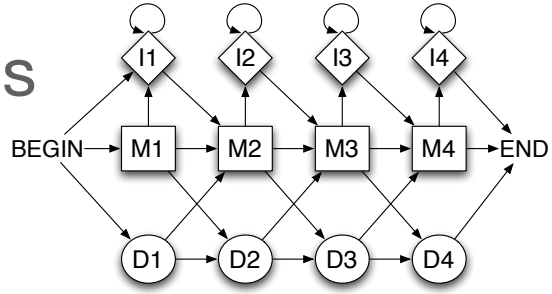
Advanced Database Searching: Sequence Patterns, Profiles & Hidden Markov Models

BI 527, Lecture #13, Fall 2011

Barry Grant
2055A Palmer Commons
Tel: 647-3113
bjgrant@umich.edu
http://thegrantlab.org



Recap on lectures 11 and 12

In previous lectures you have been introduced to:

- Common scoring matrices
  Development and application PAM & BLOSUM matrices

- Pairwise sequence alignments
  Introduction to dynamic programming
  Global alignment with Needleman-Wunsch
  Local alignment with Smith-Waterman

- BLAST database sequence searching
  A heuristic version of Smith-Waterman
  Assessing alignment Significance (Karlin-Altschul statistics, E-value, etc.)

- Multiple sequence alignments and phylogenetics
  ClustalW algorithm
  Evolutionary trees (UPGMA, NJ, MP, ML and Bayesian methods)

Outline of lectures 13 and 14

In the next two lectures we will cover:

- Sequence motifs and patterns
  Finding functional cues from conservation patterns
  Defining and using patterns and their limitations

- Sequence profiles and **p**osition **s**pecific **s**coring **m**atrices (PSSMs)
  Building and searching with profiles
  Their advantages and limitations

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

- **H**idden **M**arkov **m**odels (HMMs)
  More versatile probabilistic model for detection of remote similarities
  Defining HMMs, searching with HMMs and generating MSAs
  PFAM, SMART, GENSCAN, Developing and applying your own HMMs

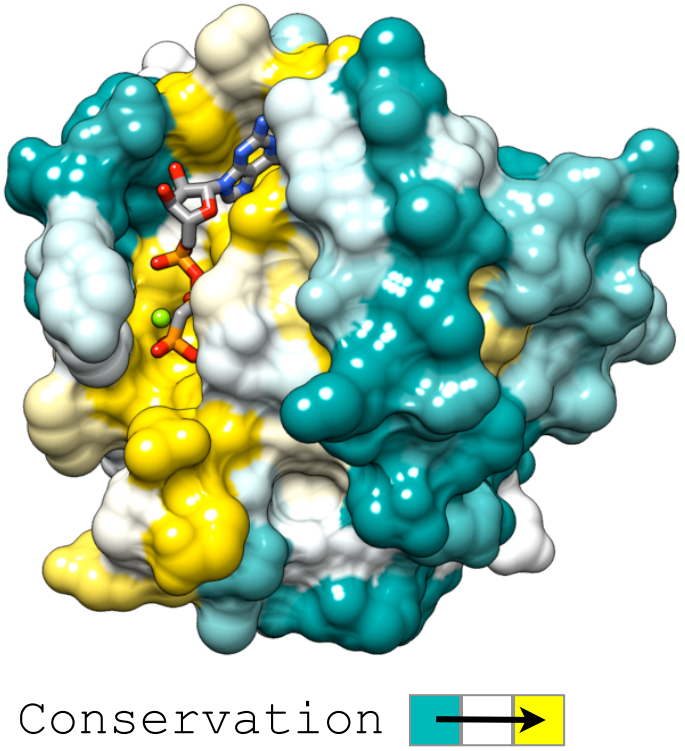
- Summary and example problems

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.

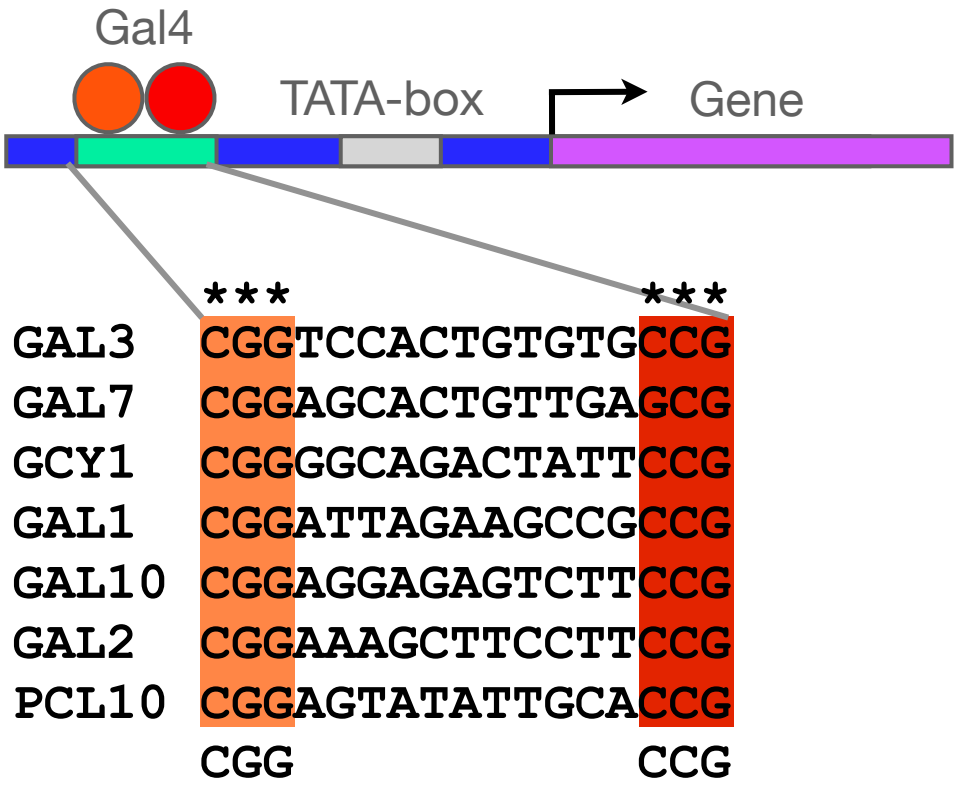
ATP/GTP-binding proteins: G-x(4)-G-K-T



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



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.

