Basic Bioinformatics, Sequence Alignment, and Homology

Biochemistry Boot Camp 2022
Session #11

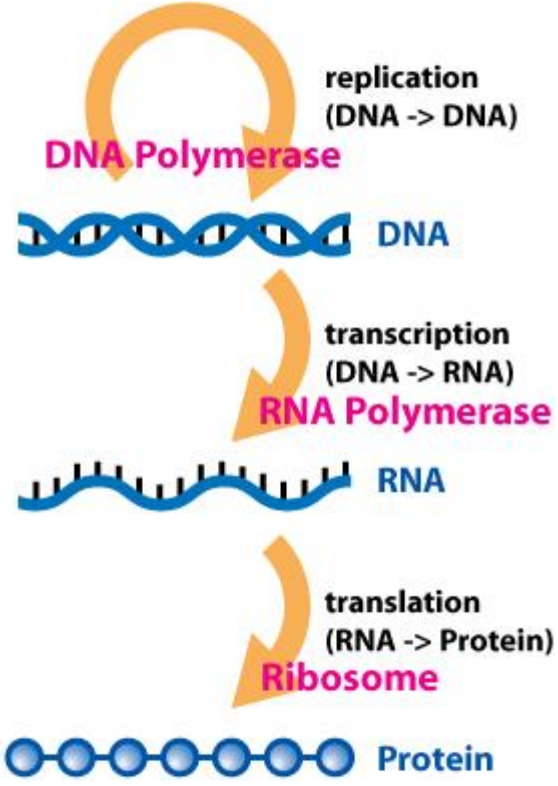
Nick Fitzkee
nfitzkee@chemistry.msstate.edu

* 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)

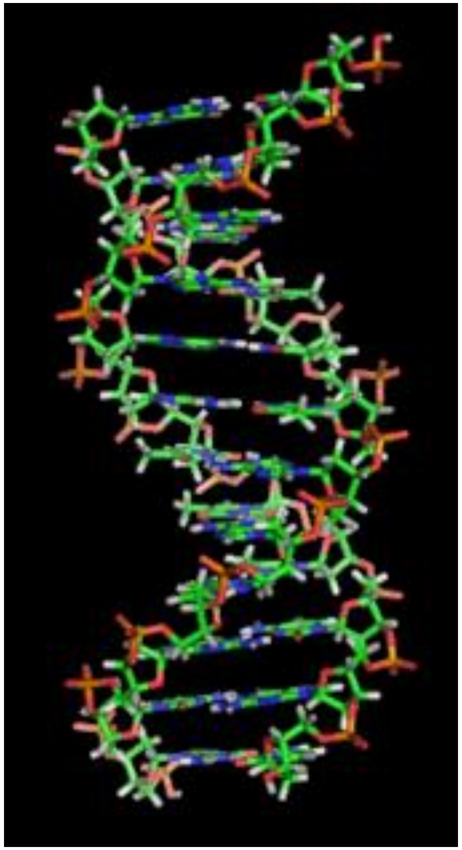
- Central Dogma:
  - DNA → RNA → Protein



The Central Dogma of Molecular Biology

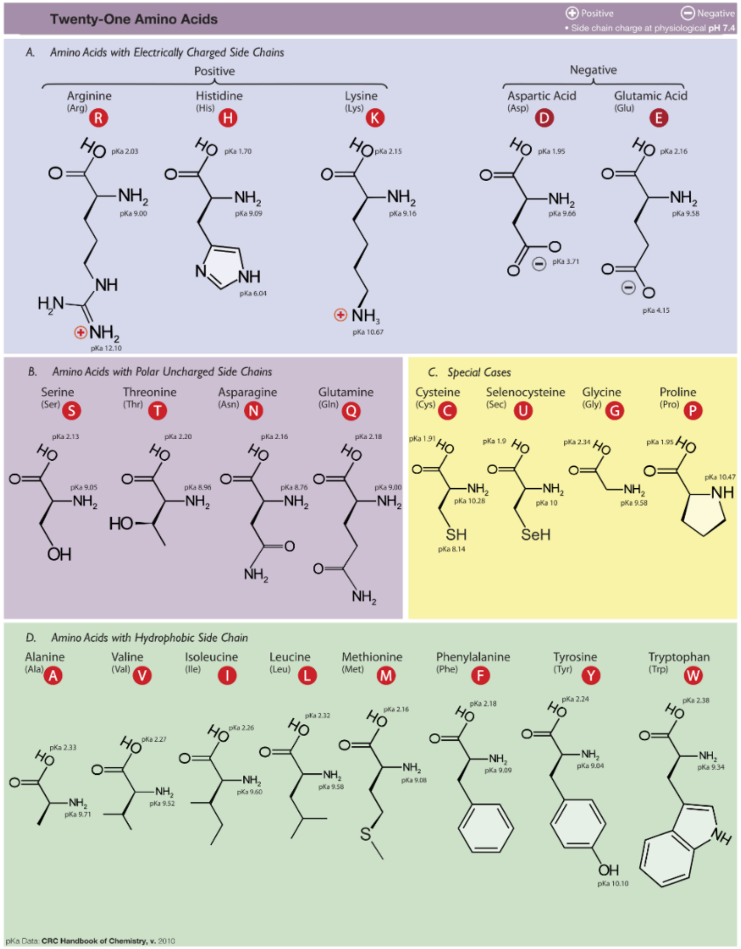
Primary Structure (Sequence)

- **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



pKa Data: **CRC Handbook of Chemistry, v.** 2010

Dan Cojocari, Department of Medical Biophysics, **University of Toronto** 2011

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

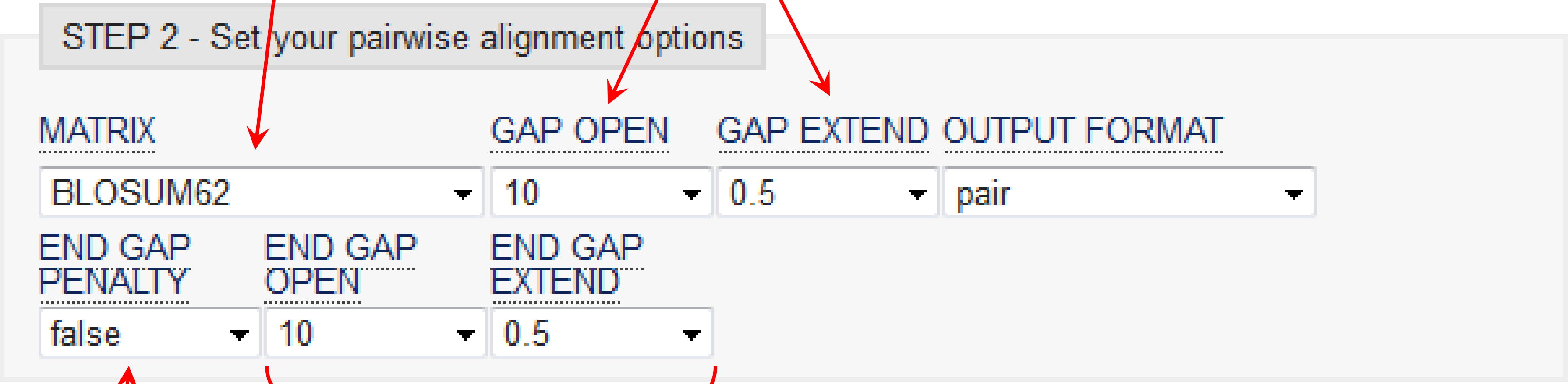
- EMBOSS Needle
  http://www.ebi.ac.uk/Tools/psa/emboss_needle/
  - Command line version also available

- 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:



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

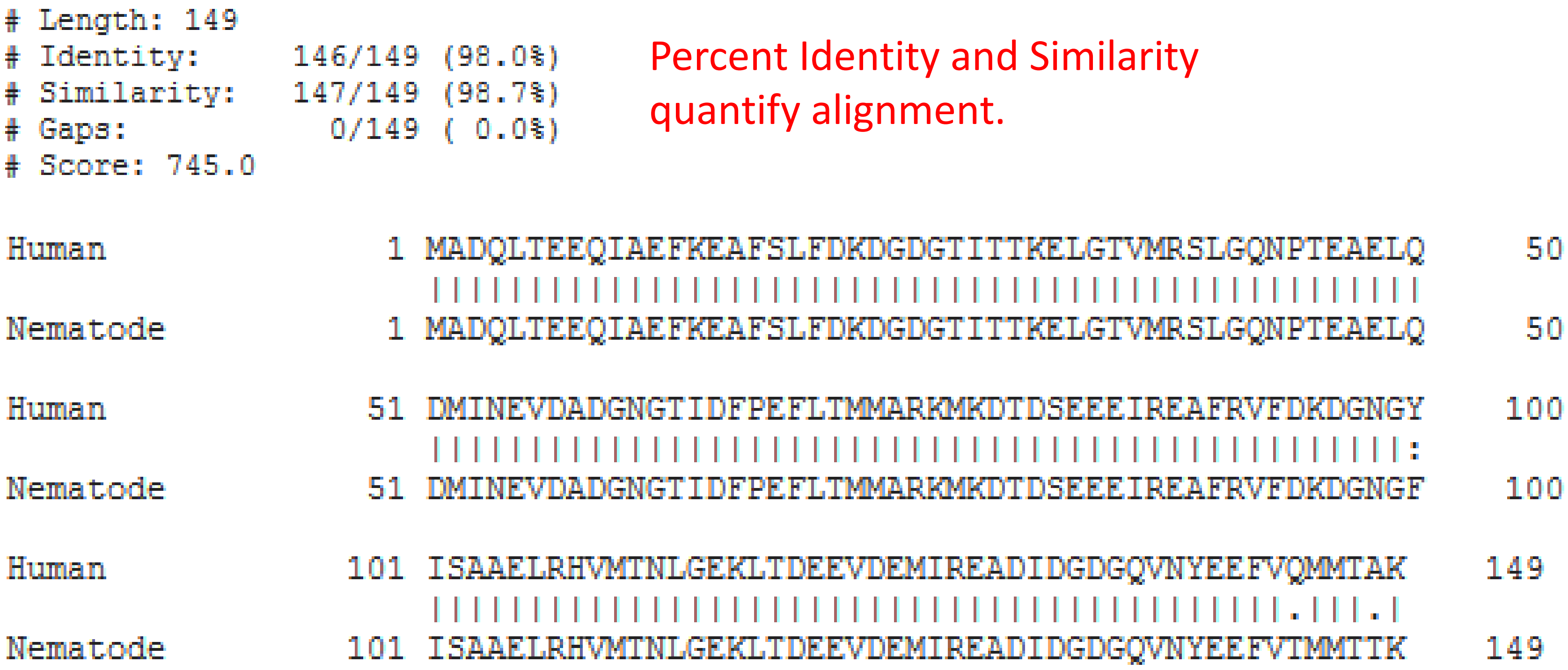
How to compare residues?

