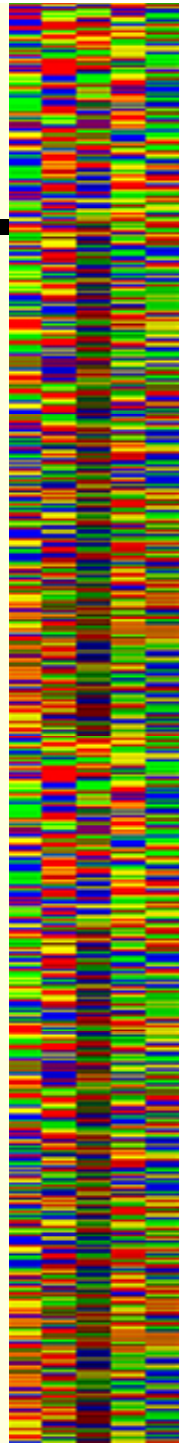
15 – 19 September

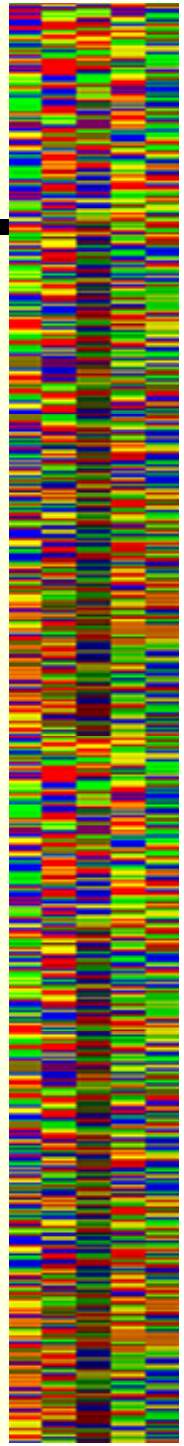
9	M 15	MG	Pairwise Alignment		Ch 9 and Handout
10	W 17	MG	Genome Sequencing		
11	F 19	MG	Gene Finding/Annotation	Hw3	

- ***if you don't know anything about DNA sequencing read also pages 33-40***
- ***grad section – be prepared to be presenter, meet in Lilly G-230***

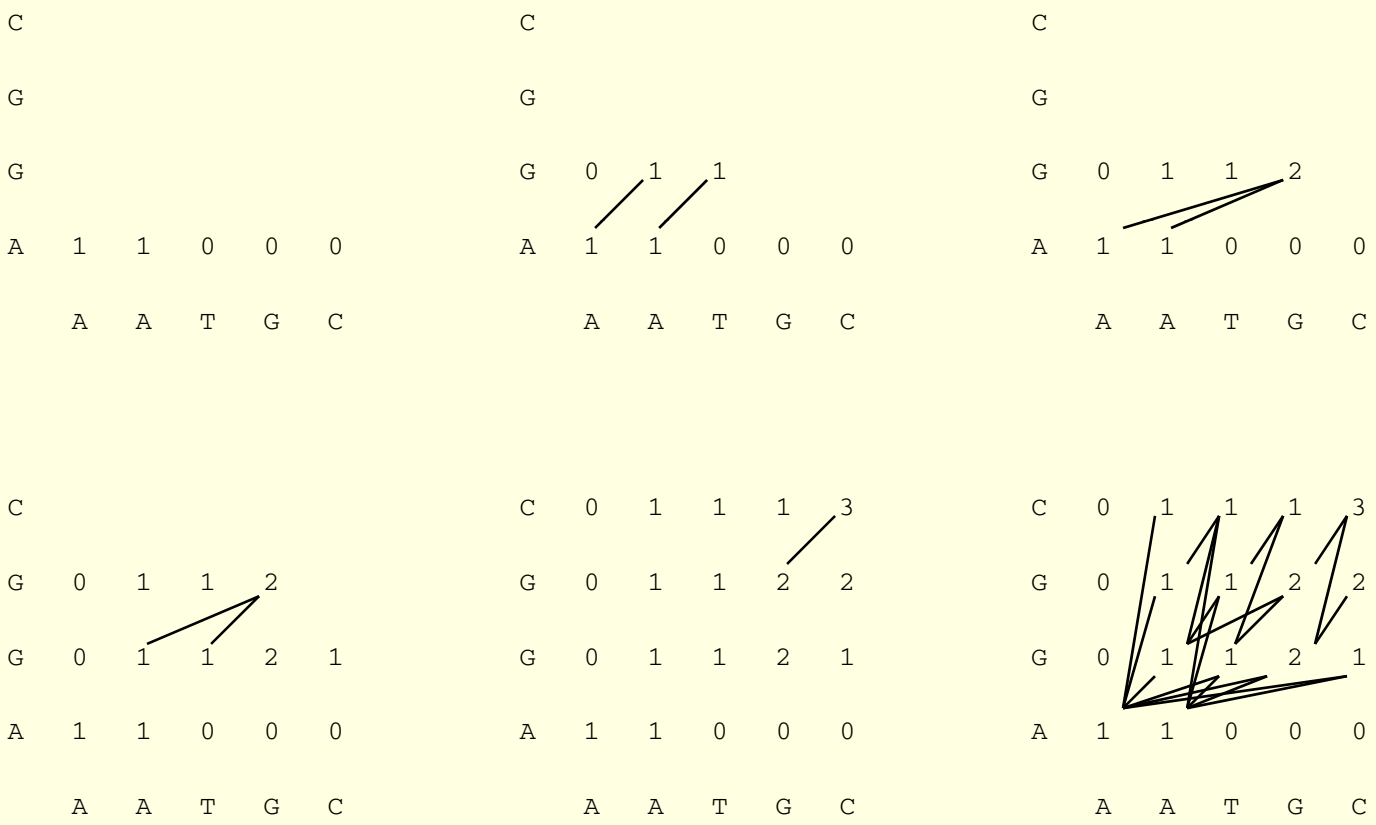


Dynamic Programming Alignment

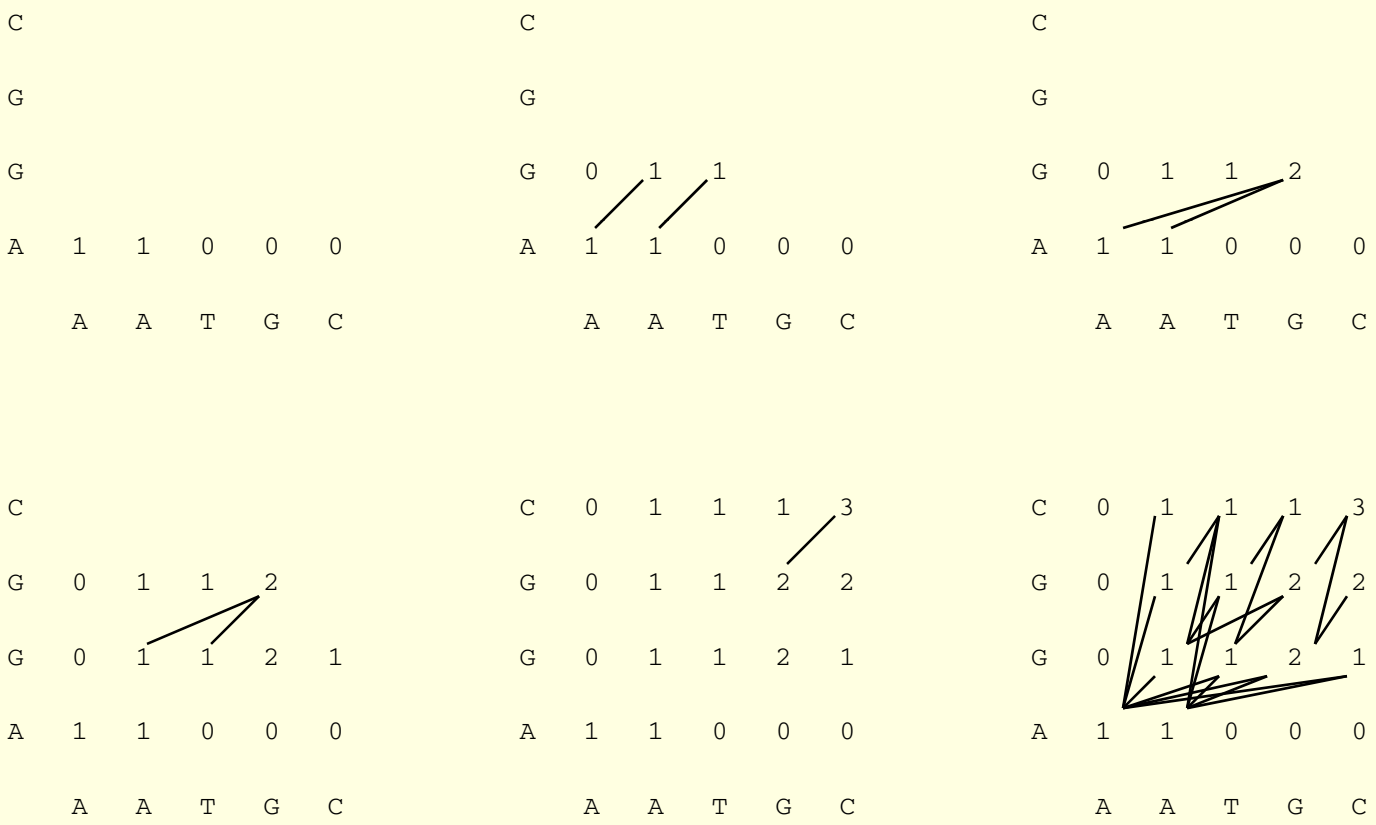
- ***A simple example, alignment of AATGC and AGGC***
- ***Scoreing system***
  - Match = +1
  - Mismatch = 0
  - Gaps = 0
- ***Global alignment (all sequence characters used from both sequences)***



