Basic Bioinformatics, Sequence Alignment, and Homology

Biochemistry Boot Camp 2023
Session #11

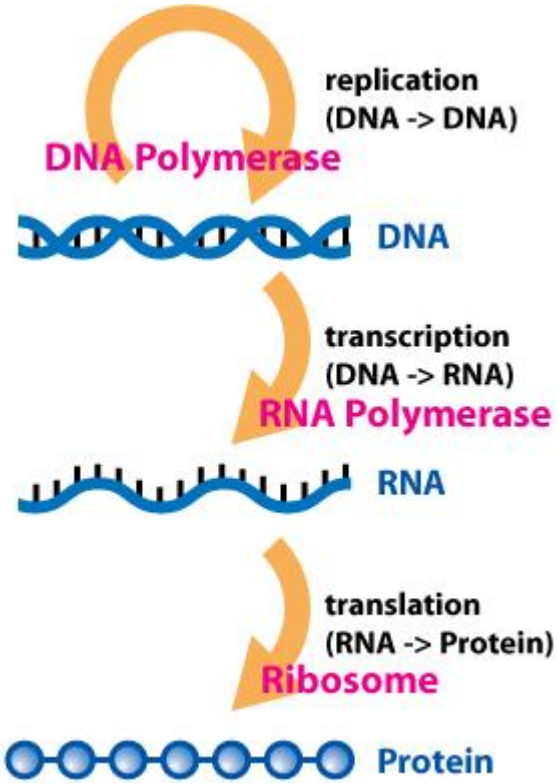
Nick Fitzkee
nfitzkee@chemistry.msstate.edu

* BLAST slides have been adapted from an earlier presentation by W. Shane Sanders.

Biology Review

- Genome is the genetic material of an organism, normally DNA but RNA possible (viruses)

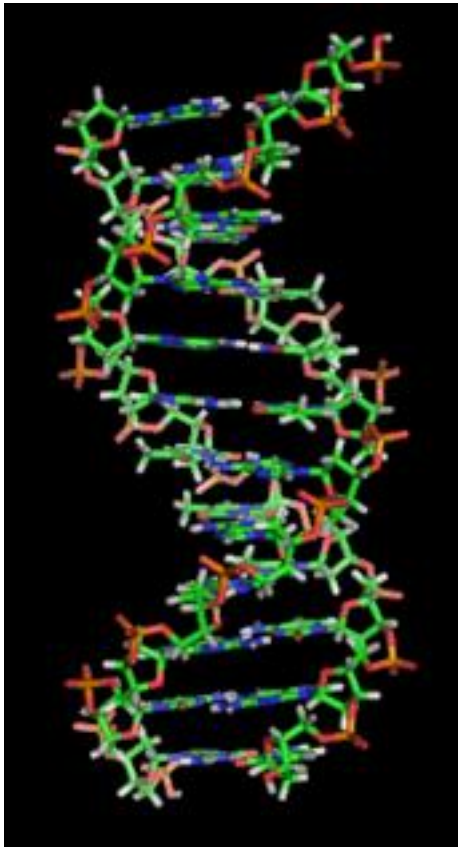
- Central Dogma:
  - DNA → RNA → Protein



The Central Dogma of Molecular Biology

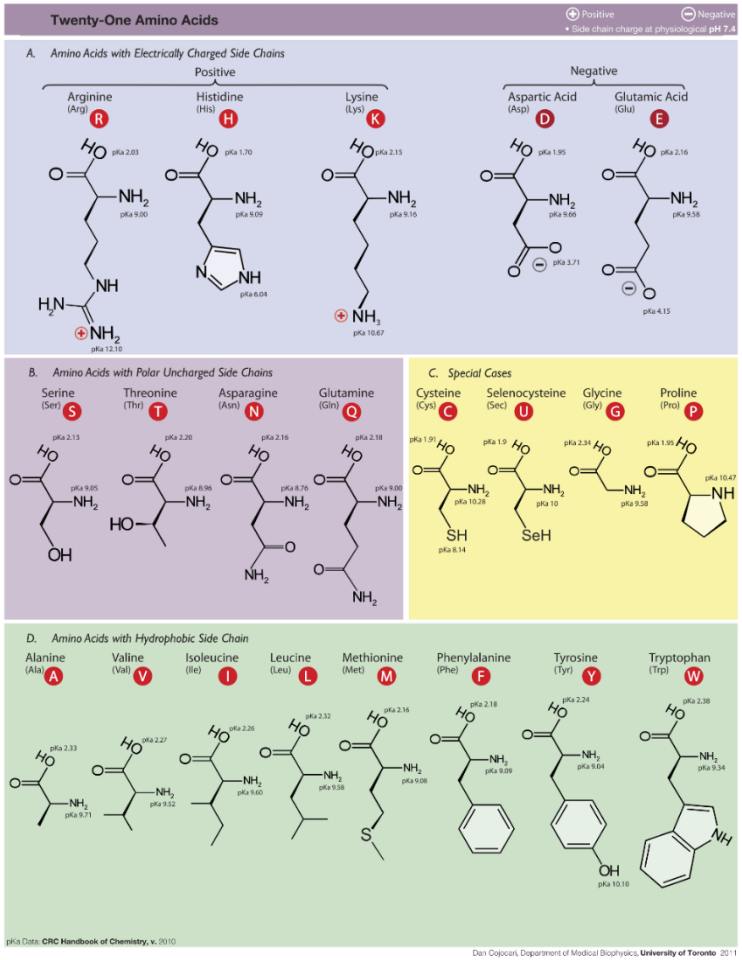
Primary Structure (Sequence)

- **DNA and Proteins are chemically complex**, but their “alphabets” are rather simple.
  - 4 nucleobases (A, C, T, G)
  - 20 amino acids
- DNA sequences are represented from 5’ to 3’



Primary Structure (Sequence)

- **DNA and Proteins are chemically complex**, but their "alphabets" are rather simple.
  - 4 nucleobases (A, C, T, G)
  - 20 amino acids
- Protein sequences are represented from NT to CT



Storing Sequences

- GenBank (*.gb| *.genbank)
  - National Center for Biotechnology's (NCBI) Flat File Format (text)
  - Provides a large amount of information about a given sequence record
  - http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
  - We've seen this before! (Remember NCBI Protein result?)

- **FASTA (*.fasta | *.fa)**
  - Pronounced "FAST-A"
  - Simple text file format for storing nucleotide or peptide sequences
  - Each record begins with a single line description starting with ">" and is followed by one or more lines of sequence

- FASTQ (*.fastq | *.fq)
  - Pronounced "FAST-Q"
  - Text based file format for storing nucleotide sequences and their corresponding quality scores
  - Quality scores are generated as the nucleotide is sequenced and correspond to a probability that a given nucleotide has been correctly sequenced by the sequencer

- **Text files are also okay in many cases.**

Storing Sequences

- FASTA format
- Can represent nucleotide sequences or peptide sequences using single letter codes

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

- FASTQ format
- Represents nucleotide sequences and their corresponding quality scores

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Sequence Alignment

Sequence alignment is the procedure of comparing two (pairwise) or more (multiple) sequences and searching for a series of individual characters or character patterns that are the same in the set of sequences.

- **Global alignment** – find matches along the entire sequence (use for sequences that are quite similar)

- **Local alignment** – finds regions or islands of strong similarity (use for comparing less similar regions [finding conserved regions])

Sequence Alignment

Sequence 1: `GARVEY`

Sequence 2: `AVERY`

Global Alignment:

```
GARVE-Y
-A-VERY
```

Global Sequence Alignment

- EMBOSS Needle
  http://www.ebi.ac.uk/Tools/psa/emboss_needle/
  - Command line version also available

- Alternative: Biopython (library for the python programming language)

- **Example:** Human vs. Nematode Calmodulin (download `sequences.txt` global sequence #1 and #2)

Global Sequence Alignment

- EMBOSS Needle Options:

How much penalty to open a gap in the sequence?

How to compare residues?

STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	OUTPUT FORMAT
BLOSUM62	10	0.5	pair

END GAP PENALTY	END GAP OPEN	END GAP EXTEND
false	10	0.5

Worry about the ends?

How much penalty to have overhang at each end?

Global Sequence Alignment

```
# Length: 149
# Identity:     146/149 (98.0%)
# Similarity:   147/149 (98.7%)
# Gaps:           0/149 ( 0.0%)
# Score: 745.0

Human               1 MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQ     50
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
Nematode            1 MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQ     50

Human              51 DMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGY    100
                      ||||||||||||||||||||||||||||||||||||||||||||||||:
Nematode           51 DMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGF    100

Human             101 ISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK    149
                      |||||||||||||||||||||||||||||||||||||||||.|||.|
Nematode          101 ISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVTMMTTK    149
```

Percent Identity and Similarity quantify alignment.

Identical residues shown with |, similar residues with : and ., and blanks represent dissimilar residues.

- Pretty darn similar!

