Sequencing similarity search and functional prediction

Vitor Pavinato
correapavinato.1@osu.edu

Plan for today

- Sequence alignment
  - Pieces we need to put together to understand how it is done (how it works);
- Office hours - introduction standalone Blast (in the afternoon and tomorrow during the lab);

- Answer some questions raised during the lab (14/Jan/2022);
  - **Biological significance of using different values for a nucleotide scoring matrix;**
  - **Sequence homology (*Caenorhabditis elegans* superoxide dismutase);**
  - **Is there a magic cutoff to decide if an alignment is significant?**

Sequence alignment

Why do we align sequences?

Sequence alignment

Why do we align sequences?

Genome assembly

Find **homologous**

Find functional domains

Find Target for gene silencing

Motivation

You are conducting a differential gene expression analysis and, for example, you found an isoform (more about this next class) of a mRNA with unknown function associated with one of your treatments.

Then you look for the drosophila db to find the function of this gene

What is the rationality behind it?

“Nature is a tinkerer and not an inventor” (Jacob 1977)

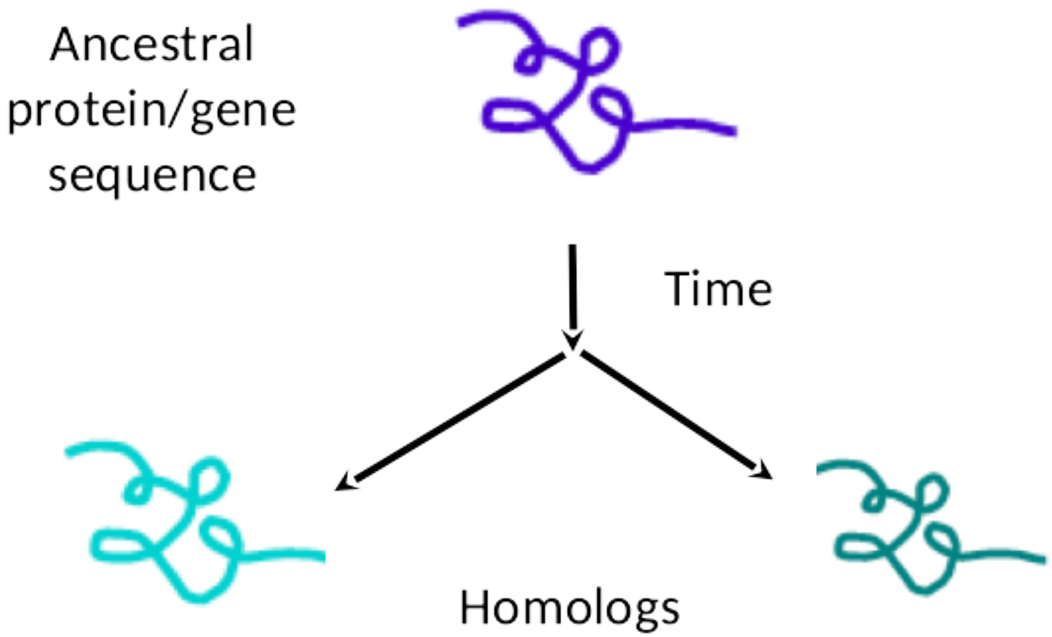
“New sequences are adapted from pre-existing sequences rather than invented *de novo.* This is very fortunate for computational sequence analysis. We can often recognize a **significant sequence similarity** between a new sequence and a sequence about which something is already known;”

“When we do this, we transfer the information about structure and/or function to the new sequence. We say that the two related sequences are **homologous** and we transfer information by **homology**”

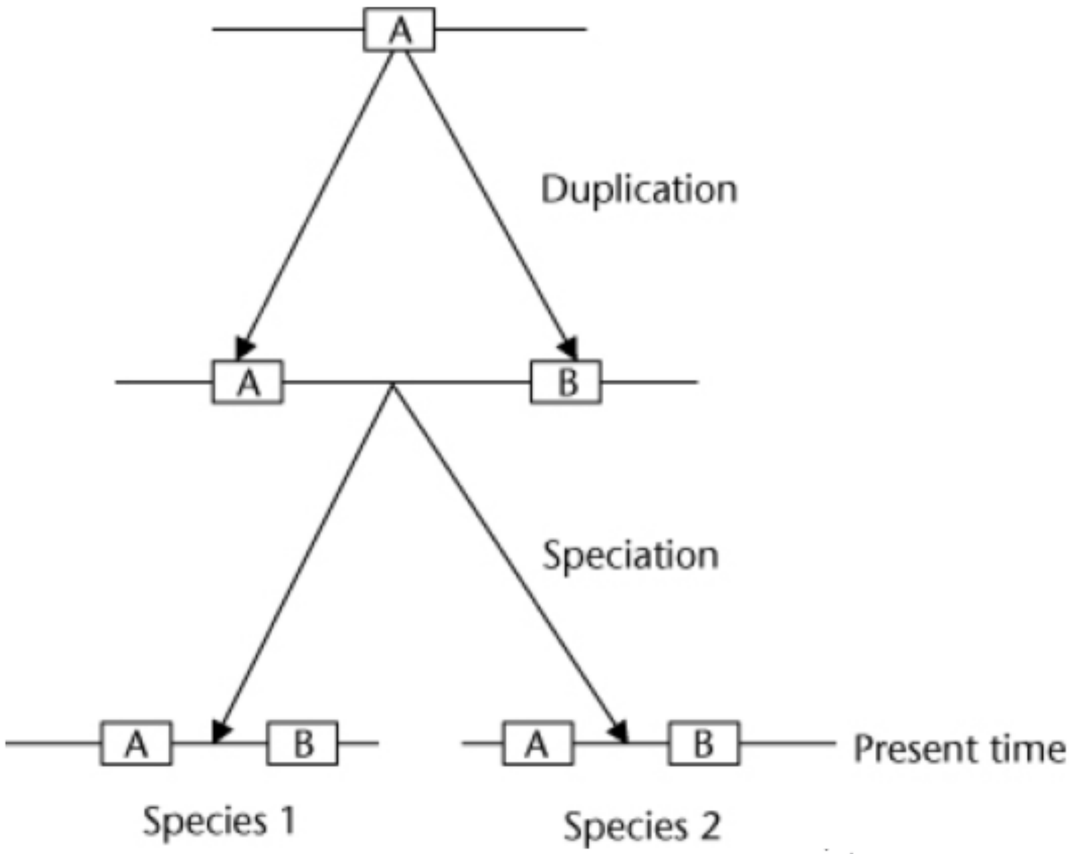
Durbin et al.1998

What is homology?

"Two regions of DNA that evolved from the same sequence (through processes of duplication of genomic regions and separation of two species) are homologous, or homologs of one another.



The Biostar Handbook: 2nd Edition.



Paralogous sequences or paralogs were separated by duplication of a genomic region within the same genome.”

Regions in the genomes of two species that are descended from the same area in a common ancestor’s genome are orthologs. These regions are said to be orthologous.

The concept of ***sequence alignment is crucial***

Before the similarity between two sequence can be evaluated, we begin by founding a plausible alignment between them

Almost all alignment methods will eventually find the best alignment between two sequences under a score scheme

It then a matter of deciding whether the alignment had occur because the sequences were related or just by chance

