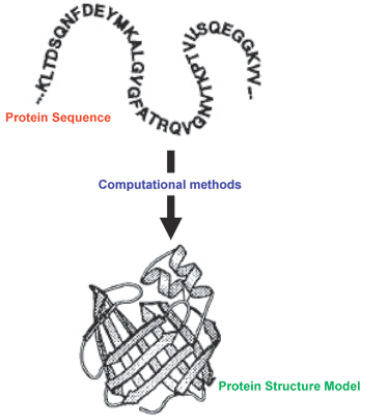
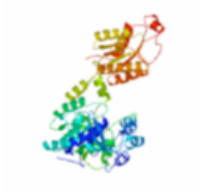
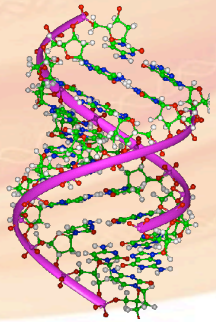




Bioinformatics Computational Methods 1 - BIOL 6308



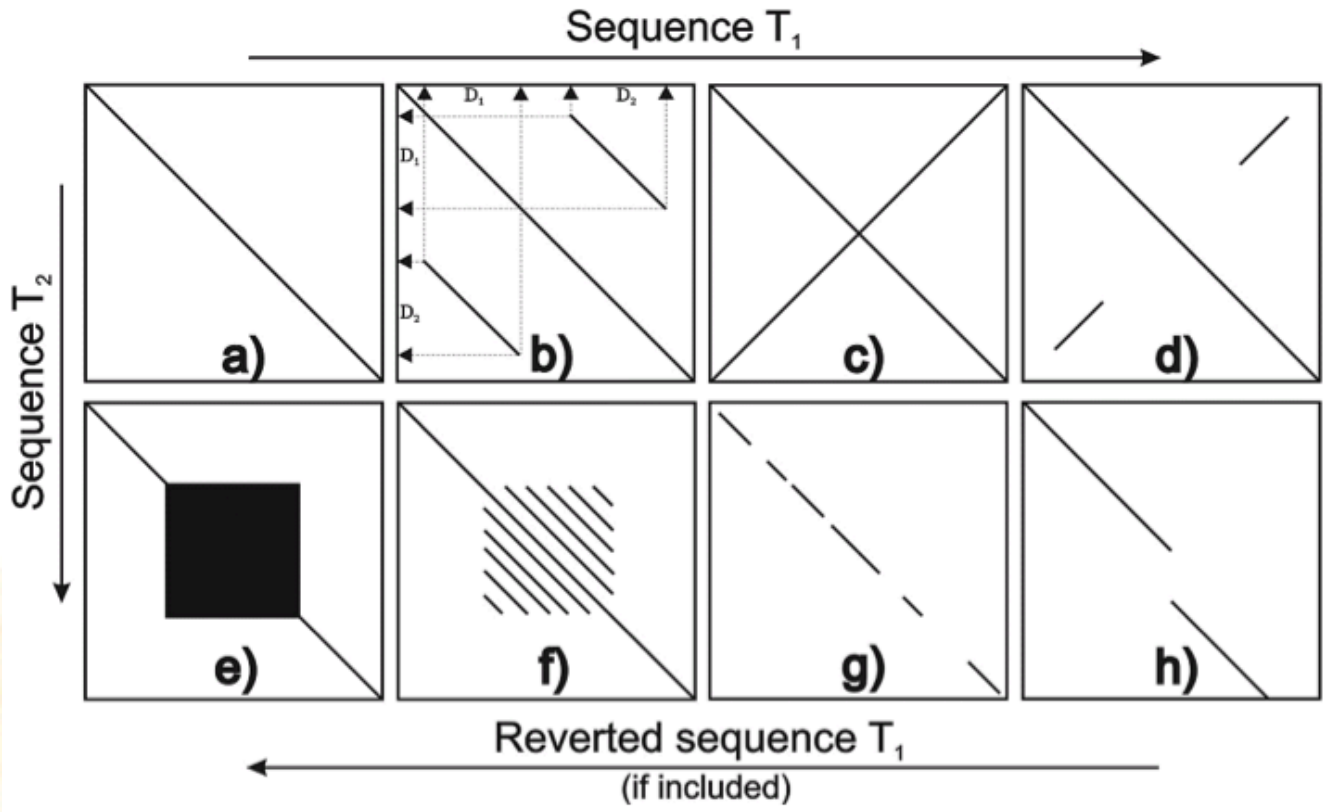
November 5th 2013

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

Last Time

- MeSH Vocabulary
  - Getting More Out of Your PubMed Searches
  - Boolean Logic
- Entrez and the Power of NCBI's Vocabulary
- Sequence Formats
- Sequence Alignments
  - Why Do We Align Sequences?
  - The Characters in an Alignment
  - Manual Alignments
  - **Dot Plots**
    - How:
      - Generate Dot Plots
      - To Analyze a Dot Plot
  - Global Vs Local

Good Stuff!

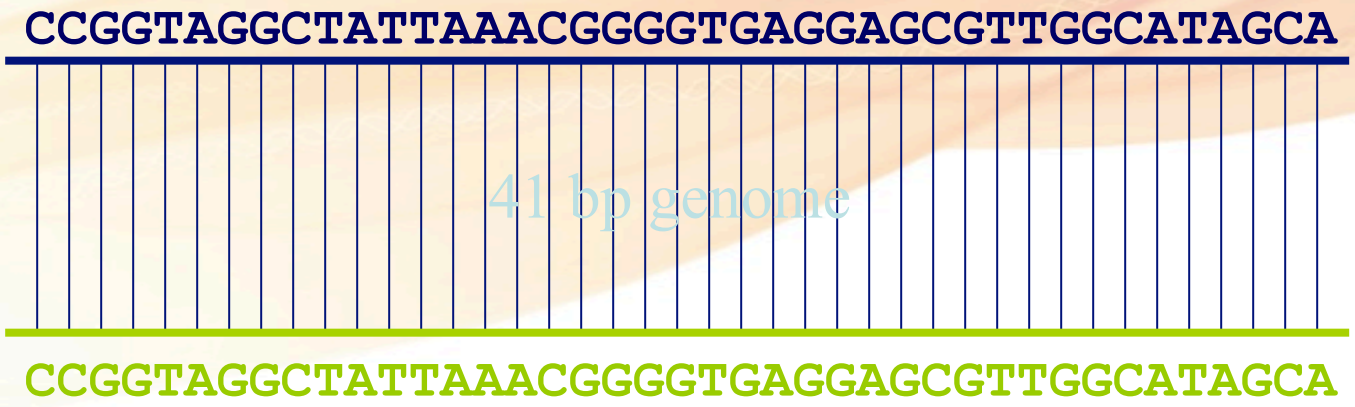


PART OF LAB
http://www.code10.info/index.php?option=com_content&view=article&id=64:inroduction-to-dot-plots

Jan Schulz, Florian Leese and Christoph Held

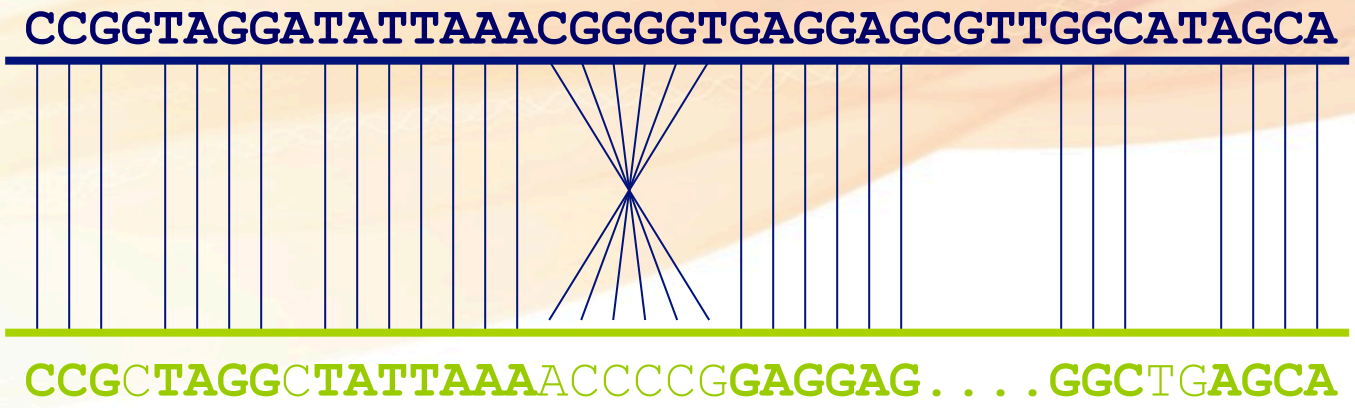
Goal of Whole Genome Alignments (WGA)

- For two genomes, *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*



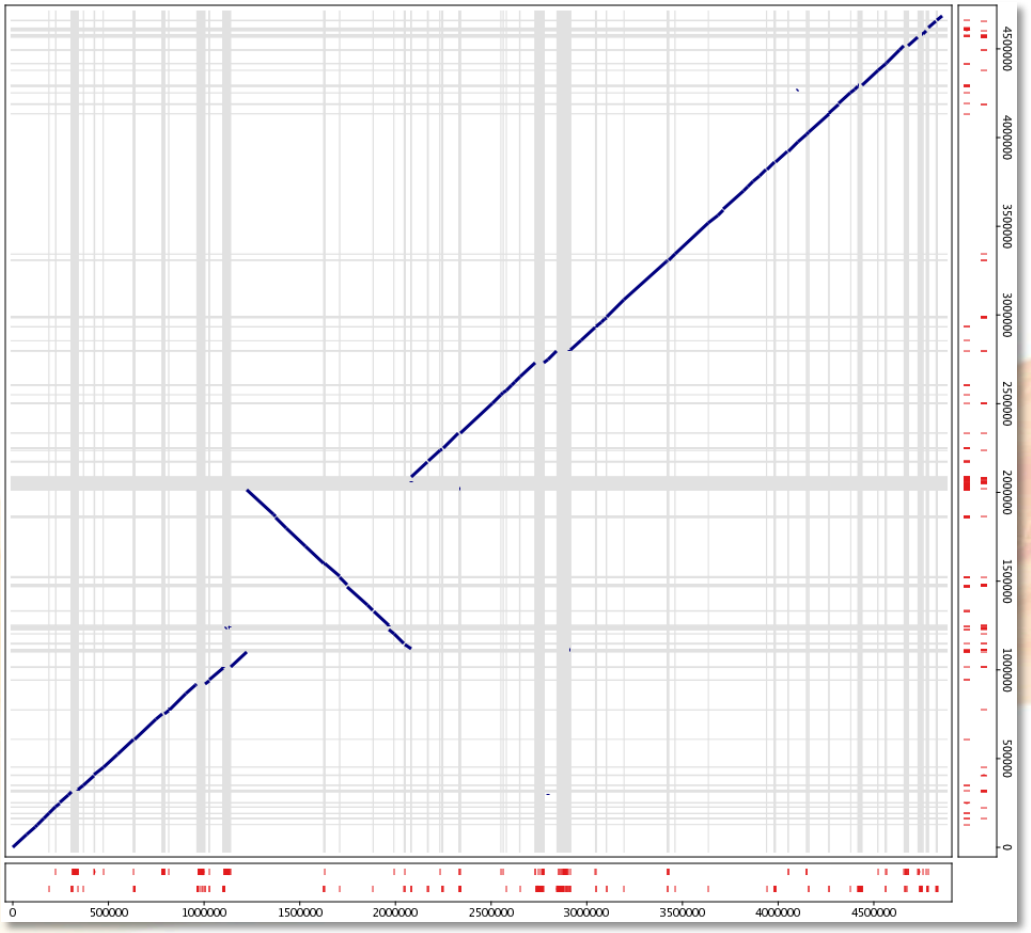
Not so fast...

- Genome A may have:
  - Insertions, deletions, translocations, inversions, duplications or SNPs with respect to B (sometimes all of the above)
  - Ideal for a Dotplot



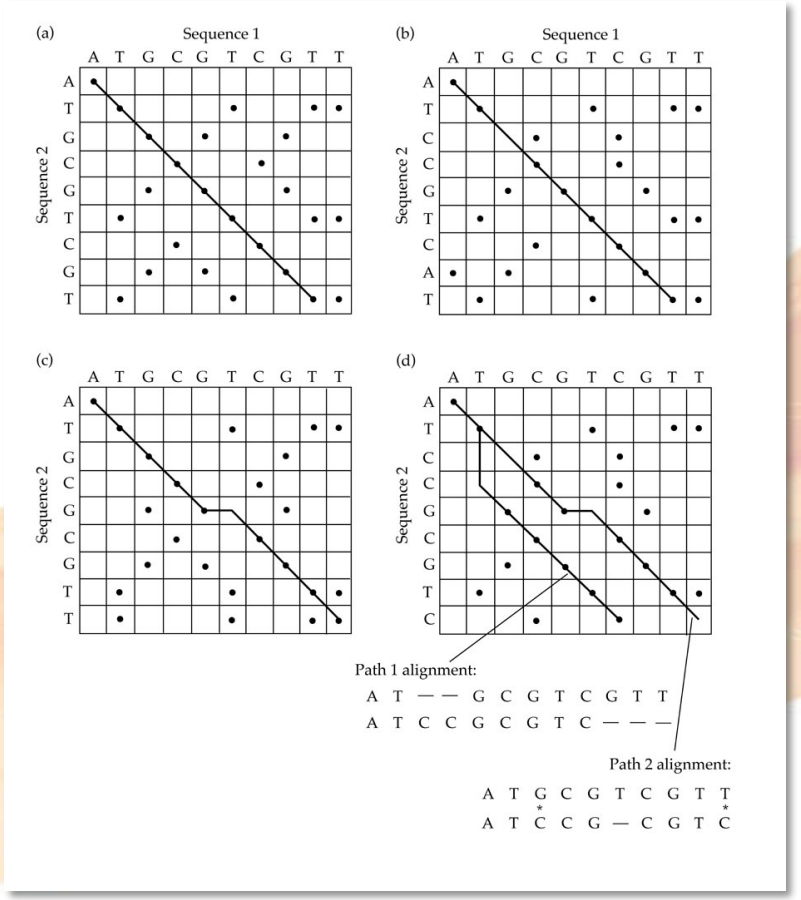
Adam M Phillippy

Two Genomes Compared with a Dot Plot



Dot Plot

- Remember one of the Disadvantages:
  - May not identify the “best alignment”
  - **How do we find the best alignment?**
- **We need a more sensitive Dot Plot line**
- **Before we can do this, we need to know**
  - **The Characters**
  - **Sliding Window**
  - **Scoring pairs**



The Characters!

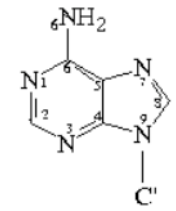
Nucleotide Bases

DNA

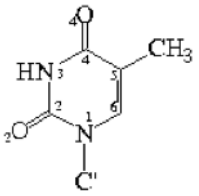
- A = Adenine
- T = Thymine
- C = Cytosine
- G = Guanine

RNA

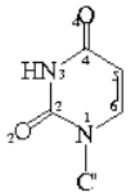
- A = Adenine
- U = Uracil
- C = Cytosine
- G = Guanine



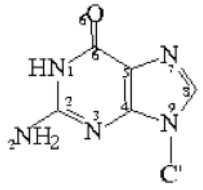
Adenine



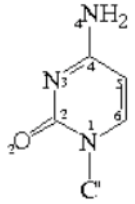
Thymine



Uracil



Guanine



Cytosine

Purines

Pyrimidines

IUPAC Nucleic Acid Codes

Know these!

DNA:

Nucleotide Code:	Base:
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

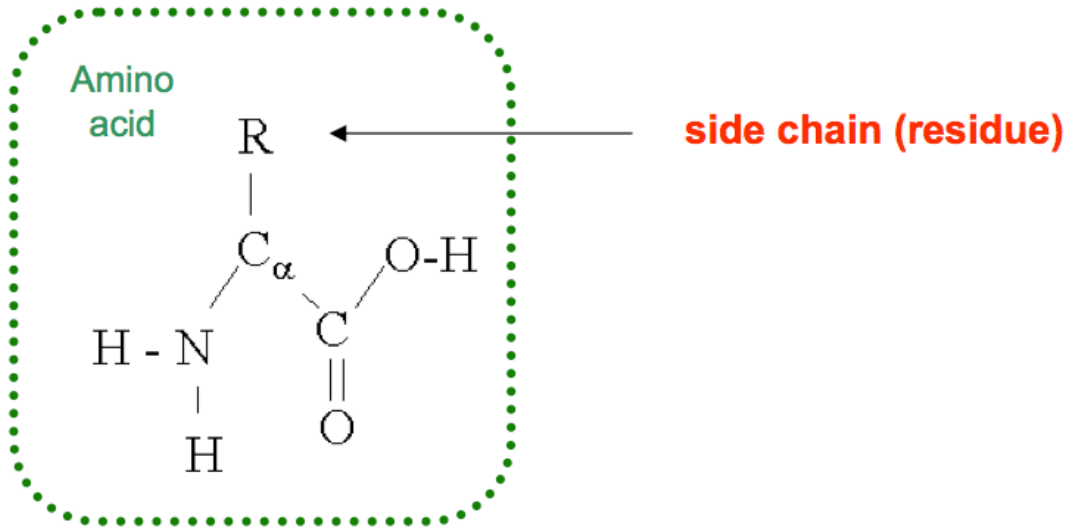
Amino Acids

Proteins are composed of amino acids

Composition

- **Core + side chain (residue)**
- **20 different residues → 20 unique amino acids**

A C D E F G H I K L M N P Q R S T V W Y



IUPAC Amino Acid Codes

Protein:

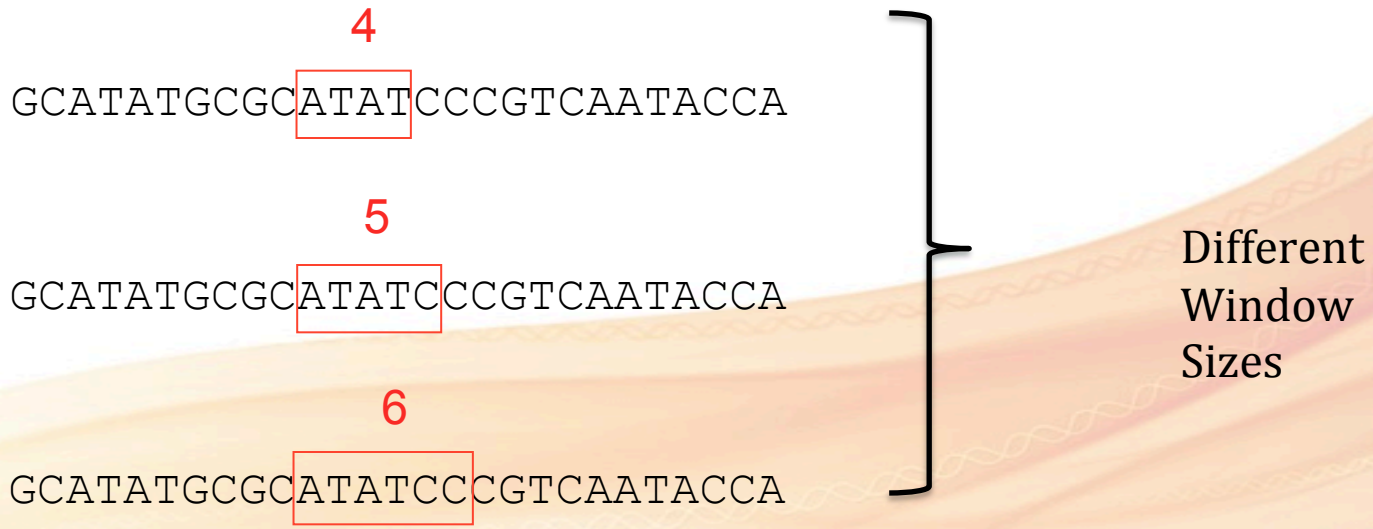
Amino Acid Code:	Three letter Code:	Amino Acid:
A	Ala	Alanine
B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
X	Xaa	Any amino acid
Y	Tyr	Tyrosine

Know all of them **except for B**!

