Biol 478/595 Intro to Bioinformatics

September					
	M 1		Labor Day		
4	W 3	MG	Database Searching		Ch. 6
5	F 5	MG	Database Searching	Hw1	
6	M 8	MG	Scoring Matrices		Ch 3 and Ch 4
7	W 10	MG	Pairwise Alignment		
8	F 12	MG	Pairwise Alignment	Hw2	
9	M 15	MG	Pairwise Alignment		Ch 9
10	W 17	MG	Genome Sequencing		
11	F 19	MG	Gene Finding/Annotation	Hw3	
12	M 22	MG	Gene Finding/Annotation		Ch. 11
13	W 24	MG	Quiz – Gene Finding/Annotation		Ch. 7
14	F 26	MG	Sequence Motifs	No Hw	
X	M 29	Both	Exam		

- No new homework this week to allow time to study for exam (or not)
- Homowork 4 is not due until 10/3
- Quiz today
- Midterm next Monday

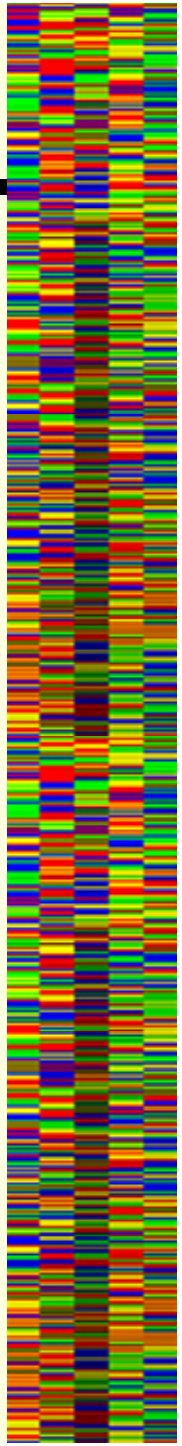
Biol478/595 Intro to Bioinformatics

Quiz – Wednesday 9/22

Covers through 9/15 (lecture 9)

- *database searching*
- *dynamic programming alignment*

- *Four-five mostly short answer questions*
- *Single sheet of notes for reference*
- *closed book, closed notes*
- *no calculators or computers.*

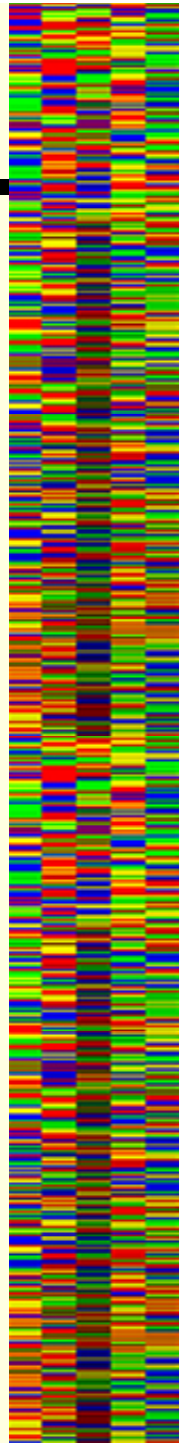


Biol478/595 Intro to Bioinformatics

Midterm exam

Monday September 29

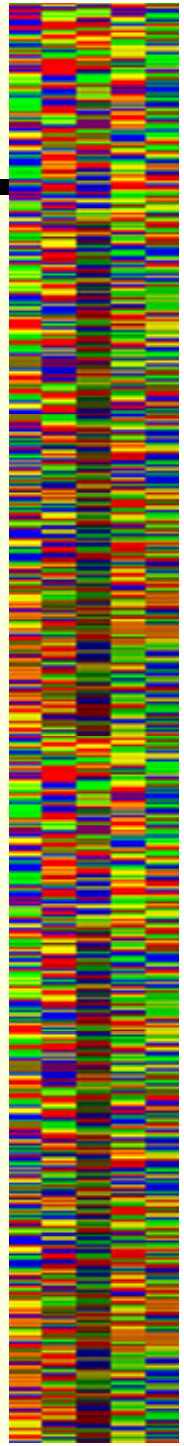
- ***Covers material through Friday 9/26***
- ***Example exams will be posted***
- ***Location TBA***



