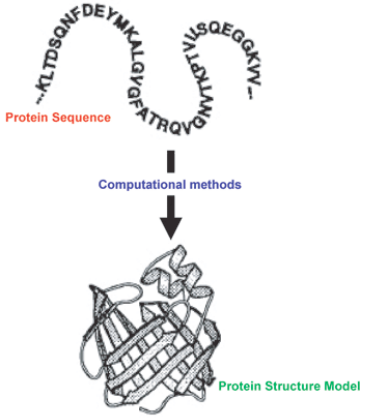
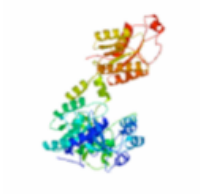
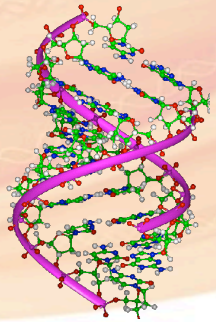




Bioinformatics Computational Methods 1 - BIOL 6308



November 12th 2013

http://155.33.203.128/cleslin/home/teaching6308F2013.php

Last Time

- The Characters in Sequence Alignments
- Sliding Window
- Modular Nature of Proteins
- Evolutionary Basis of Sequence Alignment
  - When is Homology Real?
  - Conserved Regions and Similar Biological Functions
- Scoring and Gaps
- Quantitative Measures of Sequence Similarity and Differences
- Scoring Matrices
- PAM and BLOSUM
- Why Do We Use the Log of the Odd Scores

Evolutionary Basis of Sequence Alignment

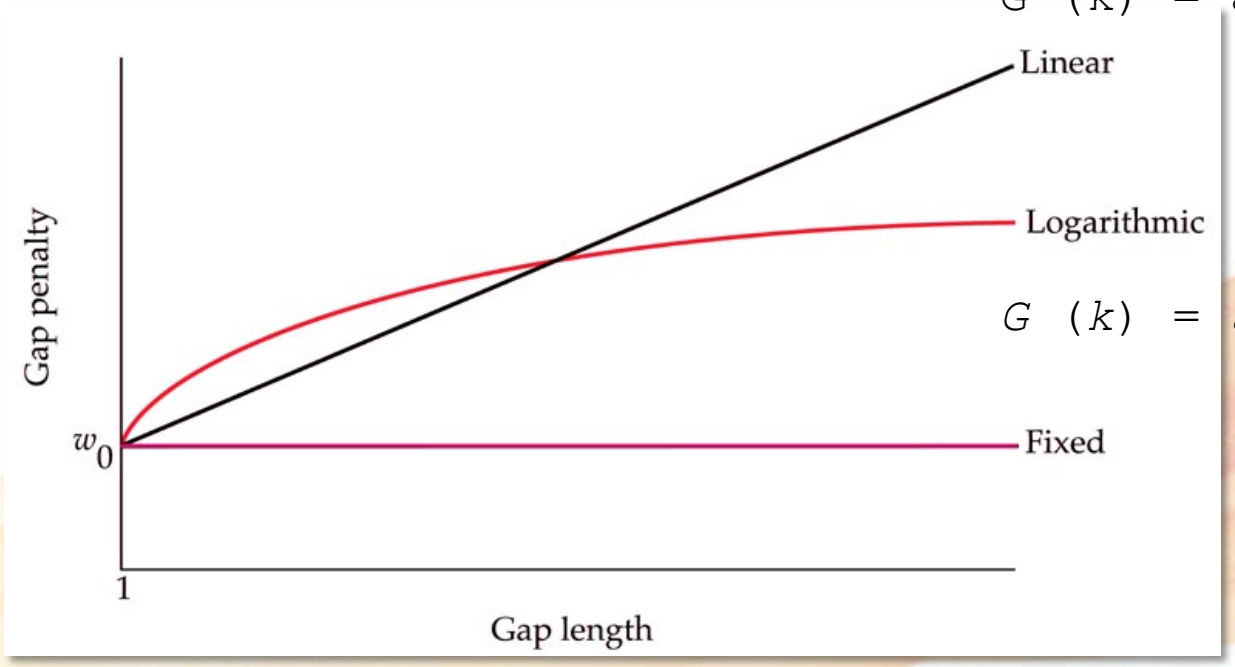
- **Identity** is a m*easure* made on an alignment
  - Quantity that describes how much two sequences are alike in the strictest terms
  - Sequence A can be "32 % identical to" Sequence B
- **Similarity (Positives)** is a measure of how close two amino acids are to identical
  - Quantity that relates how much two amino acid sequences are alike
  - DNA does not have similarity
  - For instance, isoleucine and leucine are similar
- **Homology** is a *property* that exists or does not exist
  - Proteins or genes are defined as homologous if they can be said to have shared an ancestor
  - Sequence A **IS** or **IS NOT** homologous to Sequence B
    - Sequence A cannot be "40% homologous to" B
  - Is a **conclusion** drawn from data suggesting that two genes share a common evolutionary history
    - Established on the basis of measured **similarity** or **identity**

Three Main Systems for Gap Extension Penalties:

- **Fixed gap-penalty system**
  - 0 gap-extension costs
- **Linear gap-penalty system**
  - Gap-extension cost is calculated by:
    - multiplying the(gap length -1)
    - by a **constant** representing the gap-extension penalty for increasing the gap by 1
    - Alignment with fewer gaps is favored over the alignment with more gaps
    - Overall penalty for one large gap is the same as for many small gaps that add up to the same penalty as the large one
- **Logarithmic gap-penalty system**
  - Gap-extension penalty increases with the logarithm of the gap length, i.e., slower
  - Remember, loss of domains or large chunks of a sequence are common between distantly related proteins

Gap Penalty is Function of Gap Length

$G(k) = a + bk$

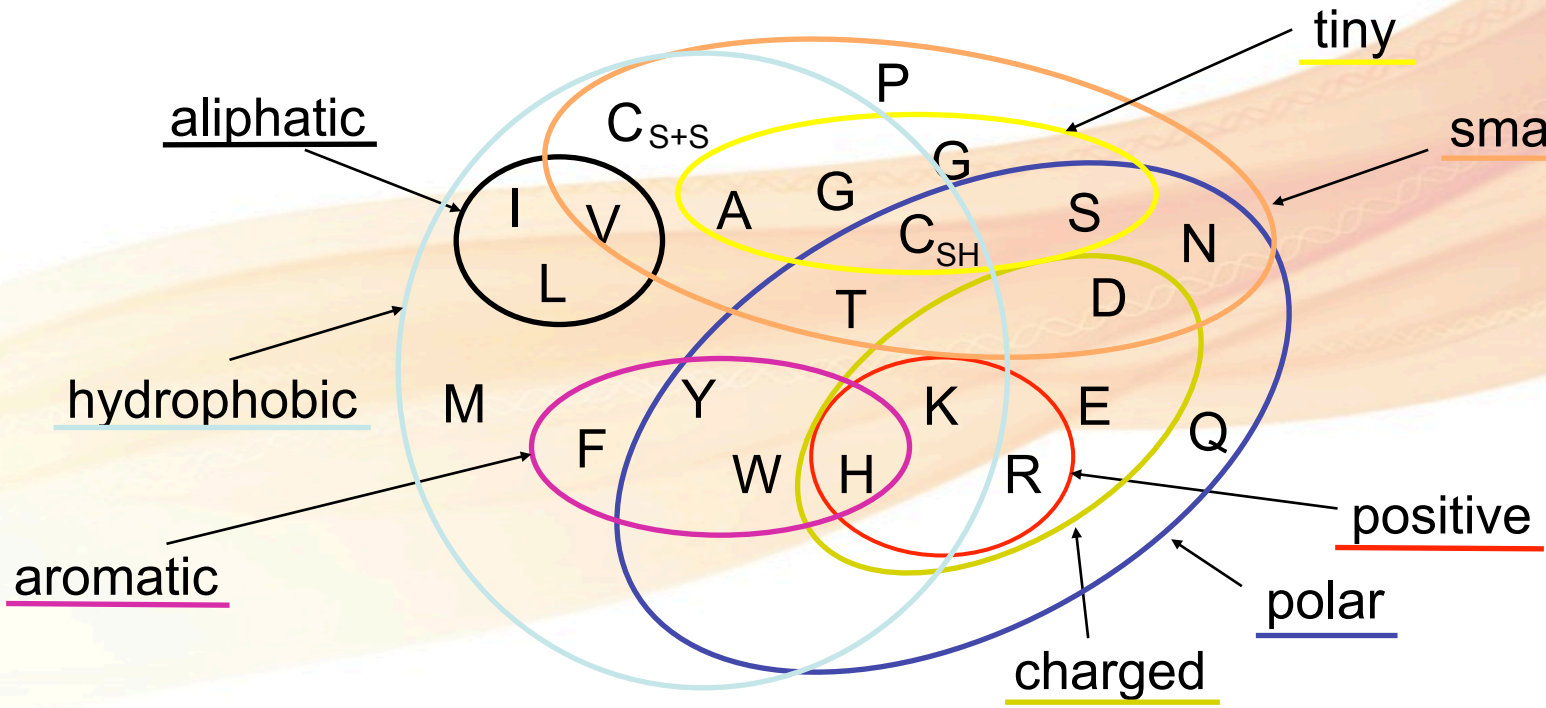


$G(k) = a + c \ln k$

Measures of Sequence Similarity

Scoring Matrix Inherently Based On:

- Conservative a.a. substitution due to similar physicochemical properties
  - Isoleucine for Valine (both small, aliphatic)
  - Serine for Threonine (both polar, small)
  - ...



BLOSUM62 and PAM120 Matrices

The colors represent different physiochemical properties.

Note that some substitutions are positive, which indicates that they occur more frequently than chance

The average value is negative: it is more likely than an a.a. will stay the same than change

The diagonal values are unchanged amino acids, all of which have positive values. Some are less changeable than others: tryptophan and cysteine especially.

(A)

BLOSUM62

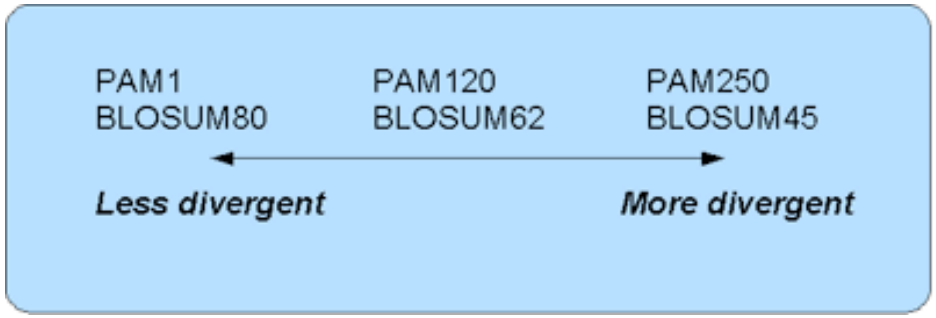
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5							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

(B)

PAM120

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

Which Matrix to Use?



- BLOSUM62 is default, and probably the best choice
- The selection of a wrong scoring matrix will strongly influence the outcome of BLAST
  - Closely related sequences choose BLOSUM80 or PAM1
    - Created for highly similar alignments
  - Distantly related sequence use BLOSUM45 or PAM250

Matrix Limitations

- The probability of a mutation in one position of a sequence is only dependent on which AA is in the position
- Do not take into account long range interactions between residues
- They assume that identical residues are equal
  - Whereas in real life
    - A residue in the active site has other evolutionary constraints than the same **type of residue** outside of the active site

Adopted from Volker Flegel

PAM vs BLOSUM

- PAM properties:
  - Based on an explicit evolutionary model
  - Assumes that more distant changes are reflection of repeated short-term changes
  - And therefore can work over a wide range of divergences
- PAM limitations:
  - Assumptions of model clearly violated
  - Each position is context dependent
    - Rates of substitution vary across and within proteins
    - Local 3-D environments vary
  - Rare changes more prone to sampling error (changes in similar sequences occur at sites that are less constrained)

PAM versus BLOSUM

- BLOSUM properties:
  - Not based on an explicit evolutionary model
    - purely empirically derived
  - Based on sequence comparisons covering a broad range of divergences
- BLOSUM limitations:
  - Restricted to a subset of conserved domains

Review

- What does the 1 in PAM1 mean?
- What does the 62 in BLOSUM62 mean?
- Why do:
  - Leucine and isoleucine get a BLOSUM62 score of 2
  - Whereas leucine and aspartic acid get a score of -4?
- What is the difference between a gap open and a gap extension penalty?

Review

- **What does the 1 in PAM1 mean?**
  - PAM1 matrix estimates what rate of substitution would be expected if 1% of the amino acids had changed
  - Used for calculating other matrices
    - by assuming that repeated mutations would follow the same pattern as those in the PAM1 matrix
    - and multiple substitutions can occur at the same site
- **What does the 62 in BLOSUM62 mean?**
  - Matrix created by clustering all sequences that were more similar than a given percentage
    - into one single sequence
    - and then comparing those sequences only; thus reducing the contribution of closely related sequences

Review

- **Why does leucine and isoleucine get a BLOSUM62 score of 2, whereas leucine and aspartic acid get a score of -4?**
  - Leucine and isoleucine are **biochemically very similar**
  - consequently, a **substitution** of one for the other is **much more likely to occur**
- **What is the difference between a gap open and a gap extension penalty?**
  - When we assign a score to an observed gap in an alignment, we charge a larger penalty for the **first** gapped position than for **subsequent** positions in the same gap

Sequence Comparison Overview

- Problem: Find the “best” alignment between a query sequence and a target sequence
- To solve this problem, we need
  - **A method for scoring alignments, and**
  - **An algorithm** for finding the alignment with the best score
    - **The algorithm** for finding the "optimal alignment" is dynamic programming
    - We will come back to this
- The alignment score is calculated using
  - A substitution matrix and gap penalties
  - Let's take a closer look at what you should know how to do:

You Should be Able To Do the Following:

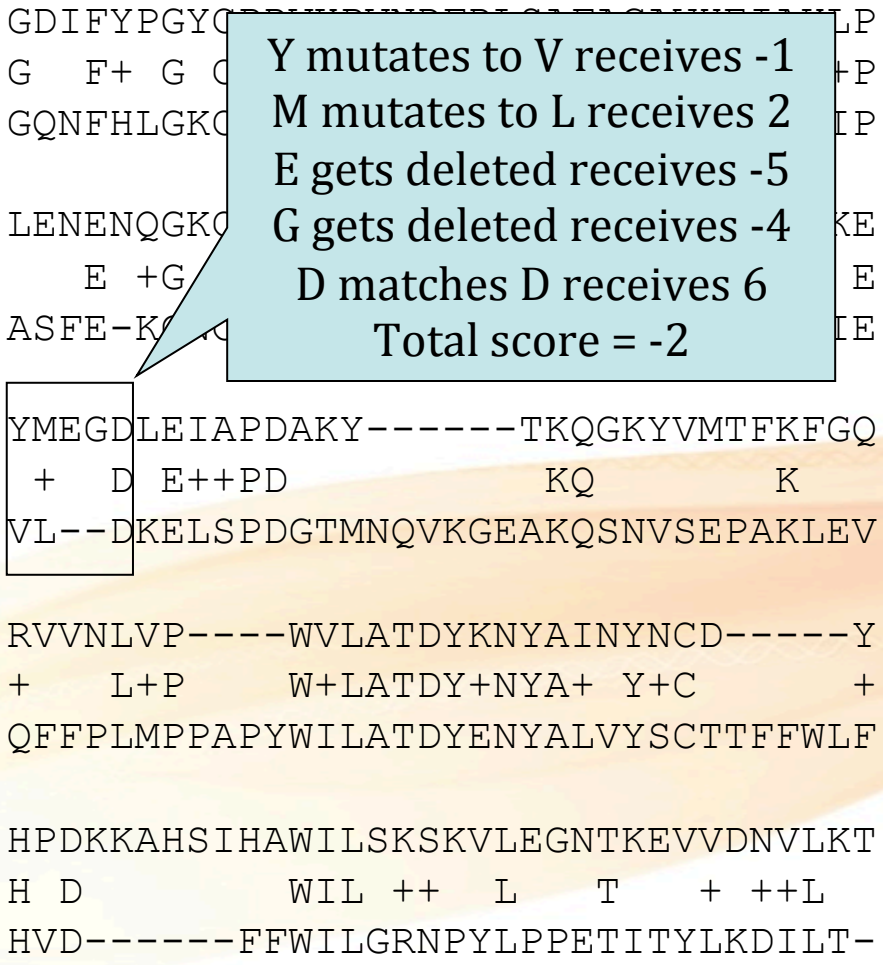
```
GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLP
G  F+ G CP      +FD+  + G W+EI K+P
GQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIP

LENENQGKCTIAEYKYDGKKASVYNSFVSNGVKE
   E +G C  A Y           S + NG  E
ASFE-KGNCIQANY----------SLMENGNIE

YMEGDLEIAPDAKY------TKQGKYVMTFKFGQ
 +  D E++PD          KQ       K
VL--DKELSPDGTMNQVKGEAKQSNVSEPAKLEV

RVVNLVP----WVLATDYKNYAINYNCD-----Y
+   L+P    W+LATDY+NYA+ Y+C      +
QFFPLMPPAPYWILATDYENYALVYSCTTFFWLF

HPDKKAHSIHAWILSKSKVLEGNTKEVVDNVLKT
H D        WIL ++  L   T   + ++L
HVD------FFWILGRNPYLPPETITYLKDILT-
```