The Sliding Window

Sliding Window Approach

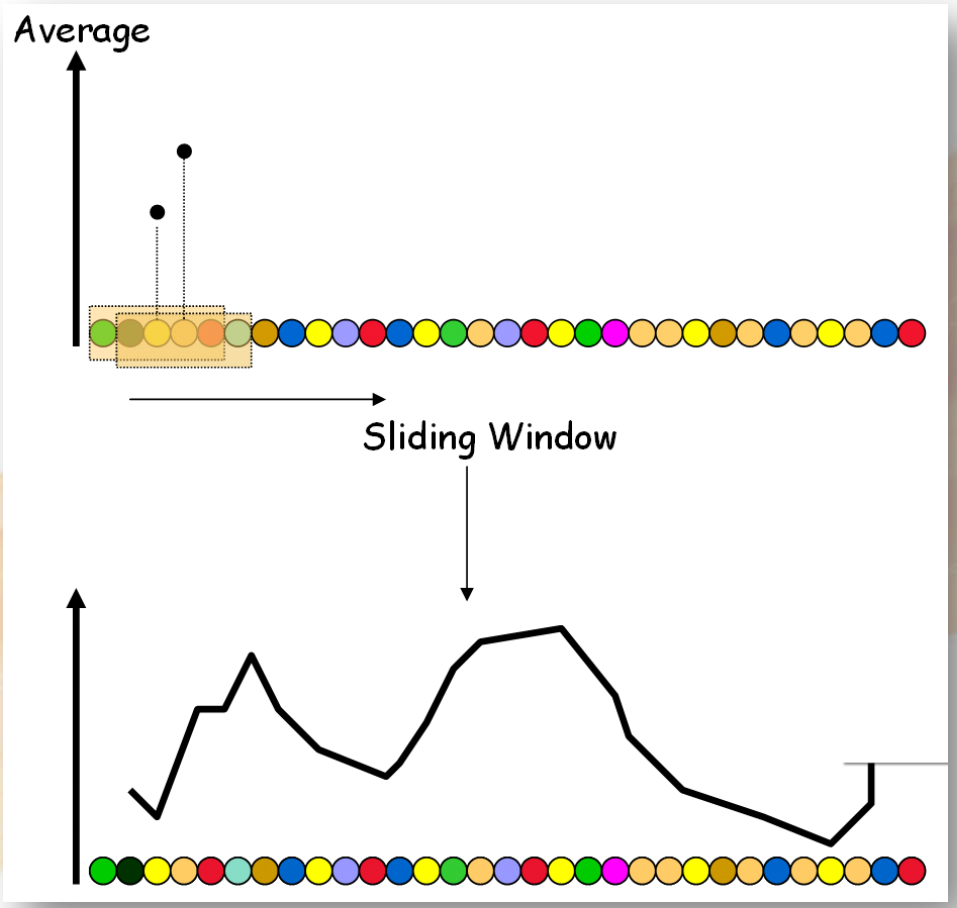
- A sliding window-gathers information about properties of nucleotides or amino acids



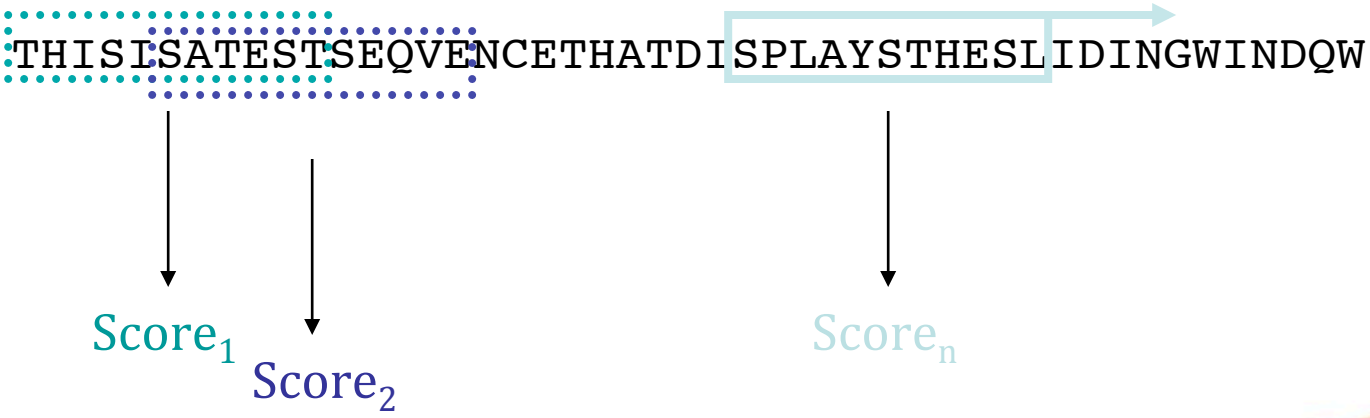
- A simple example:
  - Used to calculate the %G+C content within a window
  - Then move the window one nucleotide and repeat the calculation

Ideal for Identifying Strong Signals

- Very simple methods
  - Few artifacts
  - Not most sensitive of methods
- Make the window the same size as the feature you're looking for
- ProtScale example:
- http://web.expasy.org/protscale/

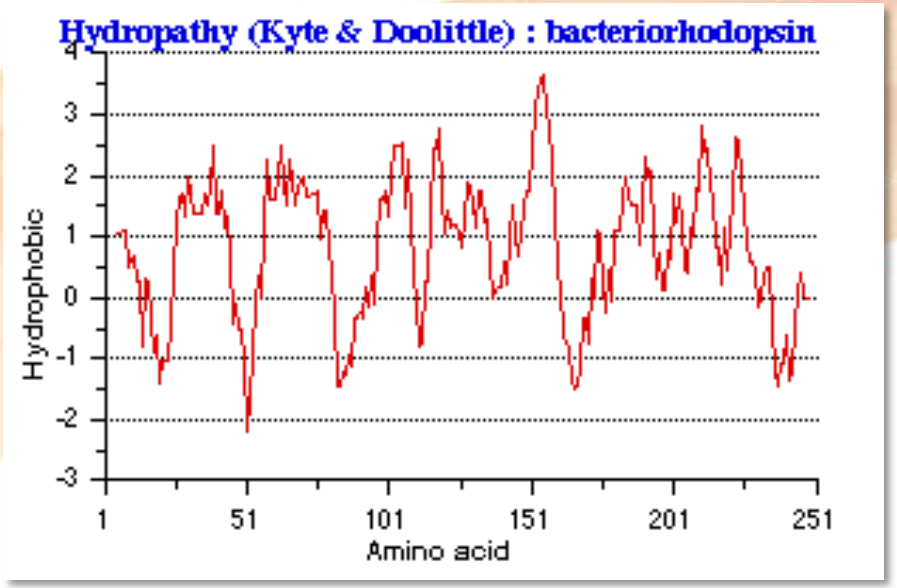


Sliding Window – Usually 1 or 2 Parameters



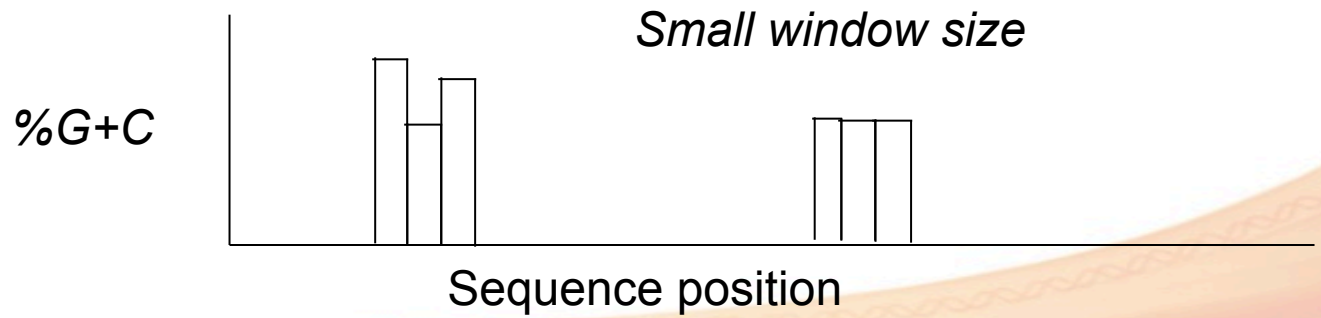
Width or Size = 11, Step = 5

Results are usually displayed as a graph, see example ->

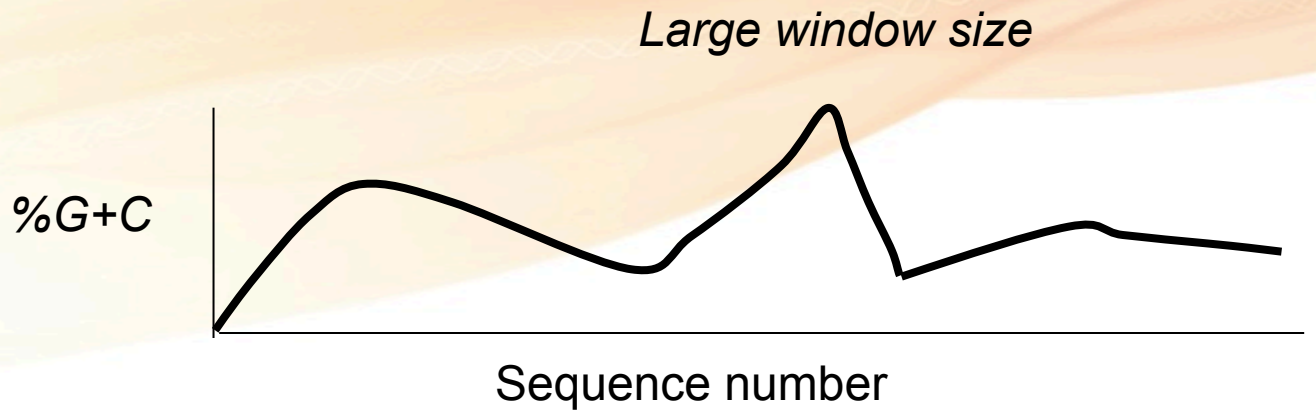


Problems with Sliding Window

- If the window is too small it is difficult to detect the trend of the measurement

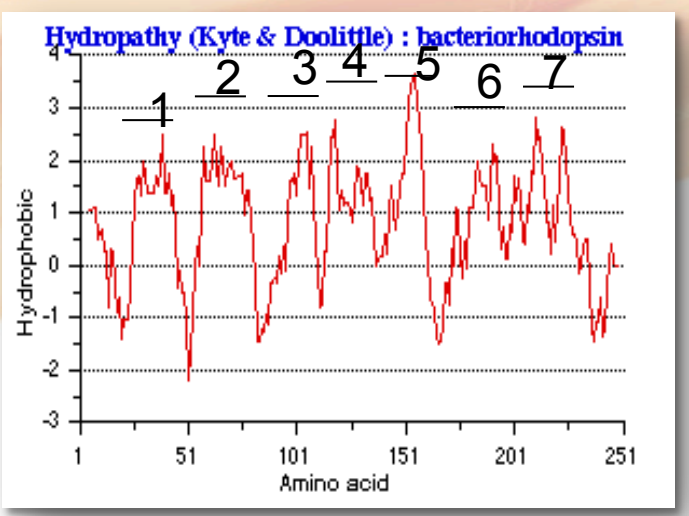


- If too large you could miss meaningful data



Hydrophobicity Profile

- Plot the variation of **hydrophobicity** along the sequence of a protein
  - Defines relative hydrophobicity of a.a. residues
  - More positive - more hydrophobic the a.a. are in that region of the protein
- These profiles have been used to predict:
  - Positions of turns between secondary structure
  - Exposed and buried residues
  - Antigenic sites
  - Transmembrane domains



Example Usage of Sliding Windows

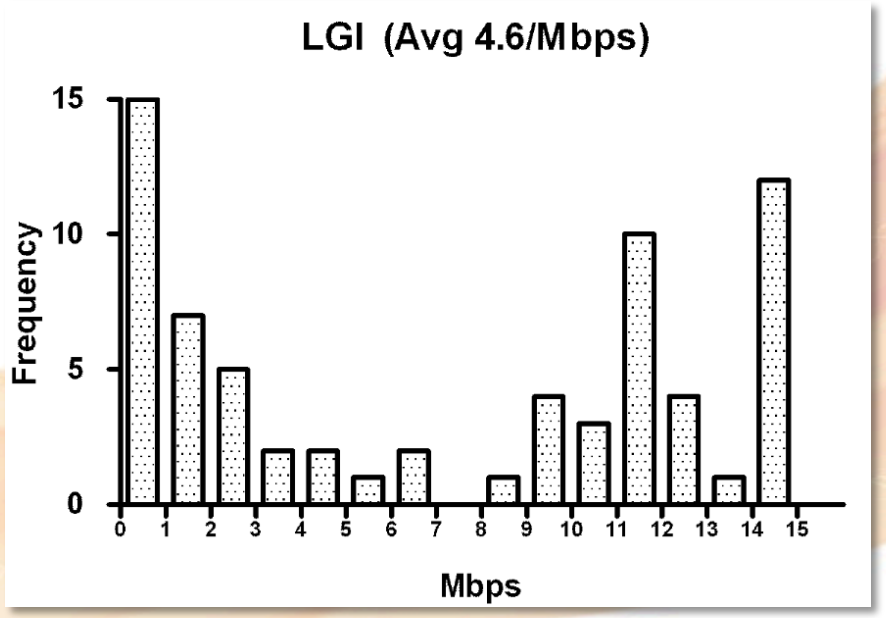
GC Tracts

- Homopolymeric poly-G/poly-C tracts (G/C tracts)
- In the genome of *Caenorhabditis elegans*
  - Exist at high frequency
  - Maintained by the activity of the DOG-1 protein
- The probability of one stretch of 18 guanines
  - By chance in 100 Mb AT-rich *C. elegans* genome (GC content 36%)
    - **~1/6 to the 18th power - 1 in 100 trillion**
  - **396** G/C tracts were found
- **Overrepresented**
  - **396** G/C tracts were found
  - Human (**200** in 3.3 Bbp) and yeast (**1** in 12 Mbp) genomes

Poly-G/poly-C tracts in the genomes of Caenorhabditis, Zhao et al, BMC Genomics. 2007 Nov 7;8:403.

Sliding Window – G/C Tracts

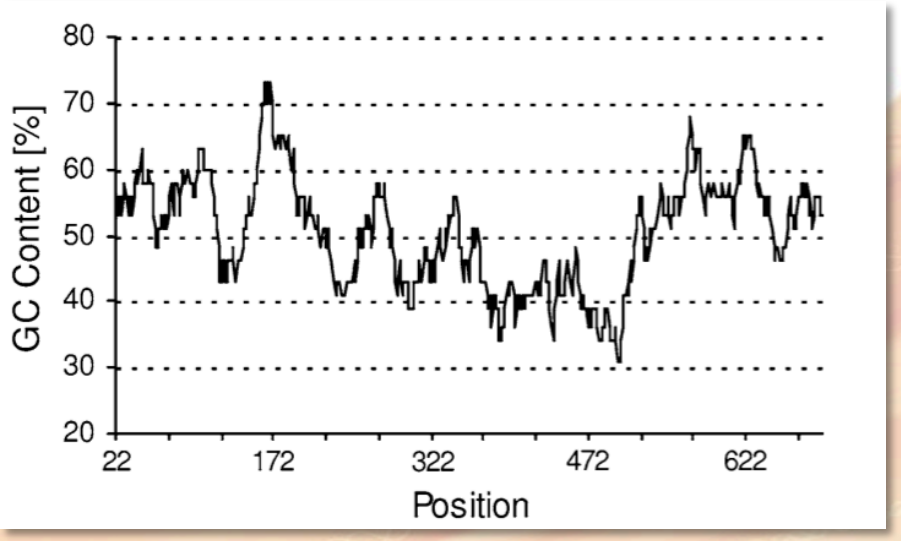
- G/C tracts range in size from **18 to 32 base pairs**
- Distribution of G/C tracts in every mega-base pair block on each chromosome of *C. elegans*
- **X** axis represents:
  - length of the chromosome that was divided by million base pairs
- **Y** axis represents:
  - Is the frequency of G/C tracts



Poly-G/poly-C tracts in the genomes of Caenorhabditis, Zhao et al, BMC Genomics. 2007 Nov 7;8:403.

Sliding Window – GC Content

- Important characteristic of DNA seq
- **Extremely high** or **low** GC content:
  - More difficult to handle with standard molecular biological techniques
  - Such as PCR or Sequencing
- Sliding window can be used to analyze the GC content of a gene of interest
- GFP gene sequence has been shown on the right



The GeneOptimizer Algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization Systems and Synthetic Biology **Vol.** 4 **Issue** 3, 2010

Transmembrane - Our Starting Material

- FREL_CANAL - http://www.uniprot.org/uniprot/P78588.html
- UniProt annotations:

Regions

Transmembrane	122 – 142	21	Helical; Potential	
Transmembrane	198 – 218	21	Helical; Potential	
Transmembrane	234 – 254	21	Helical; Potential	
Transmembrane	281 – 301	21	Helical; Potential	
Transmembrane	313 – 333	21	Helical; Potential	
Transmembrane	340 – 360	21	Helical; Potential	
Transmembrane	499 – 519	21	Helical; Potential	
Domain	239 – 373	135	Ferric oxidoreductase	
Domain	374 – 492	119	FAD-binding FR-type	
Nucleotide binding	437 – 442	6	FAD Potential	
Compositional bias	65 – 77	13	Poly-Ser	

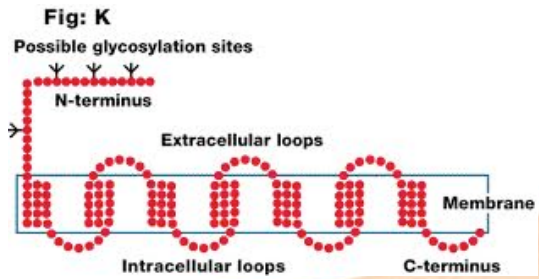
*FREL_CANAL = FREL_CANAX

Transmembrane Domains (TD)

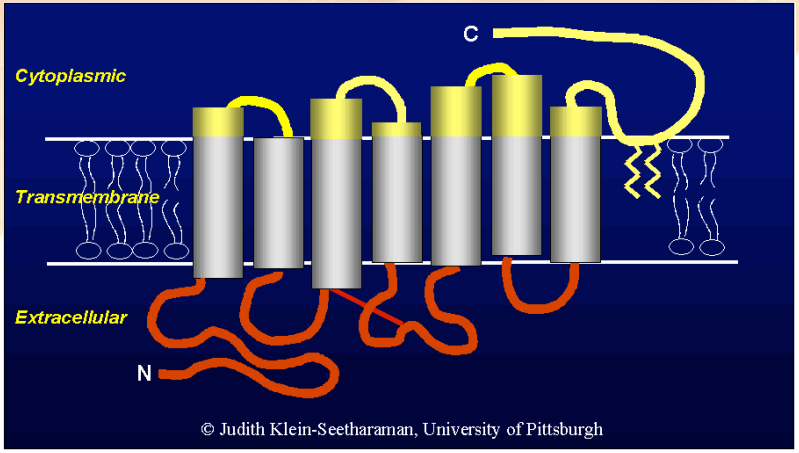
- Usually denotes a single transmembrane alpha helix of a transmembrane protein
- It is called a "domain"
  - Because an alpha-helix **in a membrane** can fold **independently** from the rest of the **protein**
  - Similar to domains of water-soluble proteins
- More broadly:
  - **TD** is any three-dimensional protein structure which is thermodynamically stable in a membrane
  - Usually about 20 a.a in length

More on Transmembrane Domains

- Many important receptors have 7 transmembrane domains
- Most of the members of 7 TM receptors are G protein-coupled receptors
- Examples of 7 TM receptors are sensory and neurotransmitter receptors
- Remember G-proteins?



- *Binds a signaling molecule on the extracellular side of the membrane and transduces a signal on the cytoplasmic side to initiate or inhibit biochemical reactions within the cell*



Using ProtScale

ProtScale

ProtScale [Reference / Documentation] allows you to compute and represent the profile produced by any amino acid scale on a selected protein.

An **amino acid scale** is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but many other scales exist which are based on different chemical and physical properties of the amino acids. This program provides 57 predefined scales entered from the literature.

Enter a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL accession number (AC) (e.g. **P05130**) or a sequence identifier (ID) (e.g. **KPC1_DROME**):

FREL_CANAL

Or you can paste your own sequence in the box below:

http://web.expasy.org/protscale/

Please choose an amino acid scale from the following list. To display information about a scale (author, reference, amino acid scale values) you can click on its name.

Hydrophobicity Profile

- ○ Molecular weight
- ○ Bulkiness
- ○ Polarity / Grantham
- ○ Recognition factors
- ○ Hphob. OMH / Sweet et al.
- ◉ Hphob. / Kyte & Doolittle
- ○ Hphob. / Abraham & Leo
- ○ Number of codon(s)
- ○ Polarity / Zimmerman
- ○ Refractivity
- ○ Hphob. / Eisenberg et al.
- ○ Hphob. / Hopp & Woods
- ○ Hphob. / Manavalan et al.
- ○ Hphob. / Black

.....

Window size: 19

Relative weight of the window edges compared to the window center (in %): 100

Weight variation model (if the relative weight at the edges is < 100%): ◉ linear ○ exponential

Do you want to normalize the scale from 0 to 1? ○ yes ◉ no

If you need more information about how to set these parameters, please click here.

