

4 - 11 April, 2003 ... Lectures 3 - 6:

Readings and Synopsis

- **Should have read chapter 2**
- **Reading for this week and next:**
 - Chapter 3 intro: pp 53-58 for Friday 4 April
 - Chapter 3: pp 59-64 for Monday 7 April
 - Chapter 3: pp 65-76 for Wednesday 9 April
 - Chapter 3: pp 77-121 for Friday 11 April
- **Quiz on Friday 11 April in Discussion Section:**
 - Cover through Wed 9 April Lecture
- **This week and Next:**
 - Homology
 - Sequence Analysis/Comparison
 - *Simple numerical methods*
 - *Dot matrix analysis*
 - *Introduction to alignments*

Lec 3: Goals of Sequence Analysis

- **Ab initio calculation ... often very difficult**
 - Calculate some property, e.g. folding based on sequence ... “folding problem”
 - Determination of Binding sites ... substrates, large molecule complexes
- **Mapping:**
 - Residue X in this sequence is "like" residue Y in another, eg F and Y
 - One of the most powerful concepts in molecular biology
 - Relies on the concept of homology (common ancestor)
 - This has been a main justification for DNA sequencing
- **Classification:**
 - Homologous sequences form a group or hierarchy (e.g. a phylogeny)
 - Family, Superfamily, etc
- **Modeling**
 - Models of cell systems, metabolism, organisms

Lec 3: Homology

What is Homology?

- **HOMOLOGY** -
the presence of a similar feature because of descent from a common ancestor
- **HOMOPLASY** -
the presence of a similar feature because of convergence
 - Homology cannot be observed. We can't actually see the ancestral organisms/molecules and trace descent.
 - Homology is an inference, a conclusion we draw based on observed similarity.
- **HOMOLOGY versus SIMILARITY** -
 - Homology is an All-or-None, Qualitative Relationship ... like Pregnancy
 - Similarity is a Quantitative Relationship

Lec 3: Homology

Why is Homology Important?

- **Homology strongly suggests that the molecules have similar structure and function**
- **There are (very) many ways to fold a polypeptide to place specific chemical groups at specific locations. There is no reason, *a priori*, why proteins with a specific function should have similar 3-D structures.**
- **Therefore, there is no reason, *a priori*, why unrelated sequences should have any detectable similarity in sequence. Significantly similar molecular sequences are very unlikely to arise by chance - i.e. homoplasy on the molecular level is very rare.**
- **When we see significant similarity, we infer that the sequences/structures are homologous, i.e. at some point in the past they share a common ancestor and have an identical sequence and structure.**
- **The only thing that keeps the sequences tied to each other through evolution and mutation accumulation is the commonality of structure and function arising from homology.**

Lec 3: Sequence Analysis

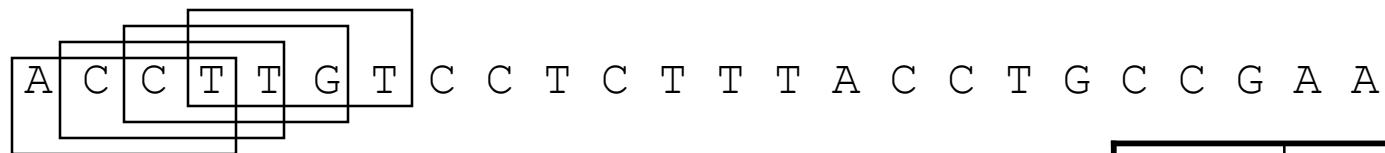
Windowed Calculation

- **One of the simplest types of analysis is the calculation of some property within a “window” in the sequence.**
- **Examples:**
 - GC content – related to gene coding
 - CpG frequency – related to gene coding
 - Protein hydrophobicity
 - Amino acid structural propensities
 - Protein Secondary Structure features
 - *Alpha Helices*
 - *Beta Sheet*
 - Amphipathicity (alpha helix, hydrophobic residues on one side, hydrophilic residues on the other ... membrane pore)

Lec 3: Sequence Comparison

Windowed Calculations

- Simple plots of properties are very noisy. One solution is to plot a running sum or average over a window.
- Simplest example is base composition, e.g., percent G+C
 - _ Can think of this as using a scoring table of 0 and 1



Window₁=0.5

Window₂=0.5

Window₃=0.5

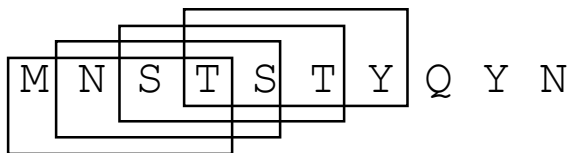
Window₄=0.25

<i>Base</i>	<i>Score</i>
<i>A</i>	<i>0</i>
<i>C</i>	<i>1</i>
<i>G</i>	<i>1</i>
<i>T</i>	<i>0</i>

Lec 3: Sequence Comparison

Windowed Calculations

- Proteins require a more complicated scoring system, e.g., for protein hydrophobicity
- A Distance Matrix is often used here (matrix of numbers relating one residue to another)