So If I was to give you the matrix, gap opening, and gap extension penalties… You should be able to calculate the score of the boxed region

BLOSUM62

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

You Should be Able To Do the Following:

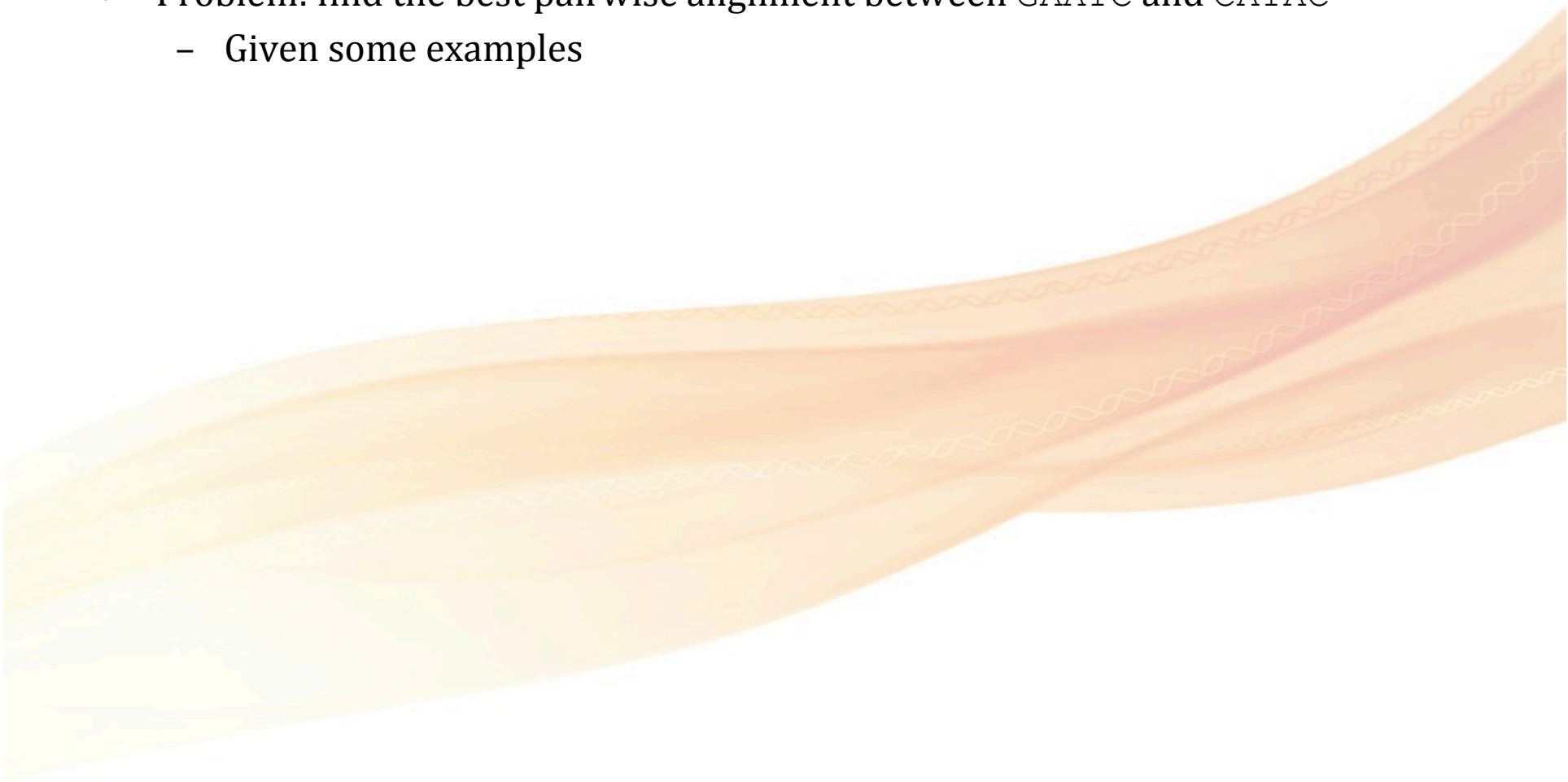


Given:
gap opening penalty of -5
gap extension penalty of -4

Or, I might give you a simple Alignment problem

A Simple Alignment Problem

- Problem: find the best pairwise alignment between `GAATC` and `CATAC`
  - Given some examples

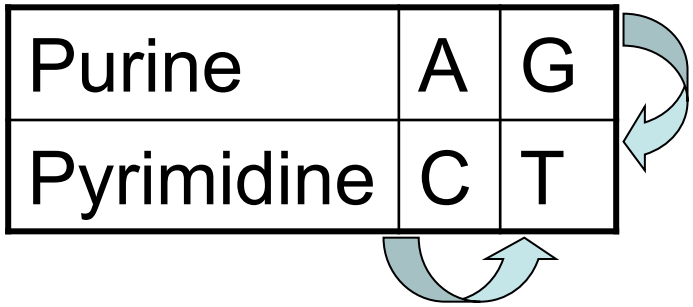


Scoring Alignments

```
GAATC      GAAT-C      -GAAT-C
CATAC      C-ATAC      C-A-TAC

GAATC-     GAAT-C      GA-ATC
CA-TAC     CA-TAC      CATA-C
```

- We need way to objectively measure the quality of an alignment
- Alignment scores consist of two parts: a substitution matrix, and a gap penalty
- Which one scores the highest?

Scoring Aligned Bases



GAATC

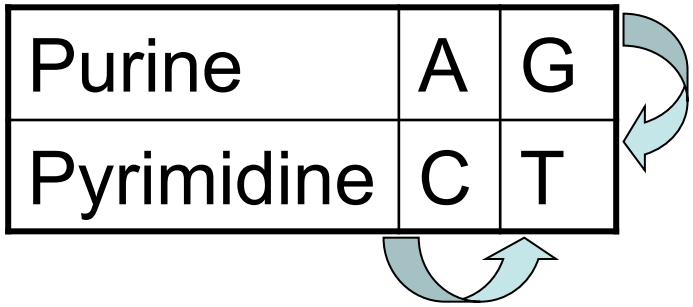
CATAC

-5 + 10 + -5 + -5 + 10 = 5

Given a hypothetical substitution matrix:

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Scoring Aligned Bases



GAAT–C

CA–TAC

-5 + 10 + ? + 10 + ? + 10 = ?

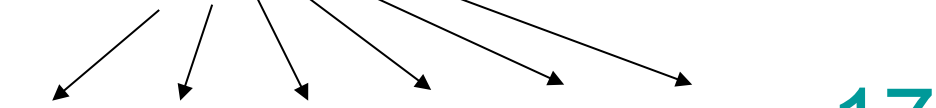
Given a hypothetical substitution matrix:

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Scoring Gaps

- Fixed gap penalty: every gap receives a score of **d**

```
GAAT-C      d=-4
CA-TAC
```



-5 + 10 + -4 + 10 + -4 + 10 = 17

- Linear gap penalty: opening a gap receives a score of **d**; extending a gap receives a score of **e**

```
G--AATC     d=-4
CATA--C     e=-1
```

-5 + -4 + -1 + 10 + -4 + -1 + 10 = 5

You Should Be Able to ...

- Explain why sequence comparison is useful
- Define a *substitution matrix* and how they are generated
- Define the different types of *gap penalties*
- Compute the score of an alignment, given the substitution matrix and gap penalties

How To Create Your Own Block Scoring Matrix

A Simplified Version…

Log-Odds Scores

- Remember, its just a set of Log-odds scores
  - Log(Observation/Expected)
  - The log-odd score of matching amino acid j with amino acid i in an alignment is

$$\log\left(\frac{P_{ij}}{Q_i \cdot Q_j}\right)$$

  - Where:
    - P_{ij} is the frequency of observing amino i aligned with j
    - Q_i, Q_j are the frequencies of amino acids i and j in the data set.
- The log-odd score is (in half bit units)

$$S_{ij} = 2 \cdot \log_2\left(\frac{P_{ij}}{Q_i \cdot Q_j}\right)$$

An Example

N_{AA}	= **14**
N_{AD}	= 5
N_{AV}	= 5
N_{DA}	= 5
N_{DD}	= 8
N_{DV}	= 2
N_{VA}	= 5
N_{VD}	= 2
N_{VV}	= 2

P_{AA}	= 14/**48**
P_{AD}	= 5/48
P_{AV}	= 5/48
P_{DA}	= 5/48
P_{DD}	= 8/48
P_{DV}	= 2/48
P_{VA}	= 5/48
P_{VD}	= 2/48
P_{VV}	= 2/48

```
1: VVAD
2: AAAD
3: DVAD
4: DAAA
```

MSA

Q_A	= **8/16**
Q_D	= 5/16
Q_V	= 3/16

So what does this mean?

So What Does This Mean?

$P_{AA} = 0.29$	$Q_A Q_A = 0.25$
$P_{AD} = 0.10$	$Q_A Q_D = 0.16$
$P_{AV} = 0.10$	$Q_A Q_V = 0.09$
$P_{DA} = 0.10$	$Q_D Q_A = 0.16$
$P_{DD} = 0.17$	$Q_D Q_D = 0.10$
$P_{DV} = 0.04$	$Q_D Q_V = 0.06$
$P_{VA} = 0.10$	$Q_V Q_A = 0.09$
$P_{VD} = 0.04$	$Q_V Q_D = 0.06$
$P_{VV} = 0.04$	$Q_V Q_V = 0.03$

1: VVAD
2: AAAD
3: DVAD
4: DAAA

MSA

$Q_A = 8/16$
$Q_D = 5/16$
$Q_V = 3/16$

$Q_A = 0.50$
$Q_D = 0.31$
$Q_V = 0.19$

So What Does This Mean?

$P_{AA} = 0.29$	$Q_A Q_A = 0.25$	$S_{AA} = 0.43^*$
$P_{AD} = 0.10$	$Q_A Q_D = 0.16$	$S_{AD} = -1.34^*$
$P_{AV} = 0.10$	$Q_A Q_V = 0.09$	$S_{AV} = 0.30^*$
$P_{DA} = 0.10$	$Q_D Q_A = 0.16$	$S_{DA} = -1.34^*$
$P_{DD} = 0.17$	$Q_D Q_D = 0.10$	$S_{DD} = 1.54^*$
$P_{DV} = 0.04$	$Q_D Q_V = 0.06$	$S_{DV} = -1.17^*$
$P_{VA} = 0.10$	$Q_V Q_A = 0.09$	$S_{VA} = 0.30^*$
$P_{VD} = 0.04$	$Q_V Q_D = 0.06$	$S_{VD} = -1.17^*$
$P_{VV} = 0.04$	$Q_V Q_V = 0.03$	$S_{VV} = 0.83^*$

$$S_{ij} = 2 \cdot \log_2\left(\frac{P_{ij}}{Q_i \cdot Q_j}\right)$$

Your New Scoring Matrix

	A	D	V
A	0.43	-1.34	0.30
D	-1.34	1.54	-1.17
V	0.30	-1.17	0.83

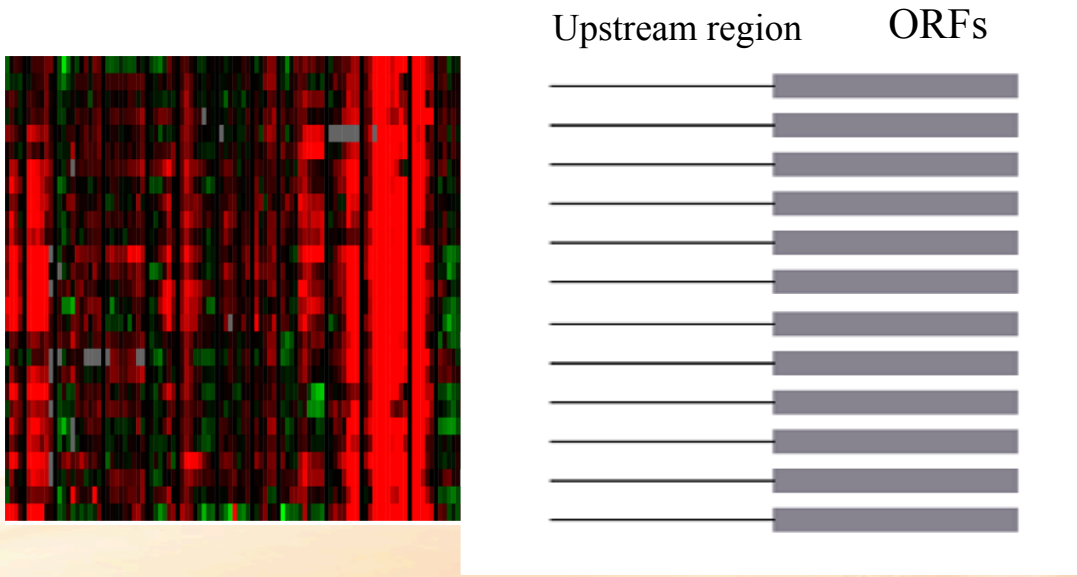
```
1: VVAD
2: AAAD
3: DVAD
4: DAAA
```

MSA

Obviously simplified here..

Now lets look a a powerful method to **identify motifs** in DNA/protein sequences

Finding Sequence Motifs Common to a Group of 'Similar' Sequences



Similar sequence found in most upstream regions

How do you identify motifs in sequence data?
How can you tell if the identified motif is 'significant'?
How do you find genomic examples of the identified motif?