Submit Reset

*FREL_CANAL = FREL_CANAX

Using ProtScale

- Transmembrane segments found using ProtScale
- FREL_CANAL - http://www.uniprot.org/uniprot/P78588.html

ProtScale

Selection of endpoints on the sequence

FREL_CANAX (P78588)

Probable ferric reductase transmembrane component (EC 1.16.1.7) (Ferric-chelate reductase) Candida albicans (Yeast).

Please select one of the following features by clicking on a pair of endpoints, and the computation will be carried out for the corresponding sequence fragment. By default, the complete sequence is used.
Note: Only the features corresponding to subsequences of at least 20 residues are highlighted.

FT	**CHAIN**	**1-669**	**Probable ferric reductase transmembrane**
FT	TRANSMEM	122-142	Helical; (Potential).
FT	TRANSMEM	198-218	Helical; (Potential).
FT	TRANSMEM	234-254	Helical; (Potential).
FT	TRANSMEM	281-301	Helical; (Potential).
FT	TRANSMEM	313-333	Helical; (Potential).
FT	TRANSMEM	340-360	Helical; (Potential).
FT	TRANSMEM	499-519	Helical; (Potential).
FT	DOMAIN	239-373	Ferric oxidoreductase.
FT	DOMAIN	374-492	FAD-binding FR-type.
FT	NP_BIND	437-442	FAD (Potential).
FT	COMPBIAS	65-77	Poly-Ser.

Or, if you wish to select a different sequence fragment (at least 20 amino acids long), you can enter the desired endpoints on the sequence here (by default, the computation will be carried out for the complete sequence).

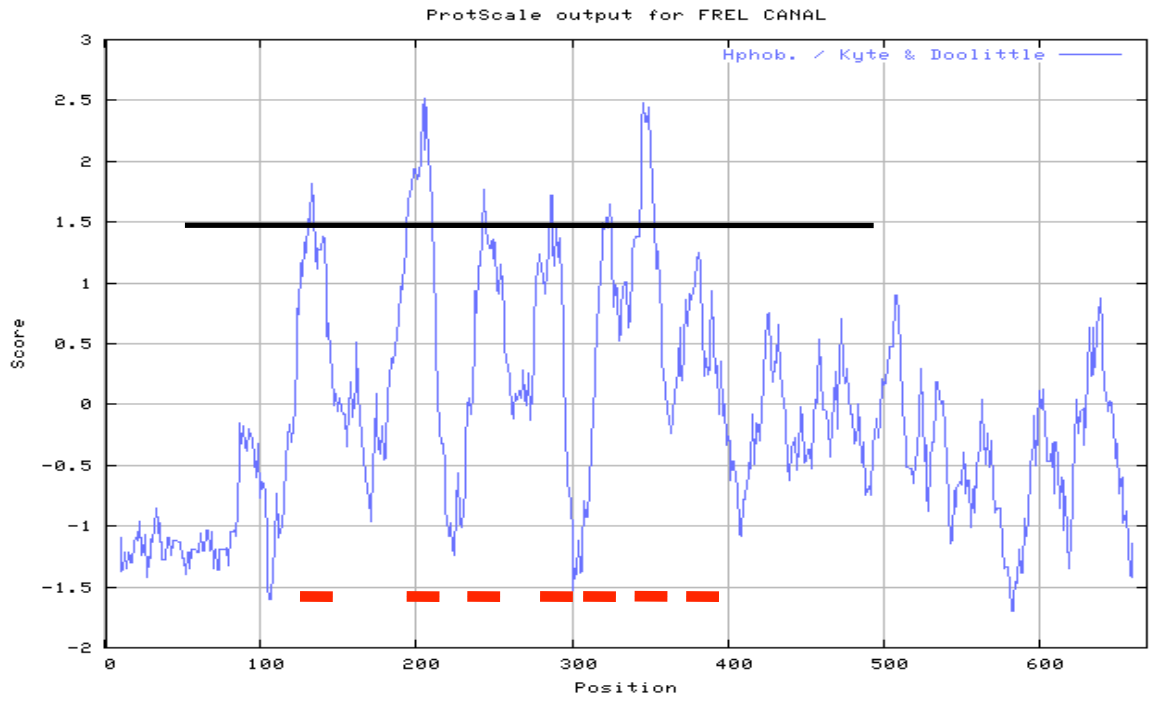
N-terminal:
C-terminal:

The sequence FREL_CANAX consists of 669 amino acids.

RESET SUBMIT

Using ProtScale

- Provides a raw signal
- Does not include interpretation of the results in terms of a score
- Consider strong signals
- In order to confirm a possible interpretation, one could slightly change the window size, or replace the scale by another similar one (e.g. two different hydrophobicity scales), and ensure that the strong signal is still present



Red = Uniprot predicitions

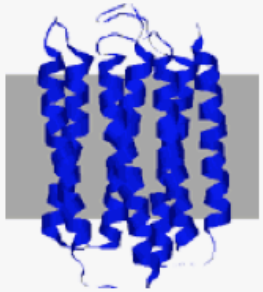
Using TMHMM

- TMHMM is the best method for predicting transmembrane domains
- TMHMM uses an HMM
  - So its very different from that of ProtScale
  - TMHMM output is a prediction
  - http://www.cbs.dtu.dk/services/TMHMM-2.0/

Using TMHMM

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins



NOTE: You can submit many proteins at once in one fasta file. Please limit each submission to at most 4000 proteins. Please tick the 'One line per protein' option. Please leave time between each large submission.

Instructions

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

Browse...

OR by pasting sequence(s) in FASTA format:

```
GHAFIHFLKPTLFLQSHPFTFTTTESNDKIVLYAKIKNGITSNIAKYLSPLPGNTATIRV
LVEGPYGEPSSAGRNCKNVVFVAGGNGIPGIYSECVDLAKKSKNQSIKLIWIIRHWKSLS
WFTEELEYLKKTNVQSTIYVTQPQDCSGLECFEHDVSFEKKSDEKDSVESSQYSLISNIK
QGLSHVEFIEGRPDISTQVEQEVKQADGAIGFVTCGHPAMVDELRFAVTQNLNVSKHRVE
YHEQLQTWA
```

Output format:

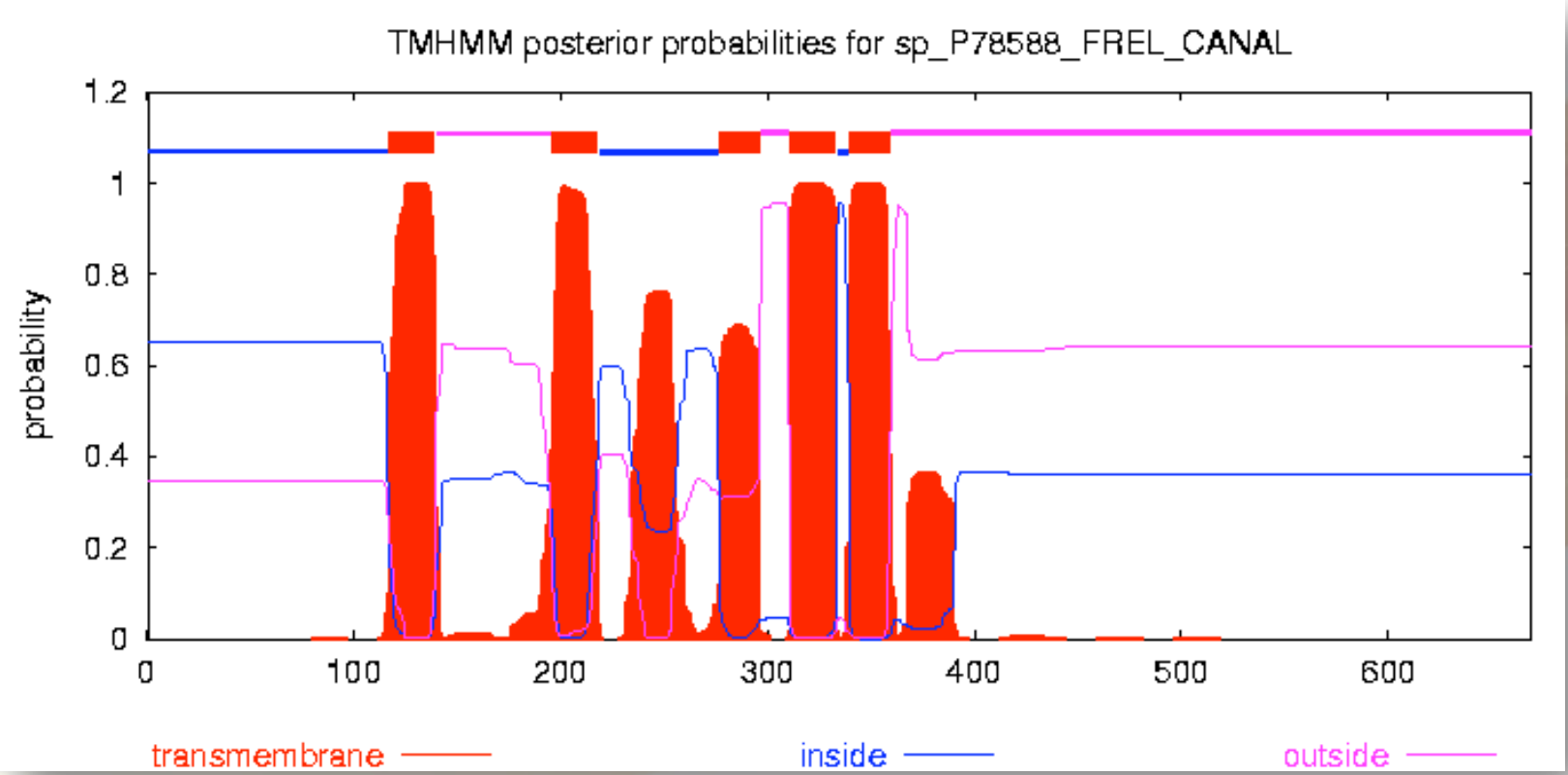
- (•) Extensive, with graphics
- ( ) Extensive, no graphics
- ( ) One line per protein

Other options:

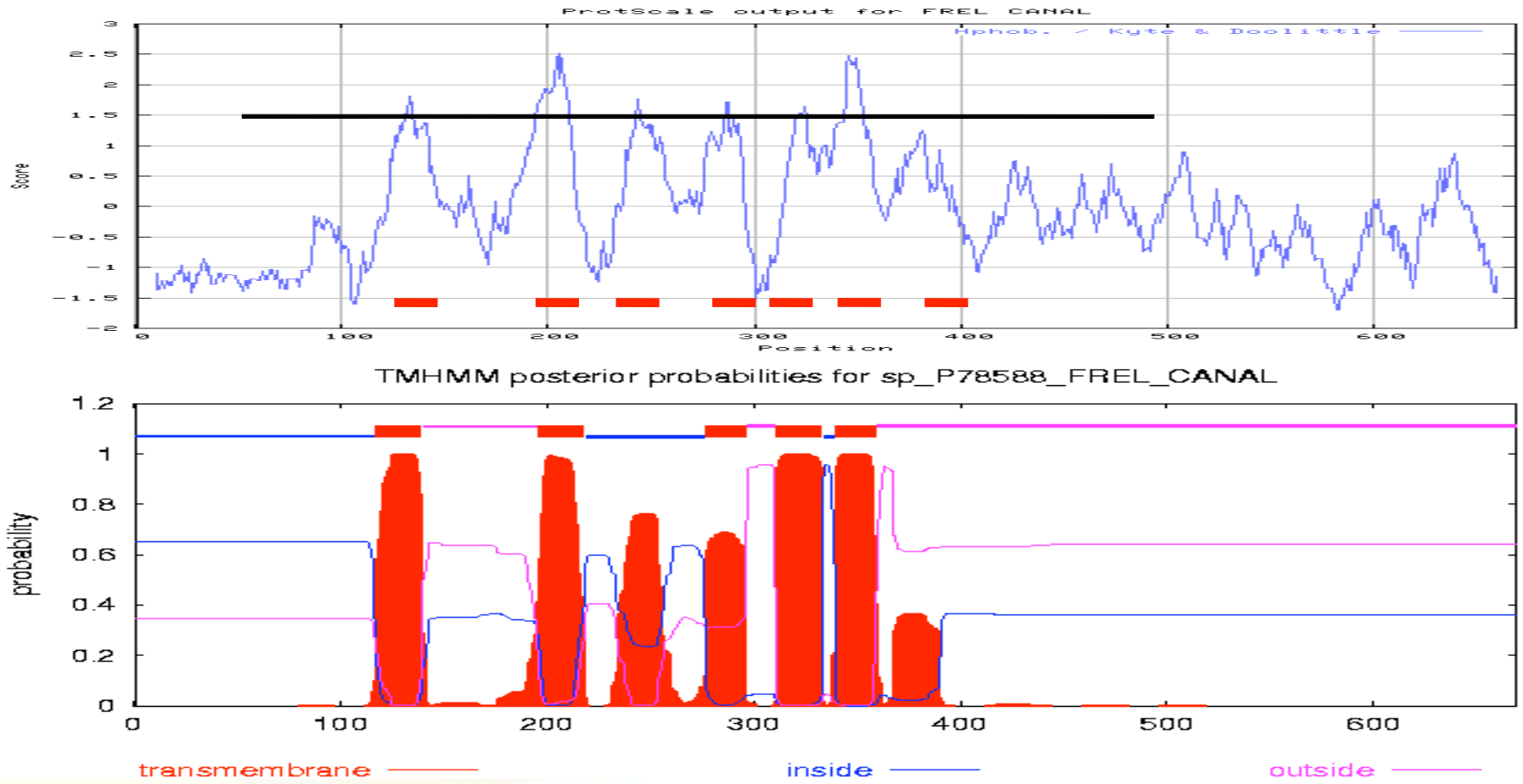
- [ ] Use old model (version 1)

Submit Clear

Using TMHMM


TMHMM posterior probabilities for sp_P78588_FREL_CANAL
probability
1.2
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
600
transmembrane
inside
outside

ProtScale vs TMHMM



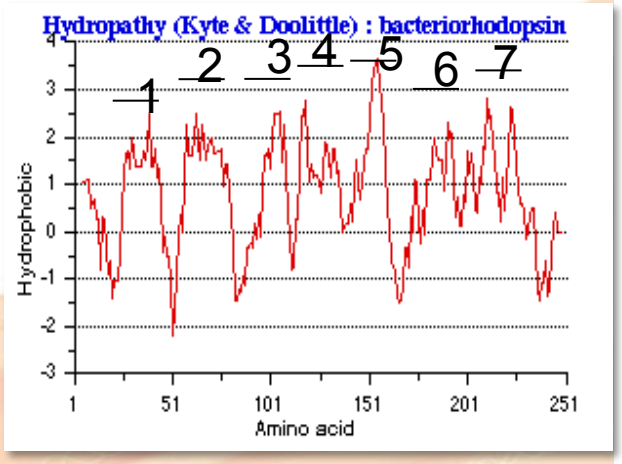
Important Point About TD

- They may be predicted on the basis of hydrophobicity scales
  - Why is this?
  - **Interior** of the **bilayer** and the **interiors** of most **proteins** of known structure are **hydrophobic**
  - **Presumed** to be a **requirement** of the amino acids that **span a membrane** be **hydrophobic** as well
- However:
  - Membrane pumps and ion channel
    - Contain numerous charged and polar residues w/in the generally non-polar transmembrane segments
  - That's why a sliding window works nicely!

http://en.wikipedia.org/wiki/Hydrophobicity_scales

End Results

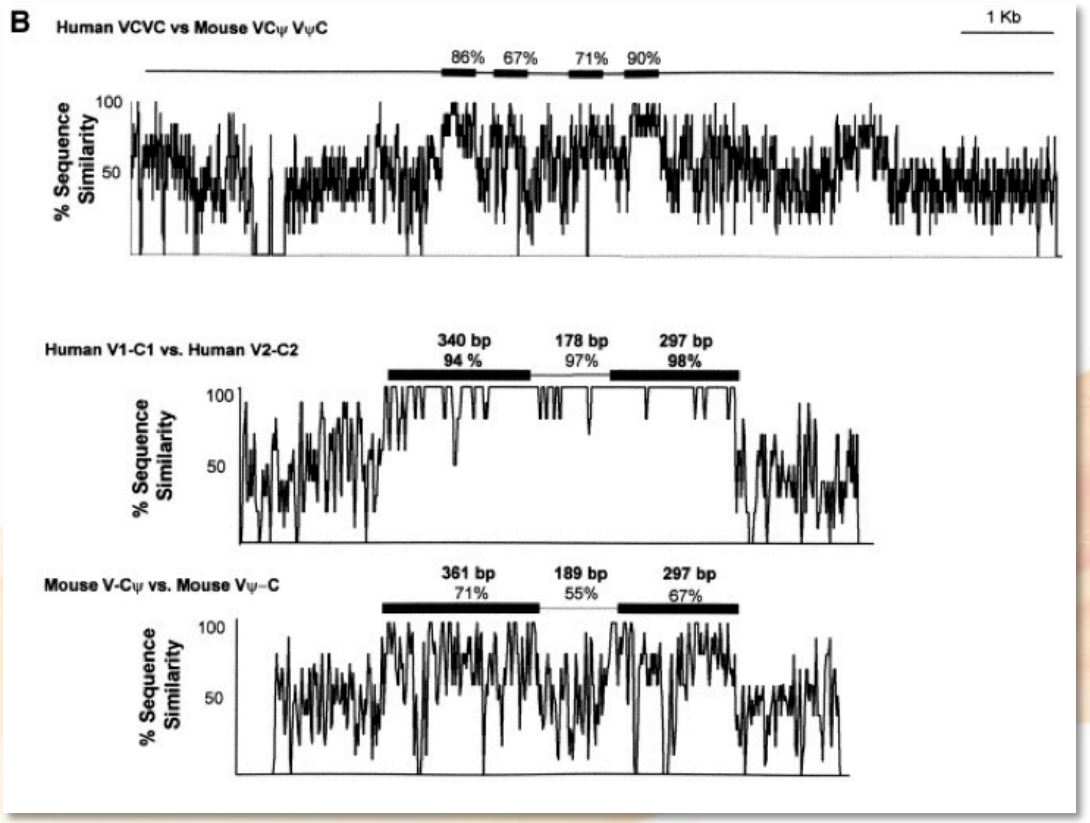
- Enables a prediction about the "transmembrane topology" of a protein
  - i.e:
    - prediction of what parts of it protrude into the cell
    - what parts protrude out
    - how many times the protein chain crosses the membrane
- It is a rough measure:
  - Additional algorithmic implementation (HMM) help with the sensitivity of the sliding window
  - THHMM - http://www.cbs.dtu.dk/services/TMHMM-2.0/



How Could I use a Sliding Window For a Pairwise Alignment?