Genomics - Gene Modeling

Search by Content

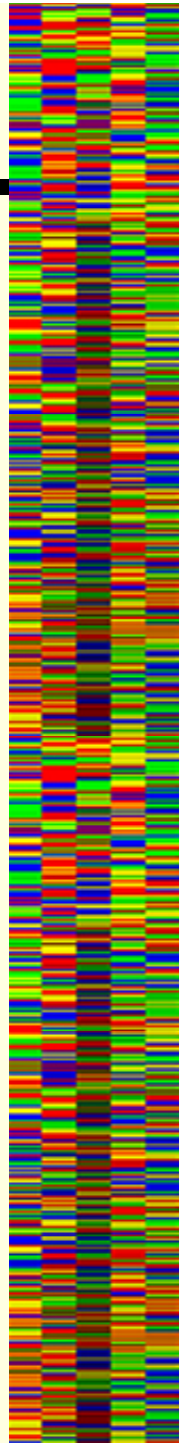
- ***Codon usage/codon preference***
  - Organisms do not use the codons for each amino acid equally, this is called *codon preference*. The primary reason appears to be that the pools of isoaccepting tRNAs differ depending on gene number and expression.
  - Rare codons (corresponding to low-level tRNAs) may also be used as a regulatory mechanism.
  - Highly expressed genes tend to use only codons corresponding to the most abundant tRNAs. This effect is stronger in prokaryotes. More weakly expressed genes use closer to equal usage and are therefore harder to detect.
  - The overall codon usage (number used) of each codon is determined by the codon preference and the amino acid preference
  - In eukaryotes, codon usage/preference may be cell, developmental stage, or tissue specific



Basic Bayesian Statistics

- ***Conditional probabilities***
  - What is the probability of x if y (or x given we know y)
    P(x|y)
- ***Predictions are usually conditional probabilities***
  - What is the probability of having tuberculosis given a positive TB test
    P(TB|test)
  - What is the probability that this DNA is an exon given GC =0.64
    P(Exon|GC=0.64)

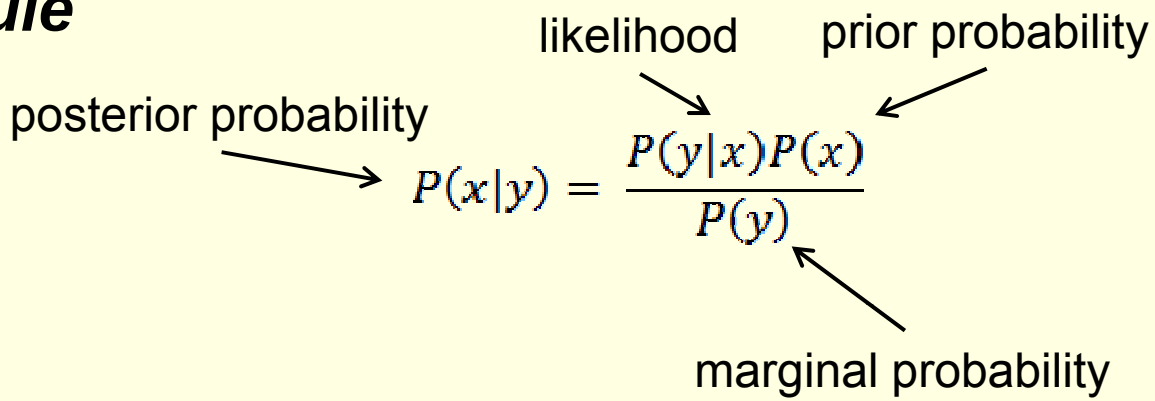
- ***Many times we cannot calculate the (conditional) probability that we want, Bayes' rule allows us to substitute other information***



Genomics - Gene Modeling

Basic Bayesian Statistics

Bayes' rule



$$P(\text{model}|\text{data}) - \frac{P(\text{data}|\text{model})P(\text{model})}{P(\text{data})}$$

- ***Prior probability is what we know in advance, e.g., probability of base A, probability of having TB. Usually this is some population average.***
- ***X is usually a hypothesis based model for which we can calculate the likelihood of different outcomes***

