

Probabilistic sequence modeling frequency and profiles

Genome and genes

- Genome: an organism's genetic material
- Gene: discrete units of hereditary information located on the chromosomes and consisting of DNA

Gene prediction: computational challenge

aatgcatgCGGctatgctaataatgcatgCGGctatgctaagctGGGatccgatgacaa
tgcatgCGGctatgctaataatgcatgCGGctatgcaagctGGGatccgatgactatgc
taagctGGGatccgatgacaataatgcatgCGGctatgctaataatgaatGGTcttGGGatt
taccttGgaatgctaagctGGGatccgatgacaataatgcatgCGGctatgctaataatgaa
tGGTcttGGGatttaccttGgaatgctaataatgcatgCGGctatgctaagctGGGat
tccgatgacaataatgcatgCGGctatgctaataatgcatgCGGctatgcaagctGGGatcc
gatgactatgctaagctgCGGctatgctaataatgcatgCGGctatgctaagctGGGat
ccgatgacaataatgcatgCGGctatgctaataatgcatgCGGctatgcaagctGGGatcct
gCGGctatgctaataatgaatGGTcttGGGatttaccttGgaatgctaagctGGGatcc
gatgacaataatgcatgCGGctatgctaataatgaatGGTcttGGGatttaccttGgaatg
ctaataatgcatgCGGctatgctaagctGGGaatgcatgCGGctatgctaagctGGGat
tccgatgacaataatgcatgCGGctatgctaataatgcatgCGGctatgcaagctGGGatcc
gatgactatgctaagctgCGGctatgctaataatgcatgCGGctatgctaagctcatgC
GGctatgctaagctGGGaatgcatgCGGctatgctaagctGGGatccgatgacaata
gcatgCGGctatgctaataatgcatgCGGctatgcaagctGGGatccgatgactatgct
aagctgCGGctatgctaataatgcatgCGGctatgctaagctcGGctatgctaataatga
aGGTcttGGGatttaccttGgaatgctaagctGGGatccgatgacaataatgcatgCGGc
tatgctaataatgaatGGTcttGGGatttaccttGgaatgctaataatgcatgCGGctat
gctaagctGGGaatgcatgCGGctatgctaagctGGGatccgatgacaataatgcatgC
GGctatgctaataatgcatgCGGctatgcaagctGGGatccgatgactatgctaagctg
CGGctatgctaataatgcatgCGGctatgctaagctcatgCGG

Gene prediction: computational challenge

aatgcatgicggctatgctaatagcatgicggctatgctaagctgggatccgatgacaa
tgcatgicggctatgctaatagcatgicggctatgcaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgicggctatgctaataaatgggtcttgggatt
taccttggaaatgctaagctgggatccgatgacaatgcatgicggctatgctaataa
tgggtcttgggatttaccttggaaatgctaatagcatgicggctatgctaagctggga
tccgatgacaatgcatgicggctatgctaatagcatgicggctatgcaagctgggatcc
gatgactatgctaagctgicggctatgctaatagcatgicggctatgctaagctgggat
ccgatgacaatgcatgicggctatgctaatagcatgicggctatgcaagctgggatcc
gicggctatgctaataaatgggtcttgggatttaccttggaaatgctaagctgggatcc
gatgacaatgcatgicggctatgctaataaatgggtcttgggatttaccttggaaat
gctaatagcatgicggctatgctaagctgggaatgcatgicggctatgctaagctggga
tccgatgacaatgcatgicggctatgctaatagcatgicggctatgcaagctgggatcc
gatgactatgctaagctgicggctatgctaatagcatgicggctatgctaagctcatgc
ggctatgctaagctgggaatgcatgicggctatgctaagctgggatccgatgacaat
gcatgicggctatgctaatagcatgicggctatgcaagctgggatccgatgactatgct
aagctgicggctatgctaatagcatgicggctatgctaagctcggctatgctaataa
tgggtcttgggatttaccttggaaatgctaagctgggatccgatgacaatgcatgicggc
tatgctaataaatgggtcttgggatttaccttggaaatgctaatagcatgicggctat
gctaagctgggaatgcatgicggctatgctaagctgggatccgatgacaatgcatgic
ggctatgctaatagcatgicggctatgcaagctgggatccgatgactatgctaagctg
cggctatgctaatagcatgicggctatgctaagctcatgicgg

Gene prediction: computational challenge

aatgcatgicggctatgctaatagcatgicggctatgctaagctgggatccgatgacaa
tgcatgicggctatgctaatagcatgicggctatgcaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgicggctatgctaataaatgggtcttgggatt
taccttggaaatgctaagctgggatccgatgacaatgcatgicggctatgctaataa
tgggtcttgggatttaccttggaaatgctaatagcatgicggctatgctaagctggga
tccgatgacaatgcatgicggctatgctaatagcatgicggctatgcaagctgggatcc
gatgactatgctaagctgicggctatgctaatagcatgicggctatgctaagctgggat
ccgatgacaatgcatgicggctatgctaatagcatgicggctatgcaagctgggatcc
gicggctatgctaataaatgggtcttgggatttaccttggaaatgctaagctgggatcc
gatgacaatgcatgicggctatgctaataaatgggtcttgggatttaccttggaaat
gctaatagcatgicggctatgctaatagcatgicggctatgctaagctggga
tccgatgacaatgcatgicggctatgctaatagcatgicggctatgcaagctgggatcc
gatgactatgctaagctgicggctatgctaatagcatgicggctatgctaagctcatgc
ggctatgctaagctgggaaatgcatgicggctatgctaagctgggatccgatgacaat
gcatgicggctatgctaatagcatgicggctatgcaagctgggatccgatgactatgct
aagctgicggctatgctaatagcatgicggctatgctaagctcggctatgctaataa
tgggtcttgggatttaccttggaaatgctaagctgggatccgatgacaatgcatgicggc
tatgctaataaatgggtcttgggatttaccttggaaatgctaatagcatgicggctat
gctaagctgggaaatgcatgicggctatgctaagctgggatccgatgacaatgcatgc
ggctatgctaatagcatgicggctatgcaagctgggatccgatgactatgctaagctg
cggctatgctaatagcatgicggctatgctaagctcatgicgg

Gene!