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

Image generated by:
weblogo.berkeley.edu

PROSITE is a protein pattern and profile database

Currently contains > 1600 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-[STAGDE]
The two Histidines are copper ligands

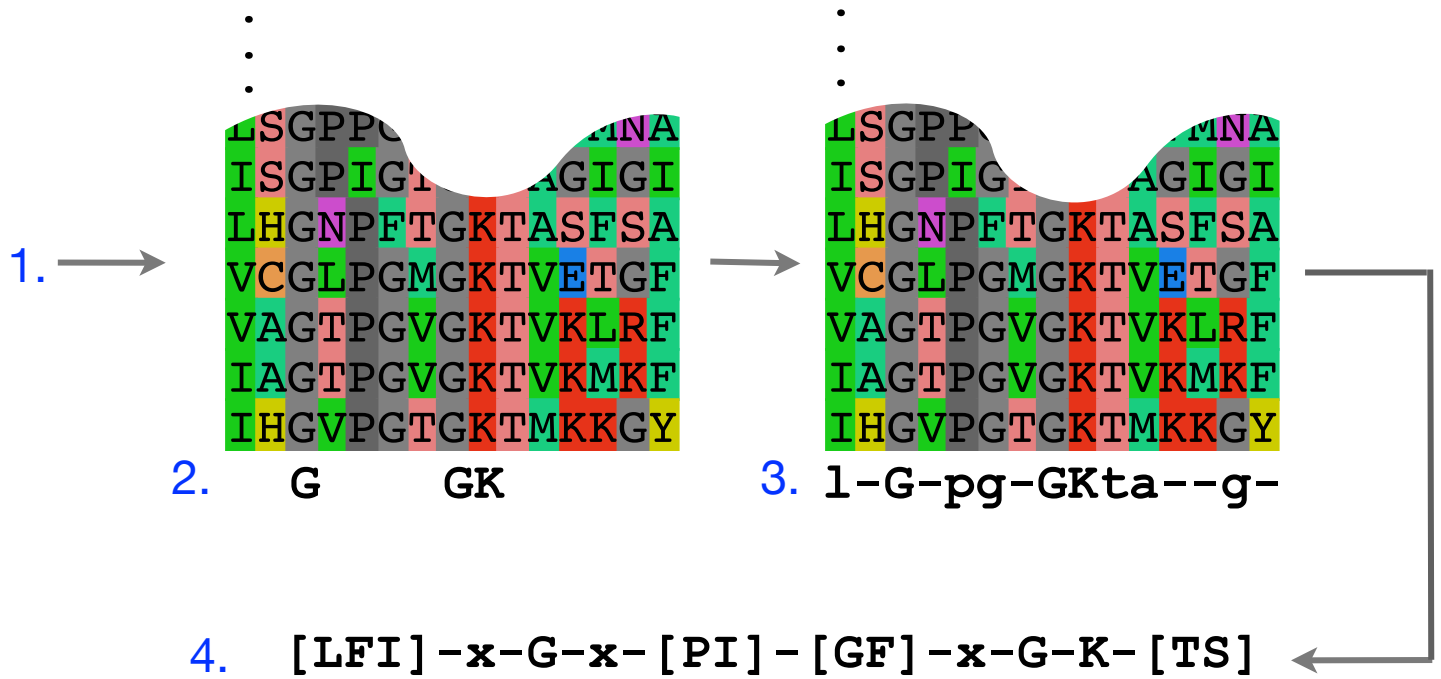
- Each position in pattern is separated with a hyphen
- x can match any residue
- [ ] are used to indicate ambiguous positions in the pattern
  e.g., [SDG] means the pattern can match S, D, or G at this position
- { } are used to indicate residues that are not allowed at this position
  e.g., {S} means 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...



Pattern advantages and disadvantages

Advantages:

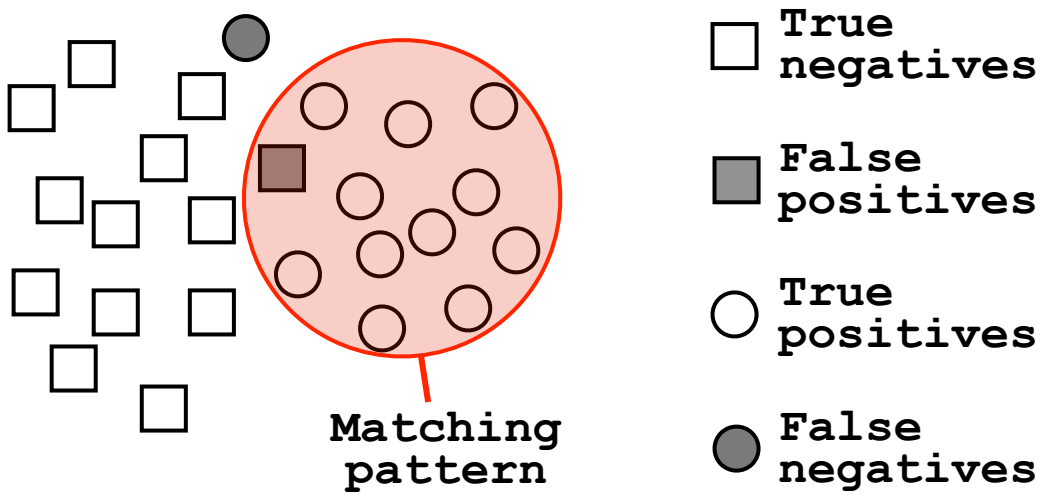
- Relatively straightforward to identify (exact pattern matching is fast)
- Patterns are intuitive to read and understand
- Databases with large numbers of protein (e.g., PROSITE) and DNA sequence (e.g., JASPER and TRANSFAC) patterns are available.

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, specificity, and PPV

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*).



Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP) **PPV = TP/(TP+FP)**

The positive predictive value (or PPV) assesses how big a proportion of the sequences matching the pattern are actually in the family of interest.
(i.e., the probability that a positive result is truly positive!)

ROC plot example

Outline of lectures 13 and 14

In the next two lectures we will cover:

- Sequence motifs and patterns
  Finding functional cues from conservation patterns
  Defining and using patterns and their limitations

- Sequence profiles and **p**osition **s**pecific **s**coring **m**atrices (PSSMs)
  Building and searching with profiles
  Their advantages and limitations

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

- **H**idden **M**arkov **m**odels (HMMs)
  More versatile probabilistic model for detection of remote similarities
  Defining HMMs, searching with HMMs and generating MSAs
  PFAM, SMART, GENSCAN, Developing and applying your own HMMs

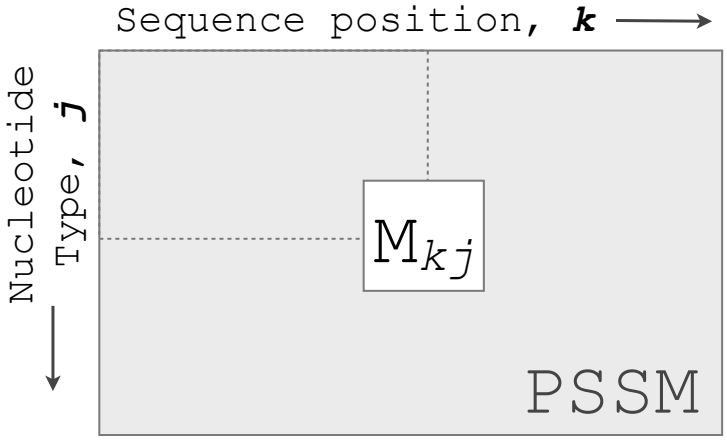
- Summary and example problems

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).



