

Basic Bioinformatics, Sequence Alignment, and Homology

Biochemistry Boot Camp 2021

Session #10

Nick Fitzkee

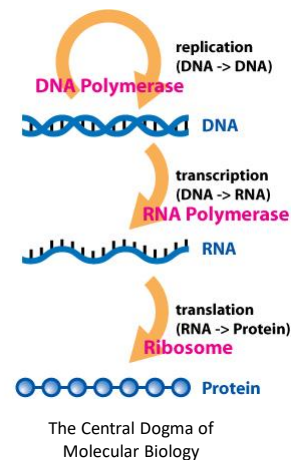
nfitzkee@chemistry.msstate.edu

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

1

Biology Review

- Genome is the genetic material of an organism, normally DNA but RNA possible (viruses)
- Central Dogma:
 - DNA → RNA → Protein



2

2

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'

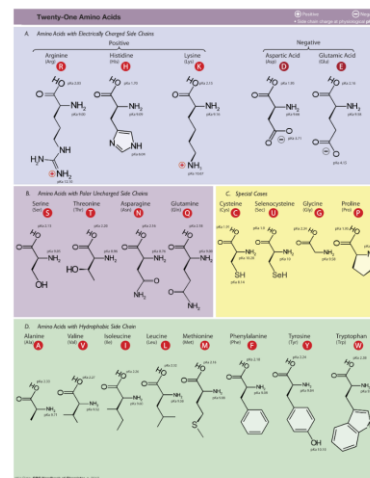


3

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



4

4

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.

5

5

Storing Sequences

- | | |
|---|--|
| <ul style="list-style-type: none"> • FASTA format • Can represent nucleotide sequences or peptide sequences using single letter codes | <ul style="list-style-type: none"> • FASTQ format • Represents nucleotide sequences and their corresponding quality scores |
|---|--|

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus]
LCLYTHIGENNIYGSVLYSETWVQIMLLITMATFNGVYLVWQMSFWGATVITWLFSAIPPYICINLV
EWINGQFSVDKATINRFAPHFILPFWALAGVHLFLHETGSSNNPLGLTSDGDKIPFHPYTIKDFLG
LLILILLLLLLALLSDFMLGDFONHMDADPLNTEHLIKEDWYFLFAYAILRSVKNLGGVLAFLSIVIL
GLMPFLHTSKHRSNMLRPLSQALEWLTMDLLTLTWIGSQFVEYPTTIIGQMASILYFSIILAFLPIAGX
IENV
```

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

6

6

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])

7

7

Sequence Alignment

Sequence 1: GARVEY

Sequence 2: AVERY

Global Alignment:

GARVE-Y

-A-VERY

8

8

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)

9

9

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?

10

10

Global Sequence Alignment

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

Percent Identity and Similarity quantify alignment.

Human	1	MADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAELQ	50
Nematode	1	MADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAELQ	50
Human	51	DMINEVDADGNGTIDFPEFLTMMARKMKDIDSEEEIREAFRVFDKDGNGY	100
Nematode	51	DMINEVDADGNGTIDFPEFLTMMARKMKDIDSEEEIREAFRVFDKDGNGF	100
Human	101	ISAAELRHVMTNLGEKLTDEEVDEMIREADIDGGQVNYEEFVQMMTAK	149
Nematode	101	ISAAELRHVMTNLGEKLTDEEVDEMIREADIDGGQVNYEEFVIMMITK	149

- Pretty darn similar!

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

11

11

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, ClustalX
 - <http://www.megasoftware.net/>
 - <http://www.clustal.org/>



12

12

MSA Example