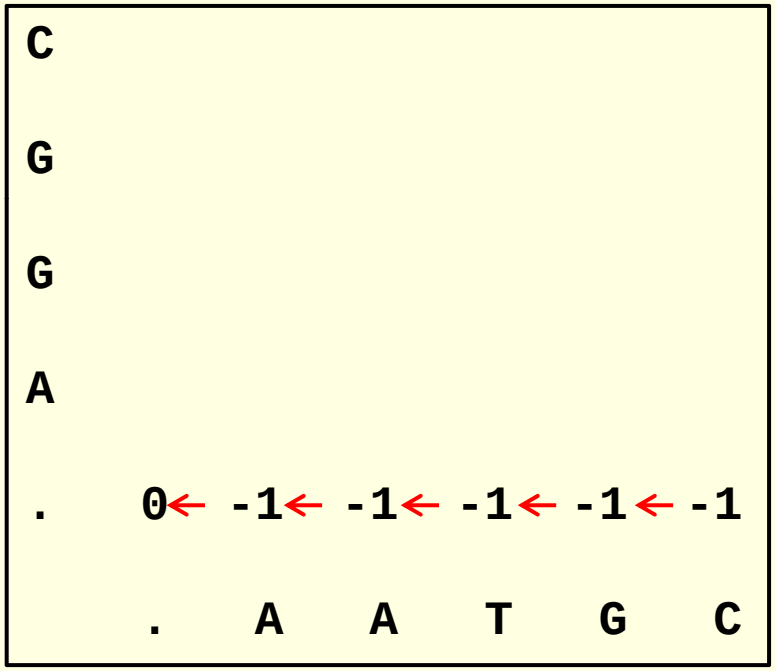
Dynamic Programming Alignment



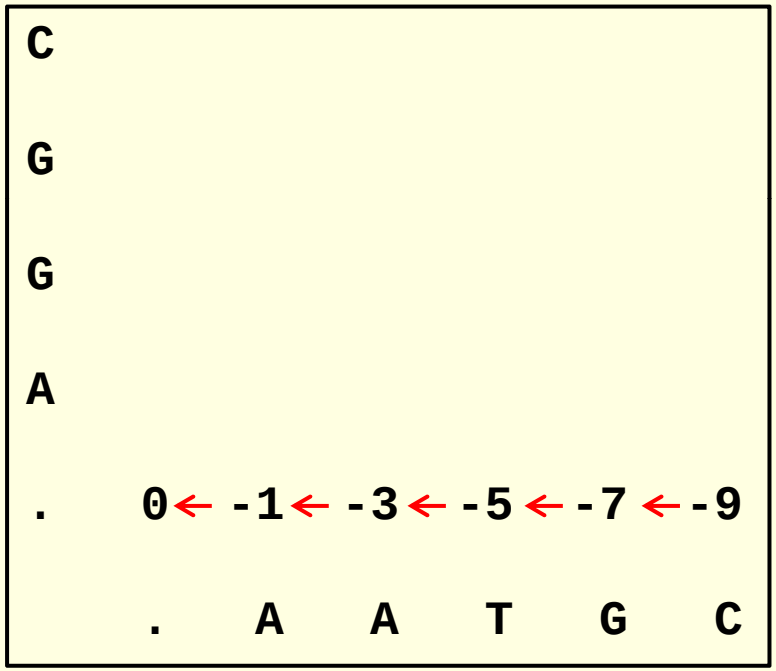
Dynamic Programming Alignment



Global Dynamic Programming Alignment

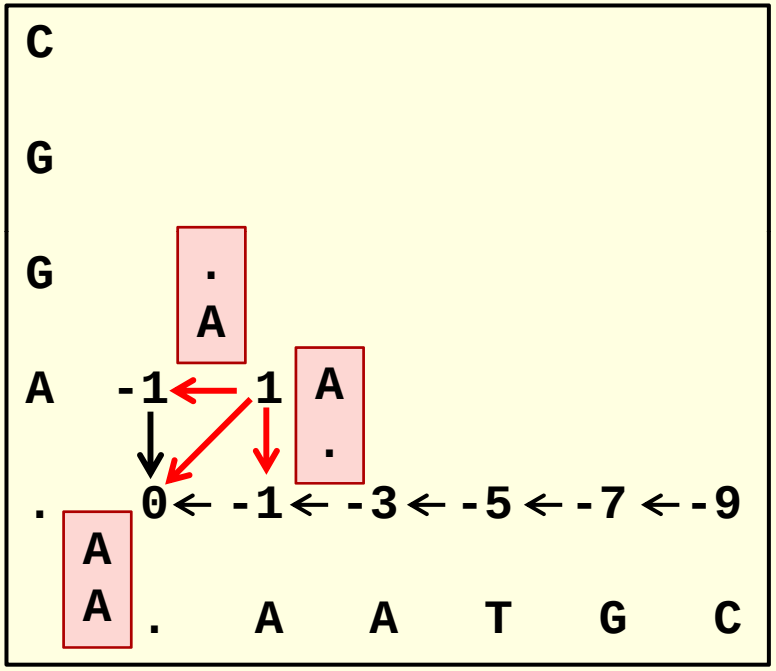


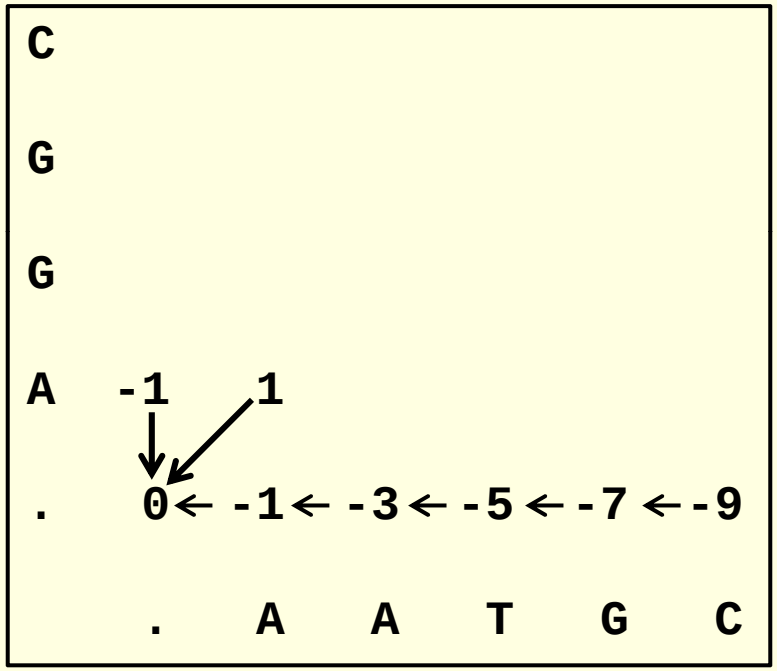
gap_{open} = -1 gap_{extend} = 0



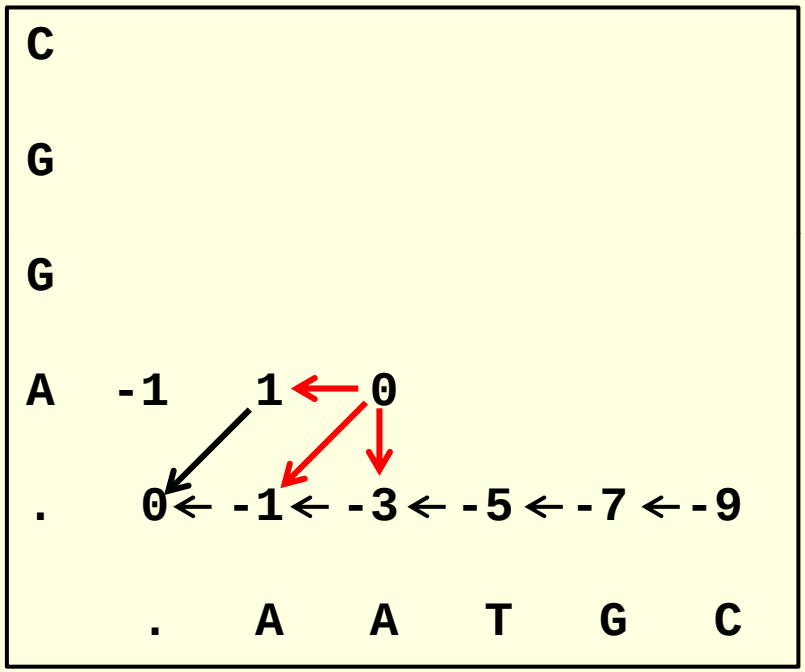
gap_{open} = -1 gap_{extend} = -2

