# BIMM 143

## Advanced Database Searching
### Lecture 3

**Barry Grant**

UC San Diego

http://thegrantlab.org/bimm143
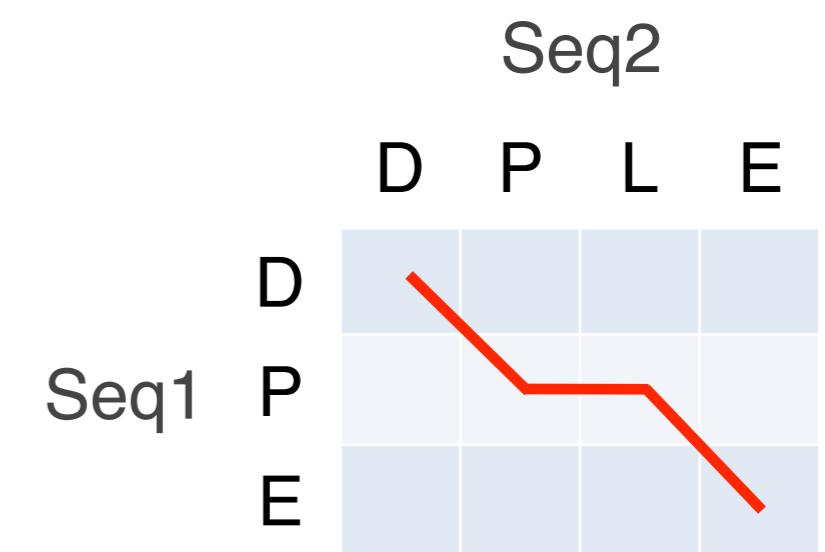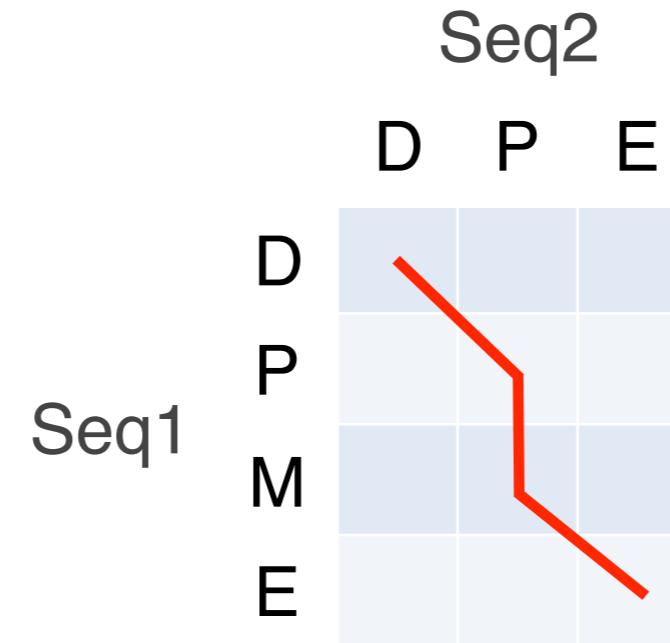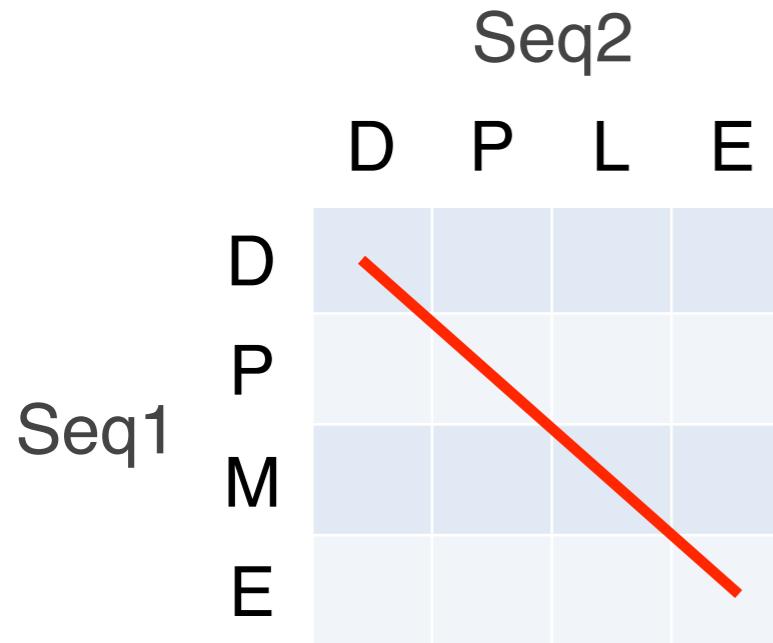
# Recap From Last Time:

- Sequence alignment is a fundamental operation underlying much of bioinformatics.

- Introduced dot matrices, dynamic programing and the BLAST heuristic approaches.

  ➡ *Key point*: Even when optimal solutions can be obtained they are not necessarily unique or reflective of the biologically correct alignment.

- Introduced classic global and local alignment algorithms (Needleman–Wunsch and Smith–Waterman) and their major application areas.

- Heuristic approaches are necessary for large database searches and many genomic applications.

**Feedback**

**HW1 Results**

# Muddy Point: Different paths represent different alignments



```
Seq1:  D P L E          Seq1:  D P M E          Seq1:  D P - E
       | | : |                 | | |                  | | |
Seq2:  D P M E          Seq2:  D P - E          Seq2:  D P L E
```

(Mis)matches are represented by <u>diagonal paths</u> &
Indels with <u>horizontal or vertical path</u> segments

# UC San Diego

## BIMM 143

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD ↗.

**Overview**

**Lectures**

**Computer Setup**

**Learning Goals**

**Assignments & Grading**

**Ethics Code**

# 3. Advanced sequence alignment and database searching

**Topics**:

Detecting remote sequence similarity, Database searching beyond BLAST, PSI-BLAST, Profiles and HMMs, Protein structure comparisons, Beginning with command line based database searches.

**Goal**:

- Be able to calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix
- Understand the limits of homology detection with tools such as BLAST
- Be able to perform PSI-BLAST, HMMER and protein structure based database searches and interpret the results in terms of the biological significance of an e-value.
- Run our first bioinformatics tool from the command line.

**Material**:

- Lecture Slides: Large PDF ↗, Small PDF ↗,
- Lab: Hands-on section worksheet ↗,
- Bonus: Alignment App ↗,
- Feedback: Muddy Point Assessment ↗.

**Homework**:

matrix values, **trace back alignment path** (in red), and overall **alignment score** change.

▶ **Details**:

Sequence 1　GATTAC

Sequence 2　GTCGACGC

| G | T | C | G | A | C | G | C |
| G | A | T | T | A | C | - | - |

Score = -4

**Match Score**　**Mismatch Score**　**Gap Score**

1　　　　-1　　　　　-2

Compute Optimal Alignment　Clear Path　Custom Path

|  |  | G | T | C | G | A | C | G | C |
|---|---|---|---|---|---|---|---|---|---|
|  |  | 0 | -2 | -4 | -6 | -8 | -10 | -12 | -14 | -16 |
| G | -2 | 1 | -1 | -3 | -5 |  |  |  |  |
| A | -4 | -1 | 0 | -2 | -4 |  |  |  |  |
| T | -6 | -3 | 0 | -1 | -3 | -5 | -5 | -7 | -9 |
| T | -8 | -5 | -2 | -1 | -2 | -4 | -6 | -6 | -8 |
| A | -10 | -7 | -4 | -3 | -2 | -1 | -3 | -5 | -7 |
| C | -12 | -9 | -6 | -3 | -4 | -3 | 0 | -2 | -4 |

**Score from Diagonal cell**
-6 + 1 (Due to a match between G & G) = -5

**Score from Upper cell**
-8 + -2 (The Gap score) = -10

**Score from Side cell**
-3 + -2 (The Gap score) = -5

Winning (max) score is -5

▼ **Reference**:
See the lecture and hands-on session for class 2 for a full discussion of Global, Local, and various Heuristic approaches to biomolecular sequence alignment.
Barry J Grant.

# Todays Menu

- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns

- **Sequence profiles** and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations

- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity

- **Hidden Markov models** (HMMs): More versatile probabilistic model for detection of remote similarities

Side Note:

**Q.** Where do our alignment match and mis-match scores typically come from?

# By default BLASTp match scores come from the **BLOSUM62** matrix

| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **C** | 9 | | | | | | | | | | | | | | | | | | | |
| **S** | -1 | 4 | | | | | | | | | | | | | | | | | | |
| **T** | -1 | 1 | 5 | | | | | | | | | | | | | | | | | |
| **P** | -3 | -1 | -1 | 7 | | | | | | | | | | | | | | | | |
| **A** | 0 | 1 | 0 | -1 | 4 | | | | | | | | | | | | | | | |
| **G** | -3 | 0 | -2 | -2 | 0 | 6 | | | | | | | | | | | | | | |
| **N** | -3 | 1 | 0 | -2 | -2 | 0 | 6 | | | | | | | | | | | | | |
| **D** | -3 | 0 | -1 | -1 | -2 | -1 | 1 | 6 | | | | | | | | | | | | |
| **E** | -4 | 0 | -1 | -1 | -1 | -2 | 0 | 2 | 5 | | | | | | | | | | | |
| **Q** | -3 | 0 | -1 | -1 | -1 | -2 | 0 | 0 | 2 | 5 | | | | | | | | | | |
| **H** | -3 | -1 | -2 | -2 | -2 | -2 | 1 | -1 | 0 | 0 | 8 | | | | | | | | | |
| **R** | -3 | -1 | -1 | -2 | -1 | -2 | 0 | -2 | 0 | 1 | 0 | 5 | | | | | | | | |
| **K** | -3 | 0 | -1 | -1 | -1 | -2 | 0 | -1 | 1 | 1 | -1 | 2 | 5 | | | | | | | |
| **M** | -1 | -1 | -1 | -2 | -1 | -3 | -2 | -3 | -2 | 0 | -2 | -1 | -1 | 5 | | | | | | |
| **I** | -1 | -2 | -1 | -3 | -1 | -4 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | 1 | 4 | | | | | |
| **L** | -1 | -2 | -1 | -3 | -1 | -4 | -3 | -4 | -3 | -2 | -3 | -2 | -2 | 2 | 2 | 4 | | | | |
| **V** | -1 | -2 | 0 | -2 | 0 | -3 | -3 | -3 | -2 | -2 | -3 | -3 | -2 | 1 | 3 | 1 | 4 | | | |
| **F** | -2 | -2 | -2 | -4 | -2 | -3 | -3 | -3 | -3 | -3 | -1 | -3 | -3 | 0 | 0 | 0 | -1 | 6 | | |
| **Y** | -2 | -2 | -2 | -3 | -2 | -3 | -2 | -3 | -2 | -1 | 2 | -2 | -2 | -1 | -1 | -1 | -1 | 3 | 7 | |
| **W** | -2 | -3 | -2 | -4 | -3 | -2 | -4 | -4 | -3 | -2 | -2 | -3 | -3 | -1 | -3 | -2 | -3 | 1 | 2 | 11 |

**Blo**cks **Su**bstitution **M**atrix. Scores obtained from observed frequencies of substitutions in blocks of aligned sequences with no more than 62% identity.
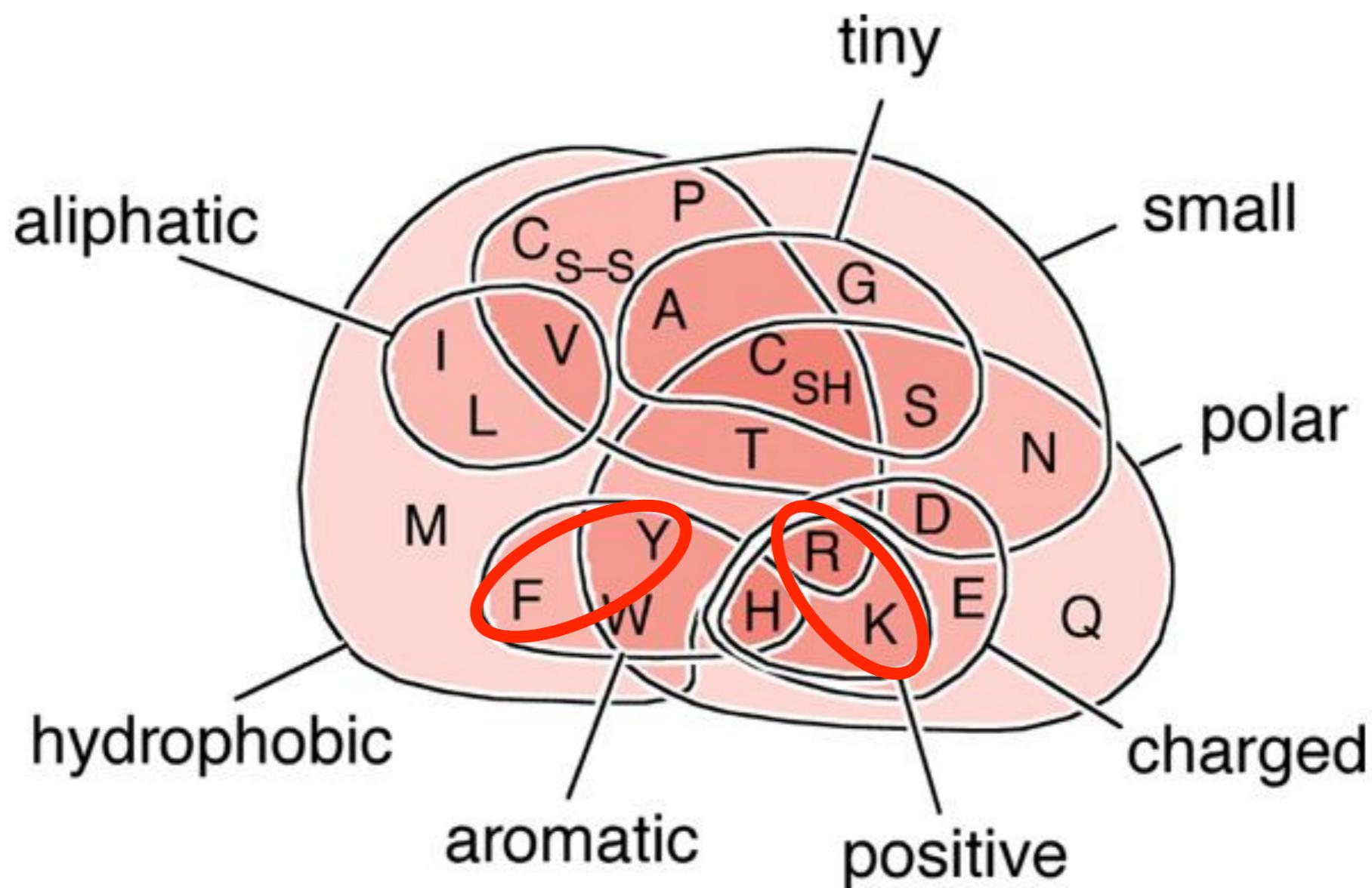
**Note**. Some amino acid mismatches have positive scores (highlighted in red) reflecting the shared physicochemical properties of these amino acids

Not all matches score equally (blue highlighted values)

# Protein scoring matrices reflect the properties of amino acids

# Protein scoring matrices reflect the properties of amino acids



**Key Trend**: High scores for amino acids in the same "biochemical group" and low scores for amino acids from different groups.

**N.B.** BLOUSM62 does not take the local context of a particular position into account

(*i.e.* all like substitutions are scored the same regardless of their location in the molecules).

**We will revisit this later…**

# Todays Menu
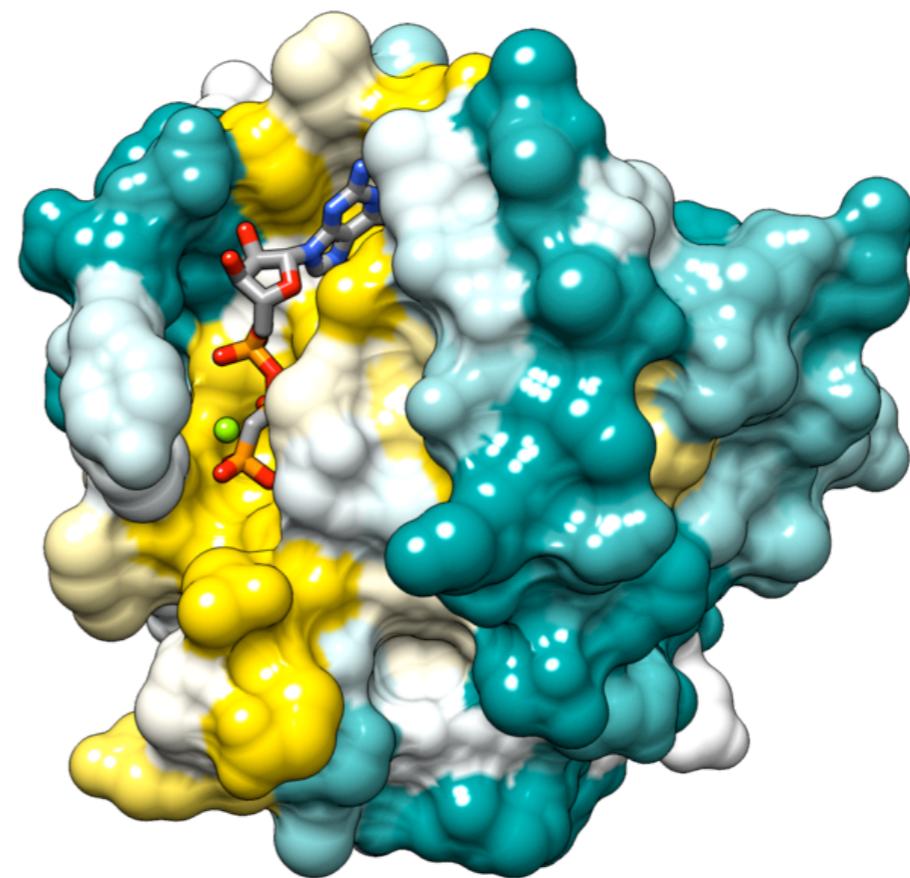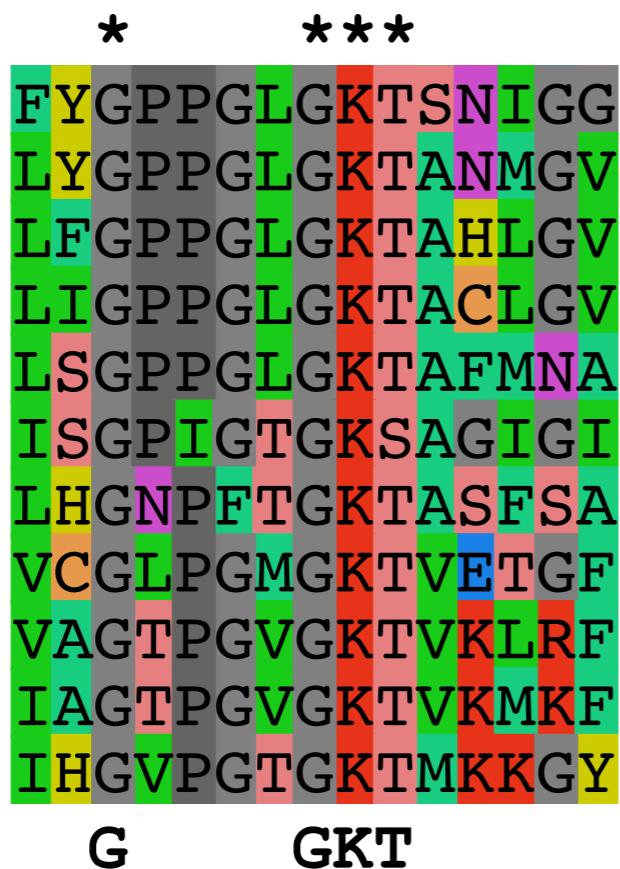
- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns

- Sequence profiles and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations

- PSI-BLAST algorithm: Application of iterative PSSM searching to improve BLAST sensitivity

- Hidden Markov models (HMMs): More versatile probabilistic model for detection of remote similarities

# Functional cues from conservation patterns

Within a protein or nucleic acid sequence there may be a small number of characteristic residues that occur consistently. These conserved "sequence fingerprints" (or **motifs**) usually contain functionally important elements

- E.g., the amino acids that are consistently found at enzyme active sites or the nucleotides that are associated with transcription factor binding sites.

