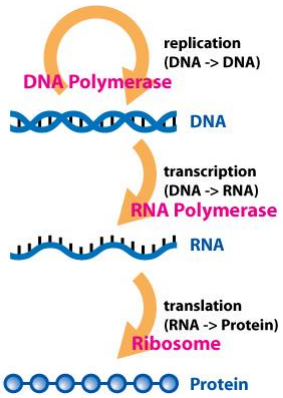
Basic Bioinformatics, Sequence Alignment, and Homology

Biochemistry Boot Camp 2017
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

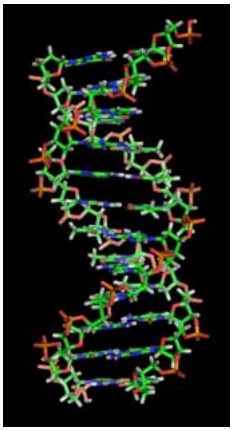


The Central Dogma of Molecular Biology

2

Primary Structure (Sequence)

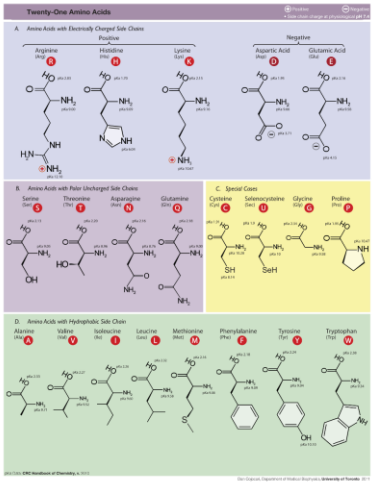
- **DNA and Proteins are chemically complex**, but their "alphabets" are rather simple.
  - 4 nucleobases (A, C, T, G)
  - 20 amino acids
- DNA sequences are represented from 5' to 3'



3

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

- 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
LLILILLLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

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

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

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

One Stop Shop for Many Tools

- Lots of tools are available as stand-alone packages online
- So far, our emphasis has been on these tools; however several "multiple tool" solutions also exist
- **Demo:** Biology Workbench http://workbench.sdsc.edu/

SDSC
SAN DIEGO SUPERCOMPUTER CENTER
Biology WorkBench

9

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)

10

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.

- Pretty darn similar!

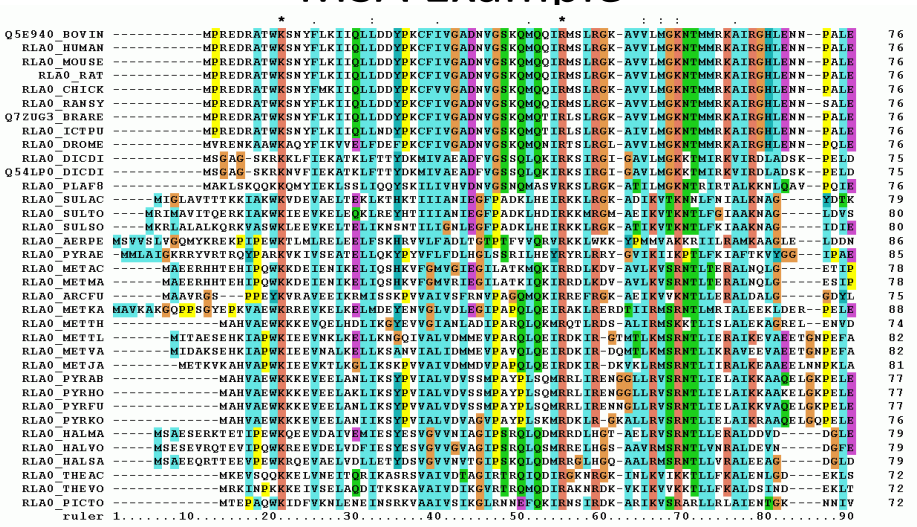
Identical residues shown with |, similar residues with : and ., and blanks represent dissimilar residues.