Basic Bayesian Statistics

- *marginal probability of y, P(y)*

$$P(y) = \sum_{x} P(y|x)P(x)$$

$$P(\text{positivetest}) = \sum_{\text{TB}\in(+,-)} P(\text{positivetest}|\text{TB})P(\text{TB})$$

Genomics - Gene Modeling

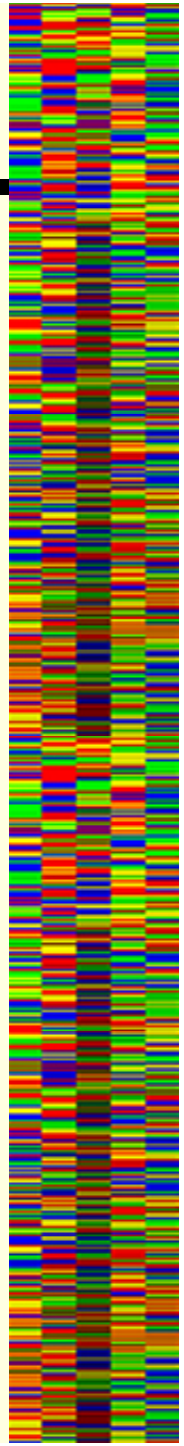
Basic Bayesian Statistics – Bayes' rule

$$(\mathrm{TB}|\mathrm{positivetest}) = \frac{P(\mathrm{positivetest}|\mathrm{TB})P(\mathrm{TB})}{P(\mathrm{positivetest})}$$

Lets say…

- ***P(TB) = 0.0001 = 1 in 10,000***
- ***P(positivetest|TB) = 0.99***
- ***P(positivetest) = 0.001 = 1 in 1000***

P(TB|positivetest) = 0.0099 ~ 1%



Genomics - Gene Modeling

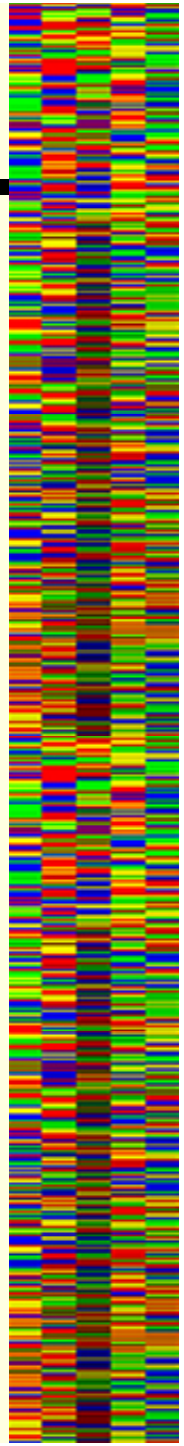
Basic Bayesian Statistics – Bayes' rule

- *Bayes' Rule is at the heart of much predictive software*
- *In the simplest example, we can simply compare two models, and reduce it to a log-odds ratio*
- *For instance, the probability that two proteins are homologous given the comparison of amino acid I and j, A_{ij}*

$$P(\text{homology}|A_{ij}) \qquad P(\text{not\_homology}|A_{ij})$$

$$\log\frac{P(\text{homology}|A_{ij})}{P(\text{not\_homology}|A_{ij})}$$

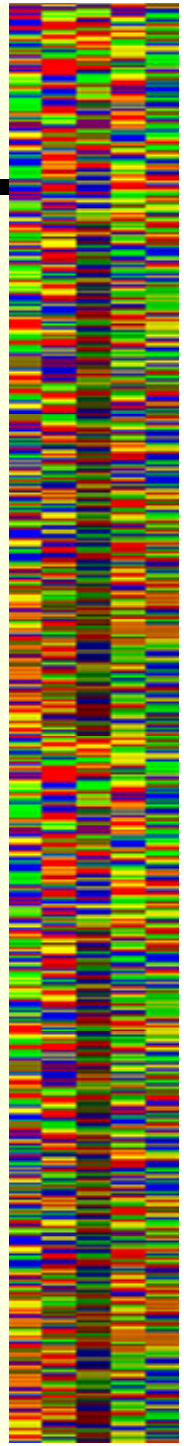
- *this is the log-odds ratio we saw before with PAM and BLOSUM*