Describing Features Using Frequency Matrices

- Goal:
  - Describe a sequence feature (or motif) more quantitatively than possible using a **consensus sequences** or a **pattern**
- Need:
  - To describe how often particular bases are found in particular positions in a sequence feature

- Definition:
  - For a feature of length **m** using an alphabet of **n** characters
    - a frequency matrix is an **n** by **m** matrix in which each element contains the frequency at which a given member of the alphabet is observed at a given position
    - in an **aligned set of sequences** containing the feature

Position Specific Score Matrix (PSSM)

- PSSMs also called **Position Weight Matrixes** (PWMs) or **Profiles**
- Three common uses of frequency matrices
  - Describe a sequence feature
  - Calculate probability of occurrence of feature in a random sequence
  - Calculate degree of match between a new sequence and a feature

The Basis of a PSSM

- A frequency matrix can be converted to PSSM - by converting frequencies to scores (e.g., by taking logs)
- A PSSM is matrix is:
  - Based on the amino acid frequencies (or nucleic acid frequencies)
  - At every position of a multiple alignment
- From these **frequencies**:
  - The PSSM that will be calculated will result in a matrix
  - That will assign **positive scores** to residues that appear **more often** than by **chance at a certain position**

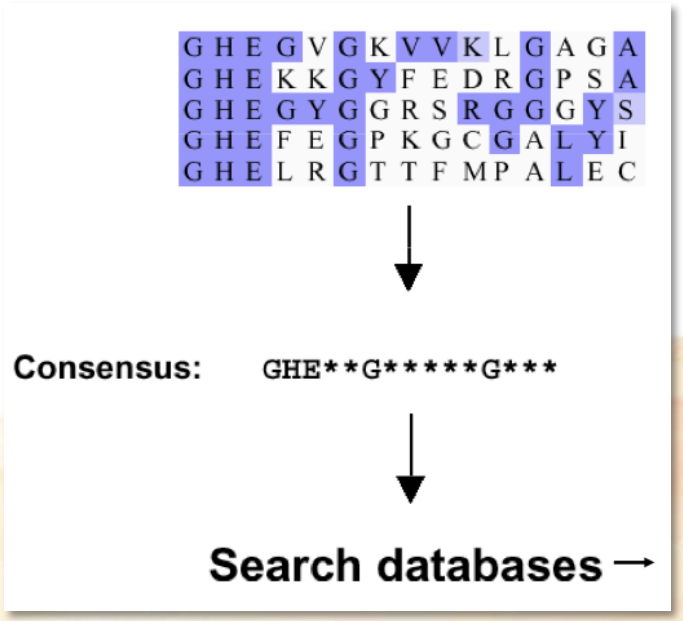
What Does a PSSM Provide

- A method to score how close any sequence is to a **collected set sequences** using a generated matrix
  - Training sequences - ***collected set sequences***
  - Compare against candidate DNA or AA sequences
- Expect that a sequence close to the training sequences tend to have higher scores at each position
  - A good assumption?
  - Yes!
    - Since, product (or Sum) of scores at each position, should be higher than that of most other sequences of similar length

PSSM vs Consensus Sequence

- There are similarities b/t PSSM and **consensus sequence**
  - As with finding occurrences of a **consensus sequence**, we consider all positions in the target sequence as candidate matches
- But with PSSM:
  - For each position
  - Calculate a score by “looking up” the value corresponding to the base at that position
- **So why not "just" use a consensus or pattern?**

Knowledge-Based Motif Identification – Consensus Sequences



- A part of MSA that contains motif positions
- For each column keep the most frequent character (or frequency above a certain threshold)
- MA**GHELEGPVFCGGGYT**AAYF contains the consensus
- Great, it works! But…

Knowledge-Based Motif Identification – Consensus Sequences

- Contains **no information** about **sequence variability** in each **column**
- Results are **very dependent** on the set of sequences used **to derive consensus**
- Provides only **binary answer** (MATCH or NO MATCH)
- Does not provide you information about the quality of the match (no score)
- Useful for finding **very conserved regions** in DNA (restriction sites)
- **DNA consensus sequence** can be constructed using an **extended alphabet** in which each letter corresponds to a combination of nucleotides
  - For instance R=[A or G], Y=[T or C], N=[A or G or T or C], etc

PROSITE

PROSITE

- The first pattern DB
  - Based on the idea that a family can be characterized by a pattern of conserved residues in a single motif
- Sequence information in motifs is reduced to regular expressions & the seed regex used to search for other sequences
- Some families can't be characterized by single motifs
  - Here, additional regexs are created until an optimal set is achieved that captures most or all of the family
  - Results are then manually annotated for inclusion in the db

http://expasy.org/prosite/prosuser.html

PROSITE

- Protein Families and Domains
  - Described as patterns or profiles
  - Release 20.70, of 08-Feb-2011:
    - 1603 documentation entries, 1308 patterns, 916 profiles and 902 ProRules
- Protein Families and Domains
  - Described as patterns or profiles
  - Release 20.96, of 10-Oct-2013:
    - 1671 documentation entries, 1308 patterns, 1054 profiles and 1062 ProRules

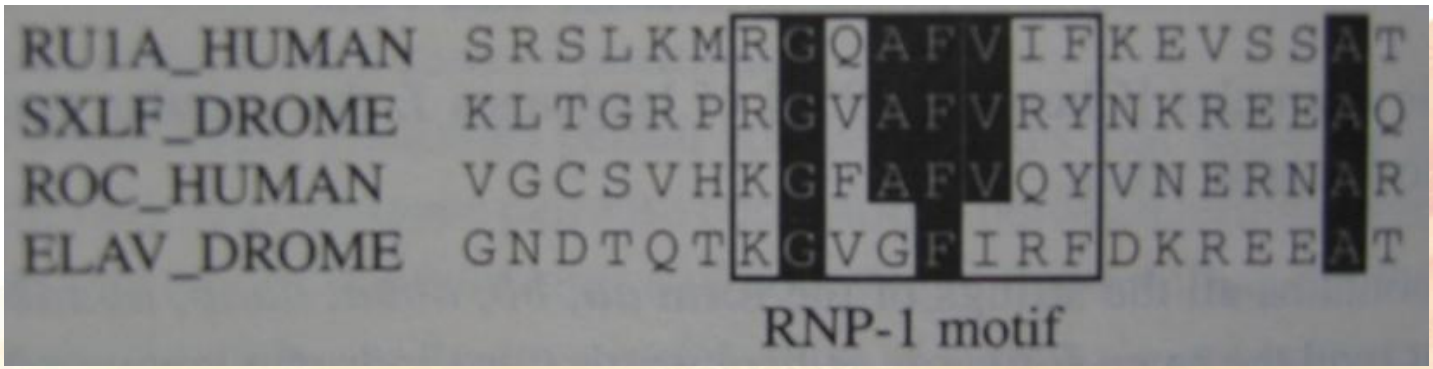
http://expasy.org/prosite/prosuser.html

PROSITE Patterns

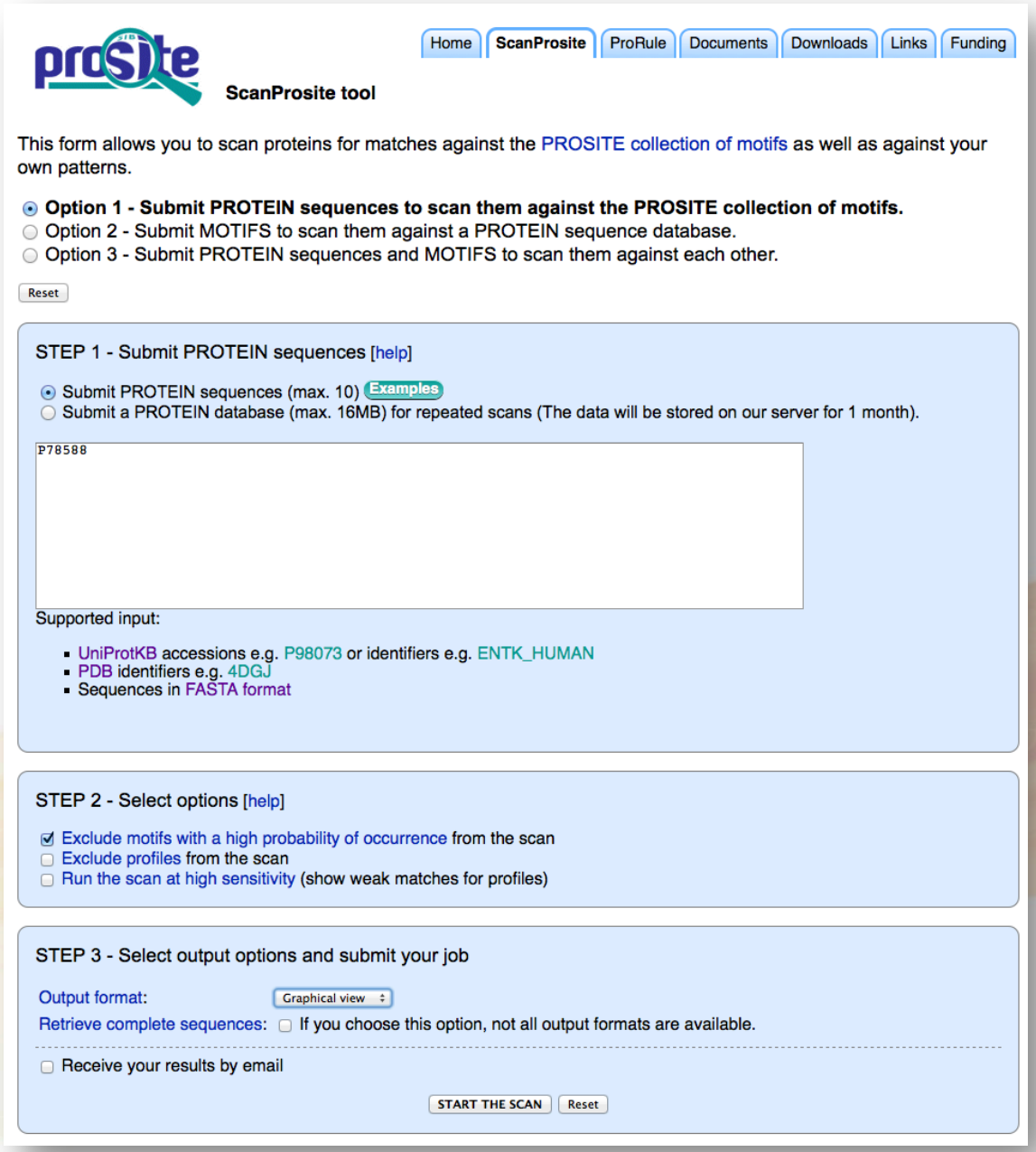
- Unlike methods which assign scores to alignments, PROSITE patterns either match a sequence or do not
- Consists of a string of pattern elements separated by dashes and terminated by a period
  - Pattern Element – single letter
  - [ ] - any one letter
  - { } – anything but enclosed letters
  - X – any residue can occur
  - X(y) – any letter of length y

PROSITE Patterns

RNP-1 Motif



`[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]`



http://prosite.expasy.org/scanprosite/

PROSITE Output

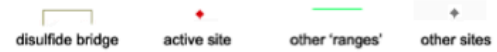
Hits for all PROSITE (release 20.96) motifs on sequence P78588 [UniProtKB/Swiss-Prot (release 2013_10 of 16-Oct-13: 541561 entries)]:

found: 1 hit in 1 sequence

P78588 **FREL_CANAX** (669 aa)
RecName: Full=Probable ferric reductase transmembrane component; EC=1.16.1.7; AltName: Full=Ferric-chelate reductase;. *Candida albicans (Yeast)*

```
MTESKFHAKYDKIQAEFKTNGTEYAKMTTKSSSGSKTSTSASKSSKSTGSSNASKSSTNAHGSNSS
TSSTSSSSSKSGKGNSGTSTTETITTPLLIDYKKFTPYKDAYQMSNNNFNLSINYGSGLLGYWAGI
LAIAIFANMIKKMFPSLTNNLSGSISNLFRKHLFLPATFRKKKAQEFSIGVYGFFDGLIPTRLETI
IVVIFVVLTGLFSALHIHHVKDNPQYATKNAELGHLIADRTGILGTFLIPLLILFGGRNNFLQWLT
GWDFATFIMYHRWISRVDVLLIIVHAITFSVSDKATGKYKNRMKRDFMIWGTVSTICGGFILFQAM
LFFRRKCYEVFFLIHIVLVVFFVVGGYYHLESQGYGDFMWAAIAVWAFDRVVRLGRIFFFGARKAT
VSIKGDDTLKIEVPKPKYWKSVAGGHAFIHFLKPTLFLQSHPFTFTTTESNDKIVLYAKIKNGITS
NIAKYLSPLPGNTATIRVLVEGPYGEPSSAGRNCKNVVFVAGGNGIPGIYSECVDLAKKSKNQSIK
LIWIIRHWKSLSWFTEELEYLKKTNVQSTIYVTQPQDCSGLECFEHDVSFEKKSDEKDSVESSQYS
LISNIKQGLSHVEFIEGRPDISTQVEQEVKQADGAIGFVTCGHPAMVDELRFAVTQNLNVSKHRVE
YHEQLQTWA
```

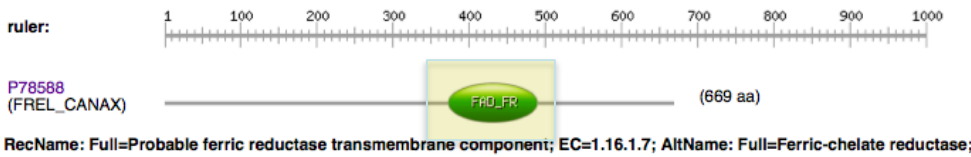
Legend:



Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not intended to indicate homology or shared function.
For more information about how these graphical representations are constructed, go to http://prosite.expasy.org/mydomains/.

hits by profiles: [1 hit (by 1 profile) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.
Hits by PS51384 **FAD_FR** *Ferredoxin reductase-type FAD binding domain profile* :



RecName: Full=Probable ferric reductase transmembrane component; EC=1.16.1.7; AltName: Full=Ferric-chelate reductase;. *Candida albicans (Yeast)*

374 - 492: **score** = 10.203

```
AVWAFDRVVRLGRIfffgarkaTvsikgDDTLKIEVPKP---KyWKSVAGGHAFIHFLKP
TLFLQSHPFTFTTTEsNDK--IVLYAKIKnGITSNIAKYlsPLPGNtatiRVLVEGPYGE
PSSA
```

Predicted feature:

DOMAIN	374	492	FAD-binding FR-type	[condition: none]

Ferredoxin Reductase-Type FAD-Binding Domain Profile

Description

Flavoenzymes have the ability to catalyse a wide range of biochemical reactions. They are involved in the dehydrogenation of a variety of metabolites, in electron transfer from and to redox centres, in light emission, in the activation of oxygen for oxidation and hydroxylation reactions [1,2]. About 1% of all eukaryotic and prokaryotic proteins are predicted to encode a flavin adenine dinucleotide (FAD)-binding domain [3]. According to structural similarities and conserved sequence motifs, FAD-binding domains have been grouped in three main families: (i) the ferredoxin reductase (FR)-type FAD-binding domain, (ii) the FAD-binding domains that adopt a Rossmann fold and (iii) the PCMH-type FAD-binding domain [4].

The FAD cofactor consists of adenosine monophosphate (AMP) linked to flavin mononucleotide (FMN) by a pyrophosphate bond. The AMP moiety is composed of the adenine ring bonded to a ribose that is linked to a phosphate group. The FMN moiety is composed of the isoalloxazine-flavin ring linked to a ribitol, which is connected to a phosphate group. The flavin functions mainly in a redox capacity, being able to take up two electrons from one substrate and release them two at a time to a substrate or coenzyme, or one at a time to an electron acceptor. The catalytic function of the FAD is concentrated in the isoalloxazine ring, whereas the ribityl phosphate and the AMP moiety mainly stabilize cofactor binding to protein residues [1,2].

The structural core of all FR family members is well conserved. The FAD-binding fold characteristic of the FR family is a cylindrical β-domain with a flattened six-stranded antiparallel β-barrel organized into two orthogonal sheets (B1-B2-B5 and B4-B3-B6) separated by one α-helix (see for example <PDB:1A8P>) [5]. The cylinder is open between strands B4 and B5 which makes space for the isoalloxazine and ribityl moieties of the FAD. One end of the cylinder is covered by the only helix of the domain, which is essential for the binding of the pyrophosphate groups of the FAD. The FR family contains two conserved motifs, one (R-x-Y-[ST]) located in B4 where the invariant positively charge Arg residue forms hydrogen bonds to the negative pyrophosphate oxygen atom. The other conserved sequence motif is G-x(2)-[ST]-x(2)-L-x(5)-G-x(7)-P-x-G, which is part of H1-B6 and is known as the phosphate-binding motif [4,5].