```

Q5E940 BOVIN -----MREDRATWKSNYFLKTIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 HUMAN -----MREDRATWKSNYFLKTIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 MOUSE -----MREDRATWKSNYFLKTIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 RAT -----MREDRATWKSNYFLKTIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 CHICK -----MREDRATWKSNYFMKTIQLDDYPKCFVVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 RANSY -----MREDRATWKSNYFLKTIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--SALE 76
Q7ZUG3 BRARE -----MREDRATWKSNYFLKTIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 ICTPU -----MREDRATWKSNYFLKTIQLDDYPKCFIVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 DROME -----MVRENKRAAKAOYEIKVVELEDFPKCFIVGADNVGSKMQQIRMSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 DICDI -----MSGAG--SKRKKLFTEKATKLFTTDMKIVAEADFVGSLSLQKIRKSIRGI--GAVLMGKNTMMRKAIRGHLENN--PALE 75
Q54LP0 DICDI -----MSGAG--SKRKNVFTEKATKLFTTDMKIVAEADFVGSLSLQKIRKSIRGI--GAVLMGKNTMMRKAIRGHLENN--PALE 75
RLA0 PLAF8 -----MAKLSQKKQKQMYTEKLSSLIQQSKILIVHYDNGVSNMASVRKSLRGK--AVVLMGKNTMMRKAIRGHLENN--PALE 76
RLA0 SULAC -----HICLAVITTTKTIKKRVDEVAELTEKLTETILLIANISGFPADKLHEIRKKLRGK--ADIKVTNNLNFNIALKNAG--VDIK 79
RLA0 SULFO -----MKIMAVITQERKIAKKWLEKLEKLEETILLIANISGFPADKLHDIRKKMRGM--AEIKVTNTTLEIAKNAG--LDVS 80
RLA0 SULSO -----MKRLALALKQRVSSWGLEVEKELTEKLTETILLIANISGFPADKLHEIRKKLRGK--ADIKVTNTTLEIAKNAG--LDVS 80
RLA0 AERPE MSYVSEVQMYKREKPEPKETLMIRELEELFSKRVVLFADITGTFYVYRVRKKLWKK--YPMVAKKRIILEAMKAGGLE--LDNN 86
RLA0 PYRAE MMLAIGKRRYVTRQTPARKVIVSEATELLQKTPYVFLFDLHGLSSRIHEVRYRLRRY--GVIKIIPKTLFKIAFTKVVGG--IPAE 85
RLA0 METAC MAEERHHTTEHIPQWKDEIENIKELIQSKVFGMVGIEGILATKMKIRRDLDKV--AVLKVRNTLTERALNQLG--ETIP 78
RLA0 METMA MAEERHHTTEHIPQWKDEIENIKELIQSKVFGMVRIEGILATKMKIRRDLDKV--AVLKVRNTLTERALNQLG--ESIP 78
RLA0 ARCFU MAAVRGS--PEPKYRAVEEIKRMISSEPVVAIVSRNVPAGCMKIRREFRGK--AEIKVTNTLTERALDAG--GDL 75
RLA0 METKA MAVKAKGQPSGYSYKVAEKKRREVKELKEMDETEVGLVDLSLPAPQLAEIRAKLERD--ELIRMSNTLMRIALEEKLDER--PELE 80
RLA0 METH MAEERHHTTEHIPQWKDEIENIKELIQSKVFGMVRIEGILATKMKIRRDLDKV--AVLKVRNTLTERALNQLG--ESIP 78
RLA0 METTL MITAESEHKIAPWKIEEVNALKKELKLNQIIVALVDMMEVPARLQETRDKIT--ETMLKMSNTLTERALKEVAEETGNPEFA 82
RLA0 METVA MIDAKSEHKIAPWKIEEVNALKKELKLNQIIVALVDMMEVPARLQETRDKIT--DQMLKMSNTLTERALKEVAEETGNPEFA 82
RLA0 METJA METKYKAVHAPWKIEEVNALKKELKLNQIIVALVDMMEVPAPQLQETRDKIT--DKVKLRMSNTLTERALKEVAEETGNPEFA 81
RLA0 PYRAB MAHVAEWKKEVEELANLIKSPVIALVDYSSMPAYPLSQMRRLIRENGCLLRVSRNTLTERALKEVAEETGNPEFA 77
RLA0 PYRHO MAHVAEWKKEVEELANLIKSPVIALVDYSSMPAYPLSQMRRLIRENGCLLRVSRNTLTERALKEVAEETGNPEFA 77
RLA0 PYRFO MAHVAEWKKEVEELANLIKSPVIALVDYSSMPAYPLSQMRRLIRENGCLLRVSRNTLTERALKEVAEETGNPEFA 77
RLA0 PYRKO MAHVAEWKKEVEELANLIKSPVIALVDYSSMPAYPLSQMRRLIRENGCLLRVSRNTLTERALKEVAEETGNPEFA 76
RLA0 HALMA MSESEERKTETIPKQKEVEEVAIVHIESVESVGVVNTACIP--RLQDMRDLHET--AFLVSRNTLTERALDDVD--DELE 79
RLA0 HALVO MSESEVRQTEVIPQWKREVEEVDLVDIESVESVGVVYVACIP--RQLQDMRDLHES--RAVMSRNTLTERALDEVN--DGEF 79
RLA0 HALSA MSEEQRTTEVIPQWKREVEEVDLVDIESVESVGVVNTACIP--RQLQDMRDLHES--RAVMSRNTLTERALDEVN--DGLD 72
RLA0 THEAC MKEVVSQKKELVNEITIRIKASRSVAIVDAGIRRTQIDIRGKNRQK--INLVIKTLTLLKALENLGD--EKIS 79
RLA0 THEVO MRKINPKKKEIVSELAADITKSKAVALVDIKGVRSRQMODIRAKNRQK--VKIKVVKTLTLLKALDSIND--EKIT 72
RLA0 PICTO MTEPKKKIDFVKNLENEINSRKVAALVSIKGLRNNFQKIRNSIRDK--ARIKVTARLLRLALENLGK--NNIV 72
ruler 1. .... 10. .... 20. .... 30. .... 40. .... 50. .... 60. .... 70. .... 80. .... 90