Multiple Sequence Alignment

- Align many sequences simultaneously, normally from multiple organisms
- Mathematically much more challenging, and requires assumptions about data analysis
- Results can be used to generate phylogenetic tree
  - https://www.ebi.ac.uk/Tools/msa/clustalo/
- Example software: MEGA, https://www.megasoftware.net/

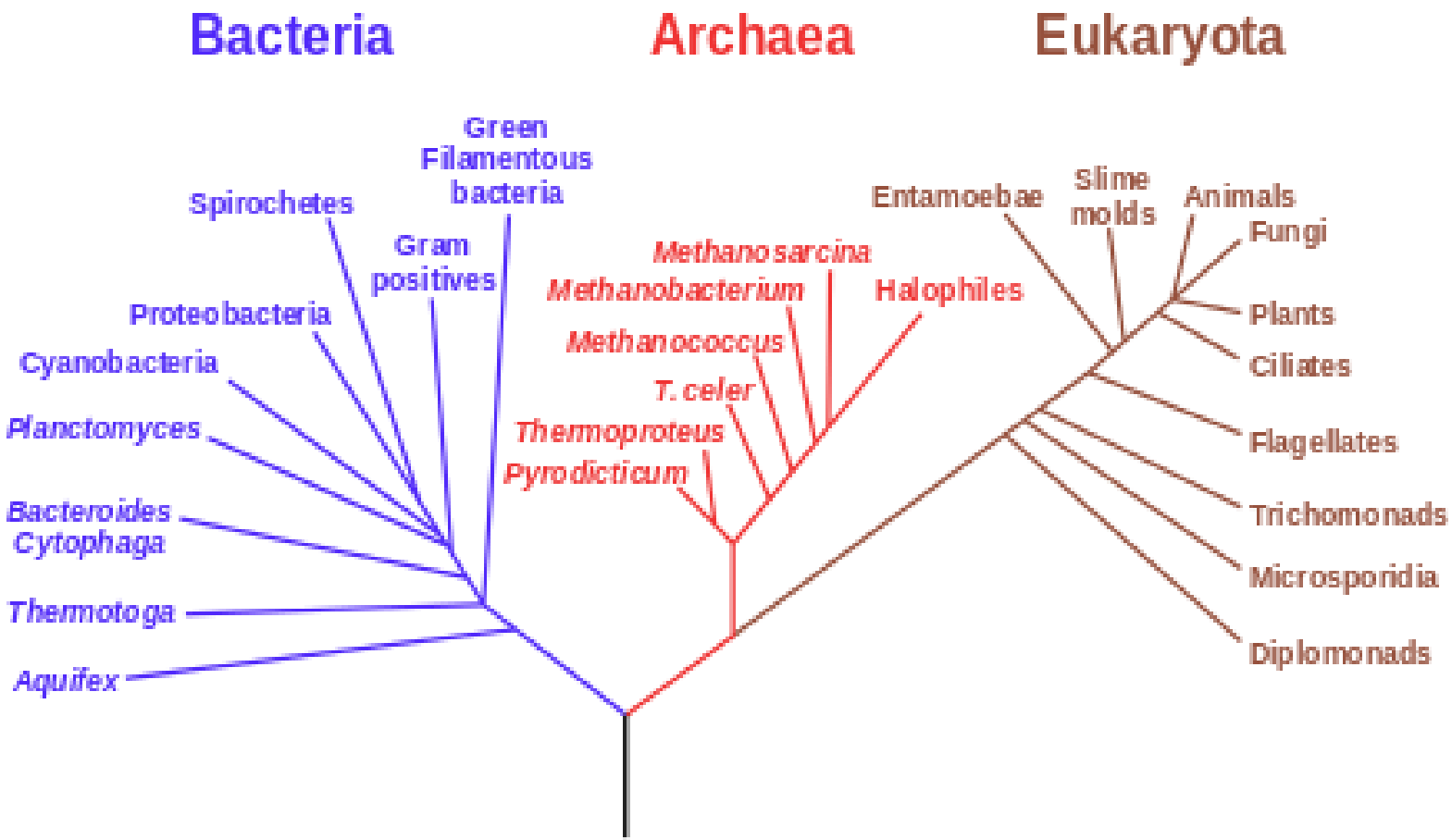
MEGA Molecular Evolutionary Genetics Analysis

MSA Example

```
                                   *   .    :     .         .   *       : : :     .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE    76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE    76
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE    76
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK    79
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS    80
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE    80
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN    86
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE    85
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP    78
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP    78
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL    75
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE    88
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD    74
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA    82
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA    82
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA    81
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE    77
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE    77
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE    77
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE    76
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE    79
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE    79
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD    79
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS    72
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT    72
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV    72
       ruler 1.......10........20........30........40........50........60........70........80........90
```

MSA of Ribosomal Protein P0 from Wikipedia, “Multiple Sequence Alignment”

MSA-Derived Phylogenetic Tree



Phylogenetic Tree derived from ribosomal proteins, Wikipedia “Phylogenetic Tree”

Why Sequence Alignment?

1. To determine possible functional similarity.
2. For 2 sequences:
   a. If they're the same length, are they almost the same sequence? (global alignment)
3. For 2 sequences:
   a. Is the prefix of one string the suffix of another? (contig assembly)
4. Given a sequence, has anyone else found a similar sequence?
5. To identify the evolutionary history of a gene or protein.
6. To identify genes or proteins.

BLAST:
Basic Local Alignment Search Tool

- A tool for determining sequence similarity
- Originated at the National Center for Biotechnology Information (NCBI)
- Sequence similarity is a powerful tool for identifying unknown sequences
- BLAST is fast and reliable
- BLAST is flexible

http://blast.ncbi.nlm.nih.gov/

Flavors of BLAST

- **blastn** – searches a nucleotide database using a nucleotide query
  DNA/RNA sequence searched against DNA/RNA database

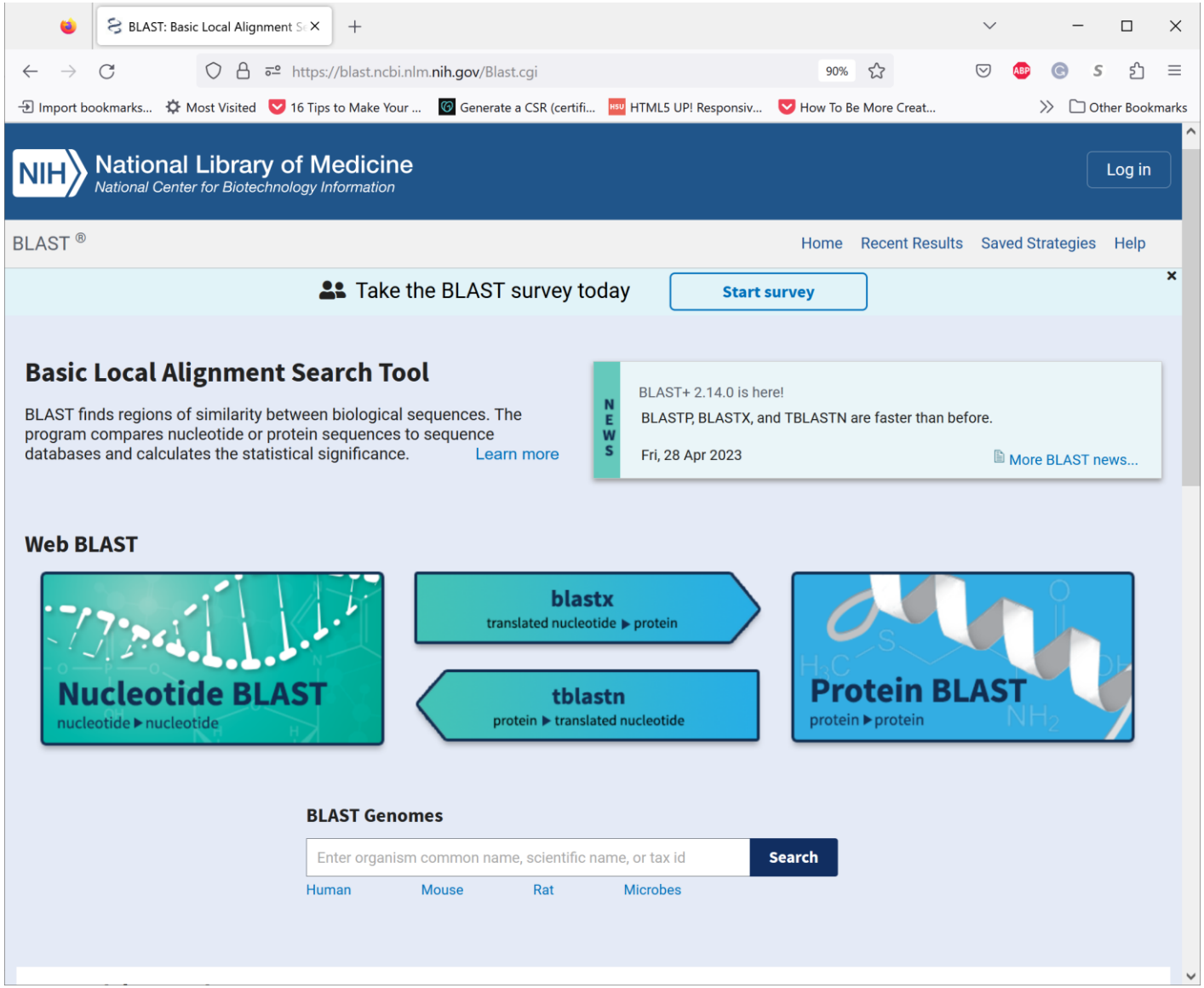
- **blastp** – searches a protein database using a protein query
  Protein sequence searched against a Protein database

- **blastx** – search a protein database using a translated nucleotide query
  DNA/RNA sequence -> Protein sequence searched against a Protein database

- **tblastn** – search a translated nucleotide database using a protein query
  Protein sequence searched against a DNA/RNA sequence database -> Protein sequence database

- **tblastx** – search a translated nucleotide database using a translated nucleotide query
  DNA/RNA sequence -> Protein sequence searched against a DNA/RNA sequence database -> Protein sequence database

BLAST Main Page



Nucleotide BLAST: Search nucle

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blast

Import bookmarks... Most Visited 16 Tips to Make Your ... Generate a CSR (certifi... HTML5 UP! Responsiv... Other Bookmarks

BLAST ® » blastn suite

Home Recent Results Saved Strategies Help

Take the BLAST survey today Start survey

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

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

Reset page

Bookmark

Enter Query Sequence

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

Query subrange

From

To

Or, upload file Browse... No file selected.