ATP/GTP-binding proteins: G-x(4)-G-K-T
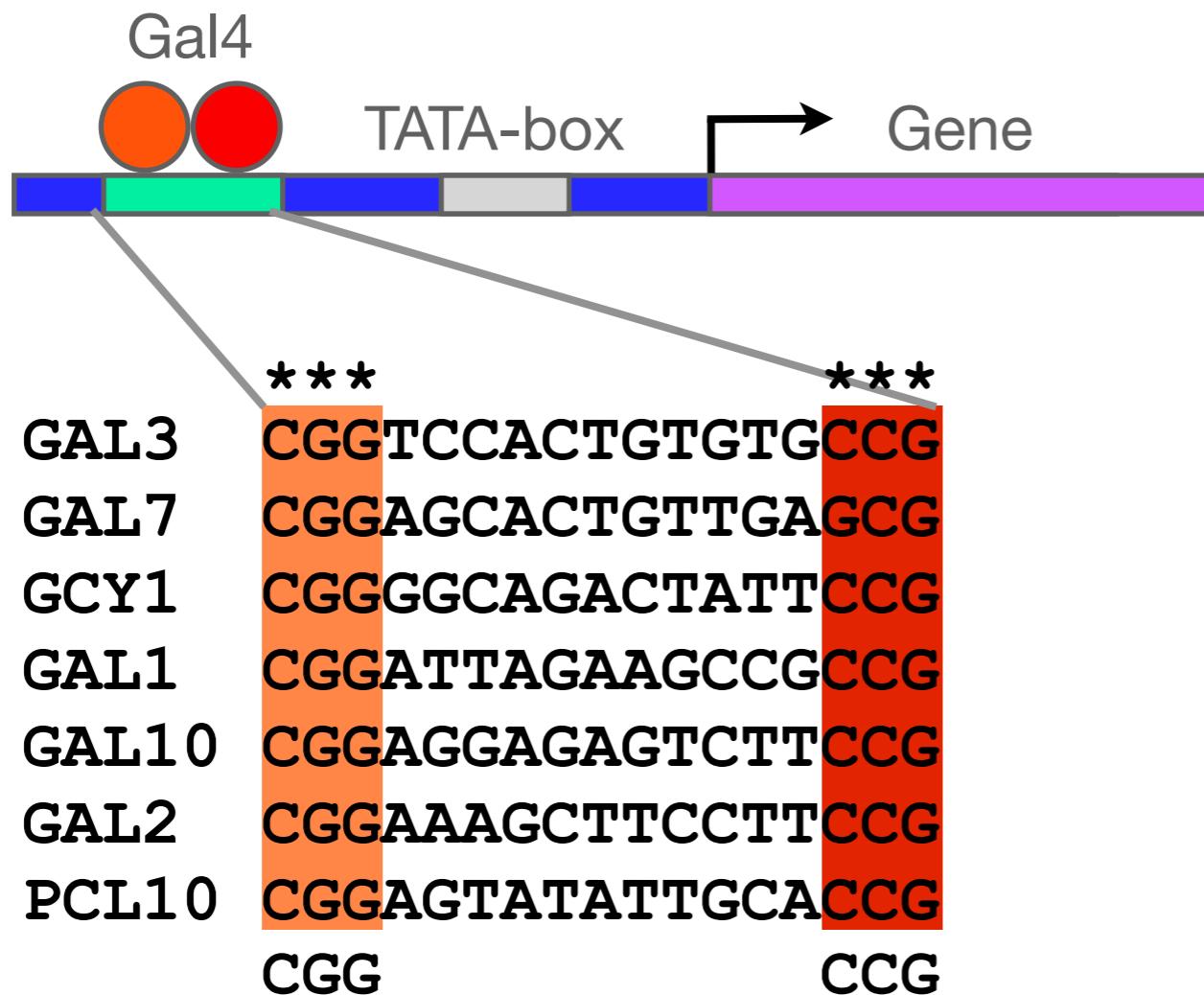


Conservation

# Functional cues from conservation patterns...

Many DNA patterns are binding sites for Transcription Factors.

- E.g., The Gal4 binding sequence
  C-G-G-N(11)-C-C-G



```
            ***            ***
GAL3    CGGTCCACTGTGTGCCG
GAL7    CGGAGCACTGTTGAGCG
GCY1    CGGGGCAGACTATTCCG
GAL1    CGGATTAGAAGCCGCCG
GAL10   CGGAGGAGAGTCTTCCG
GAL2    CGGAAAGCTTCCTTCCG
PCL10   CGGAGTATATTGCACCG
        CGG            CCG
```

# Representing recurrent sequence patterns

Beyond knowledge of invariant residues we can define **position-based** representations that highlight the range of permissible residues per position.

- **Pattern**: Describes a motif using a qualitative consensus sequence (e.g., IUPAC or regular expression). N.B. Mismatches are not tolerated!

  [LFI]-x-G-[PT]-P-G-x-G-K-[TS]-[AGSI]

- **Profile**: Describes a motif using quantitative information captured in a position specific scoring matrix (weight matrix).
  Profiles quantify similarity and often span larger stretches of sequence.
  More on this shortly...

- **Logos**: A useful visual representation of sequence motifs.
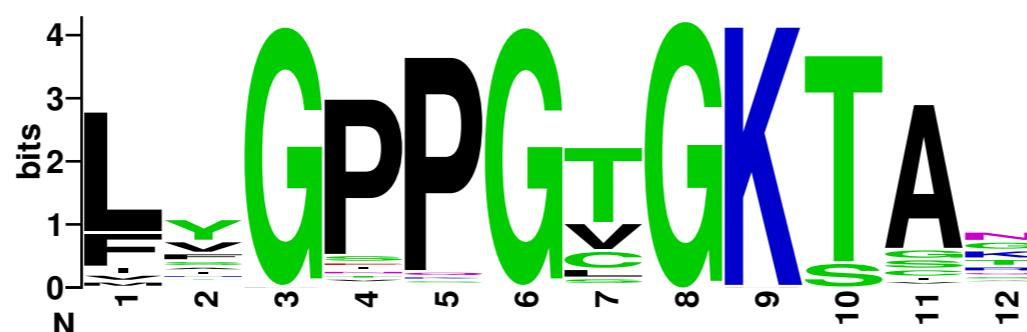


Image generated by:
weblogo.berkeley.edu

# PROSITE is a popular protein pattern and profile database

Currently contains > 1790 patterns and profiles: http://prosite.expasy.org/
Example PROSITE patterns:

> PS00087; SOD_CU_ZN_1
> [GA]-[IMFAT]-H-[LIVF]-H-{S}-x-[GP]-[SDG]-x(2)-[STAGDE]
> The two **H**istidines coordinate important copper ligands
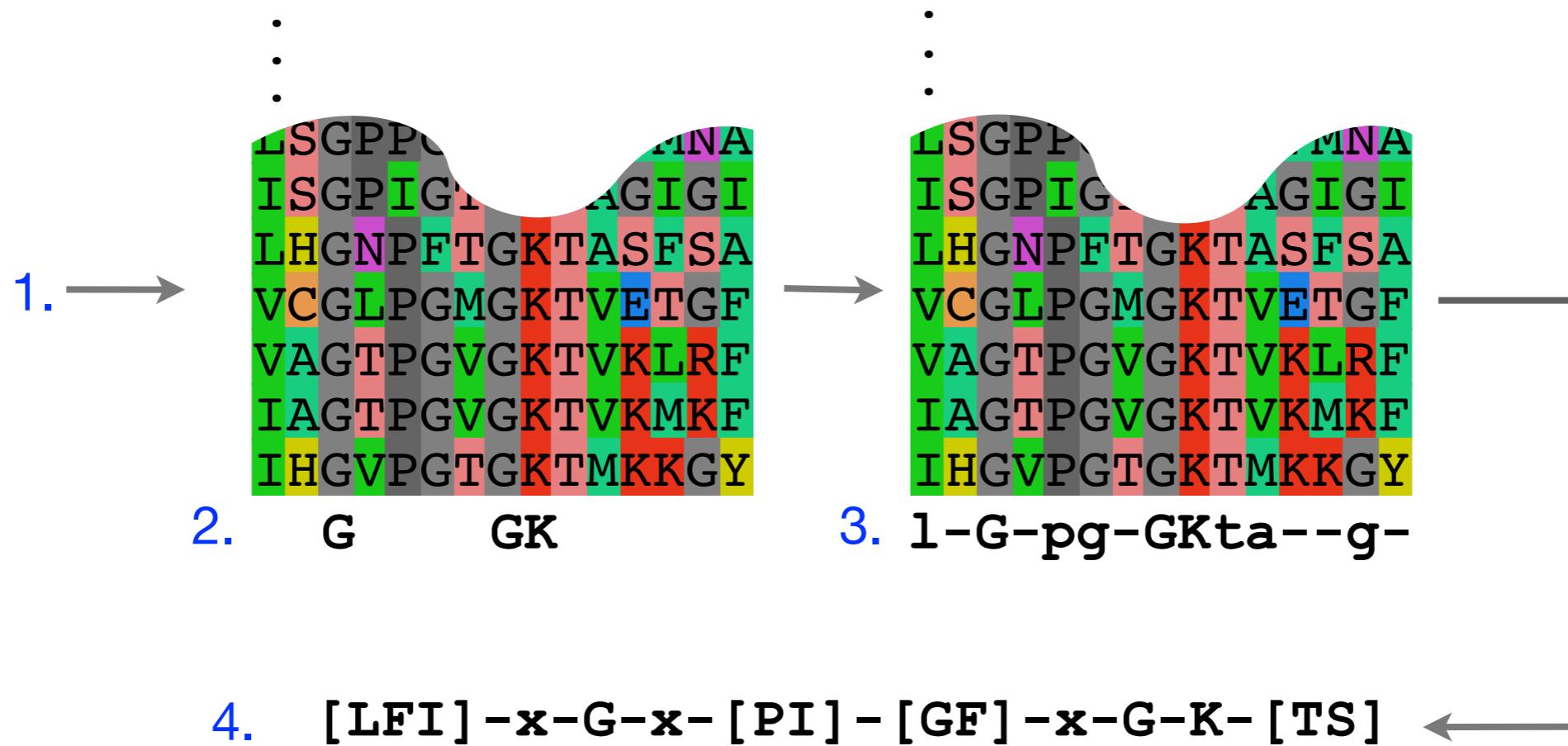
- Each position in the pattern is separated with a hyphen

- x can match any residue

- [ ] are used to indicate ambiguous positions in the pattern
  e.g., [SDG] means the pattern can match S, D, or G at this position

- { } are used to indicate residues that are not allowed at this position
  e.g., {S} means NOT S (not Serine)

- ( ) surround repeated residues, e.g., A(3) means AAA

Information from http://ca.expasy.org/prosite/prosuser.html

# Defining sequence patterns

There are four basic steps involved in defining a new PROSITE style pattern:
1. Construct a multiple sequence alignment (MSA)
2. Identify conserved residues
3. Create a core sequence pattern (i.e. *consensus sequence*)
4. Expand the pattern to improve **sensitivity** and **specificity** for detecting desired sequences - more on this shortly...



2. **G        GK**

3. **l-G-pg-GKta--g-**

4. **[LFI]-x-G-x-[PI]-[GF]-x-G-K-[TS]**

# Pattern advantages and disadvantages

**Advantages:**
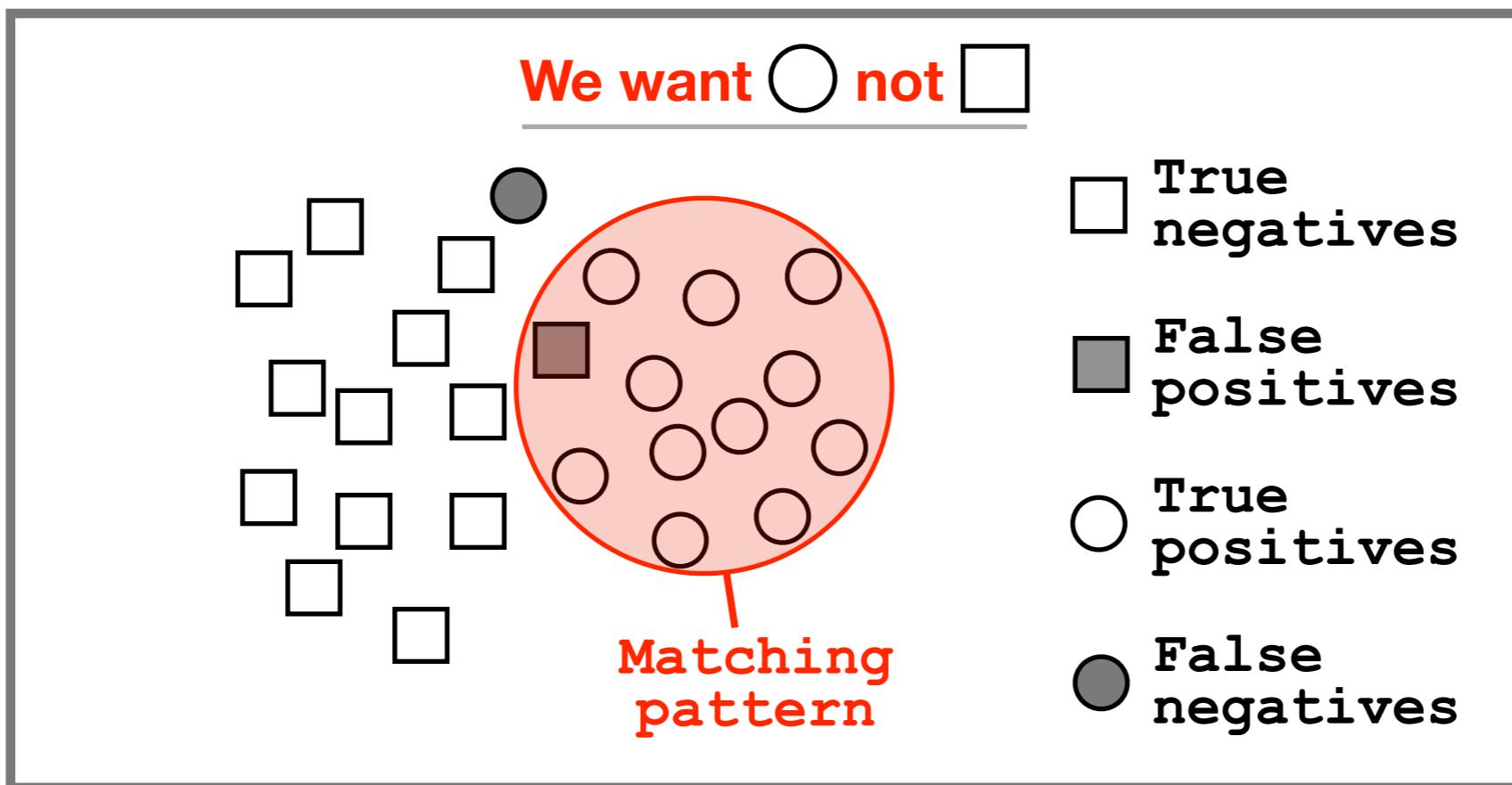
- Relatively straightforward to identify (exact pattern matching is fast)

- Patterns are intuitive to read and understand

- Databases with large numbers of protein (e.g., PROSITE) and DNA sequence (e.g., JASPER and TRANSFAC) patterns are available.

**Disadvantages:**

- Patterns are qualitative and *deterministic*
  (i.e., either matching or not!)

- We lose information about relative frequency of each residue at a position
  E.g., [GAC] *vs* 0.6 G, 0.28 A, and 0.12 C

- Can be difficult to write complex motifs using regular expression notation

- Cannot represent subtle sequence motifs

# Side note: pattern sensitivity and specificity
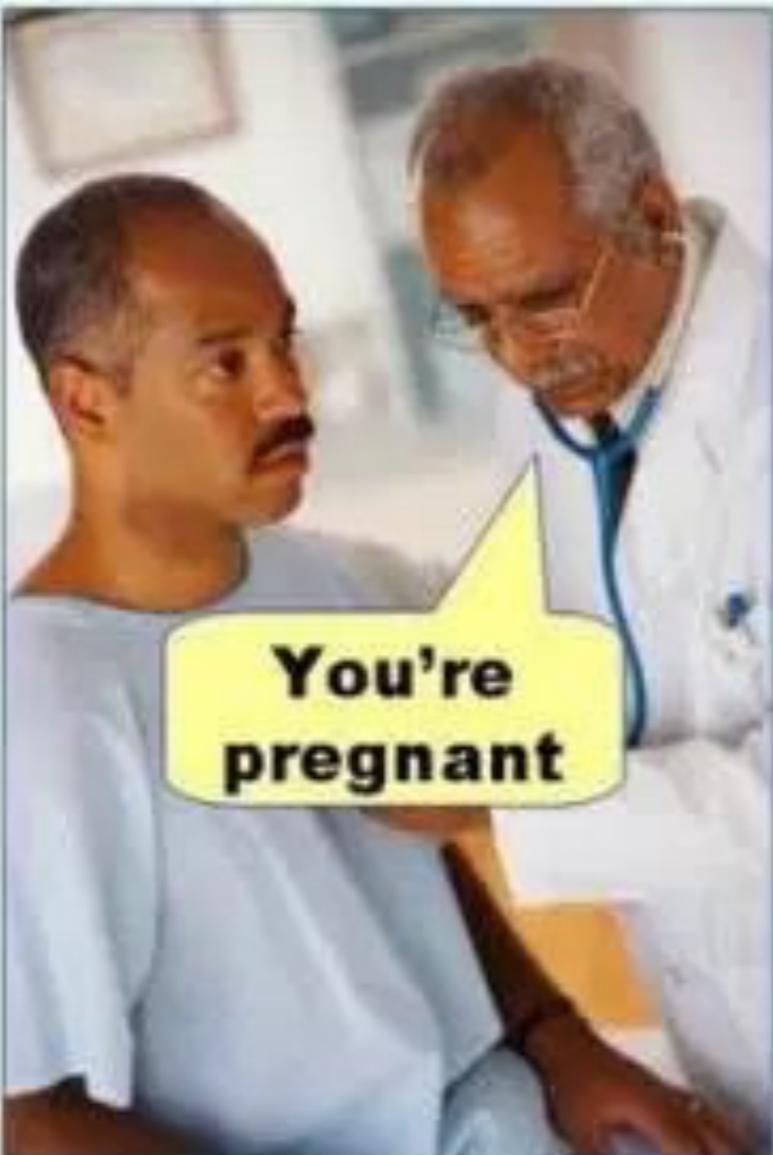
In practice it is not always possible to define one single regular expression type pattern which matches all family sequences (*true positives* ○) while avoiding matches in unrelated sequences (*true negatives* □).



