

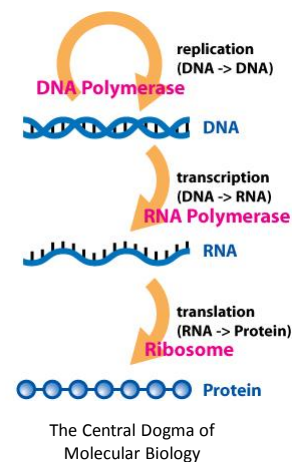
Basic Bioinformatics, Sequence Alignment, and Homology

Biochemistry Boot Camp
Session #9
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



Primary Structure (Sequence)

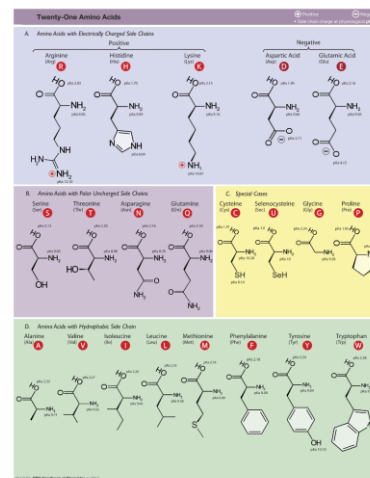
- **DNA and Proteins are chemically complex**, but their “alphabets” are rather simple.
 - 4 nucleobases (A, C, T, G)
 - 20 amino acids
- DNA sequences are represented from 5' to 3'



3

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

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

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 maximus]
LCLYTHIGENIYGSILYSEWNIQIMLLITMATAFGVYLVWQMSFWGATVITWLSAIPFYICNLY
EWINGQFSVDKATINRFAPHEILFFWVALAGVHILFHEGSGNDELGLTSDSKIIFNHPYTIKDFLG
LL.L.L.L.L.L.L.L.L.L.L.S.D.M.L.G.D.F.N.H.M.R.A.D.P.I.N.T.L.H.I.K.E.W.Y.F.A.Y.A.L.L.S.V.N.K.L.G.V.L.A.L.F.L.S.V.I.L
GLMPFLHTSKHRSNMLRPLSQLEWTLTMDLLETWTWIGSQVEYPTTIIIGQNASILYFSIILAFLIAGK
IENV
```

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTCAACTCACAGTTT
+
!''*(((****+))%%++) (%%%) .1***-+*')**55CCF>>>>>CCCCC65
```

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

Sequence Alignment

Sequence 1: GARVEY

Sequence 2: AVERY

Global Alignment:

GARVE-Y

-A-VERY

8

Global Sequence Alignment

- Many tools available, including Biology Workbench (ALIGN tool)
- EMBOSS Needle
http://www.ebi.ac.uk/Tools/psa/emboss_needle/
- **Example:** Human vs. Nematode Calmodulin (global sequence #1 and #2)

9

Global Sequence Alignment

- EMBOSS Needle Options:

How to compare residues?

How much penalty to open a gap in the sequence?