How much penalty to open a gap in the sequence?

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.

- Pretty darn similar!

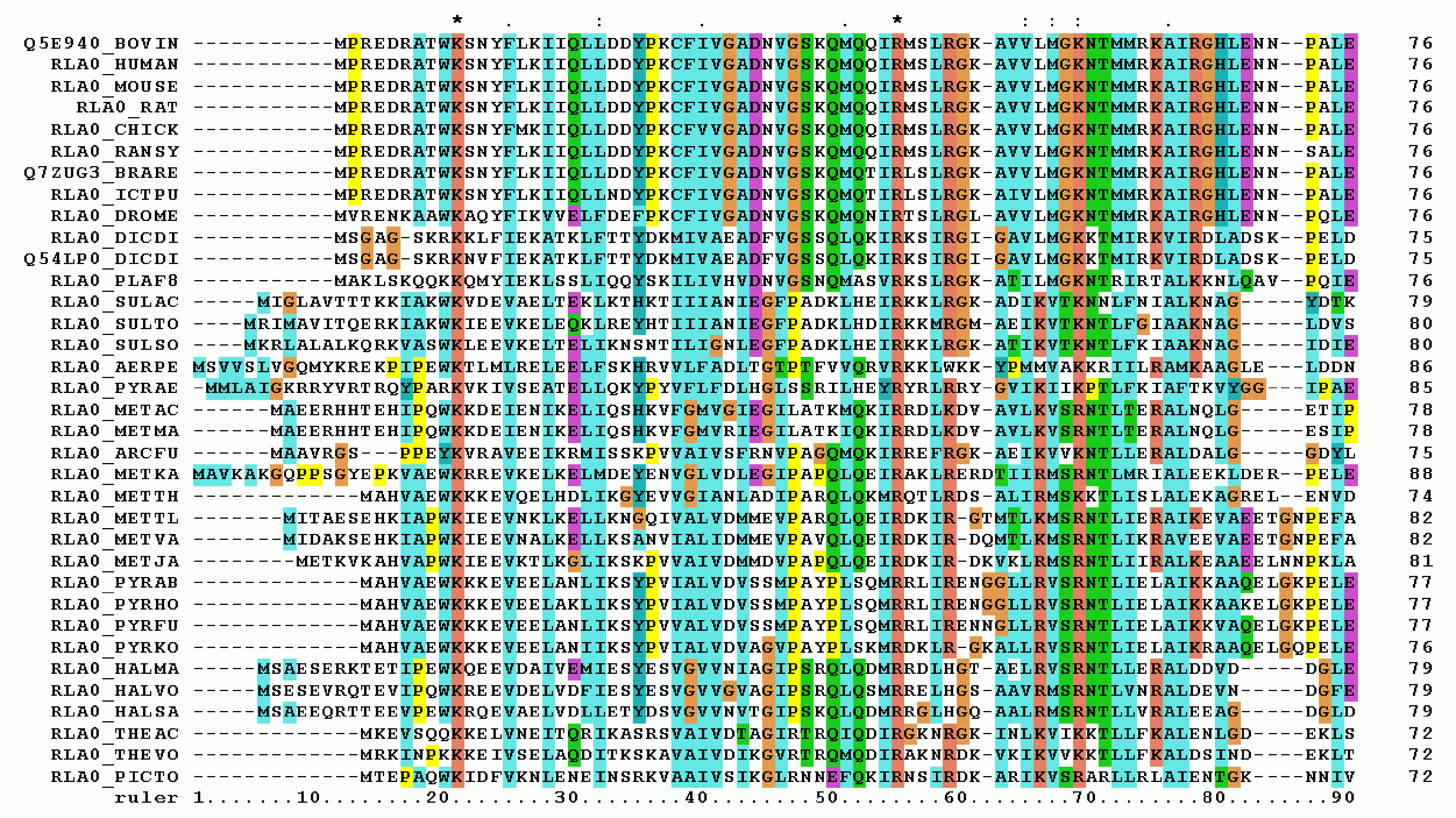
Identical residues shown with |, similar residues with : and ., and blanks represent dissimilar residues.

