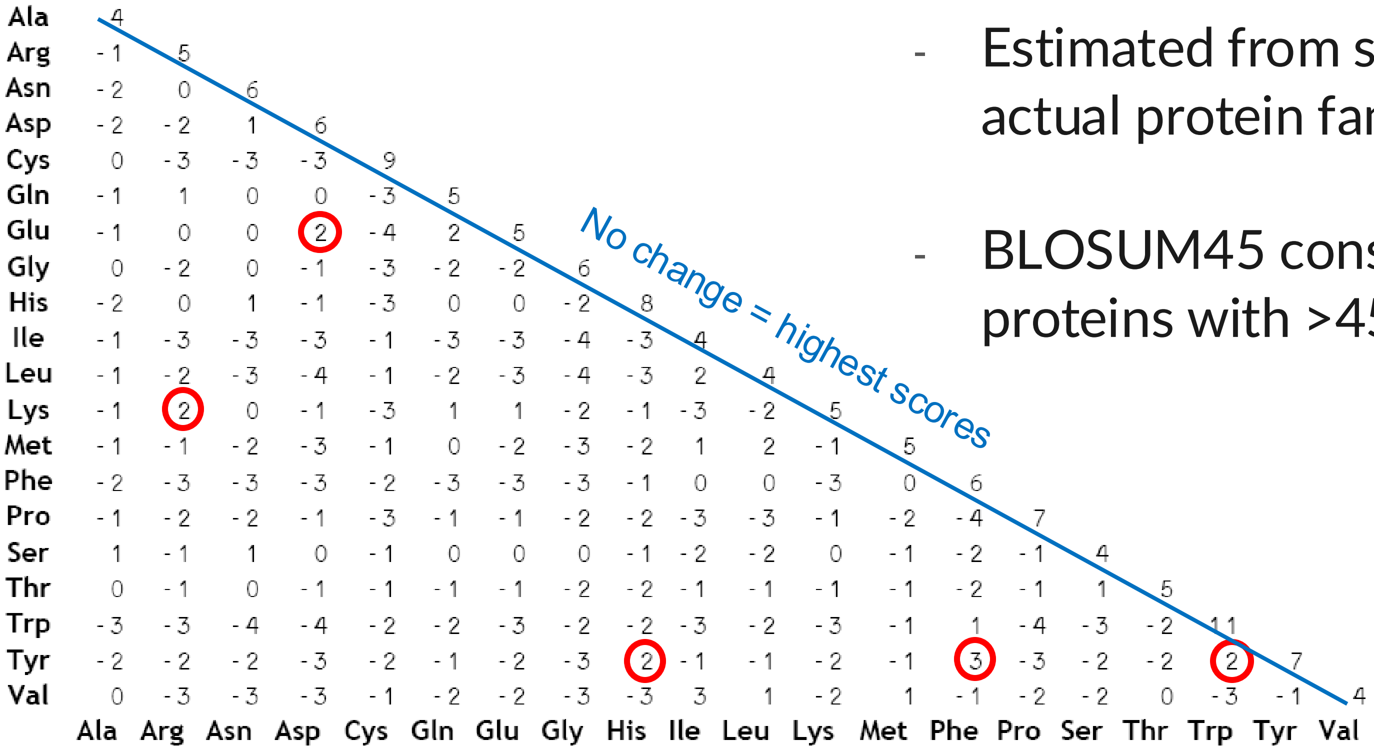
```
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQQIRMSLRGK-AVVLMGKNTMM
DNVGSKQMQTIRLSLRGK-AVVLMGKNTMM
DNVGSKQMQTIRLSLRGK-AIVLMGKNTMM
DNVGSKQMQNIRTSLRGL-AVVLMGKNTMM
DFVGSSQLQKIRKSIRGI-GAVLMGKKTMI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

wikipedia.com

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

- Alignment score must reflect these

Block Substitution Matrix (BLOSUM)



Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

No change = highest scores

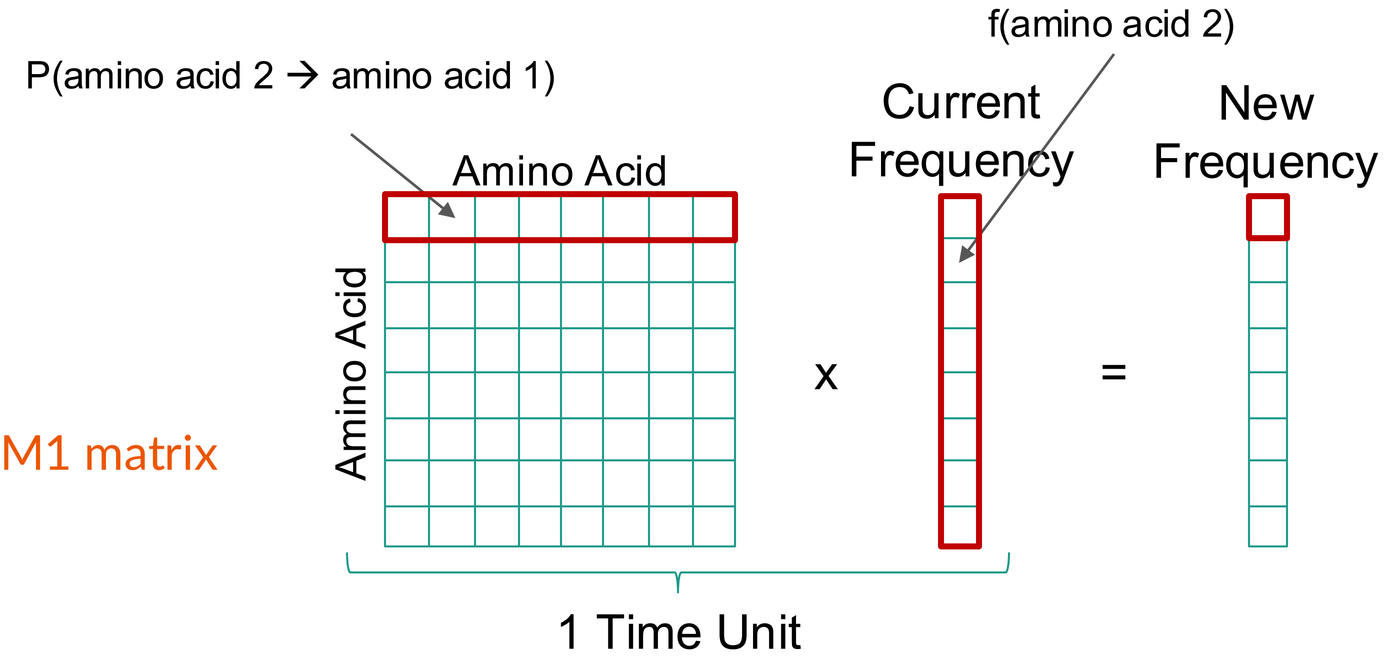
- Estimated from substitution rates of actual protein families

- BLOSUM45 constructed using proteins with >45% conservation

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

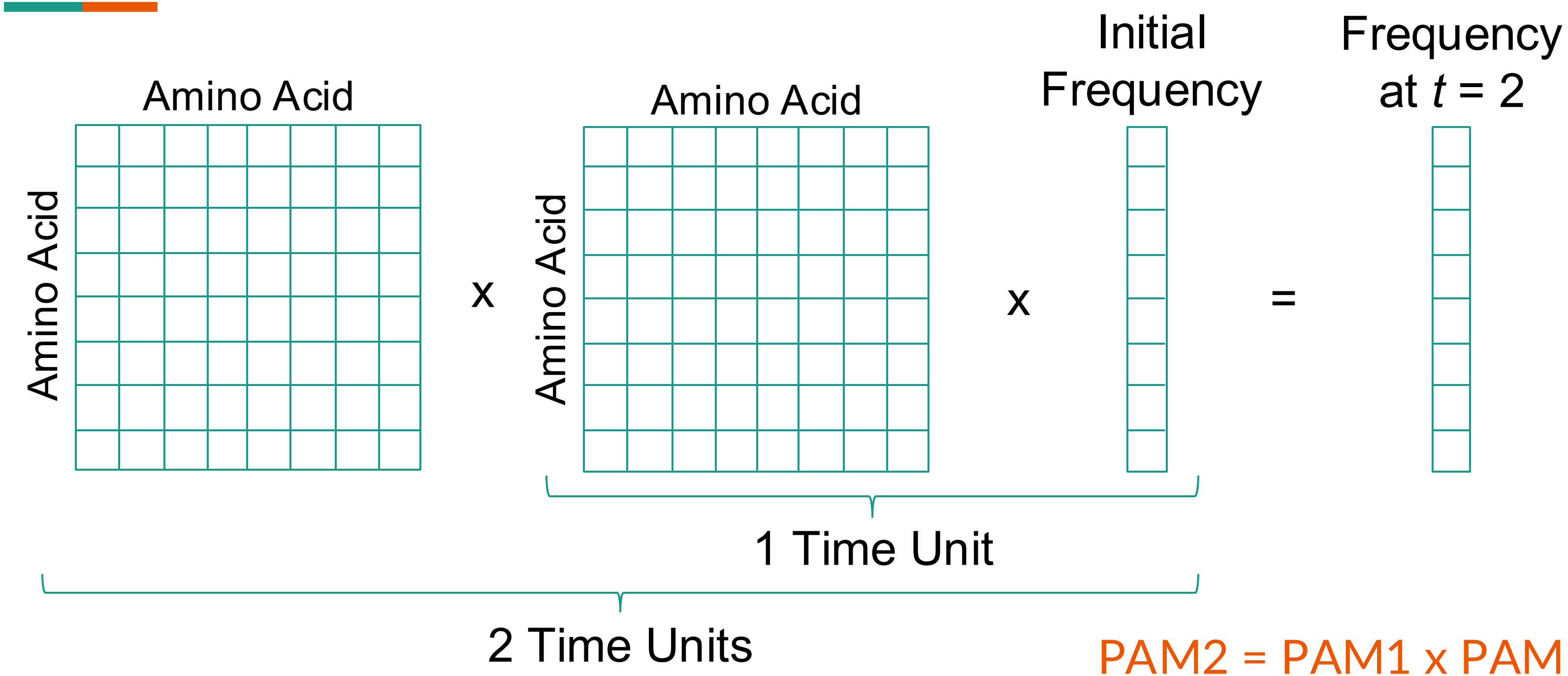
Point Accepted Mutation (PAM)

PAM1 matrix



- Mimic one step of evolution

Point Accepted Mutation (PAM)



