Drylab Discipline

- Drylab experiments are experiments
- Think about what you are about to do
- Know your tools
- A new directory for each experiment
- Meaningful file names
- Remove clutter (but make sure it **is** clutter)
- Keep a “lab notebook” 00README

Drylab Discipline: common problems

- No log: uh... what did I do there?
- Overwriting files
- Look before you leap:
  - what does that file contain?
  - How was that file created?
- Mis-identifying file content
- RTFM, RTFM, RTFM

Conserved Regions in Sequences

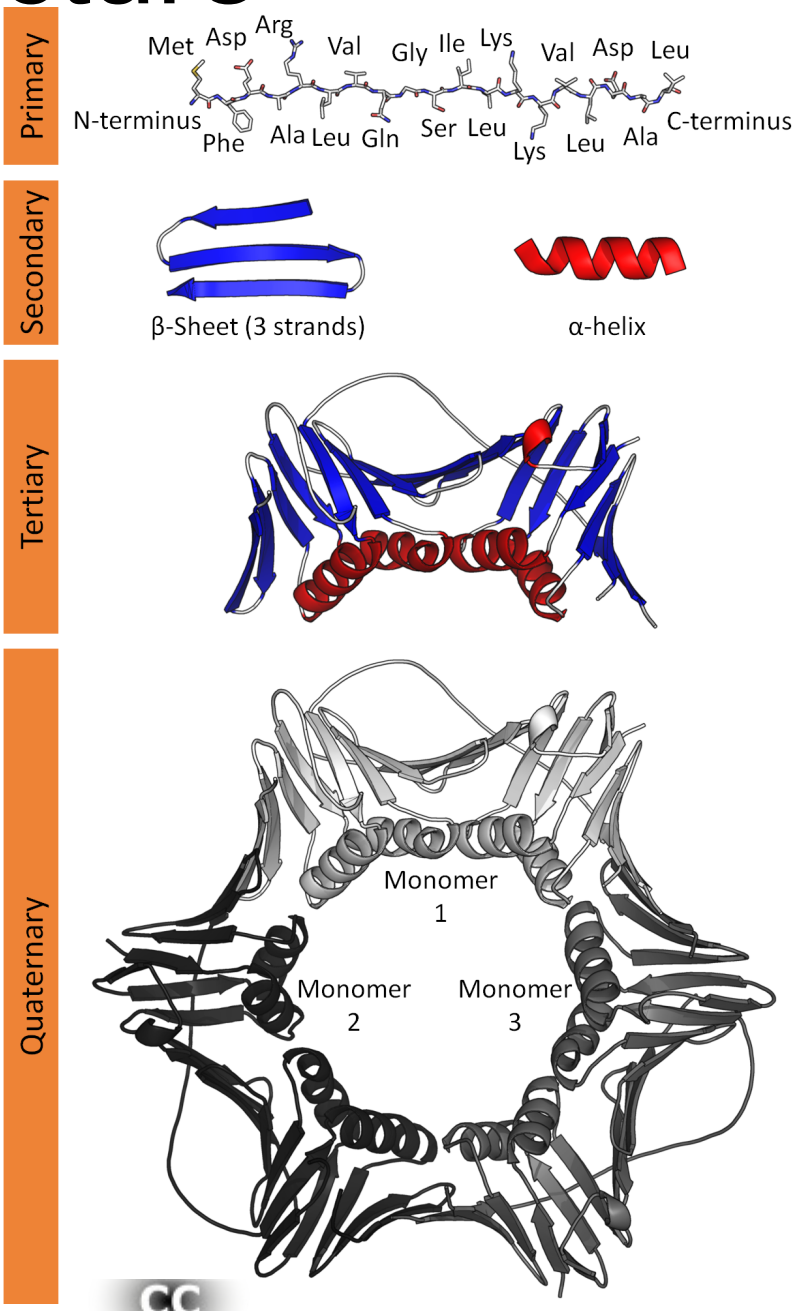
- Things that tend to stay conserved:
  - Structure
  - Function
- Conserved regions are a good place to start looking for biological importance