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

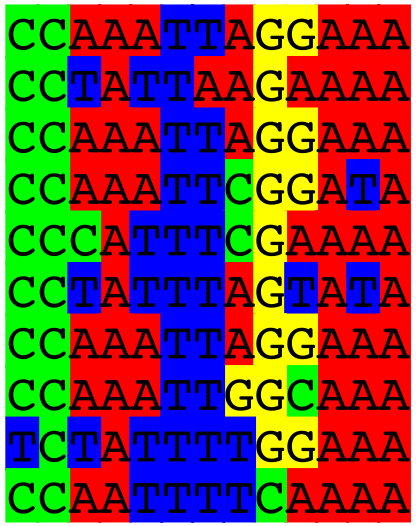
$\mathbf{M_{kj}}$ score for the *j*th nucleotide at position *k*

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

$\mathbf{p_j}$ "background" probability of nucleotide *j*

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

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**	**[ACT]**	**A**	**[AT]**	**T**	**T**	**N**	**G**	**N**	**A**	**[AT]**	**A**

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

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

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*

Adapted from Hertz and Stormo, Bioinformatics 15:563-577

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

k=1, j=C: $$M_{kj} = \log\left(\frac{C_{kj} + p_j / Z + 1}{p_j}\right) = \log\left(\frac{9 + 0.25/10 + 1}{0.25}\right) = 1.2$$

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

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**

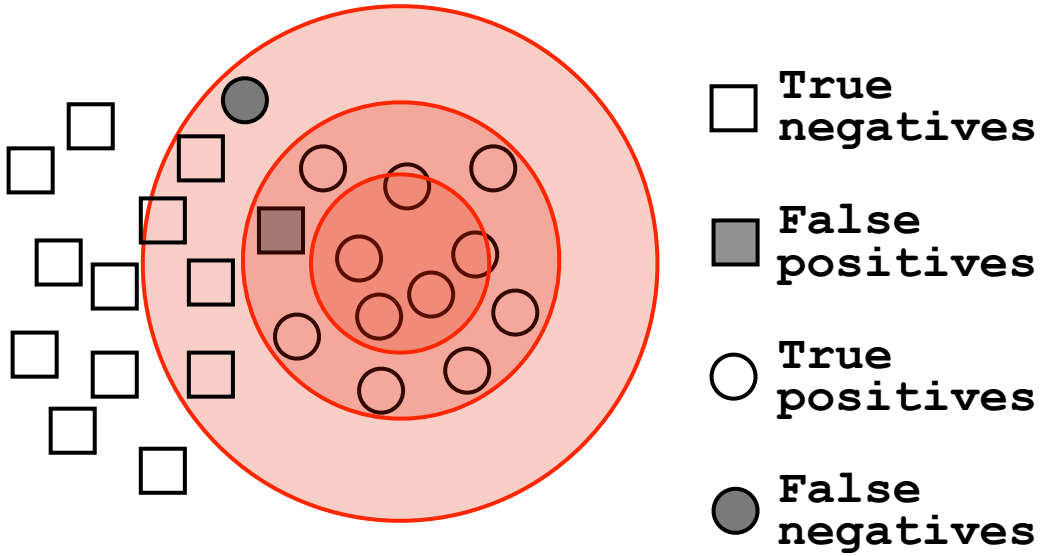
Max Score = 1.2 + 1.3 + 0.8 + 1.3 + 0.6 + 1.3 + 1.2
+ 0.6 + 1.2 + 0.6 + 1.3 + 1.1 + 1.3
= 13.8

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).

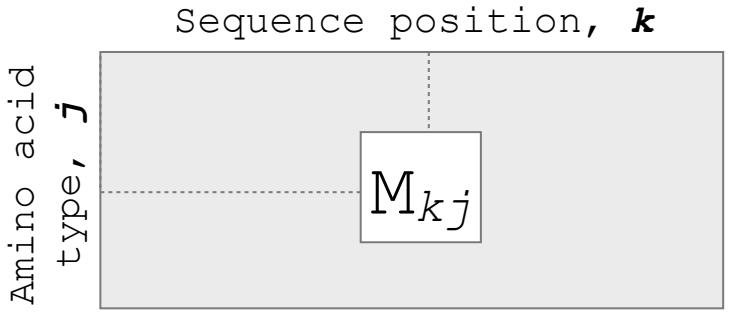



```
FP=0, FN=7, TP=5
FP=1, FN=1, TP=11
FP=5, FN=0, TP=12
```

Q. Which threshold has the best PPV (TP/(TP+FP)) ?

Protein profile calculation by the average score method

For protein profiles calculated with the *average score method* the score for a column is taken from the average of scores obtained from a substitution matrix.



$$M_{kj} = \sum_{i=1}^{20} \frac{C_{ki}}{Z} S_{ij}$$

M_{kj} Profile matrix element
(i.e. score for *j*th amino acid at the *k*th position)

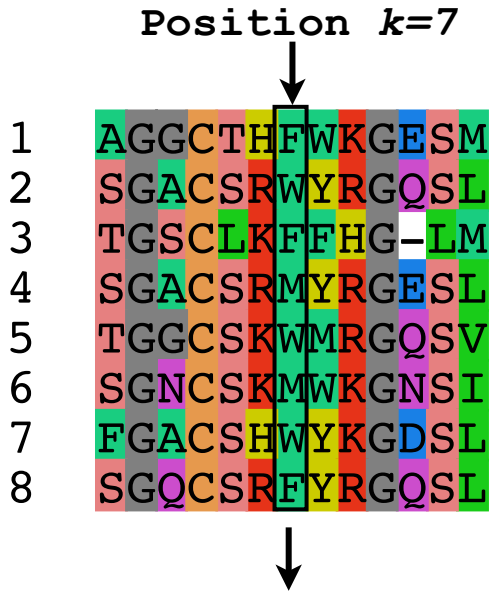
C_{ki} Number of *i*th type amino acid at position *k*

Z Total number of aligned sequences

S_{ij} Score between the *i*th and the *j*th amino acids from scoring matrix (e.g., BLOSUM62)

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

Using the average score method



F,F,F, W,W,W, M,M,

$C_{7F} = 3$, $C_{7W} = 3$, $C_{7M} = 2$,
$C_{7?} = 0$

BLOSUM62 Scores

$S_{FF} = 6$, $S_{WF} = 1$, $S_{MF} = 0$

$$M_{kj} = \sum_{i=1}^{20} \frac{C_{kj}}{Z} S_{ij}$$

$$M_{7F} = \frac{3}{8} S_{FF} + \frac{3}{8} S_{WF} + \frac{2}{8} S_{MF}$$

$$M_{7W} = \frac{3}{8} S_{FW} + \frac{3}{8} S_{WW} + \frac{2}{8} S_{MW}$$

$$M_{7M} = \frac{3}{8} S_{FM} + \frac{3}{8} S_{WM} + \frac{2}{8} S_{MM}$$

$$M_{7j} = \frac{3}{8} S_{Fj} + \frac{3}{8} S_{Wj} + \frac{2}{8} S_{Mj}$$

$M_{7F} = (3/8)(6) + (3/8)(1) + (2/8)(0) = 2.63$

Partly based on slides from K. Dunker & Z. Weng (Boston University)

Using the average score method...