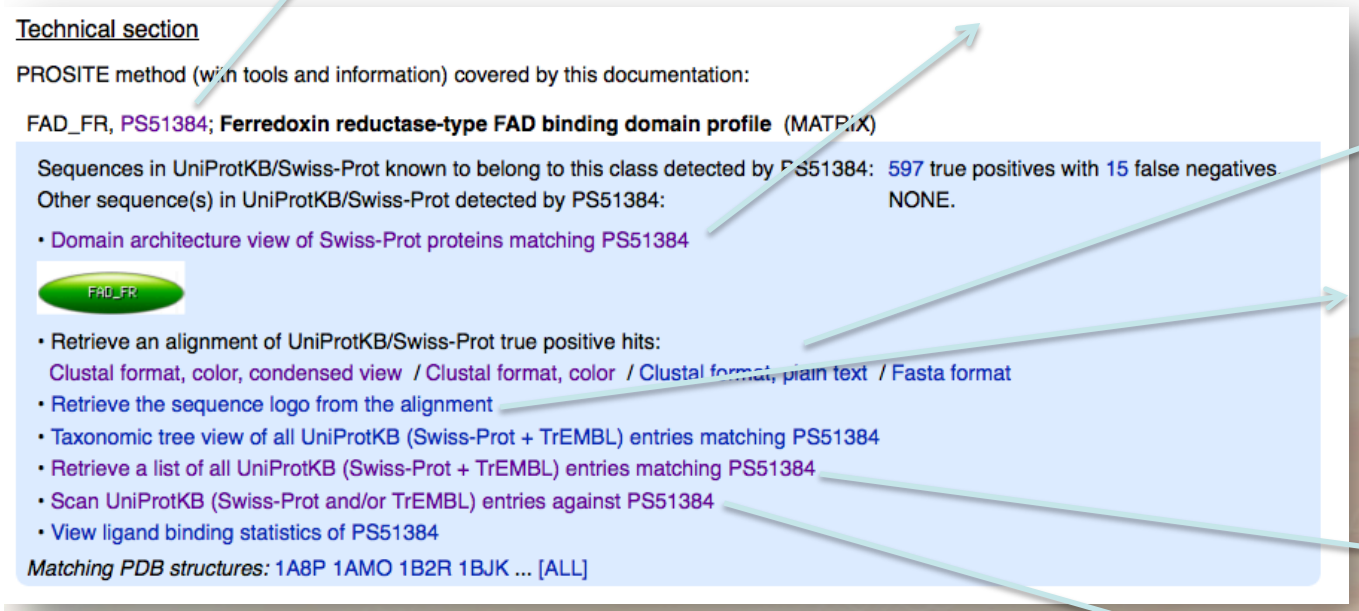
Some proteins known to contain a FR-type FAD-binding domain are listed below:

- Eukaryotic NADH-cytochrome b5 reductase. It is a membrane-bound hemoprotein which acts as an electron carrier for several membrane-bound oxygenases.
- Eukaryotic NADPH-cytochrome P450 reductase. This enzyme is required for electron transfer from NADP to cytochrome P450 in microsomes. It can also provide electron transfer to heme oxygenase and cytochrome b5.
- Nitrate reductase. It is a key enzyme involved in the first step of nitrate assimilation in plants, fungi and bacteria.
- Bacterial ferredoxin reductase. It transports electrons between flavodoxin or ferredoxin and NADPH. It is involved in the reductive activation of cobalamin-independent methionine synthase, pyruvate formate lyase and anaerobic ribonucleotide reductase.
- Bacterial flavohemoprotein. It is involved in nitric oxide detoxification in an aerobic process, termed nitric oxide dioxygenase (NOD) reaction.
- Bacterial Na(+)-translocating NADH-quinone reductase (NQR) subunit F. NQR complex catalyzes the reduction of ubiquinone-1 to ubiquinol by two successive reactions, coupled with the transport of Na(+) ions from the cytoplasm to the periplasm.

Additional Ouputs

PROSITE Entry Details

Other UniProtKB with this domain profile



Alignments can be viewed

Sequence Logo (later in lecture we'll discuss this)

Download UniProtKB sequences

Scan UniProtKB against this PROSITE domain profile (next Slide)

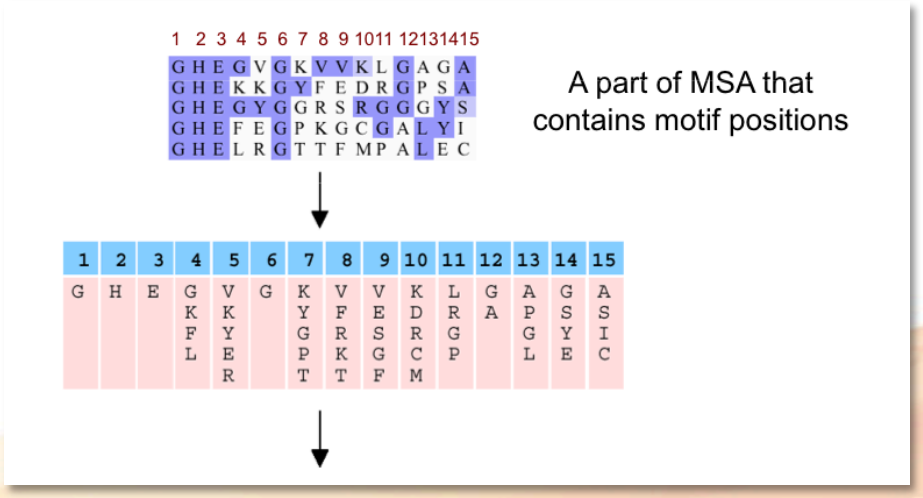
Can also search your sequence against a MOTIFS

Knowledge-Based Motif Identification - Patterns (Regular expressions)

- A pattern describes a set of alternative sequences using a single expression
- **PROSITE syntax for amino acid patterns:**
  - Each element is separated by '-'.
  - X - any amino acid
  - [ ] - OR (for instance [PG] means P or G)
  - { } - NOT (for instance {PG} means any amino acid except G or P).
  - ( ) - REPEAT (for instance, P(2,3) means PP or PPP)
  - < - pattern starts at N-terminus (left end)
  - > - pattern starts at C-terminus (right end)
- Example: **<A-x-[ST](2)-x(0,1)-{V}** means N-terminal sequence:
  - **Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-any except Val**
  - An example matching this pattern: **A-Y-S-T-P**

http://prosite.expasy.org/scanprosite/scanprosite-doc.html#pattern_syntax

Derivation of Patterns

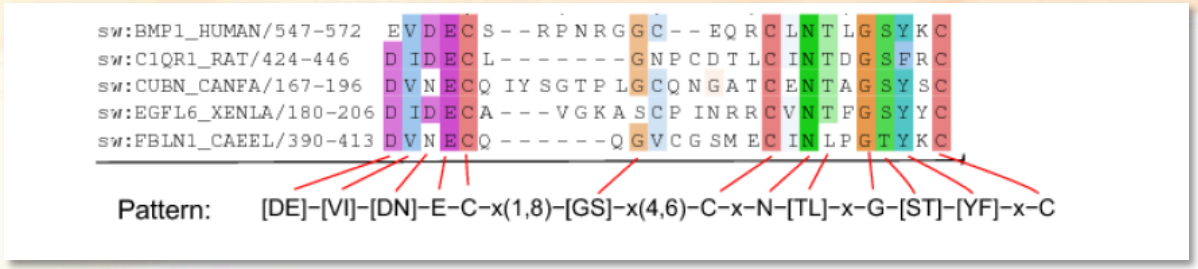


- G-H-E-[GKFL]-[VKYER]-G-[KYGPT]-[VFRKT]-[VESGF]-[KDRCM]-[LRGP]-[GA]-[APGL]-[GSYE]-[ASIC]

Search databases

Use of a Pattern

- Patterns are used to describe small functional regions:
  - Enzyme catalytic sites
  - Prosthetic group attachment sites (heme, PLP, biotin, etc.)
  - Amino acids involved in binding a metal ion
  - Cysteines involved in disulfide bonds
  - Regions involved in binding a molecule (ATP, calcium, DNA etc.) or a protein
  - N-glycosylation sites



Knowledge-Based Motif Identification - Patterns (Regular expressions)

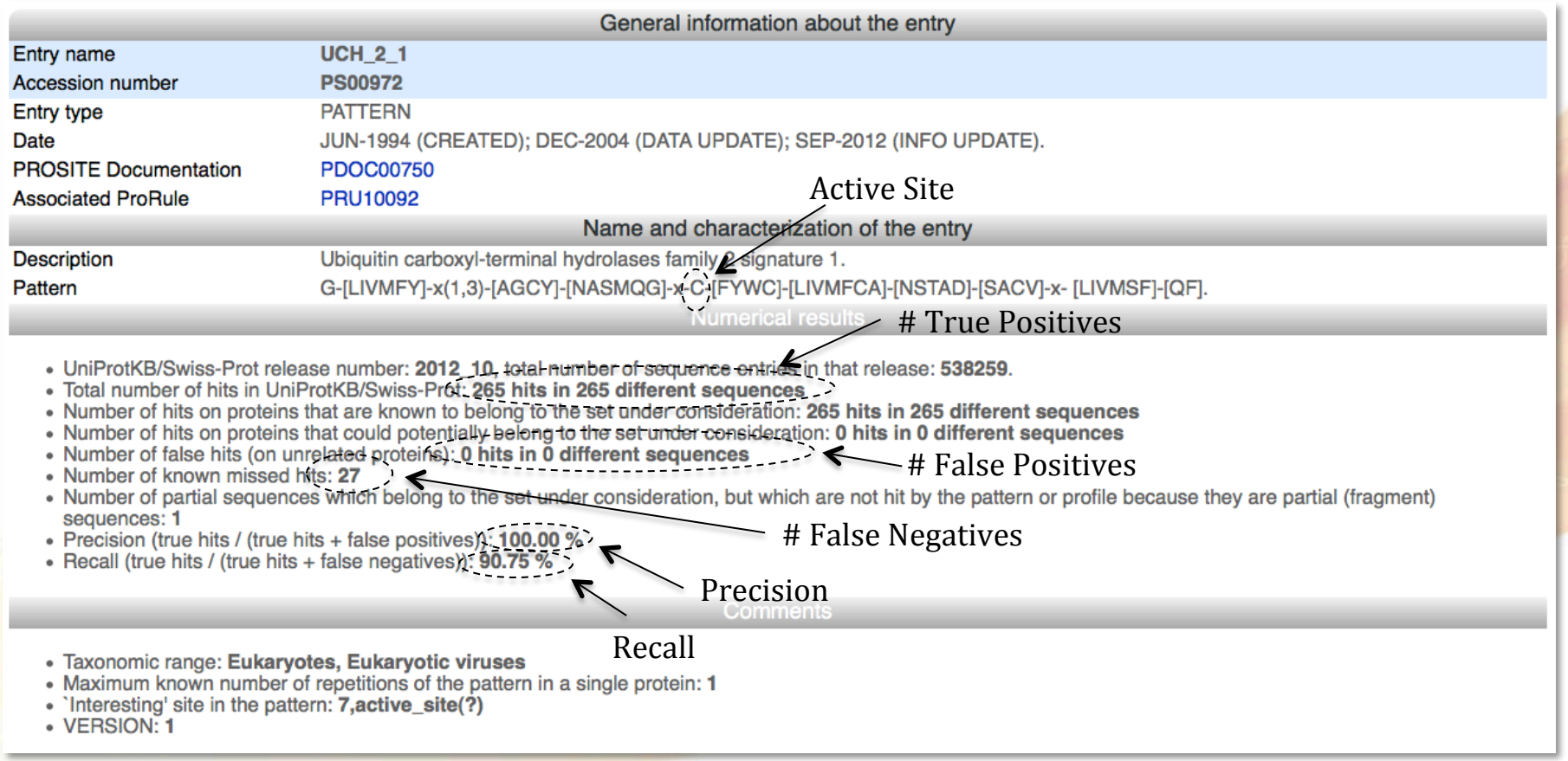
- The Good:
  - Patterns can be used to search sequence databases quickly
  - Results are very dependent on the set of sequences used to derive the pattern
- The Bad:
  - Only provide only binary answer (MATCH or NO MATCH)
  - Do not tell you anything about the quality of the match (no score)
  - Short patterns match many unrelated sequences
  - Patterns poorly model gaps

- PROSITE – a comprehensive collection of protein patterns
  - http://us.expasy.org/prosite/

Discovering Patterns

- You can also automatically build a pattern (from MSA) by using:
  - PRATT
    - Tool to discover patterns that are conserved in a set of protein sequences
    - Patterns are reported using the PROSITE format
    - http://www.expasy.org/tools/pratt/
- Word of Caution:
  - Automatic discovered patterns are usually different from those designed by a human expert with knowledge of the biochemical literature

PROSITE Entry Details - Example



General information about the entry

Entry name	**UCH_2_1**
Accession number	**PS00972**
Entry type	PATTERN
Date	JUN-1994 (CREATED); DEC-2004 (DATA UPDATE); SEP-2012 (INFO UPDATE).
PROSITE Documentation	PDOC00750
Associated ProRule	PRU10092

Name and characterization of the entry

Description	Ubiquitin carboxyl-terminal hydrolases family 2 signature 1.
Pattern	G-[LIVMFY]-x(1,3)-[AGCY]-[NASMQG]-x-C-[FYWC]-[LIVMFCA]-[NSTAD]-[SACV]-x- [LIVMSF]-[QF].

Numerical results

- UniProtKB/Swiss-Prot release number: **2012_10**, total number of sequence entries in that release: **538259**.
- Total number of hits in UniProtKB/Swiss-Prot: **265 hits in 265 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **265 hits in 265 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- Number of known missed hits: **27**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **1**
- Precision (true hits / (true hits + false positives)): **100.00 %**
- Recall (true hits / (true hits + false negatives)): **90.75 %**

Comments

- Taxonomic range: **Eukaryotes, Eukaryotic viruses**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- `Interesting' site in the pattern: **7,active_site(?)**
- VERSION: **1**

http://prosite.expasy.org/PS00972

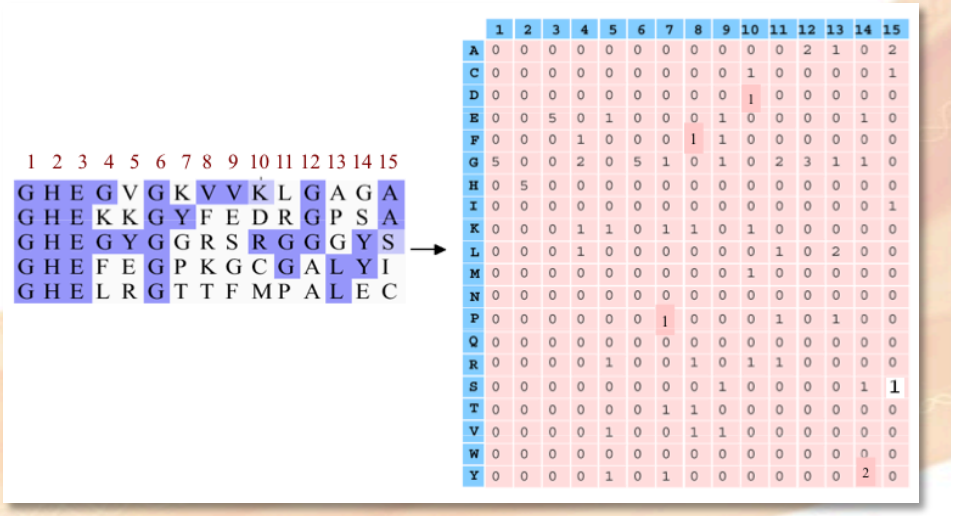
Advantage and Limitation of PROSITE Patterns

- Advantages:
  - Efficient for the identification of sites or short motifs
  - Understandable by user, you don’t need to be an expert in bioinformatic to read or build a consensus sequence
- Limitation:
  - The regular expression syntax is too rigid to represent highly divergent domains
  - One mismatch is enough to eliminate a match

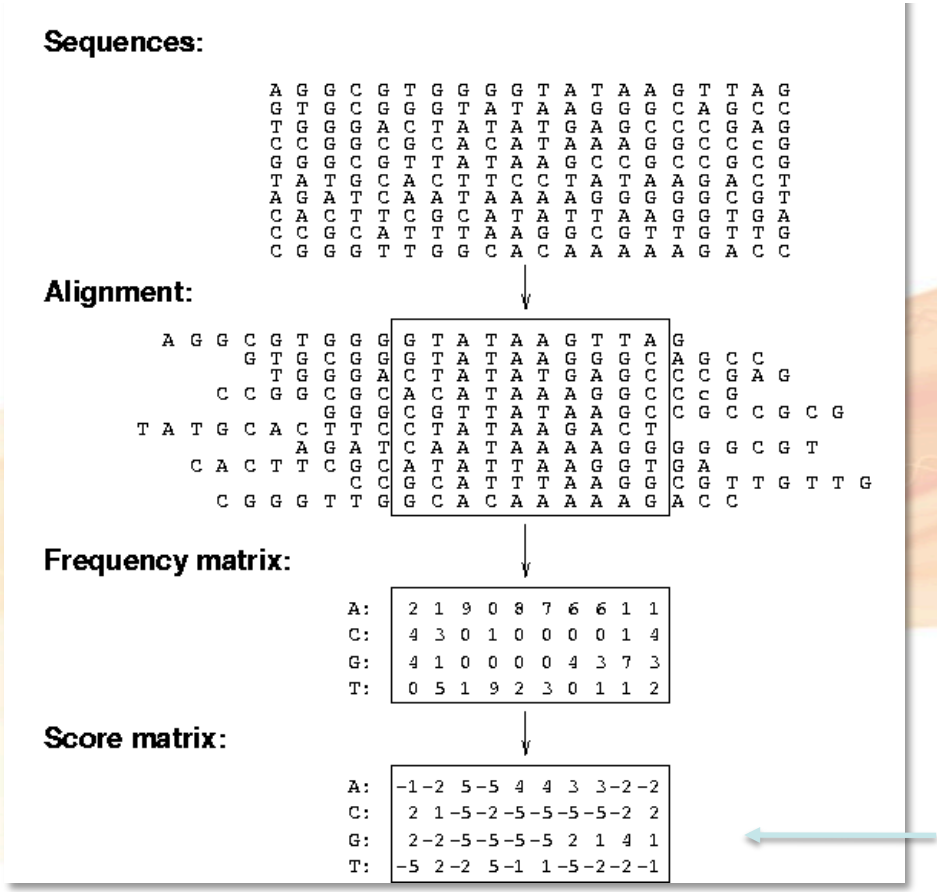
- Let's learn how a better model is generated

Knowledge-Based Motif Identification - PSSMs

- Convert MSA of character strings into a mathematical model
- Model must be suitable for alignment algorithms and sequence database searches
- PSSM gives the frequency of occurrence of each residue in each position of MSA
- Conserved positions tend to contain few different residues
- PSSM is a 20 by ***L matrix (proteins), where L is the number of columns in MSA***



Generation of a Position Specific Score Matrix from Unaligned Sequences



What's this look like?