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
- Example software: MEGA, ClustalX
  http://www.megasoftware.net/
  http://www.clustal.org/

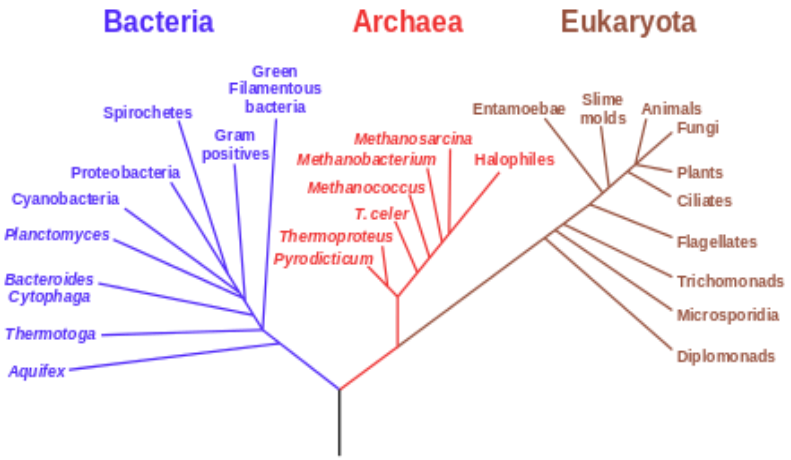
MEGA Molecular Evolutionary Genetics Analysis

MSA Example



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

MSA-Derived Phylogenetic Tree



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

15

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.

16

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/

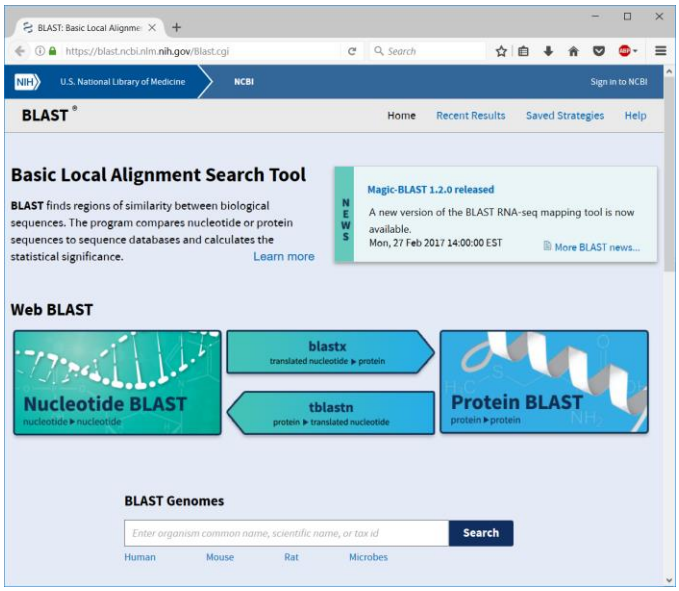
17

Flavors of BLAST

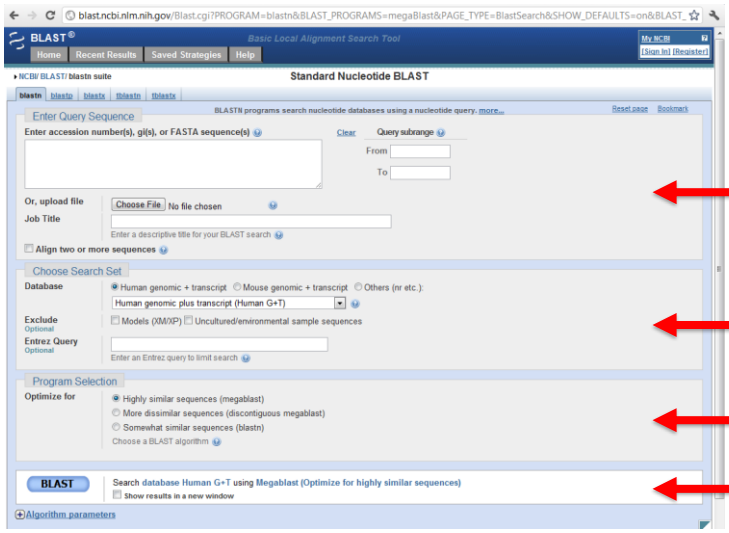
- **blastn** – searches a nucleotide database using a nucleotide query
  DNA/RNA sequence searched against DNA/RNA database
- **blastp** – searches a protein database using a protein query
  Protein sequence searched against a Protein database
- **blastx** – search a protein database using a translated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