Pieces we need

Score matrix

Type of alignment

An algorithm

Significance

Scoring matrix

$$S_{ii} = 1$$
$$S_{ij} = -1$$

Simplest nucleotide matrix

	A	T	C	G
A	1	-1	-1	-1
T	-1	1	-1	-1
C	-1	-1	1	-1
G	-1	-1	-1	1

Transition/Transversion ratio

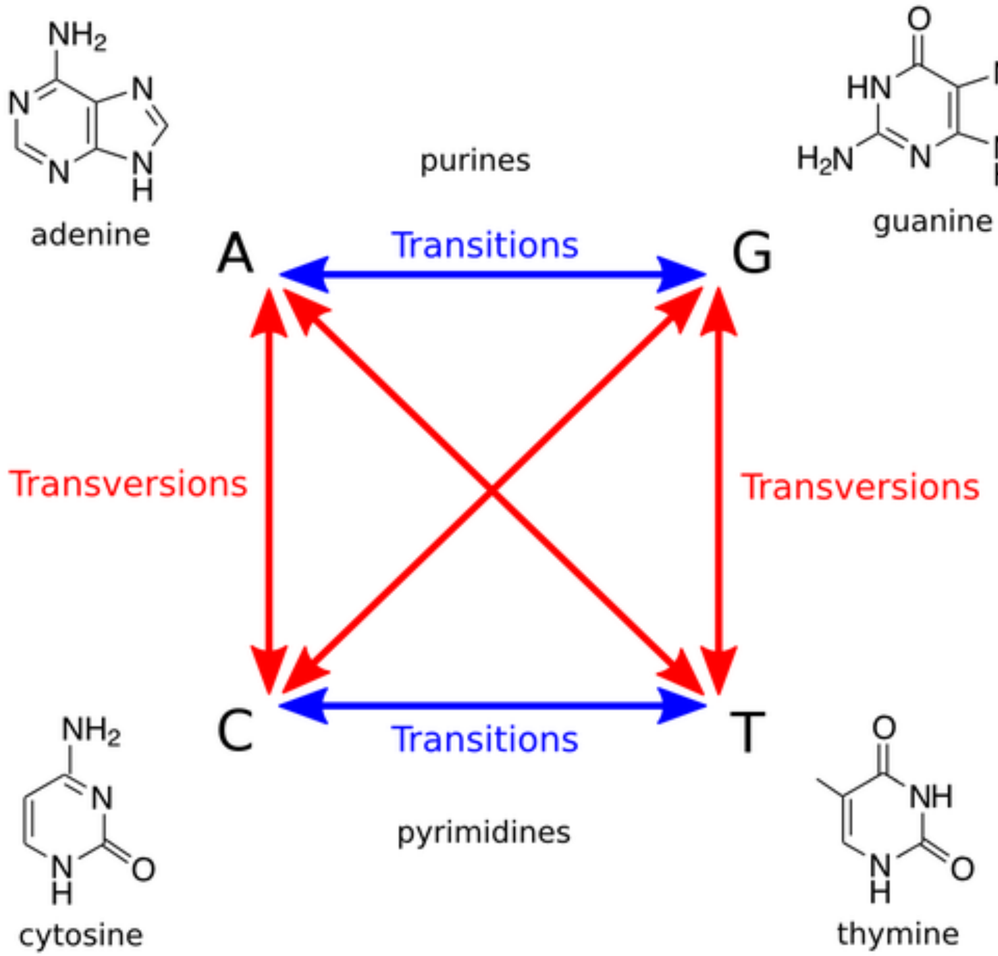
Purines	Pyrimidines
A <-> G	C <-> T

Purines and Pyrimidines are structurally more similar to each other.

Non-coding sequences Ts/Tv = 2 to 1 (3 to 1)

C-> T and G-> A occurs more frequently.

the G-T mutation is the single most common mutation in human DNA. It occurs about once in every 10,000 to 100,000 base pairs



https://rosalind.info/glossary/transitiontransversion-ratio/

Scoring system should favor matching identi or related amino acids and penalize for poor matches and for gaps

BLOSUM 62 scoring matrix

(positive values are shaded)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**4**																			
R	-1	**5**																		
N	-2	0	**6**																	
D	-2	-2	**1**	**6**																
C	0	-3	-3	-3	**9**															
Q	-1	**1**	0	0	-3	**5**														
E	-1	0	0	**2**	-4	**2**	**5**													
G	0	-2	0	-1	-3	-2	-2	**6**												
H	-2	0	**1**	-1	-3	0	0	-2	**8**											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	**2**	**4**									
K	-1	**2**	0	-1	-3	**1**	**1**	-2	-1	-3	-2	**5**								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	**1**	**2**	-1	**5**							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**					
S	**1**	-1	**1**	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	**1**	**5**			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	**1**	-4	-3	-2	**11**		
Y	-2	-2	-2	-3	-2	-1	-2	-3	**2**	-1	-1	-2	-1	**3**	-3	-2	-2	**2**	**7**	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	**3**	**1**	-2	**1**	-1	-2	-2	0	-3	-1	**4**

The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

Type of alignments

- Global alignment
- Local alignment
- Semi-global alignment (not covered in this class)
- Multiple alignment (outside the scope of this course)
- Short-read alignment (next week on RNAseq class)

Global alignment

When performing global alignments, the bases of both sequences are arranged next to one another over their entire length and all . Each base of the first sequence is matched to another base or a “gap” of the second sequence. We use global alignments when we are looking for an arrangement that maximizes the similarities over the entire length of both sequences:

Excerpt From: “The Biostar Handbook: 2nd Edition.”

Global alignment

When performing global alignments, the bases of both sequences are arranged next to one another over their entire length and all . Each base of the first sequence is matched to another base or a “gap” of the second sequence. We use global alignments when we are looking for an arrangement that maximizes the similarities over the entire length of both sequences:

```
THISLINE-
.......|-
ISALIGNED
```

```
THIS-LI-NE-
--||-||-||-
--ISALIGNED
```

Excerpt From: “The Biostar Handbook: 2nd Edition.”

Local alignment

Local alignments are used when we need to find the region of maximal similarity between two sequences. When performing local alignments, the al- gorithms look for the highest scoring (partial) interval between the two sequences :

```
NE
||
NE
```

```
LINE
..||
IGNE
```

Basic Local Alignment Search Tool (BLAST)

Excerpt From: “The Biostar Handbook: 2nd Edition.”