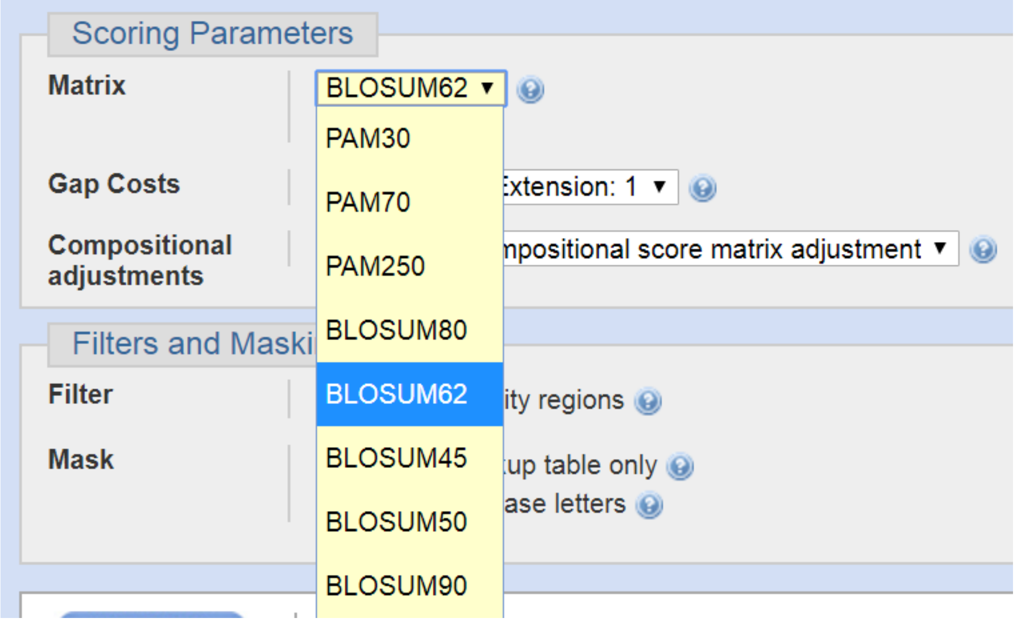
PAM2 = PAM1 x PAM1

- Extrapolate evolution via multiplication

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

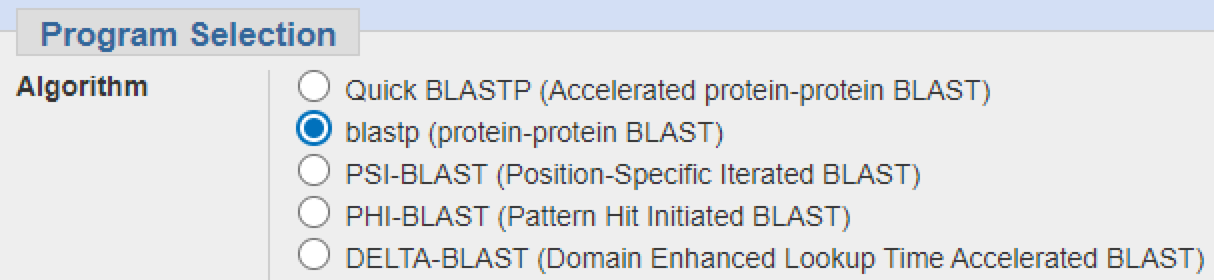
Filter: ity regions

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)

- BLASTP assumes that all amino acid residues are the same
- But each protein domain & motif evolve differently
  - Unknown pattern: PSI-BLAST
  - Known pattern: PHI-BLAST