We want ○ not □

□ True negatives

■ False positives

○ True positives

● False negatives

Matching pattern
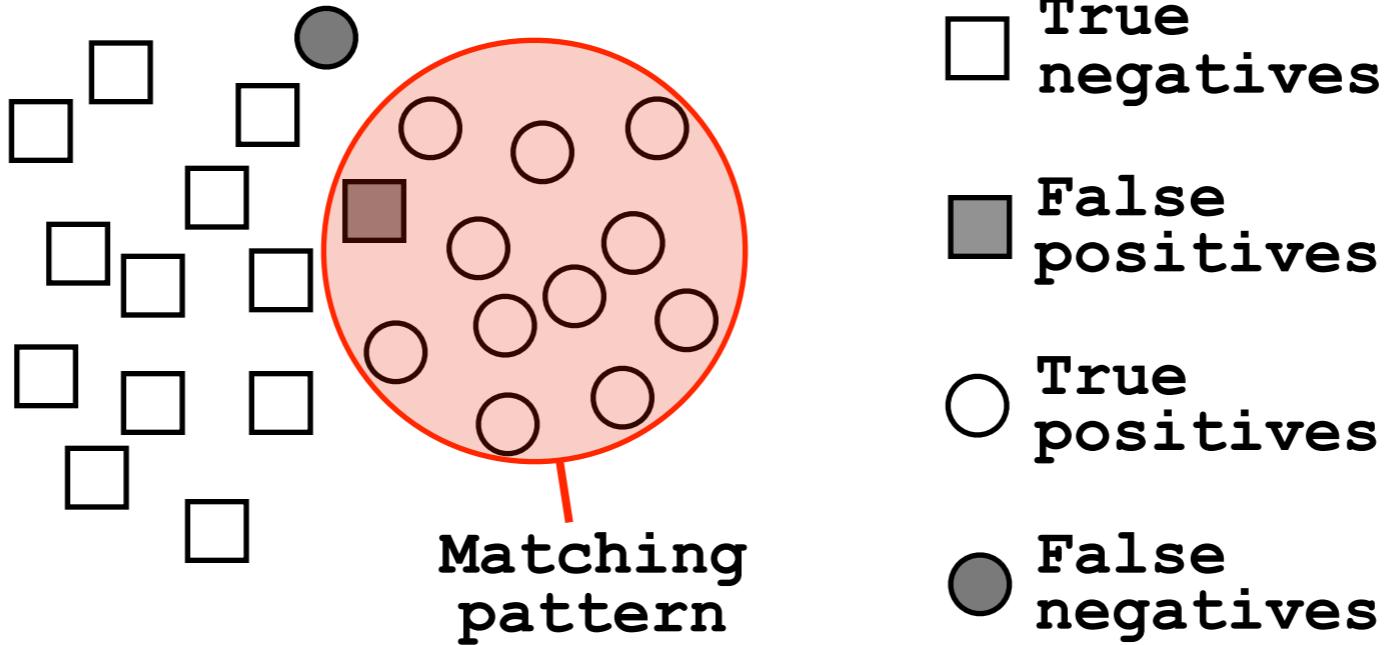
Sensitivity = TP/(TP+FN)    ○/(○ + ●)

Specificity = TN/(TN+FP)    □/(□ + ■)

# Side note: pattern sensitivity, specificity, and PPV



□ **True negatives**

■ **False positives**

○ **True positives**

● **False negatives**

**Matching pattern**
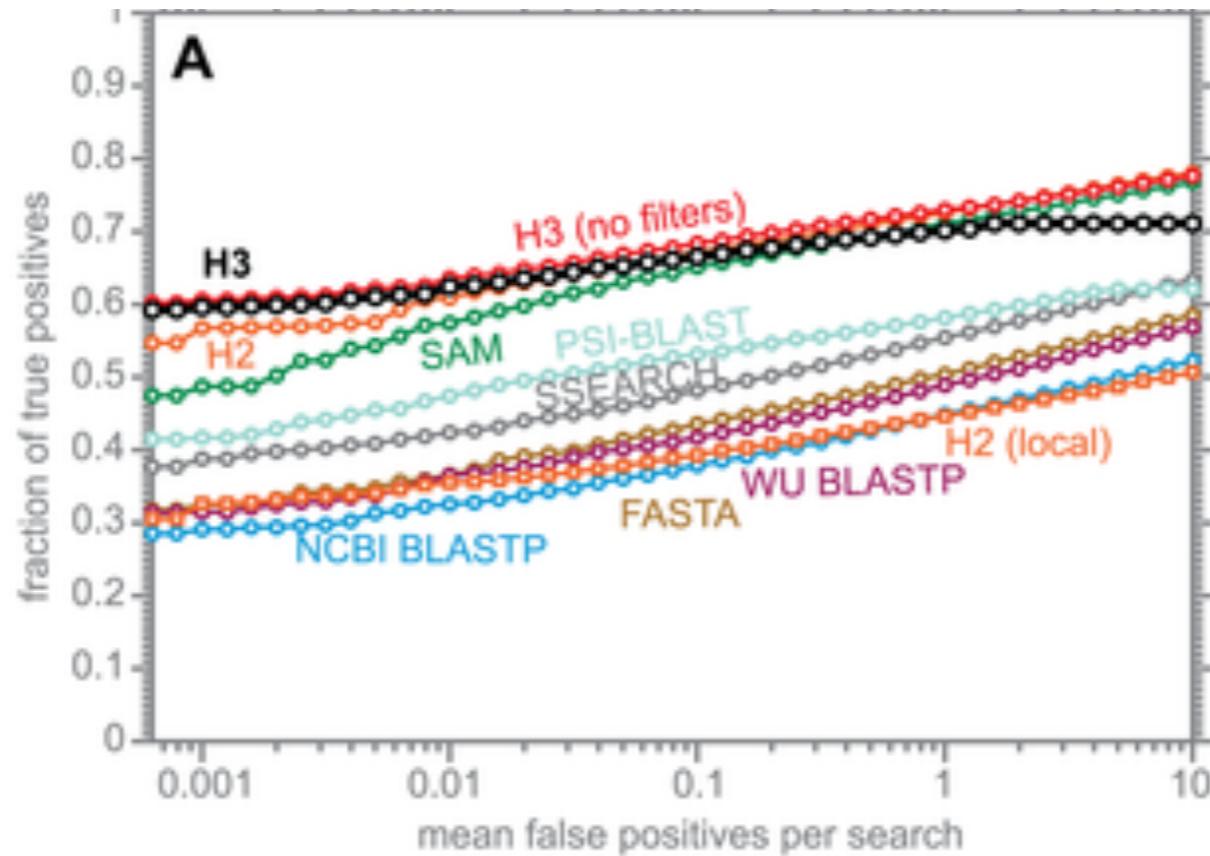
**Sensitivity = TP/(TP+FN)** = Fraction of total circles we found
(i.e. things we want!) ○**/(**○**+**●**)**

**Specificity = TN/(TN+FP)** = Fraction of total squares we missed
(i.e. things we **don't** want!) □**/(**□**+**■**)**

**PPV = TP/(TP+FP)** = Fraction of our highlighted matches that are actually circles
(i.e. proportion of the things we found that are what we want!)
○**/(**○**+**■**)**                    ROC plot example

# ROC plot of sequence searching performance...



H3 (HMMER3) has a much higher search sensitivity and specificity than BLASTp

In each benchmark, true positive subsequences have been selected to be no more than 25% identical to any sequence in the query alignment ... (see paper for details).

See: Eddy (2011) PLoS Comp Biol 7(10): e1002195

# Todays Menu
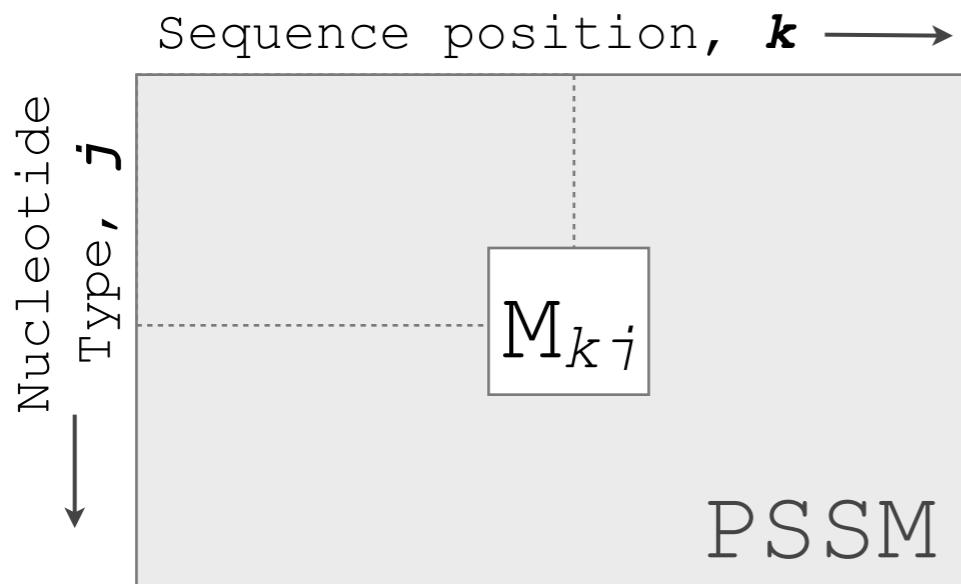
- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns

- **Sequence profiles** and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations

- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity

- **Hidden Markov models** (HMMs): More versatile probabilistic model for detection of remote similarities

# Sequence profiles

A sequence profile is a **position-specific scoring matrix** (or **PSSM**, often pronounced 'possum') that gives a *quantitative* description of a sequence motif.

Unlike deterministic patterns, profiles assign a score to a query sequence and are widely used for database searching.
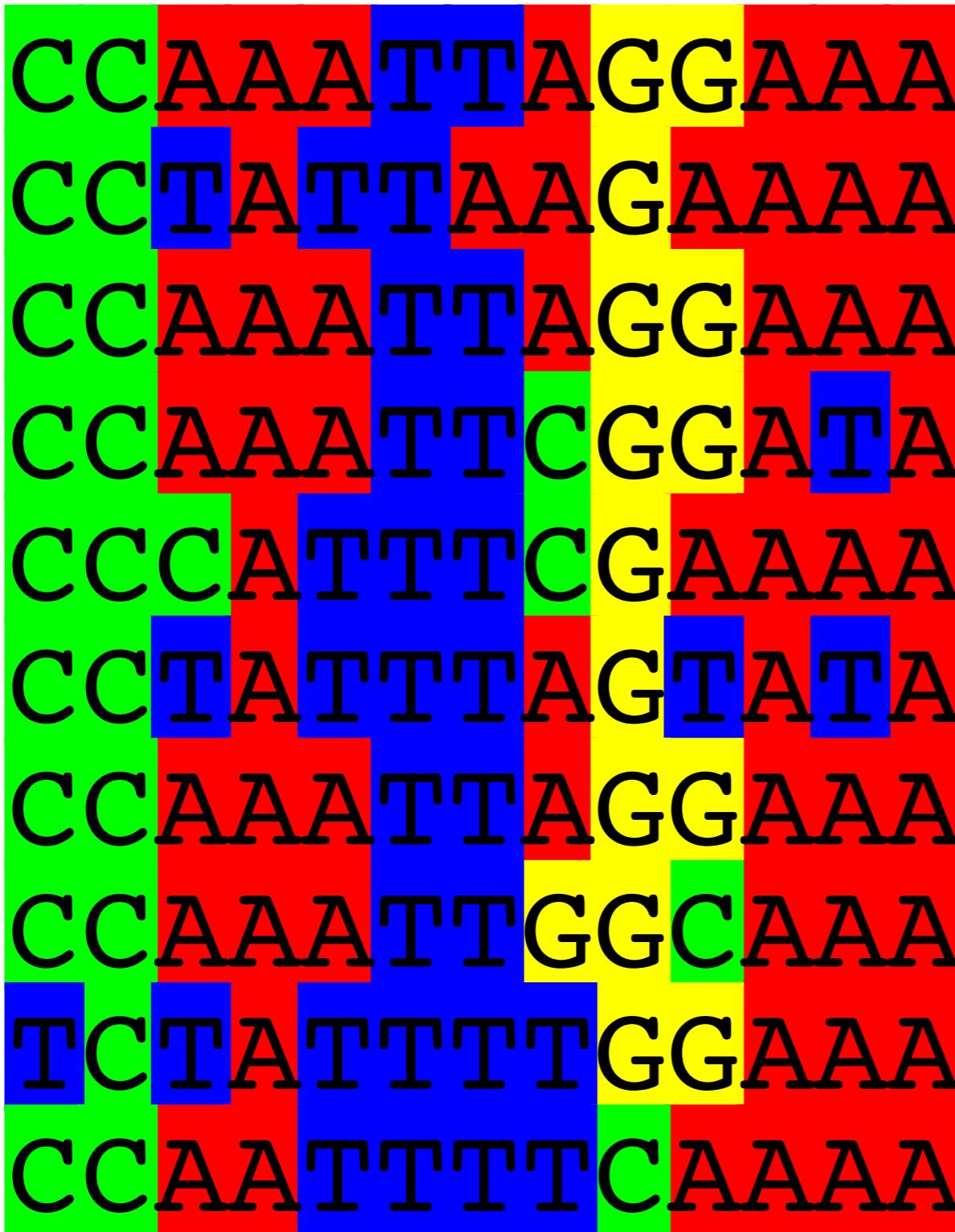
A simple PSSM has as many columns as there are positions in the alignment, and either 4 rows (one for each DNA nucleotide) or 20 rows (one for each amino acid).

Sequence position, $k$ $\longrightarrow$

Nucleotide Type, $j$

$M_{kj}$

PSSM

$$M_{kj} = \log\left(\frac{p_{kj}}{p_j}\right)$$

$M_{kj}$  score for the *j*th nucleotide at position *k*

$p_{kj}$  probability of nucleotide *j* at position *k*

$p_j$  "background" probability of nucleotide *j*

See Gibskov *et al.* (1987) PNAS 84, 4355

# Computing a transcription factor bind site PSSM



Here we have **10 aligned** transcription factor binding site nucleotide sequences

That span **13 positions** (i.e. columns of nucleotides).

We will build a 13 x 4 **PSSM** ($k$=13, $j$=4).

# Computing a transcription factor bind site PSSM



First we will build an alignment **Counts matrix**

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | | | | | | | | | | | | | |
| C: | | | | | | | | | | | | | |
| G: | | | | | | | | | | | | | |
| T: | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |

# Computing a transcription factor bind site PSSM



**Alignment Counts matrix:**

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | | | | | | | | | | | | | |
| C: | | | | | | | | | | | | | |
| G: | | | | | | | | | | | | | |
| T: | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |

Position k = 1

# Computing a transcription factor bind site PSSM



**Alignment Counts matrix:**

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | 0 | | | | | | | | | | | | |
| C: | 9 | | | | | | | | | | | | |
| G: | 0 | | | | | | | | | | | | |
| T: | 1 | | | | | | | | | | | | |
| | | | | | | | | | | | | | |

**Position k = 1**

# Computing a transcription factor bind site PSSM



**Alignment Counts matrix:**

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A:** | 0 | | | | | | | | | | | | |
| **C:** | 9 | | | | | | | | | | | | |
| **G:** | 0 | | | | | | | | | | | | |
| **T:** | 1 | | | | | | | | | | | | |
| **Consensus** | C | | | | | | | | | | | | |

# Computing a transcription factor bind site PSSM
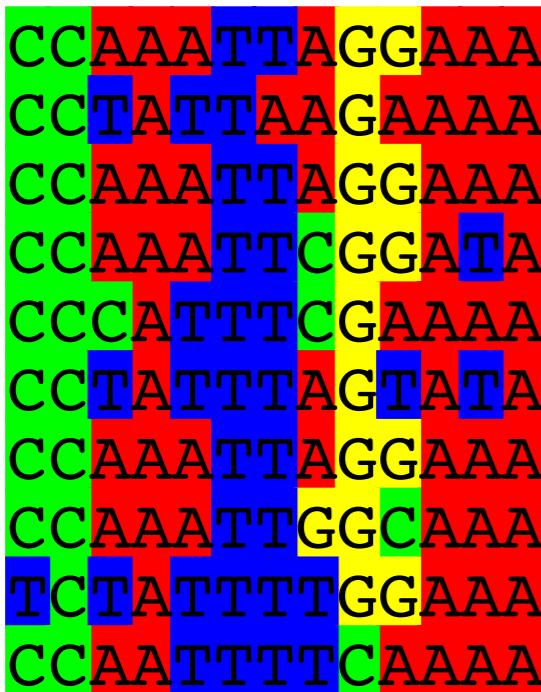


**Alignment Counts matrix:**

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---:|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A:** | 0 | 0 | | | | | | | | | | | |
| **C:** | 9 | 10 | | | | | | | | | | | |
| **G:** | 0 | 0 | | | | | | | | | | | |
| **T:** | 1 | 0 | | | | | | | | | | | |
| **Consensus** | C | C | | | | | | | | | | | |