Consensus Representation of Motifs:

Site 1	A	G	A	T	G	G	A	T	G	G
Site 2	T	G	A	T	T	G	A	T	G	T
Site 3	T	G	A	T	G	G	A	T	G	G
Site 4	A	G	A	T	T	G	A	T	C	G
Site 5	T	G	A	T	G	G	A	T	T	G
Site 6	T	G	A	T	G	G	A	T	T	G
Site 7	A	G	A	T	G	G	A	T	T	G

IUPAC consensus: **W G A T K G A T B K**
(where W = A or T)

We know how to build something like this in Perl, but a PSSM is a model

1st Step - Build the Frequency Matrix

Site 1	A	G	A	T	G	G	A	T	G	G
Site 2	T	G	A	T	T	G	A	T	G	T
Site 3	T	G	A	T	G	G	A	T	G	G
Site 4	A	G	A	T	T	G	A	T	C	G
Site 5	T	G	A	T	G	G	A	T	T	G
Site 6	T	G	A	T	G	G	A	T	T	G
Site 7	A	G	A	T	G	G	A	T	T	G

PSSM represents frequencies of each base at each position in the motif *

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

*PWM/PSSM can correspond to the frequency matrix or a likelihood matrix, in this example we'll use the frequency matrix

2nd Step - Finding Matches to a PSSM

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

- Is the sequence **A G A T T G A T C T** a match to this matrix?
- Joint probability: Assuming each position is independent,

$$P(\text{motif}) = \prod_{i} P_{b=\text{G,A,T,C}}(i)$$

$P_b(i)$ = Probability finding b at position i

Background model:
P(G,A,T,C) = 0.25

P(sequence | matrix model) = (0.4)(1.0)(1.0)(1.0)(0.3)(1.0)(1.0)(1.0)(0.2)(0.2) = **0.0048**

P(sequence | background model) = (0.25)(0.25)(0.25)(0.25)(0.25)(0.25)(0.25)(0.25)(0.25)(0.25) = **$9.54e^{-7}$**

3rd Step - Log-likelihood Ratio LLR

= log (P(sequence | matrix model) / P(sequence | background model))

- A measure of how different the likelihood of the sequence is, given the motif model vs. the background model
- In our example:
  - Score = LLR = log (0.0048 / 9.54e-7) = 3.4
- The larger the LLR, the more likely the motif fits the model
- To select motifs in real life, can define a LLR cutoff by using the score of an actual motif (lowest scoring)

$$\text{Score} = \log_k\left(\prod_i \frac{P_b(i)}{q_b} \right)_{b = G,A,T,C} = \sum_i \log_k\left(\frac{P_b(i)}{q_b} \right)_{b = G,A,T,C}$$

Simple, but.....

Finding Matches to (Instances of) a PSSM

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

- Is the sequence **A A A T T G A T C T** a match to this matrix?
- Joint Probability: Assuming each position is independent

$$P(\text{motif}) = \prod_{i} \; P_{b=G,A,T,C}(i)$$

- What's the problem?
- P(sequence | matrix model) = (0.4)(0)(1.0)(1.0)(0.3)(1.0)(1.0)(1.0)(0.2)(0.2) = 0

- **If your PSSM was trained on a small sample set, you might have missed some examples = *overfitting* of the matrix (i.e. *too* specific)**

Pseudo-counts: Protecting Against Overfitting Due to Small Sample Sizes

Add 1 count to each base at each position, then divide by $n + 4$

Site 1	A	G	A	T	G	G	A	T	G	G
Site 2	T	G	A	T	T	G	A	T	G	T
Site 3	T	G	A	T	G	G	A	T	G	G
Site 4	A	G	A	T	T	G	A	T	C	G
Site 5	T	G	A	T	G	G	A	T	T	G
Site 6	T	G	A	T	G	G	A	T	T	G
Site 7	A	G	A	T	G	G	A	T	T	G

Without pseudo-counts:

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

Pseudo-counts: Protecting Against Overfitting Due to Small Sample Sizes

Add 1 count to each base at each position, then divide by $n + 4$

Site 1	A	G	A	T	G	G	A	T	G	G
Site 2	T	G	A	T	T	G	A	T	G	T
Site 3	T	G	A	T	G	G	A	T	G	G
Site 4	A	G	A	T	T	G	A	T	C	G
Site 5	T	G	A	T	G	G	A	T	T	G
Site 6	T	G	A	T	G	G	A	T	T	G
Site 7	A	G	A	T	G	G	A	T	T	G

With pseudo-counts (rounded values):

G	0.1	0.7	0.1	0.1	0.4	0.7	0.1	0.1	0.3	0.7
A	0.3	0.1	0.7	0.1	0.1	0.1	0.7	0.1	0.1	0.1
T	0.4	0.1	0.1	0.7	0.25	0.1	0.1	0.7	0.3	0.2
C	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1

Rounded values

Same can be done with amino acids…

Creating a PSSM for Proteins/Motifs: Example

NTEGEWI

NITRGEW

NIAGECC

Amino acid frequencies at every position of the alignment:

Amino Acid	1	2	3	4	5	6	7
N	1	0	0	0	0	0	0
T	0	0.33	0.33	0	0	0	0
E	0	0	0.33	0	0.66	0.33	0
G	0	0	0	0.66	0.33	0	0
W	0	0	0	0	0	0.33	0.33
I	0	0.66	0	0	0	0	0.33
H	0	0	0	0	0	0	0
R	0	0	0	0.33	0	0	0
A	0	0	0.33	0	0	0	0
C	0	0	0	0	0	0.33	0.33
...	...	...	...	...	...	...	...

Creating a PSSM with AA: Example

- a.a. do not appear at a specific position of a multiple alignment must also be considered
  - In order to model every possible sequence and have calculable log-odds scores
  - We'll use pseudocounts again!
    - Assigns minimal scores to residues that do not appear at a certain position of the alignment according to the following equation:

$$Score(X) = \frac{Frequency + pseudocount}{N + 20(pseudocount)}$$

- Where
  - Frequency is the frequency of residue i in column j (the count of occurrences)
  - pseudocount is a number higher or equal to 1
  - N is the number of sequences in the multiple alignment

Creating a PSSM with AA :Example

In this example, N = 3 and let's use pseudocount = 1:

NTEGEWI

NITRGEW

NIAGECC

Score(N) at position 1 = 3/3 = **1**

Score(I) at position 1 = 0/3 = **0**

Becomes:

Score(N) at position 1 -> (**3**+1) / (3+20) = 4/23 = **0.174**

Score(I) at position 1 -> (**0**+1) / (3+20) = 1/23 = **0.044**

To simplify for this example we'll assume that every amino acid appears equally in protein sequences, i.e. $f_i = 0.05$ for every i:

PSSM Score(N) at position 1 = log(**0.174** / 0.05) = **0.541**

PSSM Score(I) at position 1 = log(**0.044** / 0.05) = **-0.061**

* PWM/PSSM can correspond to the frequency matrix or a likelihood matrix, in this example we'll use the log likelihood matrix

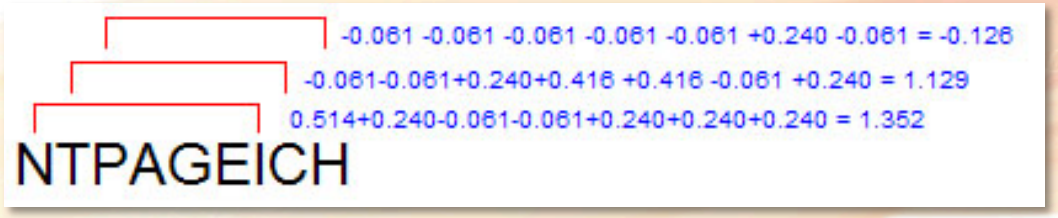
Creating a PSSM with AA :Example

- The matrix assigns positive scores to residues that appear more often than expected by chance and negative scores to residues that appear less often than expected by chance.

Amino acid	1	2	3	4	5	6	7
N	0.541	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061
T	-0.061	0.240	0.240	-0.061	-0.061	-0.061	-0.061
E	-0.061	-0.061	0.240	-0.061	0.416	0.240	-0.061
G	-0.061	-0.061	-0.061	0.416	0.240	-0.061	-0.061
W	-0.061	-0.061	-0.061	-0.061	-0.061	0.240	0.240
I	-0.061	0.416	-0.061	-0.061	-0.061	-0.061	0.240
H	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061
R	-0.061	-0.061	-0.061	0.240	-0.061	-0.061	-0.061
A	-0.061	-0.061	0.240	-0.061	-0.061	-0.061	-0.061
C	-0.061	-0.061	-0.061	-0.061	-0.061	0.240	0.240
...	...	...	...	...	...	...	...

Using a PSSM

- To search for matches to a PSSM
  - Scan along a the sequence using a sliding window technique
  - The length (**L**) of the PSSM
- The matrix is slid on a sequence one residue at a time and the scores of the residues of every region of length **L** are added
- Scores that are higher than an **empirically** predetermined threshold are reported



What's empirically determined?

DNA foot printing
CHiP-assay
CHiP-chip

Advantages of PSSM

- Weights sequence according to observed diversity specific to the motif of interest
- Minimal assumptions
- Easy to compute, for DNA
  - For proteins, as we'll see there's more to take into account
  - Can be used in comprehensive evaluations
- The score of a substring aligned with the PSSM can be interpreted as the log-likelihood
  - Which is significance, right?

Significance

- So you have a hit to your PSSM
- Is that enough?
- How could we calculate a Significance of Scores from a PSSM?
- Example:
  - A PSSM created from *E. coli* promoter sequences

A	-3.8	1.9	0.1	1.2
C	-1.5	-3.8	-0.8	-1
G	-1.3	-4.8	-0.6	-0.7
T	1.7	-3.2	0.8	-0.9

- What score does this PSSM assign the sequence **TAGA** ?

* PWM/PSSM can correspond to the frequency matrix or a likelihood matrix, in this example we'll use the log likelihood matrix

Position-Specific Scoring Matrix

- A PSSM created from *E. coli* promoter sequences

A	-3.8	1.9	0.1	1.2
C	-1.5	-3.8	-0.8	-1
G	-1.3	-4.8	-0.6	-0.7
T	1.7	-3.2	0.8	-0.9

- What score does this PSSM assign the sequence **TAGA** ?
- Score: 1.7 + 1.9 + -0.6 + 1.2 = **4.2**
- **I have my likelihood score! So we're done?**
- **Did we answer our question about significance?**
- No, we still don't know how significant 4.2 is…

So Now You Have a Score

- While probabilities are used in the construction of the PSSMs
  - Scores themselves **cannot be interpreted statistically**
  - This has led to difficulty with choosing the score cutoff values for each matrix
    - Contributes to the large numbers of false positive predictions seen in practice
- What we need is a significance value
- Any ideas what we could use?

P-values

- Determine if the motif matched your model by something **more than random**
  - **Random models** are import to determine in bioinformatics
  - So how can you do this?
    - Calculate a **p-value!**
    - What do we need
      - In order to obtain a valid p-value
      - One needs to model the background sequence properly, which serves as the "**null model**"
- P-value can be thought of as
  - Probability of obtaining a result at least as extreme as the one observed
  - The lower this probability the higher is the importance of the observed match

The Null Hypothesis

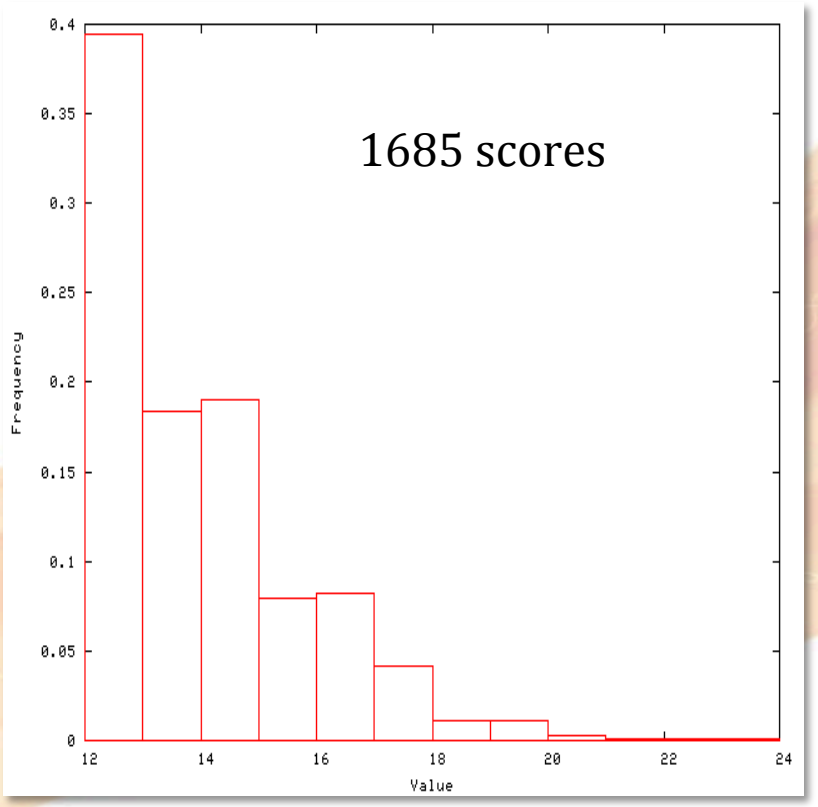
- We are interested in characterizing the distribution of scores from sequence comparison algorithms, or here a [PSSM]
- We would like to measure how surprising a given score is*, assuming that the a new sequence does not contain our motif*
- The assumption is called the **null hypothesis**
- The purpose of **most statistical tests** is to determine whether the **observed results** provide a reason to reject the hypothesis that they are merely a product of chance factors

Sequence Similarity Score Distribution

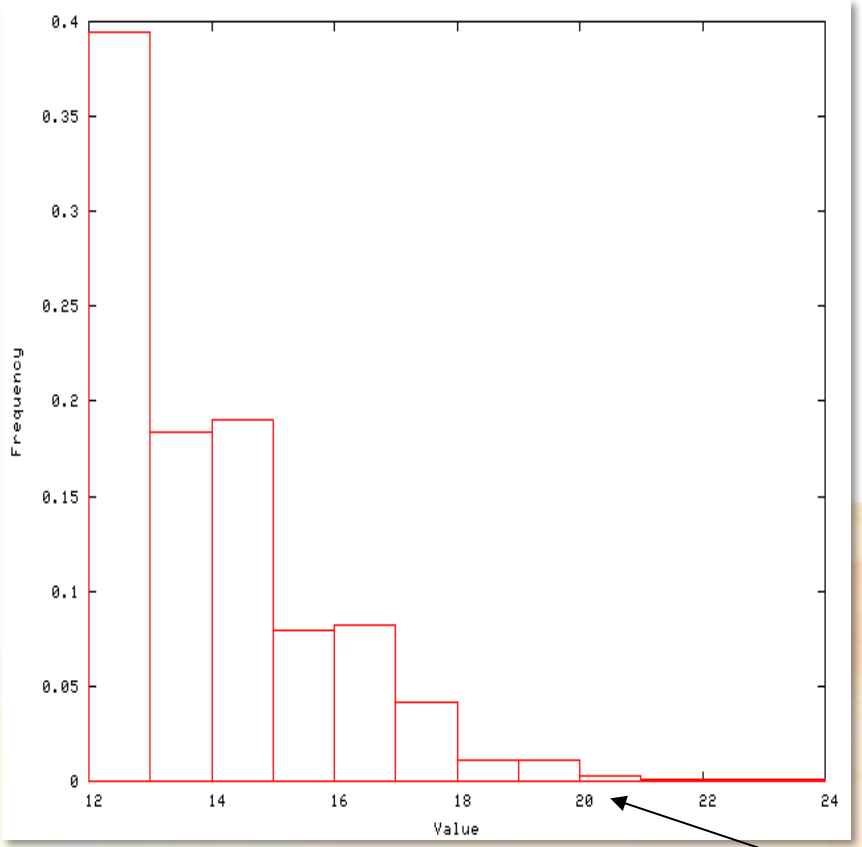
- Search a randomly generated database of DNA sequences against a PSSM
- What will be the form of the resulting distribution of PSSM comparison scores?