Pattern hit initiated (PHI-BLAST)

x = any amino acid

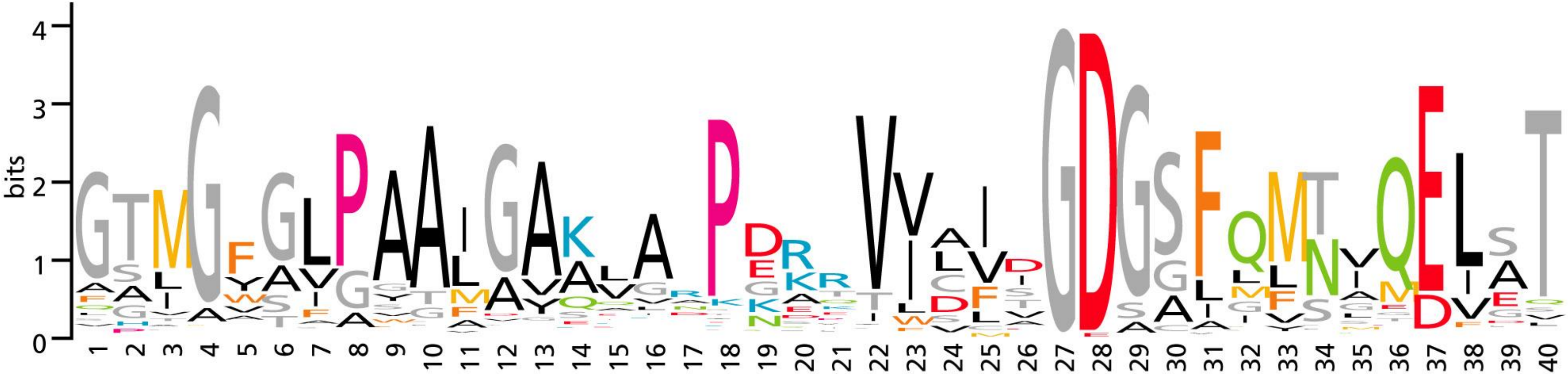
`[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]`

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 scoring matrix (PSSM)

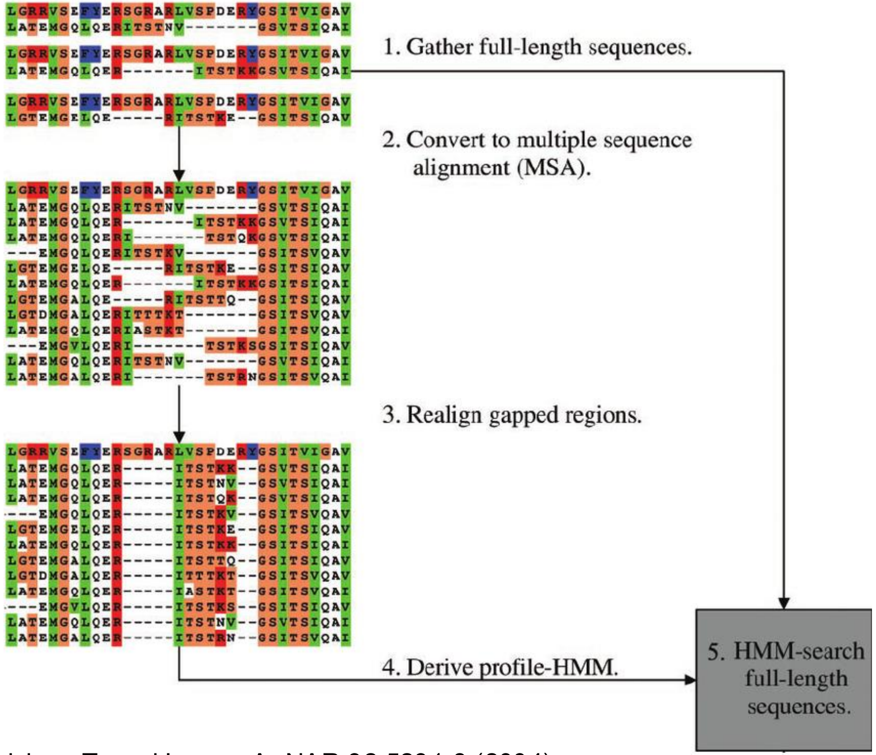


www.nemates.org/uky/520/Lecture/Lect6/BIO520_2010_Lect6.pp

weblogo.berkeley.edu

- Different scoring for each position in the motif/domain
- How do we know?