Protein Structure

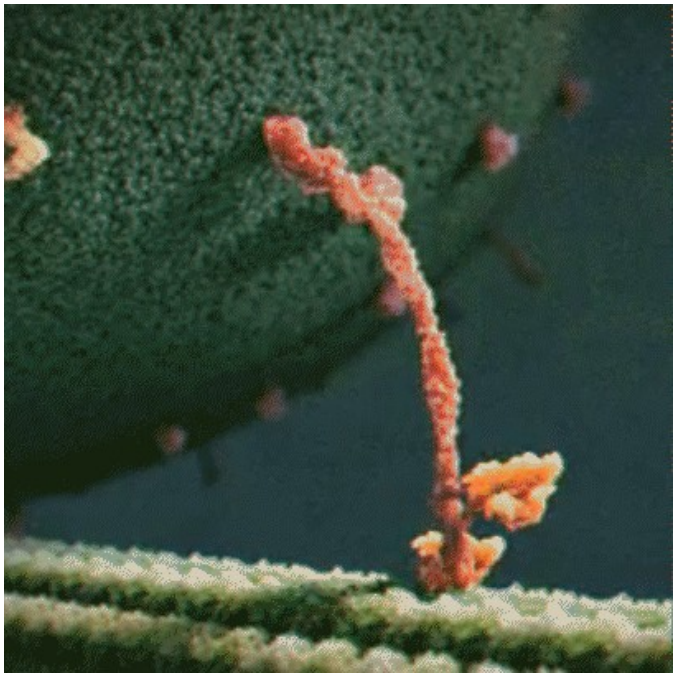
- The 3D arrangement of atoms in a protein molecule
- Proteins fold into conformations driven by
  - Hydrogen bonding
  - ionic interactions
  - hydrophobic packing
- Structure is important for understanding how proteins function
- Determining a structure takes more time & effort than a sequence, no guarantees

Protein structure

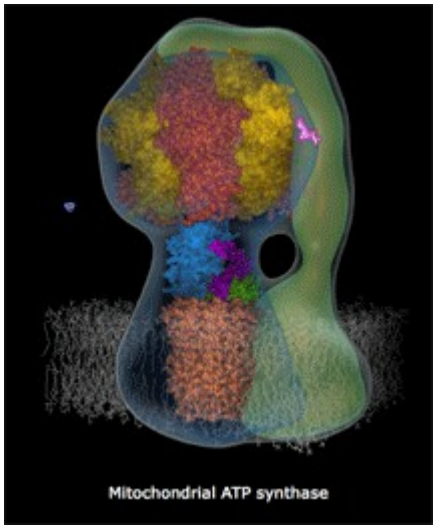
- Primary: amino-acid sequence
- Secondary: helices, sheets, turns
- Tertiary: three dimensional structure
- Quarternary: 3D structure aggregating >1 polypeptide chains



Why is Structure Important?