**Position k = 2**

# Computing a transcription factor bind site PSSM
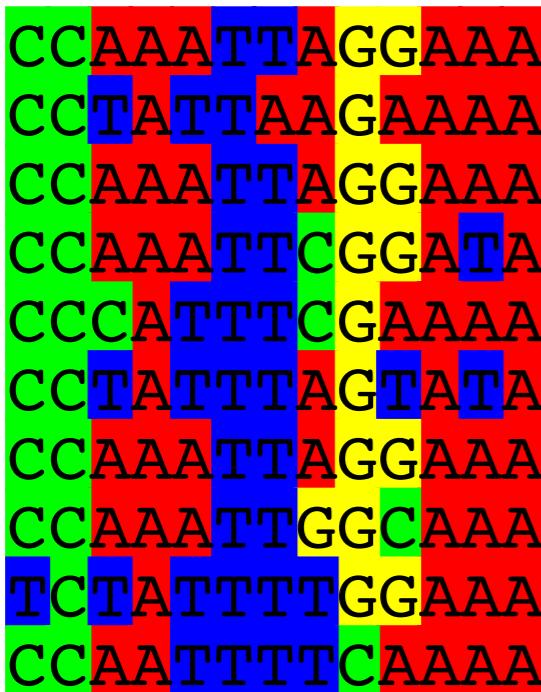


**Alignment Counts matrix:**

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---:|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A:** | 0 | 0 | 6 | | | | | | | | | | |
| **C:** | 9 | 10 | 1 | | | | | | | | | | |
| **G:** | 0 | 0 | 0 | | | | | | | | | | |
| **T:** | 1 | 0 | 3 | | | | | | | | | | |
| **Consensus** | C | C | [AT] | | | | | | | | | | |

Position k = 3
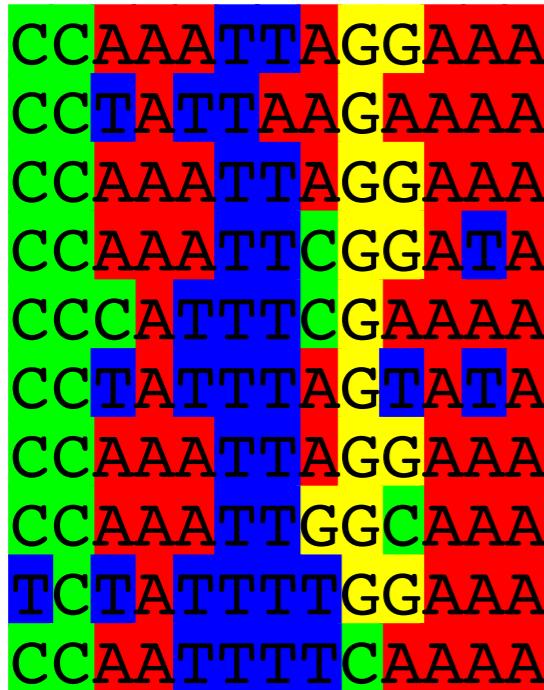
# Computing a transcription factor bind site PSSM



**Alignment Counts matrix:**

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A:** | 0 | 0 | 6 | 10 | 5 | 0 | 1 | 5 | 0 | 3 | 10 | 8 | 10 |
| **C:** | 9 | 10 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 |
| **G:** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 9 | 5 | 0 | 0 | 0 |
| **T:** | 1 | 0 | 3 | 0 | 5 | 10 | 9 | 2 | 0 | 1 | 0 | 2 | 0 |
| **Consensus** | C | C | [AT] | A | [AT] | T | T | [ACT] | G | [GA] | A | [AT] | A |

# Computing a transcription factor bind site PSSM

```
CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAATTTTCAAAA
```

## Alignment Counts matrix:

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A:** | 0 | 0 | 6 | 10 | 5 | 0 | 1 | 5 | 0 | 3 | 10 | 8 | 10 |
| **C:** | 9 | 10 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 |
| **G:** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 9 | 5 | 0 | 0 | 0 |
| **T:** | 1 | 0 | 3 | 0 | 5 | 10 | 9 | 2 | 0 | 1 | 0 | 2 | 0 |
| **Consensus** | C | C | [AT] | A | [AT] | T | T | [ACT] | G | [GA] | A | [AT] | A |

Often we will not communicate with the count matrix but rather the derived **average profile** (a.k.a. frequency matrix).
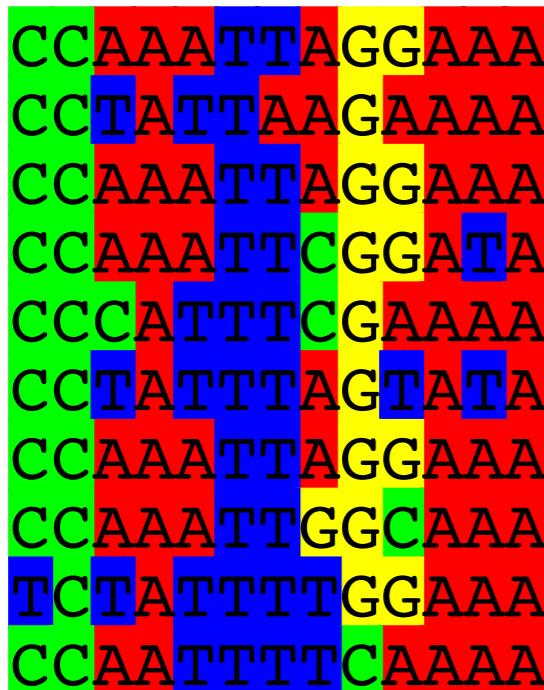
## Average Profile (Frequency) matrix:

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A:** | 0 | 0 | 0.6 | 1 | 0.5 | 0 | 0.1 | 0.5 | 0 | 0.3 | 1 | 0.8 | 1 |
| **C:** | 0.9 | 1 | 0.1 | 0 | 0 | 0 | 0 | 0.2 | 0.1 | 0.1 | 0 | 0 | 0 |
| **G:** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0.9 | 0.5 | 0 | 0 | 0 |
| **T:** | 0.1 | 0 | 0.3 | 0 | 0.5 | 1 | 0.9 | 0.2 | 0 | 0.1 | 0 | 0.2 | 0 |
| **Consensus** | C | C | [AT] | A | [AT] | T | T | [ACT] | G | [GA] | A | [AT] | A |

# Computing a transcription factor bind site PSSM



**Alignment Counts matrix:**

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | 0 | 0 | 6 | 10 | 5 | 0 | 1 | 5 | 0 | 3 | 10 | 8 | 10 |
| C: | 9 | 10 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 |
| G: | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 9 | 5 | 0 | 0 | 0 |
| T: | 1 | 0 | 3 | 0 | 5 | 10 | 9 | 2 | 0 | 1 | 0 | 2 | 0 |
| **Consensus** | C | C | [AT] | A | [AT] | T | T | [ACT] | G | [GA] | A | [AT] | A |

---

**Or the "score ($M_{kj}$) matrix" = PS$\underline{S}$M**

$C_{kj}$   Number of $j$th type nucleotide at position $k$

$Z$   Total number of aligned sequences

$P_j$   "background" probability of nucleotide $j$

$P_{kj}$   probability of nucleotide $j$ at position $k$

$$M_{kj} = \log\left(\frac{p_{kj}}{p_j}\right) \qquad p_{kj} = \frac{C_{kj} + p_j}{Z+1}$$

$$M_{kj} = \log\left(\frac{C_{kj} + p_j \, / \, Z + 1}{p_j}\right)$$

# Computing a transcription factor bind site PSSM...

Alignment Matrix: C_{kj}

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | 0 | 0 | 6 | 10 | 5 | 0 | 1 | 5 | 0 | 3 | 10 | 8 | 10 |
| C: | 9 | 10 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 |
| G: | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 9 | 5 | 0 | 0 | 0 |
| T: | 1 | 0 | 3 | 0 | 5 | 10 | 9 | 2 | 0 | 1 | 0 | 2 | 0 |

$$k=1, j=\text{A}: \quad M_{kj} = \log\left(\frac{C_{kj} + p_j / Z + 1}{p_j}\right) = \log\left(\frac{0 + 0.25 / 10 + 1}{0.25}\right) = -2.4$$

$$k=1, j=\text{C}: \quad M_{kj} = \log\left(\frac{C_{kj} + p_j / Z + 1}{p_j}\right) = \log\left(\frac{9 + 0.25 / 10 + 1}{p_j}\right)$$

$$M_{kj} = \log_e\left(\frac{p_{kj}}{p_j}\right) = \log_e\left(\frac{(C_{kj} + p_j) / (Z + 1)}{p_j}\right)$$

$$k=1, j=\text{T}: \quad M_{kj} = \log\left(\frac{C_{kj} + p_j / Z + 1}{p_j}\right) = \log\left(\frac{1 + 0.25 / 10 + 1}{0.25}\right) = -0.8$$

PSSM: M_{kj}

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | -2.4 | -2.4 | 0.8 | 1.3 | 0.6 | -2.4 | -0.8 | 0.6 | -2.4 | 0.2 | 1.3 | 1.1 | 1.3 |
| C: | 1.2 | 1.3 | -0.8 | -2.4 | -2.4 | -2.4 | -2.4 | -0.2 | -0.8 | -0.8 | -2.4 | -2.4 | -2.4 |
| G: | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -0.8 | 1.2 | 0.6 | -2.4 | -2.4 | -2.4 |
| T: | -0.8 | -2.4 | 0.2 | -2.4 | 0.6 | 1.3 | 1.2 | -0.2 | -2.4 | -0.8 | -2.4 | -0.2 | -2.4 |

# Scoring a test sequence

Query Sequence

CCTATTTAGGATA

PSSM:

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | -2.4 | -2.4 | 0.8 | **1.3** | 0.6 | -2.4 | -0.8 | **0.6** | -2.4 | 0.2 | **1.3** | 1.1 | **1.3** |
| C: | **1.2** | **1.3** | -0.8 | -2.4 | -2.4 | -2.4 | -2.4 | -0.2 | -0.8 | -0.8 | -2.4 | -2.4 | -2.4 |
| G: | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -0.8 | **1.2** | **0.6** | -2.4 | -2.4 | -2.4 |
| T: | -0.8 | -2.4 | **0.2** | -2.4 | **0.6** | **1.3** | **1.2** | -0.2 | -2.4 | -0.8 | -2.4 | **-0.2** | -2.4 |
| Test seq: | C | C | T | A | T | T | T | A | G | G | A | T | A |

**Query Score** = 1.2 + 1.3 + 0.2 + 1.3 + 0.6 + 1.3 + 1.2
+ 0.6 + 1.2 + 0.6 + 1.3 + −0.2 + 1.3
= **11.9**

# Scoring a test sequence

Query Sequence

`CCTATTTAGGATA`

PSSM:

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A:** | -2.4 | -2.4 | 0.8 | **1.3** | 0.6 | -2.4 | -0.8 | **0.6** | -2.4 | 0.2 | **1.3** | 1.1 | **1.3** |
| **C:** | **1.2** | **1.3** | -0.8 | -2.4 | -2.4 | -2.4 | -2.4 | -0.2 | -0.8 | -0.8 | -2.4 | -2.4 | -2.4 |
| **G:** | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -0.8 | **1.2** | **0.6** | -2.4 | -2.4 | -2.4 |
| **T:** | -0.8 | -2.4 | **0.2** | -2.4 | **0.6** | **1.3** | **1.2** | -0.2 | -2.4 | -0.8 | -2.4 | **-0.2** | -2.4 |
| Test seq: | **C** | **C** | **T** | **A** | **T** | **T** | **T** | **A** | **G** | **G** | **A** | **T** | **A** |

**Query Score =** 1.2 + 1.3 + 0.2 + 1.3 + 0.6 + 1.3 + 1.2
+ 0.6 + 1.2 + 0.6 + 1.3 + −0.2 + 1.3
**= 11.9**

**Q.** Does the query sequence match the DNA sequence profile?

# Scoring a test sequence...

Query Sequence

CCTATTTAGGATA

Best Possible Sequence

CCAATTTAGGAAA

PSSM:

| Position k = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A:** | -2.4 | -2.4 | **0.8** | **1.3** | 0.6 | -2.4 | -0.8 | **0.6** | -2.4 | 0.2 | **1.3** | **1.1** | **1.3** |
| **C:** | **1.2** | **1.3** | -0.8 | -2.4 | -2.4 | -2.4 | -2.4 | -0.2 | -0.8 | -0.8 | -2.4 | -2.4 | -2.4 |
| **G:** | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -2.4 | -0.8 | **1.2** | **0.6** | -2.4 | -2.4 | -2.4 |
| **T:** | -0.8 | -2.4 | 0.2 | -2.4 | **0.6** | **1.3** | **1.2** | -0.2 | -2.4 | -0.8 | -2.4 | -0.2 | -2.4 |
| **Max Score:** | **C** | **C** | **A** | **A** | **T** | **T** | **T** | **A** | **G** | **G** | **A** | **A** | **A** |
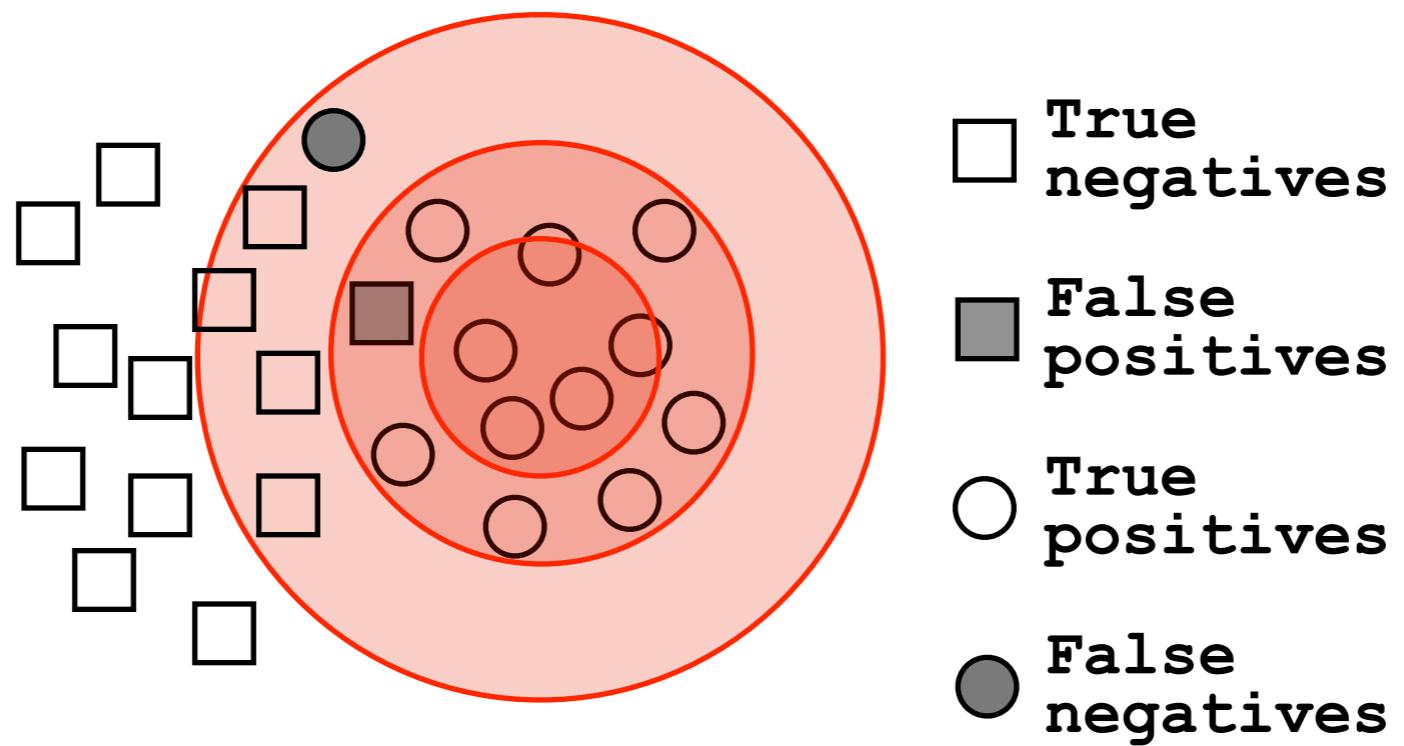
**Max Score** = 1.2 + 1.3 + 0.8 + 1.3 + 0.6 + 1.3 + 1.2
+ 0.6 + 1.2 + 0.6 + 1.3 + 1.1 + 1.3
= **13.8**

**A.** Following method in Harbison *et al.* (2004) Nature 431:99-104

Heuristic threshold for match = 60% x Max Score = (0.6 x 13.8 = 8.28); 11.9 > 8.28; Therefore our query is a potential TFBS!

# Picking a threshold for PSSM matching

Again, you want to select a threshold that **minimizes FPs** (e.g., how many shuffled or random sequences does the PSSM match with that score) and **minimizes FNs** (e.g., how many of the 'real' sequences are missed with that score).



True negatives

False positives

True positives

False negatives

```
FP=0, FN=7, TP=5          5/(5+0) = 1
FP=1, FN=1, TP=11         11/(11+1) = 0.92
FP=5, FN=0, TP=12         12/(12+5) = 0.71
```
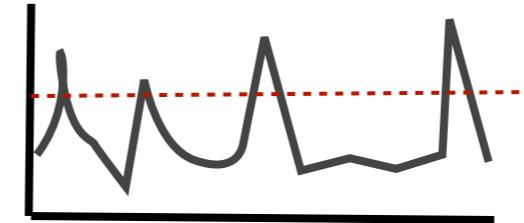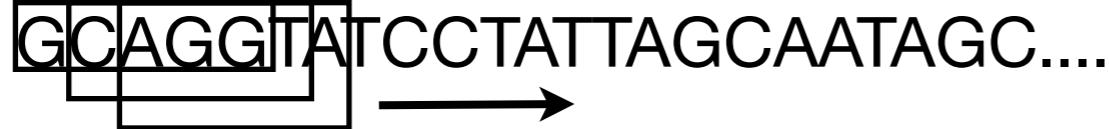
Q. Which threshold has the best PPV (TP/(TP+FP)) ?

# Searching for PSSM matches

If we do not allow gaps (i.e., no insertions or deletions):

- Perform a linear scan, scoring the match to the PSSM at each position in the sequence - the "sliding window" method

GCAGGTATCCTATTAGCAATAGC....