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

New Experimental databases Try experimental taxonomic nt databases Download

For more info see What are taxonomic nt 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 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

Sequence Input

Databases to Search Against

Program Selection

Click to Run!

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&BLAST_SPE

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp

Enter Query

Enter accessio

Or, upload file

Job Title

Align two c

Choose Se

Database

Organism Optional

Exclude Optional

Entrez Query Optional

Program S

Algorithm

BLAST

Algorithm pa

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&BLAST_PROGRAMS=blastx&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&BLAST_SPE

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastx

Translated BLAST: blastx

blastn | blastp

Enter Query S

Enter accession

Or, upload file

Genetic code

Job Title

Align two or r

Choose Sear

Database

Organism Optional

Exclude Optional

Entrez Query Optional

BLAST

Algorithm paran

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=tblastn&BLAST_PROGRAMS=tblastn&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&BLAST_SI

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ tblastn

Translated BLAST: tblastn

blastn | blastp | blastx | tblastn | tblastx

Enter Qu

Enter acce

Or, upload

Job Title

Align tw

Choose

Database

Organism Optional

Exclude Optional

Entrez Que Optional

BLAST

Algorithm

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=tblastx&BLAST_PROGRAMS=tblastx&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&BLAST_SF

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ tblastx

Translated BLAST: tblastx

blastn | blastp | blastx | tblastn | tblastx

TBLASTX search translated nucleotide databases using a translated nucleotide query. more...

Reset page | Bookmark

Enter Query Sequence

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

Query subrange

From

To

Or, upload file | Choose File | No file chosen

Genetic code | Standard (1)

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database | Nucleotide collection (nr/nt)

Organism Optional | Enter organism name or id--completions will be suggested | Exclude

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

Exclude Optional | Models (XM/XP) | Uncultured/environmental sample sequences

Same Page Organization

BLAST Example

- What gene is this?

```
>unknown_sequence_1
TGATGTCAAGACCCTCTATGAGACTGAAGTCTTTTCTACCGACTTCTCCAACATTTCTGCAGCCAAGCAG
GAGATTAACAGTCATGTGGAGATGCAAACCAAAGGGAAAGTTGTGGGTCTAATTCAAGACCTCAAGCCAA
ACACCATCATGGTCTTAGTGAACTATATTCACTTTAAAGCCCAGTGGGCAAATCCTTTTGATCCATCCAA
GACAGAAGACAGTTCCAGCTTCTTAATAGACAAGACCACCACTGTTCAAGTGCCCATGATGCACCAGATG
GAACAATACTATCACCTAGTGGATATGGAATTGAACTGCACAGTTCTGCAAATGGACTACAGCAAGAATG
CTCTGGCACTCTTTGTTCTTCCCAAGGAGGGACAGATGGAGTCAGTGGAAGCTGCCATGTCATCTAAAAC
ACTGAAGAAGTGGAACCGCTTACTACAGAAGGGATGGGTTGACTTGTTTGTTCCAAAGTTTTCCATTTCT
GCCACATATGACCTTGGAGCCACACTTTTGAAGATGGGCATTCAGCATGCCTATTCTGAAAATGCTGATT
TTTCTGGACTCACAGAGGACAATGGTCTGAAACTTTCCAATGCTGCCCATAAGGCTGTGCTGCACATTGG
TGAAAAGGGAACTGAAGCTGCAGCTGTCCCTGAAGTTGAACTTTCGGATCAGCCTGAAAACACTTTCCTA
CACCCTATTATCCAAATTGATAGATCTTTCATGTTGTTGATTTTGGAGAGAAGCACAAGGAGTATTCTCT
TTCTAGGGAAAGTTGTGAACCCAACGGAAGCGTAGTTGGGAAAAAGGCCATTGGCTAATTGCACGTGTGT
ATTGCAATGGGAAATAAATAAATAATATAGCCTGGTGTGATTGATGTGAGCTTGGACTTGCATTCCCTTA
TGATGGGATGAAGATTGAACCCTGGCTGAACTTTGTTGGCTGTGGAAGAGGCCAATCCTATGGCAGAGCA
TTCAGAATGTCAATGAGTAATTCATTATTATCCAAAGCATAGGAAGGCTCTATGTTTGTATATTTCTCTT
TGTCAGAATACCCCTCAACTCATTTGCTCTAATAAATTTGACTGGGTTGAAAAATTAAAA
```

BLAST Results

NCBI Blast:Nucleotide Sequenc

https://blast.ncbi.nlm.nih.gov/Blast.cgi 80%

Import bookmarks... Most Visited 16 Tips to Make Your ... Generate a CSR (certifi... HTML5 UP! Responsiv... Other Bookmarks

An official website of the United States government Here's how you know

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST ® » blastn suite » **results for RID-7YY6WY0C016**

Home Recent Results Saved Strategies Help

Take the BLAST survey today Start survey

< Edit Search Save Search Search Summary

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

Job Title	Nucleotide Sequence
RID	7YY6WY0C016 Search expires on 06-08 02:20 am Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_27249
Description	None
Molecule type	dna
Query Length	1110
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

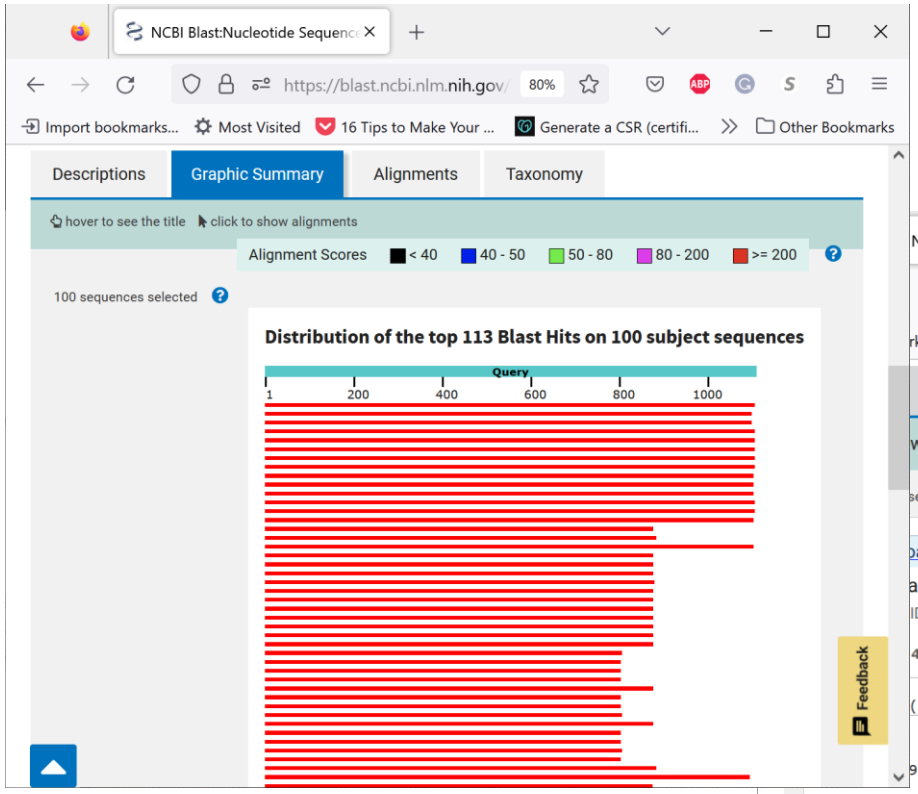
Sequences producing significant alignments Download Select columns Show 100