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

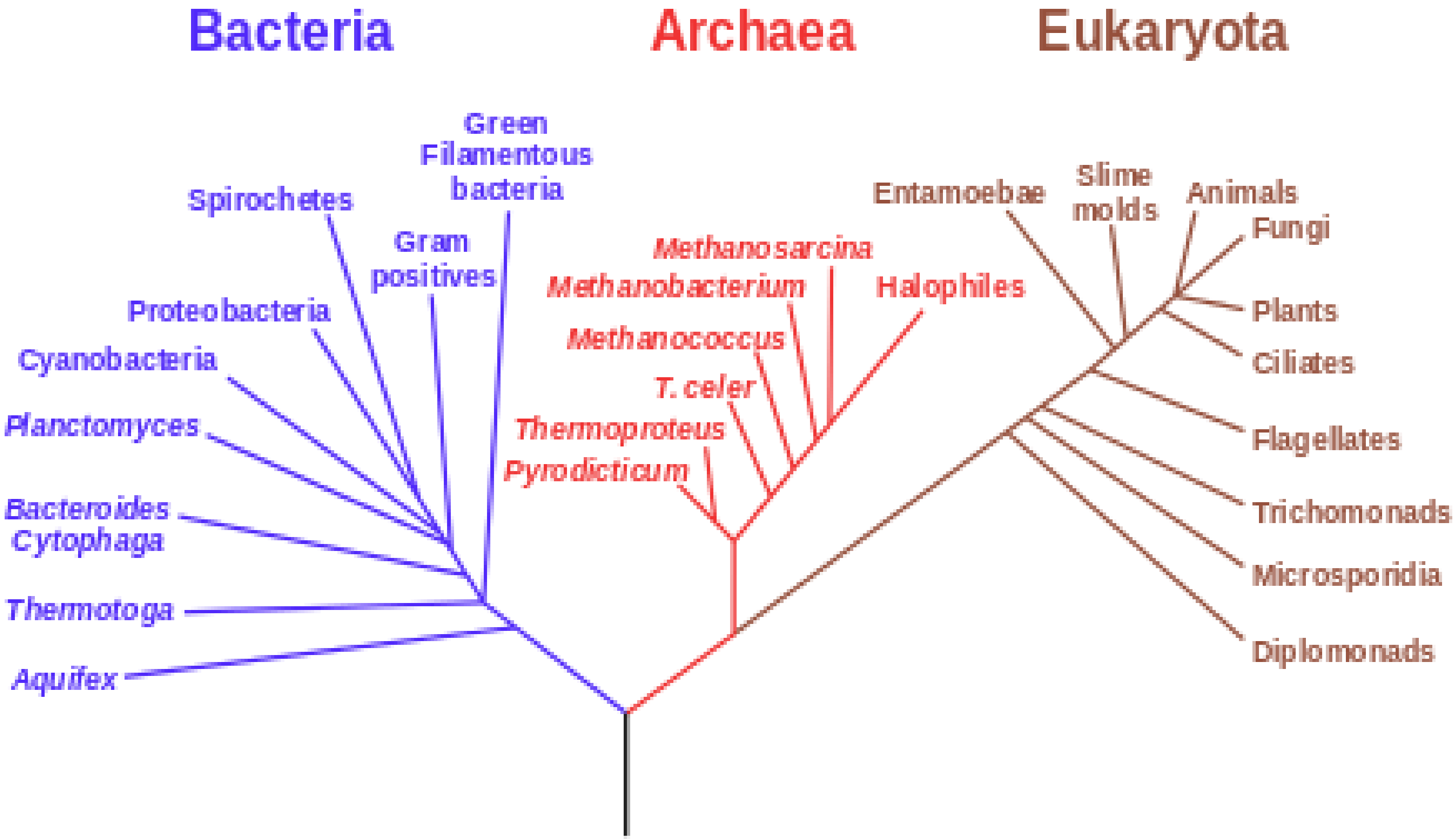
MSA Example



```
                                 *   .      :       .          .   *         : : :      .
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:
   1. If they’re the same length, are they almost the same sequence? (global alignment)
3. For 2 sequences:
   1. 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

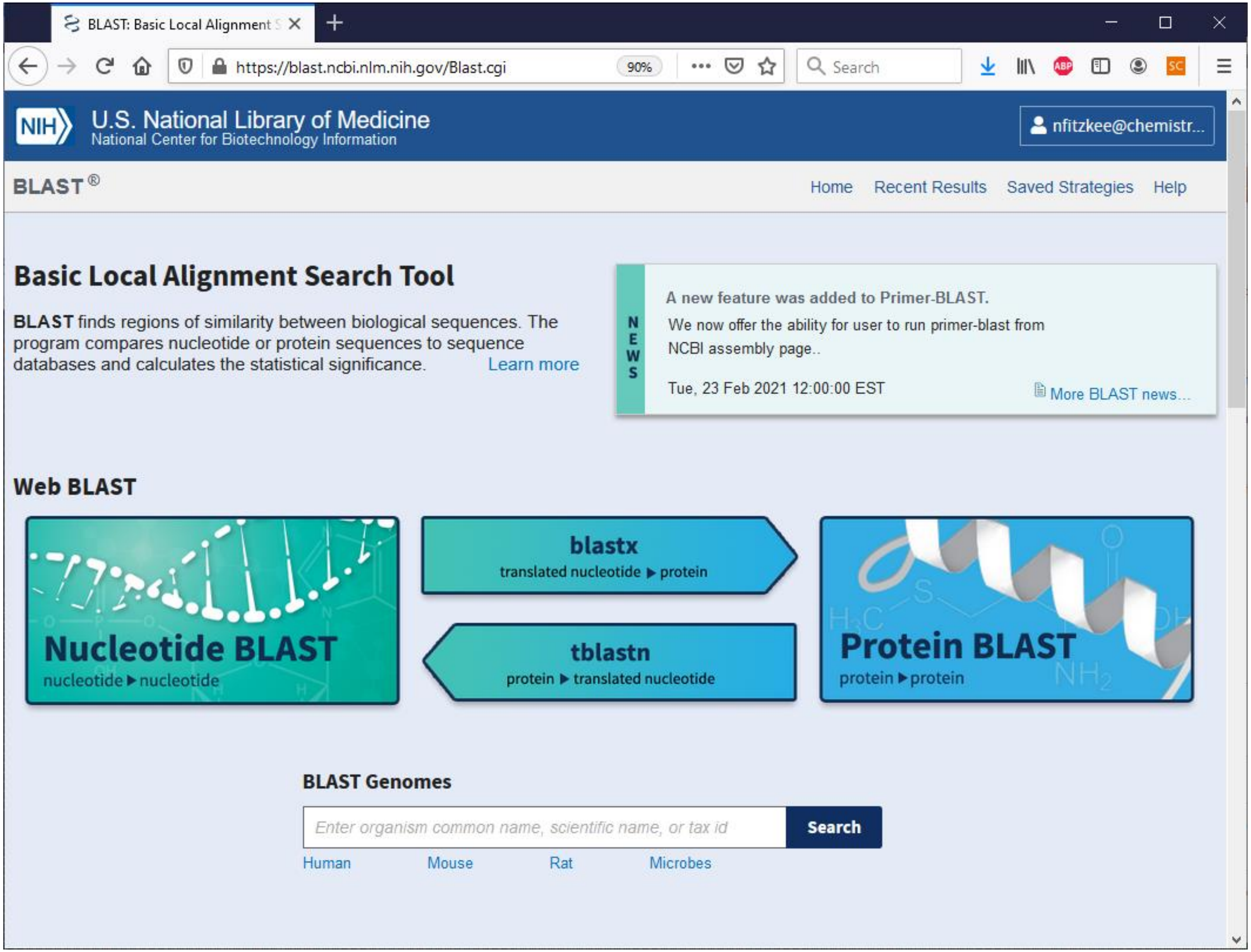
- **blastp** – searches a protein database using a protein query
  Protein sequence searched against a Protein database

- **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



BLAST: Basic Local Alignment S
https://blast.ncbi.nlm.nih.gov/Blast.cgi
90%
Search
U.S. National Library of Medicine
National Center for Biotechnology Information
nfitzkee@chemistr...
BLAST®
Home
Recent Results
Saved Strategies
Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

A new feature was added to Primer-BLAST.

We now offer the ability for user to run primer-blast from NCBI assembly page..

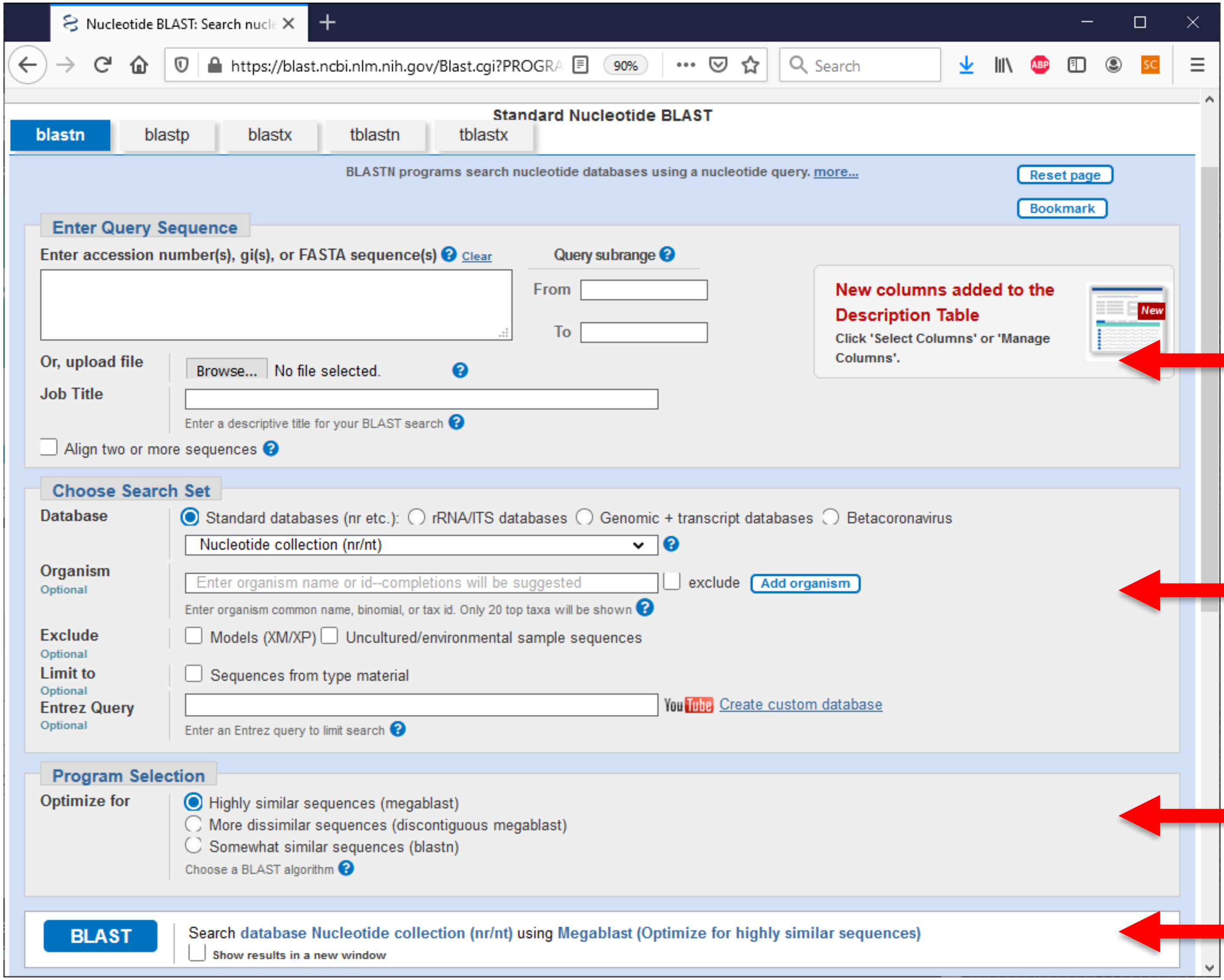
Tue, 23 Feb 2021 12:00:00 EST

More BLAST news...

Web BLAST

BLAST Genomes

Enter organism common name, scientific name, or tax id



Standard Nucleotide BLAST

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

Enter Query Sequence

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

Query subrange

From

To

New columns added to the Description Table

Click 'Select Columns' or 'Manage Columns'.

Or, upload file: 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

Nucleotide collection (nr/nt)

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

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

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

Limit to (Optional): □ Sequences from type material

Entrez Query (Optional): 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

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

□ Show results in a new window

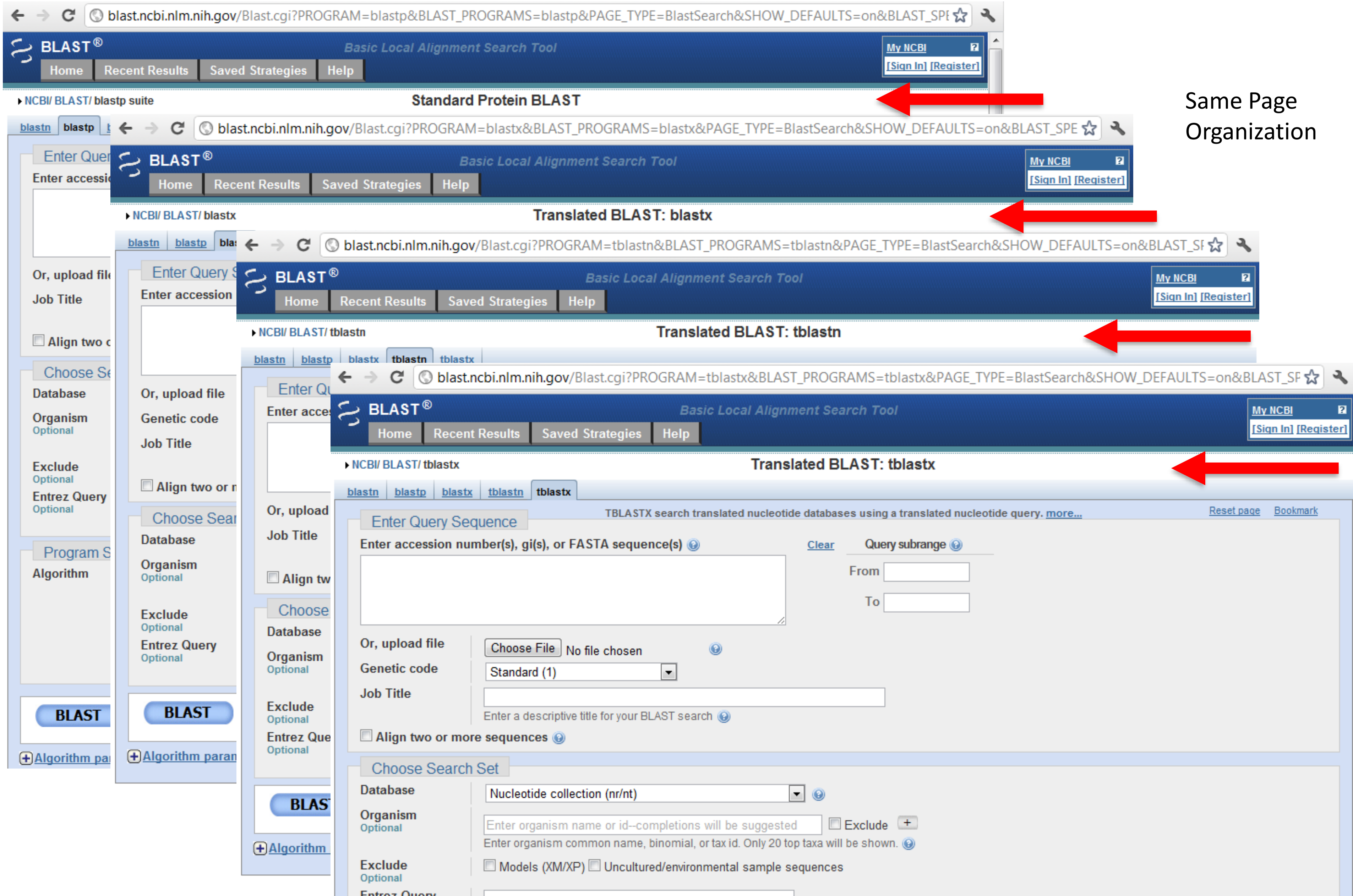
Sequence Input

Databases to Search Against

Program Selection

Click to Run!

Same Page Organization



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

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

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

NCBI/ BLAST/ blastx

Translated BLAST: blastx

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

NCBI/ BLAST/ tblastn

Translated BLAST: tblastn

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

NCBI/ BLAST/ tblastx

Translated BLAST: tblastx

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

Enter Query Sequence