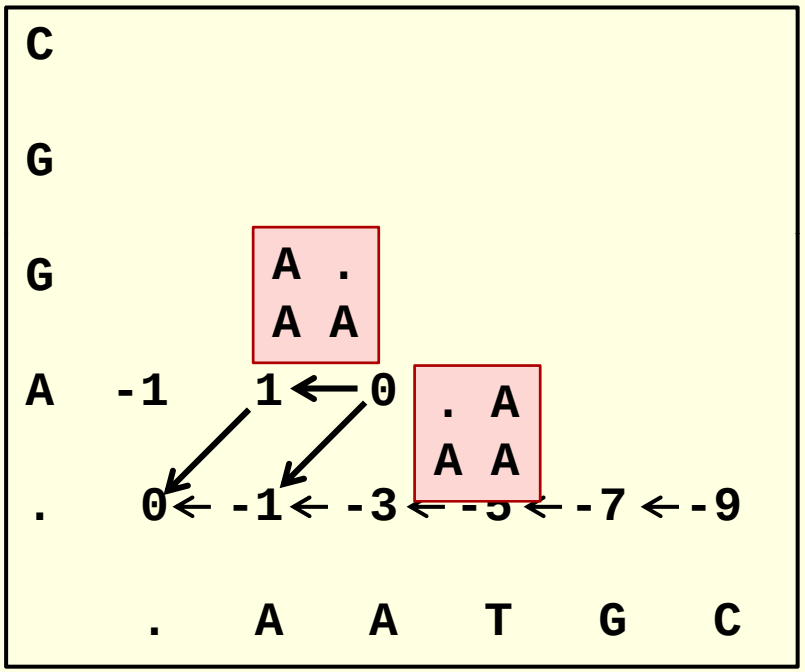
Dynamic Programming Alignment



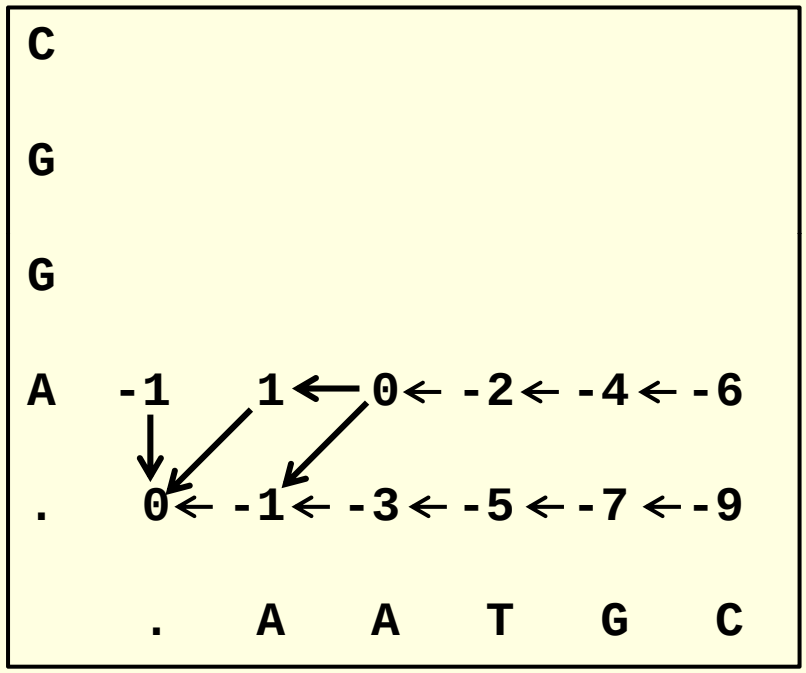


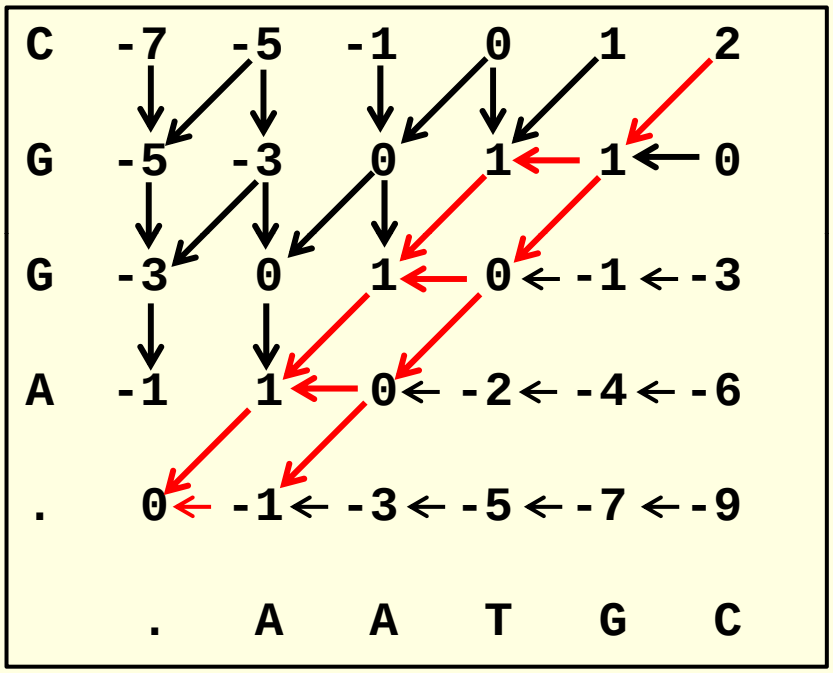
Dynamic Programming Alignment





Dynamic Programming Alignment





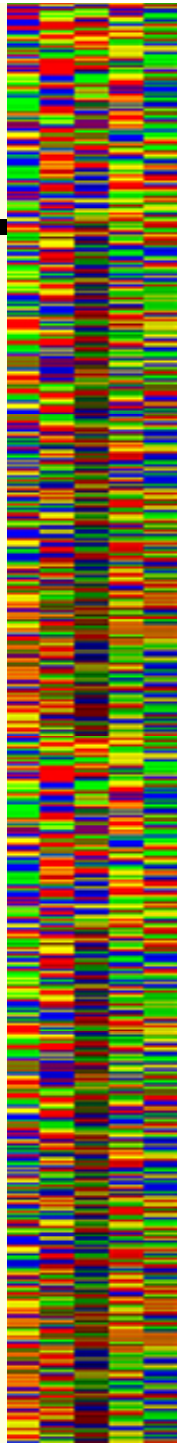
3 equivalent:

```
A G . G C      . A G G C      A . G G C
A A T G C      A A T G C      A A T G C
```