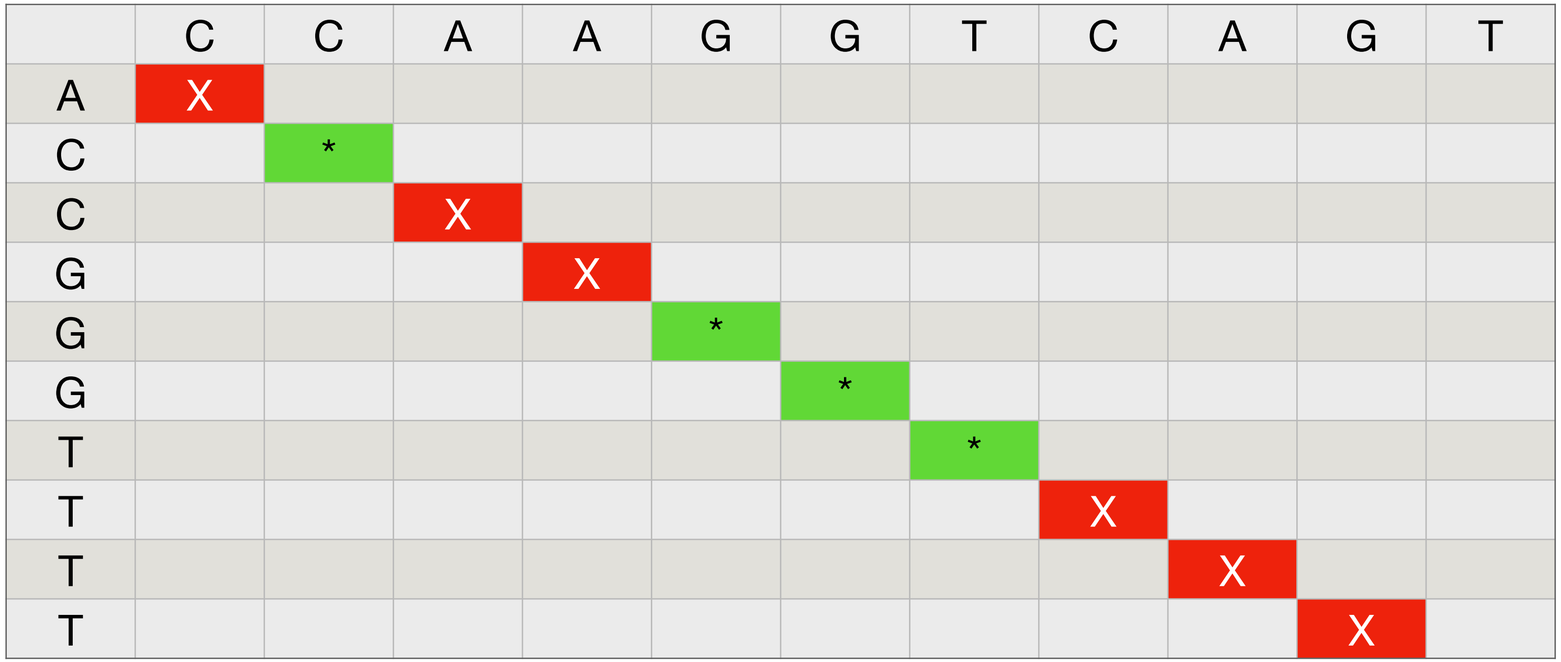
Scope

- Ungapped
- Gapped
  - Linear
  - Affine

Simplest algorithm: ungapped local alignment

$$S_{ii} = 1$$

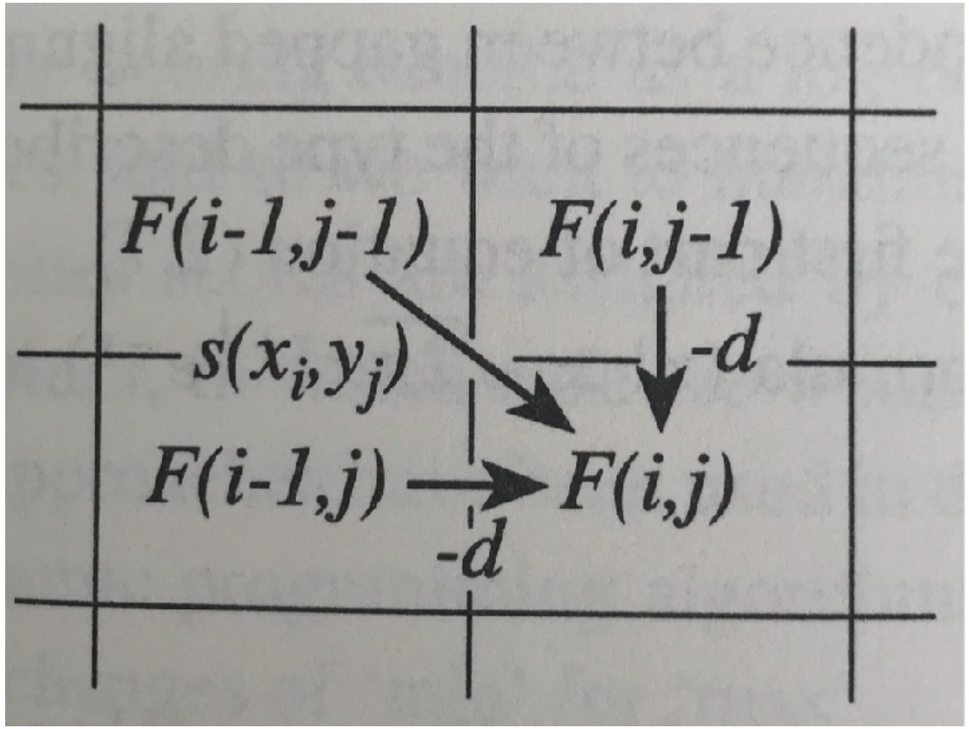
$$S_{ij} = -1$$



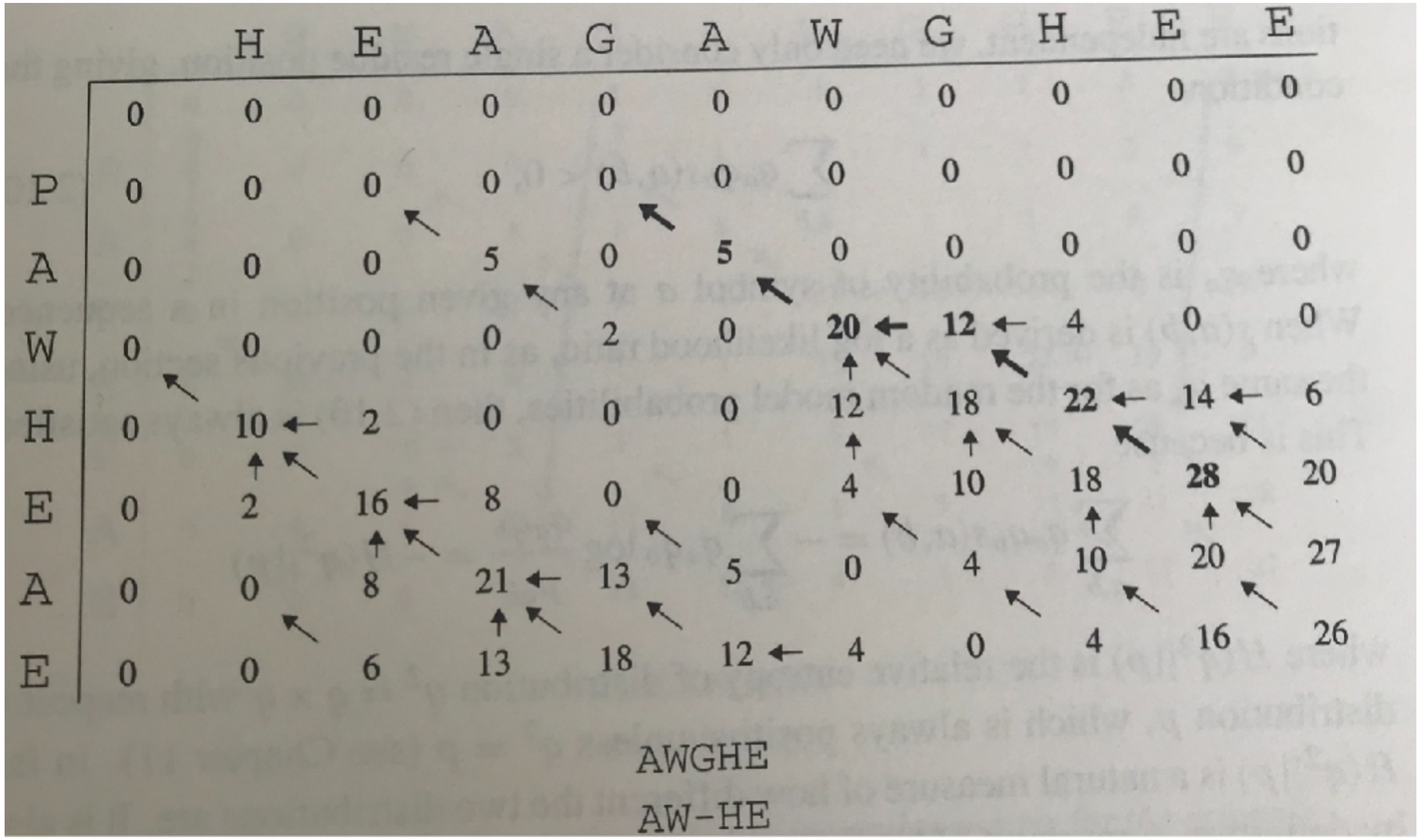
How do you “walk" through the rows and columns?

Dynamic Programming: Recursion

$$F(i,j) = max \begin{cases} 0, \\ F(i-1,j-1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d\,. \end{cases}$$



Dynamic Programming: Recursion



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	**5**	0	0	0	0	0
W	0	0	0	0	2	0	**20** ←	**12** ←	4	0	0
H	0	10 ←	2	0	0	0	12	18	**22** ←	14 ←	6
E	0	2	16 ←	8	0	0	4	10	18	**28**	20
A	0	0	8	21 ←	13	5	0	4	10	20	27
E	0	0	6	13	18	12 ←	4	0	4	16	26

AWGHE

AW-HE

Keep track of scores AND how we got them → “traceback m

Determine significant

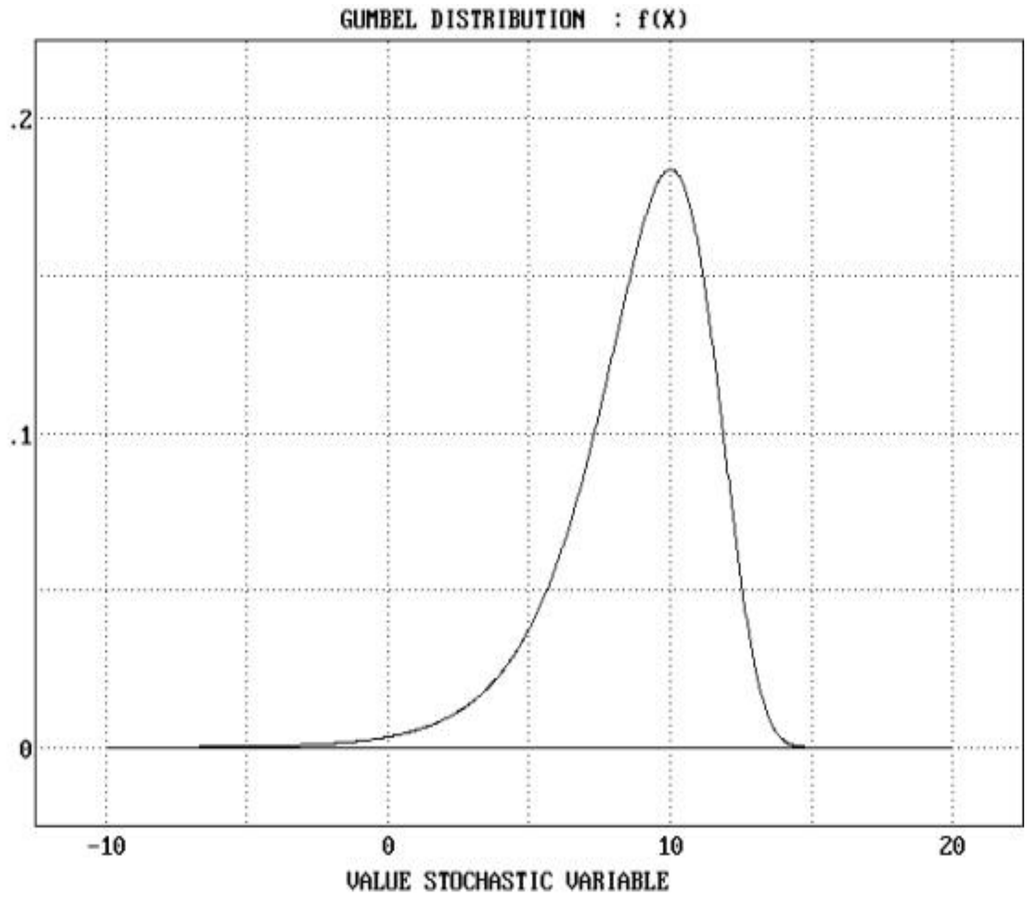
Identify the high scoring segments whose score **S** exceeds a cutoff **x** using a local alignment algorithm

$$P(S > x) = 1 - exp[-KMNe^{-\lambda x}]$$

M and **N** are the length of the sequence and the database

K and λ are positive parameters that depend on the scoring matrix and the composition of sequences being compared

Extreme Value (Gumbel) Distribution



In order to find the segment with highest score, the expected value should be negative (reason we have a score matrix with negative values), but some positive are allowed.