Gene prediction: computational challenge

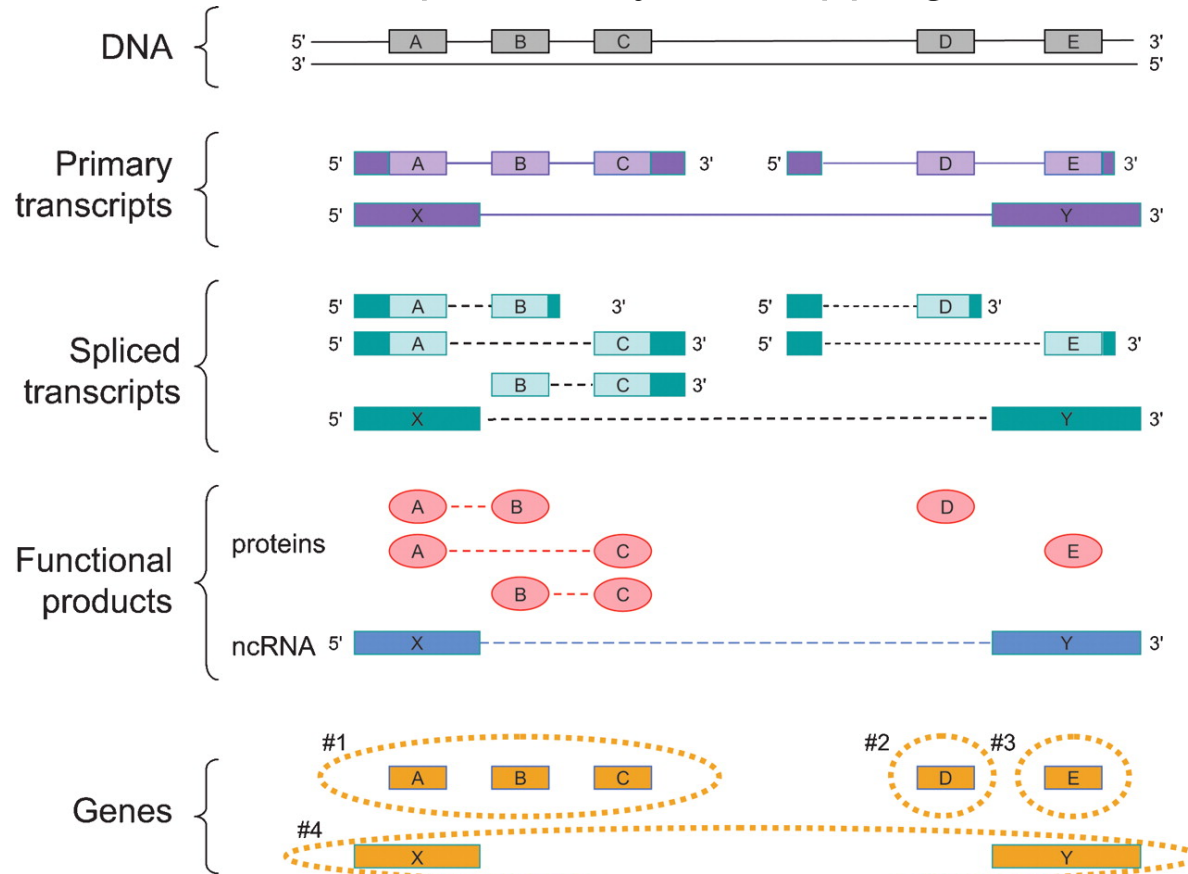
- Gene: A sequence of nucleotides coding for protein
- Gene Prediction Problem: Determine the beginning and end positions of genes in a genome

What's a gene

- What is a gene, post-ENCODE?
 - Gerstein et al. Genome Res. 2007 17: 669-681
 - ENCODE consortium: characterization of 1% of the human genome by experimental and computational techniques
- Definitions:
 - Definition 1970s–1980s: Gene as open reading frame (ORF) sequence pattern
 - Definition 1990s–2000s: Annotated genomic entity, enumerated in the databanks
 - The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products

Post-ENCODE definition

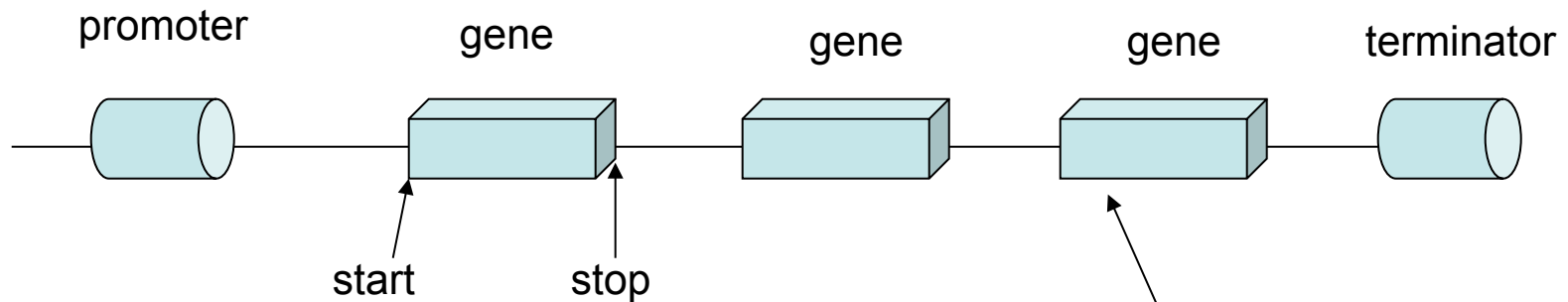
The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products