18

BLAST Main Page



19

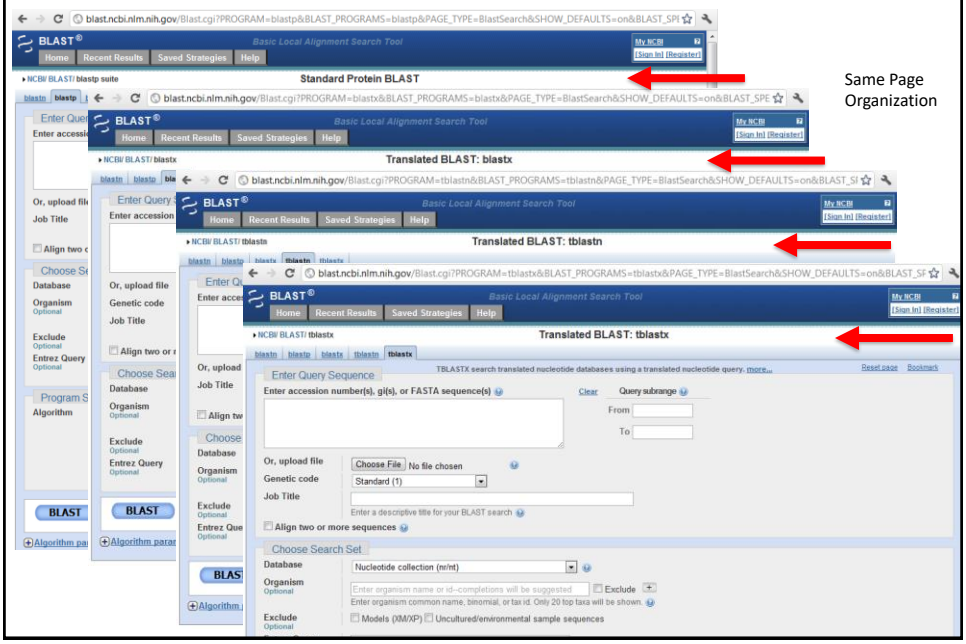


Sequence Input

Databases to Search Against

Program Selection

Click to Run!



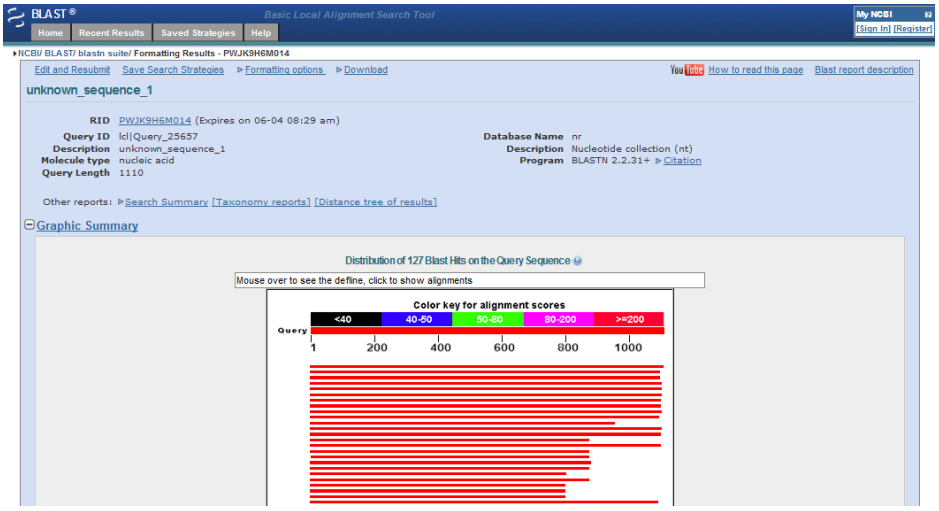
BLAST Example

- What gene is this?