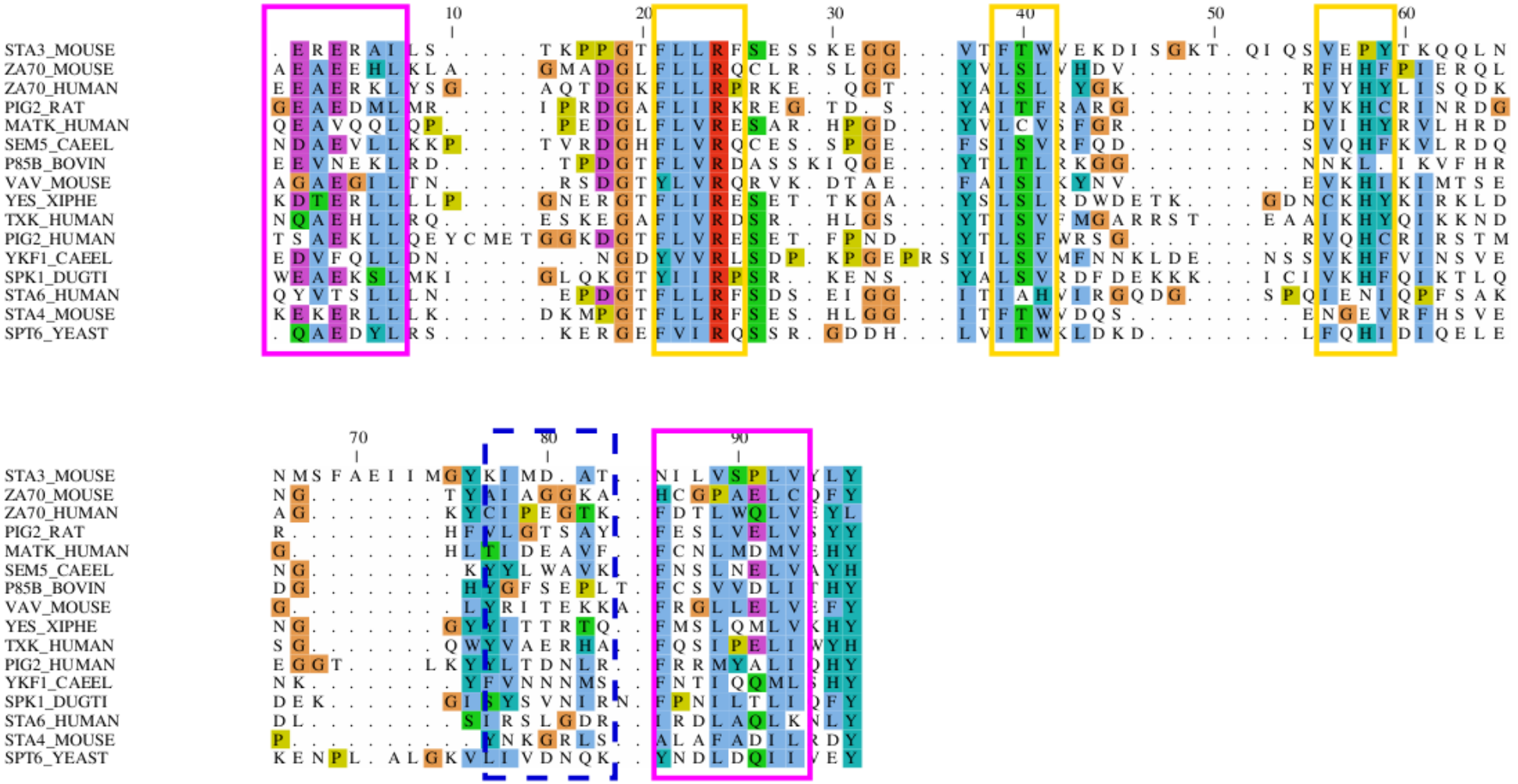
https://imgur.com/a/izZxc



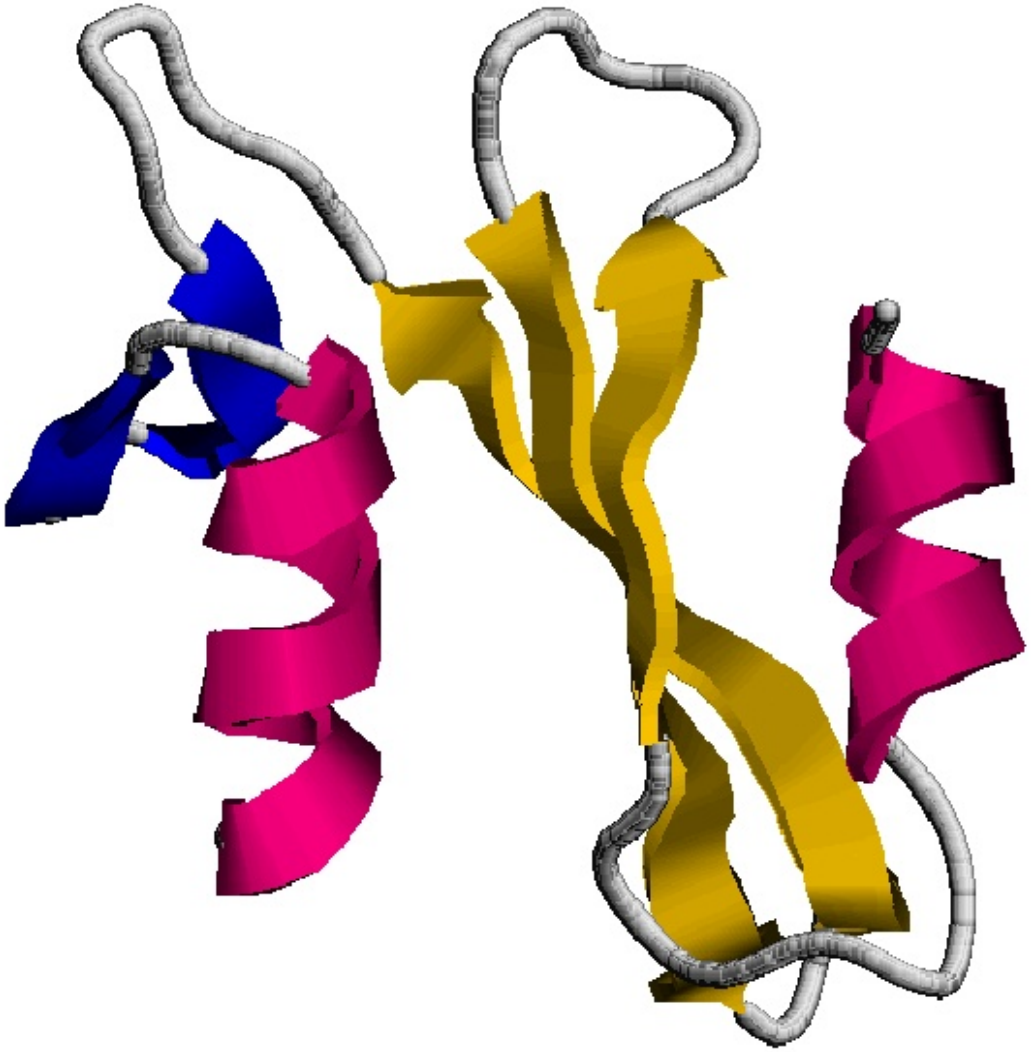
Multiple Sequence Alignments: Conserved Regions