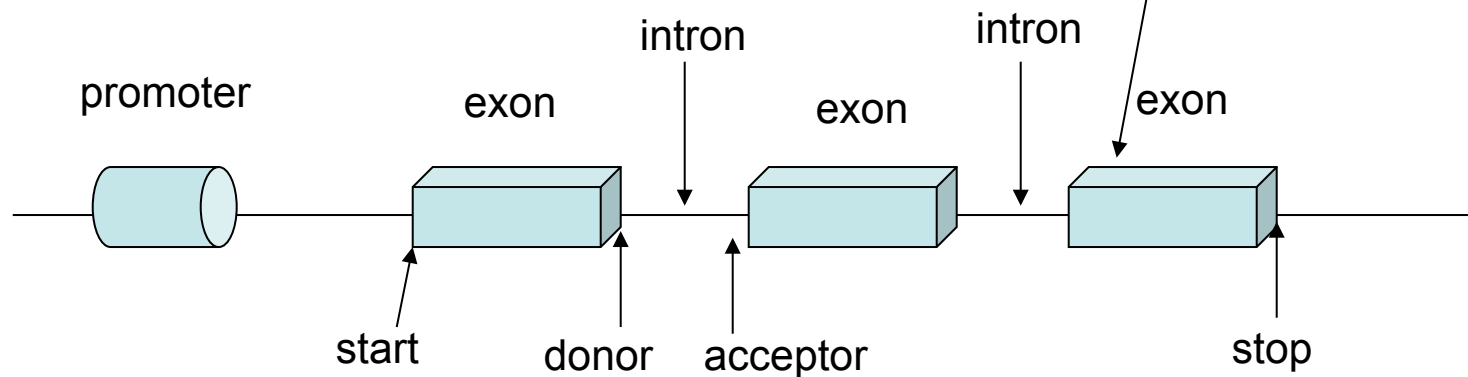
Mark B. Gerstein et al. *Genome Res.* 2007; 17: 669-681

Can we still do gene prediction?

- Prokaryotic genes



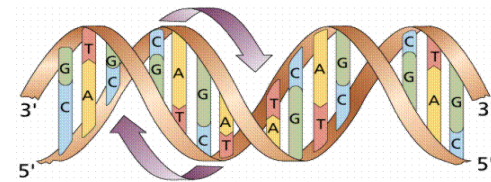
- Eukaryotic genes



Open reading frame (ORF)

A simple model for gene prediction: frequency-based DNA modeling

- DNA is a double stranded molecule
 - G-C pair → strong
 - A-T pair → weak



- Coding regions often have higher GC content than non-coding regions

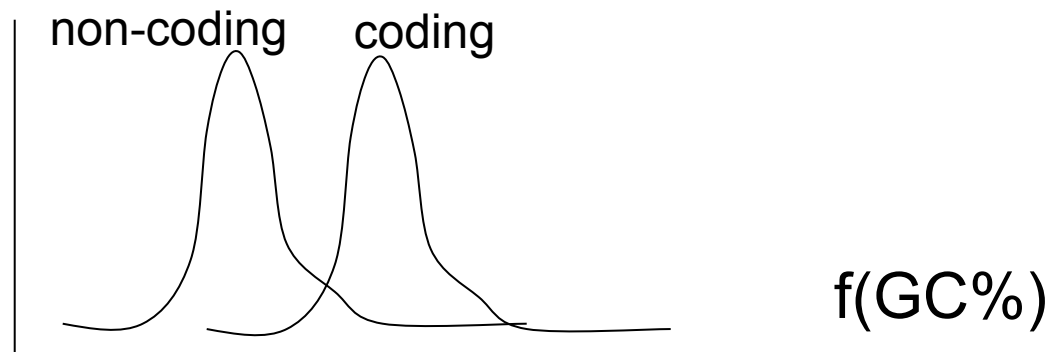
Frequency-based DNA modeling of coding vs. non-coding regions

- To predict coding regions in an organism (e.g. human), collect a set of known coding and non-coding DNA sequences from this organism (**training set**)
- Compute the frequency distribution of GC pairs in coding and non-coding regions, respectively: $f(\text{GC}\%|c)$, $f(\text{GC}\%|nc)$

An example from zygomycete *Phycomyces blakesleeanus*

Table 1. GC content of *Phycomyces* DNA.

DNA type ^a	GC	Sample size ^b	Sample length (bp) ^c
Protein-coding DNA	48 ± 0.6	56	29 862
Total non-coding DNA	30 ± 1.0	49	13 088
Introns	29 ± 1.5	28	2837
5'-end	34 ± 2.1	10	6182
3'-end	30 ± 1.1	11	4069



Model comparison

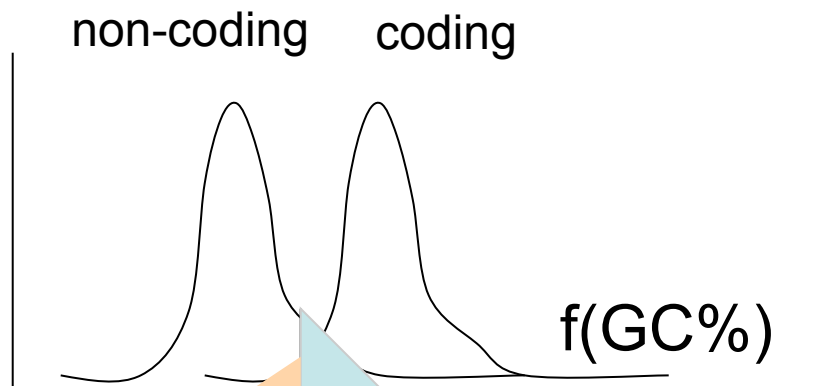
- Two models: coding vs. non-coding
- Given a DNA sequence, its likelihood of being a coding sequence, based on its GC content (GC%)