Position-specific iteratred (PSI-BLAST)



```
LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LATEMGQLQERITSTNV-------GSVTSIQAI

LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LATEMGQLQER-------ITSTKKGSVTSIQAI

LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LGTEMGELQE-----RITSTKE--GSITSIQAV
```

```
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
```

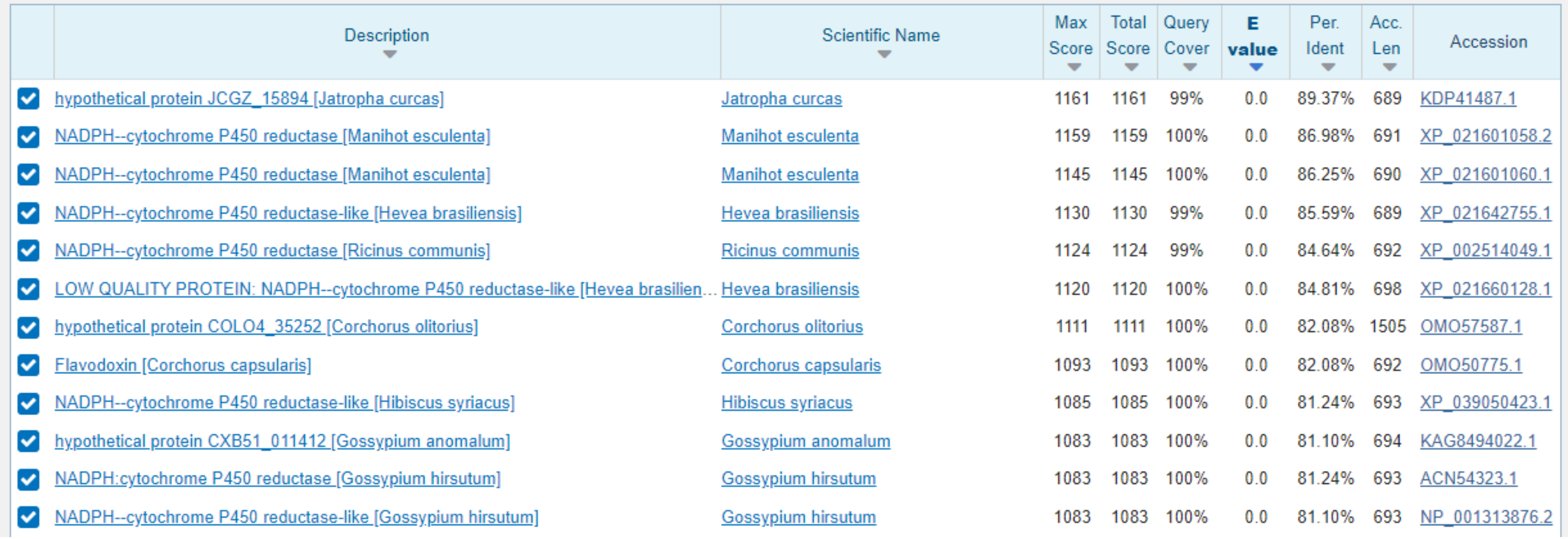
```
LGRRVSEFYERSGRARLVSPDERYGSITVIGAV
LATEMGQLQER-----ITSTKK--GSVTSIQAI
LATEMGQLQER-----ITSTNV--GSVTSIQAI
LATEMGQLQER-----ITSTQK--GSVTSIQAI
---EMGQLQER-----ITSTKV--GSITSVQAV
LGTEMGELQER-----ITSTKE--GSITSIQAV
LATEMGQLQER-----ITSTKK--GSITSIQAI
LGTEMGALQER-----ITSTTQ--GSITSIQAV
LGTDMGALQER-----ITTTKT--GSITSVQAV
LATEMGQLQER-----IASTKT--GSITSVQAI
---EMGVLQER-----ITSTKS--GSITSIQAV
LATEMGQLQER-----ITSTNV--GSVTSIQAI
LATEMGALQER-----ITSTRN--GSITSVQAI
```