If we allow gaps:

- Can use dynamic programming to align the profile to the protein sequence(s) (with gap penalties)
  We will discuss PSI-BLAST shortly...
  see Mount, Bioinformatics: sequence and genome analysis (2004)

- Can use hidden Markov Model-based methods
  We will cover HMMs at the end of today's lecture...
  see Durbin et al., Biological Sequence Analysis (1998)

# Side note: Profiles software and databases...

**InterPro** is an attempt to group a number of protein domain databases.
   http://www.ebi.ac.uk/interpro

It currently includes:
- PFAM
- PROSITE
- PRINTS
- ProDom
- SMART
- TIGRFAMs

- InterPro tries to have and maintain a high quality of annotation
- The database and a stand-alone package (**iprscan**) are available for UNIX platforms, see:
   ftp://ftp.ebi.ac.uk/pub/databases/interpro
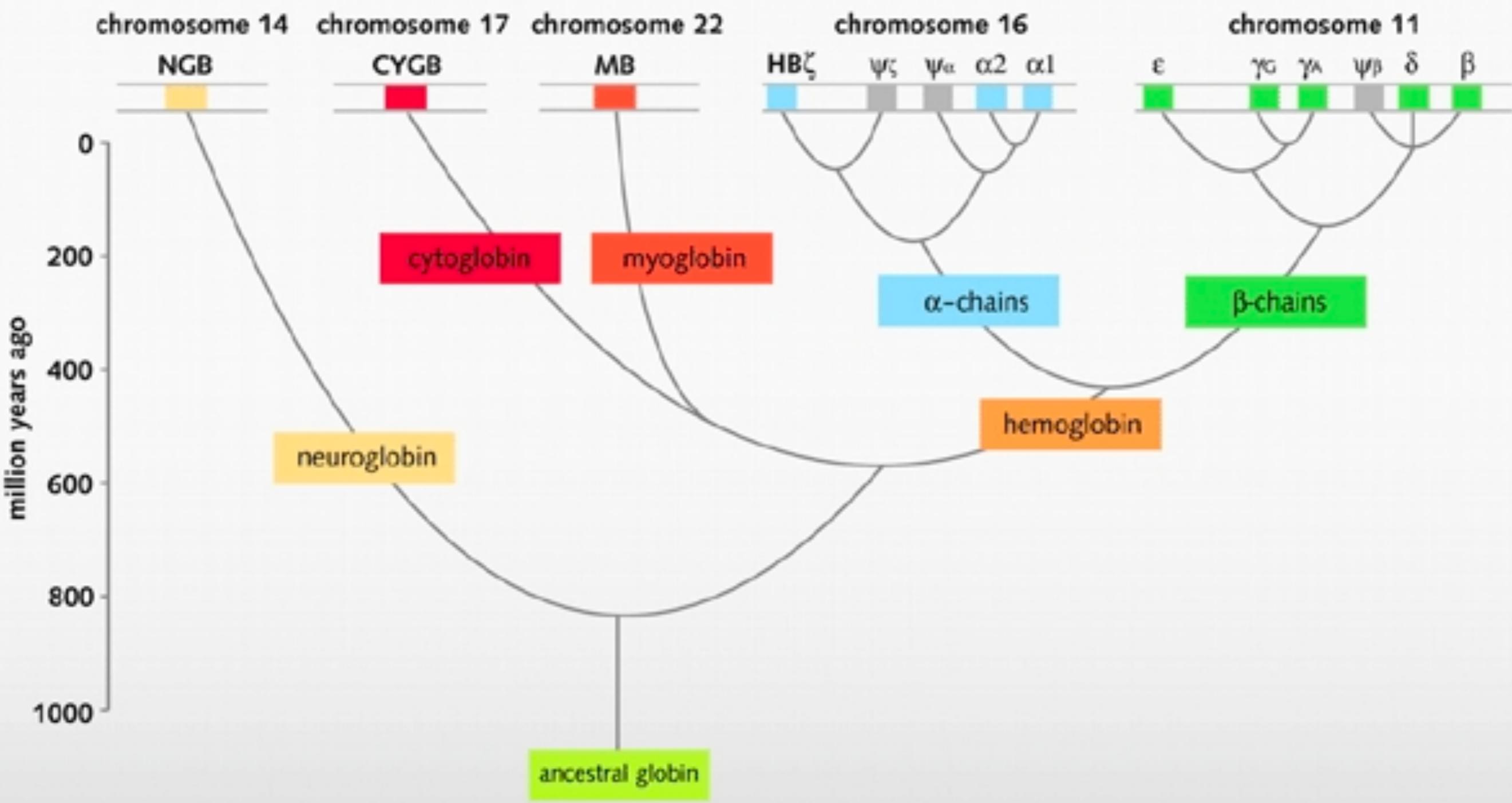
# Todays Menu

- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns

- **Sequence profiles** and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations

- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity

- **Hidden Markov models** (HMMs): More versatile probabilistic model for detection of remote similarities

**Hands-on sections 1 & 2:**
Comparing methods and the trade-off between sensitivity, selectivity and performance

**~50 mins**

# **Side Note:** Human Globins

**An evolutionary model of human globins.**
The different locations of globin genes in human chromosomes are reported at the top of the figure, distinguishing between the functional genes (in color) and the pseudogenes (in grey).

**Recall:** BLOUSM62 does not take the local context of a particular position into account

(*i.e.* all like substitutions are scored the same regardless of their location in the molecules).

# By default BLASTp match scores come from the **BLOSUM62** matrix



**Note**. All matches of Alanine for Alanine score +4 regardless of their position or context in the molecule.

# PSI-BLAST: Position specific iterated BLAST

- The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query

    – PSI-BLAST constructs a multiple sequence alignment from the results of a first round BLAST search and then creates a "profile" or specialized position-specific scoring matrix (PSSM) for subsequent search rounds

# Inspect the blastp output to identify empirical "rules" regarding amino acids tolerated at each position
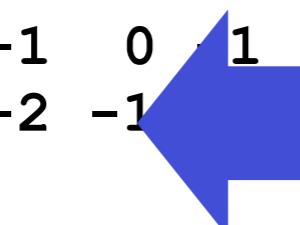
| | | |
|---|---|---|
| 730496 | 66 | FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDH 125 |
| 200679 | 63 | FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 122 |
| 206589 | 34 | FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 93 |
| 2136812 | 2 | MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 53 |
| 132408 | 65 | FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124 |
| 267584 | 44 | FSVDESGKVTATAHGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAASYLQTGNDDH 103 |
| 267585 | 44 | FSVDGSGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 103 |
| 8777608 | 63 | FTIHEDGAMTATAKGRVIILNNWEMCADMMATFETTPDPAKFRMRYWGAASYLQTGNDDH 122 |
| 6687453 | 60 | FKVEEDGTMTATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMKYWGAAAYLQTGYDDH 119 |
| 10697027 | 81 | FKVQEDGTMTATATGRVIILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTGYDDH 140 |
| 13645517 | 1 | MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 32 |
| 13925316 | 38 | FSVDGSGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 97 |
| 131649 | 65 | YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY 126 |

R,I,K   C   D,E,T   K,R,T   N,L,Y,G

|     |   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|-----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | M | -1 | -2 | 2 | 3 | 2 | 1 | 2 | 3 | 2 | 1 | 2 | 2 | 6 | 0 | 3 | 2 | 1 | 2 | 1 | 1 |
| 2 | K | -1 | 1 | 0 | 1 | -4 | 2 | 4 | -2 | 0 | -3 | -3 | 3 | -2 | -4 | -1 | 0 | -1 | -3 | -2 | -3 |
| 3 | W | -3 | -3 | -4 | -5 | -3 | -2 | -3 | -3 | | | | | | 4 | -3 | -3 | 12 | 2 | -3 | |
| 4 | V | 0 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | | | | | 3 | -2 | 0 | -3 | -1 | 4 | | |
| 5 | W | -3 | -3 | -4 | -5 | -3 | -2 | -3 | -3 | -3 | -3 | -2 | -3 | -2 | 1 | -4 | -3 | -3 | 12 | 2 | -3 |
| 6 | A | 5 | -2 | -2 | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |
| 7 | L | -2 | -2 | -4 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -3 | 2 | 0 | -3 | -3 | -1 | -2 | -1 | 1 |
| 8 | L | -1 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -3 | 2 | 2 | -3 | 1 | 3 | -3 | -2 | -1 | -2 | 0 | 3 |
| 9 | L | -1 | -3 | -4 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -3 | 2 | 0 | -3 | -3 | -1 | -2 | -1 | 2 |
| 10 | L | -2 | -2 | -4 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -3 | 2 | 0 | -3 | -3 | -1 | -2 | -1 | 1 |
| 11 | A | 5 | -2 | -2 | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |
| 12 | A | 5 | -2 | -2 | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |
| 13 | W | -2 | | | | | | | | | -3 | 2 | 1 | -3 | -3 | -2 | 7 | 0 | 0 | | |
| 14 | A | 3 | | | | | | | | | -1 | -2 | -3 | -1 | 1 | -1 | -3 | -3 | -1 | | |
| 15 | A | 2 | | | | | | | | | 0 | -2 | -3 | -1 | 3 | 0 | -3 | -2 | -2 | | |
| 16 | A | 4 | | | | | | | | | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | -1 | | |
| ... | | | | | | | | | | | | | | | | | | | | | |
| 37 | S | 2 | -1 | 0 | -1 | -1 | 0 | 0 | 0 | -1 | -2 | -3 | 0 | -2 | -3 | -1 | 4 | 1 | -3 | -2 | -2 |
| 38 | G | 0 | -3 | -1 | -2 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -4 | -2 | 0 | -2 | -3 | -3 | -4 |
| 39 | T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -3 | -2 | 0 |
| 40 | W | -3 | -3 | -4 | -5 | -3 | -2 | -3 | -3 | -3 | -3 | -2 | -3 | -2 | 1 | -4 | -3 | -3 | 9 | 2 | -3 |
| 41 | Y | -2 | -2 | -2 | -3 | -3 | -2 | -2 | -3 | 2 | -2 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| 42 | A | 4 | -2 | -2 | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |

20 amino acids

All the amino acids from position 1 to N (the end of your query protein)

|     |     | A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V  |
|-----|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1   | M   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 2   | K   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 3   | W   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 4   | V   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 5   | W   | -3 | -3 | -4 | -5 | -3 | -2 | -3 | -3 | -3 | -3 | -2 | -3 | -2 | 1  | -4 | -3 | -3 | 12 | 2  | -3 |
| 6   | A   | 5  | -2 | -2 | -2 | -1 | -1 | -1 | 0  | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1  | 0  | -3 | -2 | 0  |
| 7   | L   | -2 | -2 | -4 | -4 | -1 | -2 | -3 | -4 | -3 | 2  | 4  | -3 | 2  | 0  | -3 | -3 | -1 | -2 | -1 | 1  |
| 8   | L   | -1 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -3 | 2  | 2  | -3 | 1  | 3  | -3 | -2 | -1 | -2 | 0  | 3  |
| 9   | L   | -1 | -3 | -4 | -4 |    |    |    |    |    |    |    |    |    |    | -3 | -3 | -1 | -2 | -1 | 2  |
| 10  | L   | -2 | -2 | -4 | -4 |    |    |    |    |    |    |    |    |    |    | -3 | -3 | -1 | -2 | -1 | 1  |
| 11  | A   | 5  | -2 | -2 | -2 |    |    |    |    |    |    |    |    |    | -1 | 1  | 0  | -3 | -2 | 0  |    |
| 12  | A   | 5  | -2 | -2 | -2 |    |    |    |    |    |    |    |    |    | -1 | 1  | 0  | -3 | -2 | 0  |    |
| 13  | W   | -2 | -3 | -4 | -4 |    |    |    |    |    |    |    |    |    | -3 | -3 | -2 | 7  | 0  | 0  |    |
| 14  | A   | 3  | -2 | -1 | -2 |    |    |    |    |    |    |    |    |    | -1 | 1  | -1 | -3 | -3 | -1 |    |
| 15  | A   | 2  | -1 | 0  | 1  |    |    |    |    |    |    |    |    |    | -1 | 3  | 0  | -3 | -2 | -2 |    |
| 16  | A   | 4  | -2 | -1 |    |    |    |    |    |    |    |    |    |    | -1 | 1  | 0  | -3 | -2 | -1 |    |
| ... |     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 37  | S   | 2  | -1 | 0  | -1 |    |    |    |    |    |    |    |    |    | -1 | 4  | 1  | -3 | -2 | -2 |    |
| 38  | G   | 0  | -3 | -1 | -2 |    |    |    |    |    |    |    |    |    | -2 | 0  | -2 | -3 | -3 | -4 |    |
| 39  | T   | 0  | -1 | 0  | -1 |    |    |    |    |    |    |    |    |    | -1 | 1  | 5  | -3 | -2 | 0  |    |
| 40  | W   | -3 | -3 | -4 | -5 |    |    |    |    |    |    |    |    |    | -4 | -3 | -3 | 9  | 2  | -3 |    |
| 41  | Y   | -2 | -2 | -2 | -3 |    |    |    |    |    |    |    |    |    | -3 | -2 | -2 | 2  | 7  | -1 |    |
| 42  | A   | 4  | -2 | -2 | -2 | -1 | -1 | -1 | 0  | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1  | 0  | -3 | -2 | 0  |

The PSI-BLAST PSSM is essentially a query customized scoring matrix that is more sensitive than BLOSUM.
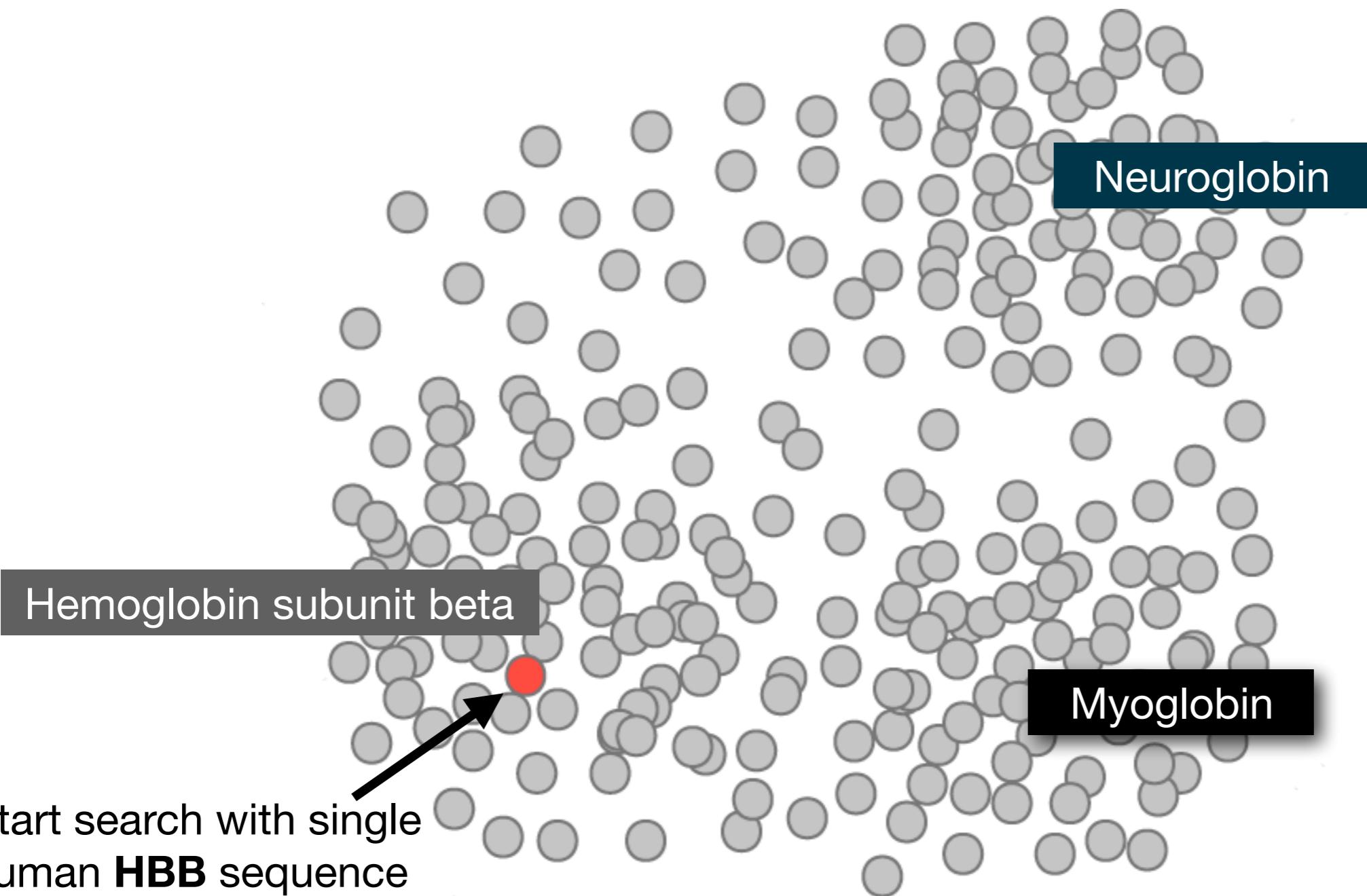
**Note:** A given amino acid (such as alanine) in your query protein can receive different scores for matching alanine depending on the position in the protein (BLOSUM $S_{AA}$ = +4)