Calculating the profile values for two unobserved amino acids - Y and E,

- where S_{FY}=3, S_{WY}=2, S_{MY}=-1 and S_{FE}=-3, S_{WE}=-3, S_{ME}=-2:

$$M_{7Y} = \frac{3}{8}S_{FY} + \frac{3}{8}S_{WY} + \frac{2}{8}S_{MY} = \frac{3}{8}(3) + \frac{3}{8}(2) + \frac{2}{8}(-1) \sim 1.6$$

$$M_{7E} = \frac{3}{8}S_{FE} + \frac{3}{8}S_{WE} + \frac{2}{8}S_{ME} = \frac{3}{8}(-3) + \frac{3}{8}(-3) + \frac{2}{8}(-2) \sim -2.8$$

From the above two equations, it is easy to predict that M7Y is much more favorable than M7E, even though neither Y nor E has been observed at this position (k = 7).

Limitation: With many aligned sequences, average scores from a substitution matrix will reduce specificity.

Q. Why?

Using the average score method...

Calculating the profile values for two unobserved amino acids - Y and E,

- where S_{FY}=3, S_{WY}=2, S_{MY}=-1 and S_{FE}=-3, S_{WE}=-3, S_{ME}=-2:

$$M_{7Y} = \frac{3}{8}S_{FY} + \frac{3}{8}S_{WY} + \frac{2}{8}S_{MY} = \frac{3}{8}(3) + \frac{3}{8}(2) + \frac{2}{8}(-1) \sim 1.6$$

$$M_{7E} = \frac{3}{8}S_{FE} + \frac{3}{8}S_{WE} + \frac{2}{8}S_{ME} = \frac{3}{8}(-3) + \frac{3}{8}(-3) + \frac{2}{8}(-2) \sim -2.8$$

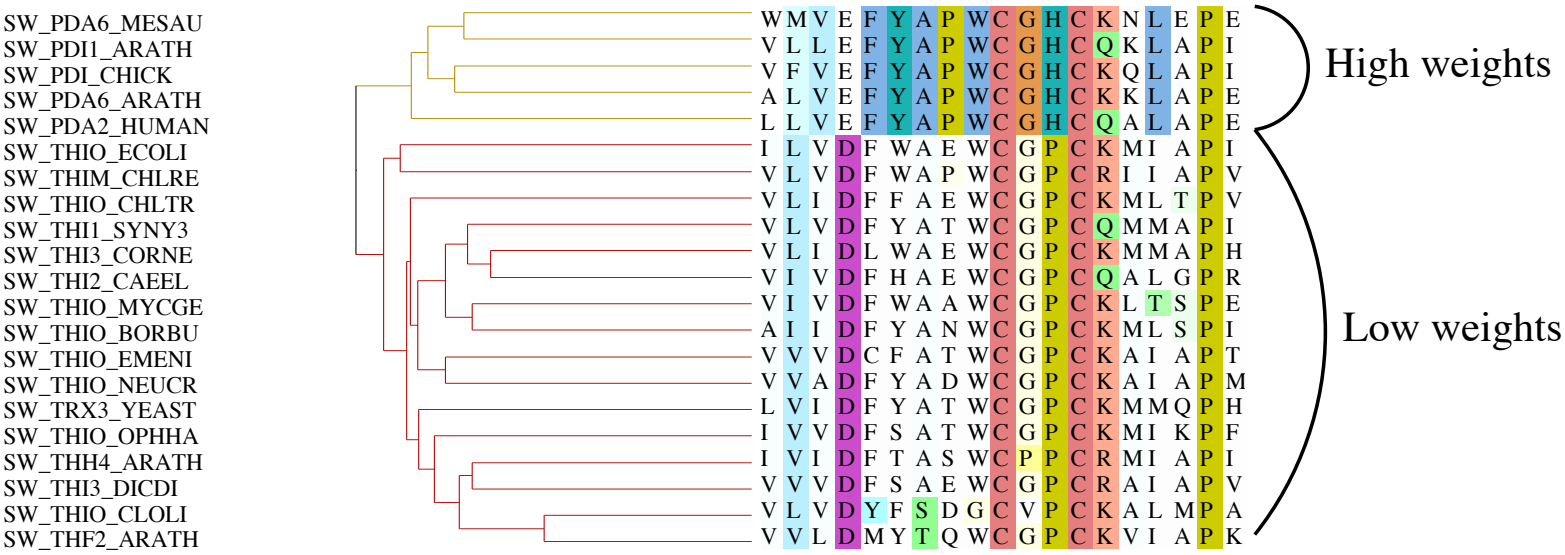
From the above two equations, it is easy to predict that M7Y is much more favorable than M7E, even though neither Y nor E has been observed at this position (k = 7).

Limitation: With many aligned sequences, scores from a substitution matrix will reduce specificity.
E.g., if alanine is in the same position in 50 diverse sequences, then substitutions of other residues are unlikely. However, the “average score” is the same as for a single sequence with alanine, and so that PSSM position will be very tolerant of non-alanines.

Sequence weighting

An MSA is often made of a few distinct sets of related sequences, or sub-families. It is not unusual that these sub-families are very differently populated, thus influencing observed residue frequencies.

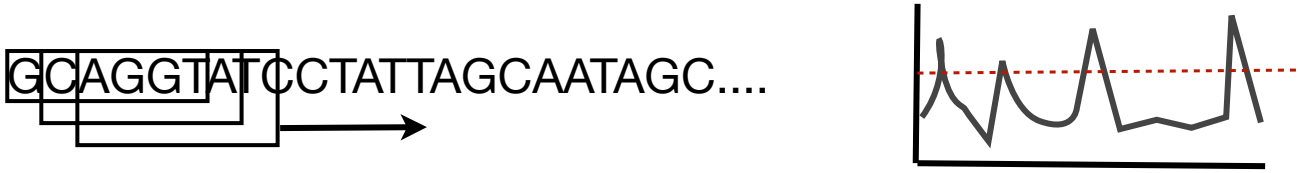
Sequences weighting attempt to compensate for this sequence sampling bias by differentially weighting sequences to reduce redundancy.