$$P(c|GC\%) = \frac{P(GC\%|c)P(c)}{P(GC\%)} = \frac{P(GC\%|c)P(c)}{P(GC\%|c)P(c) + P(GC\%|nc)P(nc)}$$

$$P(nc|GC\%) = \frac{P(GC\%|nc)P(nc)}{P(GC\%)} = \frac{P(GC\%|nc)P(nc)}{P(GC\%|c)P(c) + P(GC\%|nc)P(nc)}$$

$$P(GC\%|c) = \int_0^x f(x < GC\%|c)$$

$$P(GC\%|nc) = \int_x^1 f(x > GC\%|nc)$$



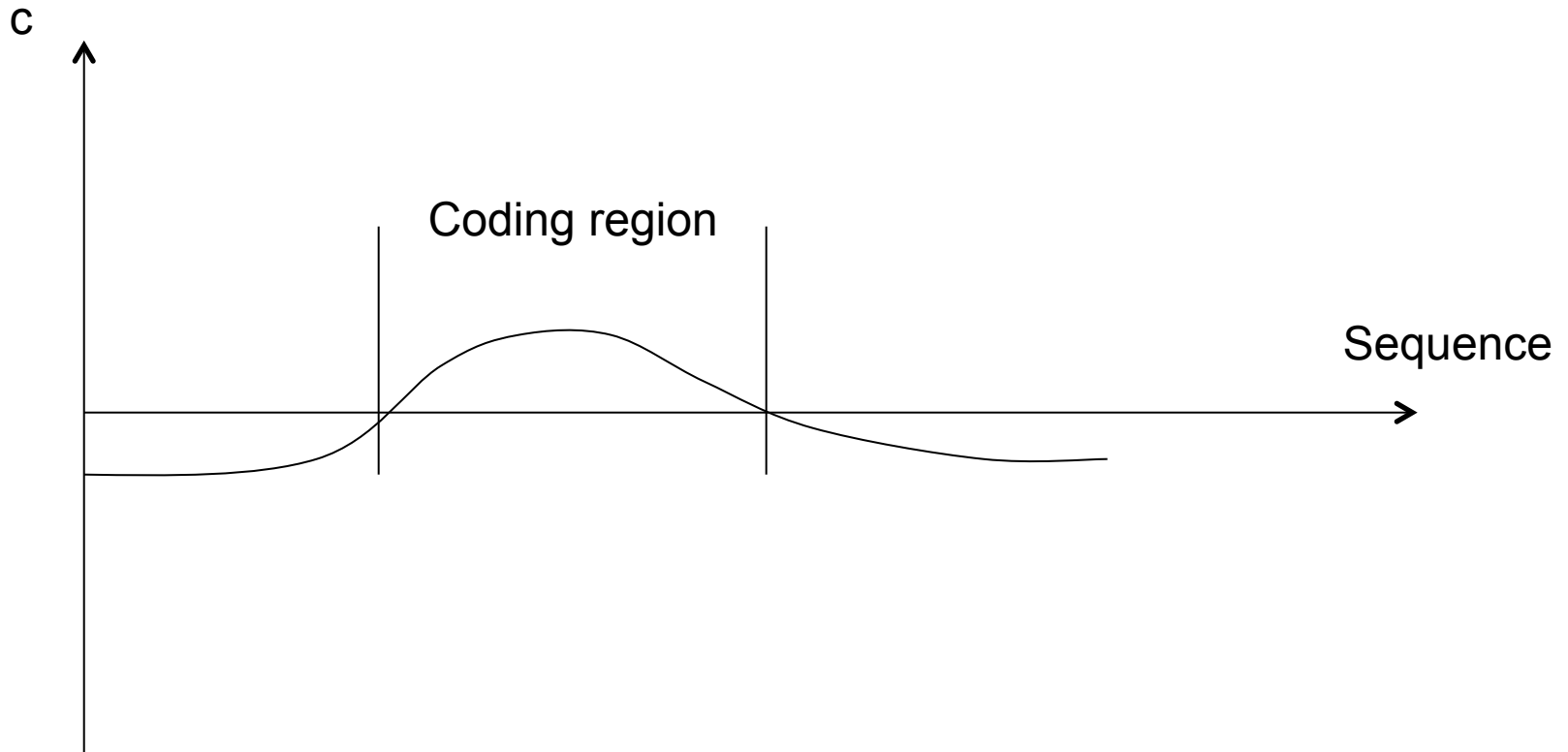
Assume $P(c)=P(nc)$,

$$R = \log\left(\frac{P(GC\%|c)}{P(GC\%|nc)}\right)$$

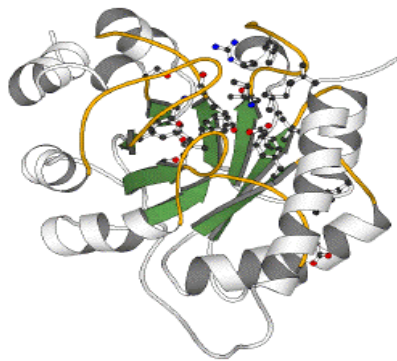
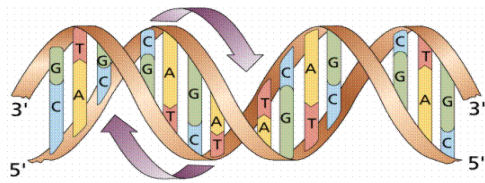
Sliding window approach