53
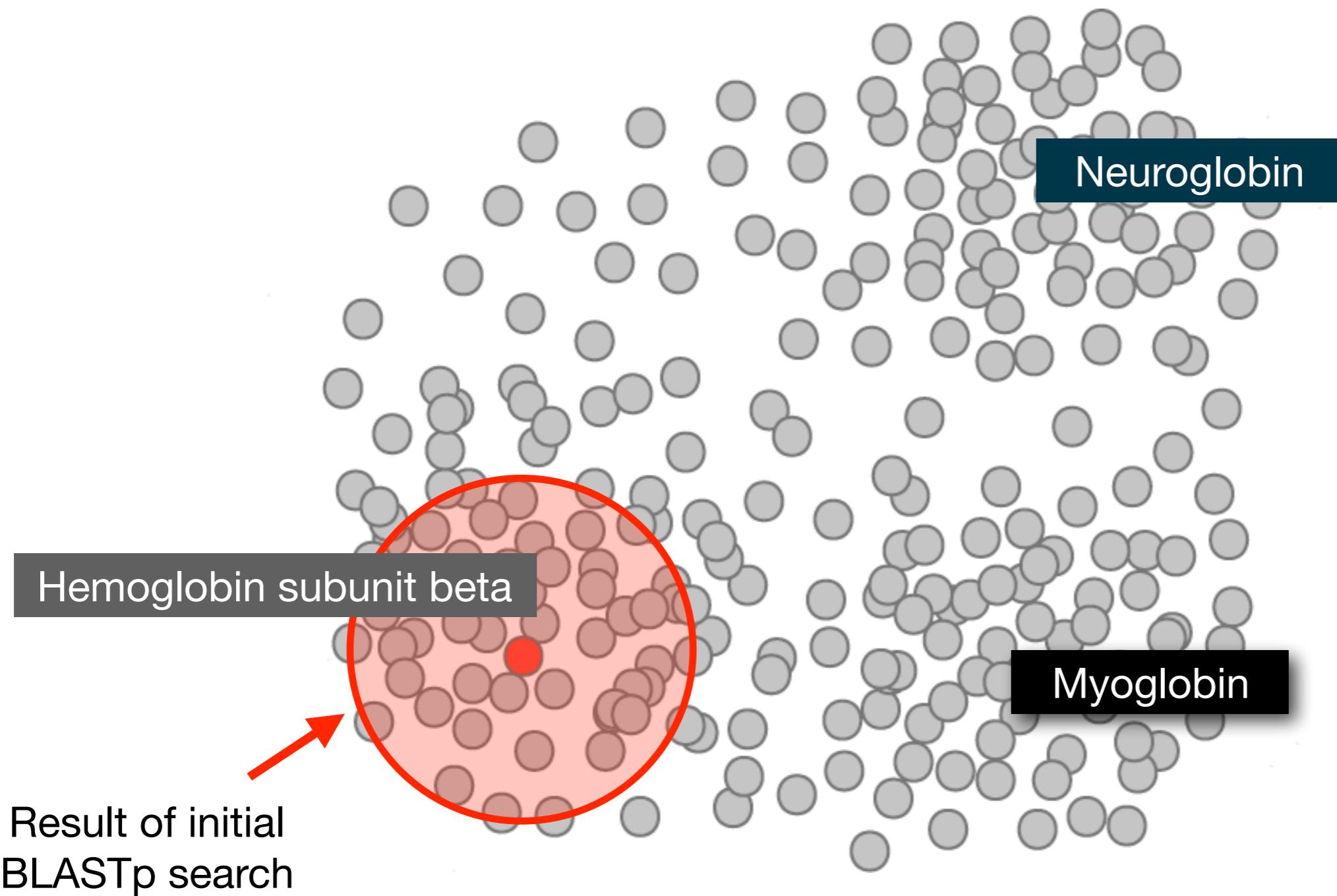
# PSI-BLAST:  Position-Specific Iterated BLAST
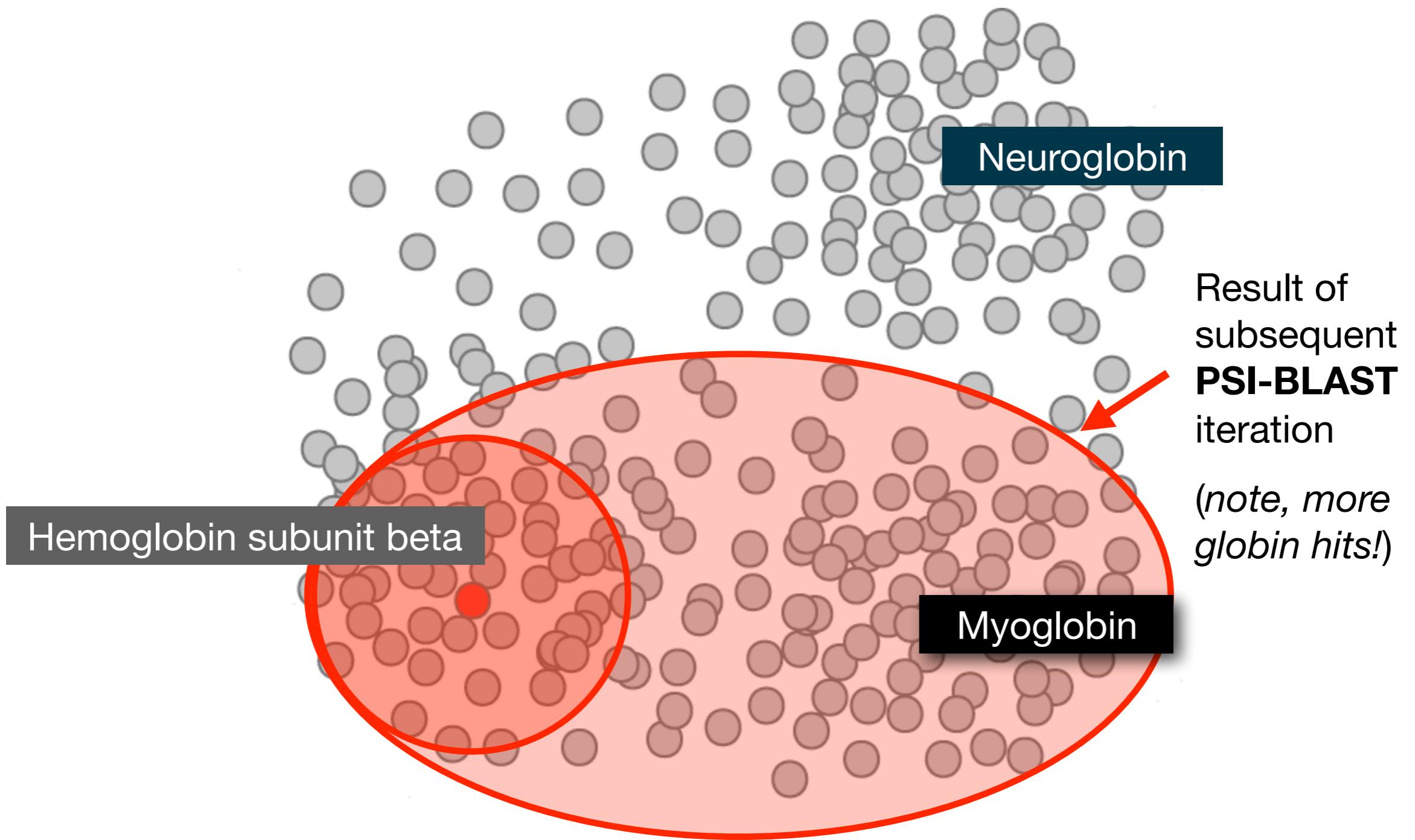
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear.  Enter the more sensitive PSI-BLAST

1. BLAST input sequence to find significant alignments

2. Construct a multiple sequence alignment (MSA)

3. Construct a PSSM

4. BLAST PSSM profile to search for new hits
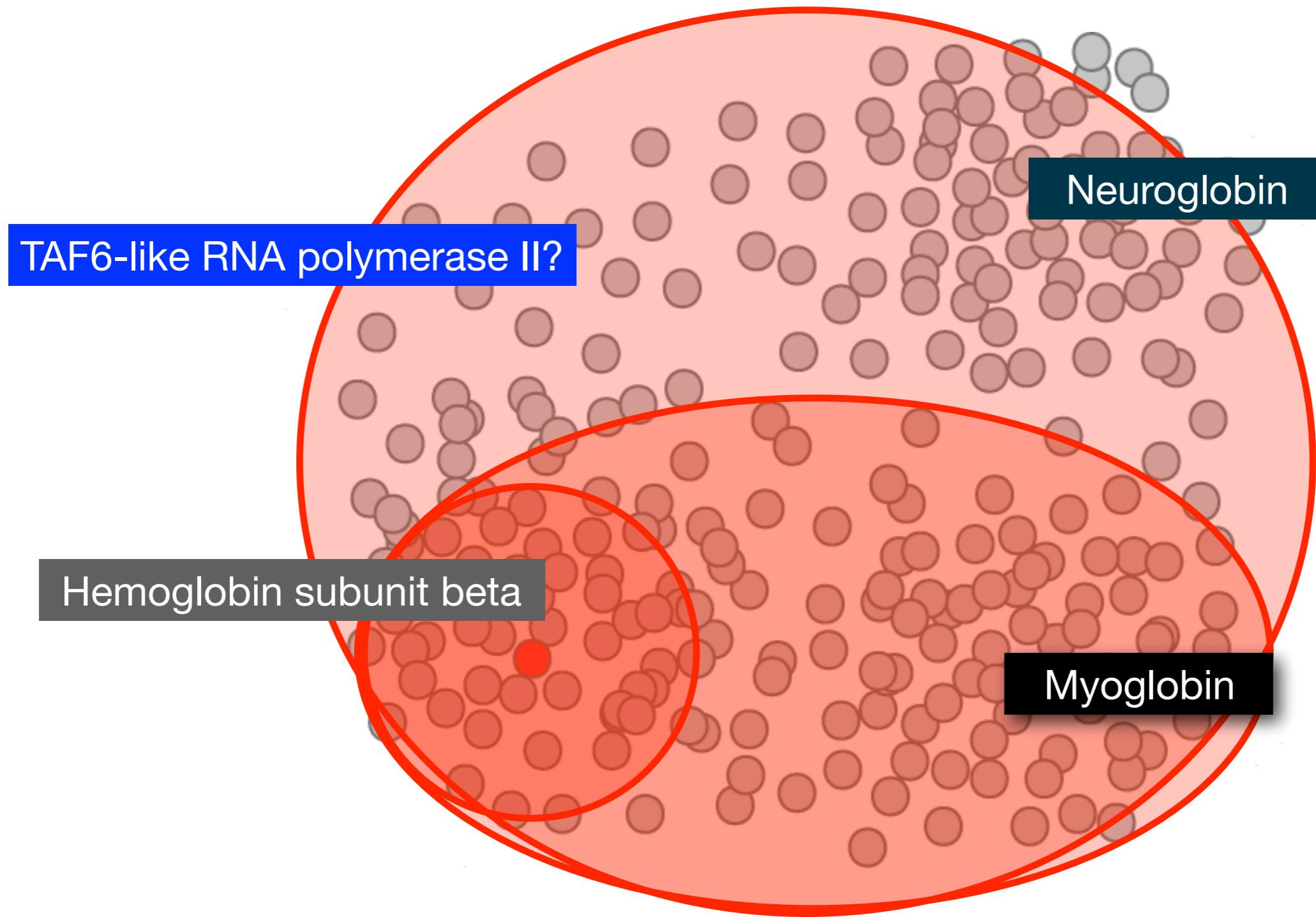
5. Iterate

(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

Neuroglobin

Hemoglobin subunit beta

Myoglobin

Start search with single
human **HBB** sequence

Neuroglobin

Hemoglobin subunit beta

Myoglobin

Result of initial
BLASTp search

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| hemoglobin subunit beta [Homo sapiens] | 301 | 301 | 100% | 2e-106 | 100% | NP_000509.1 |
| hemoglobin subunit delta [Homo sapiens] | 284 | 284 | 100% | 7e-100 | 93% | NP_000510.1 |
| hemoglobin subunit epsilon [Homo sapiens] | 240 | 240 | 100% | 2e-82 | 76% | NP_005321.1 |
| hemoglobin subunit gamma-2 [Homo sapiens] | 235 | 235 | 100% | 2e-80 | 73% | NP_000175.1 |
| hemoglobin subunit gamma-1 [Homo sapiens] | 232 | 232 | 100% | 3e-79 | 73% | NP_000550.2 |
| hemoglobin subunit alpha [Homo sapiens] | 114 | 114 | 97% | 7e-33 | 43% | NP_000508.1 |
| hemoglobin subunit zeta [Homo sapiens] | 100 | 100 | 97% | 3e-27 | 36% | NP_005323.1 |

1

| Description | Max score | Total score | Query cover | E value | Ident | Accession | |
|---|---|---|---|---|---|---|---|
| hemoglobin subunit beta [Homo sapiens] | 301 | 301 | 100% | 2e-106 | 100% | NP_000509.1 | **1** |
| hemoglobin subunit delta [Homo sapiens] | 284 | 284 | 100% | 7e-100 | 93% | NP_000510.1 | |
| hemoglobin subunit epsilon [Homo sapiens] | 240 | 240 | 100% | 2e-82 | 76% | NP_005321.1 | |
| hemoglobin subunit gamma-2 [Homo sapiens] | 235 | 235 | 100% | 2e-80 | 73% | NP_000175.1 | |
| hemoglobin subunit gamma-1 [Homo sapiens] | 232 | 232 | 100% | 3e-79 | 73% | NP_000550.2 | |
| hemoglobin subunit alpha [Homo sapiens] | 114 | 114 | 97% | 7e-33 | 43% | NP_000508.1 | |
| hemoglobin subunit zeta [Homo sapiens] | 100 | 100 | 97% | 3e-27 | 36% | NP_005323.1 | |
| myoglobin [Homo sapiens] | 80.5 | 80.5 | 97% | 2e-19 | 26% | NP_005359.1 | **2** |
| neuroglobin [Homo sapiens] | 54.7 | 54.7 | 92% | 2e-09 | 23% | NP_067080.1 | |

**New relevant globins found only by PSI-BLAST**

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| hemoglobin subunit beta [Homo sapiens] | 301 | 301 | 100% | 2e-106 | 100% | NP_000509.1 |
| hemoglobin subunit delta [Homo sapiens] | 284 | 284 | 100% | 7e-100 | 93% | NP_000510.1 |
| hemoglobin subunit epsilon [Homo sapiens] | 240 | 240 | 100% | 2e-82 | 76% | NP_005321.1 |
| hemoglobin subunit gamma-2 [Homo sapiens] | 235 | 235 | 100% | 2e-80 | 73% | NP_000175.1 |
| hemoglobin subunit gamma-1 [Homo sapiens] | 232 | 232 | 100% | 3e-79 | 73% | NP_000550.2 |
| hemoglobin subunit alpha [Homo sapiens] | 114 | 114 | 97% | 7e-33 | 43% | NP_000508.1 |
| hemoglobin subunit zeta [Homo sapiens] | 100 | 100 | 97% | 3e-27 | 36% | NP_005323.1 |

**1**

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| myoglobin [Homo sapiens] | 80.5 | 80.5 | 97% | 2e-19 | 26% | NP_005359.1 |
| neuroglobin [Homo sapiens] | 54.7 | 54.7 | 92% | 2e-09 | 23% | NP_067080.1 |

**2**

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| myoglobin [Homo sapiens] | 159 | 159 | 97% | 3e-50 | 26% | NP_005359.1 |
| hemoglobin subunit alpha [Homo sapiens] | 151 | 151 | 97% | 3e-47 | 42% | NP_000508.1 |
| hemoglobin subunit mu [Homo sapiens] | 147 | 147 | 97% | 6e-46 | 35% | NP_001003938.1 |
| hemoglobin subunit theta-1 [Homo sapiens] | 147 | 147 | 97% | 2e-45 | 37% | NP_005322.1 |
| neuroglobin [Homo sapiens] | 134 | 134 | 92% | 3e-40 | 23% | NP_067080.1 |
| PREDICTED: cytoglobin isoform X2 [Homo sapiens] | 115 | 115 | 66% | 3e-33 | 25% | XP_016879605.1 |
| PREDICTED: microtubule cross-linking factor 1 isoform X1 [Homo sapie | 46.3 | 46.3 | 27% | 7e-06 | 39% | XP_011523942.1 |
| PREDICTED: microtubule cross-linking factor 1 isoform X4 [Homo sapie | 46.3 | 46.3 | 27% | 7e-06 | 39% | XP_005258156.1 |

**3**

**?**

**Inclusion of irrelevant hits can lead to PSSM corruption**

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| hemoglobin subunit beta [Homo sapiens] | 301 | 301 | 100% | 2e-106 | 100% | NP_000509.1 |
| hemoglobin subunit delta [Homo sapiens] | 284 | 284 | 100% | 7e-100 | 93% | NP_000510.1 |
| hemoglobin subunit epsilon [Homo sapiens] | 240 | 240 | 100% | 2e-82 | 76% | NP_005321.1 |
| hemoglobin subunit gamma-2 [Homo sapiens] | 235 | 235 | 100% | 2e-80 | 73% | NP_000175.1 |
| hemoglobin subunit gamma-1 [Homo sapiens] | 232 | 232 | 100% | 3e-79 | 73% | NP_000550.2 |
| hemoglobin subunit alpha [Homo sapiens] | 114 | 114 | 97% | 7e-33 | 43% | NP_000508.1 |
| hemoglobin subunit zeta [Homo sapiens] | 100 | 100 | 97% | 3e-27 | 36% | NP_005323.1 |

**1**

| | | | | | | |
|---|---|---|---|---|---|---|
| myoglobin [Homo sapiens] | 80.5 | 80.5 | 97% | 2e-19 | 26% | NP_005359.1 |
| neuroglobin [Homo sapiens] | 54.7 | 54.7 | 92% | 2e-09 | 23% | NP_067080.1 |

**2**

| | | | | | | |
|---|---|---|---|---|---|---|
| myoglobin [Homo sapiens] | 159 | 159 | 97% | 3e-50 | 26% | NP_005359.1 |
| hemoglobin subunit alpha [Homo sapiens] | 151 | 151 | 97% | 3e-47 | 42% | NP_000508.1 |
| hemoglobin subunit mu [Homo sapiens] | 147 | 147 | 97% | 6e-46 | 35% | NP_001003938.1 |
| hemoglobin subunit theta-1 [Homo sapiens] | 147 | 147 | 97% | 2e-45 | 37% | NP_005322.1 |
| neuroglobin [Homo sapiens] | 134 | 134 | 92% | 3e-40 | 23% | NP_067080.1 |
| PREDICTED: cytoglobin isoform X2 [Homo sapiens] | 115 | 115 | 66% | 3e-33 | 25% | XP_016879605.1 |
| PREDICTED: microtubule cross-linking factor 1 isoform X1 [Homo sapie | 46.3 | 46.3 | 27% | 7e-06 | 39% | XP_011523942.1 |
| PREDICTED: microtubule cross-linking factor 1 isoform X4 [Homo sapie | 46.3 | 46.3 | 27% | 7e-06 | 39% | XP_005258156.1 |

**3**

**Score and E value depends on PSSM**

# PSI-BLAST is performed in five steps

- A normal blastp search uses a scoring matrix (e.g., BLOSUM62) to perform pairwise alignments of your query sequence (such as RBP) against the database. PSI-BLAST also begins with a protein query that is searched against a database of choice.

- PSI-BLAST constructs a multiple sequence alignment (MSA) from an initial blastp-like search. It then creates a **PSSM** based on that multiple alignment.

- This **PSSM** is then used as a query to search the database again.

- PSI-BLAST estimates the statistical significance of the database matches, essentially using the parameters we described for gapped alignments.

- The search process is continued iteratively, typically 3 to 5 times. At each step a new PSSM is built.

# PSI-BLAST returns dramatically more hits

You must decide how many iterations to perform and which sequences to include!

You can stop the search process at any point - typically whenever few new results are returned or when no new "sensible" results are found.

| Iteration | Hits with E < 0.005 | Hits with E > 0.005 |
|-----------|---------------------|---------------------|
| 1 | 34 | 61 |
| 2 | 314 | 79 |
| 3 | 416 | 57 |
| 4 | 432 | 50 |
| 5 | 432 | 50 |

Human retinol-binding protein 4 (RBP4; P02753) was used as a query in a PSI-BLAST search of the RefSeq database.