1. Gather full-length sequences.
2. Convert to multiple sequence alignment (MSA).
3. Realign gapped regions.
4. Derive profile-HMM.
5. HMM-search full-length sequences.

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

- Start from input sequences
- First BLASTP search
- Construct PSSM from hits
- Re-search using the PSSM
- Repeat

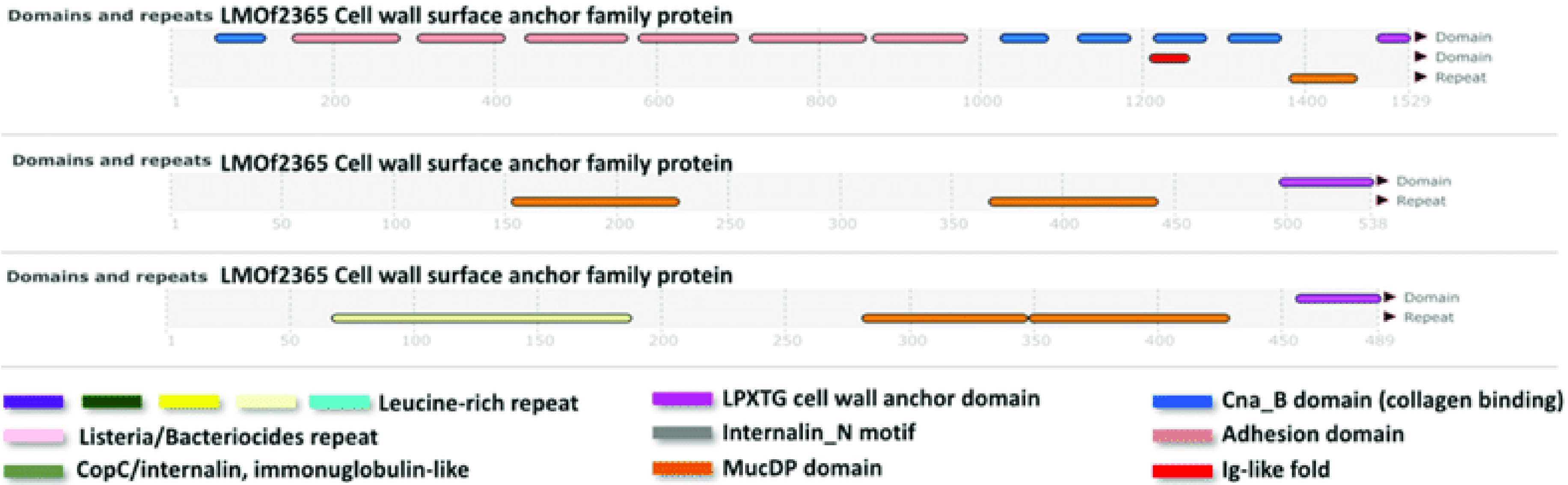
Using BLASTP to annotate protein function



	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
  - >80% similarity to known and predicted CYP reductase class I

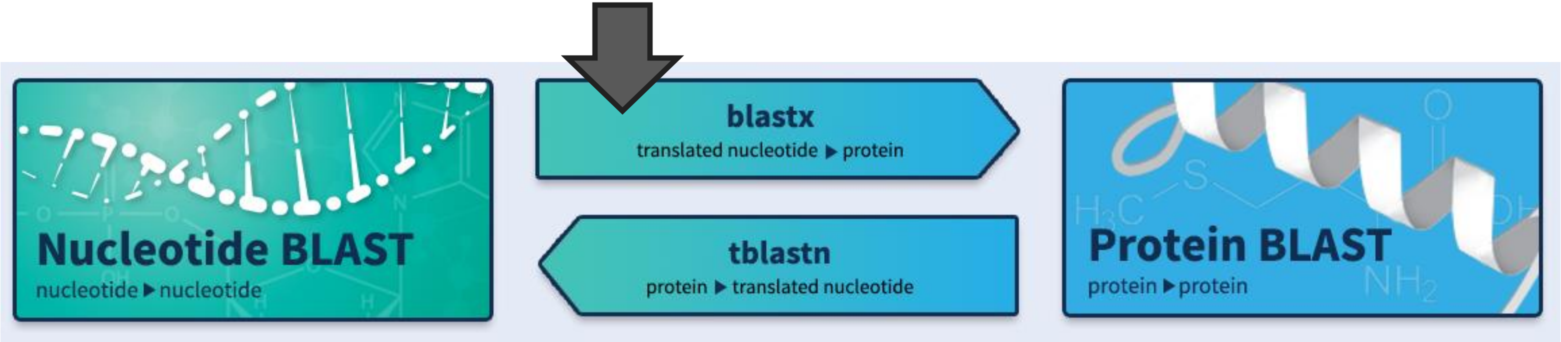
InterPro: Protein domain search


Domains and repeats
LMOf2365 Cell wall surface anchor family protein
Domain
Domain
Repeat
1
200
400
600
800
1000
1200
1400
1529
Domains and repeats
LMOf2365 Cell wall surface anchor family protein
Domain
Repeat
1
50
100
150
200
250
300
350
400
450
500
538
Domains and repeats
LMOf2365 Cell wall surface anchor family protein
Domain
Repeat
1
50
100
150
200
250
300
350
400
450
489
Leucine-rich repeat
LPXTG cell wall anchor domain
Cna_B domain (collagen binding)
Listeria/Bacteriocides repeat
Internalin_N motif
Adhesion domain
CopC/internalin, immonuglobulin-like
MucDP domain
Ig-like fold

- Use patterns from known protein functional domains

Mixed protein-nucleotide alignment

BLASTX and TBLASTN



- For coding DNA, alignment at protein level is more informative
  - Codon structure, not all amino acid change is equally likely
- **BLASTX** = align translated DNA to protein database
- **tBLASTN** = align protein to translated DNA database

Example use cases

- **BLASTX**
  - Get RNA sequence from RNA-seq
  - Unsure which **open reading frame** is translated
  - Does this RNA translate to a known protein?