aatgcatg|cggctatgcta|atgcatg|cggctatgcta|agctggg|atccgatgaca|atgcatg|cggctatg
cta|atgcatg|cggctatgca|agctggg|atccgatgact|atgcta|agctggg|atccgatgaca|atgcatg|c
gg|ctatgcta|aatgaa|tgg|tctggg|attac|cttgg|aatgcta|agctggg|atccgatgaca|atgcatg|cgg
ctatgcta|aatgaa|tgg|tctggg|attac|cttgg|aatgcta|atgcatg|cggctatgcta|agctggg|atccg
atgaca|atgcatg|cggctatgcta|atgcatg|cggctatgca|agctggg|atccgatgact|atgcta|agctg
cgg|ctatgcta|atgcatg|cggctatgcta|agctggg|atccgatgaca|atgcatg|cggctatgcta|atgcat
g|cggctatgca|agctggg|atcctg|cggctatgcta|aatgaa|tgg|tctggg|attac|cttgg|aatgcta|agct
ggg|atccgatgaca|atgcatg|cggctatgcta|aatgaa|tgg|tctggg|attac|cttgg|aatgcta|atgca
t|g|cggctatgcta|agctggg|aatgcatg|cggctatgcta|agctggg|atccgatgaca|atgcatg|cggcta
t|gcta|atgcatg|cggctatgca|agctggg|atccgatgact|atgcta|agctg|cggctatgcta|atgcatg|c
g|ctatgcta|agctcatg|cggctatgcta|agctggg|aatgcatg|cggctatgcta|agctggg|atccgatg
ca|atgcatg|cggctatgcta|atgcatg|cggctatgca|agctggg|atccgatgact|atgcta|agctg|cggc
t|atgcta|atgcatg|cggctatgcta|agctcgg|ctatgcta|aatgaa|tgg|tctggg|attac|cttgg|aatgcta
a|gctggg|atccgatgaca|atgcatg|cggctatgcta|aatgaa|tgg|tctggg|attac|cttgg|aatgctaa
t|gcatg|cggctatgcta|agctggg|aatgcatg|cggctatgcta|agctggg|atccgatgaca|atgcatg|c
g|ctatgcta|atgcatg|cggctatgca|agctggg|atccgatgact|atgcta|agctg|cggctatgcta|atgca
t|g|cggctatgcta|agctcatg|cgg

Sliding window approach



A more complicated model: codon usages



DNA

transcription

RNA

translation

Protein

CCTGAGCCAAC TATTGATGAA



CCUGAGCCAACUAUUGAUGAA



PEPTIDE

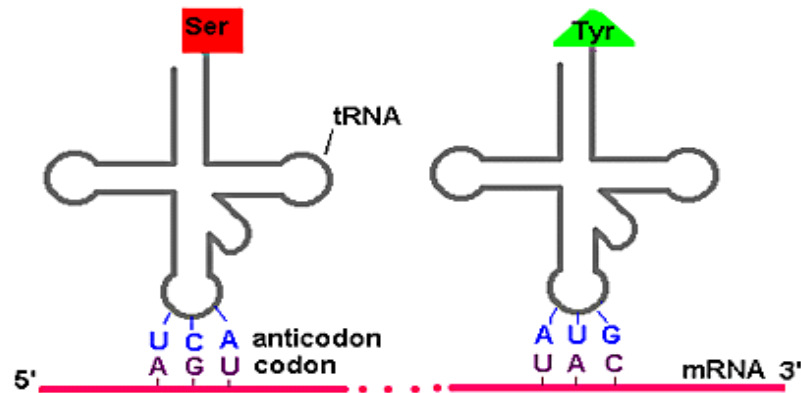
Translating nucleotides into amino acids

- Codon: 3 consecutive nucleotides
- $4^3 = 64$ possible codons
- Genetic code is degenerative and redundant
 - Includes start and stop codons
 - An amino acid may be coded by more than one codon (codon degeneracy)

Codons

- In 1961 Sydney Brenner and Francis Crick discovered **frameshift mutations**
- Systematically deleted nucleotides from DNA
 - Single and double deletions dramatically altered protein product
 - Effects of triple deletions were minor
 - Conclusion: every triplet of nucleotides, each ***codon***, codes for exactly one amino acid in a protein

Genetic code and stop codons



UAA, UAG and UGA correspond to 3 Stop codons that (together with Start codon ATG) delineate Open Reading Frames

		2nd base in codon				
		U	C	A	G	
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G
						3rd base in codon

The Genetic Code

Six frames in a DNA sequence

CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
CTGCAGACGAAACCTCTT**ATG**TAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC

→

CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG

←

GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG

- stop codons – TAA, TAG, TGA
- start codons - ATG

Testing reading frames

- Create a 64-element hash table and count the frequencies of codons in a reading frame;
- Amino acids typically have more than one codon, but in nature certain codons are more in use;
- Uneven use of the codons may characterize a coding region;

Codon usage in Human genome

	U	C	A	G
U	UUU Phe 57	UCU Ser 16	UAU Tyr 58	UGU Cys 45
	UUC Phe 43	UCC Ser 15	UAC Tyr 42	UGC Cys 55
	UUA Leu 13	UCA Ser 13	UAA Stp 62	UGA Stp 30
	UUG Leu 13	UCG Ser 15	UAG Stp 8	UGG Trp 100
C	CUU Leu 11	CCU Pro 17	CAU His 57	CGU Arg 37
	CUC Leu 10	CCC Pro 17	CAC His 43	CGC Arg 38
	CUA Leu 4	CCA Pro 20	CAA Gln 45	CGA Arg 7
	CUG Leu 49	CCG Pro 51	CAG Gln 66	CGG Arg 10
A	AUU Ile 50	ACU Thr 18	AAU Asn 46	AGU Ser 15
	AUC Ile 41	ACC Thr 42	AAC Asn 54	AGC Ser 26
	AUA Ile 9	ACA Thr 15	AAA Lys 75	AGA Arg 5
	AUG Met 100	ACG Thr 26	AAG Lys 25	AGG Arg 3
G	GUU Val 27	GCU Ala 17	GAU Asp 63	GGU Gly 34
	GUC Val 21	GCC Ala 27	GAC Asp 37	GGC Gly 39
	GUA Val 16	GCA Ala 22	GAA Glu 68	GGA Gly 12
	GUG Val 36	GCG Ala 34	GAG Glu 32	GGG Gly 15