- Multiple Sequence Alignments - detecting *conserved regions*
  - Revealing evolutionary relationships between more than 2 proteins
  - Structure
- The information represented by these conserved regions can be used to align sequences, search similar sequences in the databases or annotate new sequences.
- Different methods exist to build models of these conserved regions:
  - Consensus sequences;
  - Patterns
  - Position Specific Score Matrices (PSSMs)
  - Profiles;
  - Hidden Markov Models (HMMs),
  - and a few others.

MSA reveals secondary structures

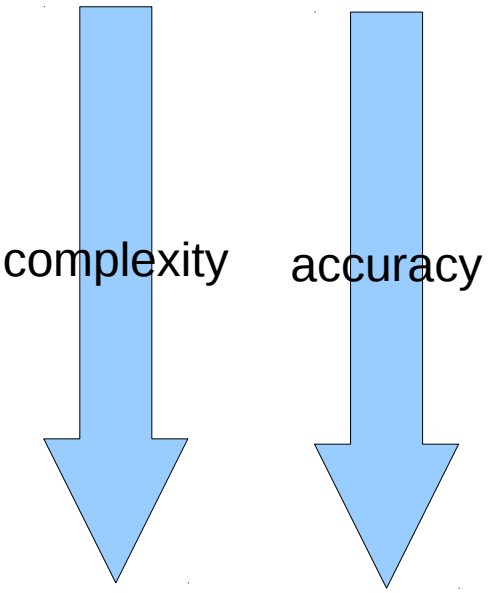


MSA reveals secondary structures



Representing Sequence Conservation

- Consensus sequence
- Sequence pattern
- Position specific score matrix
- Hidden Markov Model



Consensus Sequence

- The consensus sequence method is the simplest method to build a mode from a multiple sequence alignment.
- The consensus sequence is built using the following rules:
  - Plurality, or “majority wins”.
  - Skip too much variation.

Consensus Sequence

- The consensus sequence method is the simplest method to build a mode from a multiple sequence alignment.
- The consensus sequence is built using the following rules:
  - Plurality, or “majority wins”.
  - Skip too much variation.