Genomics - Gene Modeling

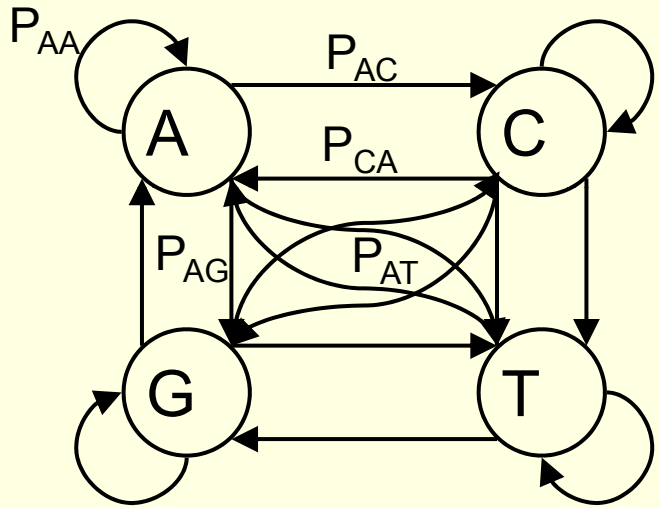
Search by Content

- ***Complex overlapping constraints can be captured by a Markov model***
- ***Look at a certain size DNA word, and determine characteristic frequencies in***
  - Exons (frame 1, 2, and 3)
  - Introns
  - Intergenic regions
  - Reverse strand of above three regions
- ***Compare to decide which model is the best fit.***
- ***Average over a fairly long window (50 – 200 bp)***



Genomics - Gene Modeling

Markov Models



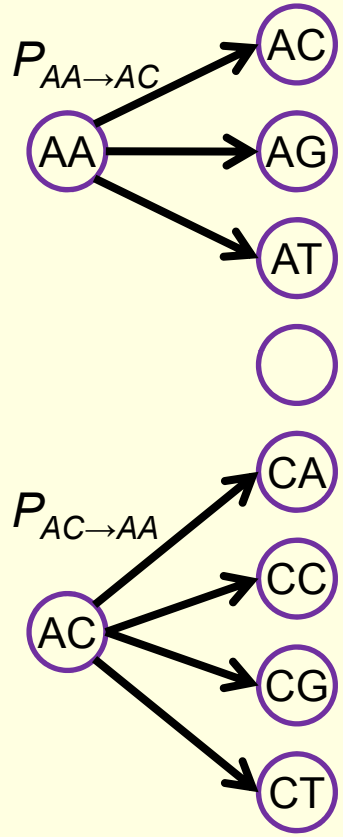
1st order Markov model
next base depends only
on the current base

- Each transition has a probability which is determined by counting how often it is seen in a large training dataset
- pseudocount must be used for zeroes

Genomics - Gene Modeling

Markov model (2^{nd} order)

- ***16 two letter words, 256 transitions***
- ***6^{th} order model, 4096 words, $1.7x10^7$ transitions***

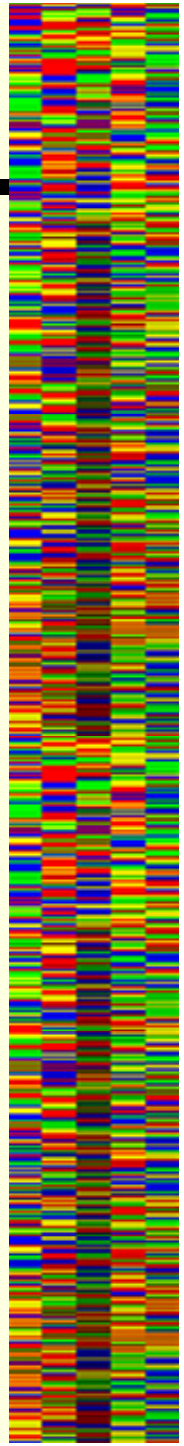


… for all possible word transitionss

Genomics - Gene Modeling

Markov models

- ***A Markov chain is a model for stochastic generation of sequential phenomena***
- ***Every position in a chain is equivalent***
- ***The order of the Markov chain is the number of previous positions on which the current position depends***
  - in nucleic acid sequence, 0-order is mononucleotide, 1st-order is dinucleotide, 2nd-order is trinucleotide, etc.
- ***The model parameters are the frequencies of the elements at each position (possibly as a function of preceding elements)***