Codon usage in Mouse genome

<u>AA</u>	<u>codon</u>	<u>/1000</u>	<u>frac</u>	<u>AA</u>	<u>codon</u>	<u>/1000</u>	<u>frac</u>
Ser	TCG	4.31	0.05	Leu	CTG	39.95	0.40
Ser	TCA	11.44	0.14	Leu	CTA	7.89	0.08
Ser	TCT	15.70	0.19	Leu	CTT	12.97	0.13
Ser	TCC	17.92	0.22	Leu	CTC	20.04	0.20
Ser	AGT	12.25	0.15				
Ser	AGC	19.54	0.24	Ala	GCG	6.72	0.10
				Ala	GCA	15.80	0.23
				Ala	GCT	20.12	0.29
Pro	CCG	6.33	0.11	Ala	GCC	26.51	0.38
Pro	CCA	17.10	0.28				
Pro	CCT	18.31	0.30	Gln	CAG	34.18	0.75
Pro	CCC	18.42	0.31	Gln	CAA	11.51	0.25

Using codon frequency to find correct reading frame

Consider sequence $x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 \dots$
where x_i is a nucleotide

$$\text{let } p_1 = p_{x_1 x_2 x_3} p_{x_4 x_5 x_6} \dots$$

$$p_2 = p_{x_2 x_3 x_4} p_{x_5 x_6 x_7} \dots$$

$$p_3 = p_{x_3 x_4 x_5} p_{x_6 x_7 x_8} \dots$$

then probability that i th reading frame is the coding frame is:

$$P_i = \frac{p_i}{p_1 + p_2 + p_3}$$

slide a window along the sequence and
compute P_i

Adding the background model: gene finding

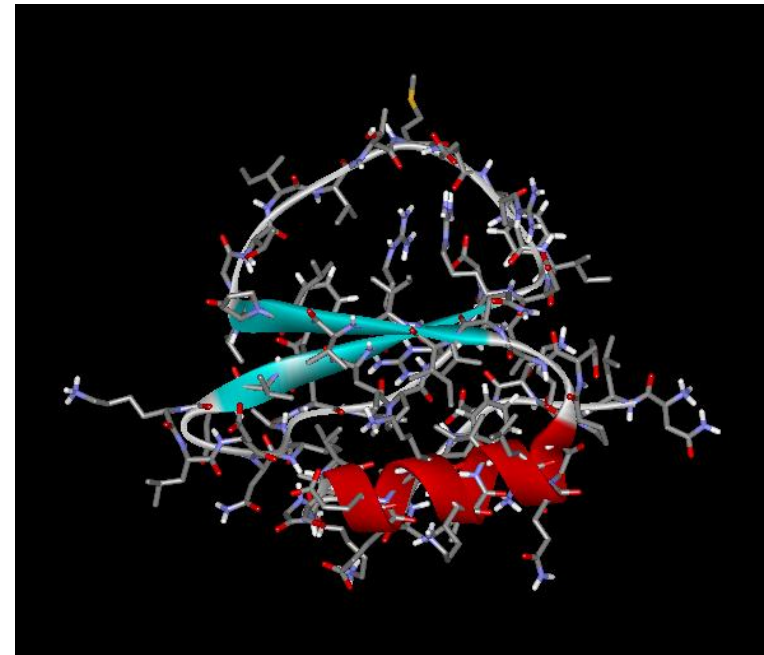
- In the previous model, we assume at least one reading frame is the codon sequence → testing reading frames, not gene finding
- Adding a background model
 - $p_0 = p_{x1}p_{x2}p_{x3}p_{x4}\dots$
 - Based on the nucleotide frequency in the non-coding sequence
 - $P_i = p_i / (p_0 + p_1 + p_2 + p_3)$
- In practice, this model should be extended to all six reading frames.