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

MSA Example

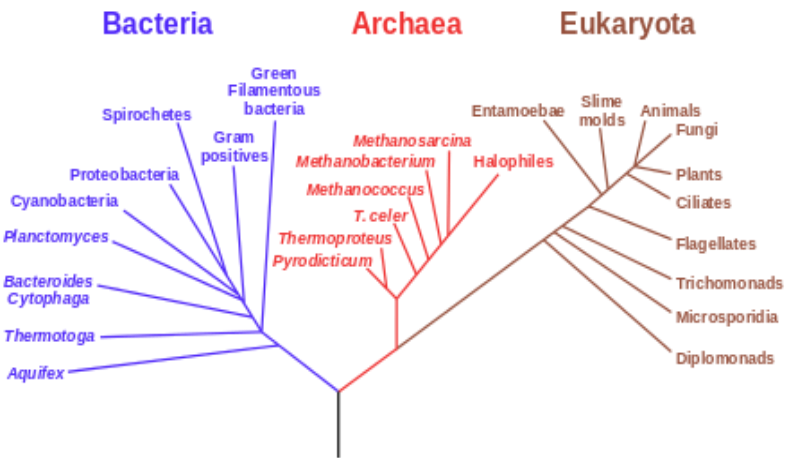
```

Q5E940_BOVIN -----M*PREDRATWKSNYELKIIQLDDYKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--PALE 76
RLA0_HUMAN -----M*PREDRATWKSNYELKIIQLDDYKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--PALE 76
RLA0_MOUSE -----M*PREDRATWKSNYELKIIQLDDYKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--PALE 76
RLA0_RAT -----M*PREDRATWKSNYELKIIQLDDYKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--PALE 76
RLA0_CHICK -----M*PREDRATWKSNYFMKIIQLDDYKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--PALE 76
RLA0_RANSY -----M*PREDRATWKSNYELKIIQLDDYKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--SALE 76
Q7ZUG3_BRARE -----M*PREDRATWKSNYELKIIQLDDYKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--PALE 76
RLA0_ICTPU -----M*PREDRATWKSNYELKIIQLDDYKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--PALE 76
RLA0_DROME -----M*VRENKRAWKAOYEIKVVELEDFKPCFIVGADNVGSKMQQIRMSLRGK-AVILMGKNTMMRKAIRGHLENN--DOLE 76
RLA0_DICDI -----M*SGAG-SKRKKLFTEKATKLFITDKRMIVAEADVFVSSQLQKIRKSIRGI-GAVILMGKNTMIRKVIKRLADSK--PELD 75
Q54LPO_DICDI -----M*SGAG-SKRKNVFEKATKLFITDKRMIVAEADVFVSSQLQKIRKSIRGI-GAVILMGKNTMIRKVIKRLADSK--PELD 75
RLA0_PLAF8 -----M*MAKLSKQKKQMYIEKLSLILQYKSKILLVHVDNVGSKMQQIRMSLRGK-AVILMGKNTMIRKVIKRLADSK--POLE 76
RLA0_SULAC -----M*HIELAVITTKKIAKRVDEVAELTEKLLKTKTILLIANTEGFPADKLEIRKRLGK-ADIKVTKNHLNFIALKNAG----VDK 79
RLA0_SULTO -----M*HIMAVITQERKIAKRVDEVAELTEKLLKTKTILLIANTEGFPADKLEIRKRLGK-ADIKVTKNHLNFIALKNAG----LDVS 80
RLA0_SULSO -----M*KNLALALKQRVSSKLEVEKELTEKLLKTKTILLIANTEGFPADKLEIRKRLGK-ADIKVTKNHLNFIALKNAG----DLE 80
RLA0_AERPE M*SVYSEVQMYKREKTEPEKTLMLRELELESKRIVVLEADITGTFVYVYKVKLWKK-VPMVAKKRLLRANKAAGLE----LDDN 86
RLA0_PYRAE M*MLAIGKRRYVRETRQYPAKRVKIVSEATELQKQYVYVLEDFLHGLSRIKLEHYRRLRY-GVIKIIPKLFKIAFTKVVGG--IPAE 85
RLA0_METAC M*MAERHHTTEHPQWKDEIENIKELIQSKVFGVGLGILATKIQKIRDLKDV-AVLKVRNTLLEKALNQLG----ETIP 78
RLA0_METMA M*MAERHHTTEHPQWKDEIENIKELIQSKVFGVGLGILATKIQKIRDLKDV-AVLKVRNTLLEKALNQLG----ESIP 78
RLA0_ARCFU M*MAAVRGS--DPEYKVRAVEEIKRMISSEVYVALVSRNVFAGCMKIRREFRQK-AEIKVYKNTLLEKALDALG----GDVL 75
RLA0_METKA M*MAVKAKGQPSYSEKVAEWRKREKLEKLEMDPEVVGELVDLELPADQLQELAKLRERDQIRMSNTLMRLALEEKLEER--PELE 80