Statistics I

Models

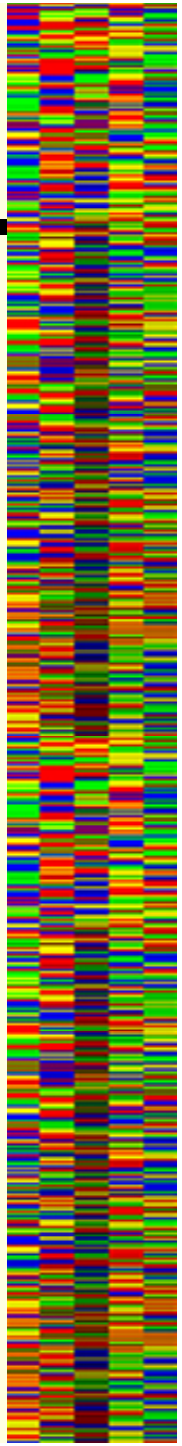
- ***Models allow us to answer the question, "How surprised am I?"***
- ***Models tell us what we expect to see***
  - If we know what to expect, we can tell if we should be surprised
  - Usually predict behavior of unrelated sequences
  - Models allow us to be quantitative
    - Statistics are a formal and quantitative way of measuring surprise
  - Models are nearly always a simplification
    - In some situations, the simplification may not be appropriate!
    - If you understand the model you are less likely to be fooled



Statistics I

Common models

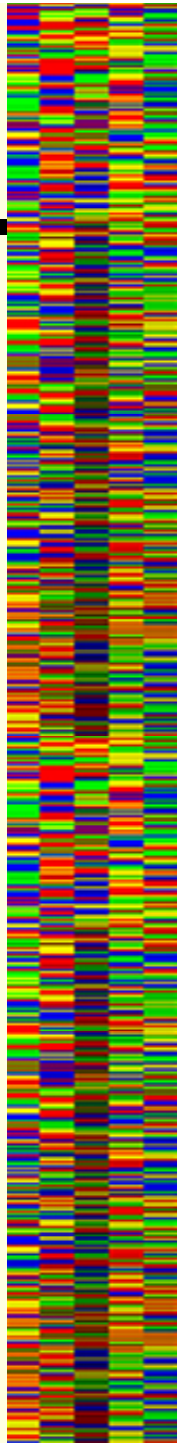
- ***Three kinds models commonly used in molecular biology***
- ***Random sequence model***
  - Assumes unrelated sequences behave as random or "scrambled" sequences
  - often evaluated by Monte Carlo approach
- ***Unrelated sequence model***
  - Assumes you can actually tell which ones are unrelated
  - FASTA, Profilesearch
- ***Theoretical models***
  - Many possibilities, many assumptions
  - Extreme value theory/BLAST



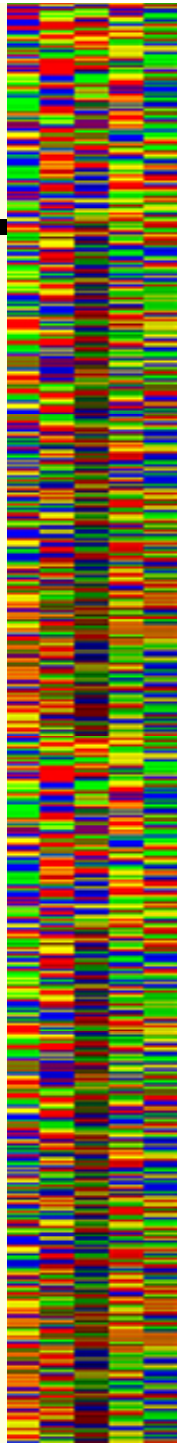
Statistics I

Random Sequence Model

- ***Assumes that unrelated sequences act like random sequences***
- ***A random sequence is typically created by sampling residues or bases a random according to the frequencies in the database***
- ***How are random sequences unrealistic***
  - Lengths?
  - Composition?
  - Patterns?



Genomics - Sequence Alignment



Monte Carlo Approach

- ***Compares result to randomized result, similarly to results generated by a roulette wheel at Monte Carlo***
- ***Typical procedure for alignments***
  - Randomize sequence A
  - Align to sequence B
  - Repeat many times (hundreds - thousands)
  - Use average as expected score to predict behavior of unrelated sequences
- ***A common statistic is the Z score (standardized score, standard normal deviate)***
  - Z = (Obs_score - Exp_score) / Std_deviation
  - Expected score depends on model
- ***Bad Rule:***
  - $Z < 3$ No evidence of homology
  - $3 < Z < 6$ Homology possible
  - $6 < Z$ Strong evidence of homology, ($Z > 8$) better ???

Genomics - Sequence Alignment

Updated Monte Carlo procedure for dynamic programming alignments

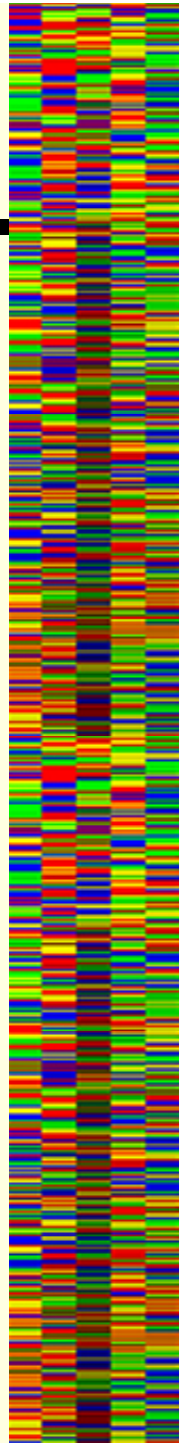
- ***Karlin-Altschul statistics are approximately correct***

  $n = KNe^{-\lambda S}$

  $\ln(n/N) = \ln K - \lambda S$ *this is a linear equation*

  $y = b + mx$

- ***Plot the log of the "observed P-value" vs score for randomized alignments of the same length and composition and determine P from the linear plot***
  - Randomize sequences and align using same parameters a large number of times, e.g. >10,000. Rank results by score. Observed P-value is rank divided by number of samples



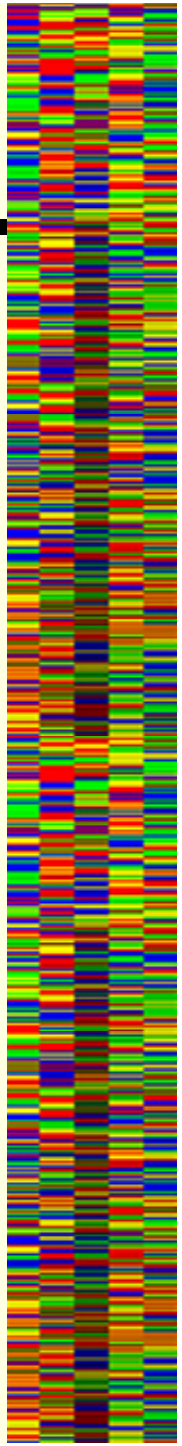
Statistics I

The big question: Is this result interesting?