- How can we visualize alignments?
- With an identity plot
  - XY plot
    - Let x = position in gene A
    - Let y = %identity of A_x to corresponding position in B
  - Plot the identity function
  - This can reveal conservation in genes

Interspecies and Intraspecies Identity Plots

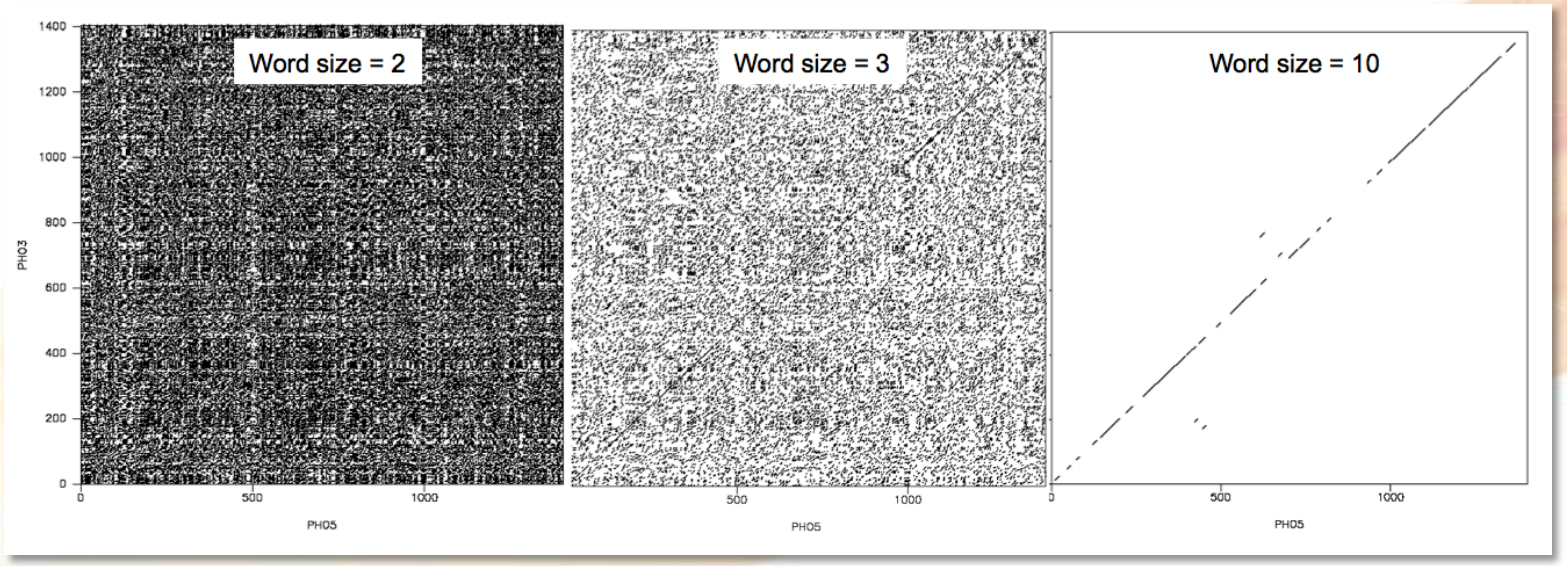


Duplication of primate and rodent B7-H3 immunoglobulin V- and C-like domains: divergent history of functional redundancy and exon loss

A More Informative Dot Plot Line Using a Model

Dot Plot with Word Matches

- With dna seqs - each residue is expected to be a match every 4 positions on average
  - A letter-based dot plot = confusing
  - Instead we use word matches



Alignment of PHO5 and PHO3 coding sequences, with different word sizes

  - But, word matches require a perfect match over the whole word length

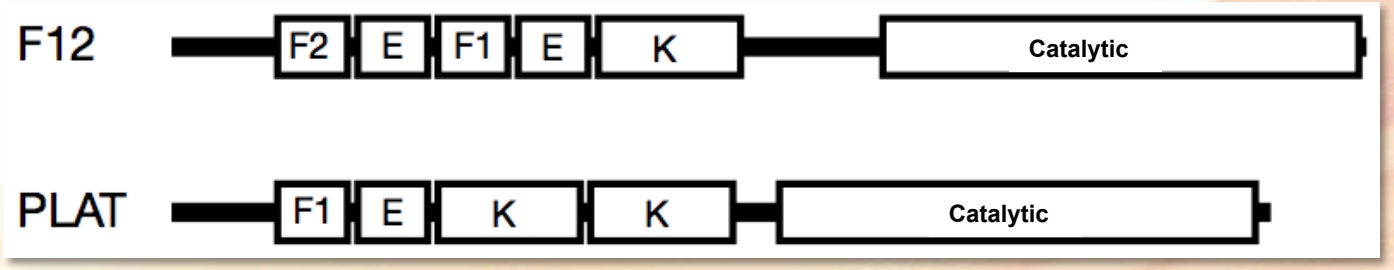
Adopted from Jacques van Helden

So, What Can We Do?

- We can use a window based approach
  - Last time we spoke about stringency
  - But is stringency enough?
    - Think about this from the protein level

Modular Nature of Proteins

- Many proteins do not display global patterns of similarity
  - Instead appear to be mosaics of modular domains
  - Tissue plasminogen activator (PLAT)
    - http://www.ncbi.nlm.nih.gov/nuccore/NM_000930.3
  - Coagulation factor XII (F12)
    - http://www.ncbi.nlm.nih.gov/nuccore/NM_000505.3



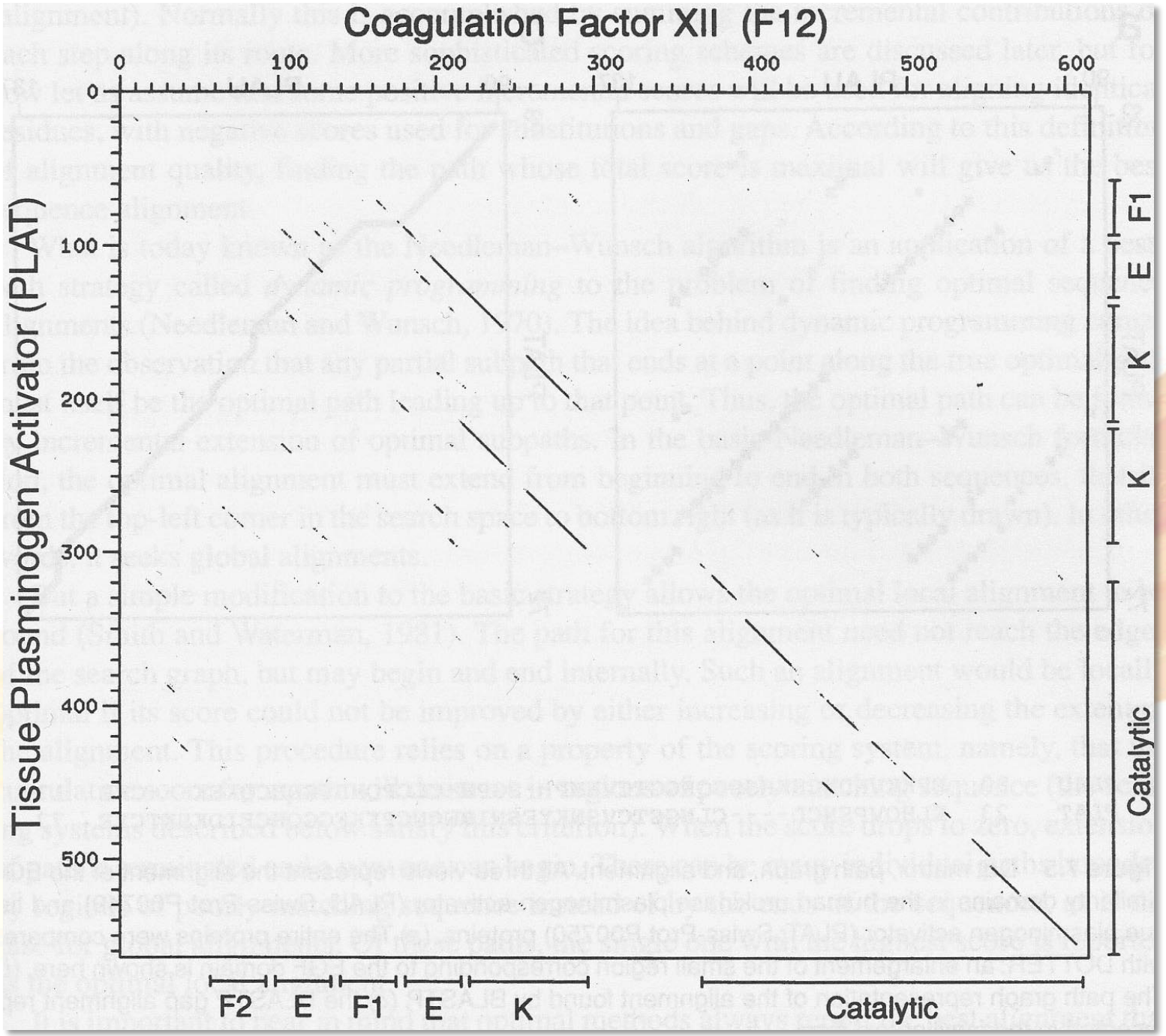
- Besides the catalytic domain - serine protease activity
- Two proteins have different numbers of other structural modules:
  - A two types of fibronectin repeats
  - A domain with similarity to epidermal growth factor
  - Module that is called a ‘‘kringle’’ domain

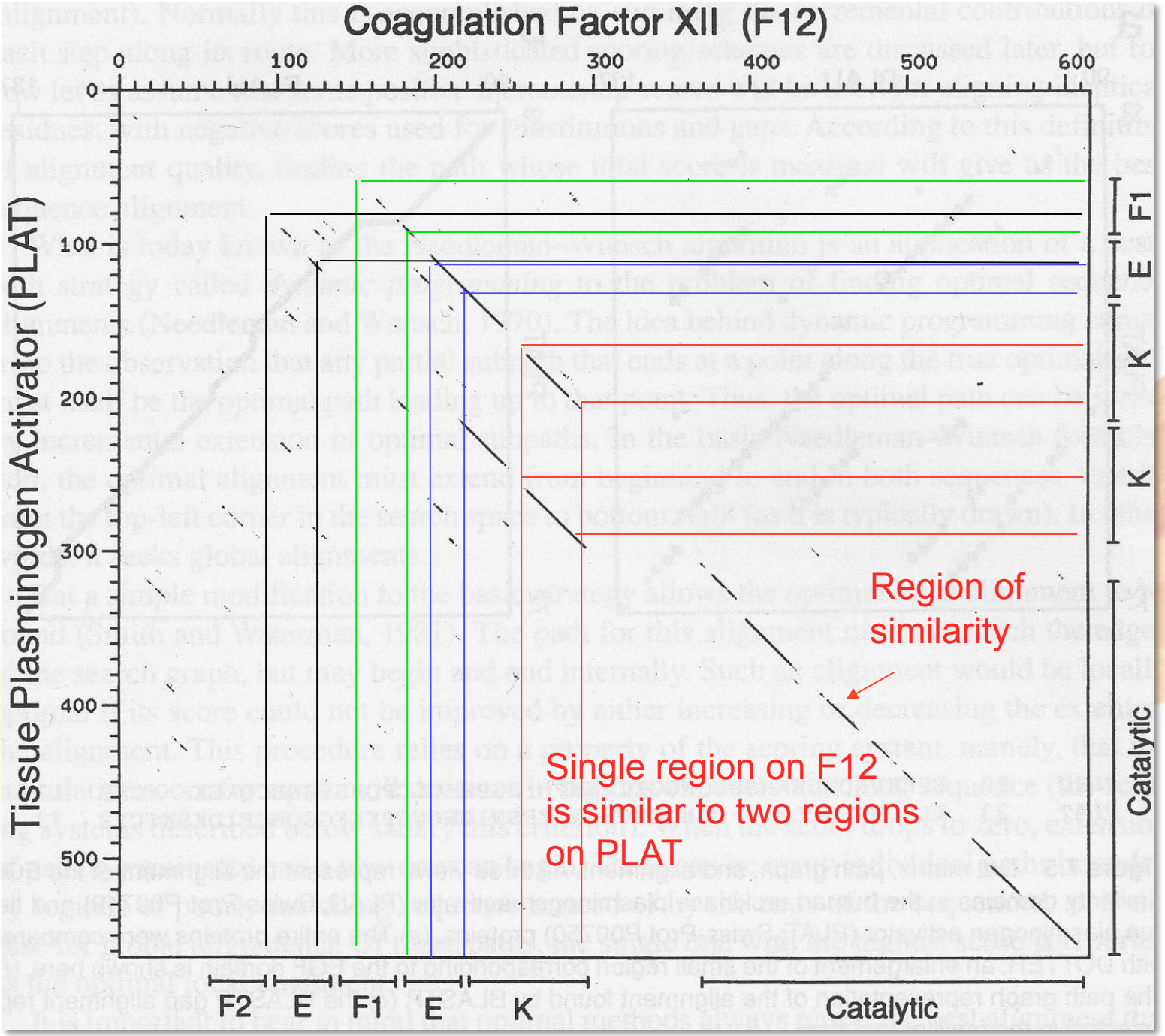
The Dotter Program

- Example of a program consists of three components:
  - The Characters
  - Sliding window
  - A table that gives a score for each amino acid match
    - **A Model (later in lecture)**

- A graph that converts the score to a dot of certain density
- The higher the dot density the higher the score

http://sonnhammer.sbc.su.se/Dotter.html


Coagulation Factor XII (F12)
0
100
200
300
400
500
600
Tissue Plasminogen Activator (PLAT)
F2
E
F1
E
K
Catalytic
F1
E
K
K
Catalytic


Coagulation Factor XII (F12)
0
100
200
300
400
500
600
Tissue Plasminogen Activator (PLAT)
F2
E
F1
E
K
Catalytic
F1
E
K
K
Catalytic
Region of similarity
Single region on F12 is similar to two regions on PLAT

What About that Model?

Well, Let's Discuss the Evolutionary Basis of Sequence Alignments

1st Some Comments/Caveat – Sequence Alignment

- When sequences are aligned - We assume they share a common ancestor
  - Poor assumption
  - **Take the time** to analyze and look for reasons as to **why** they **might** share a **common ancestor**
    - DNA sequences are less informative than protein sequences
    - Protein fold is more conserved than protein sequence
- Remember:
  - Two sequences can **always** be **aligned**
    - We need to determine what is a **meaningful** result
    - **More you know**, the better you will be at making this call
  - Genes or proteins are either **homologs** or they are not
    - There is no such thing as **percent homology**
    - There *is:* **percent identity** or **similarity**

Evolutionary Basis of Sequence Alignment

- **Identity**
  - Is a **measure** 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

Evolutionary Basis of Sequence Alignment

- **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**
    - Similarity is loosely used with DNA
    - But is should not be confused with protein similiarity
  - For instance, isoleucine and leucine are similar

Evolutionary Basis of Sequence Alignment

- **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
  - A **conclusion** drawn from data suggesting that two genes share a common evolutionary history
    - Established on the basis of measured **similarity** or **identity**

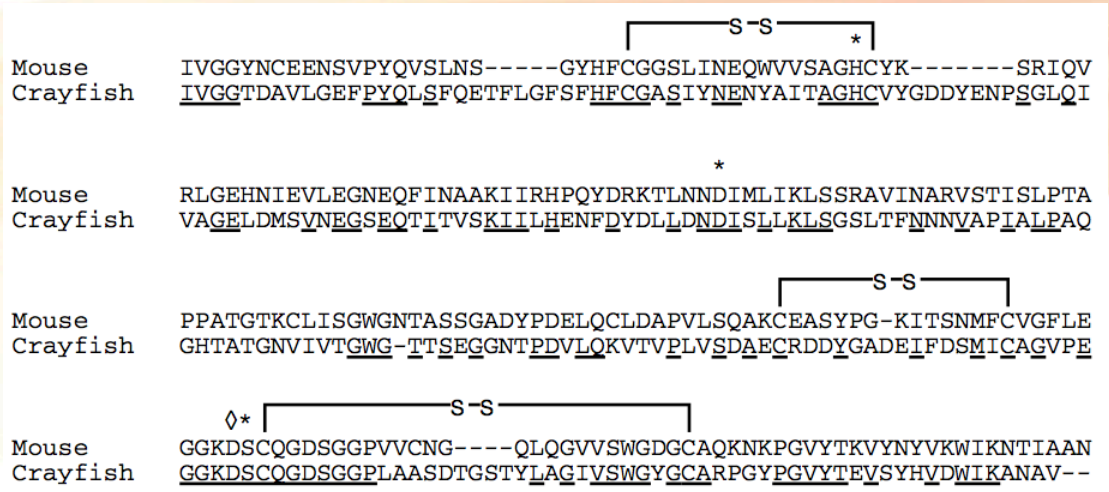
- Let's look at an example

When Is Homology Real?

- As a general rule, in a pairwise alignment:
  - Proteins – alignment over 100 a.a.
    - >25% identical a.a.
      - proteins will have similar folding pattern
      - most likely homologous
    - 18-25% identical- **twilight zone**- tantalizing
    - < 18% identical- cannot determine from alignment
  - DNA - alignment over 100 n.t.
    - >70% identical - most likely homologous
    - 65-70% identical – edge or twilight zone

Two Sequences That Share 41% Identity

- Conserved proteins often share functional importance
- Alignment between trypsin protein of mouse (P07146) and crayfish (P00765)
  - Identical a.a. underlined
  - Three disulfide Bond (-S—S-) are conserved
  - a.a. side chains involved in the charge relay system (*)
  - Active site residues governs substrate specificity (diamond)



>50% mismatches in proteins

Conserved Regions

- In a residue-by-residue alignment:
  - Often becomes apparent that:
    - certain regions of a protein, or perhaps specific amino acids
    - are more highly conserved than others
  - **May** be suggestive as to which residues - critical for maintaining a protein's structure or function
- Why do I say **may**?

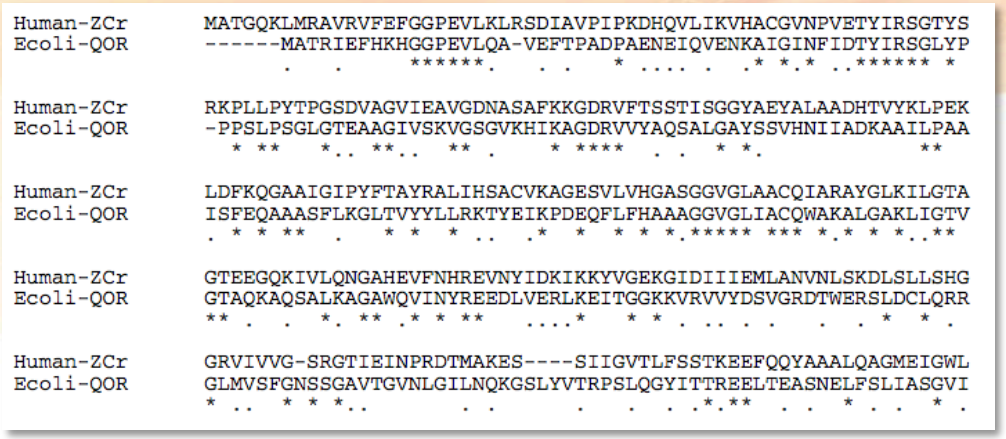
Conserved Regions

- Other positions **that do not** play a significant functional role can be identical for historical reasons
- **Particular caution** should be taken when sequences are **taken from very closely related species**
  - Similarities may be more **reflective of history than of function**
  - e.x. regions of high sequence similarity between mouse and rat homologs may simply be those **that have not had sufficient time to diverge**
- Conserved regions should always:
  - Matched against known residues of functional importance
  - If it's a protein – ***perform structural analysis***
  - Or be tested experimentally for importance

Similar Biological Functions

- Observation of a high degree of sequence similarity b/t two genes or proteins
  - Might infer:
    - share a common evolutionary history
  - From this it might be anticipated:
    - **should also have similar biological functions**
- Should be treated as **hypothetical until tested experimentally**
  - Zeta-crystallin - component of the transparent lens matrix of the vertebrate eye
  - However, on the basis of extended sequence similarity, it can be inferred that its homolog in E. coli is the metabolic enzyme quinone oxidoreductase

Don't always need high degree similarity



Evolutionary Basis of Sequence Alignment

- Identity: Quantity that describes how much two sequences are alike in the strictest terms
- Similarity (Positives): Quantity that relates how much two amino acid sequences are alike
  - DNA does not have similarity
- Homology: a conclusion drawn from data suggesting that two genes share a common evolutionary history
- **Using the above terms we can hypothesize about:**
  - **Structural information**
  - **Functional information**
  - **Evolutionary relationships**

- **In order to do this, we have to come up with kind of score**

Sequence Alignment Scoring

Two Sequences: A and B

- Pairwise Alignment – Visual Inspection
- Lengths are ***m*** and ***n***, respectively
  - The number of matched pairs is ***x***
  - The number of mismatched pairs is ***y***
  - Total number of bases in gaps is ***z***

```
TTAGACGAGTG   (Length = n)
TTGGA GC TG   (Length = m)
```

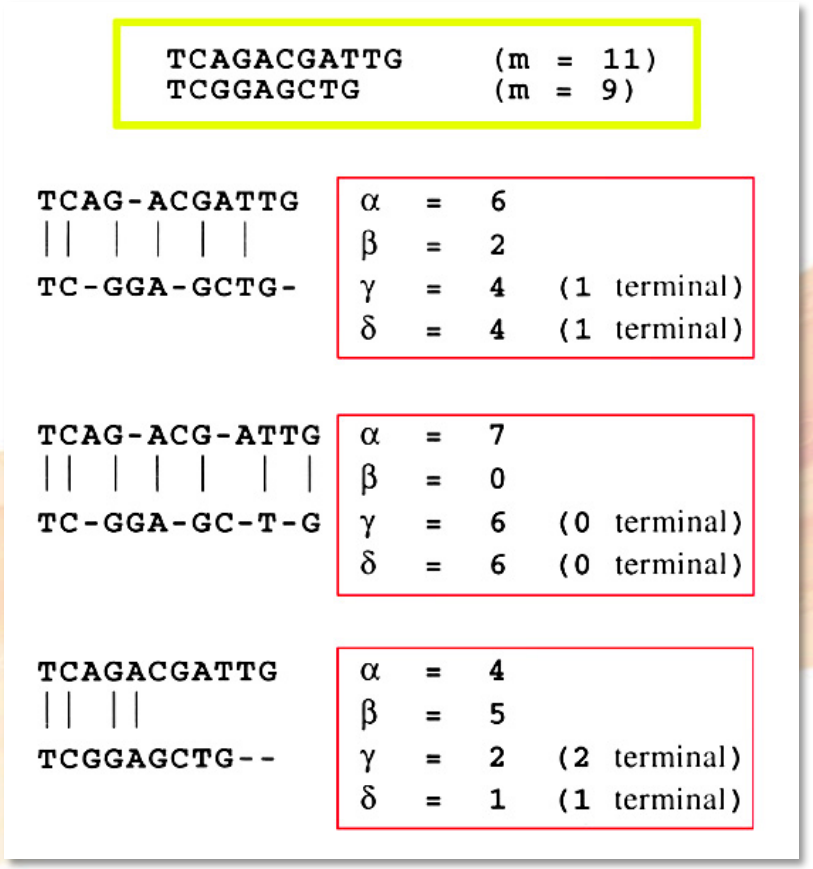
$$n + m = 2(x + y) + z$$

How Do We Find the Best Alignment?