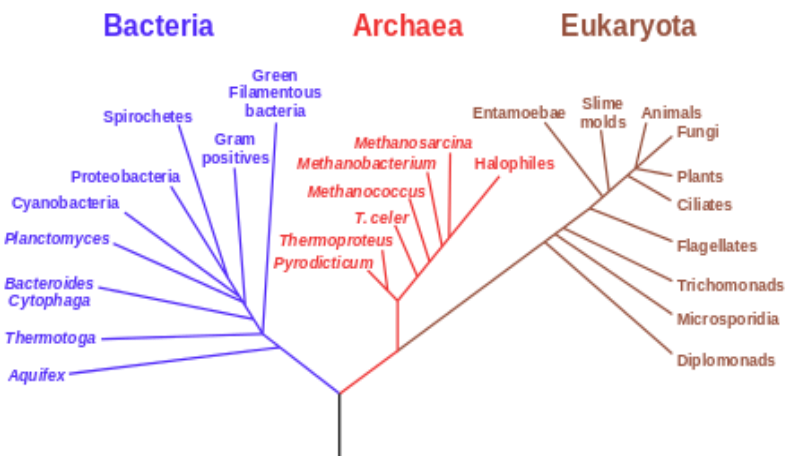
```

MSA of Ribosomal Protein P0 from Wikipedia, "Multiple Sequence Alignment"

13

13

MSA-Derived Phylogenetic Tree



Phylogenetic Tree derived from ribosomal proteins, Wikipedia "Phylogenetic Tree"

14

14

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.

15

15

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/>

16

16

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

17

17

BLAST Main Page

BLAST: 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](#)

Web BLAST

Nucleotide BLAST
nucleotide → nucleotide

blastx
translated nucleotide → protein

tblastn
protein → translated nucleotide

Protein BLAST
protein → protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Human Mouse Rat Microbes

18

18

The screenshot shows the Nucleotide BLAST Search interface. Red arrows point to specific features:

- Sequence Input:** Points to the "Enter Query Sequence" section, which includes a text box for "Enter accession number(s), g(s), or FASTA sequence(s)", a "Query subrange" section with "From" and "To" fields, and an "Or, upload file" section with a "Browse..." button.
- Databases to Search Against:** Points to the "Choose Search Set" section, which includes a "Database" dropdown menu (set to "Nucleotide collection (nr/nt)"), an "Organism" text box, and an "Exclude" section with checkboxes for "Models (MM/PP)" and "Sequences from type material".
- Program Selection:** Points to the "Program Selection" section, which includes a "Optimize for" section with radio buttons for "Highly similar sequences (megablast)", "More dissimilar sequences (discontiguous megablast)", and "Somewhat similar sequences (blastn)".
- Click to Run!:** Points to the "BLAST" button at the bottom left of the form.

A red box highlights a message: "New columns added to the Description Table. Click 'Select Columns' or 'Manage Columns'." with a link to "Select Columns".

19

The screenshot shows the BLAST interface with multiple tabs open. Red arrows point to specific features:

- Same Page Organization:** Points to the "BLAST" button in the top right corner of the interface.
- Translated BLAST: blastx:** Points to the "Translated BLAST: blastx" tab.
- Translated BLAST: tblastn:** Points to the "Translated BLAST: tblastn" tab.
- Translated BLAST: tblastx:** Points to the "Translated BLAST: tblastx" tab.

The interface shows the "Basic Local Alignment Search Tool" (BLAST) logo and navigation links: "Home", "Recent Results", "Saved Strategies", and "Help". The "Enter Query" section is visible, including "Enter accession", "Or, upload file", "Job Title", and "Align two or more sequences". The "Choose Search Set" section is also visible, including "Database", "Organism", and "Exclude" options.

20

BLAST Example

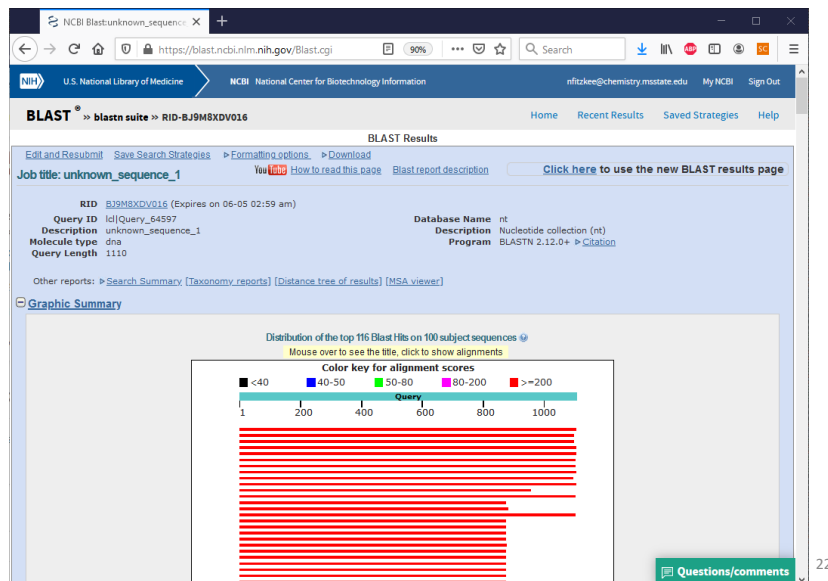
- What gene is this?