Empirical Null Score Distribution

- Distribution generated using a randomized sequence database
- Now we've modeled the background sequences, or the "**null score distribution**"
- Each PSSM score is evaluated statistically by computing its p-value
  - What is that?
  - Probability that the background model can achieve a score at least as high as that observed



Recall: p-value = Pr(data | null)



- The probability of observing a score >X is the area under the curve to the right of X
- This probability is called a p-value
- p-value = Pr(data|null)

Out of 1685 scores, 28 receive a score of 20 or better. Thus, the p-value associated with a score of 20 is approximately 28/1685 = 0.0166. So now lets model this……

P-values For PSSM Scores

- Staden gave an efficient method for calculating p-values from PSSMs
- This method involves **approximating** the distribution of scores
- The values in the PSSM must be rescaled so that all entries are between 0 and some N
  - How do we rescale?
  - Its important to understand how to rescale data in bioinformatics
    - It will allow you to set data to values in a range you would like to have them in
    - Without changing the model behind the values

P-values For PSSM Scores

A	-3.8	1.9	0.1	1.2
C	-1.5	-3.8	-0.8	-1
G	-1.3	-4.8	-0.6	-0.7
T	1.7	-3.2	0.8	-0.9

- Rescaling the matrix
  - Let's chose N = 10
  - Rescale the PSSM scores so that all values are between 0 and 10
  - Any ideas?

P-values For PSSM Scores

A	-3.8	1.9	0.1	1.2		A	1.00	6.70	4.90	6.00
C	-1.5	-3.8	-0.8	-1	→	C	3.30	1.00	4.00	3.80
G	-1.3	-4.8	-0.6	-0.7		G	3.50	0.00	4.20	4.10
T	1.7	-3.2	0.8	-0.9		T	6.50	1.60	5.60	3.90

-0.9 - -4.8 = 3.9

- Find smallest value
  - Subtract the value from every entry in the matrix
- Now we have our minimum value (0.00)
- Now what?

P-values For PSSM Scores

A	-3.8	1.9	0.1	1.2
C	-1.5	-3.8	-0.8	-1
G	-1.3	-4.8	-0.6	-0.7
T	1.7	-3.2	0.8	-0.9

→

A	1.00	6.70	4.90	6.00
C	3.30	1.00	4.00	3.80
G	3.50	0.00	4.20	4.10
T	6.50	1.60	5.60	3.90

6.50*1.4925=9.70

- Find the largest value
- Divide 10 (N) by the value
  - 10/6.70 = 1.4925
- Multiply all entries by the result

A	1.49	10.00	7.31	8.96
C	4.93	1.49	5.97	5.67
G	5.22	0.00	6.27	6.12
T	9.70	2.39	8.36	5.82

P-values For PSSM Scores

A	-3.8	1.9	0.1	1.2
C	-1.5	-3.8	-0.8	-1
G	-1.3	-4.8	-0.6	-0.7
T	1.7	-3.2	0.8	-0.9

→

A	1.00	6.70	4.90	6.00
C	3.30	1.00	4.00	3.80
G	3.50	0.00	4.20	4.10
T	6.50	1.60	5.60	3.90

↓

Round to the nearest integer

A	1.49	10.00	7.31	8.96
C	4.93	1.49	5.97	5.67
G	5.22	0.00	6.27	6.12
T	9.70	2.39	8.36	5.82

←

A	1	10	7	9
C	5	1	6	6
G	5	0	6	6
T	10	2	8	6

How could I increase the precision of the scaling?

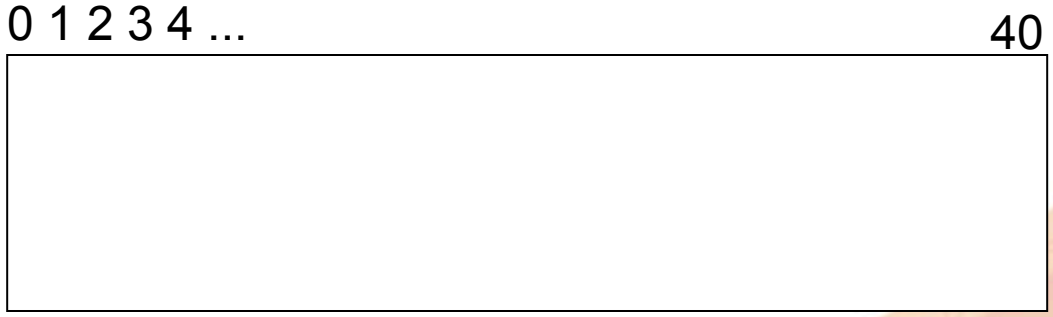
The larger N the more precise your rescaling will be -> the more accurate distribution will be -> more precise p-value estimates will be!

Now we can **approximate** the distribution of scores

Why would we want to do this?

P-values For PSSM Scores

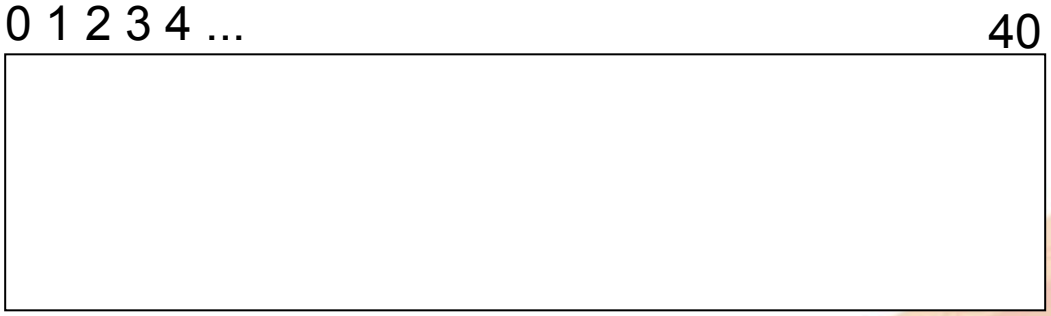
A	1	10	7	9
C	5	1	6	6
G	5	0	6	6
T	10	2	8	6



- Say that your motif has **M** columns
  - Create a matrix that has **M** rows and **N*M+1** columns (**N** here is from the previous slide = 10)
  - For this example, this is **4** rows, and **41** columns
- With this, you can determine distribution of possible scores by tabulating the score of every possible sequence?

P-values For PSSM Scores

A	1	10	7	9
C	5	1	6	6
G	5	0	6	6
T	10	2	8	6



- There are only four possible sequences of length 1
  - For each value in the first column of your motif, put a 1 in the corresponding entry in the first row of the matrix

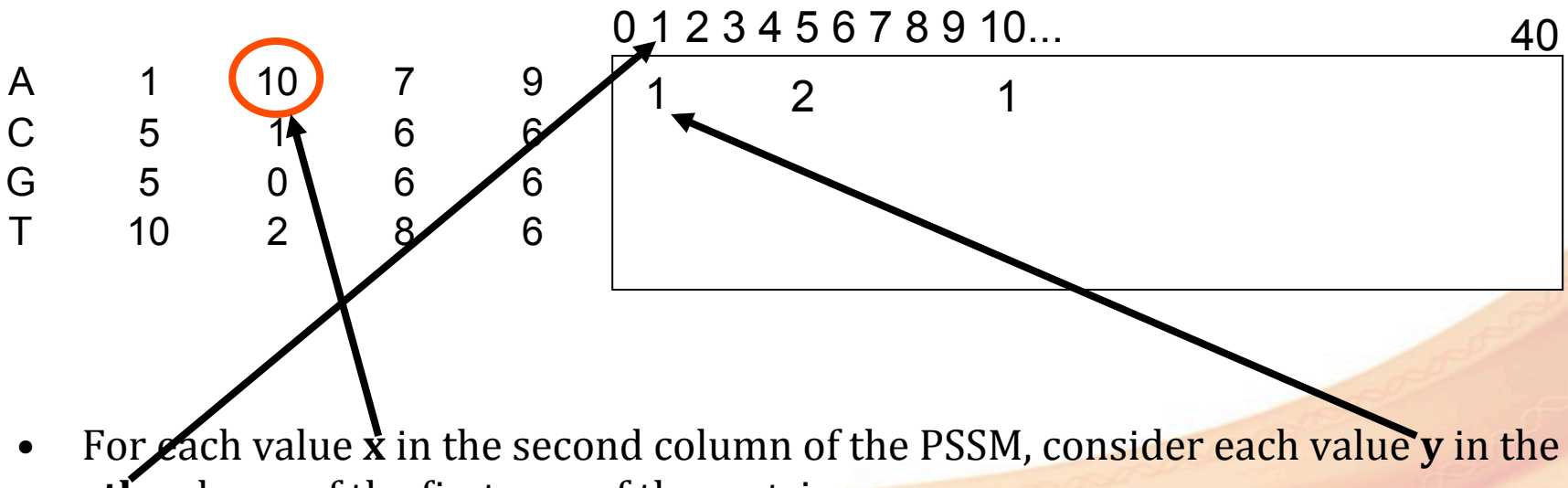
P-values For PSSM Scores

A	1	10	7	9
C	5	1	6	6
G	5	0	6	6
T	10	2	8	6

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	1				**2**					**1**										

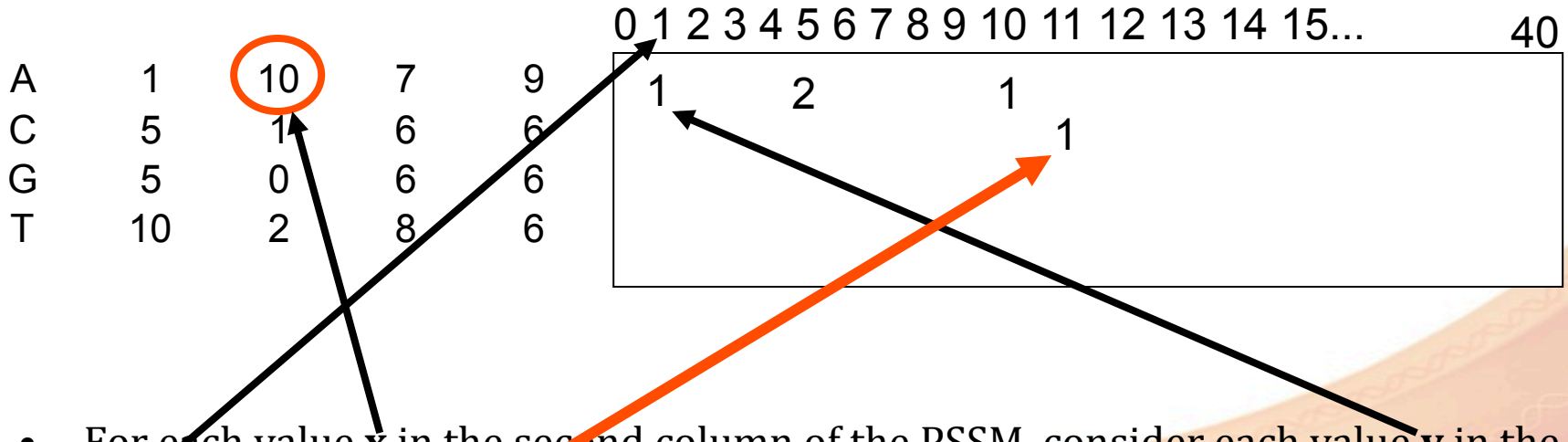
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

P-values For PSSM Scores



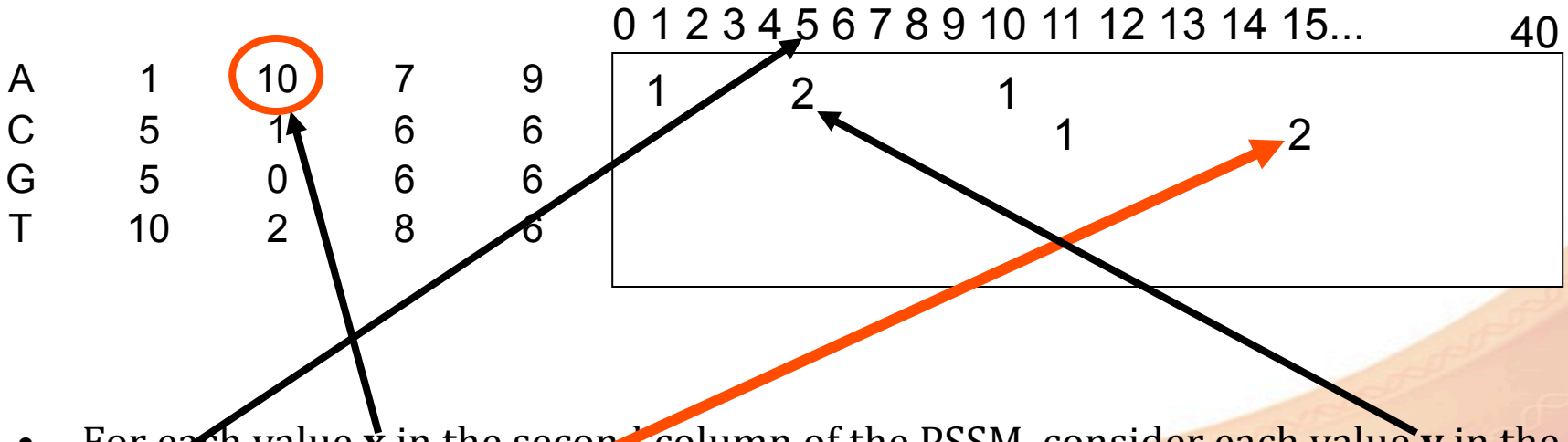
- For each value **x** in the second column of the PSSM, consider each value **y** in the **zth** column of the first row of the matrix
- Add **y** to the **x+zth column** of the matrix

P-values For PSSM Scores



- For each value **x** in the second column of the PSSM, consider each value **y** in the **zth** column of the first row of the matrix
- Add **y** to the **x+zth column** of the matrix