- We'll use different types of methods:
  - Distance and Similarity methods
- The best possible alignment:
  - known as "**Optimal Alignment**"
  - One where #'s of **mismatches** and **gaps** are **minimized** according to **certain criteria**
- Sound easy, right!
- Unfortunately, reducing the number of mismatches results in an increase in the number of gaps, and *vice versa*
- *Really, No… that can be true..?*

Maximizing Matches

α = matches
β = mismatches
γ = nucleotides in gaps
δ = gaps



See, its not so easy!

Terminology – Gap Penalty (or Cost)

- **A factor** (or a set of factors)
  - By which the gap values (numbers and lengths of gaps)
  - Are multiplied to make the gaps equivalent in value to the mismatches
- The gap penalties area based on:
  - **Our assessment of how frequent different types of insertions and deletions occur in evolution**
  - In comparison with the frequency of **occurrence of point substitutions**

Terminology – Mismatch Penalty

- The Mismatch penalty is an assessment of how frequently substitutions occur
  - DNA - simple
  - Proteins –
    - We'll use a model
      - Scoring matrices

The (Similarity) Index (S) between Two Sequences in an Alignment is:

$$S = x - \sum w_k z_k$$

Where x is the number of matches

z_k is the number of gaps of length k

and w_k is a positive number representing the penalty for gaps of length k

Similarity should be confused with DNA Identity/Protein Similarity-Identity

Gaps

- In DNA:
  - Gaps occur with roughly **1/10 the frequency of base substitutions**
  - So they are **common in most alignments**
  - Symbolized by hyphens (– – –) paired with residues:
    - like a mismatch with a blank space
- You can assign a penalty for each gap position
  - This is called a **linear gap penalty**
    - The total penalty is **proportional** to the **gap length**
  - **The problem is, what?**
    - Once you start putting them in, you can get almost anything aligned

The Gap Penalty Has Two Components:

- **gap-opening** & **gap extension** penalties
- Alignment programs usually distinguish between **creating a gap** and **extending a gap**
  - gap opening penalty and a (smaller) gap extension penalty
  - This known as an **affine gap penalty**
- Although substitutions have **a lot of theory** behind them, gap penalties are generally determined by **heuristic** means
  - **Heuristic**
    - A method or value determined by trial-and-error experiments, without a strong guiding theory
  - In this case:
    - gap opening and extension penalties are the result of trying many possibilities and seeing which ones returns the **most biologically relevant results**

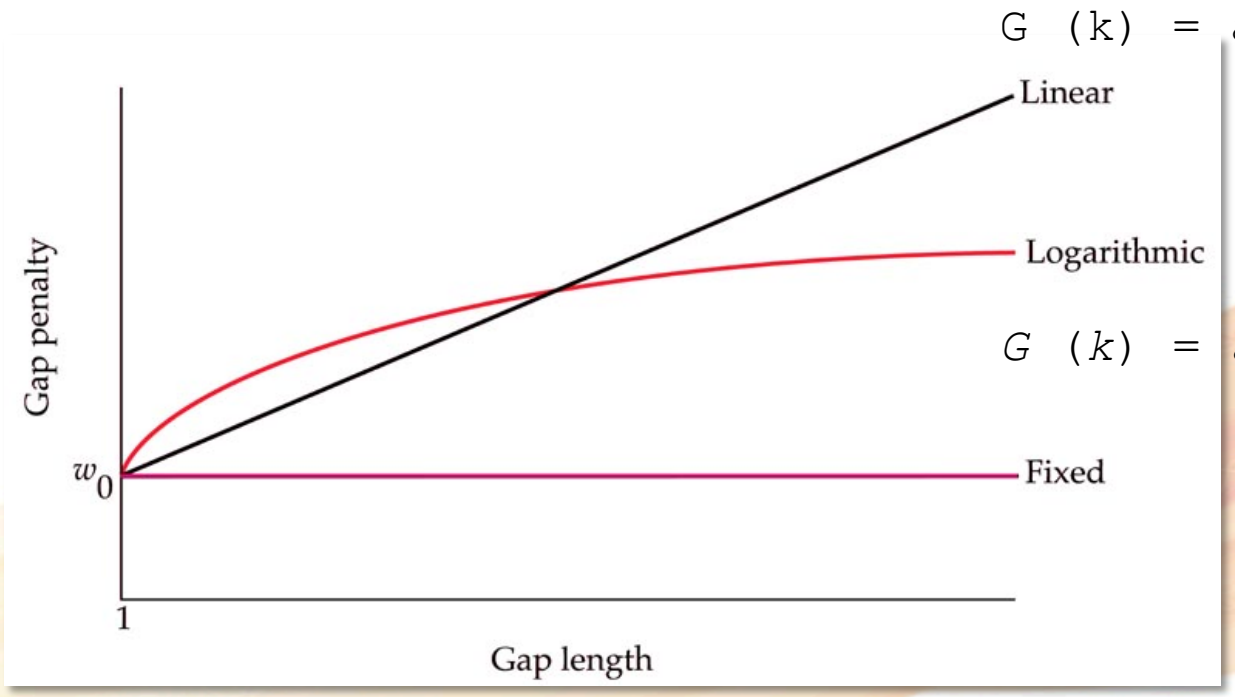
Different Gap Penalties

- Comparing 2 distantly related sequences w/ different gap penalties:
- Top has **fewer gaps** and **fewer matches**
- Bottom has **more matches overall**, but **lots of little gaps**
  - Matches near the C-terminal are absurd
- **Look at the short segment after the first gap in the lower sequence: gained 3 identities**

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



Ok, Gaps covered, but what about identity/similarity?

Simple Measures of Sequence (Similarity)

Quantitative Measures of Sequence Similarity and Differences

- Given two character strings, two measure of the distance between them are, hamming and Levenshtein distance:
- **Hamming** distance
  - Defined between two string of equal length
  - Number of position with mismatching characters

```
agtc
cgta    Hamming distance =2
```

Quantitative Measures of Sequence Similarity and Differences

- **The Levenshtein** (edit) distance
  - Defined between two string of not necessarily equal length
  - Minimal # of edit operations required to change one string into the other
  - Edit operation is a deletion, insertion or alteration of a since character in either sequence
  - A given sequence of edit operations induces a unique alignment, but not vice versa

```
ag-tcc
cgctca
```

Levenshtein distance = 3

  - These are very crude methods for measuring distance, why?

More Sophisticated Measures of Sequence (Similarity) Involve Scoring Schemes

Simple Scoring Schemes

- Must account for:
  - Residue substitutions
  - Insertion or deletions
    - indels have scores that depend on their lengths
- **Hamming** and **Levenshtein** distance measures the dissimilarity of two sequences:
  - Similar sequences give small distances
  - Dissimilar sequences give large distances
- More sophisticated measures should take a model into account

Sophisticated Scoring Schemes

- It's common in molecular biology to define scores as measure of sequence similarity
- Similar sequences give high scores - dissimilar sequence give low scores
- These are equivalent formulations
  - Algorithms for **optimal alignment** seek to either minimize a dissimilarity measure or to maximize a scoring function
  - Such as the **Similarity Index**
- **How might we score?**

Identity Matrix for DNA

- For DNA
- Common use a simple scheme for substitutions:
  - +1 for a match, 1 for a mismatch
  - Or, more complicated scheme based on higher frequency of transition mutations than transversion mutations
- How could we make this more complex?

A	1			
T	0	1		
G	0	0	1	
C	0	0	0	1
	A	T	G	C

Identity Matrix for DNA

- Transition mutations
  - purine->purine
  - pyrimidine->pyrimidine
    - a->g and t->c
  - Transversion mutations
    - purine->pyrimidine;
      - (a, g)->(t, c))
- What about protein?

A	3			
T	-2	3		
G	-1	-2	3	
C	-2	-1	-2	3
	A	T	G	C

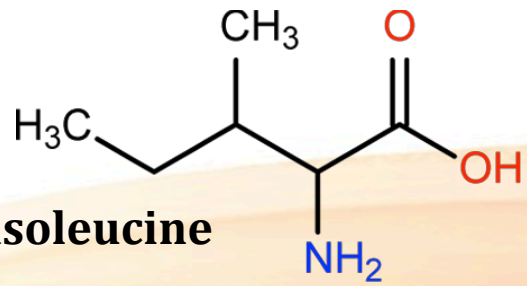
Identity Matrix for Proteins

- What problems exists with this simplified matrix for proteins?
- Certain amino acids can substitute easily for one another in related proteins
  - Because of their similar physicochemical properties
  - Examples of these "conservative substitutions"
    - Isoleucine for valine (both small and hydrophobic)
    - Serine for threonine (both polar)

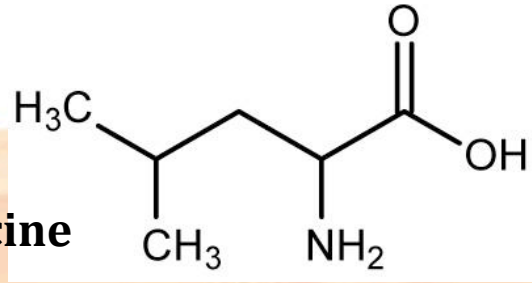
S	1			
T	0	1		
I	0	0	1	
V	0	0	0	1
	S	T	I	V

Similarity Matrix

- Not easy to give a score for amino acids that are somewhat similar
- For example:
  - A mismatched pair consisting of Leu & Ile
    - Biochemically similar to each other

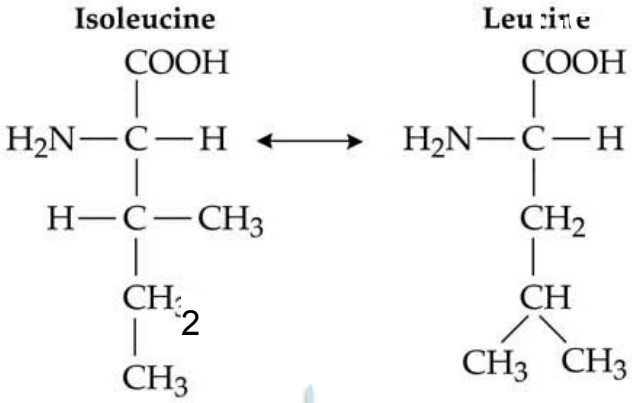


isoleucine

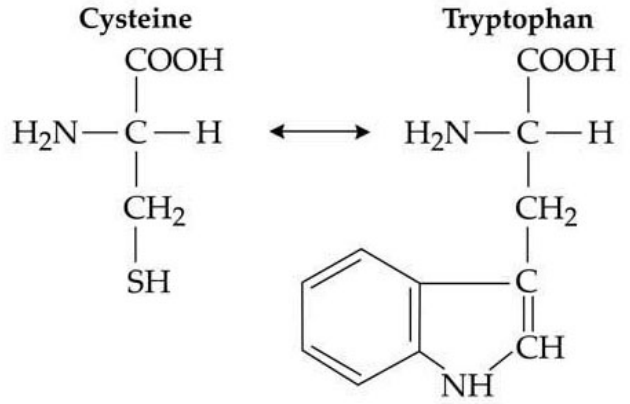


leucine

  - May be given a lesser penalty than a mismatched pair consisting of Cys & Trp
    - Very dissimilar from each other



So what can we do?



Smaller penalty than…

- Ask the question "**What's needed**"?
  - **Identical** a.a. should be given greater value than substitutions
  - **Conservative substitutions** should also be greater than non-conservative changes
  - **Different sets of values may be desired:**
    - Comparing very similar sequences (e.g., a mouse gene and its rat homolog)
    - As opposed to highly divergent sequences (e.g., mouse and yeast genes)

Consider The Following

```
(a)   AIWQH
      :  ::
      AL-QH

(b)   AIWQH
      :  ::
      A-LQH
```

- These alignments seem quite similar
  - Both contain an "indel" and 1 substitution
  - However, think of the letters as a.a. rather than elements of strings
    - Alignment **a** more accurate than **b**

- Why is it better?
  - Isoleucine and Leucine are similar sidechains
  - Tryptophan has very different structure
  - Physico-chemical measure:
    - Leucine simply substitutes for Isoleucine more frequently
    - More likely that a mutation changed **I into L (or L into I)**, and W was lost, than W changed into L and I was lost
      - **I** to **L** most likely would not affect function as much as **W** to **L** could potentially

Scoring Matrices – Alignment Perspective

- Also known **as substitution matrix**
- Biological sequences have evolved throughout time, so we should model it!
- To quantify the similarity in an alignment
  - Alignment score is the sum of the matrix's entries for each aligned a.a. pair
  - For gaps, a special gap score is necessary
- The optimal alignment is the one which **maximized the alignment score**

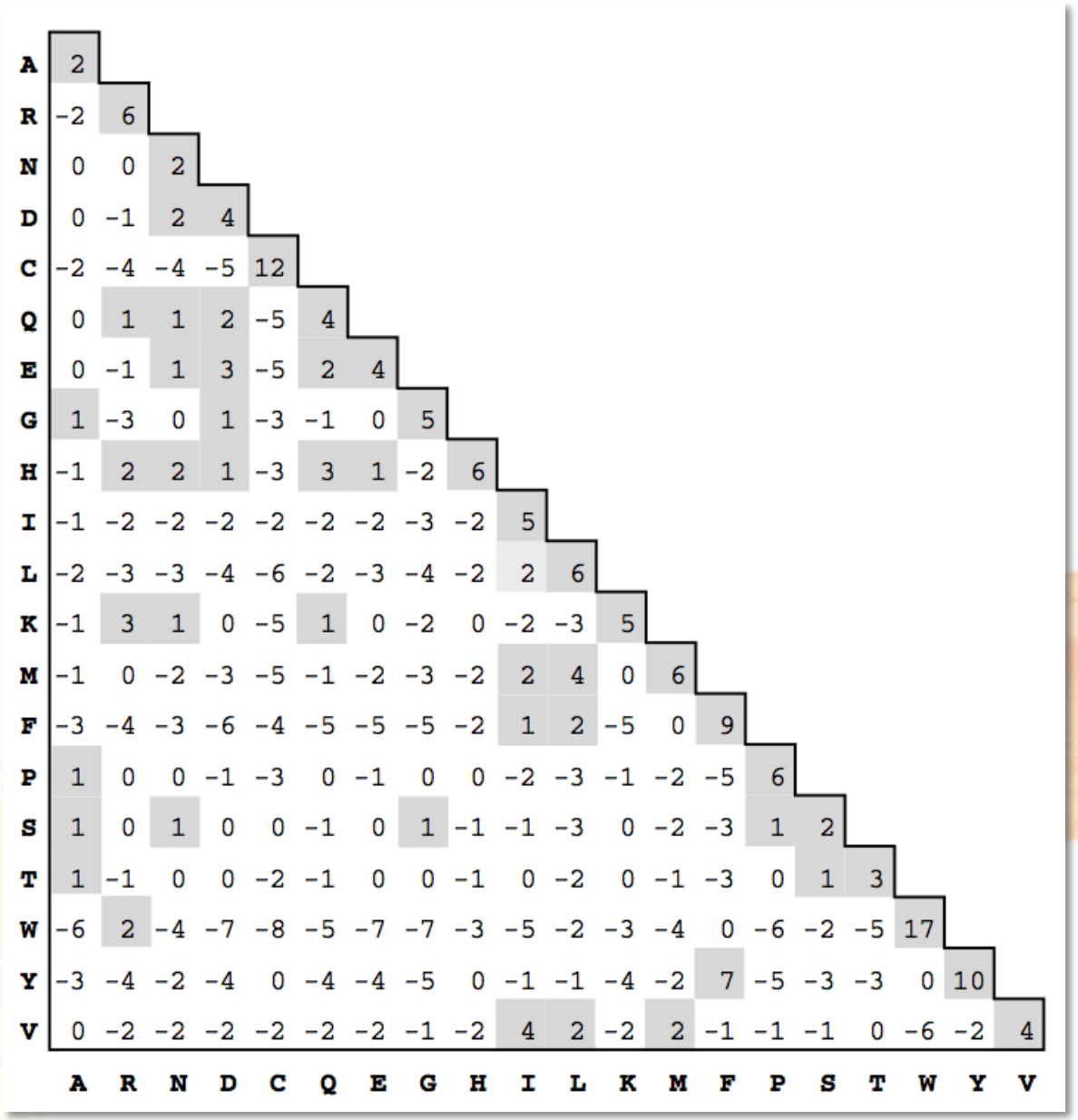
Scoring Matrices – Evolutionary Basis

- Evolution has shown that not all changes to a biological sequence are equally likely
  - Certain amino acid substitutions:
    - Happen often
    - Others are very rare
      - W -> D
- Based on the evolution of proteins:
  - Became apparent changes could be modeled by a scoring matrix
  - So, that's what was done

Scoring Matrices

- The "previous" considerations can be dealt with in a flexible manner
  - Through the use of a substitution matrix
  - The score for any pair of amino acids can be easily looked up
- Matrix lists the substitution scores of every single a.a.
  - Score for an aligned a.a. is found at the intersection of the row and column
  - The diagonal show the scores for a.a. which have not changed
  - Most changes have a negative score
- The two most widely used:
  - **PAM [Dayhoff 1978]**
  - **BLOSUM [Henikoff 1992]**

The PAM 250 Scoring Matrix



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Scoring Matrix Inherently Based On:

TABLE 1.3

General Characteristics of Alpha Amino Acids

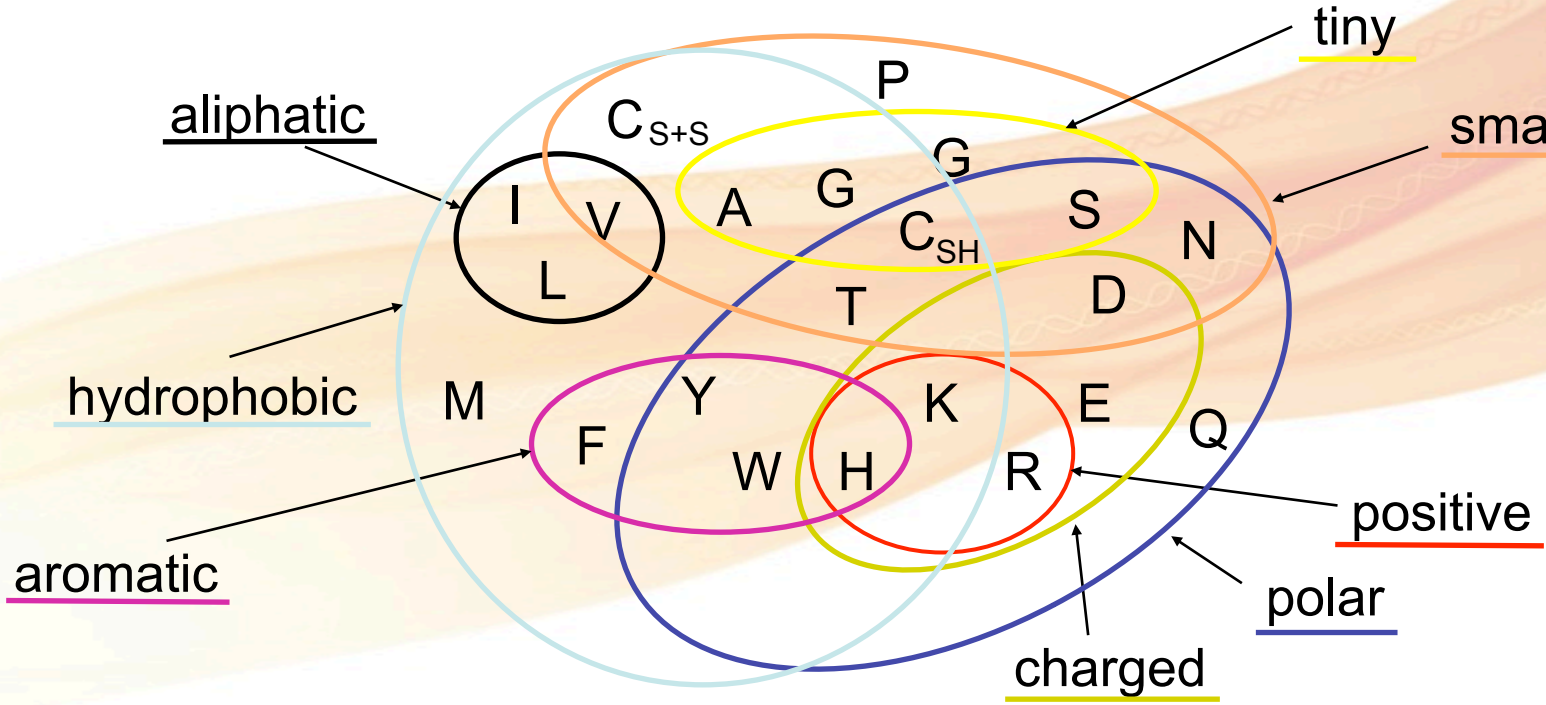
Residue	Residue (One Letter Code)	Hydrophobic	Aromatic	Aliphatic	Small	Polar	Charged
Alanine	A	√			√		
Arginine	R					√	√
Asparagine	N				√	√	
Aspartate	D				√	√	√
Cysteine	C	√			√	√	
Glutamate	E					√	√
Glutamine	Q					√	
Glycine	G	√			√	√	
Histidine	H	√	√			√	√
Isoleucine	I	√		√			
Leucine	L	√		√			
Lysine	K	√				√	√
Methionine	M	√					
Phenylalanine	F	√	√				
Proline	P				√		
Serine	S				√	√	
Threonine	T				√	√	
Tryptophan	W	√	√			√	
Tyrosine	Y	√	√			√	
Valine	V	√		√			

√ is representative of the character or partial character of the residues listed

Not a bad representation, is there a better one?