- **tBLASTN**
  - Identified novel protein, with no similarity to protein database
  - Is there a transcriptomics study that have identified the RNA of a related protein?

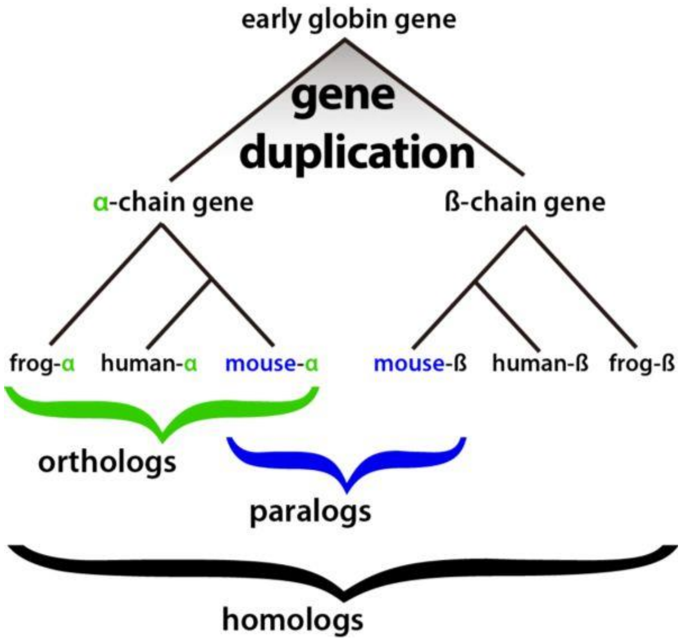
More advanced alignments

All-vs-all search for finding orthologs across genomes

- Evolutionarily related genes
  - Group 1 = {Mouse-a, Human-a}
  - Group 2 = {Mouse-b, Human-b}

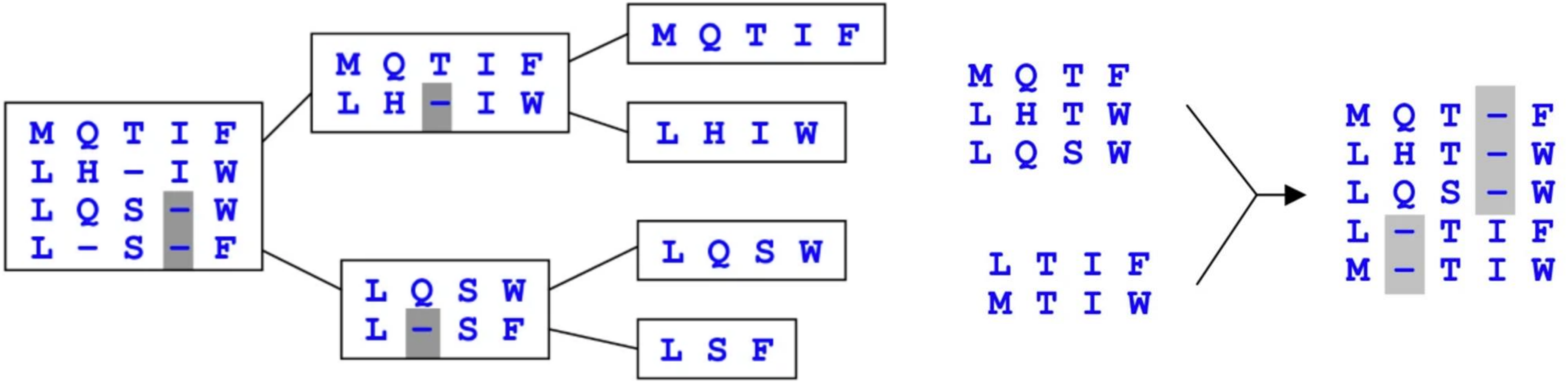
- BLAST mouse to human
- BLAST human to mouse

- Reciprocal best hit:
  - Human-a is the best hit for Mouse-a
  - Mouse-a is the best hit for Human-a



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

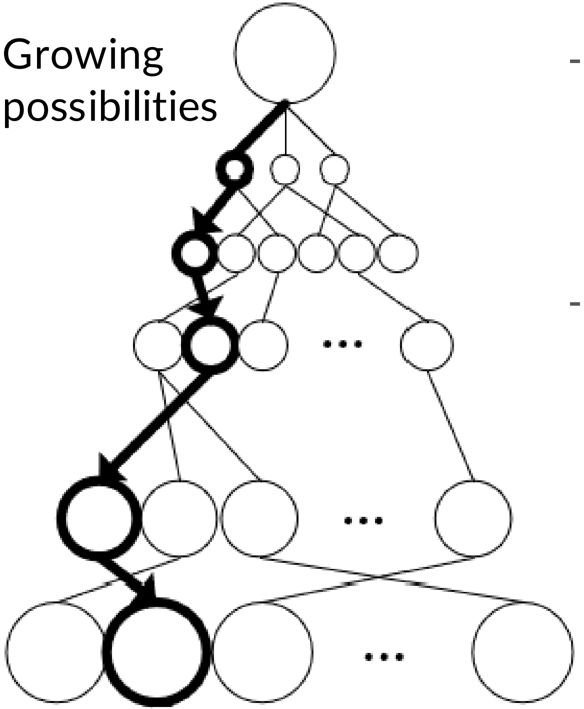
Multiple sequence alignment (MSA)



Edgar, BMC Bioinformatics, 5, 113 (2014)

- Dynamic programming is not feasible. Too many ways of grouping sequences (on top of aligning positions).