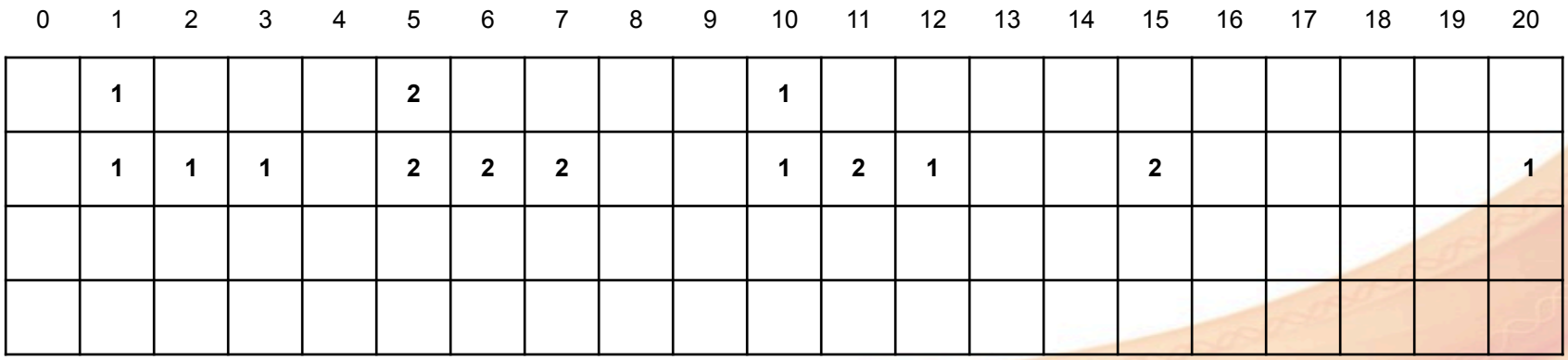
P-values For PSSM Scores

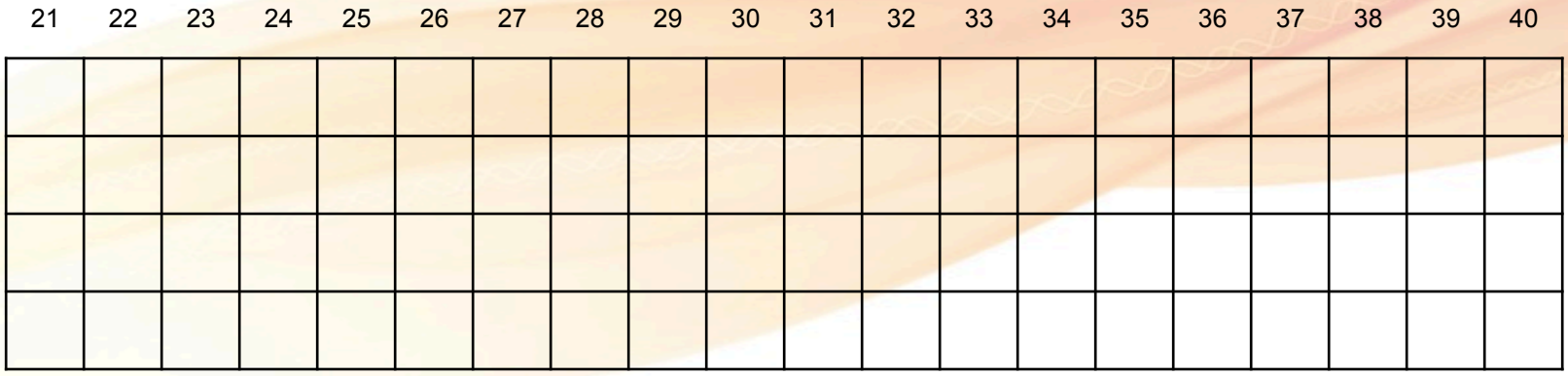


- For each value **x** in the second column of the PSSM, consider each value **y** in the **zth** column of the first row of the matrix
- Add **y** to the **x+zth column** of the matrix

A	1	10	7	9
C	5	1	6	6
G	5	0	6	6
T	10	2	8	6

The counts in row two give the score distribution for all 16 possible sequences of length 2





A	1	10	7	9
C	5	1	6	6
G	5	0	6	6
T	10	2	8	6

The counts in row three give the score distribution for all 64 possible sequences of length 3

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	1				**2**					**1**										
	1	**1**	**1**		**2**	**2**	**2**			**1**	**2**	**1**			**2**					**1**
							2	**3**	**4**	**2**	**5**	**6**	**8**	**4**	**2**	**2**	**5**	**5**	**3**	**1**

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
4	**2**	**2**				**2**	**1**	**1**											

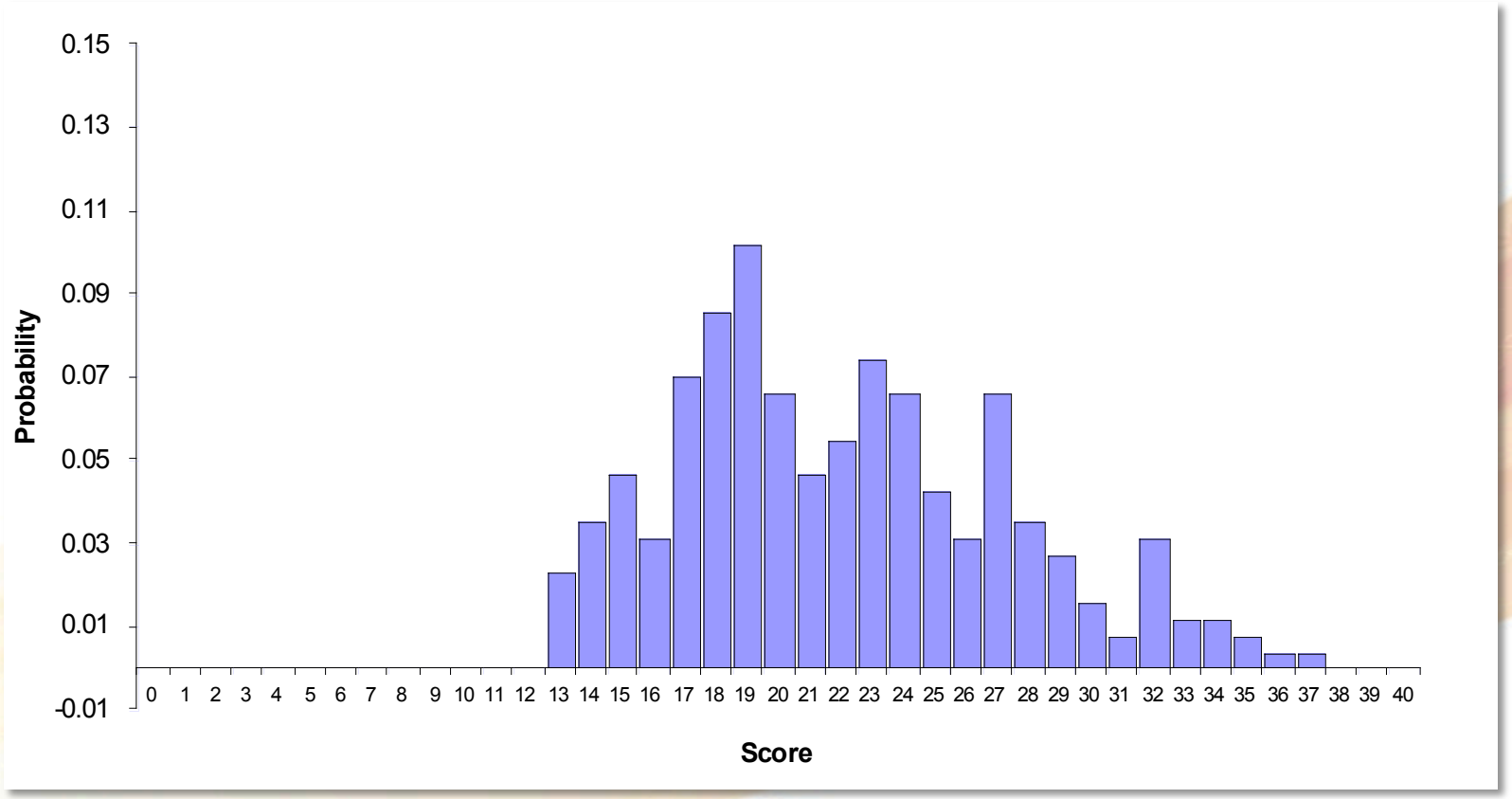
A	1	10	7	9
C	5	1	6	6
G	5	0	6	6
T	10	2	8	6

The counts in row four give the score distribution for all 256 possible sequences of length 4

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	1				2					1										
	1	1	1		2	2	2			1	2	1			2					1
							2	3	4	2	5	6	8	4	2	2	5	5	3	1
													6	9	12	8	18	22	26	17

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
4	2	2			2	1	1												
12	14	19	17	11	8	17	9	7	4	2	8	3	3	2	1	1			

Score Distribution



Why would we want to generate a distribution?

Why would we want to generate the scores like this?

Score Distribution

33	34	35	36	37	38	39	40
3	3	2	1	1			

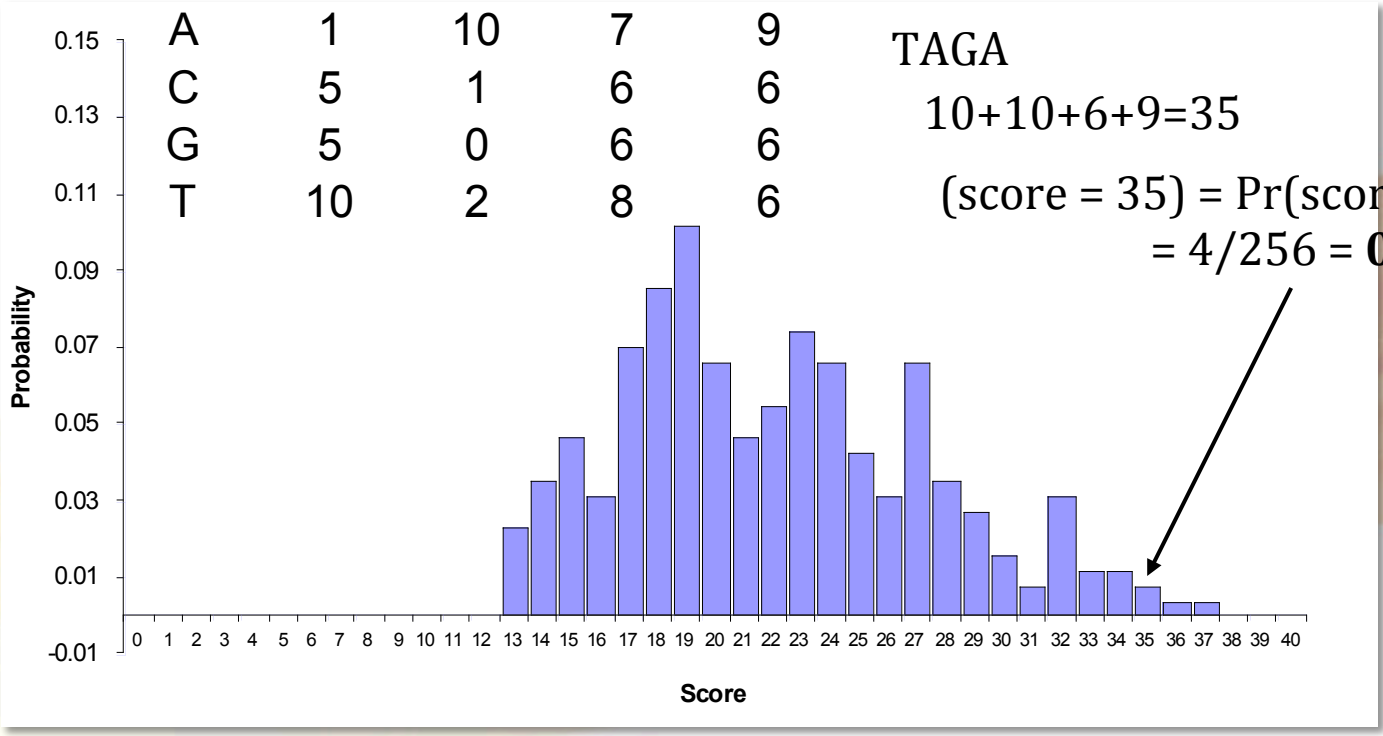
Recall,

A	-3.8	1.9	0.1	1.2
C	-1.5	-3.8	-0.8	-1
G	-1.3	-4.8	-0.6	-0.7
T	1.7	-3.2	0.8	-0.9

Before, TAGA received score of 4.2

Recalculate using the rescaled PSSM

A	1	10	7	9
C	5	1	6	6
G	5	0	6	6
T	10	2	8	6



TAGA

10+10+6+9=35

(score = 35) = Pr(score >= 35)

= 4/256 = **0.0157**

what's this?

our P-value!

This algorithm has a time complexity ~linear in the length of the PSSM, a dramatic improvement over the naive enumeration method which has time complexity exponential in the length of the PSSM!

Now We Have…

- A consensus representation and regular expression to find motifs
- But we now have:
  - A mathematical model
  - Based on probabilities
  - Let's look at some real uses of PSSM's

Prokaryotic Promoters

- Prokaryotic Promoters
  - Prokaryotic sigma 70 (σ70) promoters are characterized by a -35 consensus sequence (TTGACA) and a -10 consensus sequence (TATAAT)
  - Which are located ~35 bp and ~10 bp upstream of the transcriptional start point, respectively
    - Start point labeled +1 (no 0 exist in transcriptional nomenclature)
  - The consensus sequences vary for promoters using other sigma factors
  - Vary somewhat from species to species

Consensus

```
 -35                    -10
TTGACA................TATAAT
```

Matrix

-35 Region

									T	T	G	A	C	A	
A	11	8	8	7	8	7	3	5	5	0	1	0	14	5	9	5
C	3	4	2	4	4	3	5	2	8	1	1	2	3	11	2	5
G	3	2	4	2	4	5	5	5	5	2	1	17	1	2	3	3
T	4	7	7	8	5	6	8	9	3	17	18	2	4	3	7	9

Spacer Region

Length	9	10	11	12	13	14	15
	1	6	14	6	1	1	1

-10 Region

						T	A	T	A	A	T				
A	4	5	3	4	4	0	20	5	12	11	0	7	4	6	
C	5	4	5	4	5	2	0	3	3	4	1	2	7	6	
G	2	5	5	8	7	2	0	3	3	3	0	6	5	6	
T	10	6	8	5	6	17	1	9	3	4	20	6	5	4	

Prokaryotic Ribosomal Binding Sites

- In prokaryote
  - the AGGAGG Shine-Dalgarno consensus sequence is located 4 to 8 nucleotides 5' of the translation initiator AUG of most mRNAs
- The sequence is complementary to a CCUCCU sequence at the 3' end of 16S rRNA
- Other residues in the translation initiation region are also conserved, varying in detail from species to species

Shine-Dalgarno Consensus

AGGAGG....AUG

Plasmid RK2 Ribosome Binding Site Matrix

		A	G	G	A	G	G							C	A	A	U	G	A	A		
A	5	8	6	4	8	8	3	7	7	6	6	7	8	2	7	21	0	0	14	10	7	8
C	6	5	3	3	4	1	2	3	3	5	2	7	7	12	5	0	0	0	3	8	5	4
G	6	5	10	12	9	12	15	7	6	4	7	4	4	5	4	0	0	21	3	3	3	9
U	4	3	2	2	0	0	1	4	5	6	6	3	2	2	5	0	21	0	1	0	6	0

Prokaryotic Translation Initiation Regions

E. coli Translation Initiation Regions

	A	A	U	U	A	U	G	G	C	U	A
A	6	8	5	5	14	0	0	6	5	6	7
C	3	3	3	3	0	0	0	1	6	2	4
G	2	2	2	2	1	0	15	6	2	1	3
U	4	2	5	5	0	15	0	2	2	6	1

Plasmid RK2 Translation Initiation Regions

	C	A	C	A	A	U	G	A	A	A	G
A	7	8	2	7	21	0	0	14	10	7	8
C	7	7	12	5	0	0	0	3	8	5	4
G	4	4	5	4	0	0	21	3	3	3	9
U	3	2	2	5	0	21	0	1	0	6	0

What we have been analyzing without really knowing it, is entropy!