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

PAM

- PAM matrix published 1978
- Built using 1,572 changes in 71 groups of closely related proteins (85% similiar
  - "**P**oint **A**ccepted **M**utations per 10^8 years" - model of evolution
    - **Accepted mutations by natural selections**
  - Every change was tabulated and entered in a matrix enumerating all possible a.a. changes
  - This information was combined with the relative mutability of the a.a.:
    - -> "Mutation probability matrix"
    - Elements of the matrix:
      - **Probability a.a. in one column**, replaced by a.a. in some row after a given evolutionary interval
- The model of evolution that Dayhoff used **assumed that proteins diverged as a result of accumulated, uncorrelated mutations**

Why PAM1?

- As sequences diverge, **mutations accumulate**
- To measure the relative probability of any particular substitution:
  - For instance Serine->Threonine
    - Count # of Serine->Threonine changes in pairs of **aligned homologous sequences**
    - Use the **relative frequencies** of such changes to **form a scoring matrix for substitutions**
    - A **likely** change should **score higher** than a **rare one**
- But, what if there have been multiple substitutions at certain sites?
  - **This will bias the statistics**
  - **Avoided** by **restricting** our samples to **sequences** that are **sufficiently similar** so that we can assume that **no position** has changed **more than once**
- PAM1 matrix means probablilty of each amino acid changing into another is ~ 1% and probability of not changing is ~99%

PAM

- To derive a mutational probability matrix for a protein sequence that has undergone **N percent accepted mutations**, a **PAM-N matrix**
  - The **PAM-1 matrix** is **multiplied** by **itself N times**
  - Used to produce a matrix appropriate for **more widely divergent sequences**
- Each term now gives the probability of replacement, **j** to **i** per occurrence of residue **j**
- PAM matrix - probability any given a.a will mutate in a given time interval
  - For example
    - PAM1 gives that one amino acid out of 100 will mutate in a given time interval
    - PAM250 matrix 250 mutations in 100 amino acids

The Science Behind PAM

- One PAM is a unit of evolutionary divergence in which 1% of the amino acids have been changed
  - This does not imply that after 100 PAMs every amino acid will be different
  - Why?
    - Some positions may change several times, perhaps even reverting to the original amino acid
    - Whereas others may not change at all

The Model Behind PAM

- If there were no selection for fitness:
  - Frequencies of each possible substitution would be primarily influenced by the **overall frequencies** of different a.a (background frequencies)
  - However, in related proteins - observed substitution frequencies (called the ***target frequencies)*** are biased toward those that **do not seriously disrupt the protein's function**
  - These are point mutations that have been "**accepted**" during evolution

Some Are Common, Some Are Not

Original amino acid

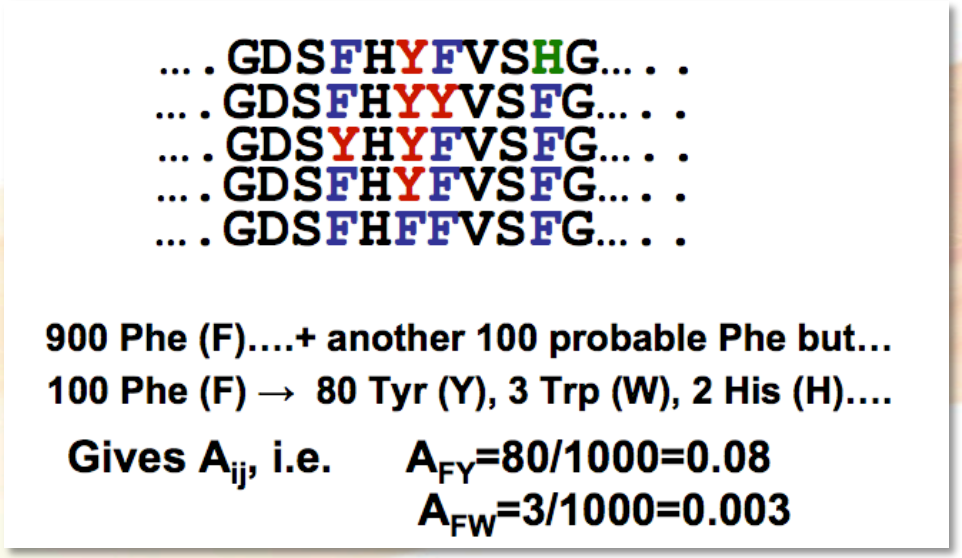
Replacement amino acid

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

Number of accepted point mutations multiplied by 10 in 1572 cases of amino acid substitutions from closely related protein sequences

1st Step - Calculate the Pair Exchange Frequencies

- Calculate the Pair Exchange Frequencies
  - A_{ij} Number of times amino acid *j* is replaced by amino acid *i* divided by all comparisons



Dan Burns

Do This for All 20 Amino Acids

	C	D	E	F	G	H	I	.....
C	A_{CC}	A_{CD}	A_{CE}					
D	A_{DC}							
E	A_{EC}							
F								
G								

Gives A_{ij} = pair exchange frequency

2nd Step - Calculate Frequencies of Occurrence

Interesting, not all a.a. are represented equally?

$$f_i = \frac{\text{Observations of amino acid i}}{\text{Observations of all amino acids}}$$

$$f_i = \frac{\text{Observations of amino acid i}}{\sum_{i=1}^{20} aa_i}$$

$$\text{i.e. } f_{Phe} = \frac{1000}{40{,}000} = 0.04$$

Amino acid frequencies:

	1978	1991
L	0.085	0.091
A	0.087	0.077
G	0.089	0.074
S	0.070	0.069
V	0.065	0.066
E	0.050	0.062
T	0.058	0.059
K	0.081	0.059
I	0.037	0.053
D	0.047	0.052
R	0.041	0.051
P	0.051	0.051
N	0.040	0.043
Q	0.038	0.041
F	0.040	0.040
Y	0.030	0.032
M	0.015	0.024
H	0.034	0.023
C	0.033	0.020
W	0.010	0.014

Bottom of page
http://www.ebi.ac.uk/uniprot/TrEMBLstats/

Dan Burns

3rd Step - Calculate the Relative Mutability

Chance on average that a given a.a. will mutate

$$m_i = f_i \times \text{(number of times amino acid i is observed to mutate)}$$

example

$$m_{Phe} = f_{Phe} \times \text{(number of times Phe is observed to mutate)}$$

$$m_{Phe} = 0.04 \times (100) = 4$$

Dan Burns

A Word About Relative Mutability

- How often each a.a likely change over a short period of time
  - Remember analysis was conducted with closely related proteins

$$m_i = f_i \times \text{(number of times amino acid i is observed to mutate)}$$

- Don't need this for calculation, but interesting
- Set value of alanine arbitrarily to 100
- Why are some a.a. more mutable than others?

Relative mutabilities of amino acids:

	1978	1991
A	100	100
C	20	44
D	106	86
E	102	77
F	41	51
G	49	50
H	66	91
I	96	103
K	56	72
L	40	54
M	94	93
N	134	104
P	56	58
Q	93	84
R	65	83
S	120	117
T	97	107
V	74	98
W	18	25
Y	41	50

4th Step – Mutation Probability Matrix

Chance that a given a.a. j will replace a.a. i

$$M_{ij} = m_i \times \frac{A_{ij}}{\sum_{i=1}^{20} A_{ij}}$$

example

$$M_{FY} = m_{Phe} \times \frac{\text{(number of times Phe} \rightarrow \text{Tyr)}}{\text{(number of times Phe} \rightarrow \text{mutates)}}$$

$$M_{FY} = 4 \times \frac{(0.08)}{(0.10)} = 3.2$$

What is it: The probability that an a.a. in row *i* of the matrix will replace the a.a. in column *j*:

How calculation is done: the mutability of a.a *j*, multiplied by the pair exchange frequency for *ij* divided by the sum of all pair exchange frequencies for a.a. *i*

Dan Burns

5th Step – Calculate Evolutionary Distance Scale

- So that only 1/100 amino acids change (PAM1)
- M_{ii} reflects amino acid conservation

$$M_{ii} \alpha \quad 1- \sum_{i=1}^{20} M_{ij}$$

example

$$M_{FF} \alpha \quad 1- \text{(frequency of Phe mutations)}$$

- Then Scale (Next Slide)

Use a scale factor **λ**

- so that M_{ii} is ~ 0.99 i.e. chance of it mutating is ~ 1%
- i.e. This Defines a PAM1 matrix

$$M_{ii} = 1 - \lambda \sum_{i=1}^{20} M_{ij} = \sim 0.99$$

- **λ** is our evolutionary scale factor
- For any particular mutation probability
- λM_{ij} reflects the normalized measure of how likely amino acid **j** will replace amino acid **i** over 1 PAM

Real PAM1 Value

Amino Acid Change	PAM 1 Probability Score
F→A	0.0002
F→R	0.0001
F→N	0.0001
F→D	0.0000
F→C	0.0000
F→Q	0.0000
F→E	0.0000
F→G	0.0001
F→H	0.0002
F→I	0.0007
F→L	0.0013
F→K	0.0000
F→M	0.0001
F→F	0.9946
F→P	0.0001
F→S	0.0003
F→T	0.0001
F→W	0.0001
F→Y	0.0021
F→V	0.0001

SUM = 1.0

Note – this is really just 1 column in a much bigger probability matrix

	..... E	F	G .....
A		0.0002	
C		0.0000	
D		0.0000	
E		0.0000	
F		0.9946	
G		0.0001	

How To Derive Additional PAM Matrices

- Assume mutations at each site are independent of previous mutations (Markov)
- Therefore, calculate changes predicted for more distantly related proteins that have undergone:
  - N mutations/100 amino acid
  - By multiplying the PAM1 matrix against itself N times

Example: PAM2 matrix:

	aa1	aa2	aa3
aa1	a	b	c
aa2	d	e	f
aa3	g	h	i

X

	aa1	aa2	aa3
aa1	a	b	c
aa2	d	e	f
aa3	g	h	i

	aa1	aa2	aa3
aa1	A	B	C
aa2	D	E	F
aa3	G	H	I

$A=a^2+bd+cg+\ldots$

$B=ab+be+ch+\ldots$

$C=ac+bf+ci+\ldots$

$D=da+ed+fg+\ldots$

Dan Burns

PAM 250

- Multiply PAM1 matrix by itself 250 times!

(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)
(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)(PAM1)

These are the M_{ij} values

Amino Acid Change	PAM 1 Score	PAM 250 Score
F→A	0.0002	0.04
F→R	0.0001	0.01
F→N	0.0001	0.02
F→D	0.0000	0.01
F→C	0.0000	0.01
F→Q	0.0000	0.01
F→E	0.0000	0.01
F→G	0.0001	0.03
F→H	0.0002	0.02
F→I	0.0007	0.05
F→L	0.0013	0.13
F→K	0.0000	0.02
F→M	0.0001	0.02
F→F	0.9946	0.32
F→P	0.0001	0.02
F→S	0.0003	0.03
F→T	0.0001	0.03
F→W	0.0001	0.01
F→Y	0.0021	0.15
F→V	0.0001	0.05

SUM = 1.0

These are the M_{ij} values! i.e. the chance that one amino acid will replace another at 250 PAMs in two proteins that are evolutionarily related to each other!

PAM 250 matrix – 250% Expected Change

- Sequences still ~ 15-30 % similar, i.e.:
  - Phe will match Phe ~ 32% of the time
  - Ala will match Ala ~ 13% of the time

6^{th} Step – Calculate Relatedness Odds Matrix

- Chance that two amino acids in a sequence alignment come from related proteins via evolution **versus** the chance that they are from two unrelated proteins aligned by chance
- Take the elements of M_{ij} and divide each term by the frequency of the replacement residue
  - Relative odds of evolution rather than chance:
  - $\mathbf{R_{ij} = M_{ij} / f_i}$

M_{ij} = prob. that j replaces i in related proteins

-vs –

P_i^{ran} = prob. that j replaces i because the proteins are completely unrelated…i.e. i was there by chance

Now, $P_i^{ran} = f_i$, the frequency of occurrence of amino acid i

Where Do the Numbers in the PAM250 Matrix Come From?

- **Then take the log!!!**
  - **Calculate log odds** (relatedness odds) and multiply by 10 to clear fractional values

Example: Phe →Tyr (which must = Tyr →Phe)

$$R_{ij} = \frac{M_{ij}}{f_i}$$

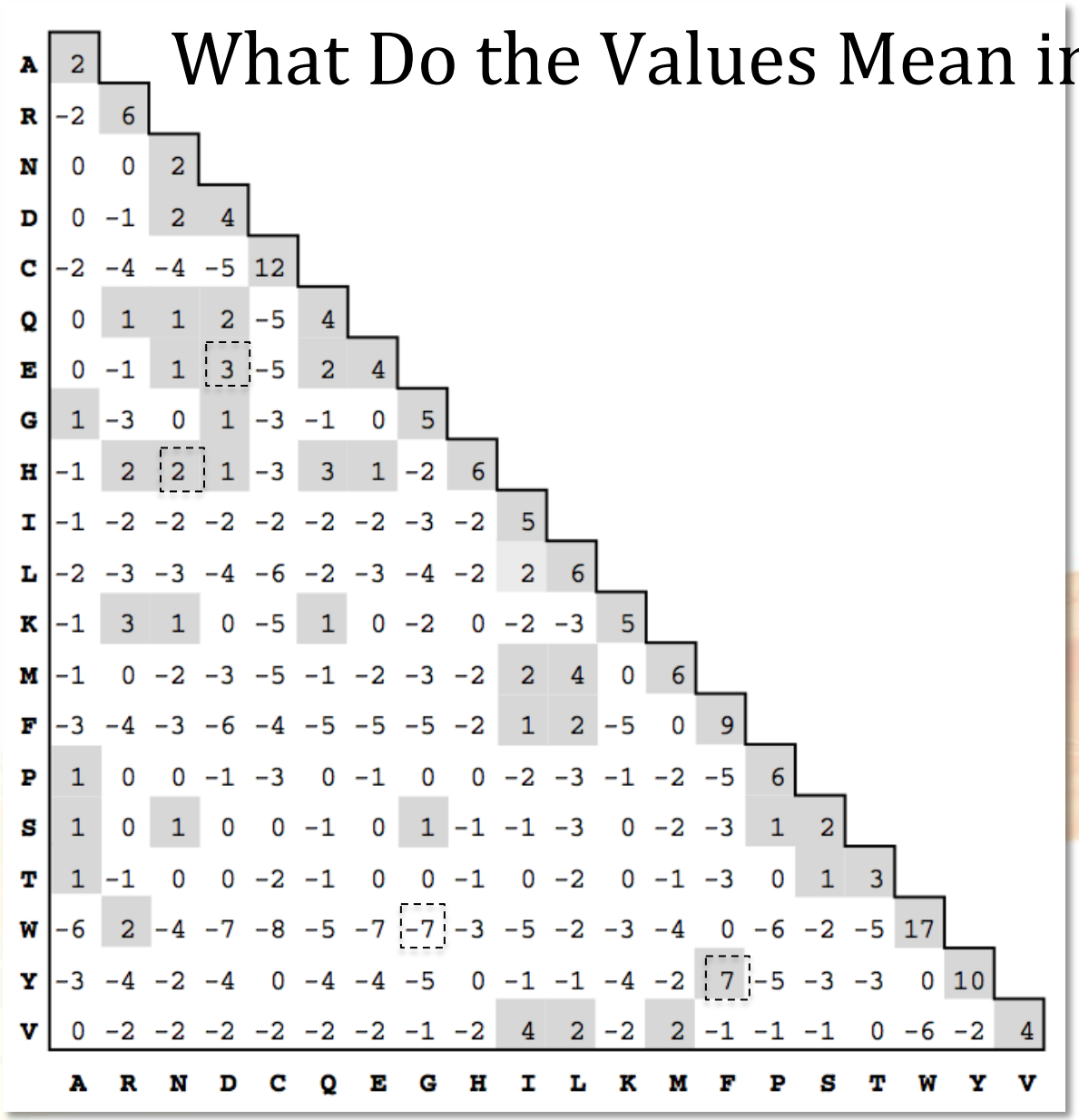
$M_{FY} = 0.15$
$f_{Phe}=0.04$
So $R_{FY}=0.15/0.04 = 3.75$
$Log_{10}R_{FY}=Log_{10}(3.75)=0.57$
$10 \times 0.57 = 5.7$

Likewise
$M_{YF} = 0.20$
$f_{Tyr}=0.03$
So $R_{YF}= 6.7$
$Log_{10}(6.7)=0.83$
$10 \times 0.57 = 8.3$

So average = (5.7+8.3)/2 = 7….the number in the PAM250 table!

Dan Burns

What Do the Values Mean in PAM250?