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

Query subrange: From, To

Or, upload 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.

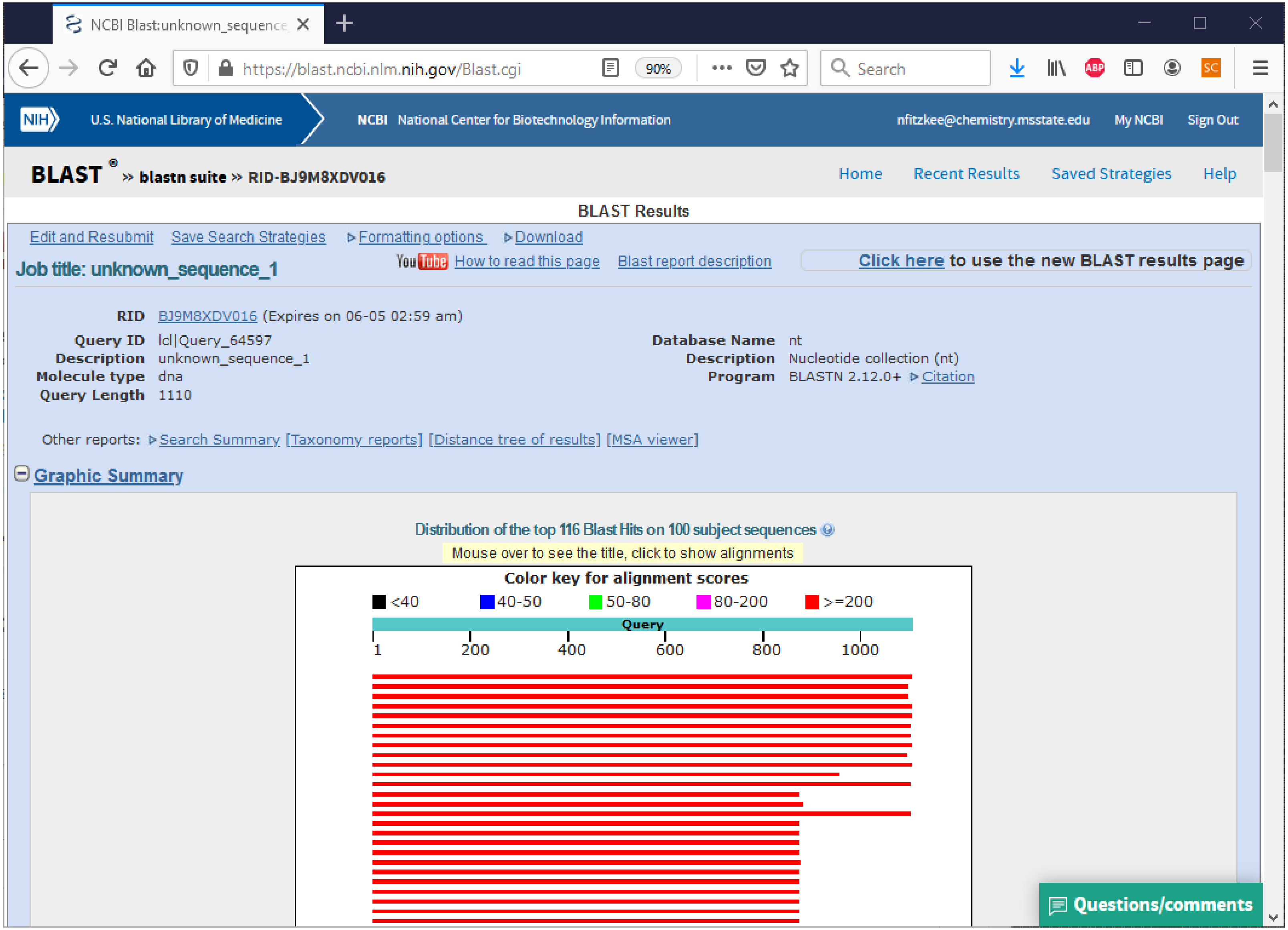
Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

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:unknown_sequence
https://blast.ncbi.nlm.nih.gov/Blast.cgi
90%
Search
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
nfitzkee@chemistry.msstate.edu
My NCBI
Sign Out
BLAST ® » blastn suite » RID-BJ9M8XDV016
Home
Recent Results
Saved Strategies
Help
Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to read this page
Blast report description
Click here to use the new BLAST results page
Questions/comments

BLAST Results

Job title: unknown_sequence_1

RID	BJ9M8XDV016 (Expires on 06-05 02:59 am)
Query ID	lcl|Query_64597
Description	unknown_sequence_1
Molecule type	dna
Query Length	1110

Database Name	nt
Description	Nucleotide collection (nt)
Program	BLASTN 2.12.0+ Citation

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

Graphic Summary

Distribution of the top 116 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

■ <40 🟦 40-50 🟩 50-80 ■ 80-200 🟥 >=200

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

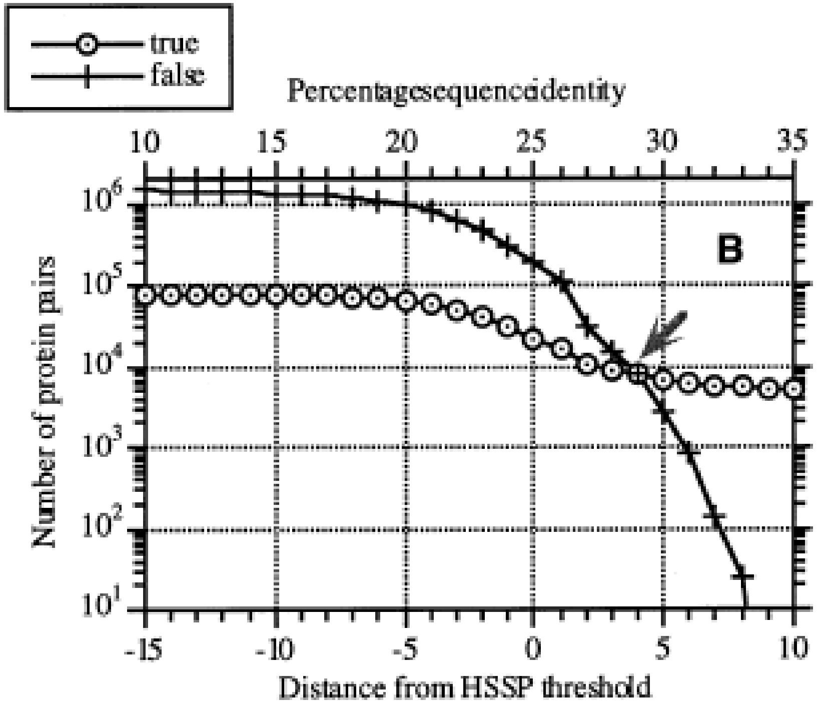
- So far we've focused on *sequence* alignment: looking at the primary (DNA or protein) sequence

- 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 sequences tend to have similar structures.
- 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 dissimilar aligned pairs explodes.

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

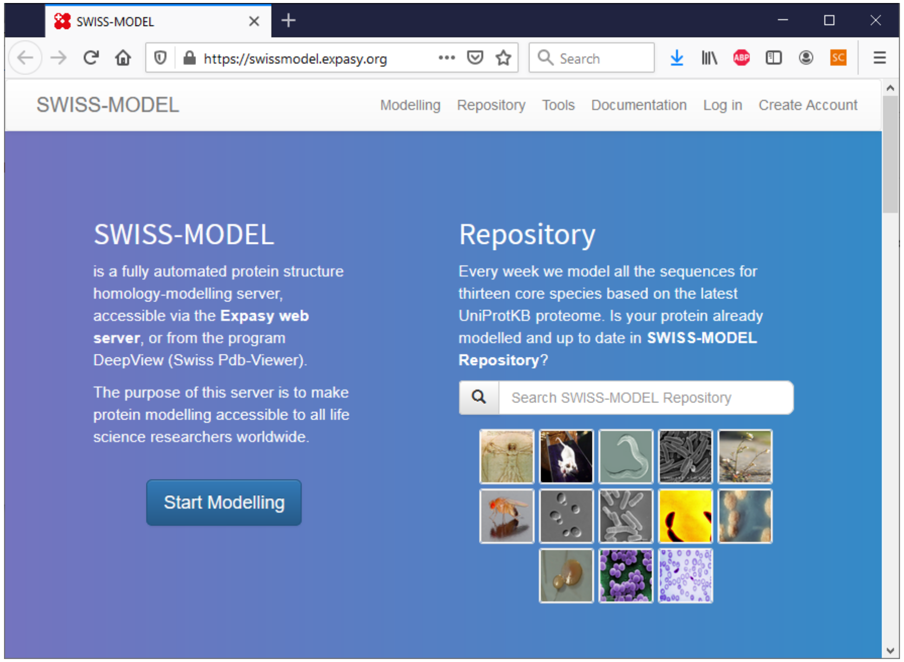
What is Homology Modeling?

- **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**



SWISS-MODEL

is a fully automated protein structure homology-modelling server, accessible via the **Expasy web server**, or from the program DeepView (Swiss Pdb-Viewer).

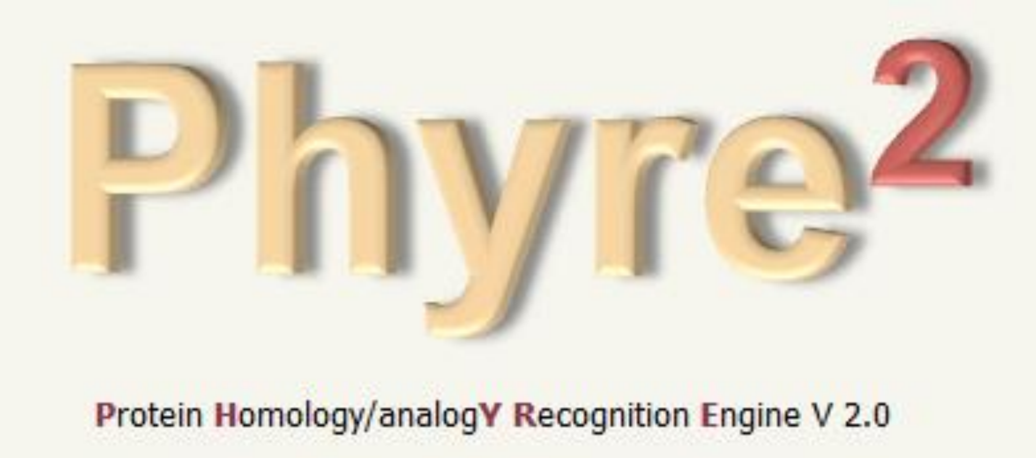
The purpose of this server is to make protein modelling accessible to all life science researchers worldwide.

Repository

Every week we model all the sequences for thirteen core species based on the latest UniProtKB proteome. Is your protein already modelled and up to date in **SWISS-MODEL Repository**?

- 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**

I-TASSER

Protein Structure & Function Predictions

(The server completed predictions for 621243 proteins submitted by 149610 users from 158 countries or regions)

(The template library was updated on 2021/05/23)

- 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

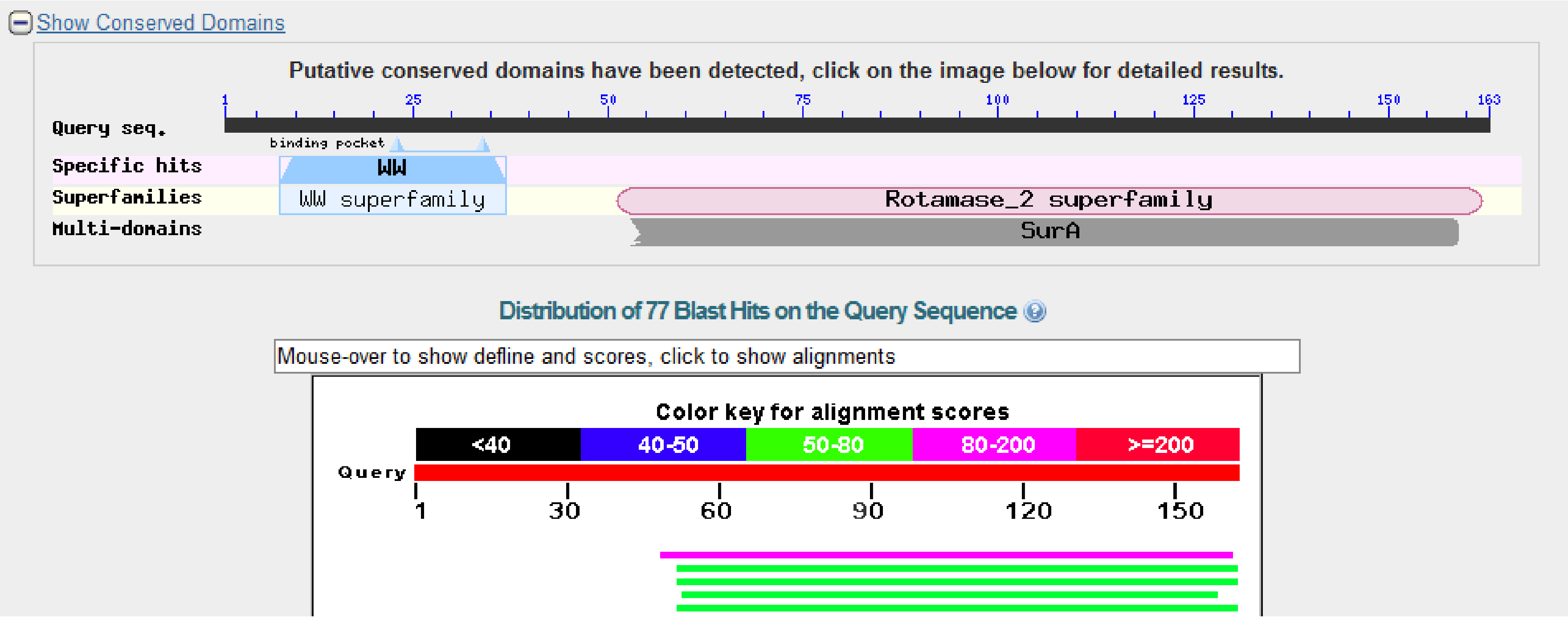
- Sequence for Pin1 protein:

```
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:



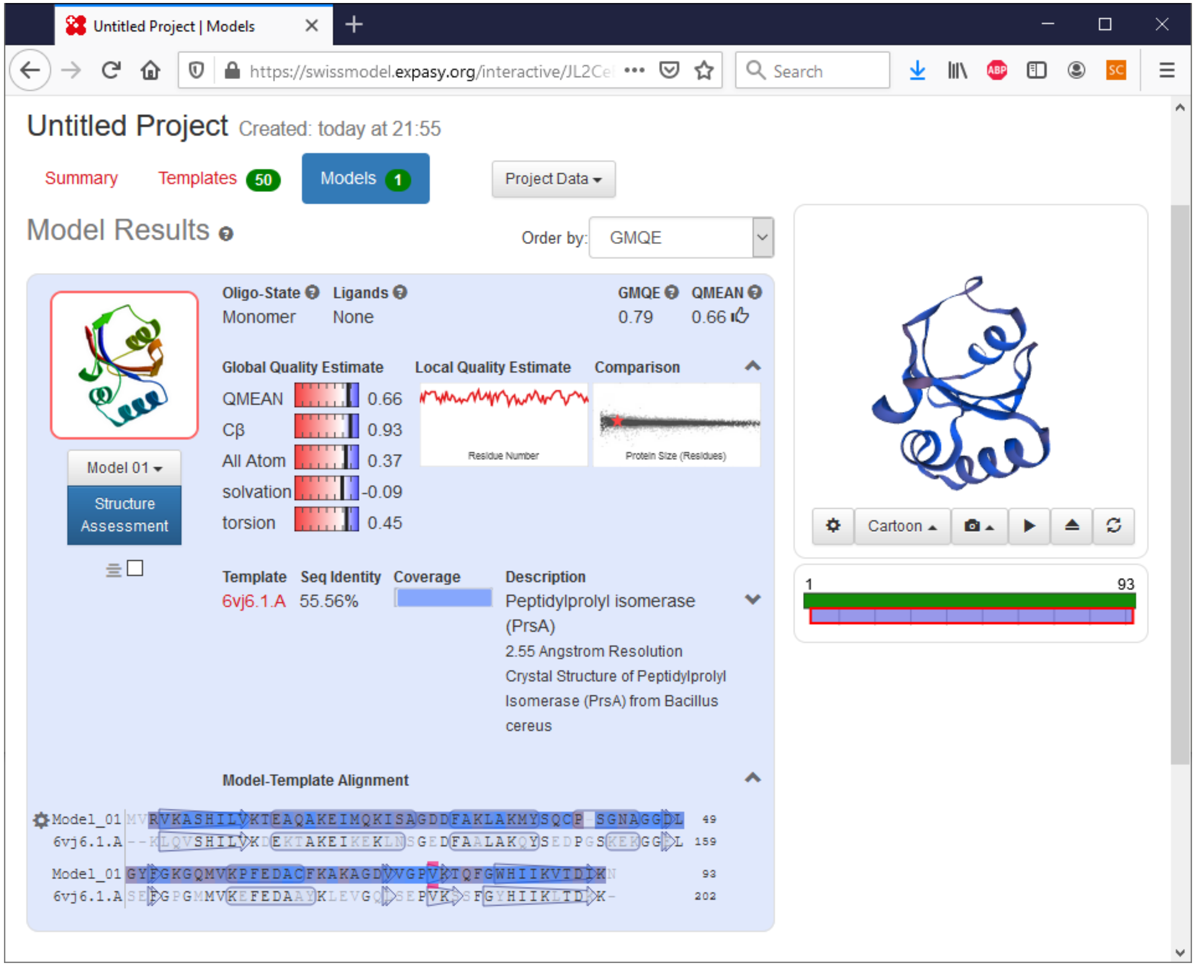
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
1
25
50
75
100
125
150
163
Query seq.
binding pocket
Specific hits
WW
Superfamilies
WW superfamily
Rotamase_2 superfamily
Multi-domains
SurA
Distribution of 77 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
30
60
90
120
150

- Sequence (only second domain found):

MVRVKASHIL VKTEAQAKEI MQKISAGDDF AKLAKMYSQC PSGNAGGDLG

YFGKGQMVKP FEDACFKAKA GDVVGPVKTQ FGWHIIKVTD IKN

Result: SWISS-MODEL



Untitled Project | Models
https://swissmodel.expasy.org/interactive/JL2Ce
Search

Untitled Project Created: today at 21:55

Model Results

Order by: GMQE

Oligo-State	Ligands	GMQE	QMEAN
Monomer	None	0.79	0.66

Global Quality Estimate

QMEAN	0.66
Cβ	0.93
All Atom	0.37
solvation	-0.09
torsion	0.45

Local Quality Estimate

Comparison

Template	Seq Identity	Coverage	Description
6vj6.1.A	55.56%		Peptidylprolyl isomerase (PrsA) 2.55 Angstrom Resolution Crystal Structure of Peptidylprolyl Isomerase (PrsA) from Bacillus cereus

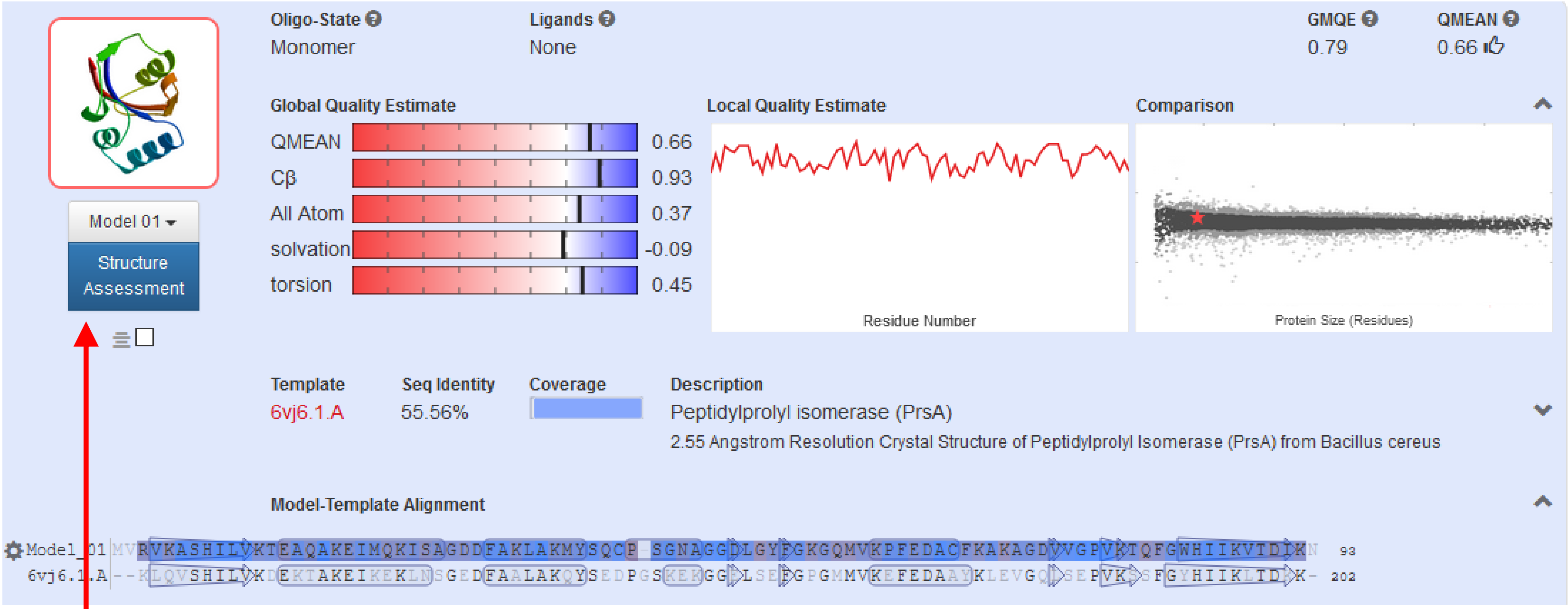
Model-Template Alignment

```
Model_01  MVRVKASHILVKTEAQAKEIMQKISAGDDFAKLAKMYSQCP-SGNAGGDL  49
6vj6.1.A  --KLQVSHILVKDEKTAKEIKEKLNSGEDFAALAKQYSEDPGSKEKGGEL  159

Model_01  GYFGKGQMVKPFEDACFKAKAGDVVGPVKTQFGWHIIKVTDIKN  93
6vj6.1.A  SEFGPGMMVKEFEDAAYKLEVGQISEPVKSSFGYHIIKLTDIK-  202
```