Genomics - Gene Modeling

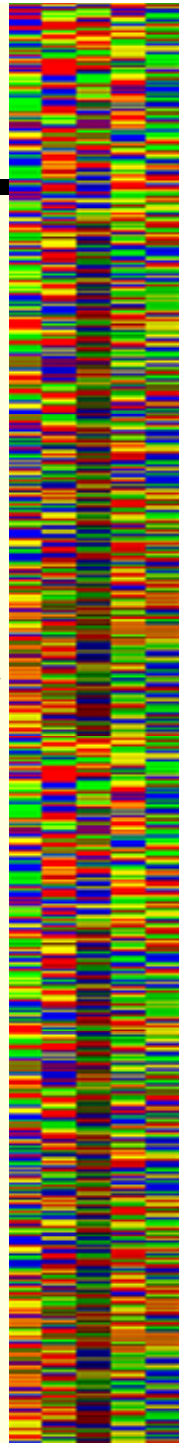
Markov models

for a sequence, *S* $S = s_1 s_2 s_3 s_4 \cdots$

$$P_0(s) = \mathrm{p}(s_1) \cdot \mathrm{p}(s_2) \cdot \mathrm{p}(s_3) \cdots = \prod_{i=1}^{N} \mathrm{p}(s_i)$$ zero order

$$P_1(s) = \mathrm{p}(s_1) \cdot \mathrm{p}(s_2 \mid s_1) \cdot \mathrm{p}(s_3 \mid s_2) \cdots = \mathrm{p}(s_1) \cdot \prod_{i=2}^{N} \mathrm{p}(s_i \mid s_{i-1})$$ 1st order

$$P_2(s) = \mathrm{p}(s_1 s_2) \cdot \mathrm{p}(s_3 \mid s_1 s_2) \cdot \mathrm{p}(s_4 \mid s_2 s_3) \cdots = \mathrm{p}(s_1 s_2) \cdot \prod_{i=3}^{N} \mathrm{p}(s_i \mid s_{i-2} s_{i-1})$$ 2nd order

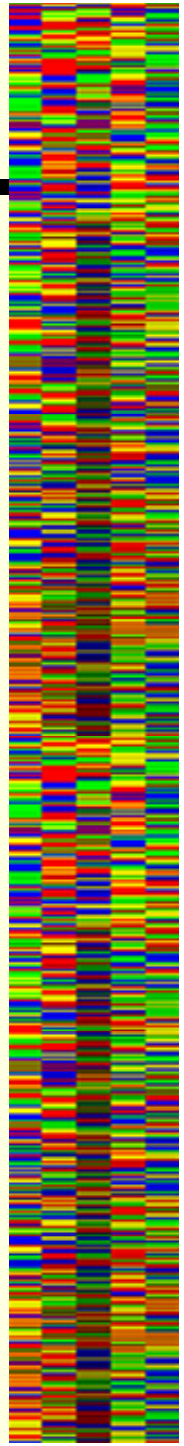


Genomics - Gene Modeling

Markov models

- ***Hexamer in frame statistic***
  - 5^{th} order Markov model
  - Given the last five bases, what is the next base.
- ***Score is sum of log-odds with appropriate coding frame as observed and non-coding as background model***

$$Score = \log (f_0 / f_n) + \log(f_1 / f_n) + \log(f_2 / f_n) + \log(f_0 / f_n)$$