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

22

BLAST Results



23

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

24

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/

25

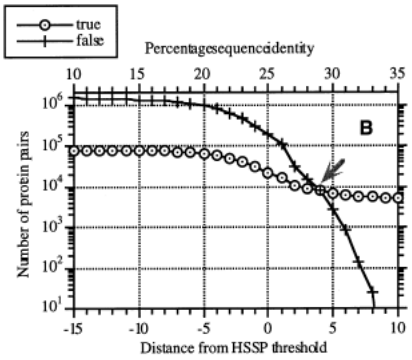
Aligning via Structure

- 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

26

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 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

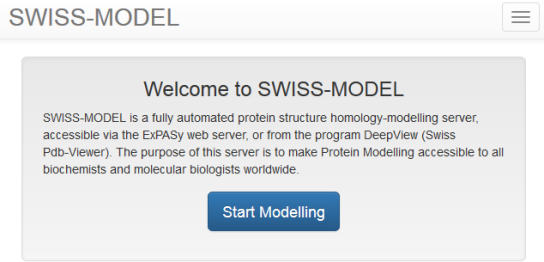
27

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!

28

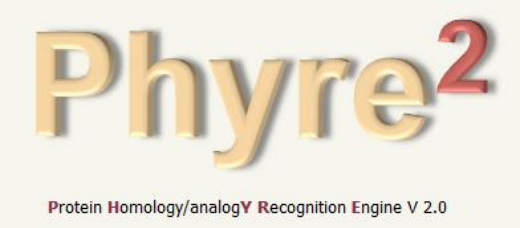
Homology Modeling Servers: **SWISS-MODEL**



- 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)

29

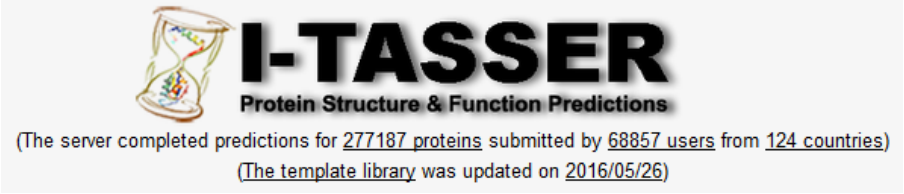
Homology Modeling Servers: **Phyre²**



- 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

30

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

31

Homology Modeling Example

- Sequence for Pin1 protein:

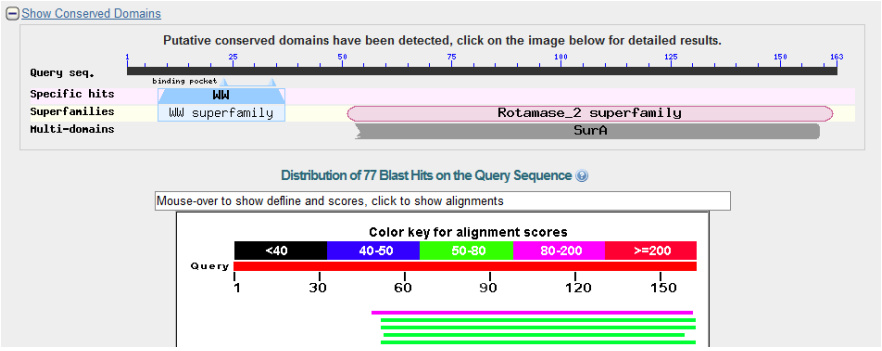
```
MADEEKLPPG WEKRMSRSSG RVYYFNHITN ASQWERPSGN SSSGGKNGQG
EPARVRCSHL LVKHSQSRRP SSWRQEKITR TKEEALELIN GYIQKIKSGE
EDFESLASQF SDCSSAKARG DLGAFSRGQM QKPFEDASFA LRTGEMSGPV
FTDSGIHIIL RTE
```

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

32

Homology Modeling Example

- Initial BLASTp result:

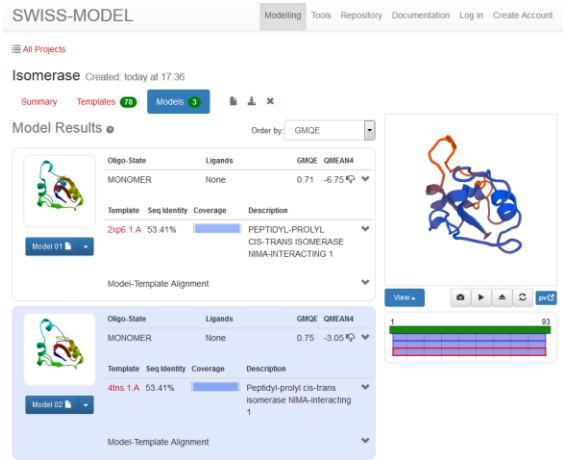


- Sequence (only second domain found):