```
>unknown_sequence_1
TGATGTCAAGACCTCTATGAGACTGAAGTCTTTTCTACCGACTTCTCCAACATTTCTGCAGCCAAGCAG
GAGATTAACAGTCATGTGGAGATGCAAACCAAGGGAAAGTTGTGGGTCTAATTCAGACCTCAAGCCAA
ACACCATCATGGTCTTAGTGAACATATTTCACTTTAAAGCCAGTGGGCAAAATCCTTTTGATCCATCCAA
GACAGAAGACAGTTCAGCTTCTTAATAGACAAGACCACCACTGTTCAGTGCCCATGATGCACCAGATG
GAACAATACTATCACCTAGTGGATATGGAATTGAAGTGCACAGTTCTGCAAAATGGACTACAGCAAGAATG
CTCTGGCACTCTTTGTTCTTCCCAAGGAGGGACAGATGGAGTCAGTGGAGCTGCCATGTCATCTAAAC
ACTGAAGAAGTGAACCGCTTACTACAGAAGGGATGGGTGACTTGTGTTGTTCCAAAGTTTCCATTTCT
GCCACATATGACCTTGAGCCACACTTTTGAAGATGGGCATTCAGCATGCCTATTCTGAAAATGCTGATT
TTTCTGGACTCAGAGGACAATGGTCTGAAACTTTCCAATGCTGCCATAAGGCTGTGCTGCACATTGG
TGAAAAGGGAAGTGAAGCTGCAGCTGTCCCTGAAGTTGAAGTTTCGGATCAGCCTGAAAACACTTTCCTA
CACCTATTATCCAAATGATAGATCTTTCATGTTGTTGATTTTGGAGAGAAGCACAAGGAGTATCTCT
TTCTAGGGAAAGTTGTGAACCAACGGAAGCGTAGTTGGGAAAAAGGCCATTGGCTAATTGCACGTGTGT
ATTGCAATGGGAAATAAATAAATATAGCCTGGTGTGATGTGATGTGAGCTTGGACTTGCATTCCTTAA
TGATGGGATGAAGATTGAACCTGGCTGAAGTTGTTGGCTGTGGAAGAGGCCAATCCTATGGCAGAGCA
TTCAGATGTCAATGAGTAATTCATTATTATCCAAAGCATAGGAAGGCTCTATGTTGTATATTCTCTT
TGTCAGAATACCCCTCAACTCATTGCTCTAATAAATTGACTGGGTGAAAAATTAAAA
```

21

21

BLAST Results



22

22

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

23

23

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/>

24

24

Aligning via Structure

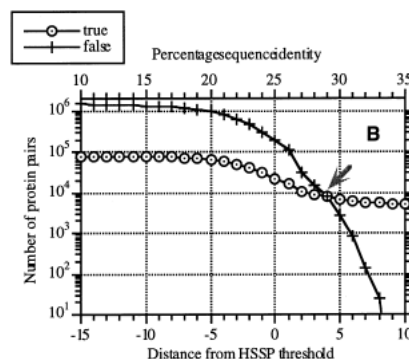
- So far we've focused on sequence alignment: looking at the primary (DNA or protein) sequence
- What about structural alignment? (Think shape or similar domains)