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



See example at http://coding.plantpath.ksu.edu/profile/

If we allow gaps:

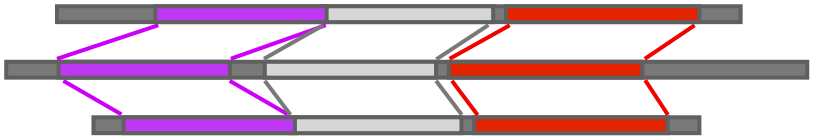
- Can use dynamic programming to align the profile to the protein sequence(s) (with gap penalties)
  We will discuss PSI-BLAST shortly...
  see Mount, Bioinformatics: sequence and genome analysis (2004)

- Can use hidden Markov Model-based methods
  We will cover HMMs in the next lecture...
  see Durbin et al., Biological Sequence Analysis (1998)

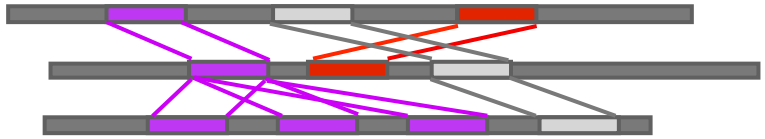
Side note: Building PSSMs from unaligned sequences

Patterns and profiles are most often built on the basis of known site equivalences (i.e. from a pre-calculated MSA).

However, a number of programs have been developed that employ local multiple alignments to search for common sequence elements in unaligned sequences.



Global similarity



Local non-consistent similarity

Gibbs *sampling* methods:
Motif Sampler - http://bayesweb.wadsworth.org/gibbs/gibbs.html
AlignAce - http://atlas.med.harvard.edu/cgi-bin/alignace.pl

Expectation maximization method:
MEME - http://meme.sdsc.edu/

See: Lawrence et al. (1993) Science. 262, 208-14

Profiles software and databases

Pftools is a package to build and search with profiles,
http://www.isrec.isb-sib.ch/ftp-server/pftools/

The package contains (among other programs):

- **pfmake** for building a profile starting from multiple alignments
- **pfsearch** to search a protein database with a profile
- **pfscan** to search a profile database with a protein

PRINTS database of PSSMs
http://bioinf.man.ac.uk/dbbrowser/PRINTS

Collection of conserved motifs used to characterize a protein

- Uses fingerprints (conserved motif groups).
- Very good to describe sub-families.

BLOCKS is another PSSMs database similar to prints
http://www.blocks.fhcrc.org

ProDom is collection of protein motifs obtained automatically using PSI-BLAST
http://prodes.toulouse.inra.fr/prodom/doc/prodom.html

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

Outline of lectures 13 and 14

In the next two lectures we will cover:

- Sequence motifs and patterns
  Finding functional cues from conservation patterns
  Defining and using patterns and their limitations

- Sequence profiles and **p**osition **s**pecific **s**coring **m**atrices (PSSMs)
  Building and searching with profiles
  Their advantages and limitations

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

- **H**idden **M**arkov **m**odels (HMMs)
  More versatile probabilistic model for detection of remote similarities
  Defining HMMs, searching with HMMs and generating MSAs
  PFAM, SMART, GENSCAN, Developing and applying your own HMMs

- Summary and example problems

Half time break...

See PSSM example at http://coding.plantpath.ksu.edu/profile/

Outline of lectures 13 and 14

In the next two lectures we will cover:

- Sequence motifs and patterns
  Finding functional cues from conservation patterns
  Defining and using patterns and their limitations

- Sequence profiles and **p**osition **s**pecific **s**coring **m**atrices (PSSMs)
  Building and searching with profiles
  Their advantages and limitations

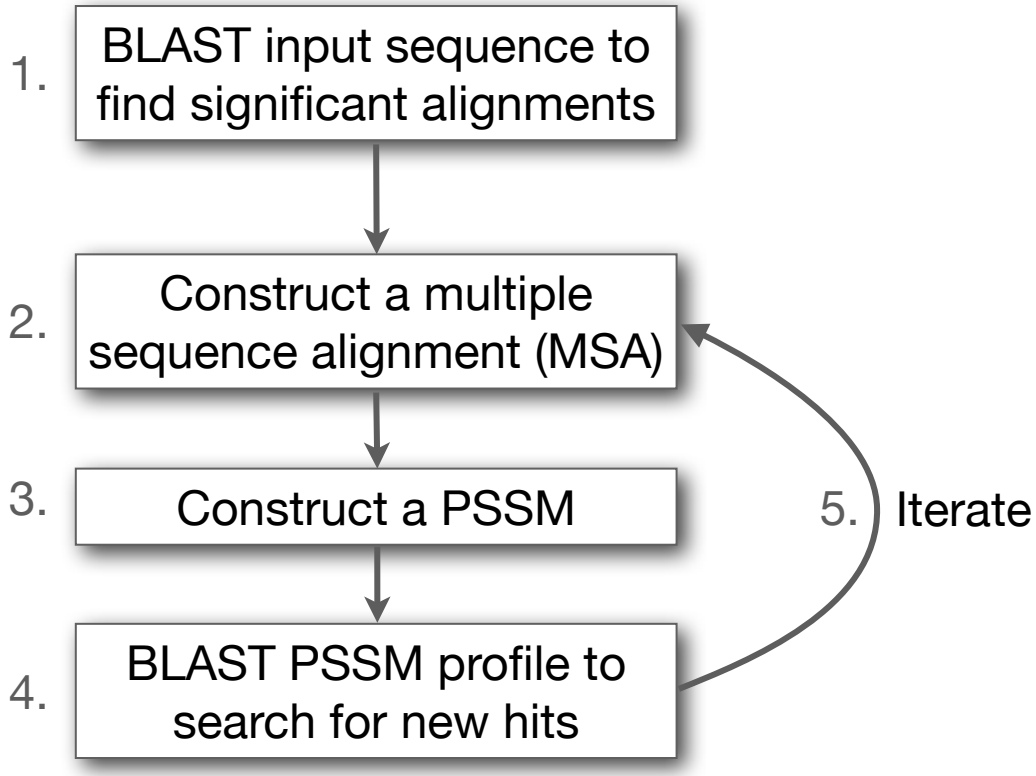
- PSI-BLAST algorithm
  Application of iterative PSSM searching to improve BLAST sensitivity

- **H**idden **M**arkov **m**odels (HMMs)
  More versatile probabilistic model for detection of remote similarities
  Defining HMMs, searching with HMMs and generating MSAs
  PFAM, SMART, GENSCAN, Developing and applying your own HMMs