MVRVKASHIL VKTEAQAKEI MQKISAGDDF AKLAKMYSQC PSGNAGGDLG
YFGKGQMVKP FEDACFKAKA GDVVGPVKTQ FGWHIIKVTD IKN

33

Result: SWISS-MODEL



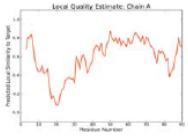
- We'll do this model in class

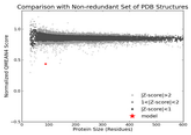
34

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	✕ - Binding site not conserved.	DODECAETHYLENE GLYCOL
4G2	✕ - Binding site not conserved.	2-(3-CHLORO-PHENYL)-5-METHYL-1H-IMIDAZOLE-4-CARBOXYLIC ACID

35

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.

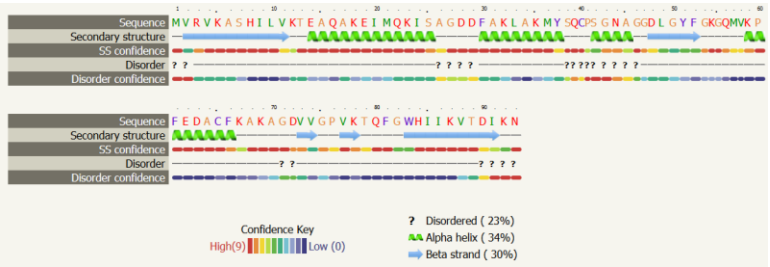
3D viewing

Interactive 3D view in JSmol

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

36

Result: Phyre^2



- 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`

37

Result: I-TASSER

Predicted Secondary Structure

```
                    20                  40                  60                  80
                     |                   |                   |                   |
Sequence    MVRVKASHILVKTEAQAKEIMQKISAGDDFAKLAKMYSQCPSGNAGGDLGYFGKGQMVKPFEDACFKAKAGDVVGPVKTQFGWHIIKVTDIKN
Prediction  CCSSSSSSSSCCHHHHHHHHHHHCCCCHHHHHHHCCCCCCCCCCCCCCCCCCCCCCHHHHHHHCCCCCCCCCCSSCCCSSSSSSSSSCC
Conf.Score  967998899988999999999988799899999868896524486455337997356999999838999978877698379999967659
            H:Helix; S:Strand; C:Coil
```

Predicted Solvent Accessibility

```
                    20                  40                  60                  80
                     |                   |                   |                   |
Sequence    MVRVKASHILVKTEAQAKEIMQKISAGDDFAKLAKMYSQCPSGNAGGDLGYFGKGQMVKPFEDACFKAKAGDVVGPVKTQFGWHIIKVTDIKN
Prediction  7643403111163574055026306736403510563173443763232330456622430250037166453362341631000030464 58
            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 `final.casp.pdb`

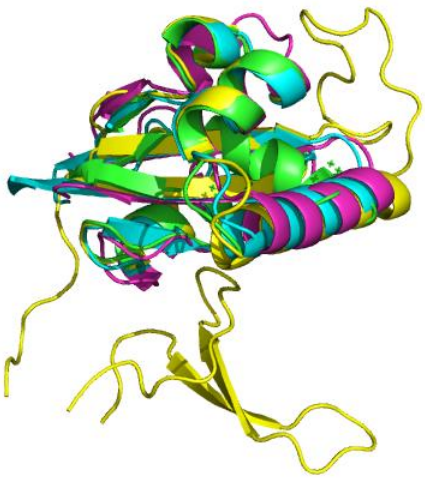
38

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
- A pre-aligned PyMOL session (pse file) is also provided
  - **Useful:** PyMOL "align" command
  - See handout on the website

39

Comparison of Results



- Colors:
  - **Original Pin1**
  - **SWISS-MODEL**
  - **Phyre^2**
  - **I-TASSER**
- **Important:** How much side chain accuracy do I need?

40

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

42