Protein secondary structure prediction

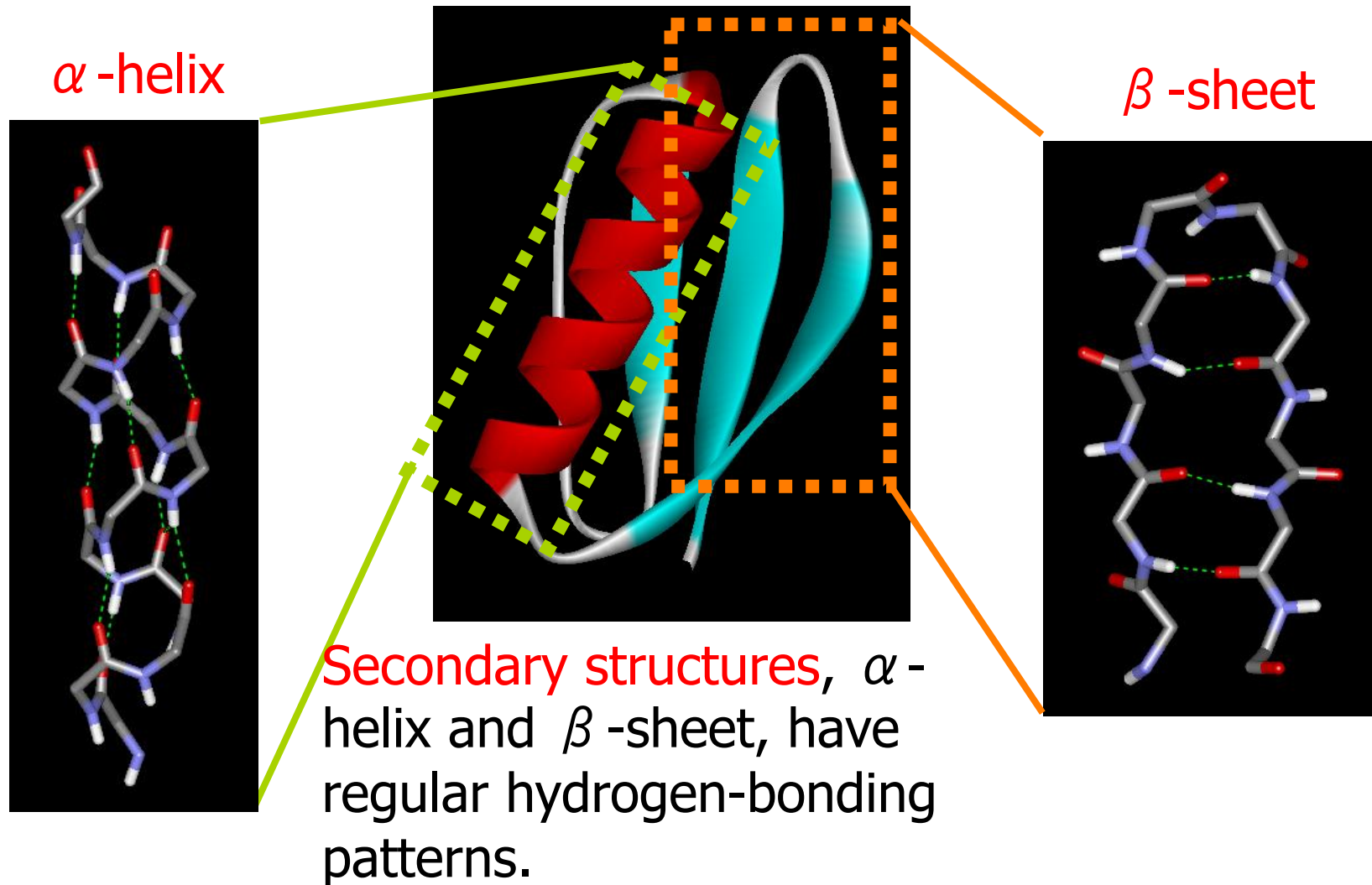
Amino acid sequence

NLKTEWPPELVGKSVEE
AKKVILQDKPEAQIIVL
PVGTVTMEYRIDRVR
LFVDKLDNIAEVPRVG

→
folding



Basic structural units of proteins: Secondary structure



Secondary structure prediction

- Given a protein sequence, secondary structure prediction aims at predicting the state of each amino acid as being either H (helix), E (extended=strand), or O (other).
- The quality of secondary structure prediction is measured with a “3-state accuracy” score, or Q_3 . Q_3 is the percent of residues that match “reality” (X-ray structure).

Chou and Fasman: a frequency model

- $P(\alpha|S) = \prod_s p(\alpha|s) = \prod_s p(\alpha|f(s))$
 - $p(\alpha|f(s)) \sim p(f(s)|\alpha) / p(f(s))$
 - Similarly for β and turn structures

Chou and Fasman: a frequency model

Amino Acid	α -Helix	β -Sheet	Turn	
Ala	1.29	0.90	0.78	Favors α -Helix
Cys	1.11	0.74	0.80	
Leu	1.30	1.02	0.59	
Met	1.47	0.97	0.39	
Glu	1.44	0.75	1.00	
Gln	1.27	0.80	0.97	
His	1.22	1.08	0.69	
Lys	1.23	0.77	0.96	
Val	0.91	1.49	0.47	Favors β -strand
Ile	0.97	1.45	0.51	
Phe	1.07	1.32	0.58	
Tyr	0.72	1.25	1.05	
Trp	0.99	1.14	0.75	
Thr	0.82	1.21	1.03	
Gly	0.56	0.92	1.64	Favors turn
Ser	0.82	0.95	1.33	
Asp	1.04	0.72	1.41	
Asn	0.90	0.76	1.23	
Pro	0.52	0.64	1.91	
Arg	0.96	0.99	0.88	

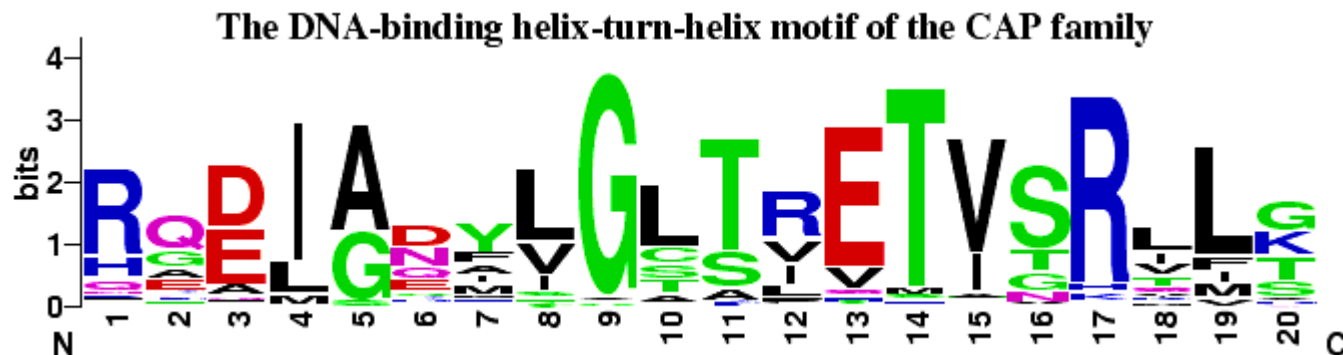
Profile model

- The frequency model does not consider the **order** of the training sequences
 - Permuting the training sequences will not change the model
- In some cases, the order is of important biological meaning, e.g. sequence motifs
- Profile model fully constrains the order of the training sequences