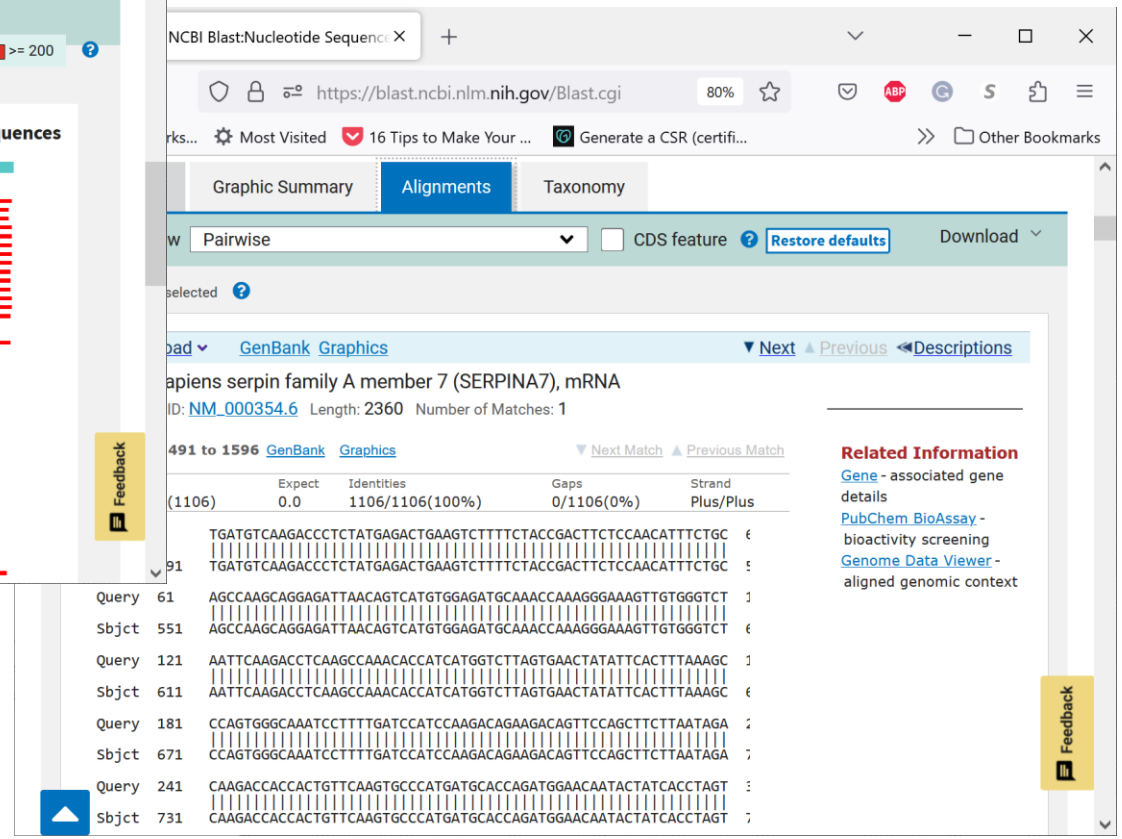
select all 100 sequences selected GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens serpin family A member 7 (SERPINA7), mRNA	Homo sapiens	2043	2043	99%	0.0	100.00%	2360	NM_000354.6
Human thyroxine-binding globulin mRNA, complete cds	Homo sapiens	2019	2019	99%	0.0	99.82%	1872	M14091.1
Pan troglodytes serpin family A member 7 (SERPINA7), mRNA	Pan troglodytes	2012	2012	99%	0.0	99.64%	1589	NM_001009109.1
PREDICTED: Pongo pygmaeus serpin family A member 7 (SERPINA7), mRNA	Pongo pygmaeus	1977	1977	99%	0.0	98.92%	2371	XM_054471959.1
PREDICTED: Pongo abelii serpin family A member 7 (SERPINA7), mRNA	Pongo abelii	1965	1965	99%	0.0	98.73%	2371	XM_002831954.5
PREDICTED: Symphalangus syndactylus serpin family A member 7 (SERPINA7), mRNA	Symphalangus ...	1943	1943	99%	0.0	98.37%	2365	XM_055269099.1
PREDICTED: Papio anubis serpin family A member 7 (SERPINA7), transcript variant X2, mRNA	Papio anubis	1908	1908	99%	0.0	97.83%	1645	XM_003918082.5
PREDICTED: Macaca mulatta serpin family A member 7 (SERPINA7), transcript variant X1, mRNA	Macaca mulatta	1897	1897	99%	0.0	97.65%	1656	XM_001088790.4

Feedback

BLAST Results – Graphical Summary and Alignments





Interpreting BLAST Results

- **Max Score** – how well the sequences match
- **Total Score** – includes scores from non-contiguous portions of the subject sequence that match the query
- **Bit Score** – A log-scaled version of a score
  - Ex. If the bit-score is 30, you would have to score on average, about 2^{30} = 1 billion independent segment pairs to find a score matching this score by chance. Each additional bit doubles the size of the search space.
- **Query Coverage** – fraction of the query sequence that matches a subject sequence
- **E value** – how likely an alignment can arise by chance
- **Max ident** – the match to a subject sequence with the highest percentage of identical bases

Installing BLAST Locally

Executables and documentation available at:

https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

Documentation:
https://www.ncbi.nlm.nih.gov/books/NBK1762/

Aligning via Structure

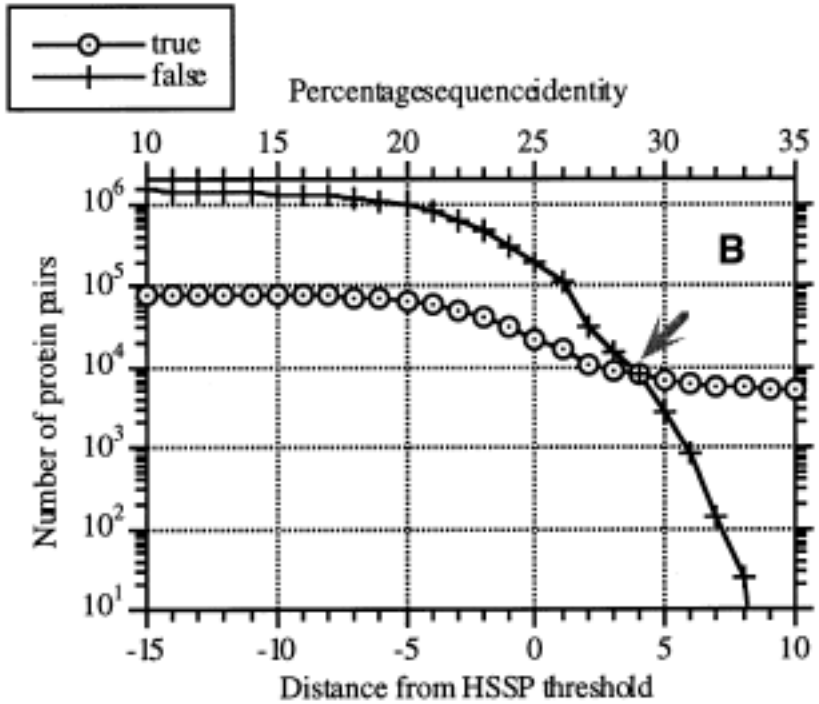
- So far we've focused on *sequence* alignment: looking at the primary (DNA or protein) sequence

- What about *structural* alignment? (Think shape or similar domains)

- VAST (Vector Alignment Search Tool) at NCBI: https://structure.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml

Homology Modeling

- Proteins with similar <u>sequences</u> tend to have similar <u>structures</u>.
- When sequence identify is greater than ~25%, this rule is almost guaranteed
  - Exception: See Lauren Perskie-Porter, Phil Bryan and "fold switching"
- Can we predict structures?



Below ~28% sequence identity, the number of structurally <u>dis</u>similar aligned pairs explodes.

Rost, *Prot. Eng.* 12(2): 85-94

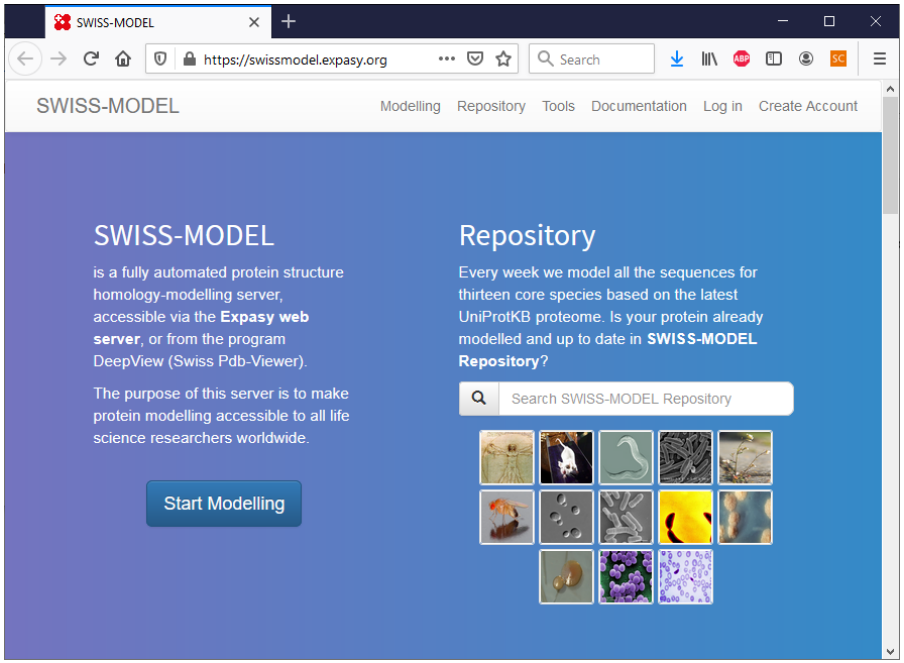
What is Homology Modeling?