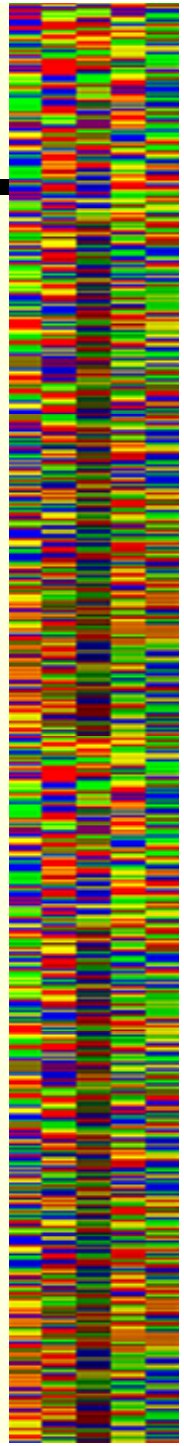


Genomics - Gene Modeling

Search by Content

- *Hexamer in frame statistic - fifth order markov model, i.e. given the last five bases, what is the next base.*
- *Four models are used : coding frame 0, coding frame 1, coding frame 2, non-coding.*
- *Score is sum of log-odds with appropriate coding frame as observed and non-coding as background model, e.g.,*

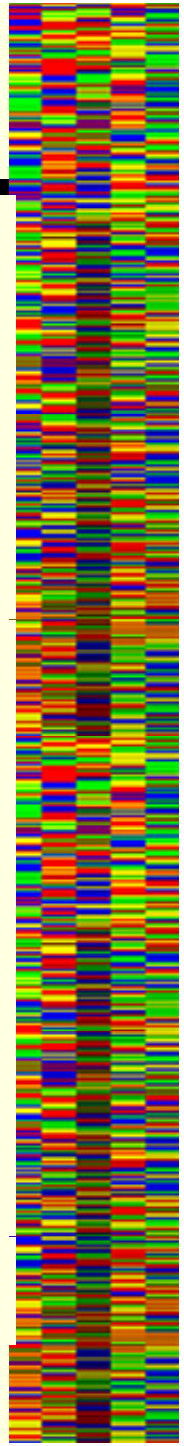
$$Score = \log (f_0 / f_n) + \log(f_1 / f_n) + \log(f_2 / f_n) + \log(f_0 / f_n)$$



Genomics - Gene Modeling

Search by Content

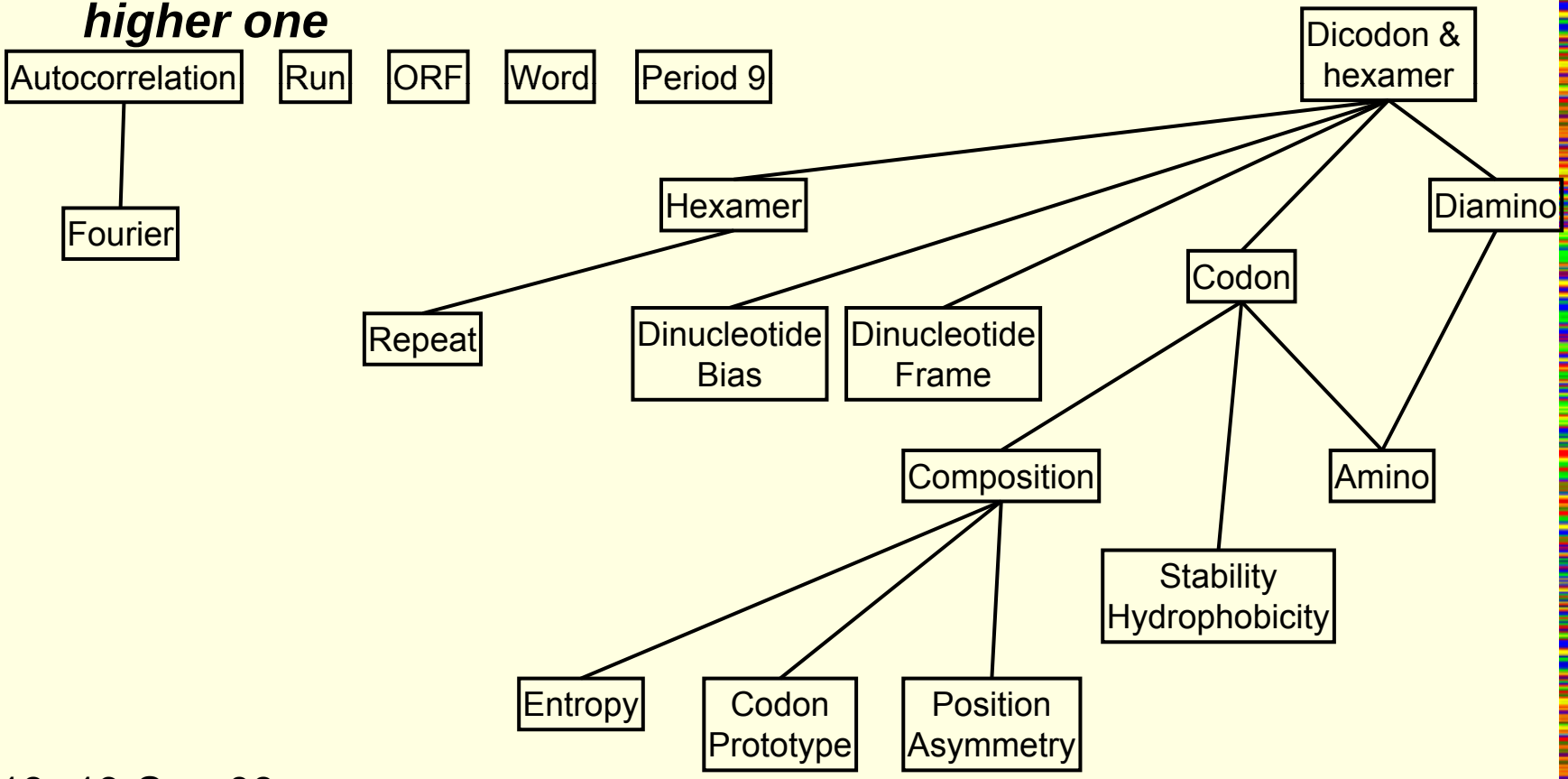
- ***Other methods (see Fickett and Tung, "Assessment of protein coding measures", Nucleic Acids Res. 20, 6441-6450, 1992)***
  - Runs (R&Y stronger in noncoding, W&S in coding)
  - N-word counts, most commonly hexamer
  - Stability (tendency to mutate to same amino acid residue)
  - Other base asymmetry measures
  - Periodicity, such as period 9
  - Global patterns (GC content, CpG islands)
  - Open reading frame (genes have longer ORF)
  - Exon length
  - Intron length