Profile / PSSM

- DNA / protein segments of the same length L
- Often represented as positional frequency matrix

```
LTMTRGDIGNYLGLTVETISRLLGRFQKSGML
LTMTRGDIGNYLGLTIETISRLLGRFQKSGMI
LTMTRGDIGNYLGLTVETISRLLGRFQKSEIL
LTMTRGDIGNYLGLTVETISRLLGRLQKMGIL
LAMSRNEIGNYLGLAVETVSRVFSRFQQNELI
LAMSRNEIGNYLGLAVETVSRVFTRFQQNGLI
LPMSRNEIGNYLGLAVETVSRVFTRFQQNGLL
VRMSREEIGNYLGLTLETVSRLFSTRFGREGLI
LRMSREEIGSYLGLKLETVSRTLKSFHQEGLI
LPMCRRDIGDYLGLTLETVSRALSQ LHTQGIL
LPMSRRDIADYLGLTVETVSRVAVSQLHTDGVL
LPMSRQDIADYLGLTIETVSRFTFKLERHGAI
```



A DNA profile (matrix)

TATAAA		1	2	3	4	5	6
TATAAT	T	8	1	6	1	0	1
TATAAA	C	0	0	0	0	0	0
TATAAA	A	0	7	1	7	8	7
TATTAA	G	0	0	1	0	0	0
TTAAAA							
TAGAAA							

		1	2	3	4	5	6
Sparse data → pseudo-counts	T	9	2	7	2	1	2
	C	1	1	1	1	1	1
	A	1	8	2	8	9	8
	G	1	1	2	1	1	1

Testing a motif

TATAAA
TATAAT
TATAAA
TATAAA
TATAAA
TATTAA
TTAAAA
TAGAAA

TCGAAT
GCATTT
ACTTAA
CGCTGC
AAACCG
CGATAC
CCAAGT
GACCTA

- Equivalent to computing the significance of a sequence motif

	1	2	3	4	5	6
T	9	2	7	2	1	2
C	1	1	1	1	1	1
A	1	8	2	8	9	8
G	1	1	2	1	1	1

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

Model comparison: relative entropy

- $H(x) = \sum_i \sum_j P(x_{ij}) \log (P(x_{ij})/P_0(x_j)) = \sum_{i=1}^L \sum_{j=1}^K (n_{ij} / N) \log \left(\frac{n_{ij} / N}{b_j} \right)$

- $b_j \rightarrow$ the random background distribution
- N sequences of length L
- $K \rightarrow 4$

	1	2	3	4	5	6
T	9	2	7	2	1	2
C	1	1	1	1	1	1
A	1	8	2	8	9	8
G	1	1	2	1	1	1

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

Probability distribution

- What is the probability $P(H|B)$ of getting a matrix with a relative entropy H from the background model $B = \{b_j\}$?
- $p(h) \rightarrow$ the probability distribution of relative entropy score for the frequency of a *single* column (can be pre-calculated)
- $P(H) = \sum_{\substack{(s_1, \dots, s_L) \\ s_1 + \dots + s_L = H}} p(s_1) \cdot p(s_2) \cdot \dots \cdot p(s_L)$
 - convolution of function $p(s)$, can be calculated only approximately by Fast Fourier Transformation (FFT)

Searching profiles: inference

- Give a sequence S of length L , compute the likelihood ratio of being generated from a profile vs. from background model:

$$R(S|P) = \prod_{i=1}^L \frac{(n_{is_i}/N)}{b_{s_i}}$$

- Searching motifs in a sequence: sliding window approach

Finding a motif

atgaccgggatactgatagaagaaagggttgggggcgtacacattagataaacgtatgaagtacgtttagactcggcgccgccg
accctatTTTTTgagcagatttagtgacctggaaaaaaaaatttgagtacaaaactTTTccgaatacaataaaacggcggga
tgagtatccctgggatgacttaaaataatggagtggtgctctcccgattTTTgaatatgtaggatcattcgccagggtccga
gctgagaattggatgcaaaaaaggattgtccacgcaatcgcgaaaccaacgcggacccaaaggcaagaccgataaaggaga
tccTTTTgcggaatgtgccgggaggctggttacgtagggaagccctaacggacttaataataaaggaagggttatag
gtcaatcatgttcttgtgaatggatttaacaataagggtgggaccgcttggcgcacccaaattcagtggtggcgagcgcaa
cggTTTTggcccttgtagaggccccgtataaacaaggaggggccaattatgagagagctaatctatcgcgtgctgttcat
aacttgagttaaaaataggagccctggggcacatacaagaggagtcttcttatcagttaatgctgtatgacactatgta
ttggccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatactaaaaggagcggaccgaaagggaag
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttactaaaaggagcgga

Motif finding is difficult



The motif finding problem

- Given a set of DNA sequences:

```
cctgatagacgctatctggctatccacgtacgtaggtcctctgtgCGAATctatgCGTTTccaacat  
agtactggtgtacatttgatacgtacgtacaccggcaacctgaaacaaacgctcagaaccagaagtgc  
aaacgtacgtgcaccctctttcttcgtggctctggccaacgagggctgatgtataagacgaaaatttt  
agcctccgatgtaagtcatacgtgtaactattacctgccaccctattacatcttacgtacgtataca  
ctgttatacaacgcgctcatggcggggatgCGTTTTggTCGTCgtacgctcgatCGTTAACgtacgTC
```

- Find the motif in each of the individual sequences

The motif finding problem

- If starting positions $\mathbf{s}=(s_1, s_2, \dots, s_t)$ are given, finding consensus is easy because we can simply construct (and evaluate) the profile to find the motif.
- But... the starting positions \mathbf{s} are usually not given. How can we find the “best” profile matrix?
 - Gibbs sampling
 - Expectation-Maximization (EM) algorithm

Conclusion

- Frequency and profile are two basic models for sequence analysis
- They represent two extreme models in terms of incorporating order information in the sequences
- Model selection should be based on biological ideas