Building a consensus sequence

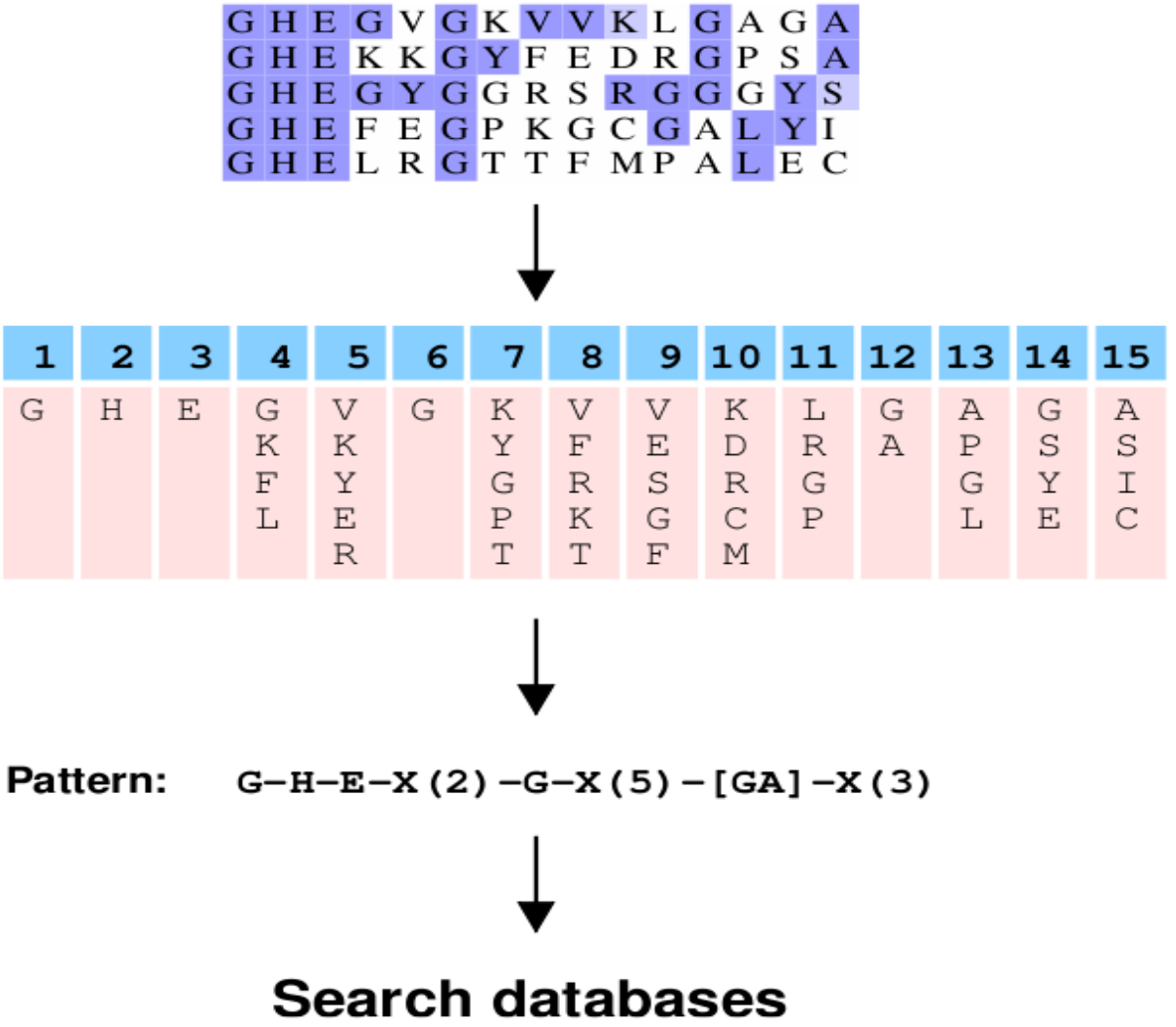
Pros and Cons of Consensus Sequence

- Pros:
  - Easy to implement
  - Fast
- Cons
  - No information about column variation
  - Very dependent on the training set.
  - No scoring, only binary result (YES/NO)
- When to use it?
  - Useful to find highly conserved signatures

Sequence Patterns

- A pattern describes a set of alternative sequences, using a single expression.
- In computer science, patterns are known as regular expressions.
- The Prosite syntax for patterns:
  - uses the standard IUPAC one-letter codes for amino acids (G=Gly, P=Pro, ...),
  - each element in a pattern is separated from its neighbor by a '-',
  - the symbol 'X' is used where any amino acid is accepted,
  - ambiguities are indicated by square parentheses '[ ]' ([AG] means Ala or Gly),
  - amino acids that are not accepted at a given position are listed between a pair of curly
  - brackets '{ }' ({AG} means any amino acid except Ala and Gly),
  - repetitions are indicated between parentheses '( )' ([AG](2,4) means Ala or Gly between
  - 2 and 4 times, X(2) means any amino acid twice),
- a pattern is anchored to the N-term and/or C-term by the symbols '<' and '>' respectively.

Building patterns



Sequence Signatures

- • Post-translational signatures:
- •Protein splicing signature:
- [DNEG]-x-[LIVFA]-[LIVMY]-[LVAST]-H-N-[STC]
- •Tyrosine kinase phosphorylation site:
- [RK]-x(2)-[DE]-x(3)-Y or [RK]-x(3)-[DE]-x(2)-Y
- • DNA-RNA interaction signatures:
- • Histone H4 signature:
- G-A-K-R-H
- • p53 signature:
- M-C-N-S-S-C-[MV]-G-G-M-N-R-R
- • Enzymes:
- L-lactate dehydrogenase active site:
- [LIVMA]-G-[EQ]-H-G-[DN]-[ST]
- • Ubiquitin-activating enzyme signature:
- P-[LIVM]-C-T-[LIVM]-[KRH]-x-[FT]-P

Patterns: conclusions
[RK]-x(3)-[DE]-x(2)-Y

- Patterns are appropriate for models of short sequence signatures.
- Pros:
  - Pattern matching is fast and easy to implement.
  - Models are easy to design & understand
- Cons:
  - Poor model for insertions/deletions (indels).
  - Small patterns find a lot of false positives. Long patterns are very difficult to design.
  - Poor predictors that tend to recognize only the sequence of the training set.
  - No scoring system, only binary response (YES/NO).
- When I use patterns?
  - To search for small signatures or active sites.
  - To communicate with others