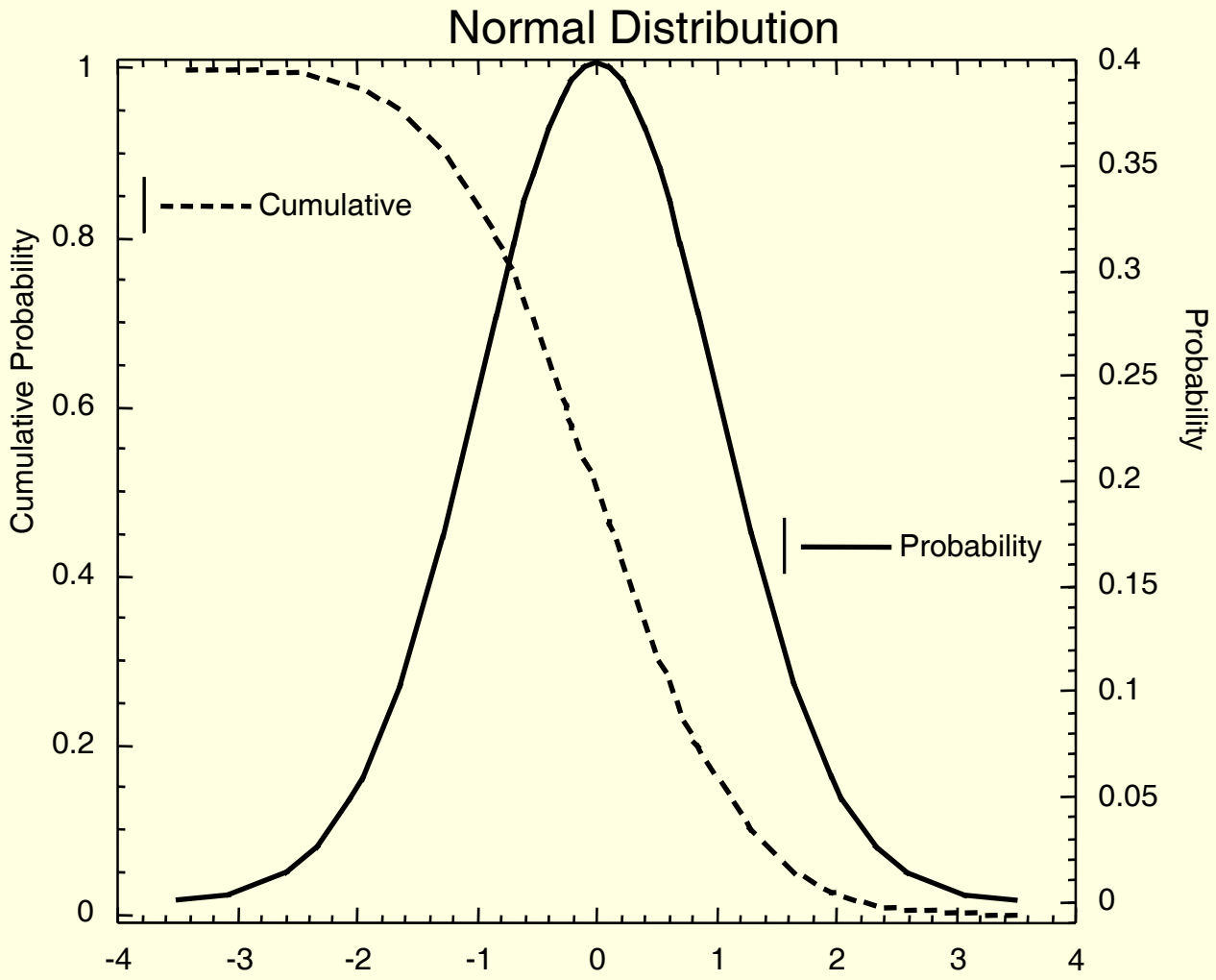
- ***Is this %GC indicative of gene coding?***
- ***Is number of points in a dotplot interesting?***
- ***Is this alignment score interesting?***
- ***Is a database search with a specific score interesting?***

- ***How do you answer these questions, which are fundamental to deciding if you are looking at homology, in a reliable way?***

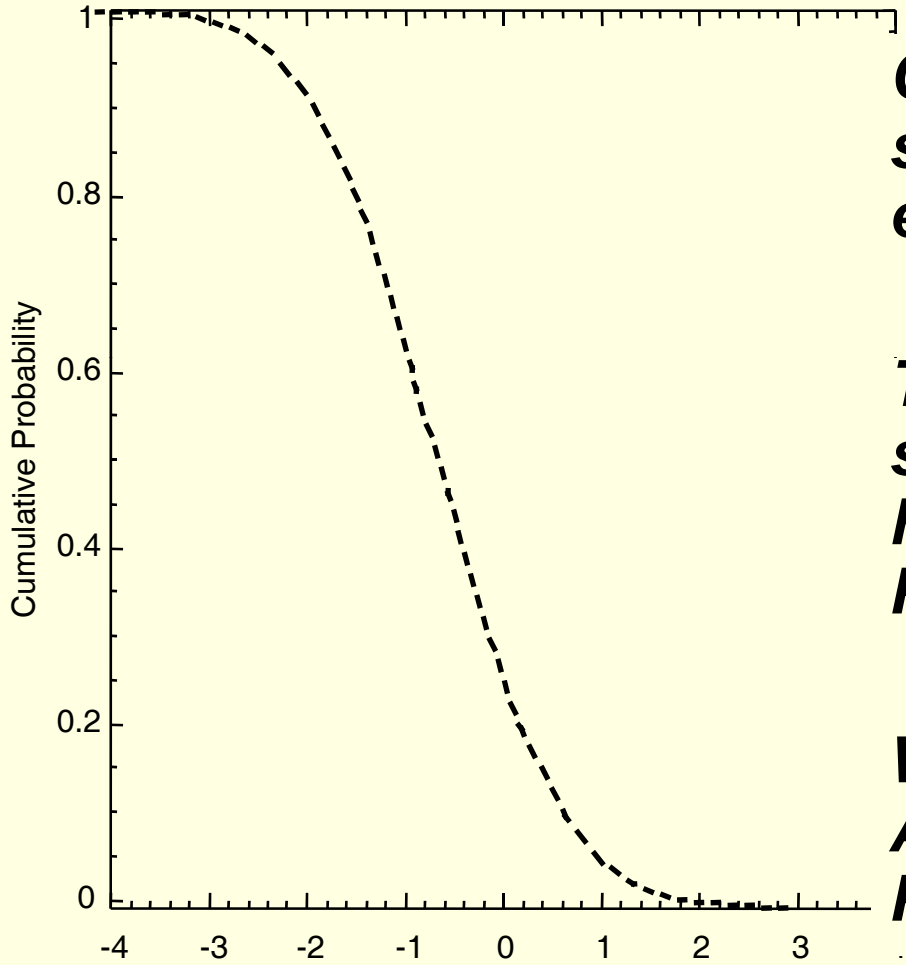
- ***Suprisal: Is this event surprising, or is it unusual?***
- ***The more surprising or an unusual an event is, the more interesting it is.***
- ***We don't know the distribution of the unusual events, but we often do know the distribution of the usual ones.***



Statistics I



Statistics I



Count the fraction of scores equal or higher to each value

This is the probability of seeing a score equal or higher than each value $P(S \geq x)$