Genomics - Gene Modeling

Search by Content

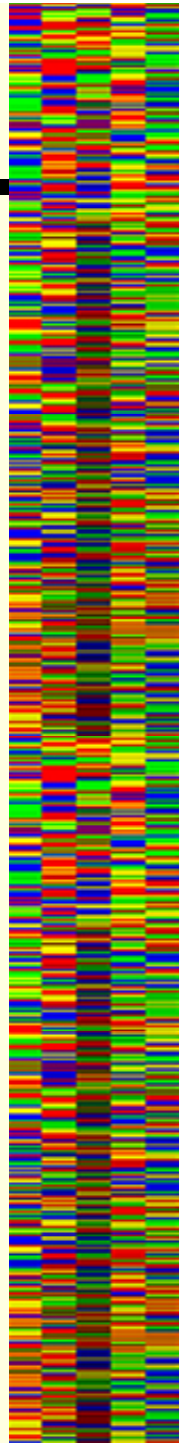
- ***Relationship between coding measures (from Fickett and Tung), lower function connected by lines is a function of the higher one***



Genomics - Gene Modeling

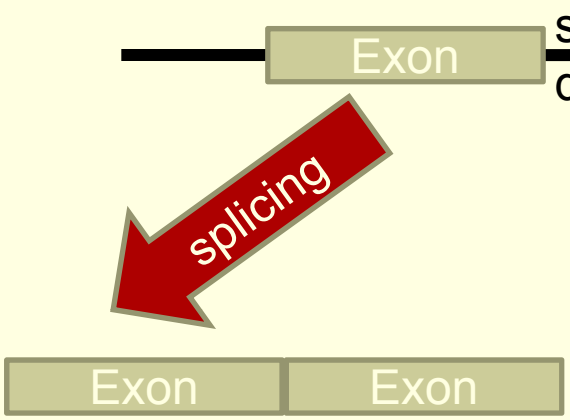
Extrinsic methods (search by signal)