# Example PSI-BLAST PSSM at iteration 3

The PSI-BLAST PSSM is essentially a query customized scoring matrix that is more sensitive than BLOSUM (e.g. BLOSUM $S_{AA}$ = +4)

20 amino acids types

| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | M | -1 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | -2 | 1 | 2 | -2 | 6 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| 2 | K | -1 | 1 | 0 | 1 | -4 | 2 | 4 | -2 | 0 | -3 | -3 | 3 | -2 | -4 | -1 | 0 | -1 | -3 | -2 | -3 |
| 3 | W | -3 | -3 | -4 | -5 | -3 | -2 | -3 | -3 | -3 | -3 | -2 | -3 | -2 | 1 | -4 | -3 | -3 | 12 | 2 | -3 |
| 4 | V | 0 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -4 | 3 | 1 | -3 | 1 | -1 | -3 | -2 | 0 | -3 | -1 | 4 |
| 5 | W | -3 | -3 | -4 | -5 | -3 | -2 | -3 | -3 | -3 | -3 | -2 | -3 | -2 | 1 | -4 | -3 | -3 | 12 | 2 | -3 |
| 6 | A | 5 | -2 | -2 | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |
| 7 | L | -2 | -2 | -4 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -3 | 2 | 0 | -3 | -3 | -1 | -2 | -1 | 1 |
| 8 | L | -1 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -3 | 2 | 2 | -3 | 1 | 3 | -3 | -2 | -1 | -2 | 0 | 3 |
| 9 | L | -1 | -3 | -4 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -3 | 2 | 0 | -3 | -3 | -1 | -2 | -1 | 2 |
| 10 | L | -2 | -2 | -4 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -3 | 2 | 0 | -3 | -3 | -1 | -2 | -1 | 1 |
| 11 | A | 5 | -2 | -2 | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |
| 12 | A | 5 | -2 | -2 | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |
| 13 | W | -2 | -3 | -4 | -4 | -2 | -2 | -3 | -4 | -3 | 1 | 4 | -3 | 2 | 1 | -3 | -3 | -2 | 7 | 0 | 0 |
| 14 | A | 3 | -2 | -1 | -2 | -1 | -1 | -2 | 4 | -2 | -2 | -2 | -1 | -2 | -3 | -1 | 1 | -1 | -3 | -3 | -1 |
| 15 | A | 2 | -1 | 0 | -1 | -2 | 2 | 0 | 2 | -1 | -3 | -3 | 0 | -2 | -3 | -1 | 3 | 0 | -3 | -2 | -2 |
| 16 | A | 4 | -2 | -1 | -2 | -1 | -1 | -1 | 3 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | -1 |
| ... | | | | | | | | | | | | | | | | | | | | | |
| 37 | S | 2 | -1 | 0 | -1 | -1 | 0 | 0 | 0 | -1 | -2 | -3 | 0 | -2 | -3 | -1 | 4 | 1 | -3 | -2 | -2 |
| 38 | G | 0 | -3 | -1 | -2 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -4 | -2 | 0 | -2 | -3 | -3 | -4 |
| 39 | T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -3 | -2 | 0 |
| 40 | W | -3 | -3 | -4 | -5 | -3 | -2 | -3 | -3 | -3 | -3 | -2 | -3 | -2 | 1 | -4 | -3 | -3 | 12 | 2 | -3 |
| 41 | Y | -2 | -2 | -2 | -3 | -3 | -2 | -2 | -3 | 2 | -2 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| 42 | A | 4 | -2 | -2 | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |
| ... | | | | | | | | | | | | | | | | | | | | | |

Query residues/positions

# PSI-BLAST errors: the corruption problem

The main source of error in PSI-BLAST searches is the spurious amplification of sequences that are unrelated to the query.

There are three main approaches to stopping corruption of PSI-BLAST queries:
- Perform multi-domain splitting of your query sequence
  If a query protein has several different domains PSI-BLAST may find database matches related to both individually. One should not conclude that these hits with different domains are related.
  - Often best to search using just one domain of interest.

- Inspect each PSI-BLAST iteration removing suspicious hits.
  E.g., your query protein may have a generic coiled-coil domain, and this may cause other proteins sharing this motif (such as myosin) to score better than the inclusion threshold even though they are not related.
  - Use your biological knowledge!

- Lower the default expect level (e.g., E = 0.005 to E = 0.0001).
  This may suppress appearance of FPs (but also TPs)

# Profile advantages and disadvantages

**Advantages:**

- Quantitate with a good scoring system

- Weights sequences according to observed diversity
  Profile is specific to input sequence set

- Very sensitive
  Can detect weak similarity

- Relatively easy to compute
  Automatic profile building tools available

**Disadvantages:**

- If a mistake enters the profile, you may end up with irrelevant data
  The corruption problem!

- Ignores higher order dependencies between positions
  i.e., correlations between the residue found at a given position and those found at other positions (e.g. salt-bridges, structural constraints on RNA etc...)

- Requires some expertise and oversight to use proficiently

# Todays Menu

- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns

- **Sequence profiles** and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations

- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity

- **Hidden Markov models** (HMMs): More versatile probabilistic model for detection of remote similarities

# Problems with PSSMs: Positional dependencies

Do not capture positional dependencies

WEIRD

WEIRD

WEIQH

WEIRD

WEIQH

| | | | | | |
|---|---|---|---|---|---|
| D | | | | | 0.6 |
| E | | I | | | |
| H | | | | | 0.4 |
| I | | | I | | |
| Q | | | | 0.4 | |
| R | | | | 0.6 | |
| W | I | | | | |
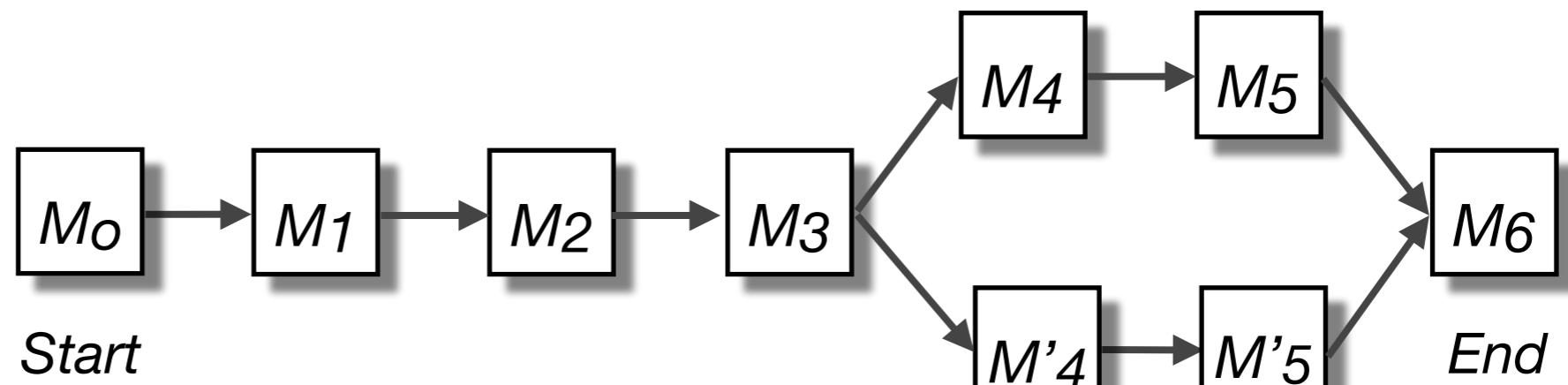
**Note**: We <u>never</u> see **QD** or **RH**, we only see **RD** and **QH**.
However, P(RH)=0.24, P(QD)=0.24, while P(QH)=0.16

# Markov chains: Positional dependencies

✓

The connectivity or **topology** of a Markov chain can easily be designed to capture dependencies and variable length motifs.

```
WEIRD
WEIRD
WEIQH
WEIRD
WEIQH
```
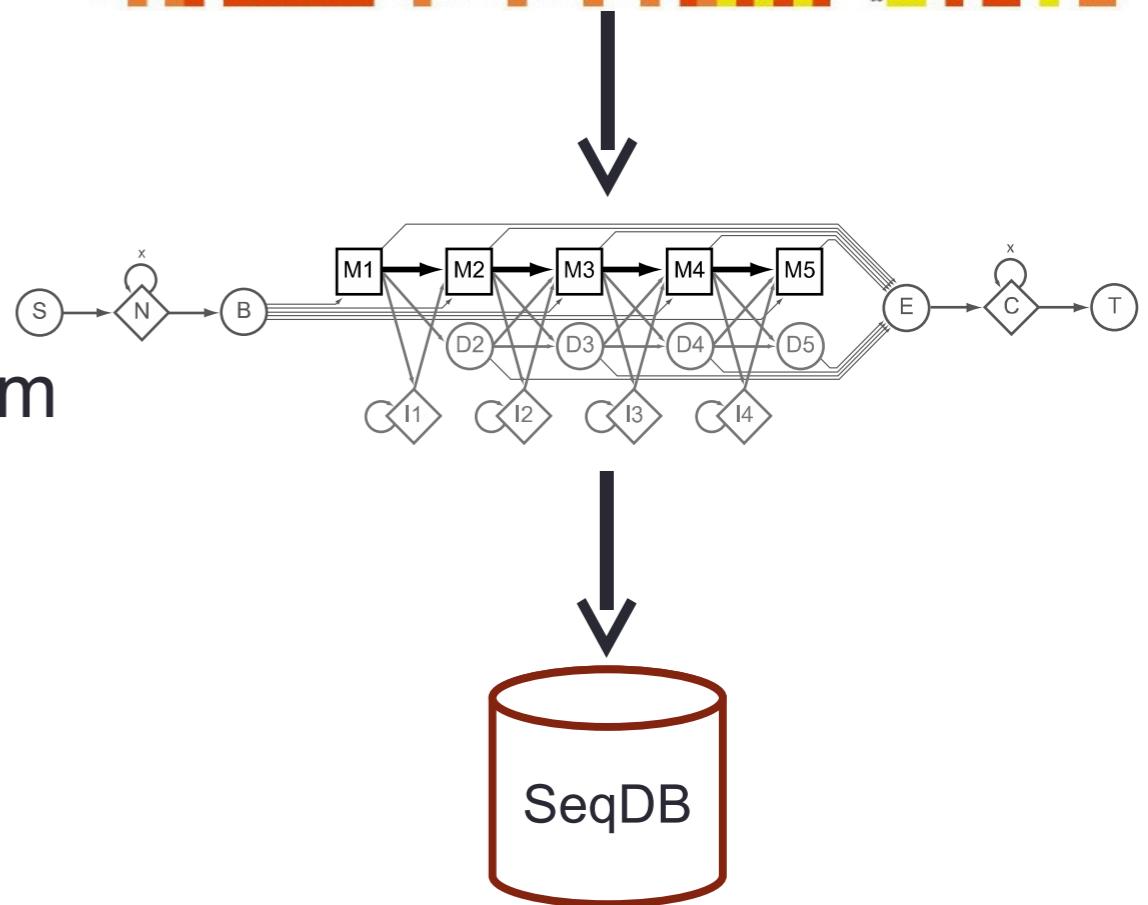


Recall that a PSSM for this motif would give the sequences **WEIRD** and **WEIRH** equally good scores even though the **RH** and **QR** combinations were not observed

# Use of HMMER
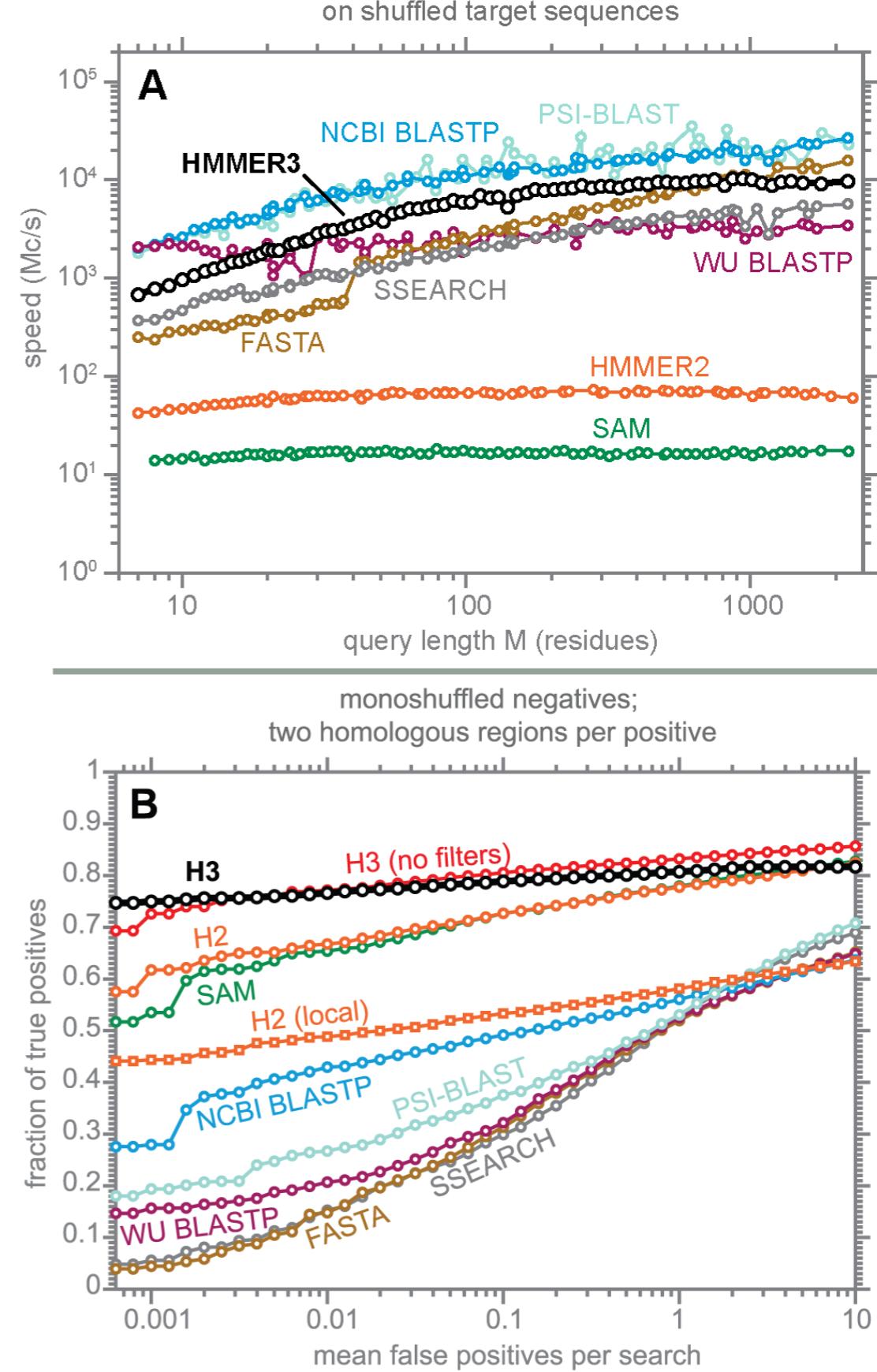
- Widely used by protein family databases
  - Use 'seed' alignments
- Until 2010
  - Computationally expensive
  - Restricted to HMMs constructed from multiple sequence alignments
- Command line application

# HMMER vs BLAST

| | HMMER | BLAST |
|---|---|---|
| Program | *PHMMER* | *BLASTP* |
| Query | Single sequence | |
| Target Database | Sequence database | |
| Program | *HMMSCAN* | *RPSBLAST* |
| Query | Single sequence | |
| Target Database | Profile HMM database, e.g. Pfam | PSSM database, e.g. CDD |
| Program | *HMMSEARCH* | *PSI-BLAST* |
| Query | Profile HMM | PSSM |
| Target Database | Sequence database | |
| Program | *JACKHMMER* | *PSI-BLAST* |
| Query | Single sequence | |
| Target Database | Sequence database | |



A — on shuffled target sequences
speed (Mc/s) vs query length M (residues): PSI-BLAST, NCBI BLASTP, HMMER3, WU BLASTP, SSEARCH, FASTA, HMMER2, SAM

B — monoshuffled negatives; two homologous regions per positive
fraction of true positives vs mean false positives per search: H3, H3 (no filters), H2, SAM, H2 (local), NCBI BLASTP, PSI-BLAST, SSEARCH, WU BLASTP, FASTA

Modified from: S. R. Eddy
**PLoS Comp. Biol.**, 7:e1002195, 2011.

# Fast Web Searches

- Parallelized searches across compute farm
  - Average query returns ~1 sec
- Range of sequence databases
  - Large Comprehensive
  - Curated / Structure
  - Metagenomics
  - Representative Proteomes
- Family Annotations
  - Pfam
- Batch and RESTful API
  - Automatic and Human interface