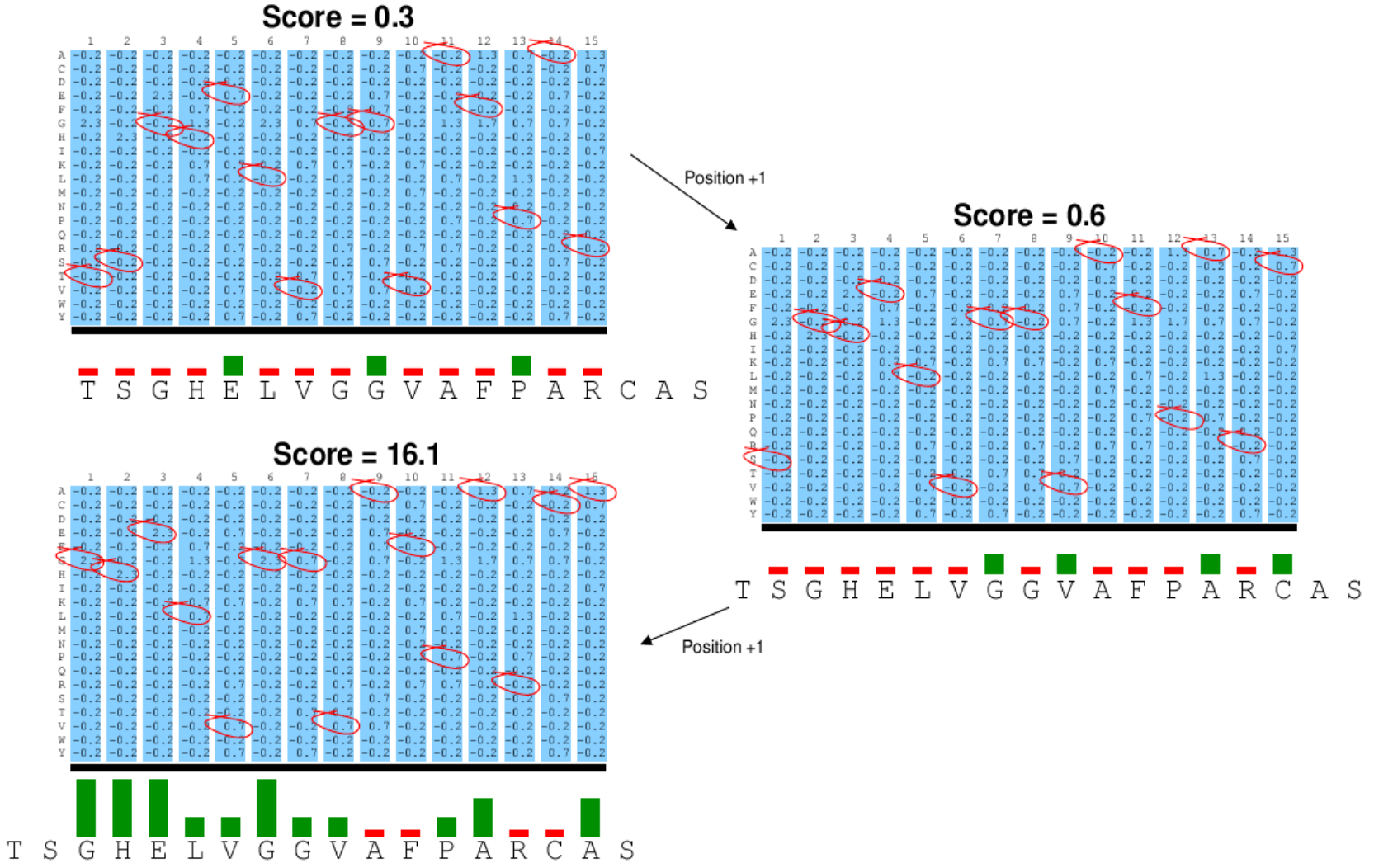
Position Specific Scoring Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0	0	0	0	0	0	0	0	0	0	0	2	1	0	2
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0	0	1	0
F	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3	1	1	0
H	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0
L	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
T	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0

Uses the *frequencies* of each residue in a specific position of a multiple alignment.