RLA0_METTH M*MAHVAEKKKWEQEHDLTKSEVVGELVAOLPARKMOTLRDS-ALIRMSKFLISALEKQREL--DND 74
RLA0_METTL M*MITAESEHKIAPWKIEEWNALKKELKLNQOIVALVDMMEVPAVQLQELRDKIR-ETMLKMSRNTLLEKRAVEEVAETGNPEFA 82
RLA0_METVA M*MIDAKSEHKIAPWKIEEWNALKKELKSNVIALIDMMEVPAVQLQELRDKIR-DQMLKMSRNTLLEKRAVEEVAETGNPEFA 82
RLA0_METJA M*METKYKAVAPWKIEEVTIKKELKSNVIALVDMMDVPAVQLQELRDKIR-DKVKLRMSRNTLLEKRAVEEVAETGNPEFA 81
RLA0_PYRAB M*MAHVAEKKKWEQEHDLTKSEVVGELVAOLPARKMOTLRDS-ALIRMSKFLISALEKQREL--DND 77
RLA0_PYRHO M*MAHVAEKKKWEQEHDLTKSEVVGELVAOLPARKMOTLRDS-ALIRMSKFLISALEKQREL--DND 77
RLA0_PYRFU M*MAHVAEKKKWEQEHDLTKSEVVGELVAOLPARKMOTLRDS-ALIRMSKFLISALEKQREL--DND 77
RLA0_PYRKO M*MAHVAEKKKWEQEHDLTKSEVVGELVAOLPARKMOTLRDS-ALIRMSKFLISALEKQREL--DND 77
RLA0_HALMA M*MSSESRKTETIPKQKQEVDAIVMIESVSYGVVNIAGIPIRDLDMRDLHET-AELRVSRNTLLEKALDDV----DDEE 79
RLA0_HALVO M*MSSESRKTETIPKQKQEVDAIVMIESVSYGVVNIAGIPIRDLDMRDLHET-AELRVSRNTLLEKALDDV----DDEE 79
RLA0_HALSA M*MSAEQRTTEVPEKQKQEVDAIVMIESVSYGVVNIAGIPIRDLDMRDLHET-AELRVSRNTLLEKALDDV----DDEE 79
RLA0_THEAC M*MKVYSQKKELVNEITDRIKASRSVAIVDAGIRTRQIDIRGKNRQK-IMLVYKIKLLFKALDNLG--EKIS 72
RLA0_THEVO M*MRKINPKKKEIVSELAQDITKSKAVAVDIKGVRRQMODIRAKNRQK-IMLVYKIKLLFKALDNLG--EKIS 72
RLA0_PICTO M*TEPRQKIDFVKNLENEINSRKYVAALVSLKGLRNNRFGKIRNSIRDK-ARIKVEARLLRLALEKNGK--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.

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

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

BLAST Main Page

The screenshot shows the NCBI BLAST main page. At the top, there is a navigation bar with tabs for Home, Recent Results, Saved Strategies, and Help. Below this, the page is divided into several sections:

- NCBI BLAST Home:** A search bar with the text "BLAST finds regions of similarity between biological sequences. [more...](#)" and a "New" alert for "Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool."
- BLAST Assembled RefSeq Genomes:** A section where users can choose a species genome to search, with a link to "list all genomic BLAST databases". It lists various species including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbas, and Apis mellifera.
- Basic BLAST:** A section where users can choose a BLAST program to run. It lists:
 - nucleotide blast:** Search a nucleotide database using a nucleotide query. Algorithms: blastn, megablast, discontinuous megablast.
 - protein blast:** Search protein database using a protein query. Algorithms: blastp, psi-blast, phi-blast, delta-blast.
 - blastx:** Search protein database using a translated nucleotide query.
 - tblastn:** Search translated nucleotide database using a protein query.
 - tblastx:** Search translated nucleotide database using a translated nucleotide query.
- Specialized BLAST:** A section where users can choose a type of specialized search (or database name in parentheses).

On the right side of the page, there are sections for "News" (Microbial BLAST) and "Tip of the Day" (Use Genomic BLAST to see the genomic context).

18

The screenshot shows the NCBI Standard Nucleotide BLAST web interface. The page title is "Standard Nucleotide BLAST". The main content area includes several sections: "Enter Query Sequence" with a text input field and "Query subrange" fields; "Or, upload file" with a "Choose File" button; "Job Title" with a text input field; "Align two or more sequences" with a checkbox; "Choose Search Set" with a "Database" dropdown menu (set to "Human genomic + transcript") and an "Exclude" section; "Program Selection" with radio buttons for "Highly similar sequences (megablast)", "More dissimilar sequences (discontiguous megablast)", and "Somewhat similar sequences (blastn)"; and a "BLAST" button. Red arrows point to the "Enter Query Sequence" field, the "Database" dropdown, the "Program Selection" radio buttons, and the "BLAST" button. Labels on the right side of the image identify these as "Sequence Input", "Databases to Search Against", "Program Selection", and "Click to Run!".

The screenshot shows the NCBI Translated BLAST web interface. The page title is "Translated BLAST: tblastx". The interface is similar to the Standard Nucleotide BLAST but includes a "Genetic code" dropdown menu (set to "Standard (1)"). Red arrows point to the "BLAST" button and the "Genetic code" dropdown. A label on the right side of the image says "Same Page Organization".

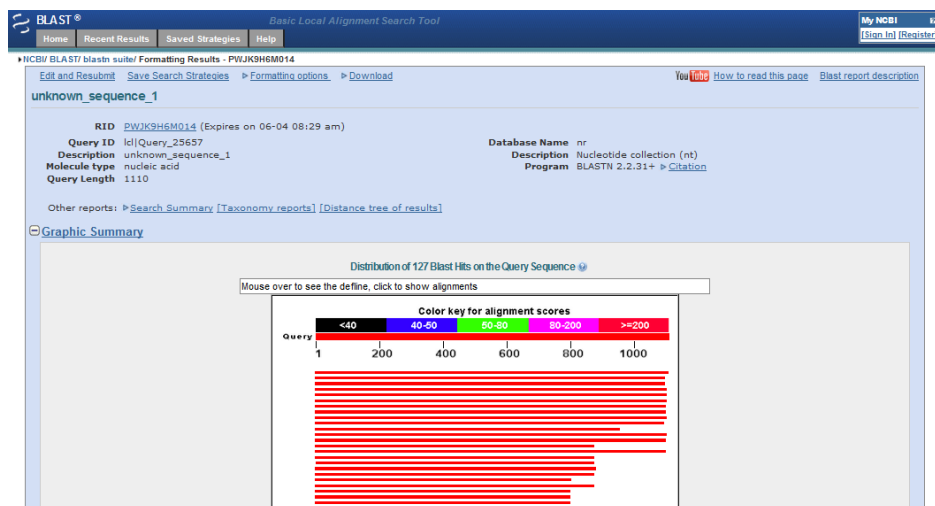
BLAST Example

- What gene is this?