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

What The Scores for Substitutions Really Mean

- Are a Direct Comparison to Random Substitutions
  - Really, How ?
- **His ↔ Asn** Score (S) = 2 → S = 2/10 = 0.2
  - Divided by 10 b/c of the scaling that occurred
  - 0.2 is log10 of the relative expectation value of the mutation
- $10^{0.2}$ = **1.6**
  - The mutation His ↔ Asn would be expected to occur **1.6** more frequently than random
  - The largest number of observed changes (83) was between Asp (D) and Glu (E) → log odds score **+3**
- But many changes were not observed :
  Gly (G) and Trp (W) → log odds score **-7**

The Scores for Substitutions

- How Do You Understand the Score for Mutation:
  Ala ↔ Ala is Equal 2?
  - See, the Dayhoff PAM250 matrix
- Ala substitutes to Ala:
  - Would be expected that Ala does not mutate 1.6 more frequently than any random Ala substitution

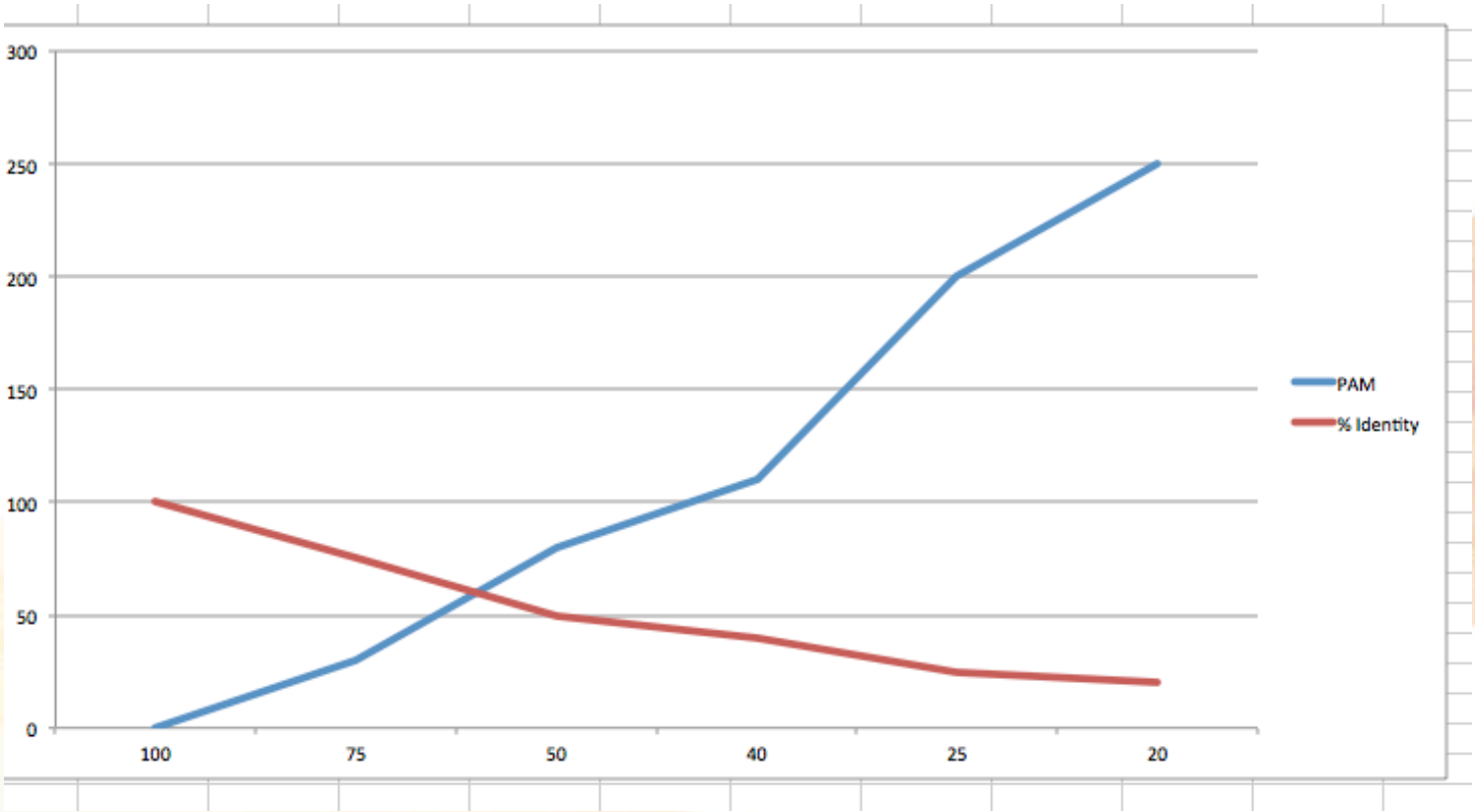
- Take a minute, sit back and appreciate this!
  - Dayhoff
    - First to explicitly use a log-odds approach
    - In which the substitution scores in the matrix are proportional to the natural log of the ratio of target frequencies to background frequencies

Which PAM Matrix to Use

- The PAM250 level, corresponding to ~20% overall sequence identity
  - ~lowest sequence similarity for which we can hope to produce a correct alignment by **sequence analysis alone**
  - Appropriate level to choose for **TYPICAL USAGE**
- The occurrence of reversions, either directly or via one or more other changes
  - Produces an apparent slowdown in mutation rates as sequences progressively diverge
    - Next slide
  - The relation between PAM Matrix and % sequence identity is:

PAM	**0**	**30**	**80**	**110**	**200**	**250**
% Identity	100	75	50	40	25	20

Relation between PAM Matrix and % Sequence Identity



Imperfect Assumptions

- PAM Model Built is built upon assumptions that have faults
- The model:
  - Replacement of any site depends on the amino acid at that site and the probability given by the table
    - This is imperfect representation of evolution
    - Replacement is not equally probable over an entire sequence
      - Think….. locally conserved regions
- Each amino acid position is equally mutable is incorrect
  - Sites vary considerably in their degree of mutability
- Many sequences depart from average amino acid composition
- Errors in PAM1 are magnified in extrapolation of PAM250

Let's Take 5…

BLOSUM

- BLOcks SUbtitution Matrix
- 1992, 14 years after the PAM matrices were published
- Henikoff wanted to model more divergent proteins
  - Locally aligned sequence where none of the aligned sequences share less then 62% identity
  - No gaps were used, just BLOCKS
- No evolutionary model in this case
  - It is thought to be more advantageous to have data generated by direct observation
    - Rather than extrapolation

BLOSUM62

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

BLOSUM

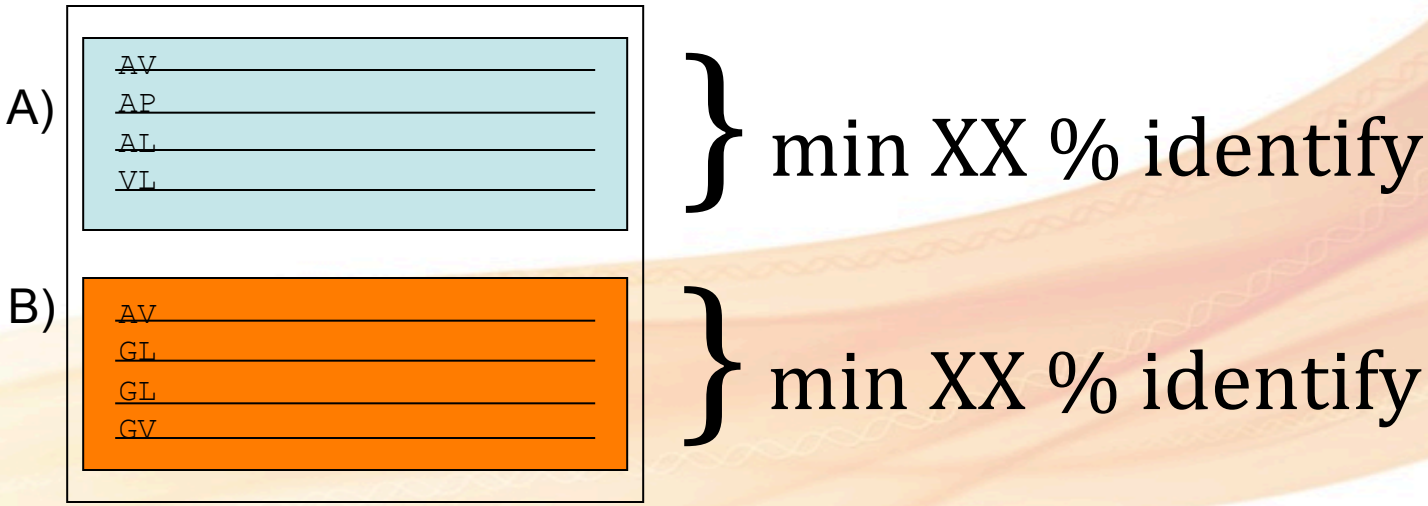
- Uses some of the basic concepts of Dayhoff model
- **Starting data is conserved blocks**
  - Aligned in order to represent the distant relationships more **explicitly**
    - More common substitutions should represent a closer relationships between two a.a.
    - Conversely, radical substitutions should be favored less
- Transition frequencies observed directly by identifying blocks that are a least:
  - 60% identical patterns - blosum60
  - 80% identical patterns - blosum80
- **PAM designed to track evolutionary origins**
- **BLOSUM designed find conserved domains**
  - "blocks" of sequence fragments represent structurally conserved regions

Why is BLOSUM Better?

- To generate matrices
  - Many sequences from aligned families were used
  - Any potential bias introduced by counting multiple contributions from identical residues pairs is removed by:
    - Clustering sequence segments on the basis of minimum percentage
    - Clusters are treated as single sequences
      - BLOSUM62:
        - » Sequences having at least 62% identity are merged into a single sequence
  - **Substitution frequencies are more heavily influenced by sequences that are more divergent than this cutoff**
- Derived from data (observed alignments) representing highly conserved sequence segments from divergent proteins
  - Opposed data based very similar sequences

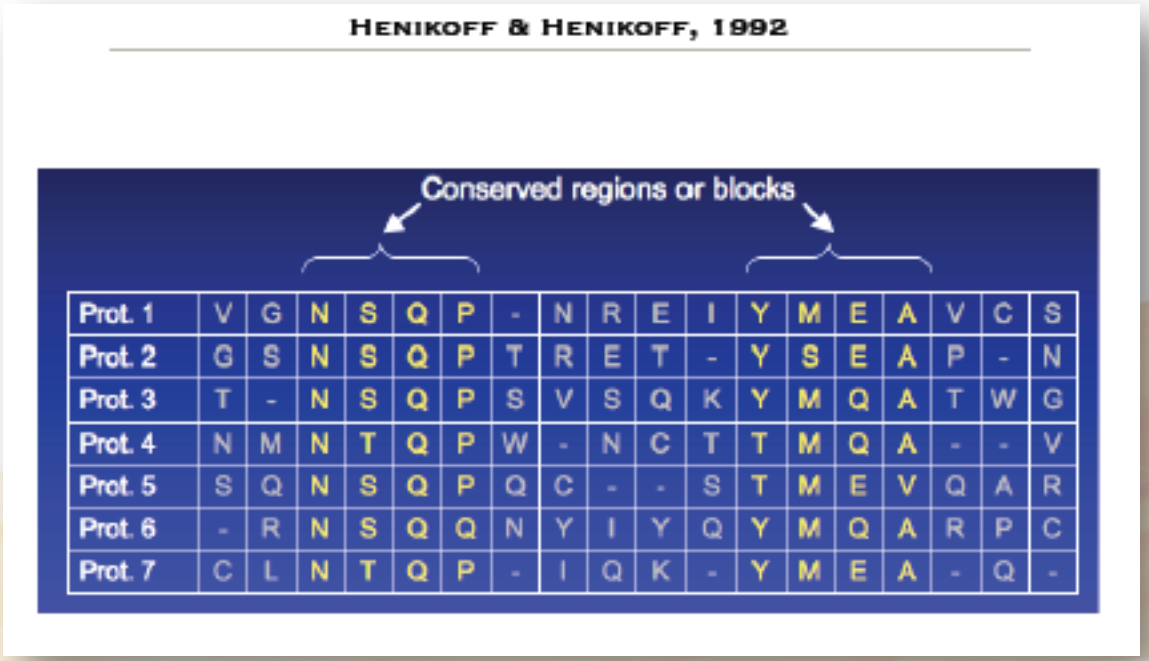
Clustering?

- Cluster sequence Blocks at XX% identity
- Do statistics only across clusters



- Normalize statistics according to cluster size

Construction of BLOSUM Scoring Matrices



BLOSUM

$$s(a,b) = \frac{1}{\lambda} ln\left(\frac{P_{ab}}{f_a f_b}\right)$$

- p_{ab}
  - Probability of two amino acids *a* and *b* replacing each other in a homologous sequence
- f_a and f_b (**Also seen as Q_i, Q_j**)
  - The background probabilities of finding the amino acids *a* and *b* in any protein sequence at random
- λ
  - Is a scaling factor, set such that the matrix contains easily computable integer values
- The logarithm for the ratio of the likelihood of two amino acids appearing with a biological sense and the likelihood of the same amino acids appearing by chance

How Do We Find the Random Mutation Rate?

- The random mutation rate
- Compute the overall occurrence of an amino acid in a protein database

UniProtKB/TrEMBL PROTEIN DATABASE RELEASE 40.10 STATISTICS

2.1 Composition in percent for the complete database

Ala (A) 8.53	Gln (Q) 3.89	Leu (L) 9.81	Ser (S) 6.75
Arg (R) 5.47	Glu (E) 6.15	Lys (K) 5.31	Thr (T) 5.61
Asn (N) 4.18	Gly (G) 7.06	Met (M) 2.44	Trp (W) 1.31
Asp (D) 5.28	His (H) 2.20	Phe (F) 4.04	Tyr (Y) 3.06
Cys (C) 1.30	Ile (I) 6.00	Pro (P) 4.76	Val (V) 6.70
Asx (B) 0.000	Glx (Z) 0.000	Xaa (X) 0.05	

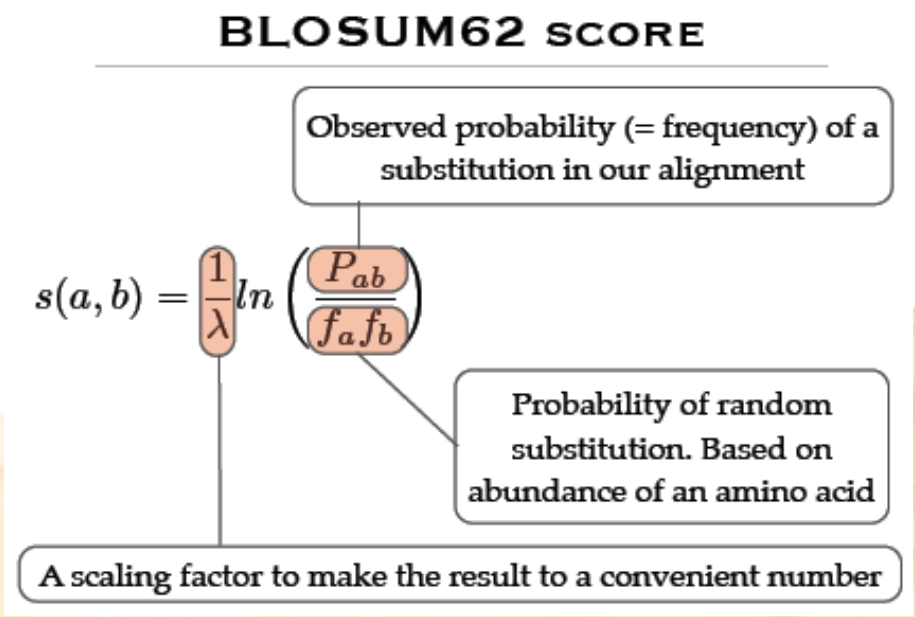
5. AMINO ACID COMPOSITION

5.1 Composition in percent for the complete database

Ala (A) 8.60	Gln (Q) 3.90	Leu (L) 9.86	Ser (S) 6.72
Arg (R) 5.46	Glu (E) 6.15	Lys (K) 5.25	Thr (T) 5.61
Asn (N) 4.12	Gly (G) 7.11	Met (M) 2.47	Trp (W) 1.31
Asp (D) 5.30	His (H) 2.20	Phe (F) 4.02	Tyr (Y) 3.03
Cys (C) 1.28	Ile (I) 5.98	Pro (P) 4.75	Val (V) 6.74
Asx (B) 0.000	Glx (Z) 0.000	Xaa (X) 0.04	

http://www.ebi.ac.uk/swissprot/sptr_stats/index.html

Calculating BLOSUM62 Scores



Breaking Down the Terms

BLOSUM62 SCORE

$$s(a,b) = \frac{1}{\lambda} ln\left(\frac{P_{ab}}{f_a f_b}\right)$$

?

?

P_{ab} – The likelihood of the hypothesis, we want to test: residues correlated b/c they are homologous

f_a x f_b – likelihood of a null hypothesis: residues unrelated

Getting Started, Step 1

$$s(a,b) = \frac{1}{\lambda} ln \left(\frac{P_{ab}}{f_a f_b} \right)$$

BLOSUM62 SCORE

Observed probability (= frequency) of a substitution in our alignment

$$s\,(a,b) = \frac{1}{\lambda} ln \left(\frac{0.0371}{f_a f_b} \right)$$

Leucine to Leucine substitution:

$$P_{ab} = 0.0371$$

Getting Started, Step 2

BLOSUM62 SCORE

Abundance of Leucin in nature, $f_a = 0.099$
Probability of a random substitution, $f_a \cdot f_b = f_L \cdot f_L = 0.099 \cdot 0.099$

$$s(a,b) = \frac{1}{\lambda} ln \left(\frac{0.0371}{0.099 \cdot 0.099} \right)$$

Probability of random substitution. Based on abundance of an amino acid

Scaling the Scores

BLOSUM62 SCORE

$\lambda =$ **0.0347**

$$s(a,b) = \frac{1}{0.0347} ln\left(\frac{0.0371}{0.099 \cdot 0.099}\right)$$

A scaling factor to make the result to a convenient number

Actual BLOSUM62 Scores

BLOSUM62 SCORE

$$s(a,b) = \frac{1}{\lambda} ln\left(\frac{P_{ab}}{f_a f_b}\right)$$

$$s\,(L,L) = \frac{1}{0.347} ln\left(\frac{0.0371}{0.099 \cdot 0.099}\right) = +3.8$$

$$s\,(W,W) = \frac{1}{0.347} ln\left(\frac{0.0065}{0.013 \cdot 0.013}\right) = +10.5$$

Rounded to +4 and +11

BLOSUM62

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

BLOSUM62 SCORE

$$s(a,b) = \frac{1}{\lambda} ln\left(\frac{P_{ab}}{f_a f_b}\right)$$

$$s\,(L,L) = \frac{1}{0.347} ln\left(\frac{0.0371}{0.099 \cdot 0.099}\right) = +3.8$$

$$s\,(W,W) = \frac{1}{0.347} ln\left(\frac{0.0065}{0.013 \cdot 0.013}\right) = +10.5$$

Rounded to +4 and +11

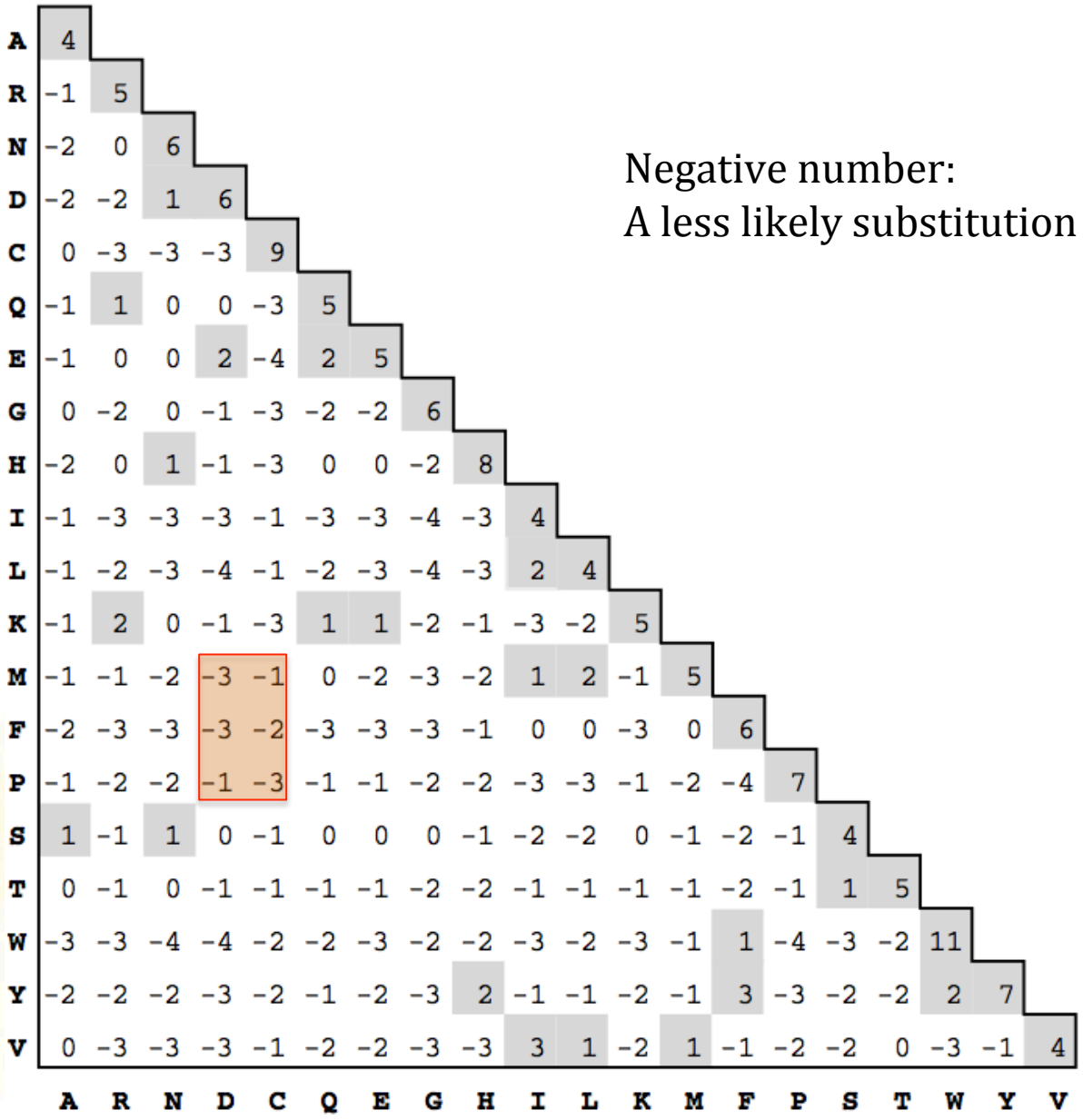
BLOSUM62

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

Positive Number:
Biologically meaningful substitution. The two a.a. replace each other more often than expected by chance alone. i.e. they are functionally interchangeable in the alignments studied -> More likely substitutions