- We'll do this model in class

Result: SWISS-MODEL



Oligo-State: Monomer

Ligands: None

GMQE: 0.79

QMEAN: 0.66

Model 01

Structure Assessment

Global Quality Estimate

QMEAN	0.66
Cβ	0.93
All Atom	0.37
solvation	-0.09
torsion	0.45

Local Quality Estimate

Comparison

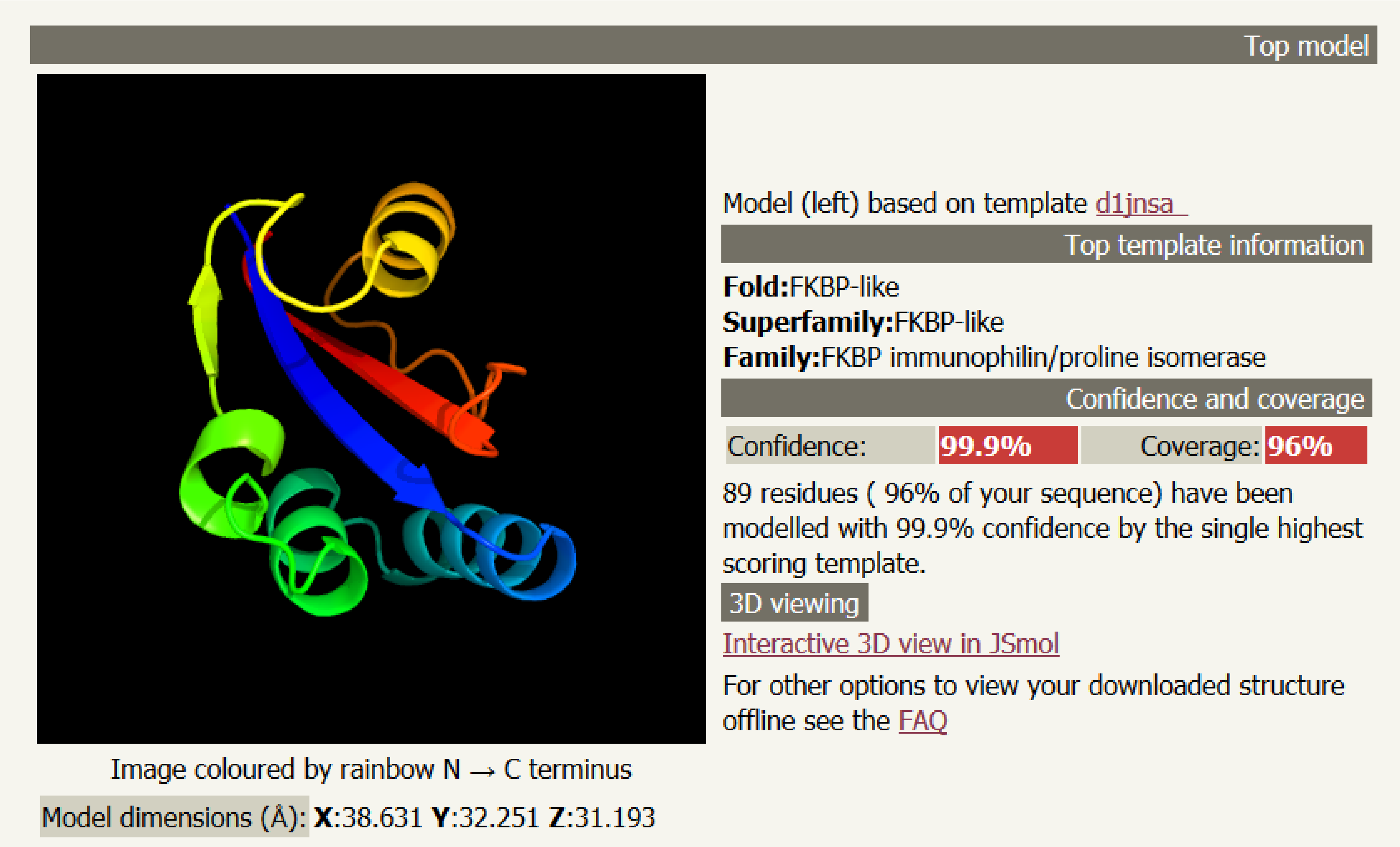
Template	Seq Identity	Coverage	Description
6vj6.1.A	55.56%		Peptidylprolyl isomerase (PrsA) 2.55 Angstrom Resolution Crystal Structure of Peptidylprolyl Isomerase (PrsA) from Bacillus cereus

Model-Template Alignment

```
Model_01 MVRVKASHILVKTEAQAKEIMQKISAGDDFAKLAKMYSQCP-SGNAGGDLGYFGKGQMVKPFEDACFKAKAGDVVGPVKTQFGWHIIKVTDIKN  93
6vj6.1.A --KLQVSHILVKDEKTAKEIKEKLNSGEDFAALAKQYSEDPGSKEKGGELSEFGPGMMVKEFEDAAYKLEVGQLSEPVKSSFGYHIIKLTDKK- 202
```

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

Result: Phyre^2



Top model

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

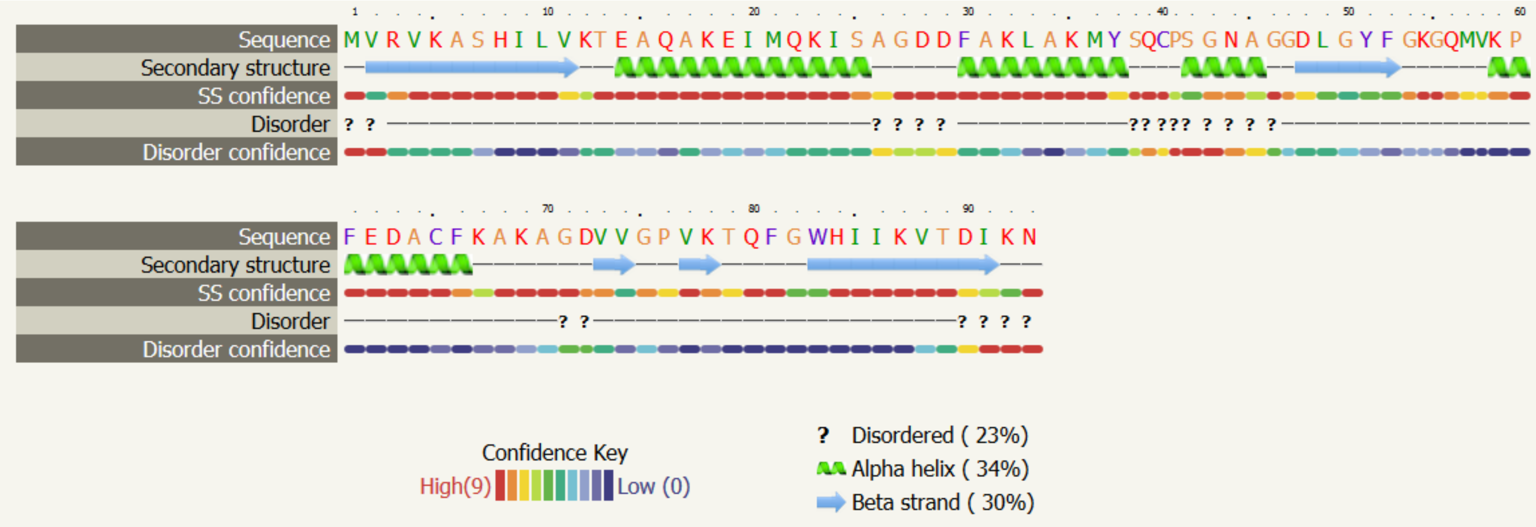
Interactive 3D view in JSmol

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

Image coloured by rainbow N → C terminus

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

Result: Phyre^2



```
1         10        20        30        40        50        60
Sequence              MVRVKASHILVKTEAQAKEIMQKISAGDDFAKLAKMYSQCPSGNAGGDLGYFGKGQMVKP
Secondary structure
SS confidence
Disorder              ??                        ????              ?????????
Disorder confidence

                    70        80        90
Sequence              FEDACFKAKAGDVVGPVKTQFGWHIIKVTDIKN
Secondary structure
SS confidence
Disorder                        ??                  ????
Disorder confidence
```

Confidence Key

High(9) Low (0)

? Disordered (23%)

Alpha helix (34%)