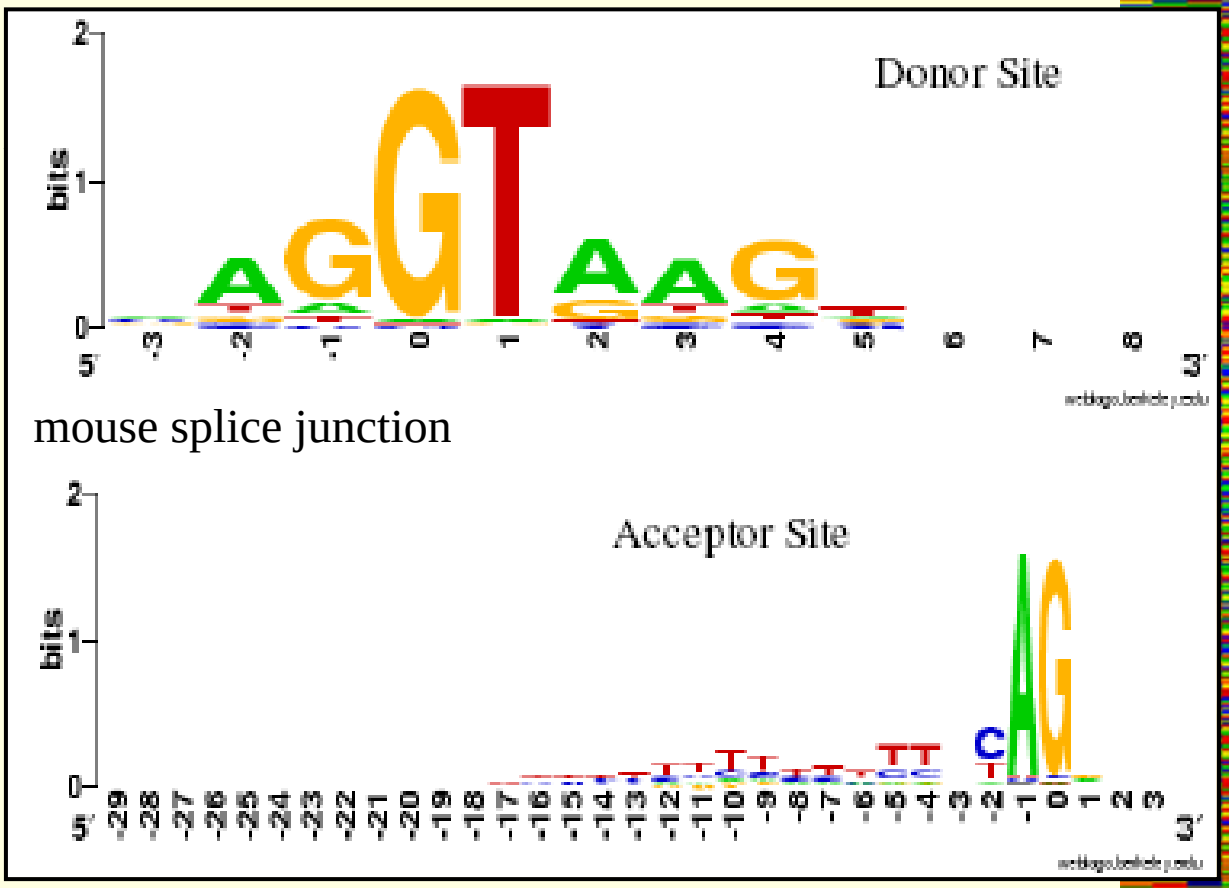
- ***Try to identify sequence signals relevant to the presence, absence, frame, and content of genes***
- ***Signals***
  - promoters
  - terminators
  - polyA sites
  - Cap signals
  - splice junctions
- ***Sequence matches***
  - expressed genes (ESTs)
  - protein databases
  - closely related genomes (translated DNA vs translated DNA)



Genomics - Gene Modeling

Splice signals





Genomics - Gene Modeling

Things to Remember about gene modeling

- *It is, in general, organism-specific*
- *It works best on genes that are reasonably similar to something seen previously*
- *It finds protein coding regions far better than non-coding regions*
- *In the absence of external (direct) information, alternative forms will not be identified*
- *It is imperfect! (It's biology, after all…)*