- **Consider:** Protein with known sequence, but unknown structure

- Use sequence alignment (protein BLAST) to identify similar sequences with known structures
  - These are termed "template structures"

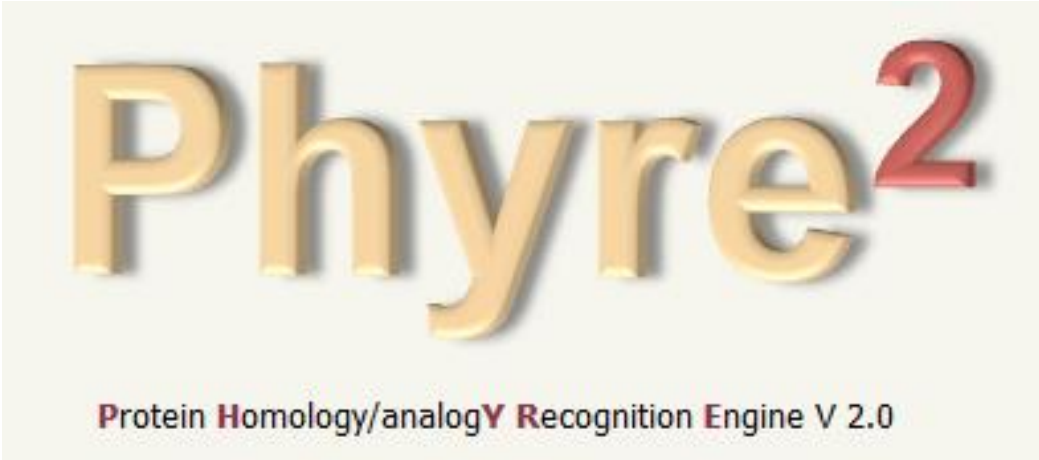
- "Map" unknown sequence onto known backbone
  - Side chains may be more ill-defined: <u>it's a model!</u>

Homology Modeling Servers: **SWISS-MODEL**



- Web page: https://swissmodel.expasy.org/
- Fastest option, can take less than 5 minutes
- Final model typically based on a single template (users can upload their own)

Homology Modeling Servers: **Phyre^2**



- Web page: http://www.sbg.bio.ic.ac.uk/phyre2/
- Trade off: can take 1-2 hours depending on server demand, but better structures
- Uses multiple templates, users can exclude files

Homology Modeling Servers: **I-TASSER**

- Web page: https://zhanggroup.org/I-TASSER/
- Slowest option by far; can take a day or more
- Uses multiple templates and performs sophisticated refinement

Homology Modeling Example

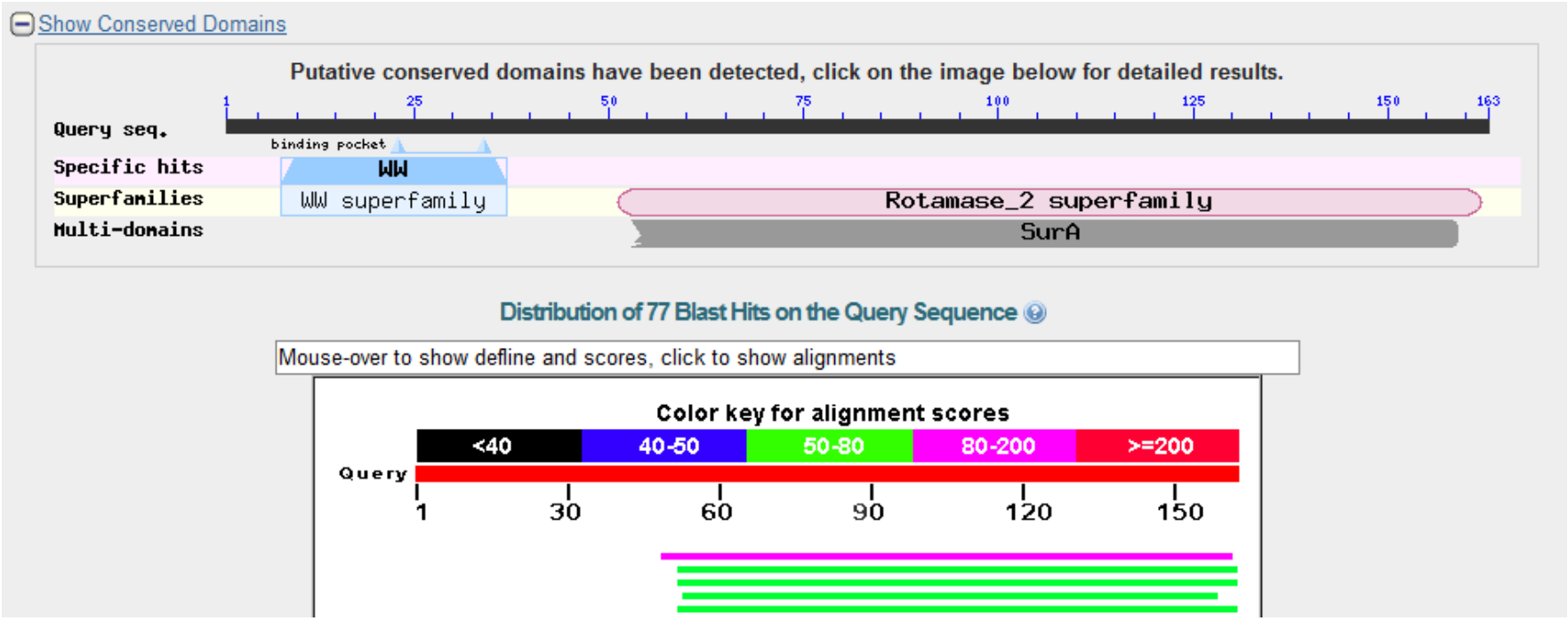
- Sequence for Pin1 protein:

```
MADEEKLPPG WEKRMSRSSG RVYYFNHITN ASQWERPSGN SSSGGKNGQG
EPARVRCSHL LVKHSQSRRP SSWRQEKITR TKEEALELIN GYIQKIKSGE
EDFESLASQF SDCSSAKARG DLGAFSRGQM QKPFEDASFA LRTGEMSGPV
FTDSGIHIIL RTE
```

- Use BLAST to identify a homologous cis-trans prolyl isomerase in *Methanocorpusculum labreanum*

Homology Modeling Example

- Initial BLASTp result:



- Sequence (only second domain found):

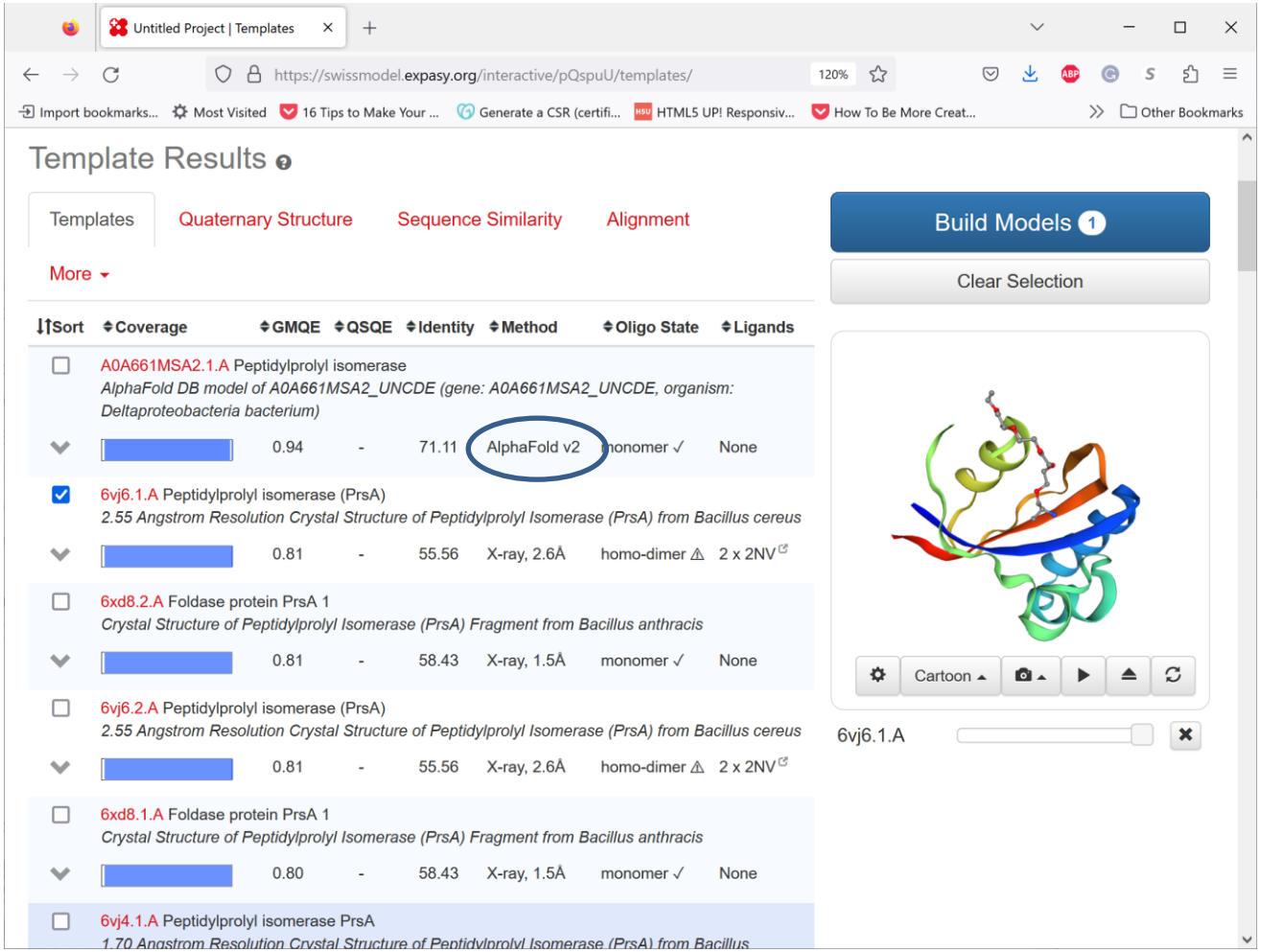
MVRVKASHIL VKTEAQAKEI MQKISAGDDF AKLAKMYSQC PSGNAGGDLG
YFGKGQMVKP FEDACFKAKA GDVVGPVKTQ FGWHIIKVTD IKN

Result: SWISS-MODEL

"Searching for templates" lets you select which structure(s) are used to build your homology model.