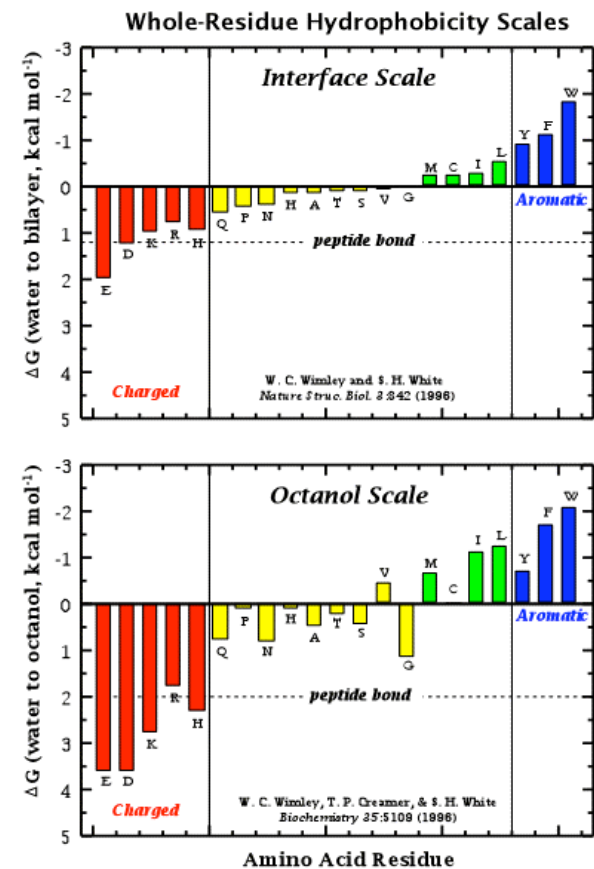


$$\text{Window}_1 = 0.78$$

$$\text{Window}_2 = 1.45$$

$$\text{Window}_3 = 0.75$$

$$\text{Window}_4 = 0.90$$

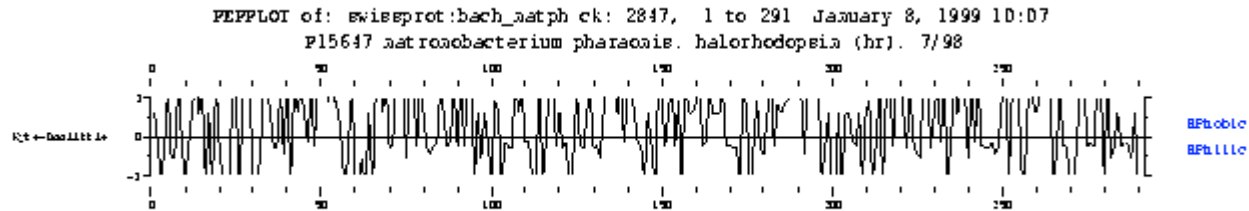


Lec 3: Sequence Comparison

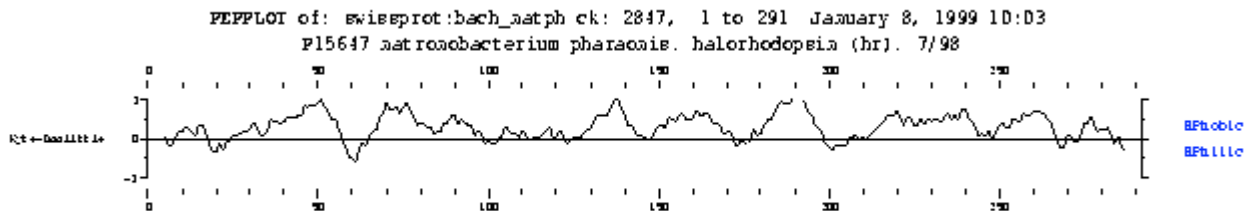
Windowed Calculations:

Plot of Average at Midpoint of Window ... Window moved One Residue at a Time

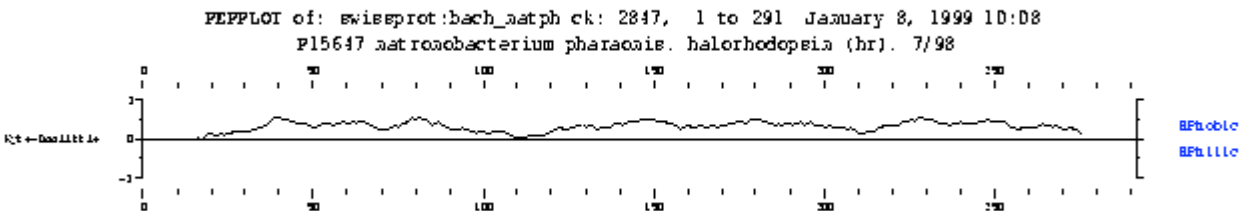
Window Size = 1:



Window Size = 10:



Window Size = 30:



Choose Window Size: 1) Large Enough to Minimize Noise; 2) Small Enough to show Features