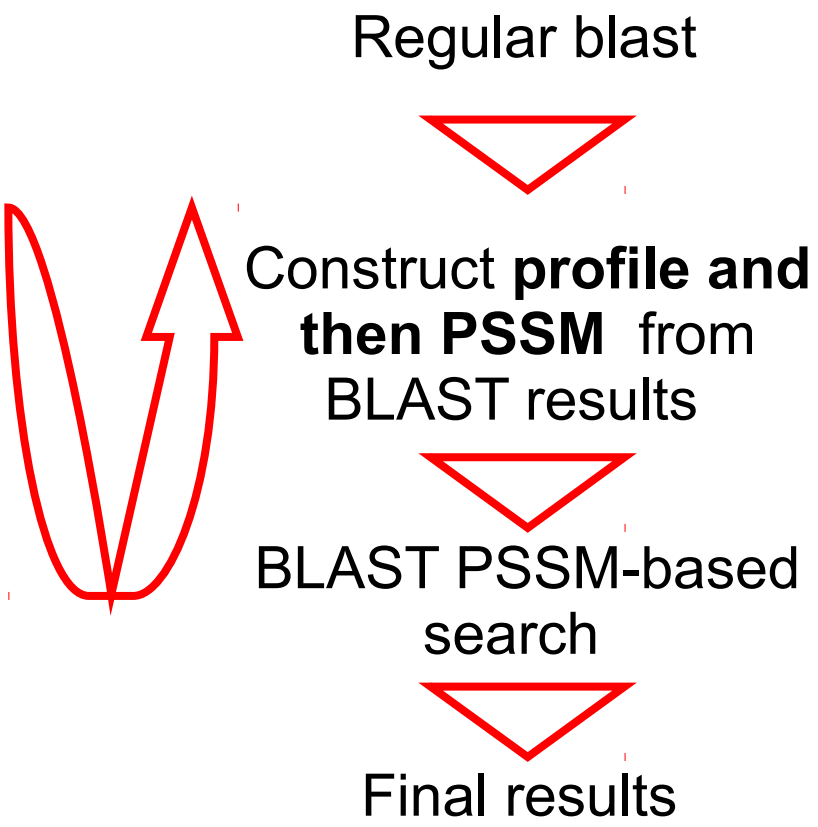
Position Specific Scoring Matrix

- Create a PSSM from “other” sequences
- Sliding window: slide across a target sequence, note high scoring matches