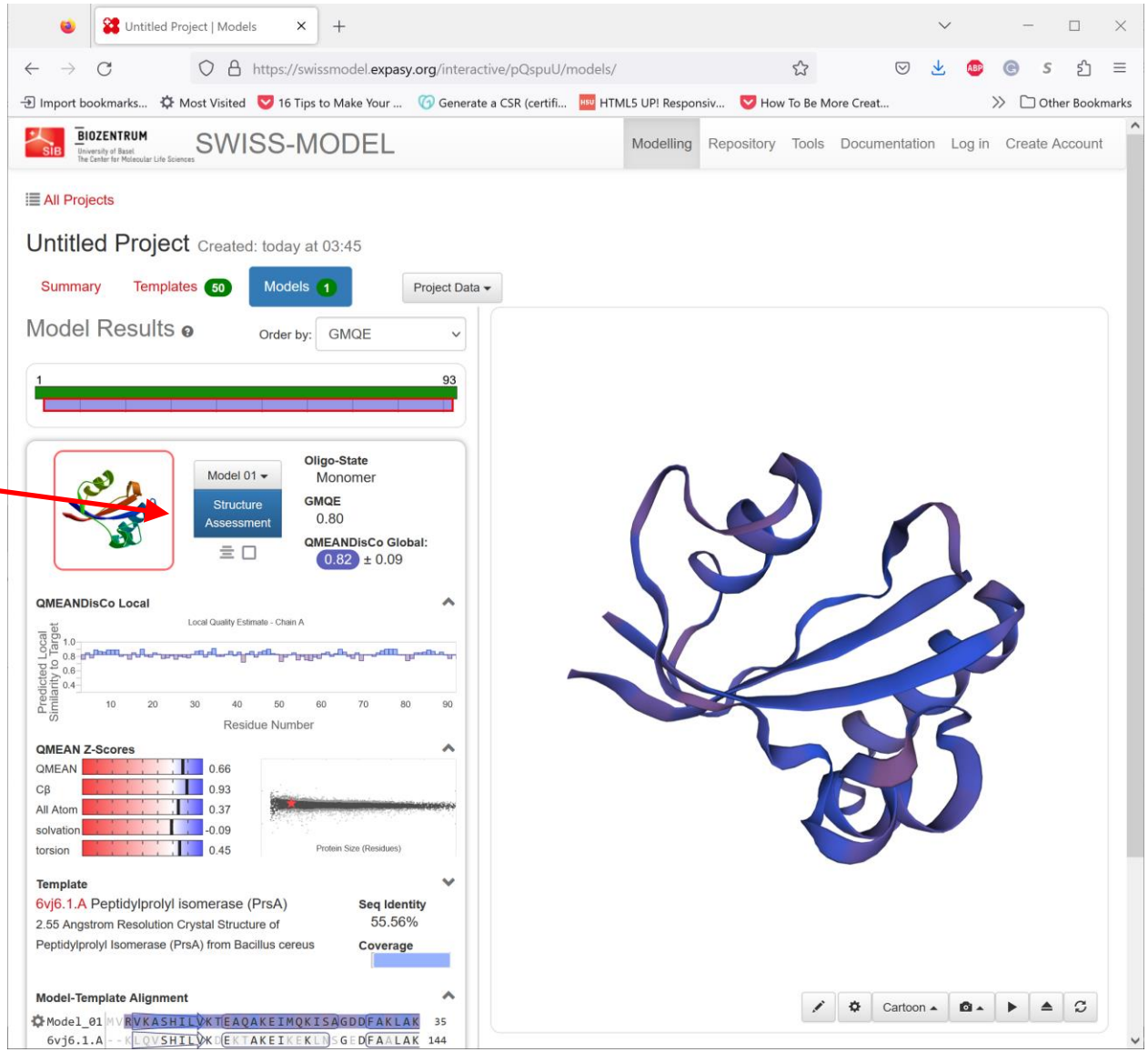
If you aren't comfortable using AlphaFold structures, you can deselect them!

Alphafold have higher sequence similarity, but you're be building a model derived from a model



Result: SWISS-MODEL

Click here to view Ramachandran plots, structure quality by residue, etc.



Click structure to download PDB file

Result: Phyre^2

Top model

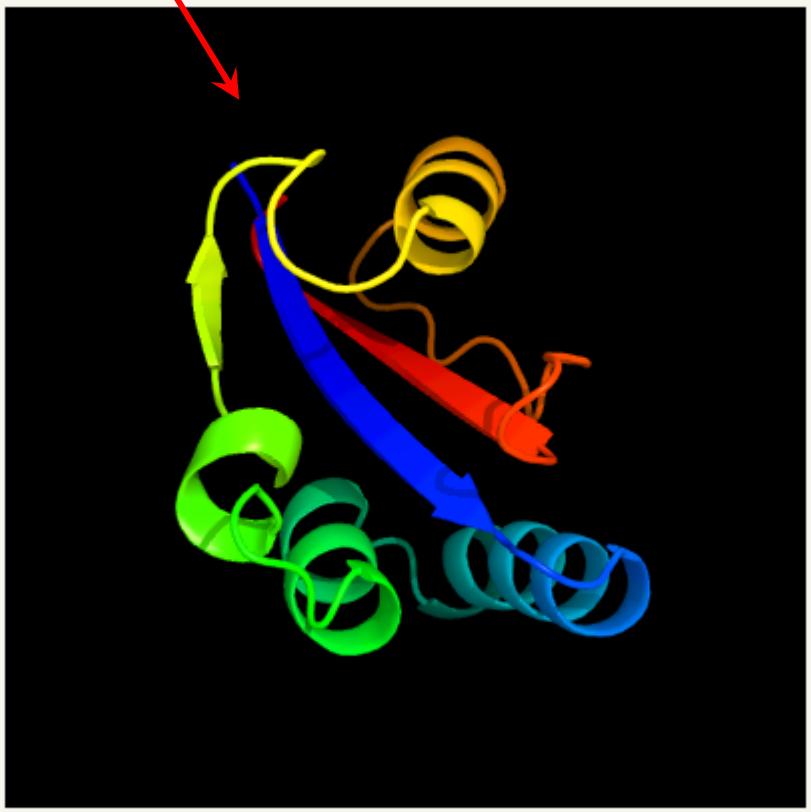


Image coloured by rainbow N → C terminus

Model dimensions (Å): **X**:38.631 **Y**:32.251 **Z**:31.193

Model (left) based on template d1jnsa

Top template information

Fold:FKBP-like
Superfamily:FKBP-like
Family:FKBP immunophilin/proline isomerase

Confidence and coverage

Confidence:	99.9%	Coverage:	96%

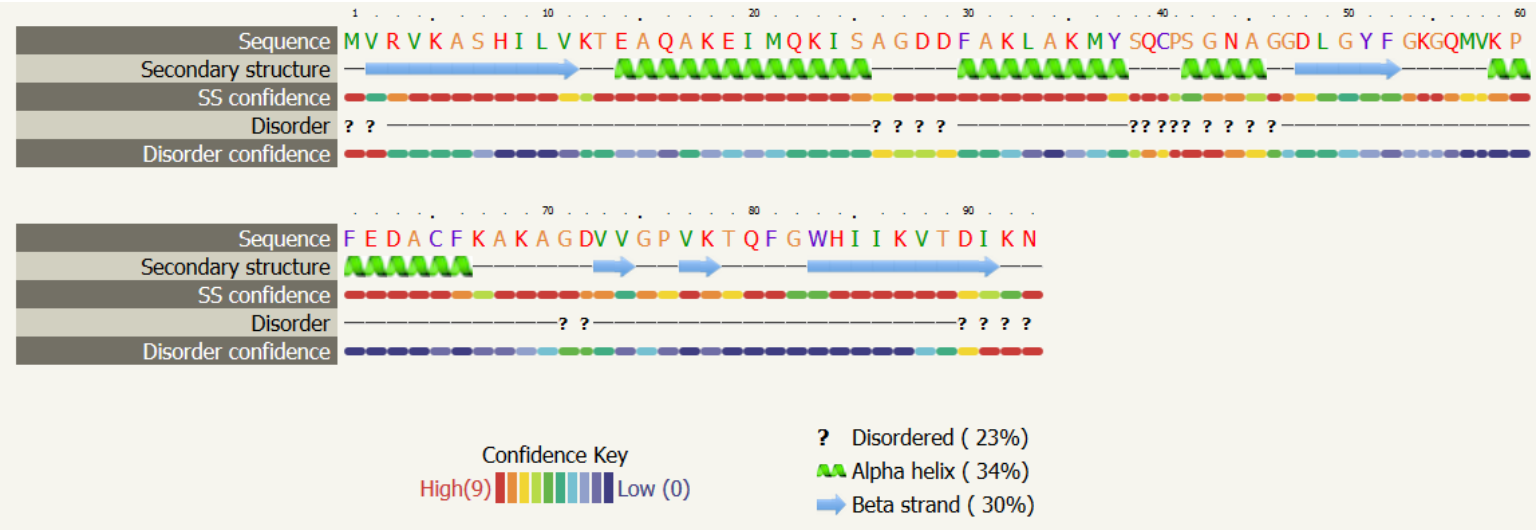
89 residues (96% of your sequence) have been modelled with 99.9% confidence by the single highest scoring template.