How the subject size N affects P(S>x)?

K = 10
M = 100
N = 1,000
lambda = 2

$$P(S > 10) = 0.002059031$$

K = 10
M = 100
N = 1,000,000
lambda = 2

$$P(S > 10) = 0.872693$$

BLAST ® » blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleoti

Enter Query Sequence

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

NM_001026785.1

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: Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Nucleotide collection (nr/nt)

Database size

Organism Optional: Enter organism name or id--completions will be suggested exclude Add organism

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:

YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for:

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

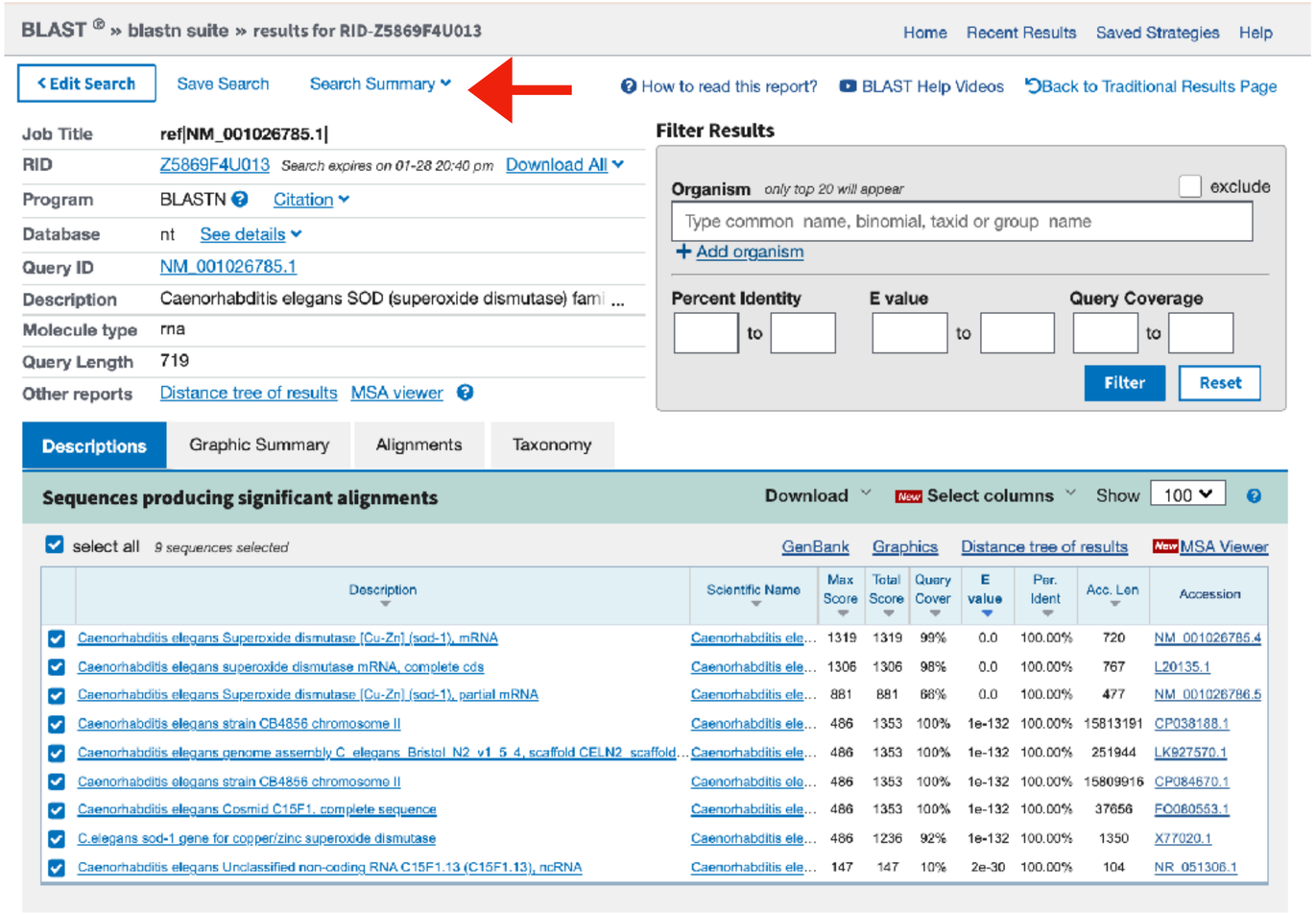
Choose a BLAST algorithm

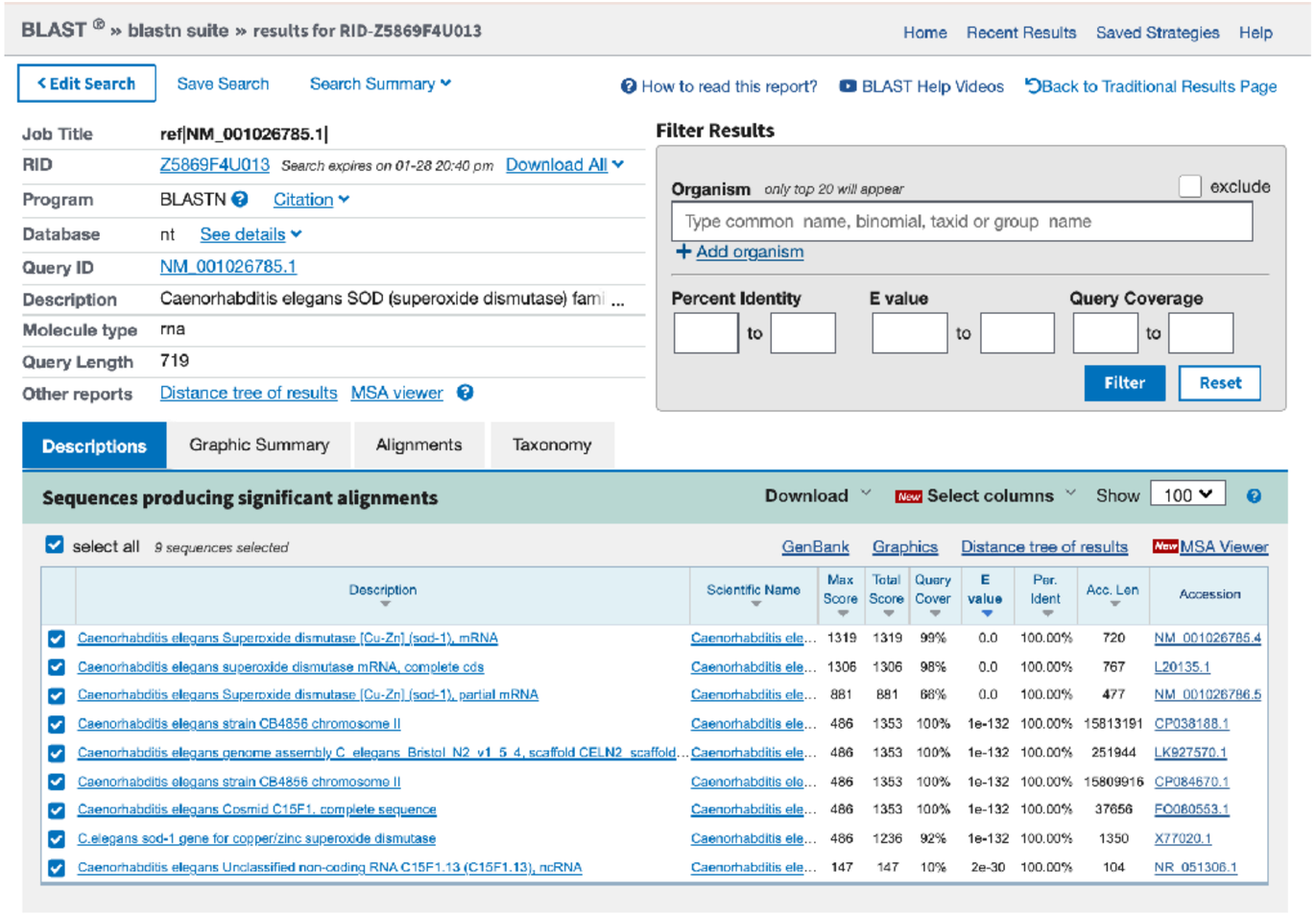
BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