What is surprising? A traditional value is $P(S \geq x) < 0.05$

Statistics I

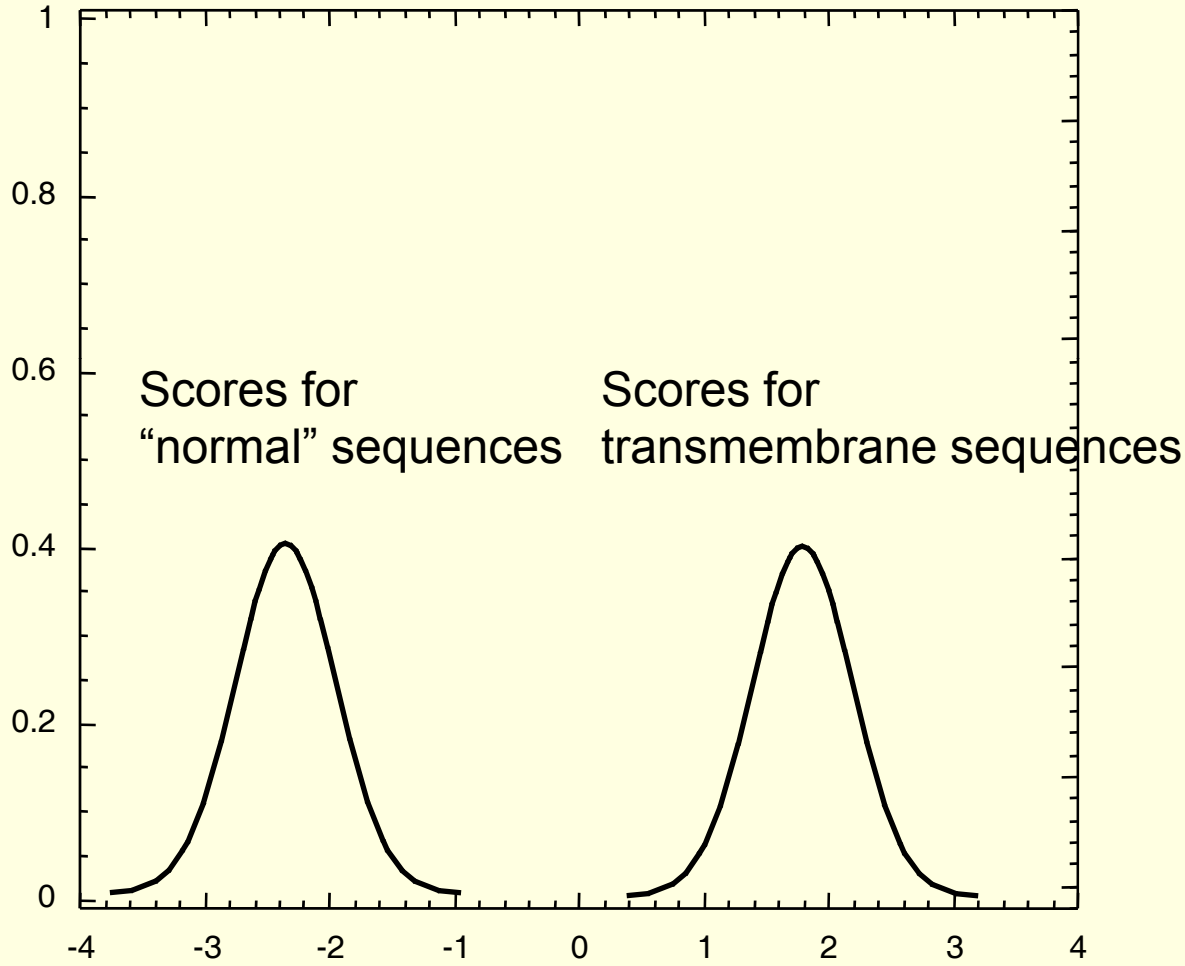
P-values

- *Probability that a random comparison will score at or above a given threshold, e.g. probability of a window scoring greater than 100*
- $P(score_{win} \geq 100) = 0.2 \times 10^{-3}$

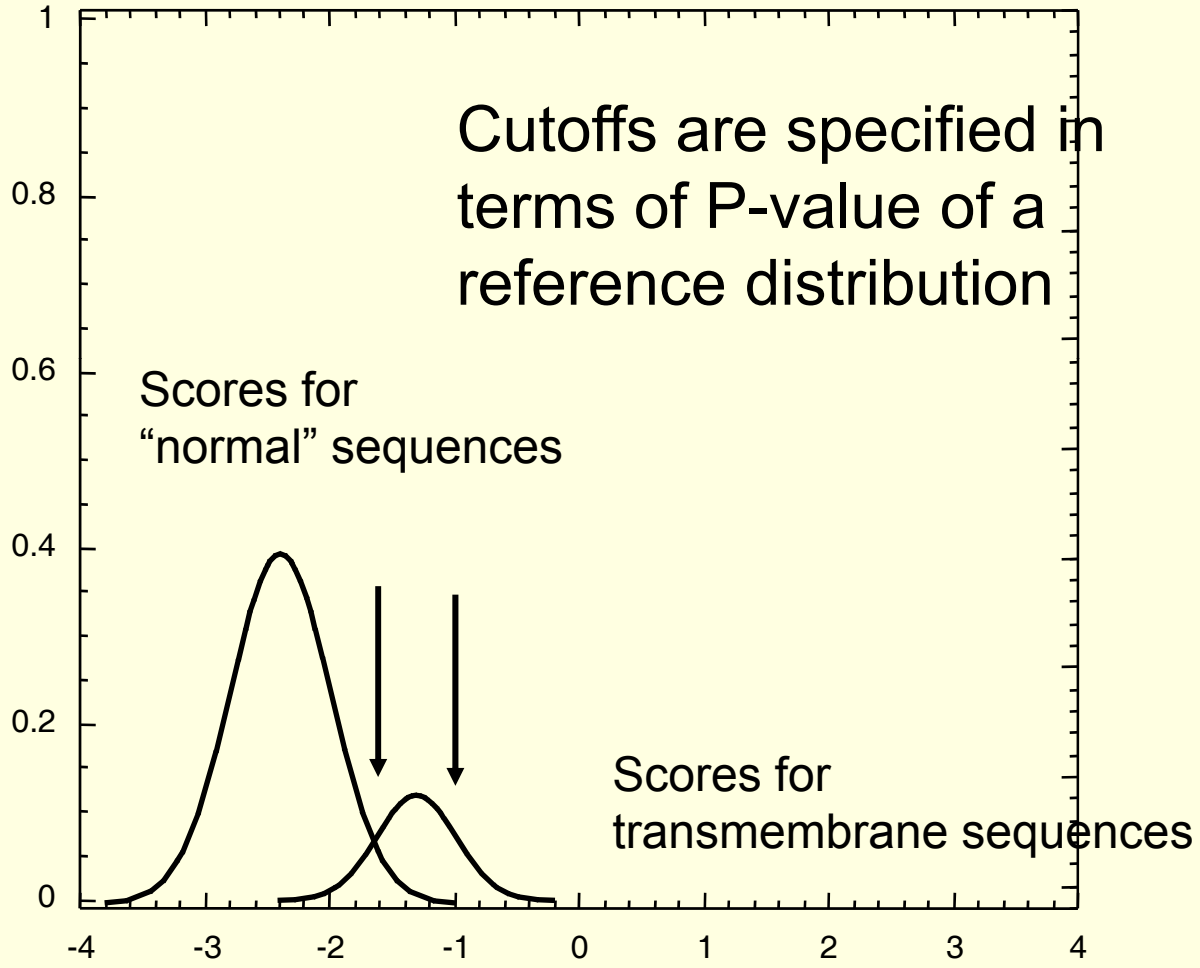
E-value

- *Expected number of random comparisons scoring above a given threshold*
- $E(score_{win} \geq 100) = P(score_{win} \geq 100) \times \text{number of trials}$
  $= 0.2 \times 10^{-3} \times N$