3D viewing

Interactive 3D view in JSmol

For other options to view your downloaded structure offline see the FAQ

Result: Phyre^2



- Download entire result, which is a duplicate of the website, can be viewed here:

  https://fitzkee.chemistry.msstate.edu/sites/default/files/bootcamp/phyre2/summary.html

- Final result is called `final.casp.pdb`

Result: I-TASSER

Predicted Secondary Structure

```
                    20                  40                  60                  80
                     |                   |                   |                   |
Sequence    MVRVKASHILVKTEAQAKEIMQKISAGDDFAKLAKMYSQCPSGNAGGDLGYFGKGQMVKPFEDACFKAKAGDVVGPVKTQFGWHIIKVTDIKN
Prediction  CCSSSSSSSSSCCHHHHHHHHHHHCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCCHHHHHHHHCCCCCCCCCCSSCCCSSSSSSSSSSCC
Conf.Score  967998899988999999999998879989999998688965244864553379973569999998389999788777698379999967659
            H:Helix; S:Strand; C:Coil
```

Predicted Solvent Accessibility

```
                    20                  40                  60                  80
                     |                   |                   |                   |
Sequence    MVRVKASHILVKTEAQAKEIMQKISAGDDFAKLAKMYSQCPSGNAGGDLGYFGKGQMVKPFEDACFKAKAGDVVGPVKTQFGWHIIKVTDIKN
Prediction  764340311116357405502630673640351056317344376323233045662243025003716645336234163100003046458
            Values range from 0 (buried residue) to 9 (highly exposed residue)
```

- Results available at:

  https://fitzkee.chemistry.msstate.edu/sites/default/files/bootcamp/itasser/

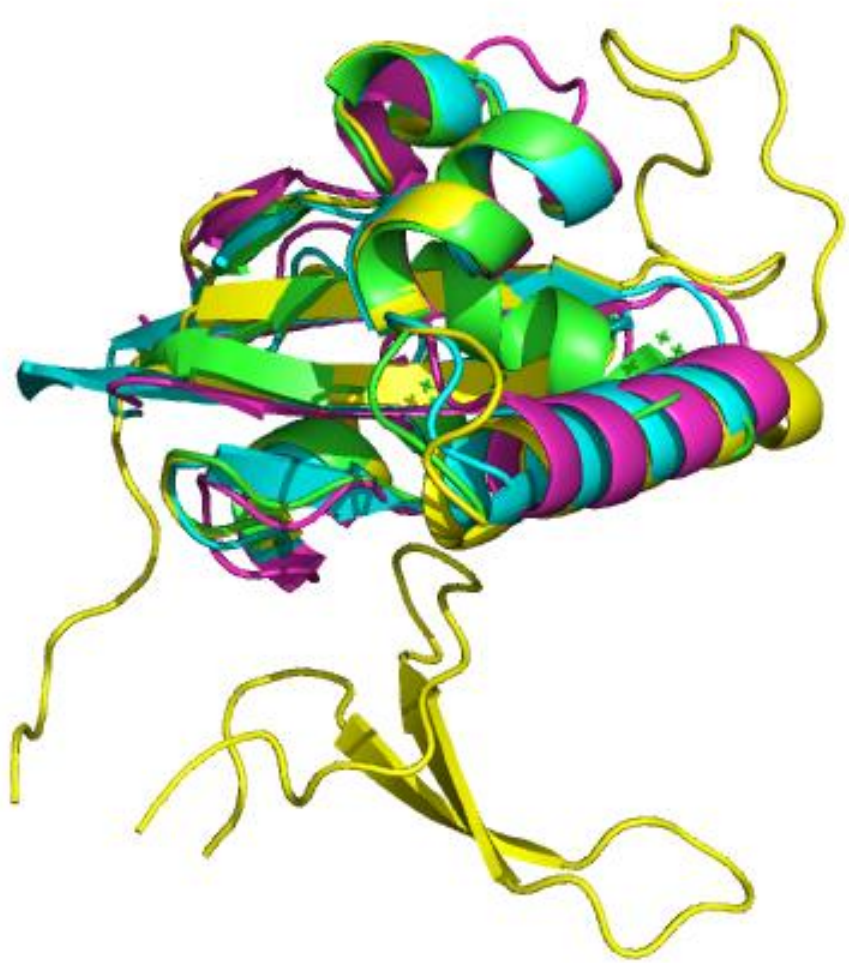
- Final result is called `model1.pdb`

Comparison of Results

- **Download the following PDBs from the Boot Camp Website:**
  - 1pin.pdb – Original Pin1 Structure
  - swiss.pdb – SWISS-MODEL Result
  - phyre2.pdb – Phyre^2 Result
  - itasser.pdb – I-TASSSER Result

- PyMOL can help us here using the “align” command (align.pse)

Comparison of Results

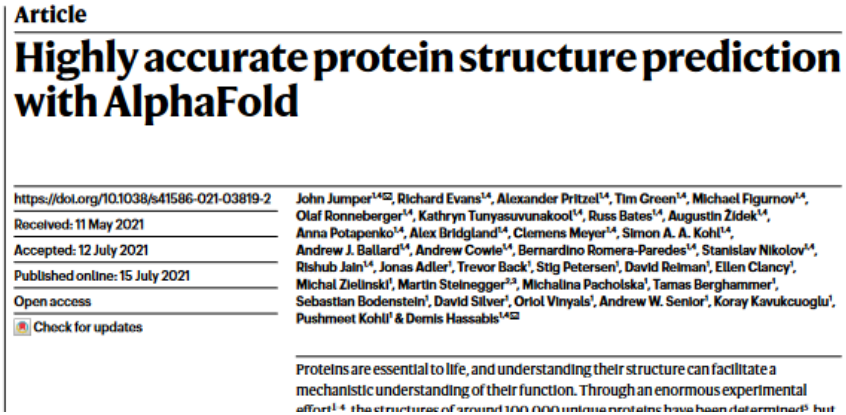


- Colors:
  - **Original Pin1**
  - **SWISS-MODEL**
  - **Phyre^2**
  - **I-TASSER**

- **Important:** How much side chain accuracy do I need?

AlphaFold2: Neural Networks

- Google Deepmind Project: Exhaustively predict protein structure based on known structure patterns



Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but

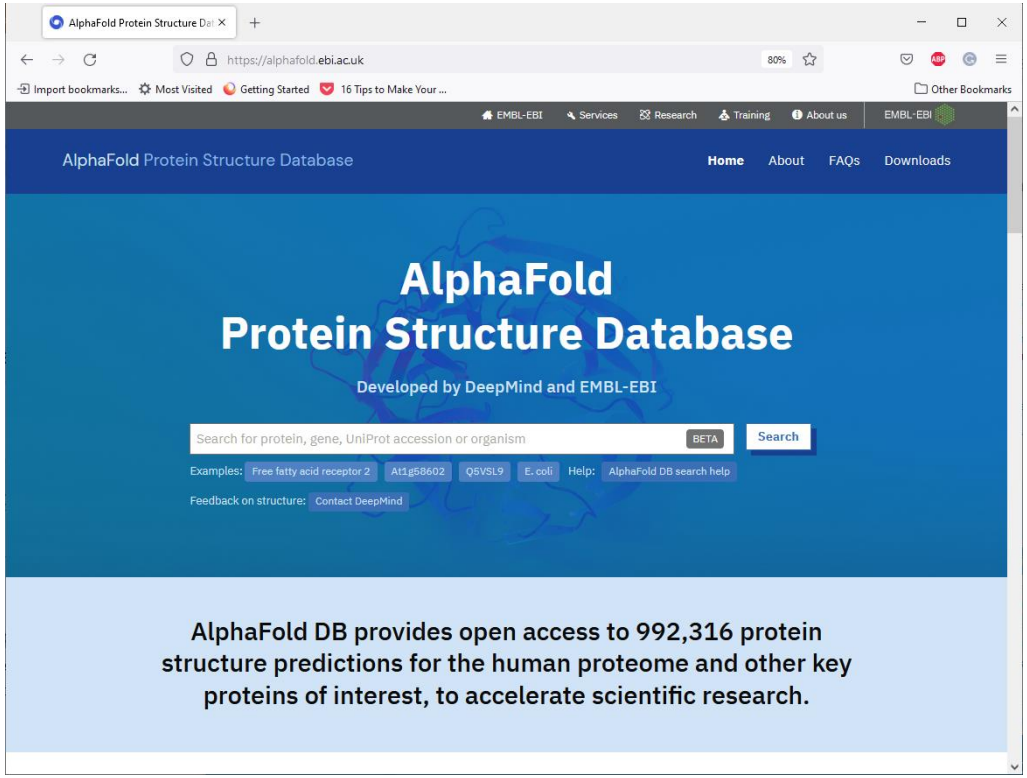
- Not really homology modeling, not really “ab initio” or physics-based

- Extremely successful!