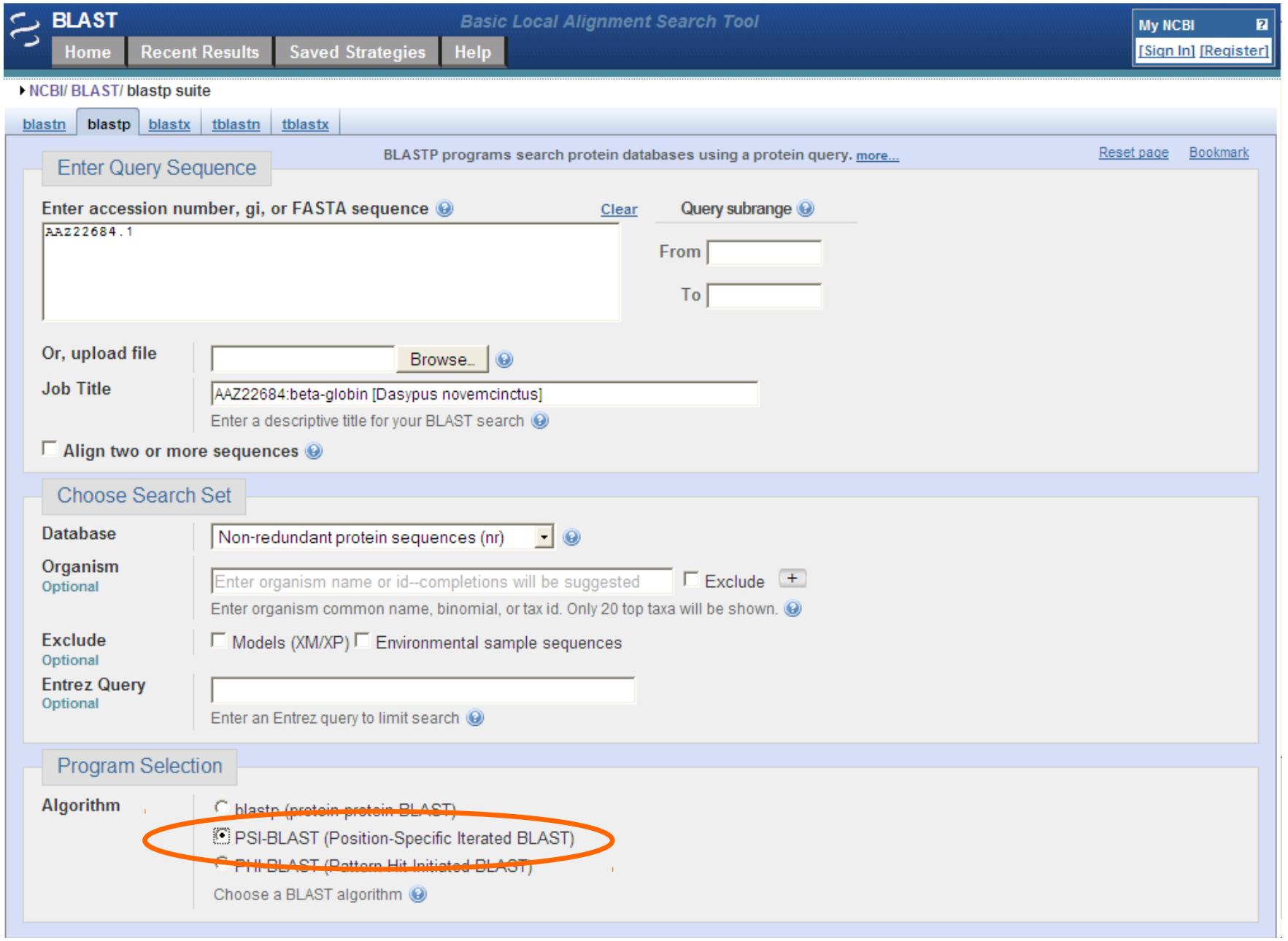


PSI-BLAST

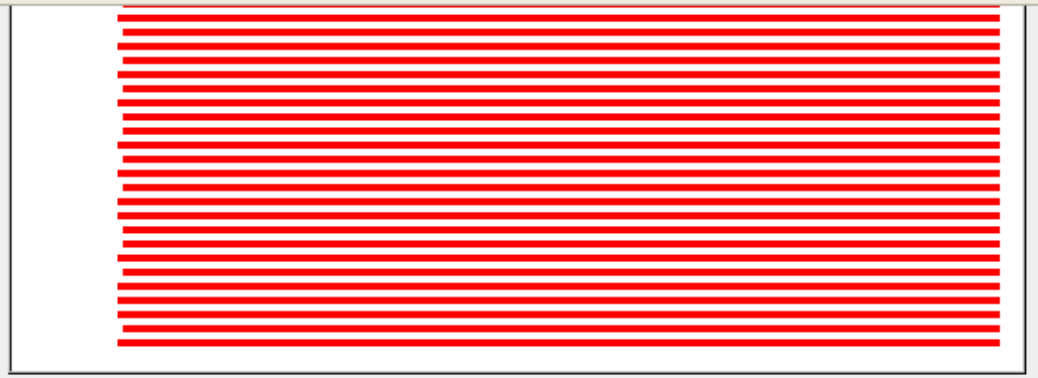
- Position Specific Iterative BLAST



BLAST – PSI-BLAST



PSI-BLAST - results



Descriptions

NEW - alignment score below the threshold on the previous iteration

- alignment was checked on the previous iteration

Run PSI-Blast iteration 2 with ma: 500 Go

Sequences with E-value BETTER than threshold

		Sequences producing significant alignments:	Score (Bits)	E Value	
NEW	☑	sp|P02087.3|HBB_DASNO RecName: Full=Hemoglobin subunit beta; ...	301	1e-80	
NEW	☑	gb|ACO88975.1| hemoglobin beta chain complex (predicted) [Das...	290	4e-77	
NEW	☑	sp|P02051.1|HBB_TARBA RecName: Full=Hemoglobin subunit beta; ...	247	3e-64	
NEW	☑	sp|P13557.2|HBB_TARSY RecName: Full=Hemoglobin subunit beta; ...	247	3e-64	
NEW	☑	ref|NP_001075729.1| hemoglobin, beta [Oryctolagus cuniculus] ...	241	2e-62	UG
NEW	☑	sp|P08535.2|HBB_LEPEU RecName: Full=Hemoglobin subunit beta; ...	241	3e-62	
NEW	☑	sp|P02036.2|HBB_SAISC RecName: Full=Hemoglobin subunit beta; ...	240	3e-62	
NEW	☑	sp|Q9TSP1.2|HBB_PAPAN RecName: Full=Hemoglobin subunit beta; ...	239	4e-62	
NEW	☑	sp|P20855.1|HBB_CTEGU RecName: Full=Hemoglobin subunit beta; ...	239	7e-62	
NEW	☑	sp|P02044.3|HBD_ATEGE RecName: Full=Hemoglobin subunit delta;...	239	9e-62	
NEW	☑	pdb|2RAO|B Chain B, X Ray Crystal Structure Of Rabbit Hemoglo...	238	1e-61	S
NEW	☑	sp|Q6WN29.3|HBB_ALOBE RecName: Full=Hemoglobin subunit beta; ...	238	1e-61	
NEW	☑	ref|NP_001157900.1| globin, beta [Macaca mulatta] >gb|ACS9404...	238	2e-61	UG
NEW	☑	sp|P19885.2|HBB_COLPO RecName: Full=Hemoglobin subunit beta; ...	238	2e-61	
NEW	☑	sp|P14526.3|HBB_BRATR RecName: Full=Hemoglobin subunit beta; ...	238	2e-61	

PSI-BLAST

- Increases *sensitivity*
- May decrease *specificity.* If we get an unrelated hit, we will get to unrelated sequences (contamination). This gets worse with each iteration. Also known as Drift
- Rule of thumb: **Drift Happens**. Within 4-7 iterations.


DRIFT HAPPENS