+ Algorithm parameters


BLAST ® » blastn suite » results for RID-Z5869F4U013
Home
Recent Results
Saved Strategies
Help
< Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title
ref|NM_001026785.1|
RID
Z5869F4U013
Search expires on 01-28 20:40 pm
Download All
Program
BLASTN
Citation
Database
nt
See details
Query ID
NM_001026785.1
Description
Caenorhabditis elegans SOD (superoxide dismutase) fami ...
Molecule type
rna
Query Length
719
Other reports
Distance tree of results
MSA viewer
Filter Results
Organism
only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
New Select columns
Show
100
select all
9 sequences selected
GenBank
Graphics
Distance tree of results
New MSA Viewer
Description
Scientific Name
Max Score
Total Score
Query Cover
E value
Per. Ident
Acc. Len
Accession
Caenorhabditis elegans Superoxide dismutase [Cu-Zn] (sod-1), mRNA
Caenorhabditis ele...
1319
1319
99%
0.0
100.00%
720
NM_001026785.4
Caenorhabditis elegans superoxide dismutase mRNA, complete cds
Caenorhabditis ele...
1306
1306
98%
0.0
100.00%
767
L20135.1
Caenorhabditis elegans Superoxide dismutase [Cu-Zn] (sod-1), partial mRNA
Caenorhabditis ele...
881
881
68%
0.0
100.00%
477
NM_001026786.5
Caenorhabditis elegans strain CB4856 chromosome II
Caenorhabditis ele...
486
1353
100%
1e-132
100.00%
15813191
CP038188.1
Caenorhabditis elegans genome assembly C_elegans_Bristol_N2_v1_5_4, scaffold CELN2_scaffold...
Caenorhabditis ele...
486
1353
100%
1e-132
100.00%
251944
LK927570.1
Caenorhabditis elegans strain CB4856 chromosome II
Caenorhabditis ele...
486
1353
100%
1e-132
100.00%
15809916
CP084670.1
Caenorhabditis elegans Cosmid C15F1, complete sequence
Caenorhabditis ele...
486
1353
100%
1e-132
100.00%
37656
FO080553.1
C.elegans sod-1 gene for copper/zinc superoxide dismutase
Caenorhabditis ele...
486
1236
92%
1e-132
100.00%
1350
X77020.1
Caenorhabditis elegans Unclassified non-coding RNA C15F1.13 (C15F1.13), ncRNA
Caenorhabditis ele...
147
147
10%
2e-30
100.00%
104
NR_051306.1



BLAST ® » blastn suite » results for RID-Z5869F4U013

Home Recent Results Saved Strategies Help

< Edit Search | Save Search | Search Summary

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	ref|NM_001026785.1|
RID	Z5869F4U013 *Search expires on 01-28 20:40 pm* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	NM_001026785.1
Description	Caenorhabditis elegans SOD (superoxide dismutase) fami ...
Molecule type	rna
Query Length	719
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to | E value to | Query Coverage to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments

Download | New Select columns | Show 100

select all *9 sequences selected*

GenBank Graphics Distance tree of results New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Caenorhabditis elegans Superoxide dismutase [Cu-Zn] (sod-1), mRNA	Caenorhabditis ele...	1319	1319	99%	0.0	100.00%	720	NM_001026785.4
Caenorhabditis elegans superoxide dismutase mRNA, complete cds	Caenorhabditis ele...	1306	1306	98%	0.0	100.00%	767	L20135.1
Caenorhabditis elegans Superoxide dismutase [Cu-Zn] (sod-1), partial mRNA	Caenorhabditis ele...	881	881	68%	0.0	100.00%	477	NM_001026786.5
Caenorhabditis elegans strain CB4856 chromosome II	Caenorhabditis ele...	486	1353	100%	1e-132	100.00%	15813191	CP038188.1
Caenorhabditis elegans genome assembly C_elegans_Bristol_N2_v1_5_4, scaffold CELN2_scaffold...	Caenorhabditis ele...	486	1353	100%	1e-132	100.00%	251944	LK927570.1
Caenorhabditis elegans strain CB4856 chromosome II	Caenorhabditis ele...	486	1353	100%	1e-132	100.00%	15809916	CP084670.1
Caenorhabditis elegans Cosmid C15F1, complete sequence	Caenorhabditis ele...	486	1353	100%	1e-132	100.00%	37656	FO080553.1
C.elegans sod-1 gene for copper/zinc superoxide dismutase	Caenorhabditis ele...	486	1236	92%	1e-132	100.00%	1350	X77020.1
Caenorhabditis elegans Unclassified non-coding RNA C15F1.13 (C15F1.13), ncRNA	Caenorhabditis ele...	147	147	10%	2e-30	100.00%	104	NR_051306.1

Search Parameters	
Program	blastn
Word size	28
Expect value	0.05
Hitlist size	100
Match/Mismatch scores	1,-2
Gapcosts	0,2.5
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Database	
Posted date	Jan 25, 2022 3:29 AM
Number of letters	644,901,701,245
Number of sequences	78,574,848
Entrez query	None

Karlin-Altschul statistics		
Lambda	1.33271	1.28
K	0.620991	0.46
H	1.12409	0.85

Results Statistics	
Length adjustment	36
Effective length of query	683
Effective length of database	642073006717
Effective search space	438535863587711

How is lambda related to the score matrix?

λ is the unique positive solution to the equation*:

$$\sum_{i,j} p_i r_j e^{\lambda s_{ij}} = 1$$

p_i = freq. of nt i in query, r_j = freq. of nt j in subject

s_{ij} = score for aligning an i,j pair

"Target frequencies"* : $q_{ij} = p_i r_j e^{\lambda s_{ij}}$

How does the % of identify affect your scoring matrix?

If you want to find regions with R% identities:

$$r = R/100$$

$$m = s_{ij} = ln(4(1 - r)/3/ln(4r)$$

Optimal mismatch penalty m for given target identity fraction r

Assuming

$$p_i, p_j = \frac{1}{4}$$

$$m = ln(4(1-r)/3/ln(4r)$$

Examples:

r	0.75	0.95	0.99
m	-1	-2	-3

r = expected fraction of identities in high-scoring BLAST hits

Could you have a 66% identity match using this scoring value?

BLAST® » blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleot

Enter Query Sequence

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

NM_001026785.1

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: Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Nucleotide collection (nr/nt)

Organism (Optional): Enter organism name or id--completions will be suggested exclude Add organism

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): YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for:

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

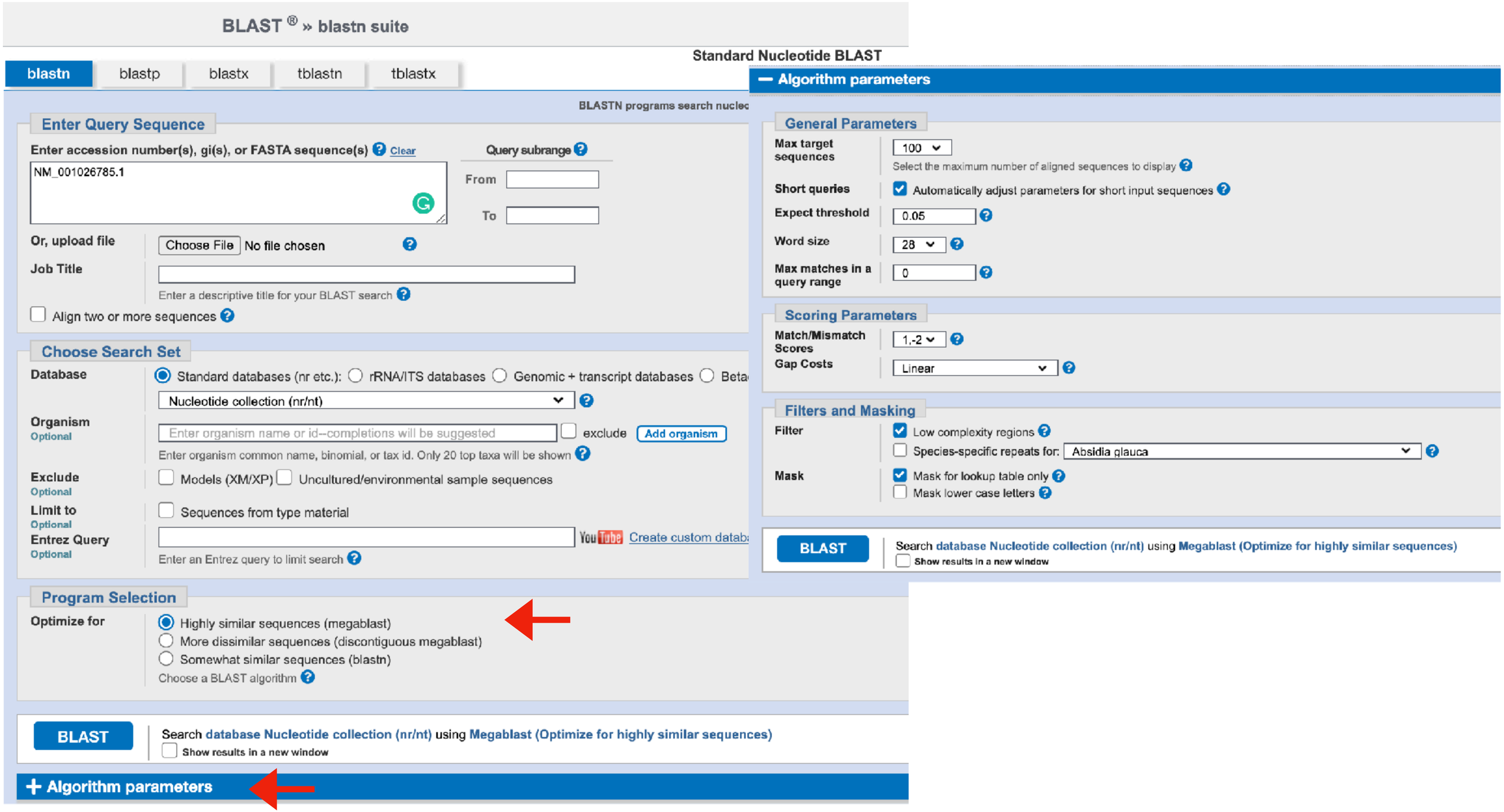
Choose a BLAST algorithm

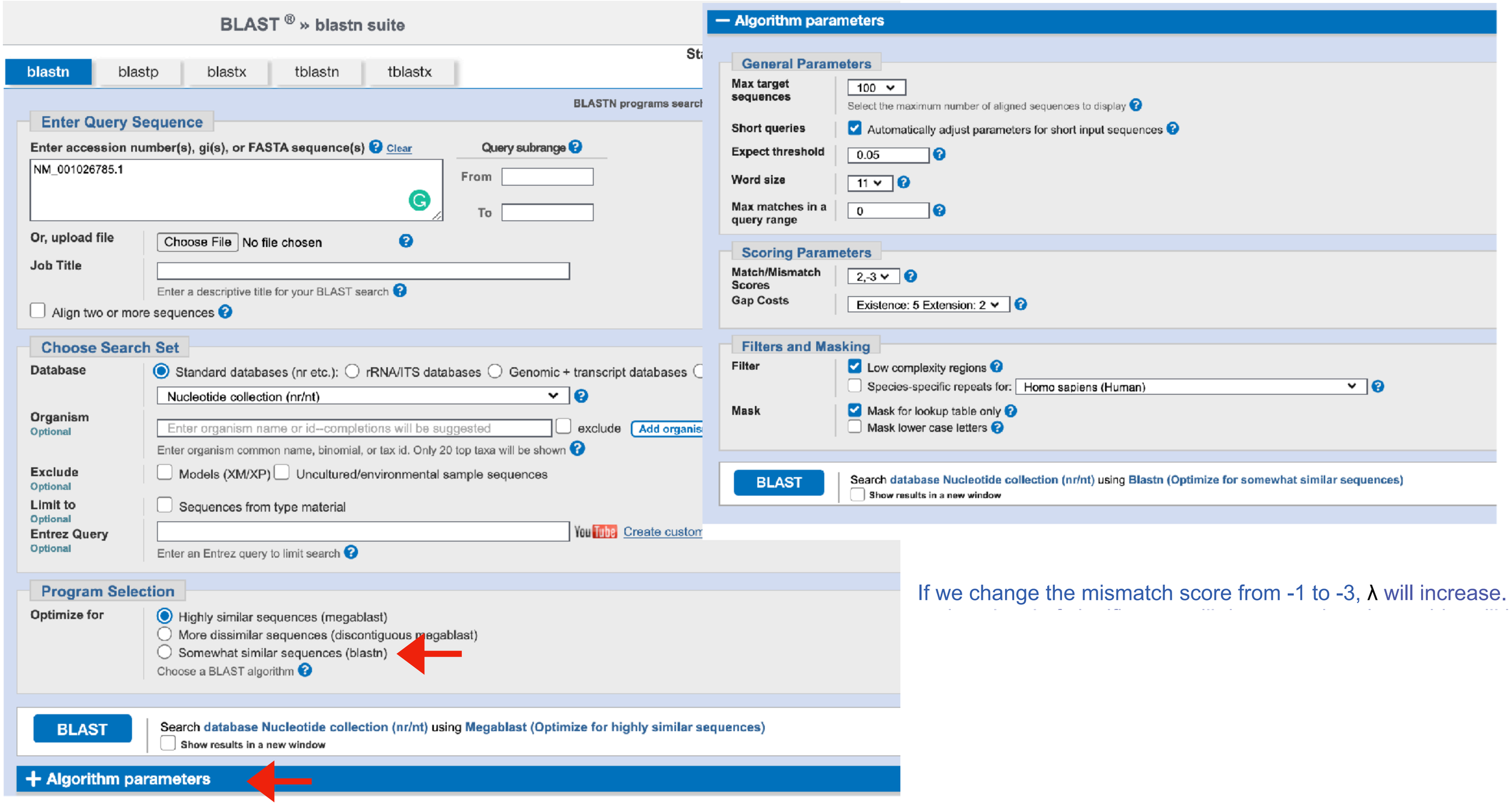
BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

+ Algorithm parameters


BLAST ® » blastn suite
Standard Nucleotide BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nuclec
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
NM_001026785.1
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
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
Add organism
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
Create custom databa
Enter an Entrez query to limit search
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
+ Algorithm parameters
— Algorithm parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
0.05
Word size
28
Max matches in a query range
0
Scoring Parameters
Match/Mismatch Scores
1,-2
Gap Costs
Linear
Filters and Masking
Filter
Low complexity regions
Species-specific repeats for:
Absidia glauca
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window



If we change the mismatch score from -1 to -3, λ will increase.

Common flavors of BLAST:

Program	Query	Database
BLASTP	aa	aa
BLASTN	nt	nt
BLASTX	nt (⇒ aa)	aa
TBLASTN	aa	nt (⇒ aa)
TBLASTX	nt (⇒ aa)	nt (⇒ aa)
PsiBLAST	aa (aa msa)	aa

msa = multiple sequence alignment

What is an E-value?

E-values were designed as the means of bringing some level of confidence to the search results, and are defined as the number of hits one can “expect” to see by chance when searching a database of a particular size. A smaller the e-value is “better”.

The e-value was intended to be used as a filtering and threshold parameter, where, supposedly by using them we could identify more “trustworthy” hits. Fundamentally the problem with e-values is that they depend on the **database size and content, hence are not a measure of validity, reliability or correctness.**

Excerpt From: “The Biostar Handbook: 2nd Edition.”

From Chapter 9: BLAST QuickStart

The alignments found by BLAST during a search are scored, as previously described, and assigned a statistical value, called the "Expect Value." The "Expect Value" **is the number of times that an alignment as good or better than that found by BLAST would be expected to occur by chance, given the size of the database searched.** An "Expect Value" threshold, set by the user, determines which alignments will be reported. A higher "Expect Value" threshold is less stringent and the BLAST default of "10" is designed to ensure that no biologically significant alignment is missed. However, "Expect Values" in the range of 0.001 to 0.0000001 are commonly used to restrict the alignments shown to those of high quality.

Excerpt From: "The Biostar Handbook: 2nd Edition."

At the end…

Since both the alignment lengths and the alignment scores are ingredients to the E-value formula, ranking by E-values typically shows a strong correlation to a ranking by score and lengths. E-values end up as a single “convenient” number that empirically seems to work well.

Excerpt From: “The Biostar Handbook: 2nd Edition.”

What is RefSeq

The NCBI Reference Sequence (RefSeq) project provides sequence records and related information for numerous organisms and provides **a baseline for medical, functional, and comparative studies.** The RefSeq database is a **non-redundant set of reference standards** derived from all the data deposited in GenBank that includes chromosomes, complete genomic molecules (organelle genomes, viruses, plasmids), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins.”

Excerpt From: “The Biostar Handbook: 2nd Edition.”

How are RefSeq sequences named?

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NS_	Genomic	Environmental sequence
NZ_[b]	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_[c]	mRNA	Predicted model
XR_[c]	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	
XP_[c]	Protein	Predicted model, associated with an XM_ accession
ZP_[c]	Protein	Predicted model, annotated on NZ_ genomic records

Remember exercise Part B (lab): the exercise ask to remove XP and XM entries

If you see "_" then the sequence is part of RefSeq

Before you go…

What is the `GENBANK` format?

GenBank format is one of the oldest bioinformatics data formats, originally invented to bridge the gap between a human-readable representation and one that can be efficiently processed by the computer. The format (defined here[1]) has a so-called fixed-width format, where the first ten characters form a column that serves as an identifier and the rest of the lines are information corresponding to that identifier.