- VAST (Vector Alignment Search Tool) at NCBI: <https://structure.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>

25

25

Homology Modeling

- Proteins with similar sequences tend to have similar structures.
- When sequence identity 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

26

26

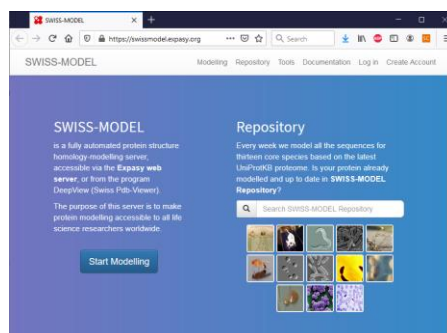
What is Homology Modeling?

- **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: it’s a model!

27

27

Homology Modeling Servers: SWISS-MODEL



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

28

28

Homology Modeling Servers: Phyre²

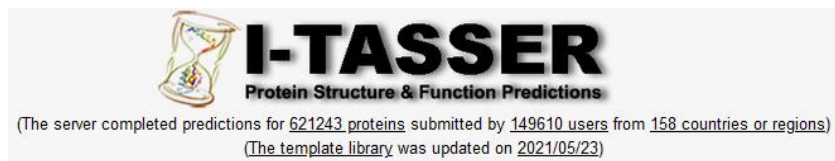


- Web page: <http://www.sbg.bio.ic.ac.uk/phyre2/>
- Trade off: can take 1-2 hours depending on server demand, but better structures
- Uses multiple templates, users can exclude files

29

29

Homology Modeling Servers: I-TASSER



- Web page: <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>
- Slowest option by far; can take a day or more
- Uses multiple templates and performs sophisticated refinement

30

30

Homology Modeling Example

- Sequence for Pin1 protein:

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

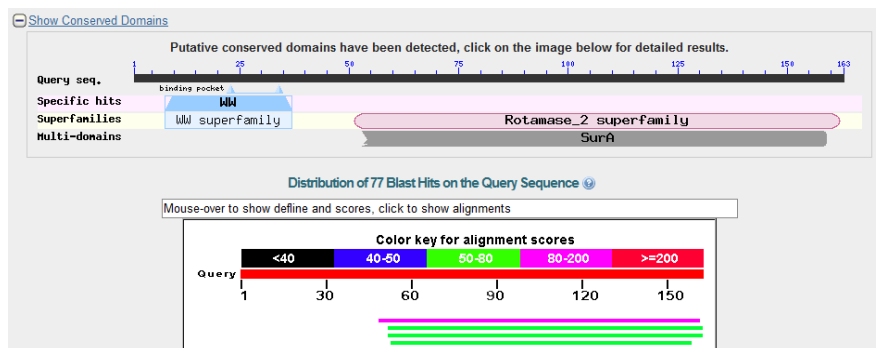
- Use BLAST to identify a homologous cis-trans prolyl isomerase in *Methanocorpusculum labreanum*

31

31

Homology Modeling Example

- Initial BLASTp result:



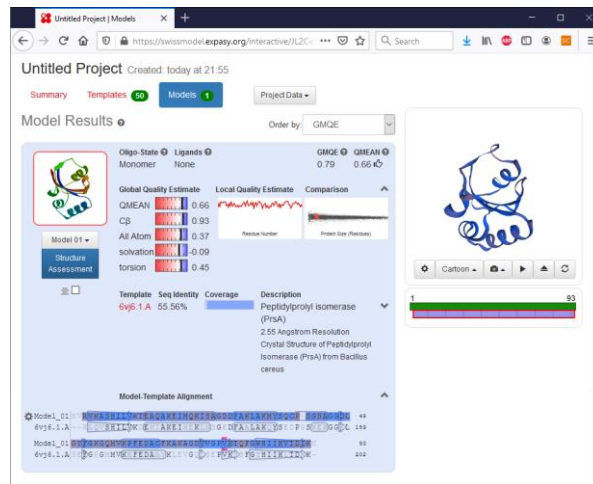
- Sequence (only second domain found):