```
>unknown_sequence_1
TGATGTCAAGACCTCTATGAGACTGAAGTCTTTTCTACCGACTTCTCCAACATTTCTGCAGCCAAGCAG
GAGATTAACAGTCATGTGGAGATGCAAACCAAAGGAAAGTTGTGGGTCTAATTC AAGACCTCAAGCCAA
ACACCATCATGGTCTTAGTGAACATATTTCACTTTAAAGCCCAGTGGGCAAATCCTTTTGATCCATCCAA
GACAGAAGCAGTTCCAGCTTCTTAATAGACAAGACCACCTGTCAAGTGCCCATGATGCACCAGATG
GAACAATACTATCACCTAGTGGATATGGAATTGAACTGCACAGTTCTGCAAAATGGACTACAGCAAGAATG
CTCTGGCACTCTTTGTTCTTCCCAAGGAGGACAGATGGAGTCAGTGGGAAGCTGCCATGTCATCTAAAC
ACTGAAGAAGTGAACCGCTTACTACAGAAGGGATGGGTTGACTTGTGTTGTTCCAAAGTTTTCCATTTCT
GCCACATATGACCTTGGAGCCACACTTTTGAAGATGGGCATTCAGCATGCCATTTCTGAAAATGCTGATT
TTTCTGGACTCACAGAGGACAATGGTCTGAAACTTTCCAATGCTGCCCATTAAGGCTGTGCTGCACATTTGG
TGAAAAGGGAACTGAAGCTGCAGCTGTCCTTGAAGTTGAACCTTTCGGATCAGCCTGAAAACACTTTCCTA
CACCCATTTATCCAAATGATAGATCTTTCATGTTGTTGATTTGGAGAGAAGCACAAAGGAGTATTTCTCT
TTCTAGGGAAAGTTGTGAACCCAACGGAAAGCGTAGTTGGGAAAAAGGCCATTGGCTAATTCACAGTGTGT
ATTGCAATGGGAAATAAATAAATAATATAGCCTGGTGTGATGATGTGAGCTTGGACTTGCATTCCTCTA
TGATGGGATGAAGATTGAACCTGGCTGAACCTTGTGGCTGTGGAAGAGGCCAATCCTATGGCAGAGCA
TTCAGAATGTCAATGAGTAATTCATTTATATCCAAAGCATAGGAAGGCTCTATGTTTGTATATTTCTCTT
TGTCAGAATACCCCTCAACTCATTTGCTCTAATAAATTTGACTGGGTTGAAAAATTAATA
```

21

BLAST Results



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

Installing BLAST Locally

Executables and documentation available at:

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

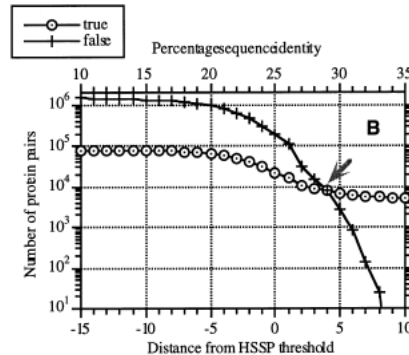
Documentation:

<http://www.ncbi.nlm.nih.gov/books/NBK1762/>

24

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 Philip Bryan and “fold switching”
- Can we use this to predict structures?



Below ~28% sequence identity, the number of structurally dissimilar aligned pairs explodes.

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

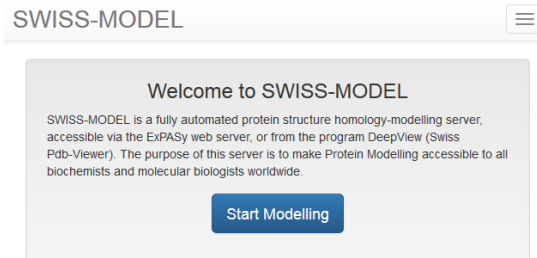
25

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!

26

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)

27

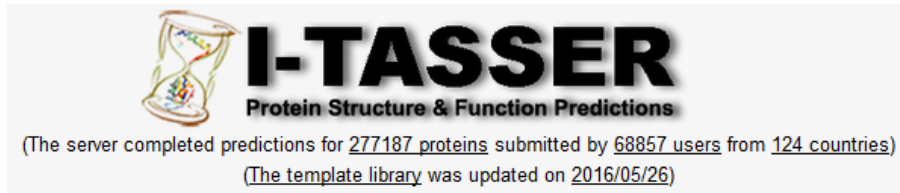
Homology Modeling Servers: **Phyre²**



- 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

28

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

29

Homology Modeling Example

- Sequence for Pin1 protein:

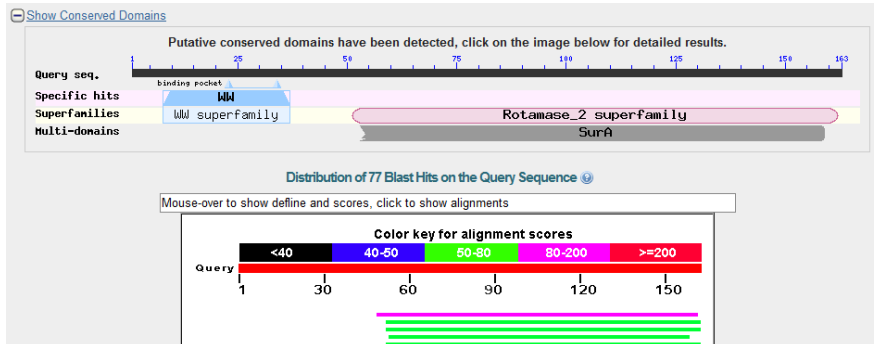