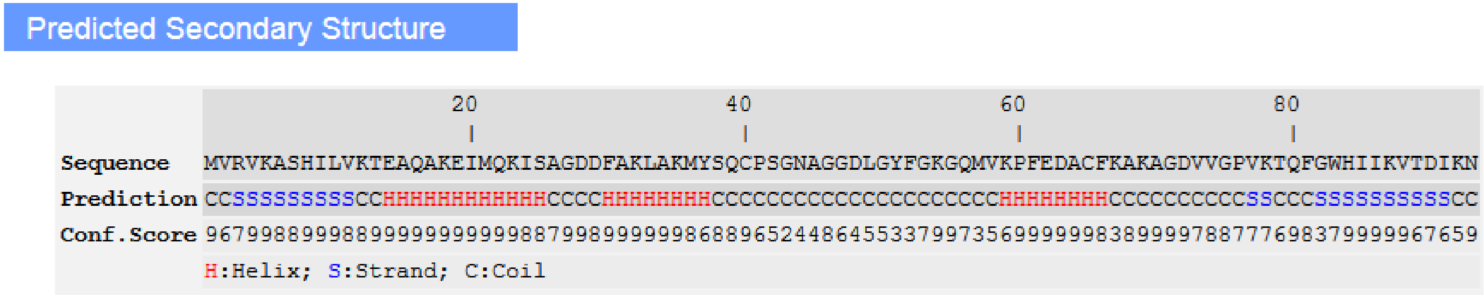
Beta strand (30%)

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

http://folding.chemistry.msstate.edu/files/bootcamp/phyre2/summary.html

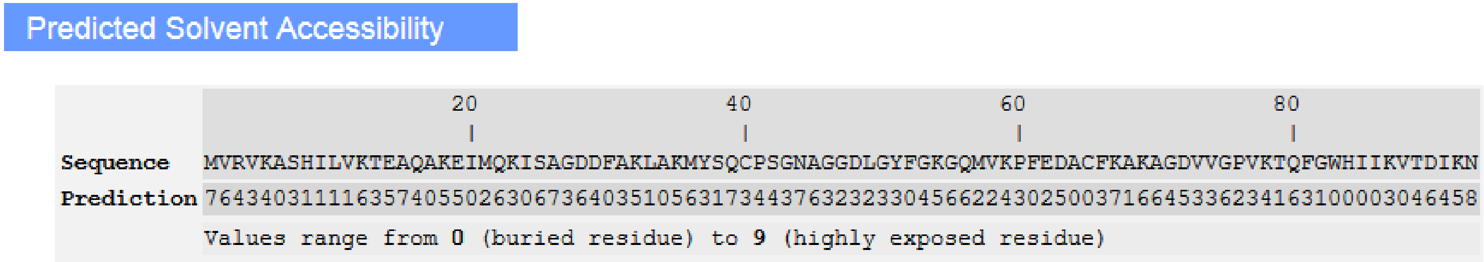
- Final result is called `final.casp.pdb`

Result: I-TASSER



Predicted Secondary Structure

```
                            20                  40                  60                  80
                             |                   |                   |                   |
Sequence   MVRVKASHILVKTEAQAKEIMQKISAGDDFAKLAKMYSQCPSGNAGGDLGYFGKGQMVKPFEDACFKAKAGDVVGPVKTQFGWHIIKVTDIKN
Prediction CCSSSSSSSSSCCHHHHHHHHHHHHCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCCHHHHHHHHCCCCCCCCCCSSCCCSSSSSSSSSSCC
Conf.Score 967998899988999999999988799899999998688965244864553379973569999998389999788777698379999967659
           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:

  http://folding.chemistry.msstate.edu/files/bootcamp/itasser/

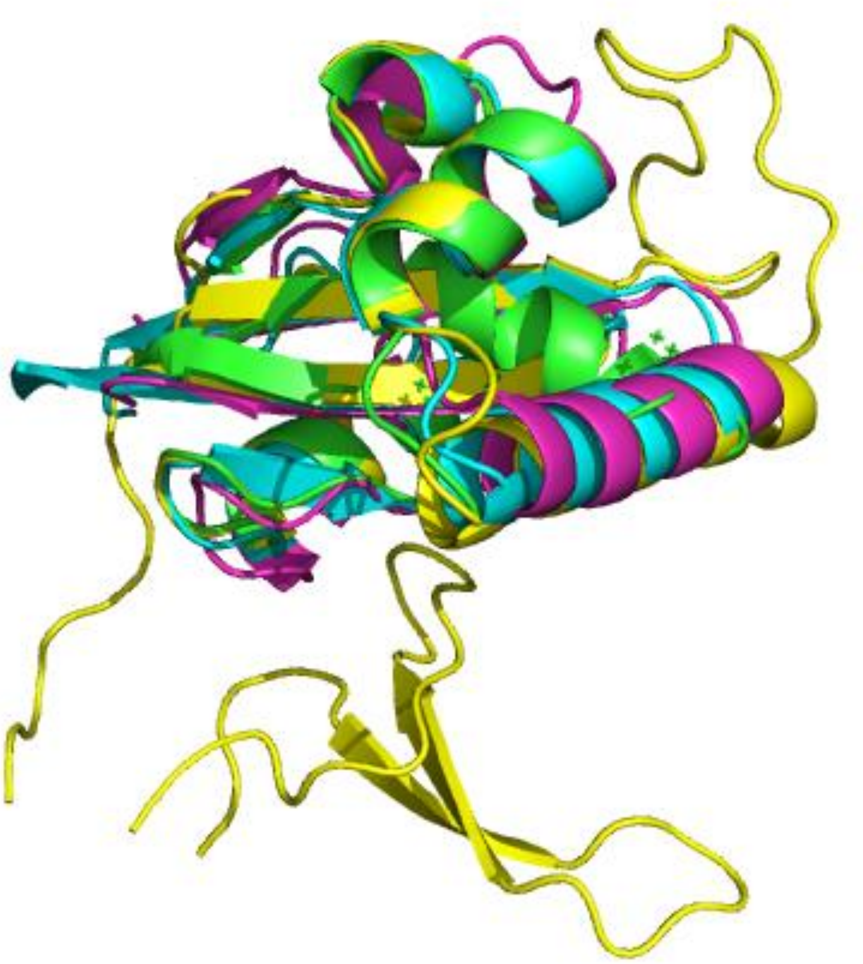
- Final result is called `model1.pdb`

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

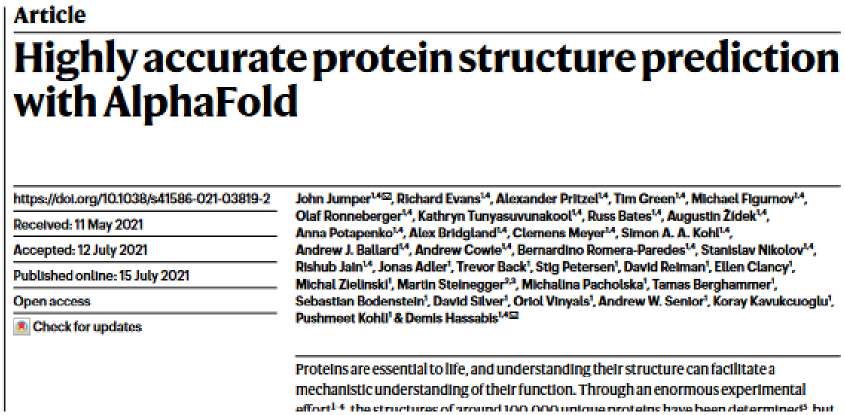


- Colors:
  - Original Pin1
  - SWISS-MODEL
  - Phyre^2
  - I-TASSER

- **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

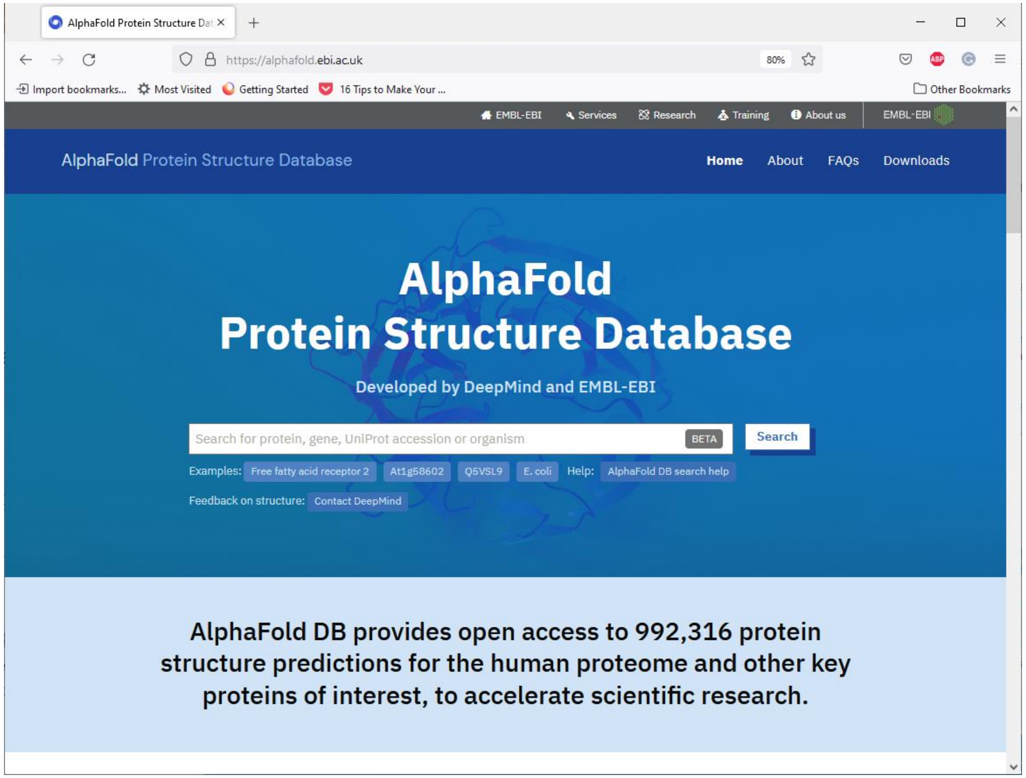
- Not really homology modeling, not really “ab initio” or physics-based

- 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/



AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA Search

Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli Help: AlphaFold DB search help

Feedback on structure: Contact DeepMind

AlphaFold DB provides open access to 992,316 protein structure predictions for the human proteome and other key proteins of interest, to accelerate scientific research.

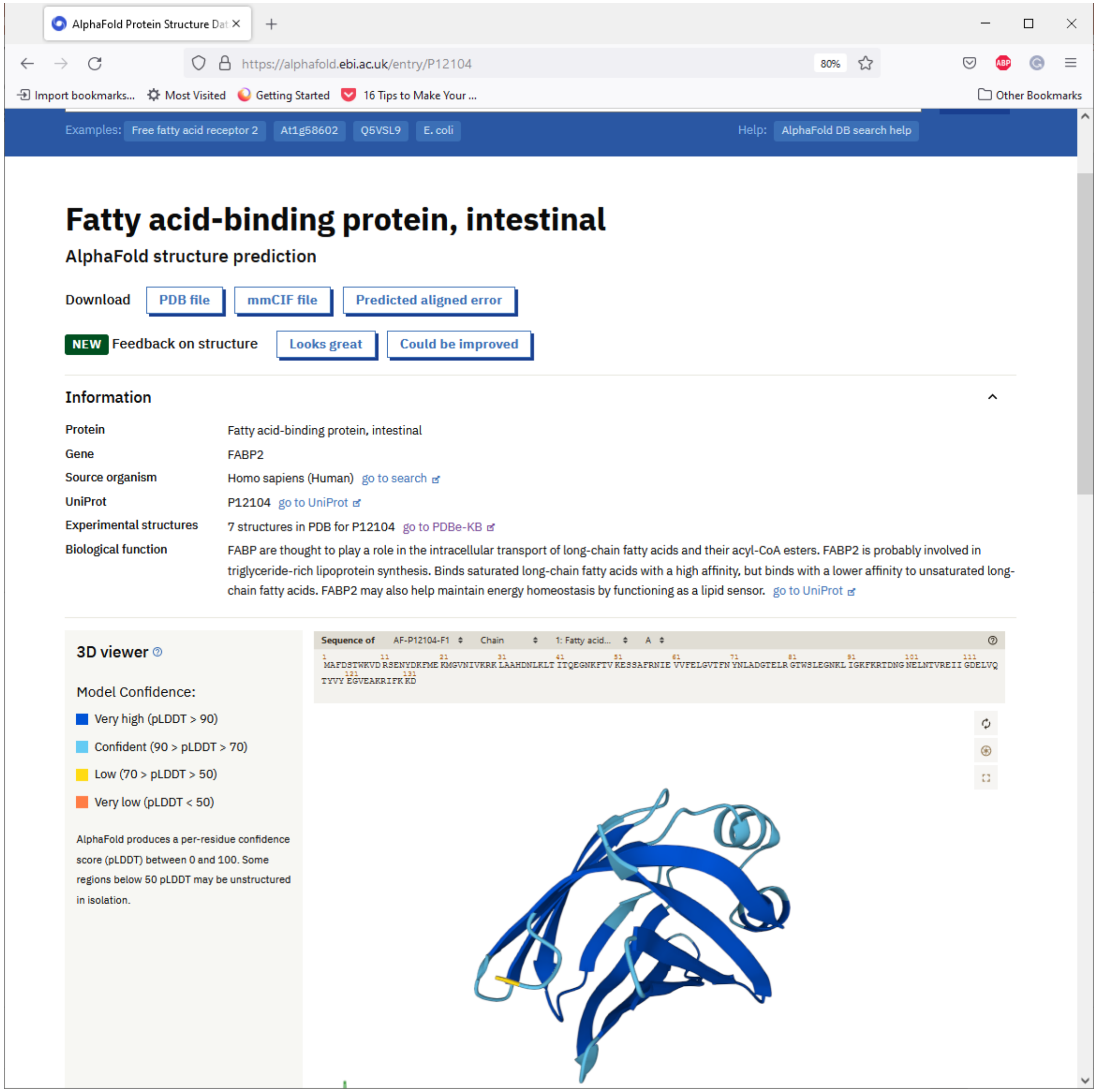
- Entry: P12104 (Human Intestinal Fatty Acid Binding Protein)

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



Fatty acid-binding protein, intestinal

AlphaFold structure prediction

Download PDB file | mmCIF file | Predicted aligned error

NEW Feedback on structure Looks great | Could be improved

Information

Protein	Fatty acid-binding protein, intestinal
Gene	FABP2
Source organism	Homo sapiens (Human) go to search
UniProt	P12104 go to UniProt
Experimental structures	7 structures in PDB for P12104 go to PDBe-KB
Biological function	FABP are thought to play a role in the intracellular transport of long-chain fatty acids and their acyl-CoA esters. FABP2 is probably involved in triglyceride-rich lipoprotein synthesis. Binds saturated long-chain fatty acids with a high affinity, but binds with a lower affinity to unsaturated long-chain fatty acids. FABP2 may also help maintain energy homeostasis by functioning as a lipid sensor. go to UniProt

3D viewer

Model Confidence:

- 🟦 Very high (pLDDT > 90)
- 🟦 Confident (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

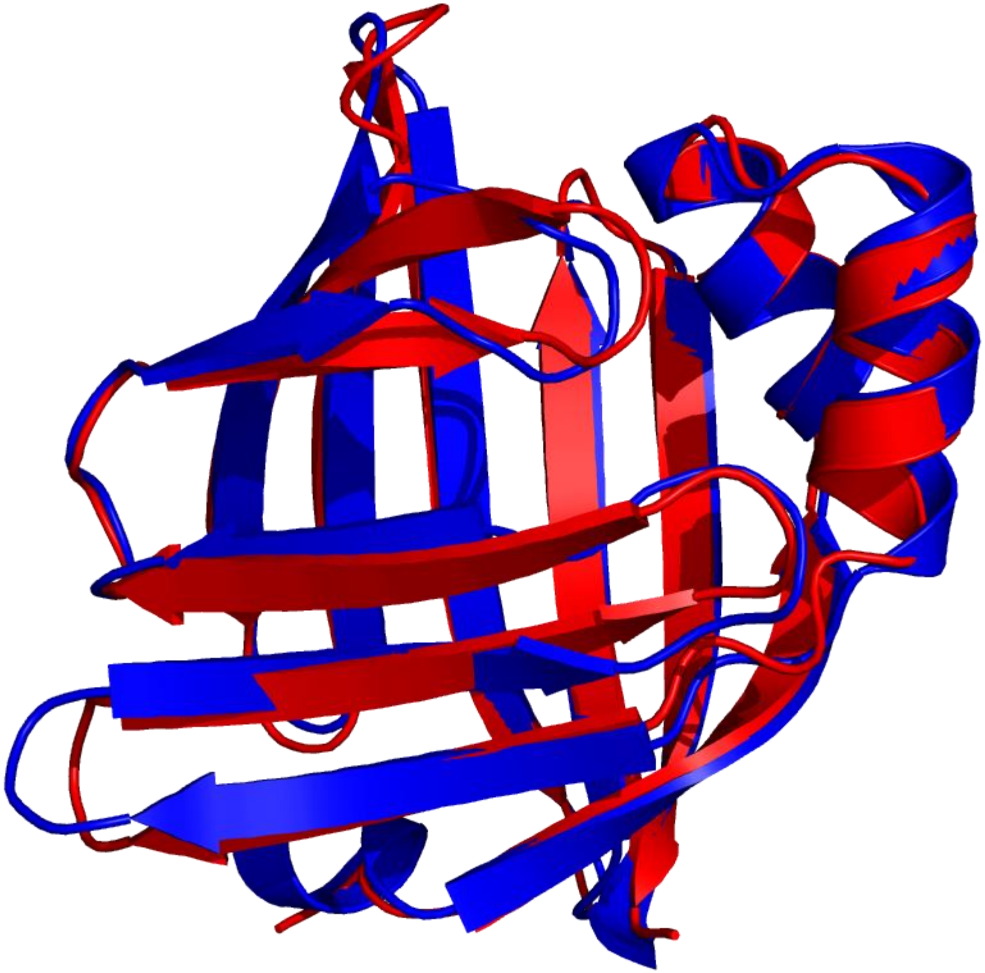
Sequence of AF-P12104-F1 Chain 1: Fatty acid... A