Jumper, *et al.* (2021) *Nature.* ***596:*** *583.* https://doi.org/10.1038/s41586-021-03819-2

AlphaFold2 Website

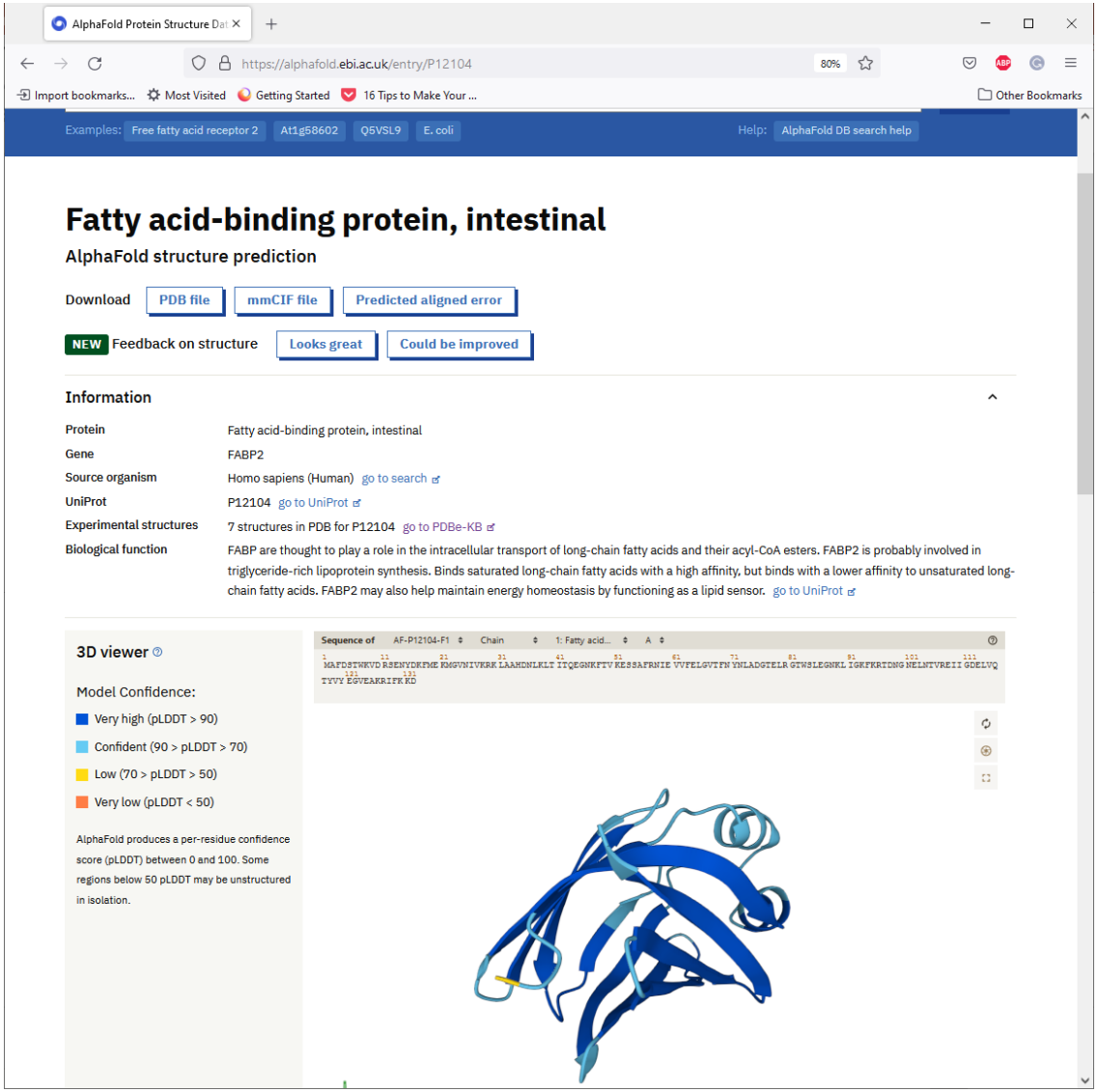
- **Prediction Database:** https://alphafold.ebi.ac.uk/



- Entry: P12104 (Human Intestinal Fatty Acid Binding Protein)

FABP Entry – P12104

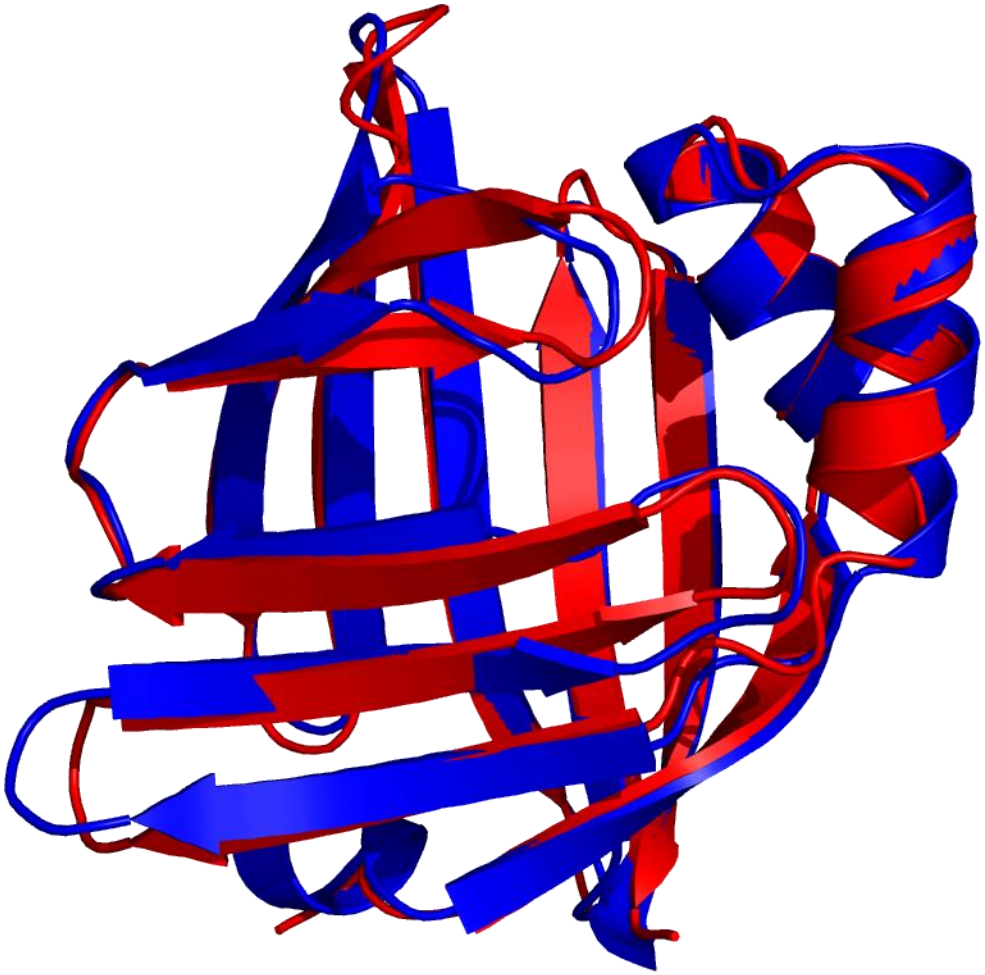
- Many entries exist, but not so easy to run this yourself on a new structure
- For more information check out the DeepMind website
- https://www.deepmind.com/research/highlighted-research/alphafold



Comparison of AlphaFold2 vs 6L9O

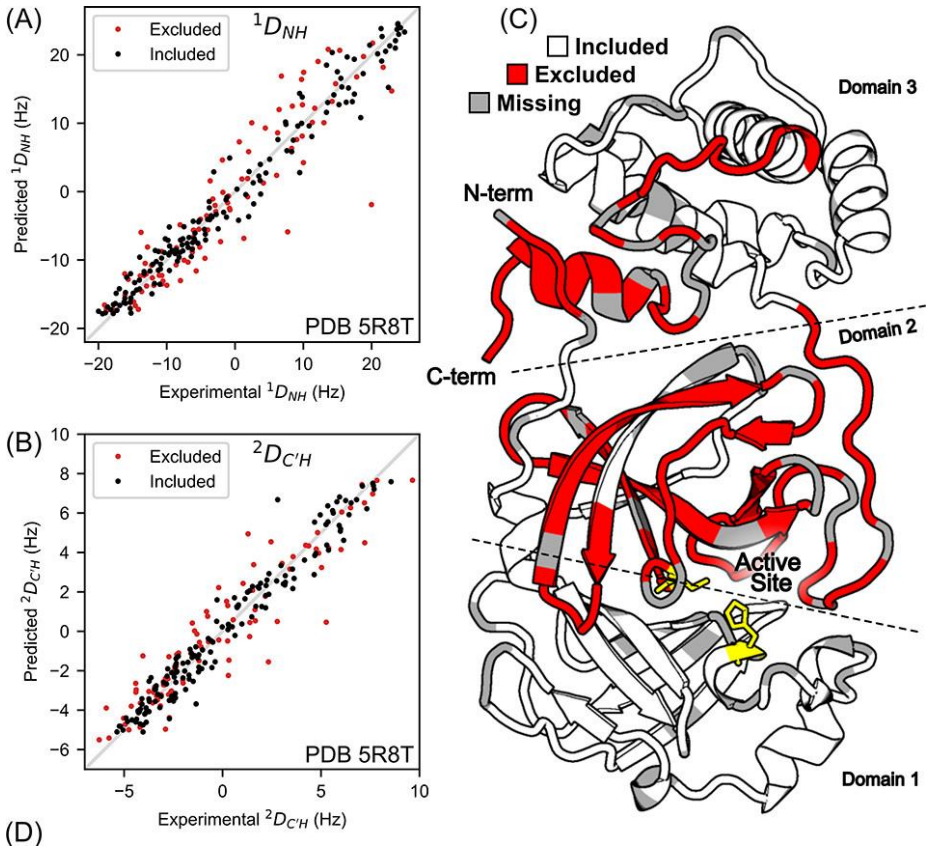
- **Red:** AlphaFold2
- **Blue:** Experimental crystal structure

- Aligned using PyMOL (align command)



AlphaFold2 Limitations

- Performs well for folded, compact regions
- Less good on loops, dynamic regions (SARS-CoV2 MPro, right)
- Very bad on disordered proteins (IDPs) → makes sense!
- **Verdict:** It's a great starting point, like many other models



Robertson, *et al.* (2021) *JACS.* ***143:*** *19306.* https://doi.org/10.1021/jacs.1c10588

Summary

- Sequence alignment is an important tool for searching and understanding how proteins are related
- BLAST can be used to search for similar sequences in large protein/DNA databases (and also works in tools like the PDB)
- Homology modeling can be helpful way to understand structures of unknown proteins
- AlphaFold2 is probably the future, but not good for disordered proteins; it’s still a model!