```
MADEEKLPPG WEKRMSRSSG RVYYFNHITN ASQWERPSGN SSSGGKNGQG
EPARVRCSHL LVKHSQSRRP SSWRQEKITR TKEEALELIN GYIQIKSGE
EDFESLASQF SDCSSAKARG DLGAFSRGQM QKPFEDASFA LRTGEMSGPV
FTDSGIHIIL RTE
```

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

30

Homology Modeling Example

- Initial BLASTp result:

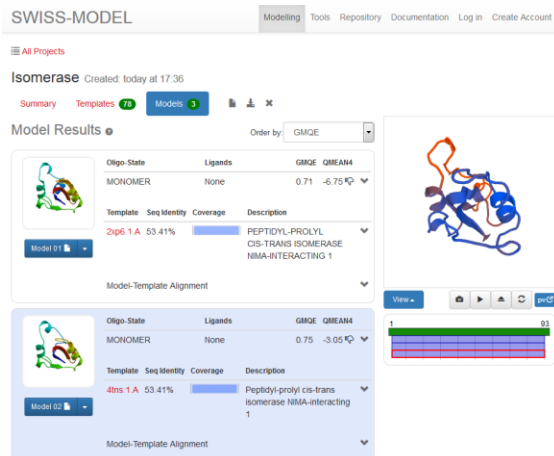


- Sequence (only second domain found):

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

31

Result: SWISS-MODEL









- View this result at:

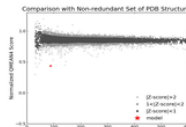
<http://swissmodel.expasy.org/interactive/wYh9Ak/models/>

32

Result: SWISS-MODEL

Model #01	File	Built with	Oligo-State	Ligands	GMQE	QMEAN4
	PDB	ProMod Version 3.70.	MONOMER	None	0.71	-6.75

QMEAN4	-6.75	
Cβ	-2.41	
All Atom	-2.34	
Solvation	-7.58	
Torsion	-2.76	



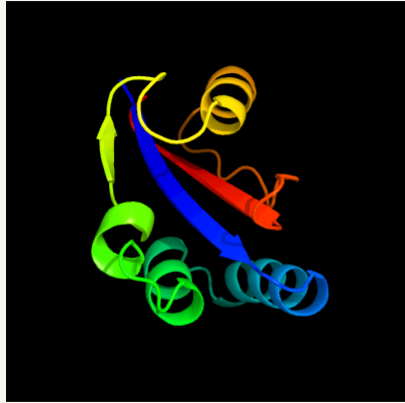
Template	Seq Identity	Oligo-state	Found by	Method	Resolution	Seq Similarity	Range	Coverage	Description
2xp6.1.A	53.41	monomer	BLAST	X-ray	1.90Å	0.45	3 - 90	0.95	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE NIMA-INTERACTING 1

Ligand	Added to Model	Description
12P	X - Binding site not conserved.	DODECAETHYLENE GLYCOL
4G2	X - Binding site not conserved.	2-(3-CHLORO-PHENYL)-5-METHYL-1H-IMIDAZOLE-4-CARBOXYLIC ACID

33

Result: Phyre²

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

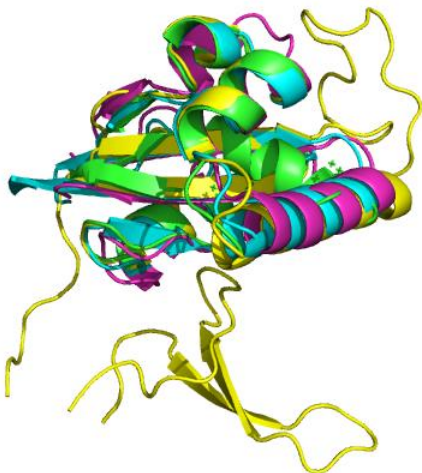
34

Comparison of Results

- **Download the following PDBs from the Boot Camp Website:**
 - 1pin.pdb – Original Pin1 Structure
 - swiss.pdb – SWISS-MODEL Result
 - phyre2.pdb – Phyre² Result
 - itasser.pdb – I-TASSER Result
- A pre-aligned PyMOL session (pse file) is also provided

37

Comparison of Results



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

38

Other Resources:

- EMBL-EBI (European Bioinformatics Institute) - <http://www.ebi.ac.uk/>
- DDBJ (DNA Data Bank of Japan) - <http://www.ddbj.nig.ac.jp/>
- NCBI's Sequence Read Archive (SRA) - <http://www.ncbi.nlm.nih.gov/sra>
- UCSC Genome Browser: <http://genome.ucsc.edu/>
- IGBB's Useful Links Page - <http://www.igbb.msstate.edu/links.php>

Many, many more available online, just search.

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