hmmer.janelia.org

# HMMER

## Biosequence analysis using profile hidden Markov Models

Home | Search | Results | Software | Help | About | Contact

phmmer | hmmscan | hmmsearch | jackhmmer

# protein sequence vs protein sequence database

**Paste a Sequence** | Upload a File | Accession Search

Paste in your sequence or use the example ❓

```
>NP_000509.1 hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

[ Submit ]  [ Reset ]

▼ **Sequence Database** ❓

Frequently used databases:  [ Reference Proteomes ] [ UniProtKB ] [ SwissProt ] [ PDB ] [ Ensembl ]

Current database selection:

[ SwissProt ▾ ]

▼ **Restrict by Taxonomy** ❓

⦿ Taxon search    ◯ Pre-defined representatives

Organism:

[                                                                    ]

## Significant Query Matches (12) in *swissprot* (v.2018_11)
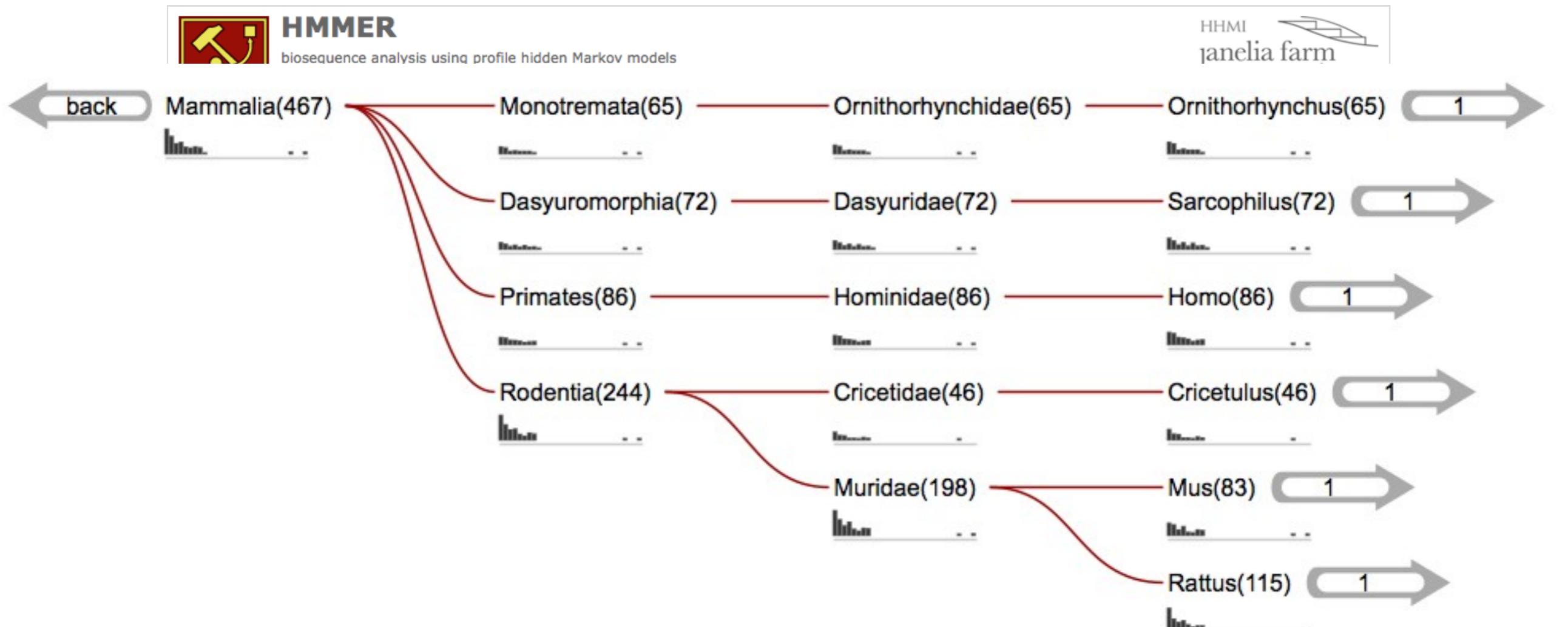
| | Target | Description | Species | Cross-references | E-value |
|---|---|---|---|---|---|
| > | HBB_HUMAN | Hemoglobin subunit beta | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 6.8e-99 |
| > | HBD_HUMAN | Hemoglobin subunit delta | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 1.6e-91 |
| > | HBE_HUMAN | Hemoglobin subunit epsilon | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 1.5e-74 |
| > | HBG2_HUMAN | Hemoglobin subunit gamma-2 | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 8.8e-73 |
| > | HBG1_HUMAN | Hemoglobin subunit gamma-1 | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 6.2e-72 |
| > | HBA_HUMAN | Hemoglobin subunit alpha | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 3.8e-29 |
| > | HBAZ_HUMAN | Hemoglobin subunit zeta | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 4.5e-23 |
| > | HBAT_HUMAN | Hemoglobin subunit theta-1 | Homo sapiens | 🟤🟣🔵  🟠🔴🔵 | 5.2e-22 |
| > | HBM_HUMAN | Hemoglobin subunit mu | Homo sapiens | 🟤🟣🔵  🟠🔴🔵 | 3.4e-19 |
| > | CYGB_HUMAN | Cytoglobin | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 3.1e-14 |
| > | MYG_HUMAN | Myoglobin | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 2.3e-06 |
| > | NGB_HUMAN | Neuroglobin | Homo sapiens | 🟤🟣🔵🟢🟢🟠🔴🔵 | 0.0017 |

(show all) alignments          Your search took: 0.06 secs          showing rows 1 - 12 of 12

**Local Link**

# Visualization of Results – By Taxonomy

# **PFAM**: Protein Family Database of Profile HMMs
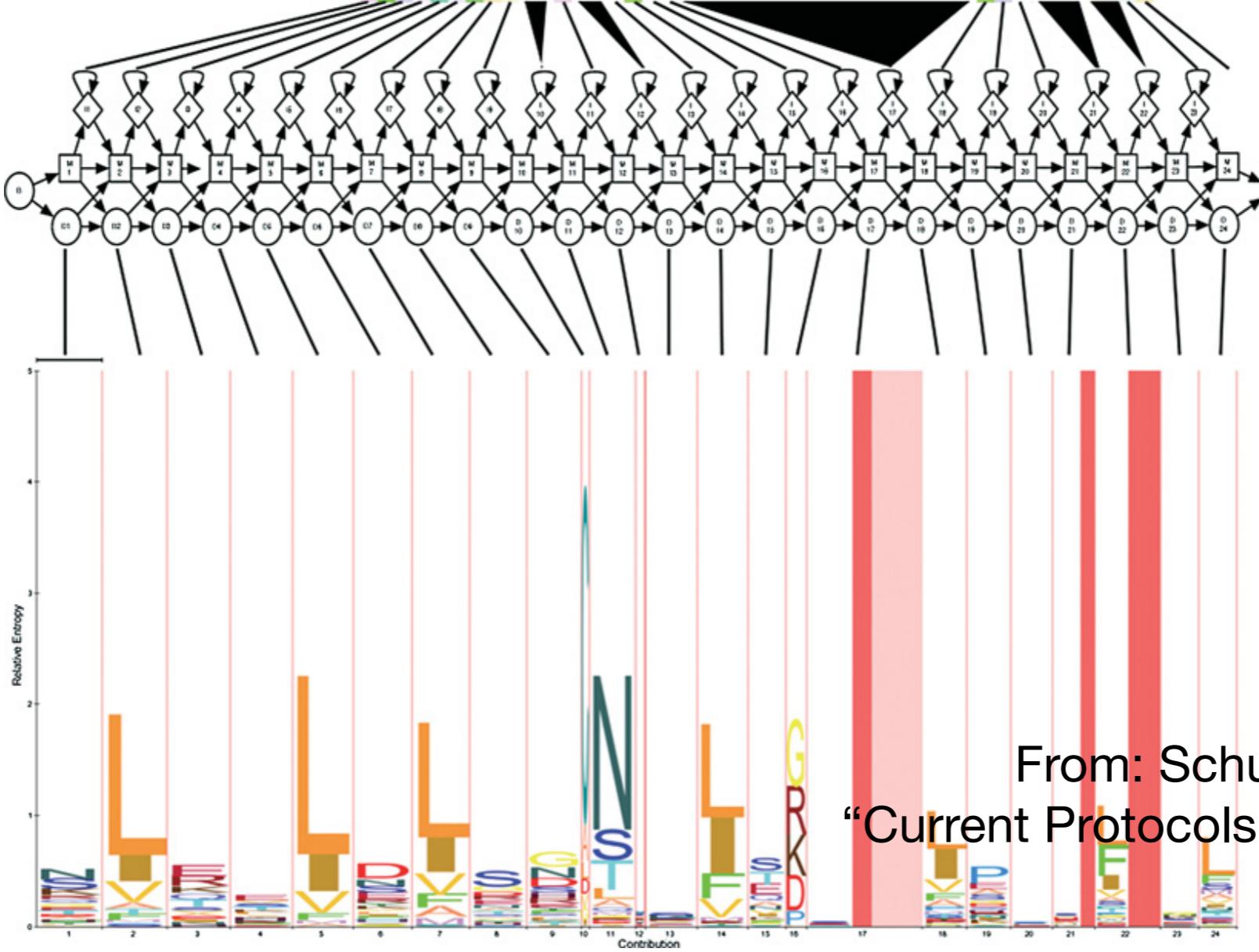
Comprehensive compilation of both multiple sequence alignments and profile HMMs of protein families.

http://pfam.xfam.org

PFAM consists of two databases:

- **Pfam-A** is a manually curated collection of protein families in the form of multiple sequence alignments and profile HMMs. HMMER software is used to perform searches.

- **Pfam-B** contains additional protein sequences that are automatically aligned. Pfam-B serves as a useful supplement that makes the database more comprehensive.

- Pfam-A also contains higher-level groupings of related families, known as **clans**

From: Schuster-Bockler *et al.*
"Current Protocols in Bioinformatics"
Supplement 18.

# Summary

- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns

- **Sequence profiles** and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations

- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity

- **Hidden Markov models** (HMMs): More versatile probabilistic model for detection of remote similarities

# Homework:
# Questions & DataCamp!

**Install R and RStudio  (see website)**

**Complete the Introduction to R course on DataCamp**
(Check your email for your DataCamp invite and sign up with your UCSD email (i.e. first part of your email address) please.

**Let me know NOW if you don't see the invite email!**

That's it!
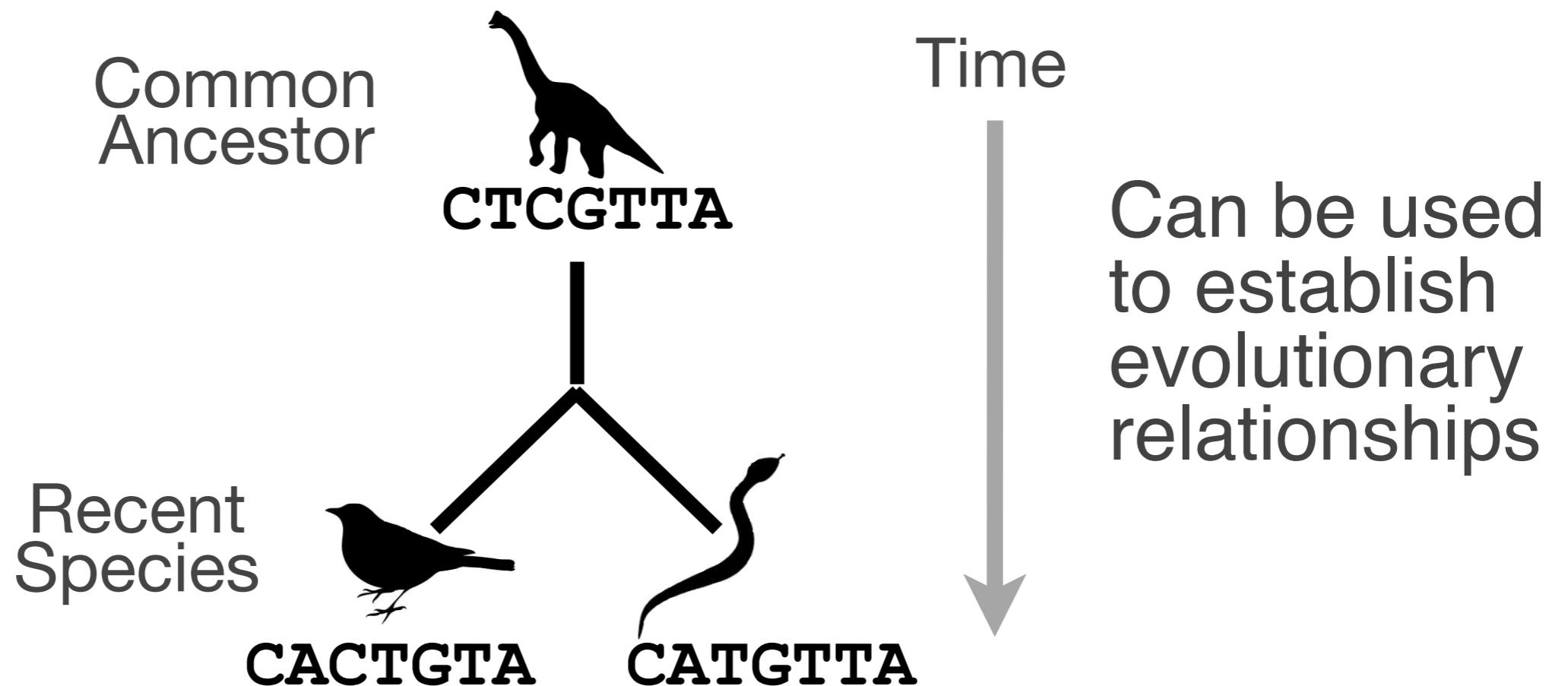
# Reference Slides:

**Side Note:** Orthologs *vs* Paralogs

# Sequence comparison is most informative when it detects homologs

Homologs are sequences that have common origins i.e. they share a common ancestor

- They may or may not have common activity

Common Ancestor

**CTCGTTA**

Time

Can be used to establish evolutionary relationships

Recent Species

**CACTGTA**   **CATGTTA**

# Key terms

When we talk about related sequences we use specific terminology.

Homologous sequences may be either:

– Orthologs or Paralogs

(Note. these are all or nothing relationships!)

Any pair of sequences may share a certain level of:
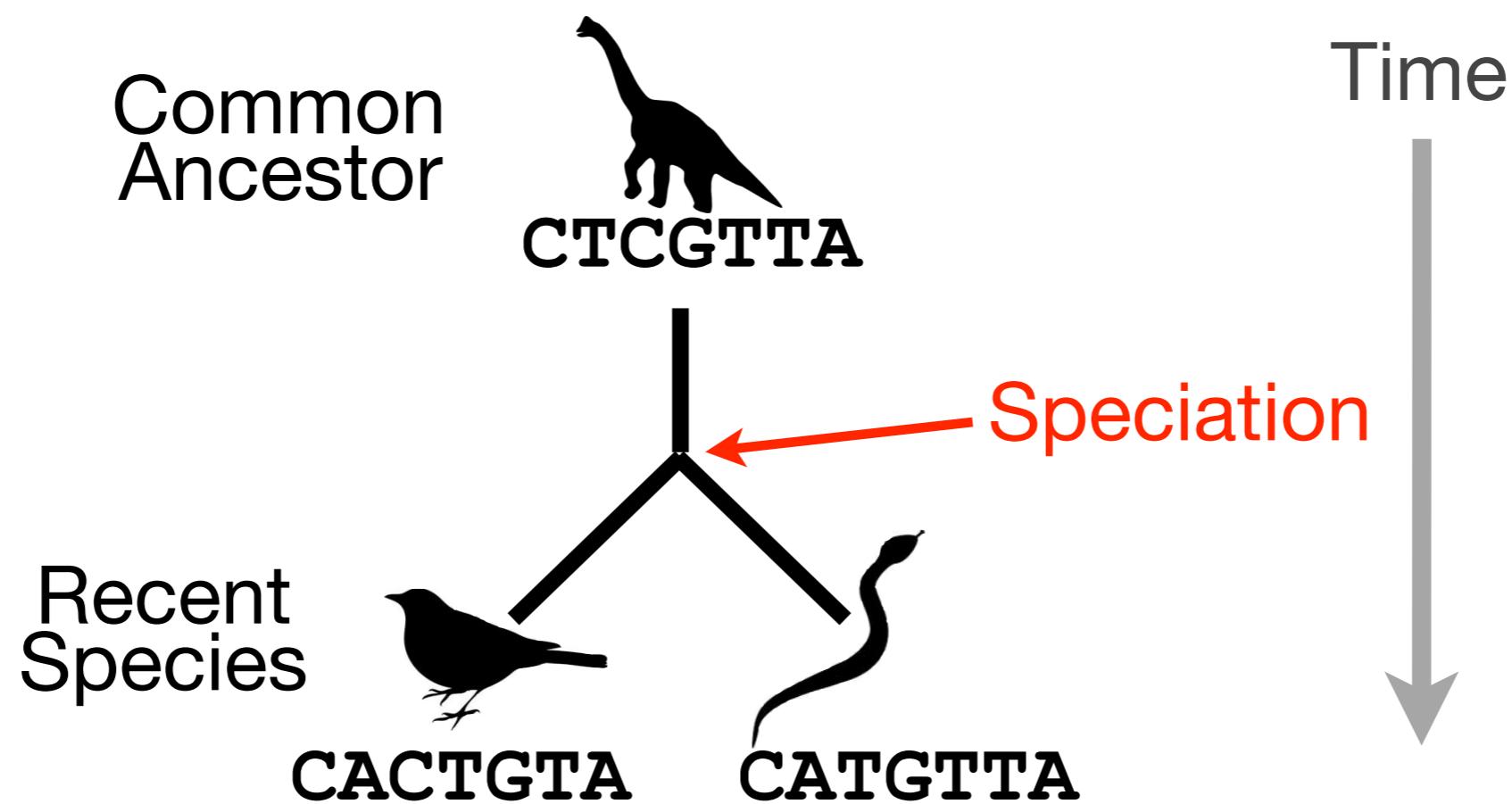
– Identity and/or Similarity

(Note. if these metrics are above a certain level we often <u>infer</u> homology)

# Orthologs tend to have similar function

Orthologs: are homologs produced by <u>speciation</u> that have diverged due to divergence of the organisms they are associated with.
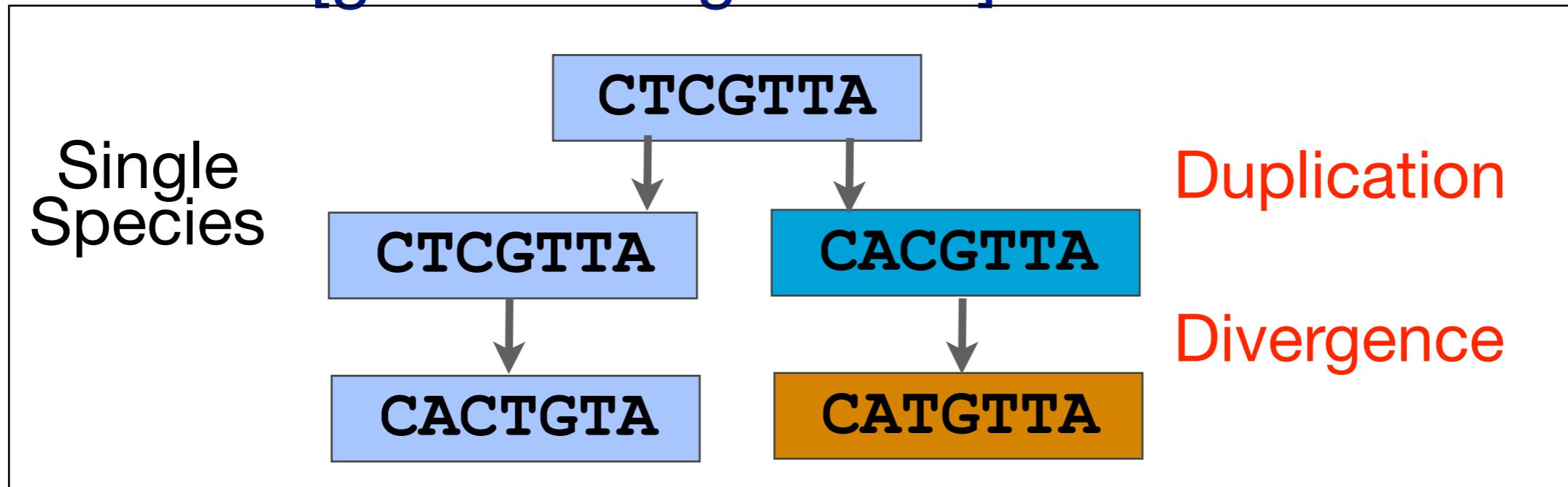
   – Ortho = [greek: straight] ... implies direct descent

Common
Ancestor

CTCGTTA

Time

Speciation

Recent
Species

CACTGTA    CATGTTA

# Paralogs tend to have slightly different functions

Paralogs: are homologs produced by gene duplication.  They represent genes derived from a common ancestral gene that <u>duplicated within an organism</u> and then subsequently <u>diverged by accumulated mutation</u>.

– Para = [greek: along side of]

# Orthologs vs Paralogs

- In practice, determining ortholog vs paralog can be a complex problem:
  – gene loss after duplication,
  – lack of knowledge of evolutionary history,
  – weak similarity because of evolutionary distance
- Homology does not necessarily imply exact same function
  – may have similar function at very crude level but play a different physiological role