5 minutes…

Information and Entropy

- The modern theory of information was developed by Claude Shannon in 1948
  - Basis for most modern communication
    - A common application is ZIP files
- Entropy is a measure of the uncertainty of the results of an event
  - Number of bits (binary, yes/no decisions) needed store or communicate the results
- The entropy of an event (H) is -1 x sum of the probability of each possible outcome (px) x the base 2 logarithm of that probability,
  - $H = -\Sigma\ px\log_2 px$

Information Theory

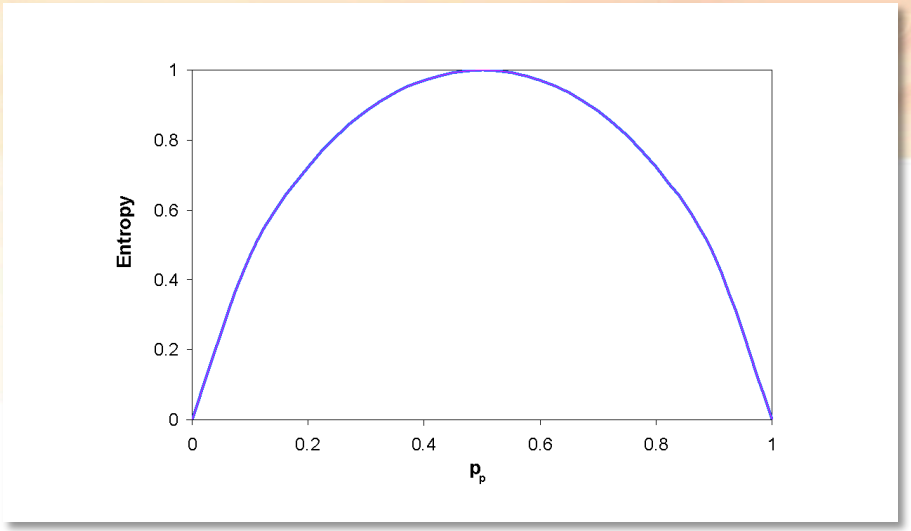
- In information theory
  - Self-information is a measure of the **information content** associated with the outcome of a random variable
  - It is expressed in a unit of information, for example:
    - **bits**
    - nats
    - hartleys
  - Depending on the base of the logarithm used in its calculation

Information Theory

- Information transmission
  - What is the minimum amount of information necessary to convey a message unambiguously?
  - How do we quantitate the uncertainty in a message?
- Finding Motifs
  - How much information is present in each position in a sequence motif?
  - How can we quantitate information content in motif sequences?
  - How can this help us understand the biological meaning of motifs?

Information Content

- Outcomes with different probabilities affect the entropy
- Entropy is maximal when all outcomes are equally likely
  - Entropy is 0 when there is only 1 possible outcome
  - Loaded die, where the probability of a 6 is 1/2 and the probability of any other number is 1/10
    - H = - (1/2log2(1/2) + 5·(1/10)log2(1/10)) = -(0.5 + 5·(1/10) ·-3.321) = 1.661 bits
  - Compare this with a fair die, which has an entropy of 2.58 bits
  - The fair die's outcome is much more uncertain than the loaded die



Information Theory for DNA

- Information content of DNA is harder to determine than merely looking at the number of base pairs and multiplying it by 2 to get the size in bits
  - Remember that each site can have up to 4 different nucleotides, or 2 bits
  - However, information theory:
    - Random sequences have the lowest information content
    - Well preserved sequences contain the maximum information content
  - In other words, the actual information content ranges:
    - From zero for totally random sequences
    - To 2 bits for conserved sequences
- How is this related to entropy?

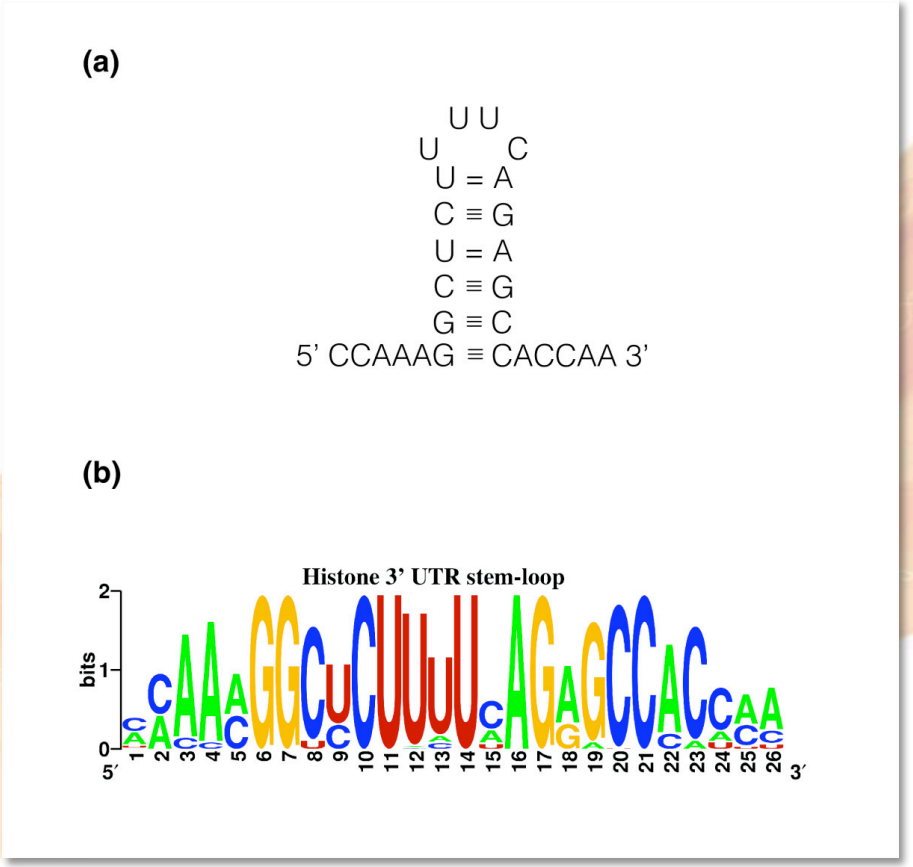
Information Content

- Uncertainty
  - Uncertainty can be thought of as the number of yes/no questions required to identify the state something is in
  - It can be measured in bits
    - A coin toss, with only 2 possibilities
    - A nucleotide, with 4 possibilities
- Maximum Uncertainty
  - **Maximum Entropy** = $\log_2(n)$ where n is the number of possible states
  - Coin $\log_2(2) = 1$ bit
  - DNA $\log_2(4) =$ **2 bits**
  - Dye $\log_2(6) = 2.585$ bits
  - Protein $\log_2(20) =$ **4.32 bits**

- **How could we visualize this?**

A Graphical Representation of Information Content

- A sequence logo is a visual representation of a IC
- Showing relative importance of different positions and which residues contribute the most
- Based on Shannon information theory
- **Height of a column equal to IC**
- **Height of the base proportional to frequency of base on that position** … more specifically known as “bits” , “information content” , or “entropy”



[Multiple independent evolutionary solutions to core histone gene regulation]

A Graphical Representation of Information Content

- A sequence logo is a visual representation of a IC
  - If there is only 1 amino acid ever found at a position
    - Completely conserved
    - There is no uncertainty about it, so its entropy is **0**
    - The information content is 4.32 bits information

Shannon Entropy

- What does this mean???
- H is a measure of entropy or randomness or disorder
  - Tells us how much uncertainty there is for the different amino acid abundances at one position in a sequence motif

Information Content of DNA

Lets say we're looking for protein-DNA interactions:
The least variable positions likely are important for specifying conservation,
i.e. the protein-DNA interaction
Therefore high information content = low sequence variation at that position

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

Information Content at position *i*:

$$\mathrm{IC}i = \log_2(4) + \sum_{b=G,A,T,C} P\mathrm{b}(i) * \log_2(P\mathrm{b}(i))$$

Where $P\mathrm{b}(i)$ is the probability of base b at position i

If using log2, the info content is in 'bits'

Maximum IC if *P* of some base is 1.0: = 2

Minimum IC if *P* is 0.25 for all bases: = 0

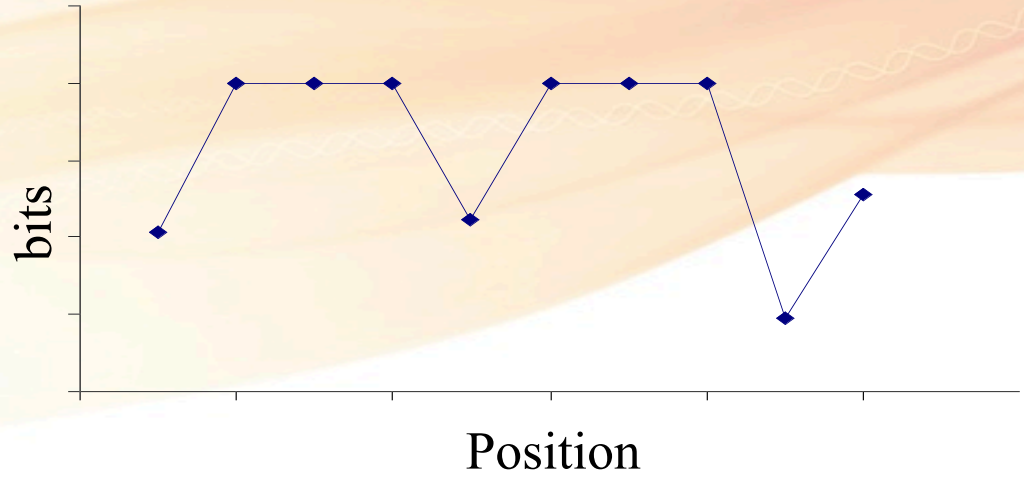
Information Content of DNA

The least variable positions likely are important for specifying the protein-DNA interaction

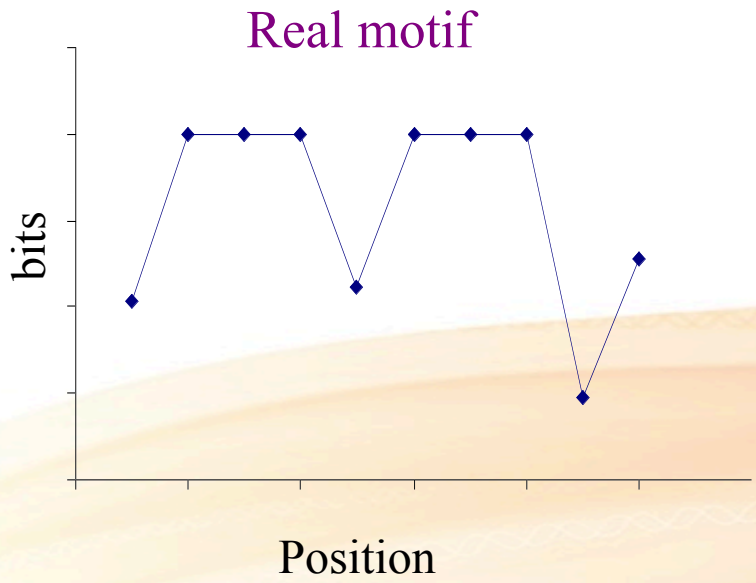
Therefore high information content = low sequence variation at that position.

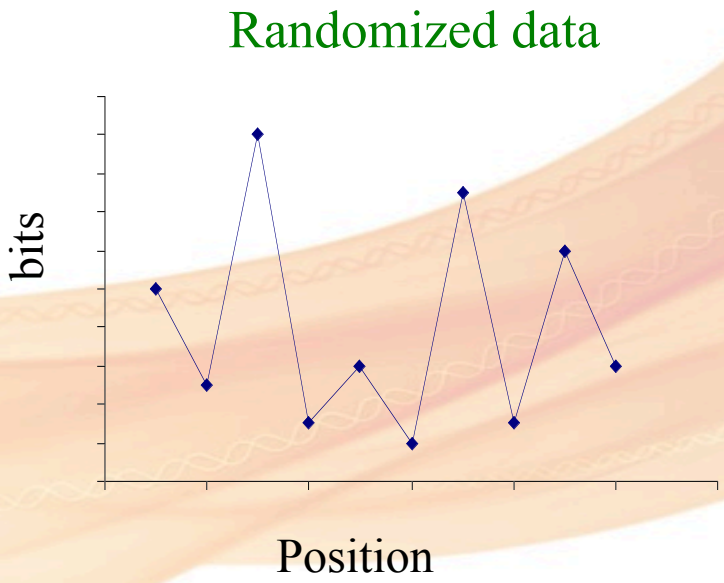
G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8	
A	0.4	0	1.0	0	0	0	1.0	0	0	0	
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2	
C	0	0	0	0	0	0	0	0	0.2	0	
IC	1.0	2.0	2.0	2.0	1.1	2.0	2.0	2.0	0.5	1.3	= bit score of 15.9

Information Profile:

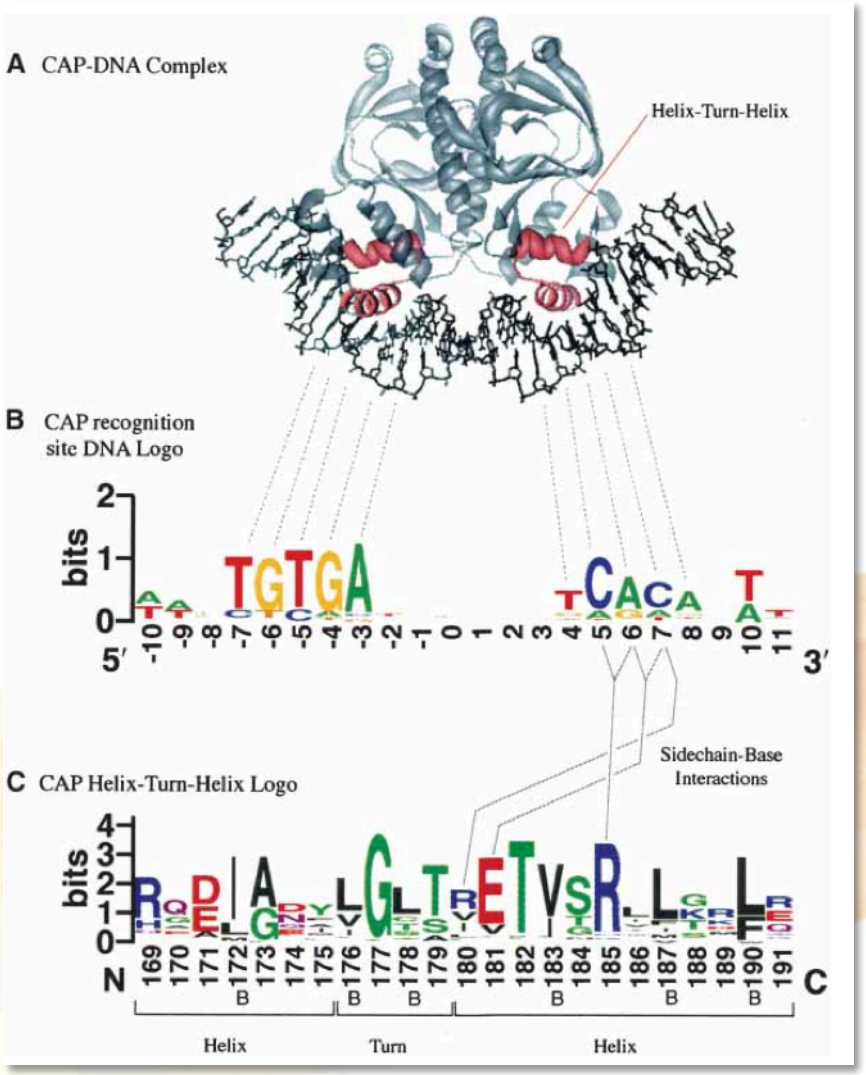


Protein-DNA Interactions





Please read Sequence_logo.pdf (Ivan Erill) from the course website:



A Combined Sequence Logo

For Thursday

- Please read Sequence_logo.pdf (Ivan Erill) from the course website:
- We will have a lecture on Thursday
- Come prepared for a quiz