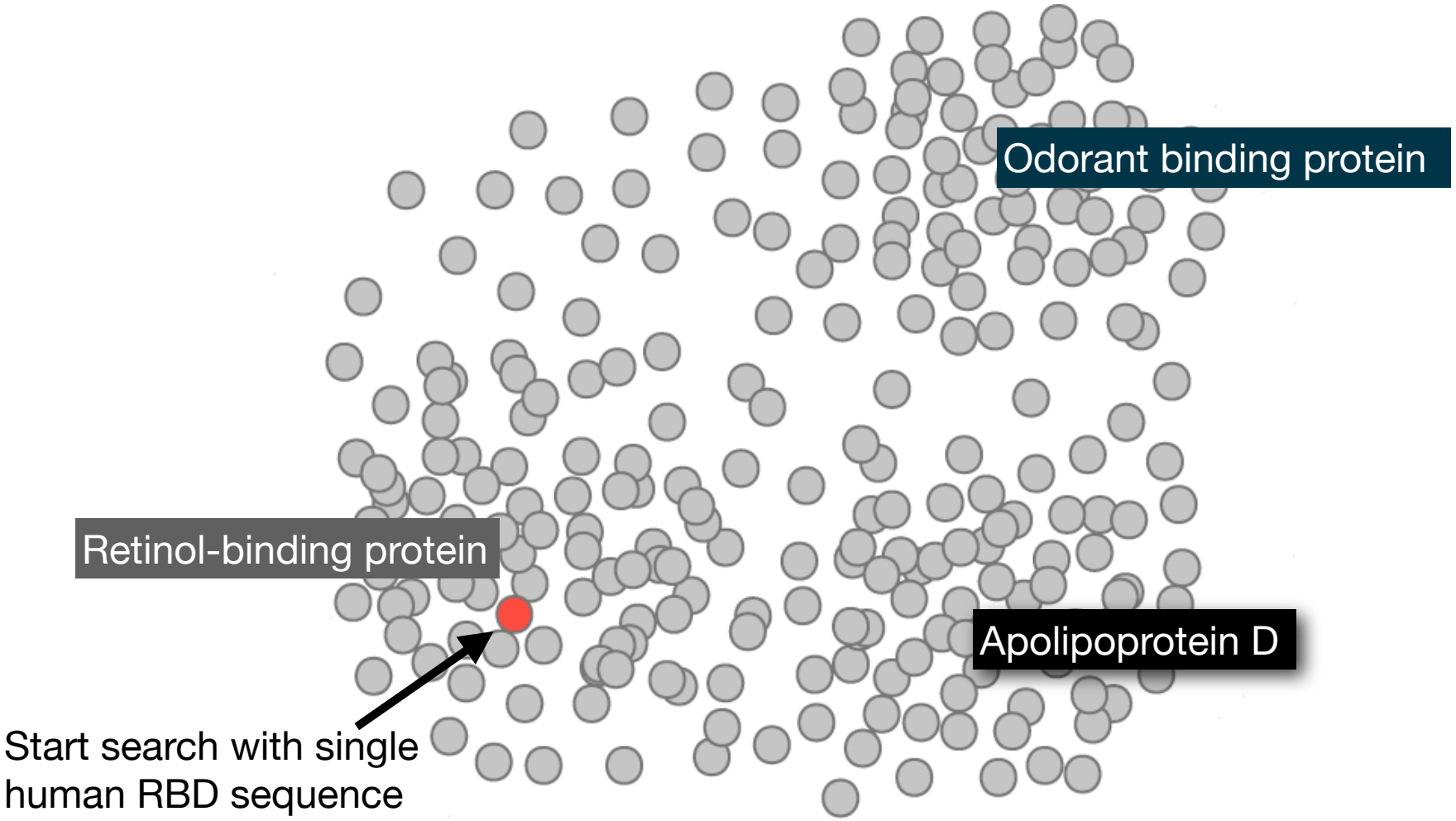
- Summary and example problems

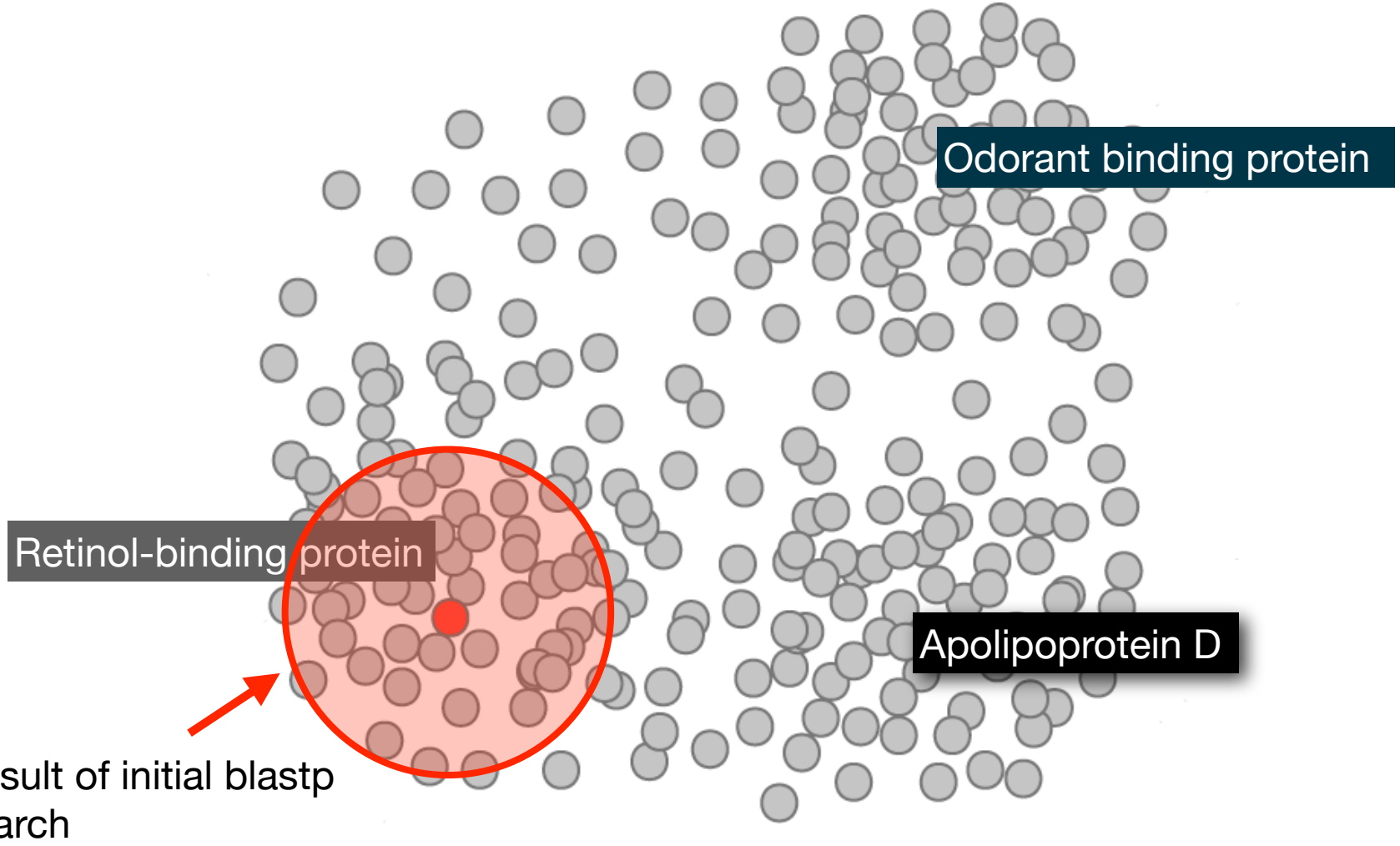
PSI-BLAST: Position-Specific Iterated BLAST