GenBank

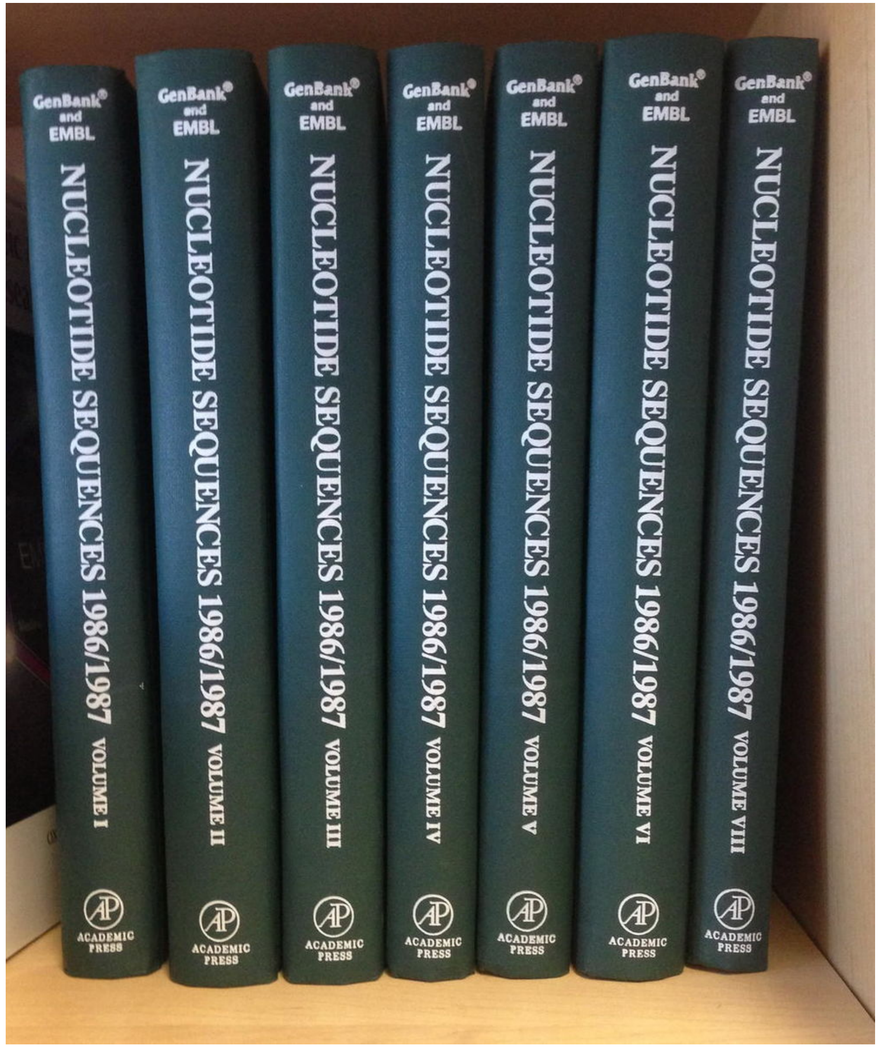


Figure 31.1: GenBank as a book

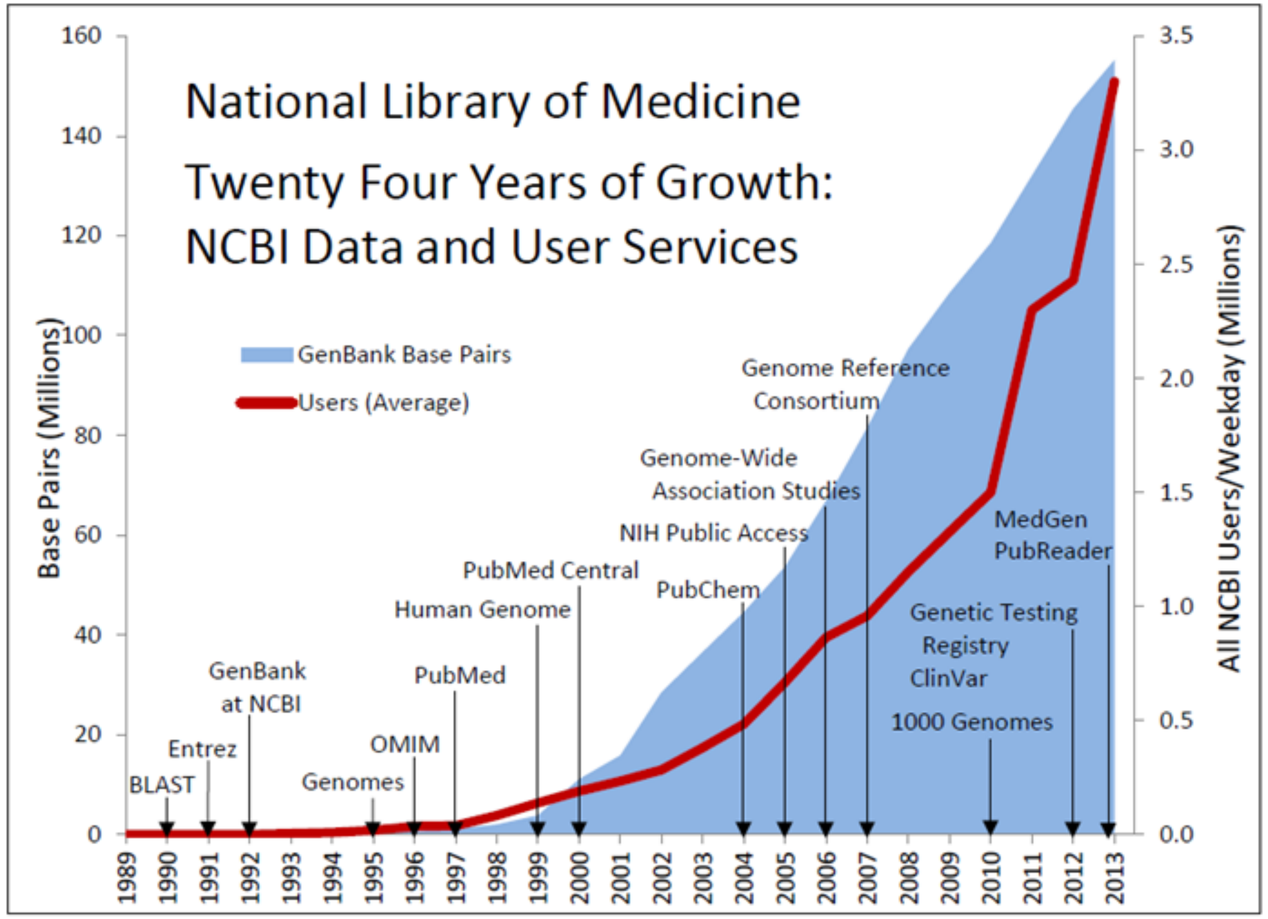
The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. It is produced and maintained by the **National Center for Biotechnology Information** (NCBI; a part of the National Institutes of Health in the United States) as part of the International Nucleotide Sequence Database Collaboration (INSDC).

The database started in 1982 by Walter Goad and Los Alamos National Laboratory. Release 242.0, produced in February 2021, contained over 12 trillion nucleotide bases in more than 2 billion sequences.[4] GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

History

Walter Goad of the Theoretical Biology and Biophysics Group at Los Alamos National Laboratory and others **established the Los Alamos Sequence Database in 1979, which culminated in 1982 with the creation of the public GenBank.**[5] Funding was provided by the National Institutes of Health, the National Science Foundation, the Department of Energy, and the Department of Defense. LANL collaborated on GenBank with the firm Bolt, Beranek, and Newman, and by the end of 1983 more than 2,000 sequences were stored in it.

In the mid 1980s, the Intelligenetics bioinformatics company at Stanford University managed the GenBank project in collaboration with LANL.[6] As one of the earliest bioinformatics community projects on the Internet, the GenBank project started BIOSCI/Bionet news groups for promoting open access communications among bioscientists. During 1989 to 1992, the GenBank project transitioned to the newly created National Center for Biotechnology Information.[7]



from 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months

Insulin was the first protein to be sequenced by F. Sanger

1945

Watson, Crick, Wilkins and Franklin discover the structure of DNA

1953

Berg, Cohen, and Boyer create "recombinant DNA"