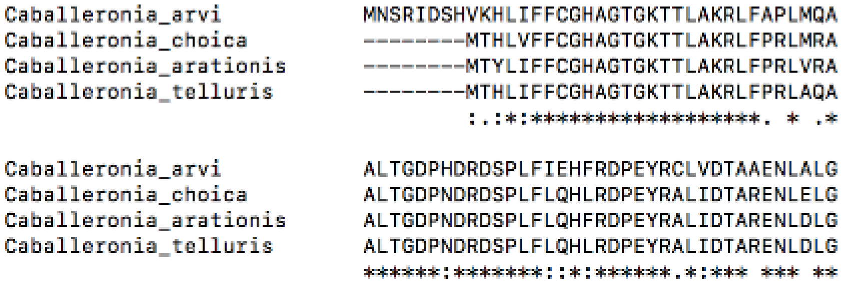
Heuristic algorithms



- Greedy algorithm optimize the next step and/or the future steps

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

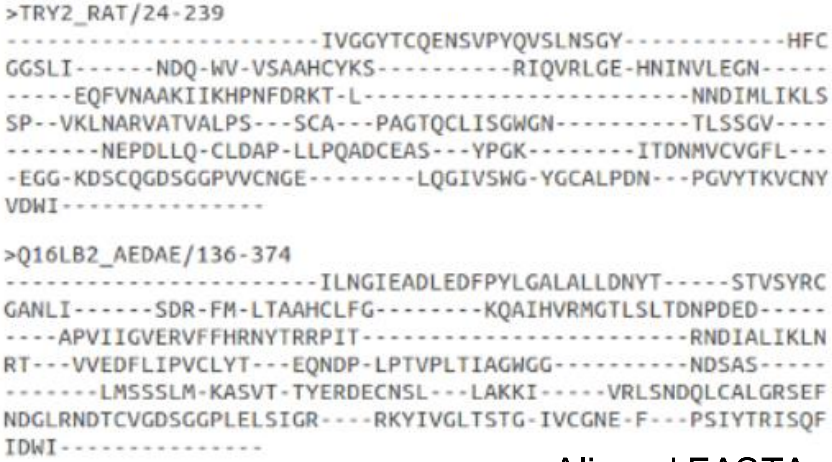
Alignment output



```
Caballeronia_arvi           MNSRIDSHVKHLIFFCGHAGTGKTTLAKRLFAPLMQA
Caballeronia_choica         --------MTHLVFFCGHAGTGKTTLAKRLFPRLMRA
Caballeronia_arationis      --------MTYLIFFCGHAGTGKTTLAKRLFPRLVRA
Caballeronia_telluris       --------MTHLIFFCGHAGTGKTTLAKRLFPRLAQA
                                    :.:*:******************. * .*

Caballeronia_arvi           ALTGDPHDRDSPLFIEHFRDPEYRCLVDTAAENLALG
Caballeronia_choica         ALTGDPNDRDSPLFLQHLRDPEYRALIDTARENLELG
Caballeronia_arationis      ALTGDPNDRDSPLFLQHFRDPEYRALIDTARENLDLG
Caballeronia_telluris       ALTGDPNDRDSPLFLQHLRDPEYRALIDTARENLDLG
                            ******:*******::*:******.*:*** *** **
```

ClustalW



```
>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---------------
```

Aligned FASTA

- Dashes “-” are added to indicate insertions/deletions
- All output sequences have the same length

Any question?

- See you next time