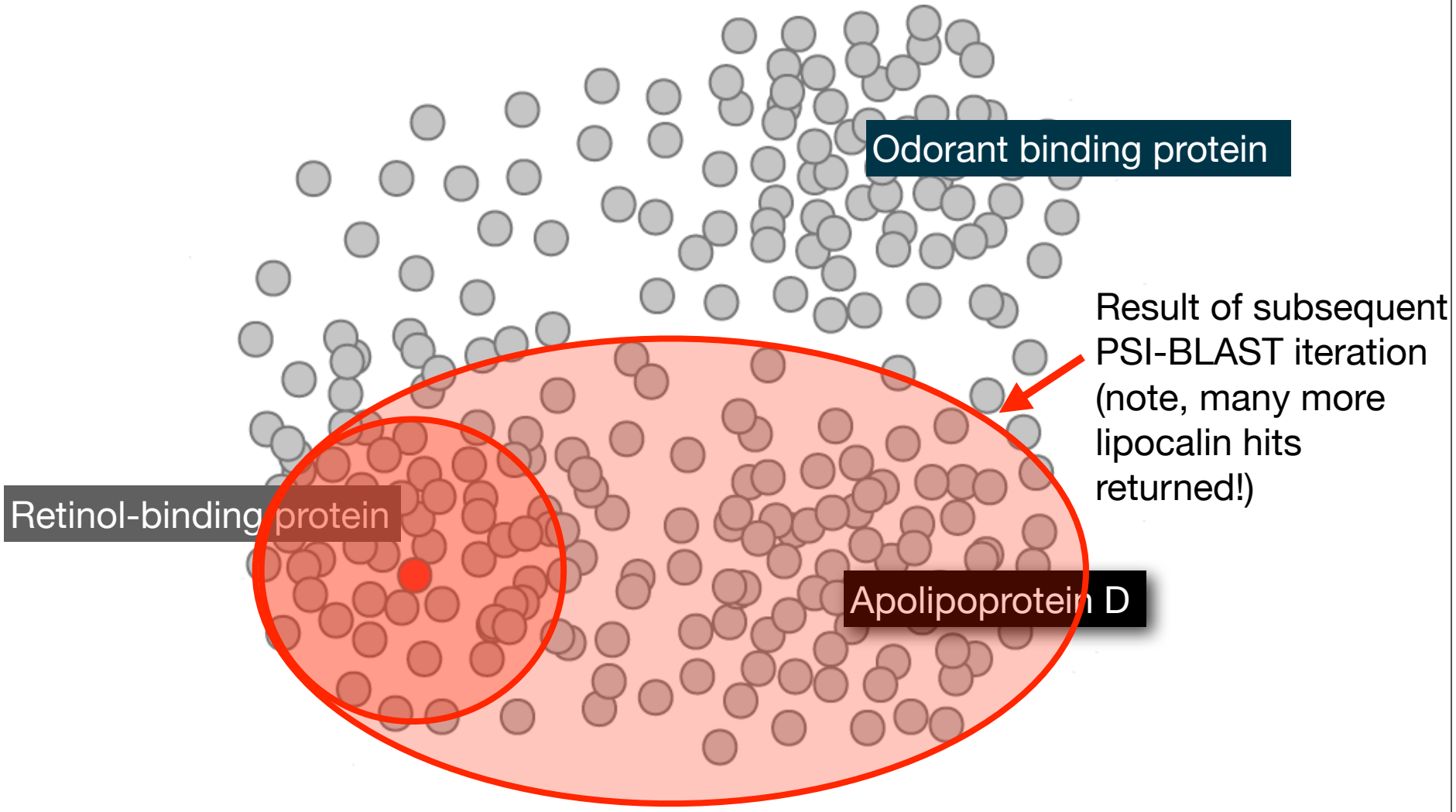
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST

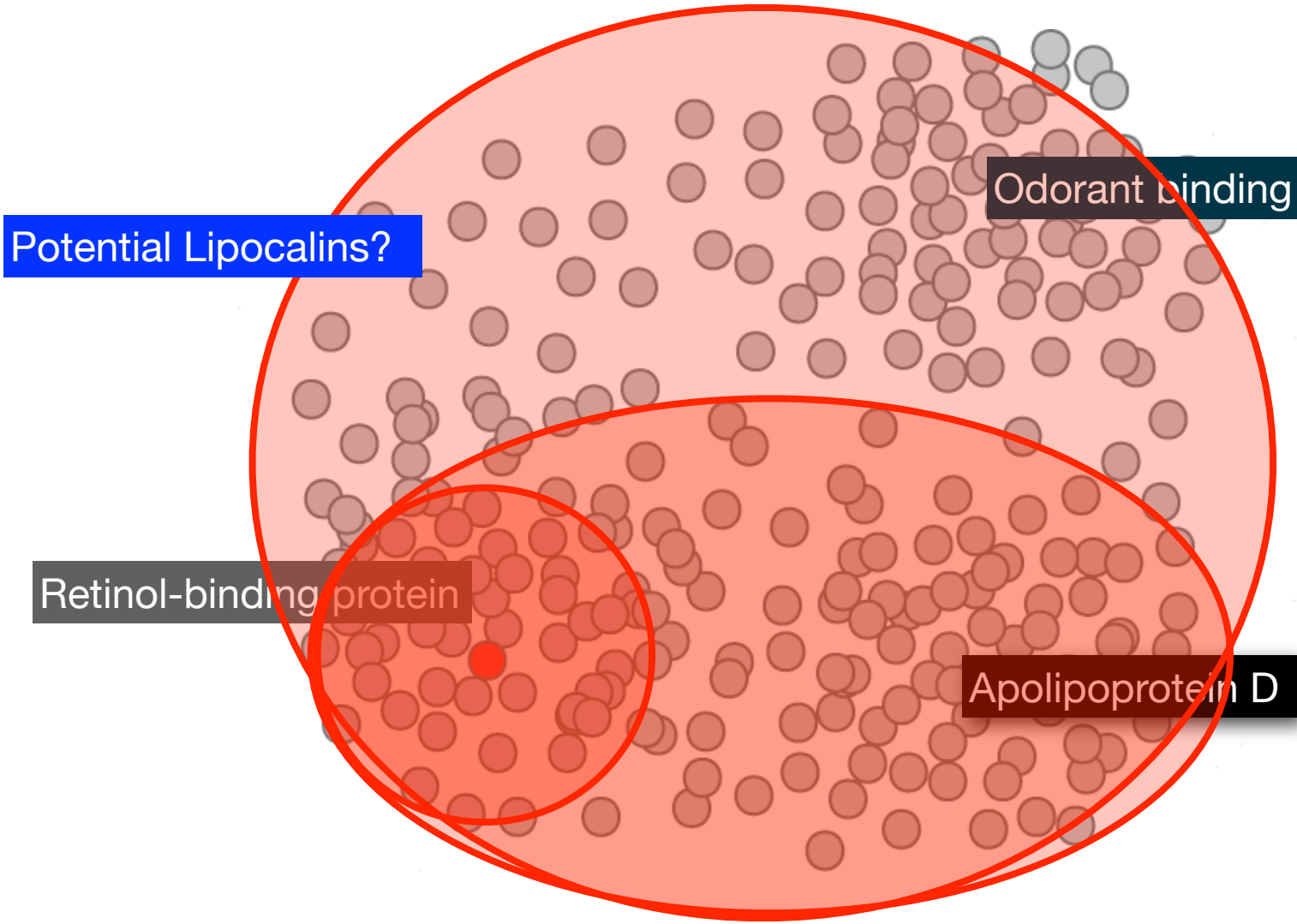


(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)


Odorant binding protein
Retinol-binding protein
Apolipoprotein D
Start search with single
human RBD sequence


Odorant binding protein
Retinol-binding protein
Apolipoprotein D
Result of initial blastp search


Odorant binding protein
Result of subsequent PSI-BLAST iteration (note, many more lipocalin hits returned!)
Retinol-binding protein
Apolipoprotein D



Result of later PSI-BLAST iteration (note, potential "corruption"!)

PSI-BLAST returns dramatically more hits

PSI-BLAST frequently returns many more hits with significant E-values than blastp

The search process is continued iteratively, typically about five times, and at each step a new PSSM is built.

- 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.

(**a**) Iteration 1

```
>ref|NP_001638.1| apolipoprotein D precursor [Homo sapiens]
Length=189

 Score = 57.4 bits (137),  Expect = 3e-07, Method: Composition-based stats.
 Identities = 47/151 (31%), Positives = 78/151 (51%), Gaps = 39/151 (25%)

Query  29   VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC  88
            V+ENFD  ++ G WY + +K P        I A +S+ E G        ++++LN  ++
Sbjct  33   VQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMENG--------KIKVLNQ-ELR  82

Query  89   ADMVGTFTDTE---------DPAKFKMKY-WGVASFLQKGNDDHWIVDTDYDTYAVQYSC  138
            AD  GT    E          +PAK ++K+ W + S        +WI+ TDY+ YA+ YSC
Sbjct  83   AD--GTVNQIEGEATPVNLTEPAKLEVKFSWFMPS------APYWILATDYENYALVYSC  134

Query  139  ----RLLNLDGTCADSYSFVFSRDPNGLPPE  165
                +L ++D      ++++ +R+PN LPPE
Sbjct  135  TCIIQLFHVD------FAWILARNPN-LPPE  158
```

(**b**) Iteration 2

```
>ref|NP_001638.1| apolipoprotein D precursor [Homo sapiens]
Length=189

 Score =  175 bits (443),  Expect = 1e-42, Method: Composition-based stats.
 Identities = 45/163 (27%), Positives = 77/163 (47%), Gaps = 31/163 (19%)

Query  14   GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSA  73
              G+A    +  +  V+ENFD  ++ G WY + +K P        I A +S+ E G++
Sbjct  18   AEGQAFHLGKCPNPPVQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMENGKIKV  76

Query  74   TAK-----GRVRLLNNWDVCADMVGTFTDTEDPAKFKMKY-WGVASFLQKGNDDHWIVDT  127
              +     G V  +           T  +  +PAK ++K+ W + S        +WI+ T
Sbjct  77   LNQELRADGTVNQIEG-------EATPVNLTEPAKLEVKFSWFMPS------APYWILAT  123

Query  128  DYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA  166
            DY+ YA+ YSC     L ++D      ++++ +R+PN LPPE
Sbjct  124  DYENYALVYSCTCIIQLFHVD------FAWILARNPN-LPPET  159
```

(**c**) Iteration 3

```
>ref|NP_000597.1| complement component 8, gamma polypeptide [Homo sapiens]
Length=202

 Score =  104 bits (260),  Expect = 2e-21, Method: Composition-based stats.
 Identities = 40/186 (21%), Positives = 74/186 (39%), Gaps = 29/186 (15%)

Query  24   VSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETG-QMSATAKGRVRLL  82
            +S+ + K NFD  +F+GTW  +A          +    AE +      Q +A A    R L
Sbjct  33   ISTIQPKANFDAQQFAGTWLLVAVGSACRFLQEQGHRAEATTLHVAPQGTAMAVSTFRKL  92

Query  83   NNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQY------  136
            +   +C  +   + DT    +F ++  G      +G     + +TDY ++AV Y
Sbjct  93   DG--ICWQVRQLYGDTGVLGRFLLQARGA-----RGAVHVVVAETDYQSFAVLYLERAGQ  145

Query  137  -SCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGR  195
             S +L       +DS    F +       EA       ++++    +Y      G+C+
Sbjct  146  LSVKLYARSLPVSDSVLSGFEQRVQ----EA----HLTEDQIFYFPKY------GFCEAA  191

Query  196  SERNLL  201
             + ++L
Sbjct  192  DQFHVL  197
```

blastp E-value for this hit was 0.27

Example PSI-BLAST PSSM at iteration 3

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

20 amino acids types

Query residues/positions

		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
...																					

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 to use proficiently

Outline of lectures 13 and 14

In the next two lectures we will cover:

- Sequence motifs and patterns
  Finding functional cues from conservation patterns
  Defining and using patterns and their limitations

- Sequence profiles and **p**osition **s**pecific **s**coring **m**atrices (PSSMs)
  Building and searching with profiles
  Their advantages and limitations

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

- **H**idden **M**arkov **m**odels (HMMs)
  More versatile probabilistic model for detection of remote similarities
  Defining HMMs, searching with HMMs and generating MSAs
  PFAM, SMART, GENSCAN, Developing and applying your own HMMs

- Summary and example problems

Homework questions

Due 10/27/11

From homework 7

B3. We know that myoglobin is homologous to alpha globin and beta globin; all are vertebrate members of a globin superfamily. Indeed myoglobin shares a very similar three-dimensional structure with alpha and beta globin.

a) Using human myoglobin (P02144) as a query in a blastp search against **human RefSeq** proteins, what E-value and score does “hemoglobin subunit alpha” and “hemoglobin subunit beta” receive?

b) Perform the same search using PSI-BLAST, what scores do these proteins receive in iteration 2?

c) How many PSI-BLAST iterations do you think are sensible for a reasonable coverage of the globin superfamily? Please explain your answer...

TIP: Find the FASTA sequence for P02144 at http://www.uniprot.org
Use NCBI blastp and PSI-BLAST from http://blast.ncbi.nlm.nih.gov/

That’s it!