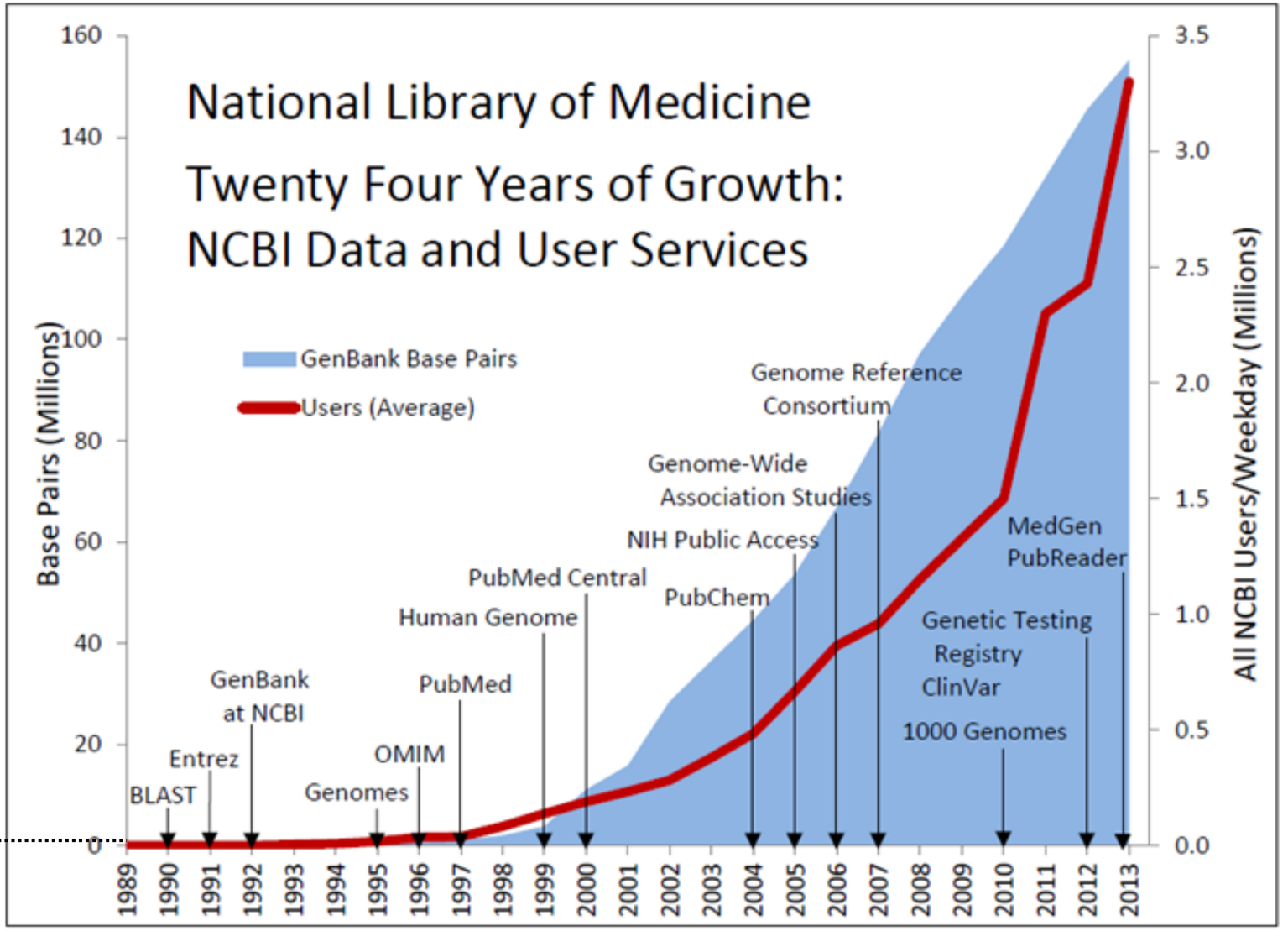
1968

1973

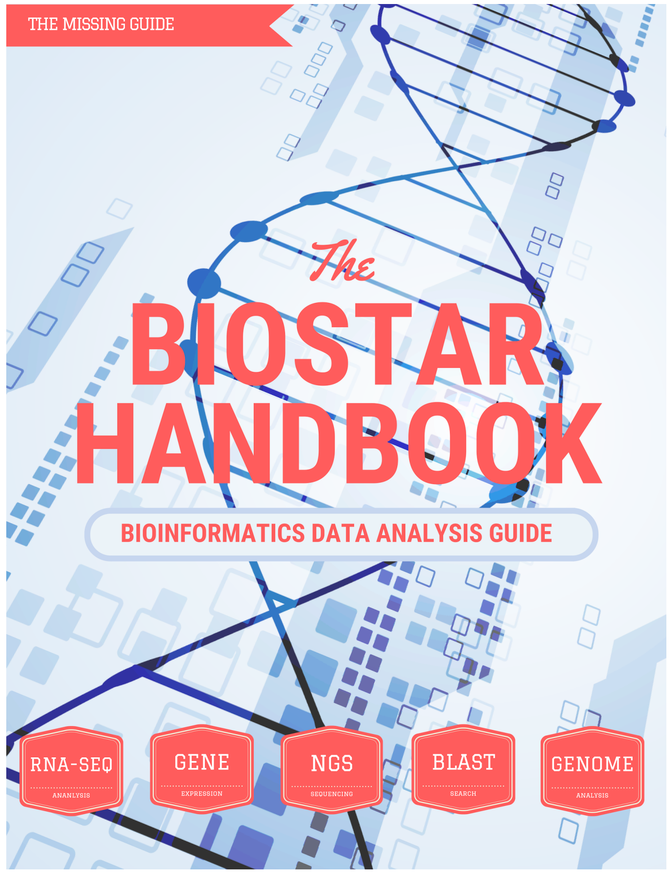
Sequence of DNA, F. Sanger, "Sanger-method"

1977

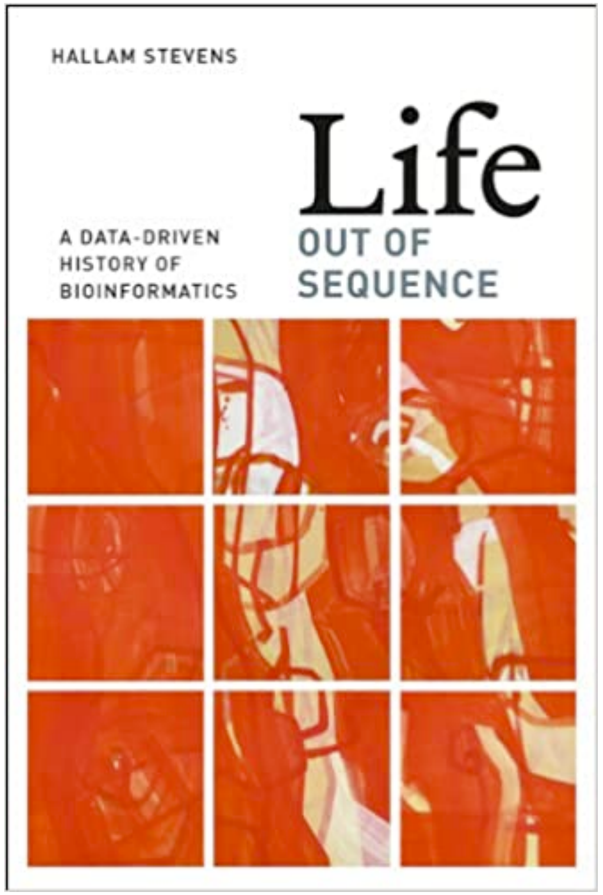
K. Mullis invented PCR



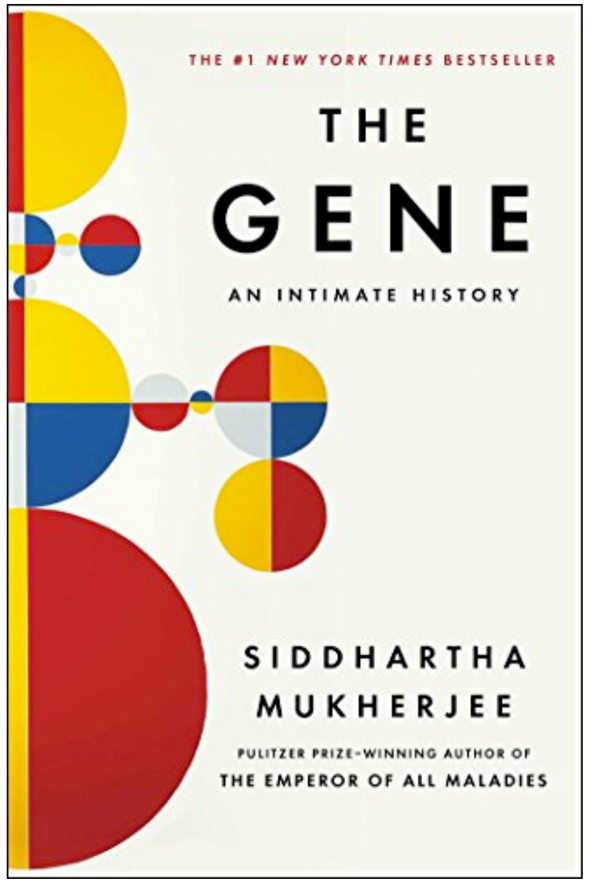
Another thing before you go…



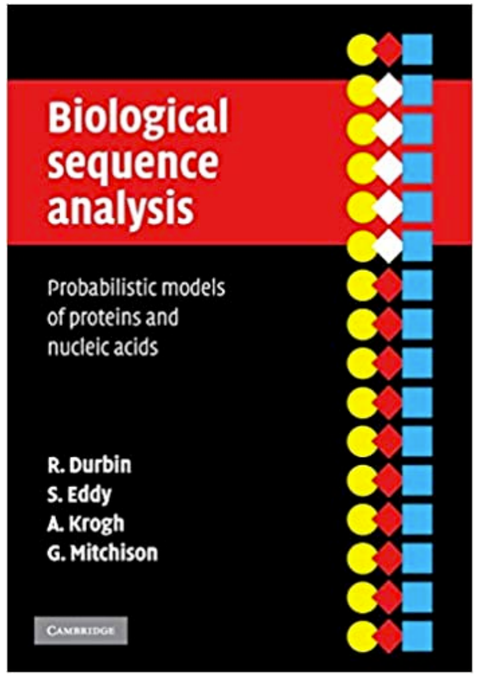
Bioinformatician companion



History of Bioinformatics



History of Genetics
Molecular Biology
Crushing course in Genetics



Hidden Markov Models
Computational Biology Algorithms