BLOSUM62

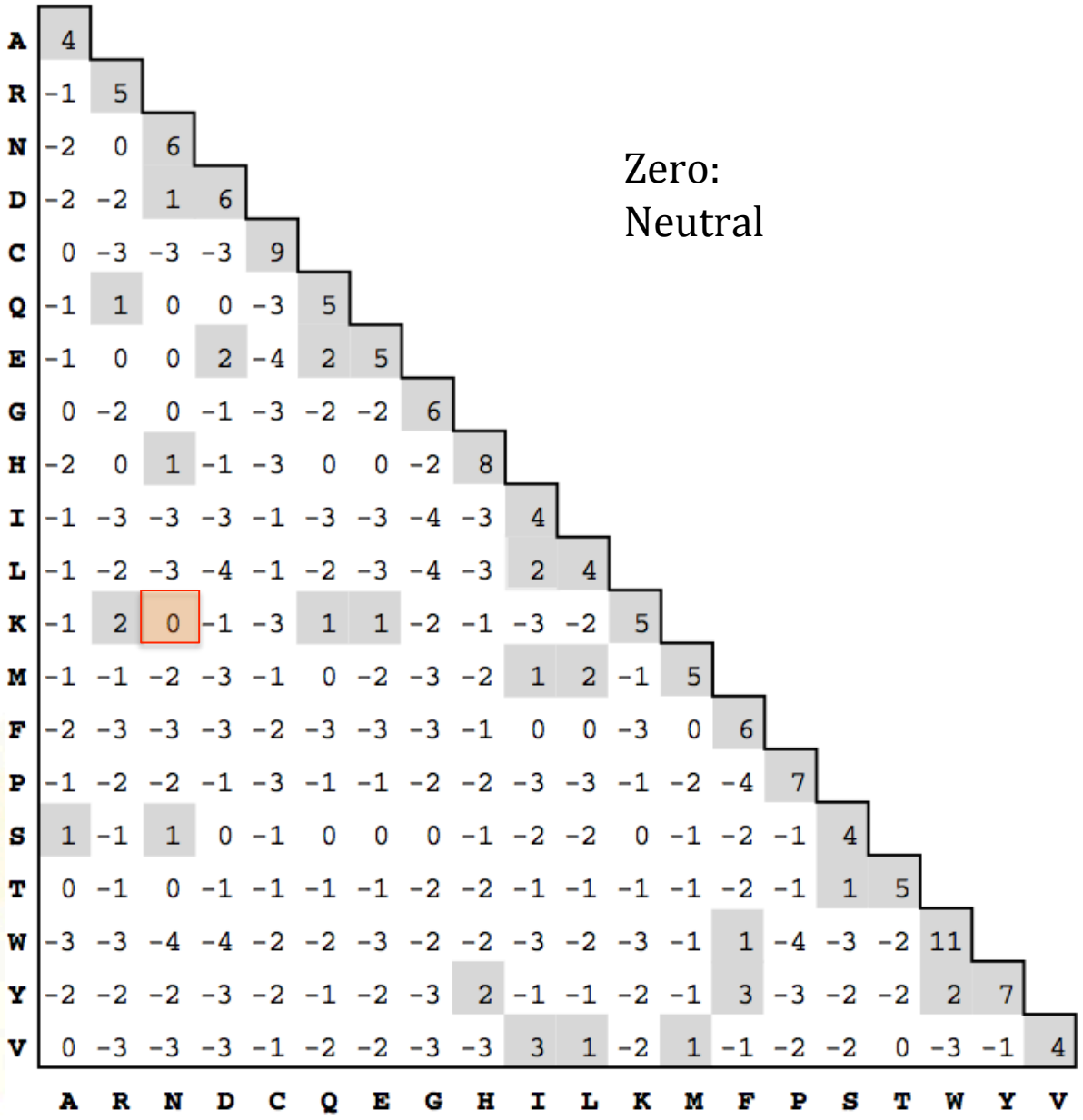
Negative number:
A less likely substitution



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

BLOSUM62

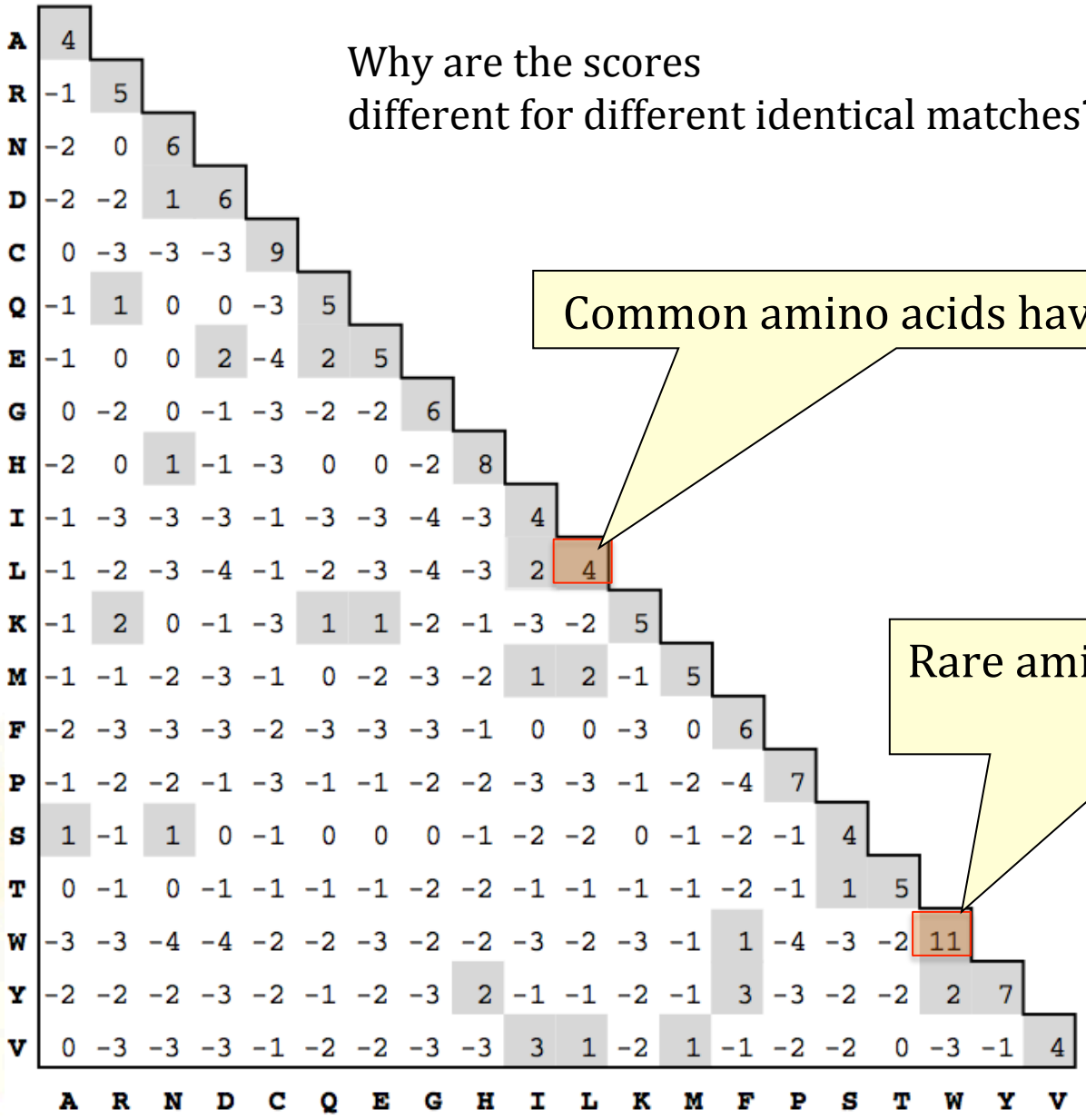
Zero:
Neutral



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

BLOSUM62

Why are the scores
different for different identical matches?



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

Common amino acids have low weights

Rare amino acids have high weights

What Does the BLOSUMXX mean?

- The take home:
  - High BLOSUM values mean **high similarity between** clusters
    - **Conserved substitution dominate**
  - Low BLOSUM values mean **low similarity between** clusters
    - **Less conserved substitutions dominate**

BLOSUM80

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7	-3	-3	-3	-1	-2	-2	0	-3	-3	-3	-1	-2	-4	-1	2	0	-5	-4	-1
R	-3	9	-1	-3	-6	1	-1	-4	0	-5	-4	3	-3	-5	-3	-2	-2	-5	-4	-4
N	-3	-1	9	2	-5	0	-1	-1	1	-6	-6	0	-4	-6	-4	1	0	-7	-4	-5
D	-3	-3	2	10	-7	-1	2	-3	-2	-7	-7	-2	-6	-6	-3	-1	-2	-8	-6	-6
C	-1	-6	-5	-7	13	-5	-7	-6	-7	-2	-3	-6	-3	-4	-6	-2	-2	-5	-5	-2
Q	-2	1	0	-1	-5	9	3	-4	1	-5	-4	2	-1	-5	-3	-1	-1	-4	-3	-4
E	-2	-1	-1	2	-7	3	8	-4	0	-6	-6	1	-4	-6	-2	-1	-2	-6	-5	-4
G	0	-4	-1	-3	-6	-4	-4	9	-4	-7	-7	-3	-5	-6	-5	-1	-3	-6	-6	-6
H	-3	0	1	-2	-7	1	0	-4	12	-6	-5	-1	-4	-2	-4	-2	-3	-4	3	-5
I	-3	-5	-6	-7	-2	-5	-6	-7	-6	7	2	-5	2	-1	-5	-4	-2	-5	-3	4
L	-3	-4	-6	-7	-3	-4	-6	-7	-5	2	6	-4	3	0	-5	-4	-3	-4	-2	1
K	-1	3	0	-2	-6	2	1	-3	-1	-5	-4	8	-3	-5	-2	-1	-1	-6	-4	-4
M	-2	-3	-4	-6	-3	-1	-4	-5	-4	2	3	-3	9	0	-4	-3	-1	-3	-3	1
F	-4	-5	-6	-6	-4	-5	-6	-6	-2	-1	0	-5	0	10	-6	-4	-4	0	4	-2
P	-1	-3	-4	-3	-6	-3	-2	-5	-4	-5	-5	-2	-4	-6	12	-2	-3	-7	-6	-4
S	2	-2	1	-1	-2	-1	-1	-1	-2	-4	-4	-1	-3	-4	-2	7	2	-6	-3	-3
T	0	-2	0	-2	-2	-1	-2	-3	-3	-2	-3	-1	-1	-4	-3	2	8	-5	-3	0
W	-5	-5	-7	-8	-5	-4	-6	-6	-4	-5	-4	-6	-3	0	-7	-6	-5	16	3	-5
Y	-4	-4	-4	-6	-5	-3	-5	-6	3	-3	-2	-4	-3	4	-6	-3	-3	3	11	-3
V	-1	-4	-5	-6	-2	-4	-4	-6	-5	4	1	-4	1	-2	-4	-3	0	-5	-3	7

$\bar{x}\, S_{ii} = 9.4$

$\bar{x}\, S_{ij} = -2.9$

BLOSUM30

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	0	0	-3	1	0	0	-2	0	-1	0	1	-2	-1	1	1	-5	-4	1
R	-1	8	-2	-1	-2	3	-1	-2	-1	-3	-2	1	0	-1	-1	-1	-3	0	0	-1
N	0	-2	8	1	-1	-1	-1	0	-1	0	-2	0	0	-1	-3	0	1	-7	-4	-2
D	0	-1	1	9	-3	-1	1	-1	-2	-4	-1	0	-3	-5	-1	0	-1	-4	-1	-2
C	-3	-2	-1	-3	17	-2	1	-4	-5	-2	0	-3	-2	-3	-3	-2	-2	-2	-6	-2
Q	1	3	-1	-1	-2	8	2	-2	0	-2	-2	0	-1	-3	0	-1	0	-1	-1	-3
E	0	-1	-1	1	1	2	6	-2	0	-3	-1	2	-1	-4	1	0	-2	-1	-2	-3
G	0	-2	0	-1	-4	-2	-2	8	-3	-1	-2	-1	-2	-3	-1	0	-2	1	-3	-3
H	-2	-1	-1	-2	-5	0	0	-3	14	-2	-1	-2	2	-3	1	-1	-2	-5	0	-3
I	0	-3	0	-4	-2	-2	-3	-1	-2	6	2	-2	1	0	-3	-1	0	-3	-1	4
L	-1	-2	-2	-1	0	-2	-1	-2	-1	2	4	-2	2	2	-3	-2	0	-2	3	1
K	0	1	0	0	-3	0	2	-1	-2	-2	-2	4	2	-1	1	0	-1	-2	-1	-2
M	1	0	0	-3	-2	-1	-1	-2	2	1	2	2	6	-2	-4	-2	0	-3	-1	0
F	-2	-1	-1	-5	-3	-3	-4	-3	-3	0	2	-1	-2	10	-4	-1	-2	1	3	1
P	-1	-1	-3	-1	-3	0	1	-1	1	-3	-3	1	-4	-4	11	-1	0	-3	-2	-4
S	1	-1	0	0	-2	-1	0	0	-1	-1	-2	0	-2	-1	-1	4	2	-3	-2	-1
T	1	-3	1	-1	-2	0	-2	-2	-2	0	0	-1	0	-2	0	2	5	-5	-1	1
W	-5	0	-7	-4	-2	-1	-1	1	-5	-3	-2	-2	-3	1	-3	-3	-5	20	5	-3
Y	-4	0	-4	-1	-6	-1	-2	-3	0	-1	3	-1	-1	3	-2	-2	-1	5	9	1
V	1	-1	-2	-2	-2	-3	-3	-3	-3	4	1	-2	0	1	-4	-1	1	-3	1	5

$\bar{x} S_{ii} = 8.3$

$\bar{x} S_{ij} = -1.16$

BLOSUM80

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7	-3	-3	-3	-1	-2	-2	0	-3	-3	-3	-1	-2	-4	-1	2	0	-5	-4	-1
R	-3	9	-1	-3	-6	1	-1	-4	0	-5	-4	3	-3	-5	-3	-2	-2	-5	-4	-4
N	-3	-1	9	2	-5	0	-1	-1	1	-6	-6	0	-4	-6	-4	1	0	-7	-4	-5
D	-3	-3	2	10	-7	-1	2	-3	-2	-7	-7	-2	-6	-6	-3	-1	-2	-8	-6	-6
C	-1	-6	-5	-7	13	-5	-7	-6	-7	-2	-3	-6	-3	-4	-6	-2	-2	-5	-5	-2
Q	-2	1	0	-1	-5	9	3	-4	1	-5	-4	2	-1	-5	-3	-1	-1	-4	-3	-4
E	-2	-1	-1	2	-7	3	8	-4	0	-6	-6	1	-4	-6	-2	-1	-2	-6	-5	-4
G	0	-4	-1	-3	-6	-4	-4	9	-4	-7	-7	-3	-5	-6	-5	-1	-3	-6	-6	-6
H	-3	0	1	-2	-7	1	0	-4	12	-6	-5	-1	-4	-2	-4	-2	-3	-4	3	-5
I	-3	-5	-6	-7	-2	-5	-6	-7	-6	7	2	-5	2	-1	-5	-4	-2	-5	-3	4
L	-3	-4	-6	-7	-3	-4	-6	-7	-5	2	6	-4	3	0	-5	-4	-3	-4	-2	1
K	-1	3	0	-2	-6	2	1	-3	-1	-5	-4	8	-3	-5	-2	-1	-1	-6	-4	-4
M	-2	-3	-4	-6	-3	-1	-4	-5	-4	2	3	-3	9	0	-4	-3	-1	-3	-3	1
F	-4	-5	-6	-6	-4	-5	-6	-6	-2	-1	0	-5	0	10	-6	-4	-4	0	4	-2
P	-1	-3	-4	-3	-6	-3	-2	-5	-4	-5	-5	-2	-4	-6	12	-2	-3	-7	-6	-4
S	2	-2	1	-1	-2	-1	-1	-1	-2	-4	-4	-1	-3	-4	-2	7	2	-6	-3	-3
T	0	-2	0	-2	-2	-1	-2	-3	-3	-2	-3	-1	-1	-4	-3	2	8	-5	-3	0
W	-5	-5	-7	-8	-5	-4	-6	-6	-4	-5	-4	-6	-3	0	-7	-6	-5	16	3	-5
Y	-4	-4	-4	-6	-5	-3	-5	-6	3	-3	-2	-4	-3	4	-6	-3	-3	3	11	-3
V	-1	-4	-5	-6	-2	-4	-4	-6	-5	4	1	-4	1	-2	-4	-3	0	-5	-3	7

$S_{II} = 7$

$S_{ID} = -7$

$S_{IC} = -2$

BLOSUM30

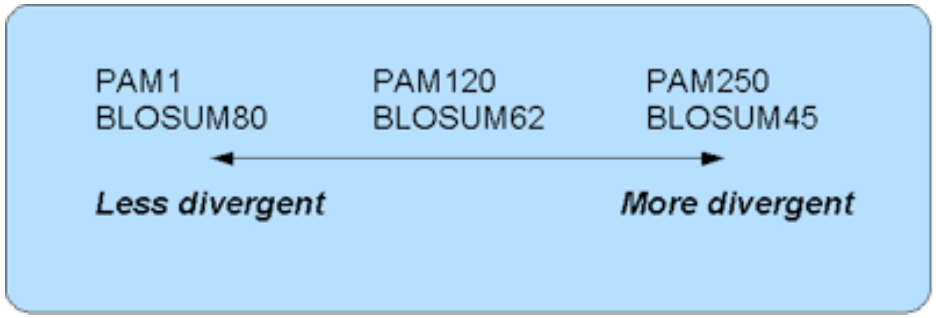
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	0	0	-3	1	0	0	-2	0	-1	0	1	-2	-1	1	1	-5	-4	1
R	-1	8	-2	-1	-2	3	-1	-2	-1	-3	-2	1	0	-1	-1	-1	-3	0	0	-1
N	0	-2	8	1	-1	-1	-1	0	-1	0	-2	0	0	-1	-3	0	1	-7	-4	-2
D	0	-1	1	9	-3	-1	1	-1	-2	-4	-1	0	-3	-5	-1	0	-1	-4	-1	-2
C	-3	-2	-1	-3	17	-2	1	-4	-5	-2	0	-3	-2	-3	-3	-2	-2	-2	-6	-2
Q	1	3	-1	-1	-2	8	2	-2	0	-2	-2	0	-1	-3	0	-1	0	-1	-1	-3
E	0	-1	-1	1	1	2	6	-2	0	-3	-1	2	-1	-4	1	0	-2	-1	-2	-3
G	0	-2	0	-1	-4	-2	-2	8	-3	-1	-2	-1	-2	-3	-1	0	-2	1	-3	-3
H	-2	-1	-1	-2	-5	0	0	-3	14	-2	-1	-2	2	-3	1	-1	-2	-5	0	-3
I	0	-3	0	-4	-2	-2	-3	-1	-2	6	2	-2	1	0	-3	-1	0	-3	-1	4
L	-1	-2	-2	-1	0	-2	-1	-2	-1	2	4	-2	2	2	-3	-2	0	-2	3	1
K	0	1	0	0	-3	0	2	-1	-2	-2	-2	4	2	-1	1	0	-1	-2	-1	-2
M	1	0	0	-3	-2	-1	-1	-2	2	1	2	2	6	-2	-4	-2	0	-3	-1	0
F	-2	-1	-1	-5	-3	-3	-4	-3	-3	0	2	-1	-2	10	-4	-1	-2	1	3	1
P	-1	-1	-3	-1	-3	0	1	-1	1	-3	-3	1	-4	-4	11	-1	0	-3	-2	-4
S	1	-1	0	0	-2	-1	0	0	-1	-1	-2	0	-2	-1	-1	4	2	-3	-2	-1
T	1	-3	1	-1	-2	0	-2	-2	-2	0	0	-1	0	-2	0	2	5	-5	-1	1
W	-5	0	-7	-4	-2	-1	-1	1	-5	-3	-2	-2	-3	1	-3	-3	-5	20	5	-3
Y	-4	0	-4	-1	-6	-1	-2	-3	0	-1	3	-1	-1	3	-2	-2	-1	5	9	1
V	1	-1	-2	-2	-2	-3	-3	-3	-3	4	1	-2	0	1	-4	-1	1	-3	1	5

$S_{II} = 6$

$S_{ID} = -2$

$S_{IC} = -2$

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

BTW - Why Log Odds Form

- BLOSUM and PAM matrices start as a likelihood of substitution
- Conversion to **odds form** yields a matrix that gives the odds that a change is evolutionarily significant versus purely random by simple conversion
- So why the conversion to **log odds form?**
  - When scoring an alignment:
    - You can **add** the values
    - If you did not take the logarithm, we would need to multiply the ratios at all the aligned positions, so what?

This is computationally cumbersome!!

http://www.mathsisfun.com/algebra/logarithms.html

For Thursday

- Go over lecture, and understand the theory
- Select a protein from the list of proteins at our local page
  - http://155.33.203.128/teaching/BIOL6308-Fall2013/local/local_BIOL6308_fall2013.html#LIST
  - Get the translated protein from the list
  - Blast at NCBI with defaults against nr database, then:
    - Change the matrix used
    - Change the Gap opening and Gap extensions penalties
    - How does it change the results?
- Post Lab for Thursday later tonight
- Read "***Where did the BLOSUM62 alignment score matrix come from*?"**
  - **Sean Eddy – Local page**
  - http://155.33.203.128/teaching/BIOL6308-Fall2013/local/Literature/Eddy.pdf