```
1          11         21         31         41         51         61         71         81         91         101        111
MAFDSTWKVD RSENYDKFME KMGVNIVKRK LAAHDNLKLT ITQEGNKFTV KESSAFRNIE VVFELGVTFN YNLADGTELR GTWSLEGNKL IGKFKRTDNG NELNTVREII GDELVQ
     121        131
TYVY EGVEAKRIFK KD
```

Comparison of AlphaFold2 vs 6L9O

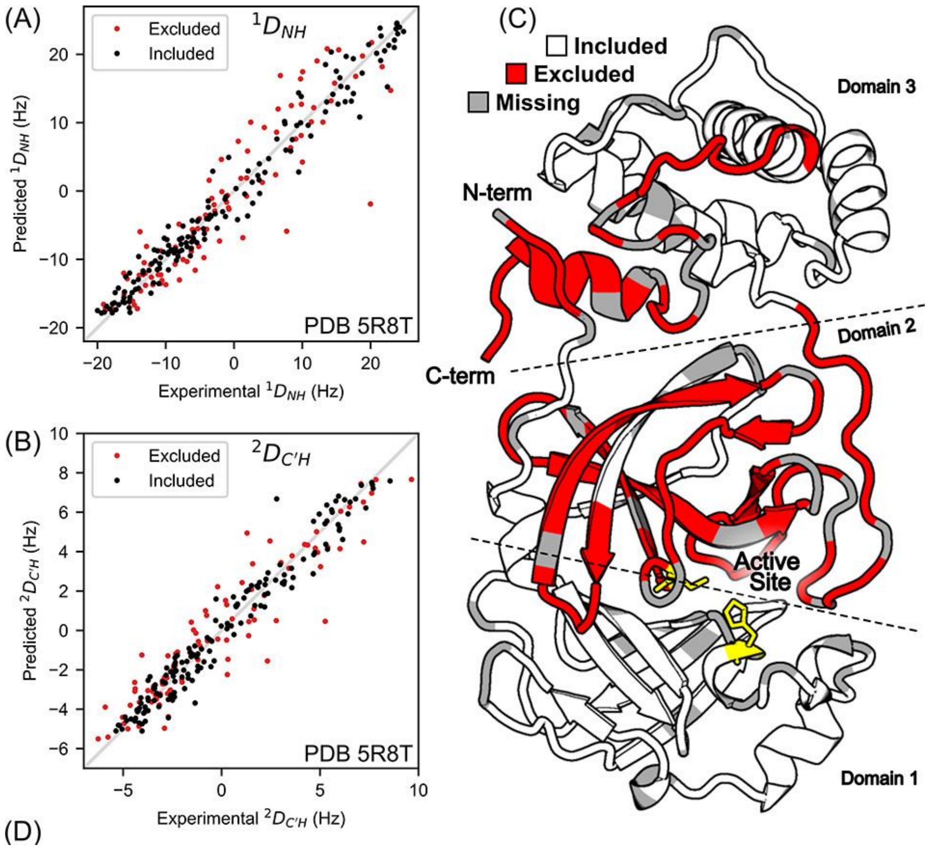
- **Red:** AlphaFold2
- **Blue:** Experimental crystal structure

- 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!