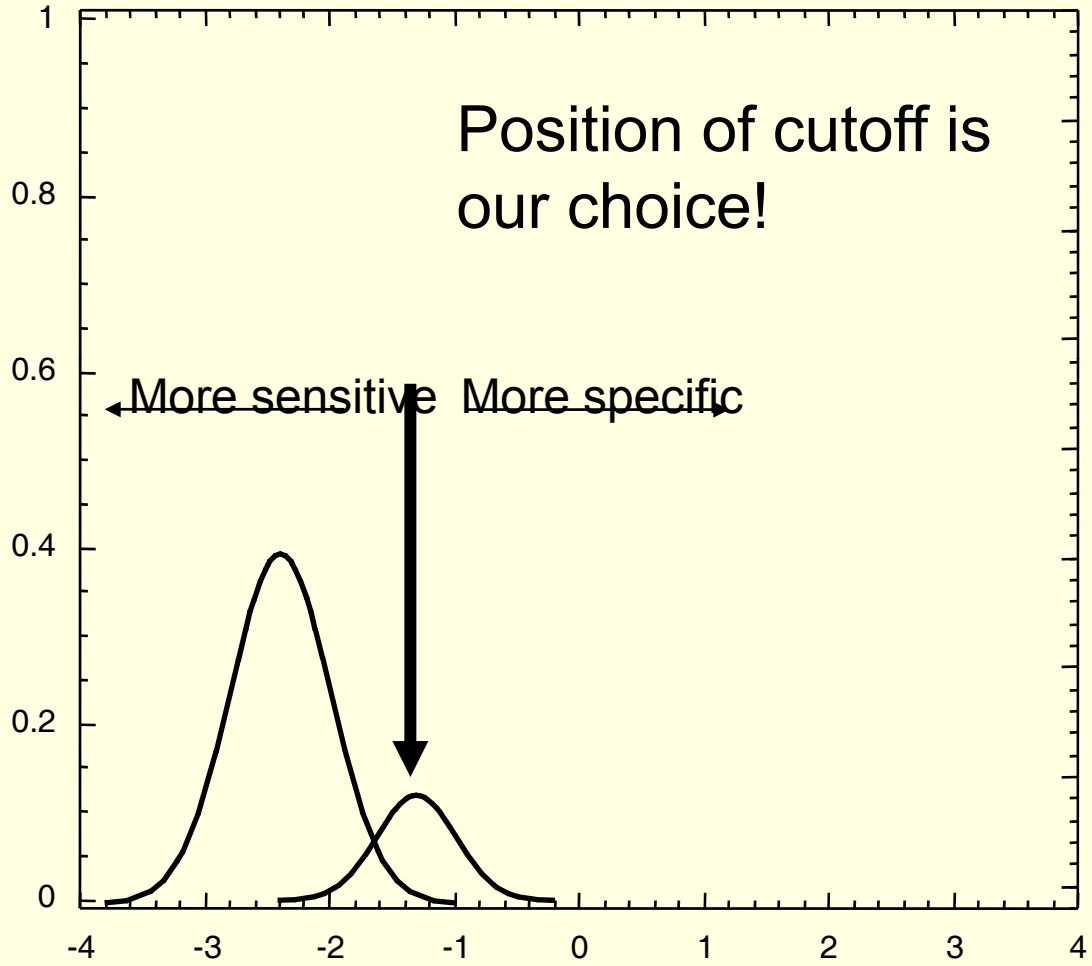
Statistics I



Statistics I



Statistics I



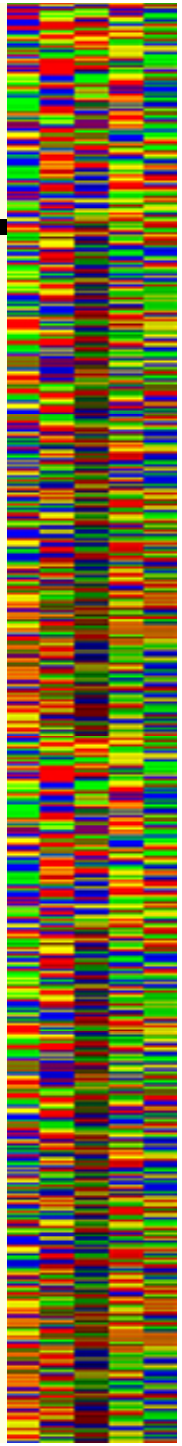
Statistics I

How surprising do we want something to be?

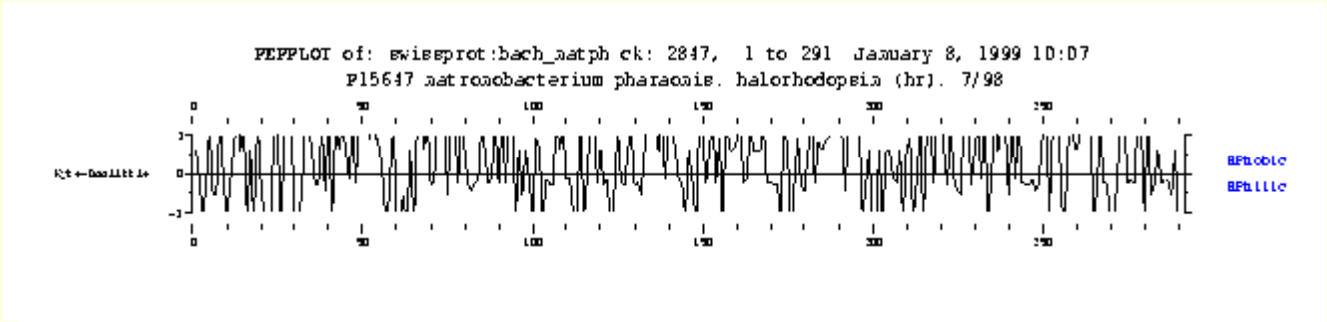
- ***How do you decide how high the score needs to be to be significant?***
  - Is 95% confidence the magic number?

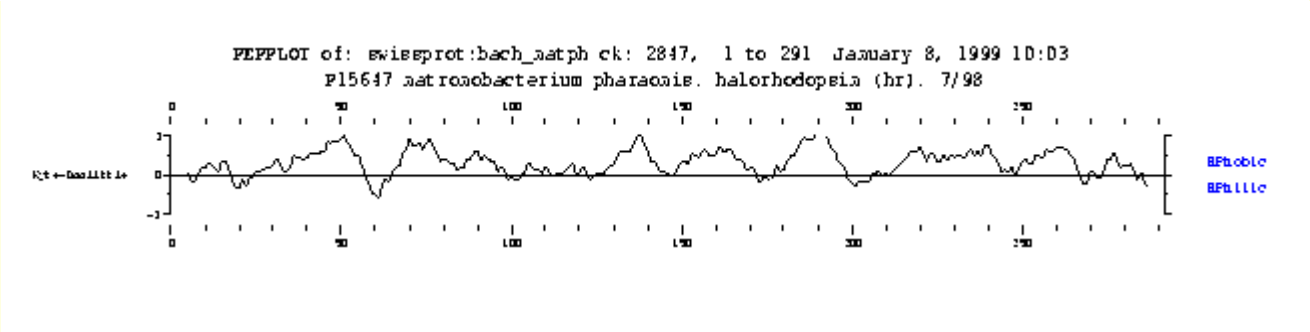
- ***Need some experience of what scores to expect***
- ***Need a rough idea of how common true interesting events are***
- ***Need to know how costly mistakes are, or what is the cost function***

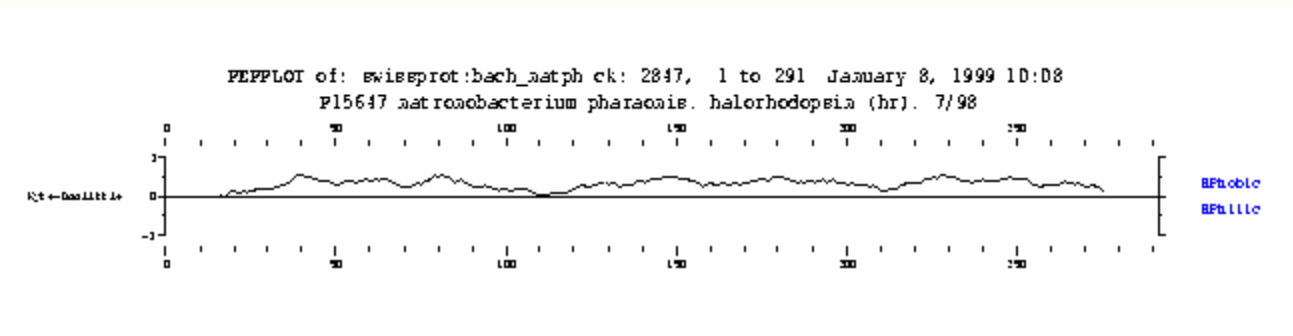
- ***Consider a test for a genetic disease***
  - Occurs at about 1/10,000
  - When found, parents are likely to abort the fetus
  - Is a 95% confidence (accuracy) in the test sufficient?

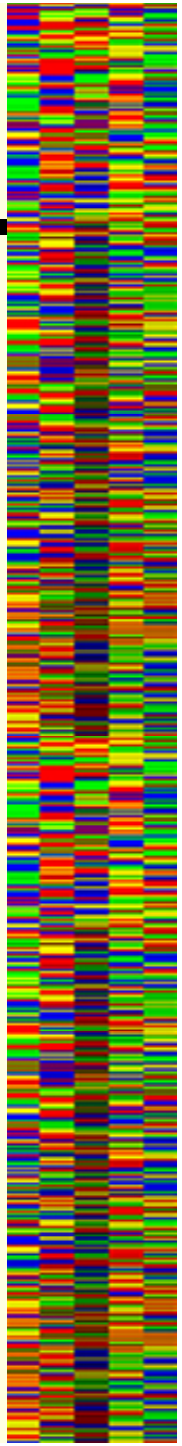


Statistics I





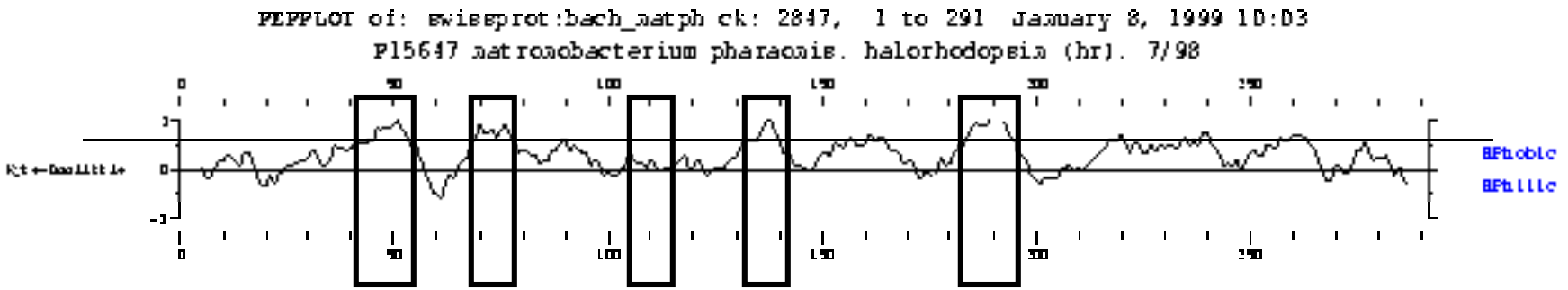




Statistics I

Classified training set

- ***A set of data for which we know what group each measurement is drawn from, and the scores for each measurement, e.g.,***
  - Positive - when there is a signal (boxed)