```
MVRVKASHIL VKTEAQAKEI MQKISAGDDF AKLAKMYSQC PSGNAGGDLG
YFGKGQMVKP FEDACFKAKA GDVVGPKVTQ FGWHIIKVTD IKN
```

32

32

Result: SWISS-MODEL

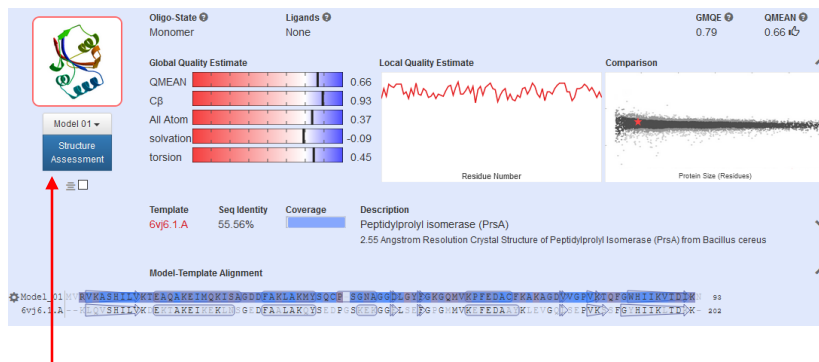


- We'll do this model in class

33

33

Result: SWISS-MODEL

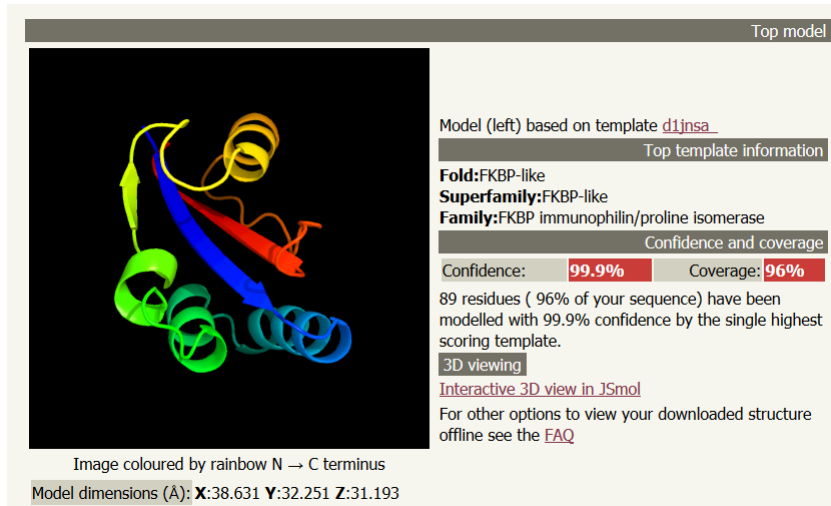


Click here to view Ramachandran plots, structure quality by residue, etc.

34

34

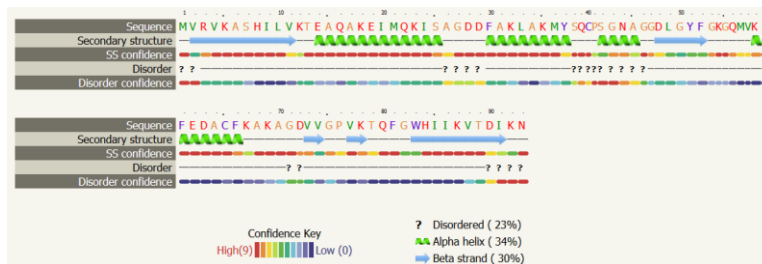
Result: Phyre²



35

35

Result: Phyre²

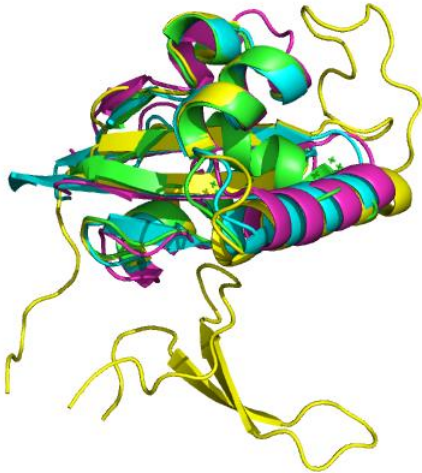


36

36

- 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`

Comparison of Results



- Colors:
 - Original Pin1
 - SWISS-MODEL
 - Phyre²
 - I-TASSER
- **Important:** How much side chain accuracy do I need?

39

39

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

40

40