Statistics I

Terminology

- ***Positive/Negative - label produced by a method, e.g. in a dotplot the points are positive, non-points are negative, in the example on the previous page, the boxed region is positive, everything else is a negative***
- ***True/False - whether the labeling is correct. In a dotplot the true points are the ones that are correctly assigned as positive (homologous) or negative (not homologous). True or false depends on the threshold***
- ***Four possibilities: True positive, True Negative, False Positive, False Negative ($P^+ N^+, P^-, N^-$)***

		Correct Classification	
		Positive	**Negative**
Method Classification	**Positive**	P^+ True Positive	N^- False Positive
	Negative	P^- False Negative	N^+ True Negative

Statistics I

Terminology

- ***Positive/Negative - label produced by a method, e.g. in a dotplot the points are positive, non-points are negative, in the example on the previous page, the boxed region is positive, everything else is a negative***
- ***True/False - whether the labeling is correct. In a dotplot the true points are the ones that are correctly assigned as positive (homologous) or negative (not homologous). True or false depends on the threshold***
- ***Four possibilities: True positive, True Negative, False Positive, False Negative ($P^+ N^+, P^-, N^-$)***

		Correct Classification	
		Positive	**Negative**
Method Classification	**Positive**	P^+ True Positive	N^- False Positive
	Negative	P^- False Negative	N^+ True Negative

Statistics I

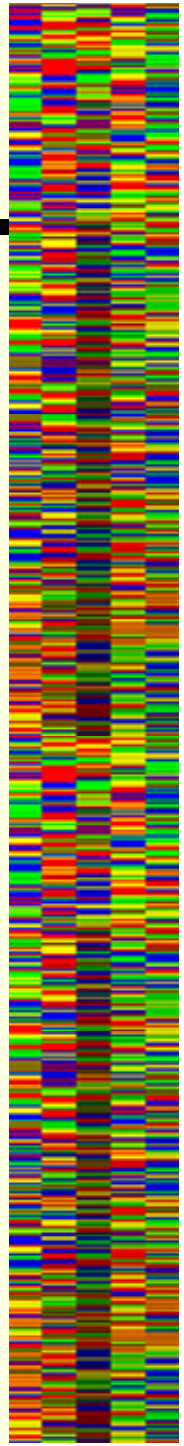
Terminology

		Correct Classification	
		Positive	**Negative**
Method Classification	**Positive**	P^+ True Positive	N^- False Positive
	Negative	P^- False Negative	N^+ True Negative

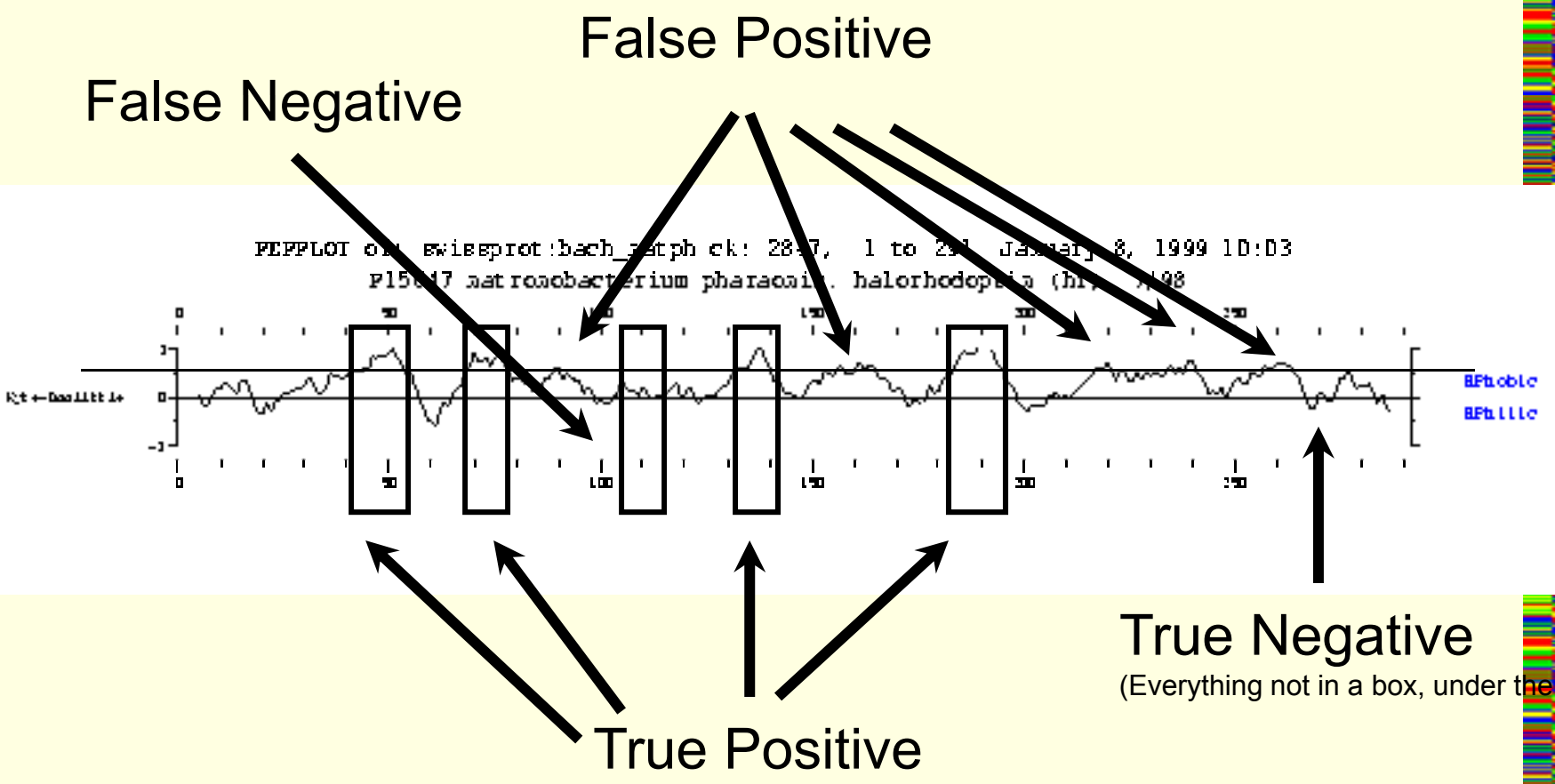
Statistics I

Sensitivity vs. Specificity

- ***Problem: maximize the number of "positive" results above the threshold while minimizing the number of "negative" results that score above the threshold***
- ***Terminology***
  - Specificity (Precision,Selectivity) = P+ / (P+ + N-)
    - Fraction of labeled positive points that are true positives
  - Sensitivity (Recall) = P+ / (P+ + P-)
    - Fraction of positives that are labeled as positive, true positive fraction
- ***Nearly always a tradeoff between sensitivity and specificity***
  - Dotplots
    - Increasing threshold (stringency) reduces numbers of both true positive and false positive windows, therefore although it is likely to improve specificity, it degrades the sensitivity



Statistics: Part I



Statistics: Part II

Is there homology?

- *We can answer, "yes," if the result is surprising when compared to unrelated sequences*

What do we mean by surprising?

- *We are surprised when an event is very unlikely to happen by chance. In this case, we are surprised when the observed level of similarity is very unlikely between unrelated sequences*

This requires a model for unrelated sequences

- *Most common choice is a random sequence model*