Gene Prediction - II

Content Based Methods

Approaches for Gene Prediction

- **Pattern-based Methods**
  - **Finding Open Reading Frames (ORFs)**
  - **Signal-based methods:**
    - **CpG islands**
    - **Finding promoter regions, poly adenylation sites, intron/exon splice sites**
- **Homology-based Method**
- **Content-based Methods**
  - **Coding statistics, *viz*., codon usage bias, periodicity in base occurrence, etc.**
- **Integration of these methods**

Open Reading Frames

A long sequence between two stop codons devoid of stop codons in-between is called an ORF.

An ORF may code for a gene if it:

- Contains a homolog in the database
- Contains signal sequence patterns for translation initiation
- Contains gene-specific features, *viz*., 3-base periodicity, high G+C content, bias in codon usage, etc.
- Has codon usage similar to other genes of the same organism?

Homology-based Approach

- Analysis by sequence similarity (homology-based approach) can reliably identify ~ 30% of the protein-coding genes in a genome
- 50-80% of new genes identified have a partial, marginal, or unidentified homolog
- Frequently expressed genes tend to be more easily identifiable by homology than rarely expressed genes

Identifying Signal Sequence Patterns

Many signals are associated with genes - each of which suggests but does not prove the existence of a gene:

CpG islands, transcription start site, start/stop codons, intron/exon splice sites, promoters,ribosome binding sites, histone binding sites, enhancers, motifs, poly adenylation sites, etc.

Approach used is statistically based

- most of these signals are modeled by position specific scoring matrices (PSSM), or Hidden Markov models (HMM)

Content-based Methods

Content-based gene finding methods rely on statistical information derived from known sequences to predict unknown genes.

Gene-Specific Features:

- such as 3-base periodicity, high G+C content, bias in codon usage (species-specific), inter dependencies in codon positions, etc.

At the core of all gene identification programs there exist one or more **coding measures**

Classification of Coding Measures

Coding statistic – a function that computes the likelihood that the sequence is coding for a protein

Coding statistics measure

- codon usage bias
- base compositional bias between codon positions
- periodicity in base occurrence

Main distinction is between

- measures dependent on model of coding DNA
- measures independent of such a model

Classification of Coding Measures

Model dependent coding statistics capture "specific" features of coding DNA

- require a representative sample of coding DNA from the species under consideration to estimate the model's parameters

Model independent coding statistics capture only the "universal" features of coding DNA.

- do not require a sample of coding DNA

Classification of Coding Measures

Model of coding DNA is probabilistic

Given a query sequence, compute the probability of sequence under

- model of coding DNA, and
- alternative model of non-coding (random) DNA

Logarithm of the ratio of these two probabilities, called the log-likelihood ratio

- gives the score of the coding statistic in the query sequence.

Measures Dependent on a Model of Coding DNA

Codon usage table is used to compute the **coding potential** of a nucleotide sequence, given by the **log-likelihood ratio:**

$$LP^i(S) = \log \frac{P^i(S)}{P_0(S)}$$

$P^i(S)$ - prob. of sequence *S*, given that *S* is coding in frame i (i=1,2,3)

$P^0(S)$ – prob. of *S* given a model of **non-coding** DNA. If

$$LP^i(S) > 0$$

$\Rightarrow$ prob. that *S* is coding in frame *i* is higher than *S* being non-coding.

Measures Dependent on a Model of Coding DNA

Compute log-likelihood ratio in the 3 frames:

$$LP^i(S) = \log \frac{P^i(S)}{P_0(S)}$$

If the sequence is coding, $LP^i(S)$ will be larger for one of the frames.

Non-coding DNA - random DNA sequence with nucleotide equiprobability & independence between positions.

Measures Dependent on a Model of Coding DNA

Measures may be based on

- Oligonucleotide counts – codon usage, amino acid usage, codon preference, hexamer usage
- Base compositional bias between codon positions
- Dependence between nucleotide positions

Codon Usage Bias

Unequal usage of codons in the coding regions is a universal feature of the genomes. This bias occurs mainly due to:

- uneven usage of amino acids in existing proteins,
- uneven usage of synonymous codons

This is due to unequal abundance of tRNA in a given genome

Codon Usage:

We need to know the frequency of codons in

- a coding (gene) sequence,
- a non-coding (random) sequence

Table – I: Human codon usage & codon preference table

The Human Codon Usage Table

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Rel. freq. of each codon among synonymous codons

Freq. of usage of each codon (per thousand)

Table – I: Human codon usage & codon preference table

The Human Codon Usage Table

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Note: preferred codon has either a G or C on the 3rd position.

Codon Usage

F(C) - frequency of codon C in genes of the species under consideration (from codon usage table)

For a given sequence of codons

$$C = C_1 C_2 \ldots C_m$$

the probability of the sequence of codons C coding for a protein is given by

$$P(C) = F(C_1)F(C_2)\ldots F(C_m)$$

Codon Usage

For e.g., if *S* is the sequence *S* = AGGACG, when read in frame 1, it results in the sequence

$$C_1^1 = \mathtt{AGG},\ C_2^1 = \mathtt{ACG}.$$

$$P^1(S) = P(C^1) = F(\mathtt{AGG})F(\mathtt{ACG})$$

Substituting appropriate values from Table-I to compute $P^i(C)$, *i* = 1, 2,3:

$$P^1(S) = P(C^1) = 0.0121 \times 0.007 = 0.0000847$$

Codon Usage

Probability of finding sequence *S* if C is non-coding:

$$P_0(S) = P_0(C) = F_0(C_1)F_0(C_2)\cdots F_0(C_m)$$

$F_0(C)$ - frequency of codon C in a non-coding sequence, and for all codons,

$$F_0(c) = 1/64 = 0.0156$$

Assuming random model of DNA, probability, P_0 for the above sequence of codons C would be

$$P_0(C) = 0.0156 \times 0.0156 = 0.000244$$

Codon Usage

Log-likelihood ratio for *S* coding in frame 1, LP^1,

$$LP^1(S) = \log\left(\frac{P(C)}{P_0(C)}\right)$$

$$LP^1(S) = \log(0.000836/0.000244) = \log(3.43) = 0.53$$

Compute log-likelihood ratio for *S* coding in frames 2 & 3 also and compare

- Frame with the largest value of LP will be the coding frame

Table – I: Human codon usage & codon preference table

G in 1st pos=
321.26/1000 = 0.32

The Human Codon Usage Table

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Sum of 1st column gives the frequency of G in the 1st position

Measures Based on Base Compositional Bias

From codon usage table, probability of each base at each codon position in coding regions can be obtained

How?

Table - II

nucleotide	codon position		
	1	2	3
A	0.27	0.31	0.18
C	0.24	0.24	0.31
G	0.32	0.20	0.29
T	0.17	0.26	0.22

Clear differences in the frequency with which different bases appear at different codon positions

Measures Based on Base Compositional Bias

According to Shepherd, most frequent codons are of the form RNY,

R = A or G, Y = C or G, N any nucleotide

Method to test for the existence and frame of a coding region by measuring the no. of differences between the sequence and the pattern

RNYRNY…RNY

nucleotide	codon position		
	1	2	3
A	0.27	0.31	0.18
C	0.24	0.24	0.31
G	0.32	0.20	0.29
T	0.17	0.26	0.22

Measures Based on Base Compositional Bias

Reason for base compositional bias:

- **Certain AA are favoured over others – amino acid bias**
- **Certain codons among synonymous codons are preferred – codon preference**
- **In the genetic code, first two positions in synonymous codons are generally the same, leading to a bias - the particular structure of the genetic code**

The Genetic Code

First Position	Second Position of Codon: T	Second Position of Codon: C	Second Position of Codon: A	Second Position of Codon: G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

Codons coding for same AA have the 1st two base conserved.

Table – I: Human codon usage & codon preference table

The Human Codon Usage Table

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

codons with C / G on the 3rd codon posn occur with a higher freq

Codon Prototype

Distribution of base frequencies at codon positions describe statistically a prototypical codon

Given a sequence *S*,

measure how similar to the prototypical distribution the observed distribution is

Closer the distributions $\Rightarrow$ more likely that *S* is coding.

These differences can be measured in a number of ways.

Codon Prototype

For coding DNA model, probability of codon *c* in coding regions

$$F(c) = f(c[1], 1) f(c[2], 2) f(c[3], 3)$$

f(b,r) – prob. of nucl. *b* at codon position *r*, assuming independence between adjacent nucleotides.

For non-coding DNA, probability of all triplets *c*

$$F_0(c) = 1/64 \ ?$$

From F and F_0 , P^i and P_0 can be computed

nucleotide	codon position		
	1	2	3
A	0.27	0.31	0.18
C	0.24	0.24	0.31
G	0.32	0.20	0.29
T	0.17	0.26	0.22

Codon Prototype

For instance, if sequence *S* is AGGACG, probability of *S*, if *S* is coding in frame 1, is computed as

$$P^1(S) = F(\mathtt{AGG})F(\mathtt{ACG}) = f(\mathtt{A},1)f(\mathtt{G},2)f(\mathtt{G},3)f(\mathtt{A},1)f(\mathtt{C},2)f(\mathtt{G},3)$$

From Table - II, we obtain

$$P^1(S) = \underbrace{0.27 \times 0.20 \times 0.29}_{F(\mathtt{AGG}) = 0.01566} \times \underbrace{0.27 \times 0.24 \times 0.29}_{F(\mathtt{ACG}) = 0.01879} = 0.0002943$$

P^2 and P^3 are computed in a similar way.

From P^i, and P_0, Codon Preference log-likelihood ratio can be obtained

Table - II

nucleotide	codon position		
	1	2	3
A	0.27	0.31	0.18
C	0.24	0.24	0.31
G	0.32	0.20	0.29
T	0.17	0.26	0.22

Measures Based on Dependence Between Nucleotide Positions

Codon Prototype and Codon Usage

- both based on the probabilities of the codons

However, the two models are very different:

From Table – II, F(AGG) = 0.01566

while from Table – I, F(AGG) = 0.01209

Measures Based on Dependence Between Nucleotide Positions

In Codon Usage:

- model described by explicit probability of each codon.

In Codon Prototype:

- **model described by the probability of occurrence of each base at each position in a codon.**

Codon Prototype and Codon Usage would be equivalent

- if codon positions were independent

Measures Based on Dependence Between Nucleotide Positions

Measures based on the frequency of usage of oligonucleotides, e.g., Codon Usage, implicitly capture dependencies between nucleotide positions.

These dependencies can be explicitly described by means of **Markov models.**

Markov Models

In Codon Prototype

- probability of a nucleotide to appear at a given codon position is constant, independent of the nucleotides in nearby positions, e.g.,

S = AGGACG, probability of occurrence of *G* at codon position 3, (S3 & S6) is the same, 0.29

In Markov Models

- probability of a nucleotide at a particular codon position depends on the nucleotide(s) preceding it.

⇒ probability of occurrence of *G* at codon position 3, (S3 & S6) is not the same, but depends on the nucleotide preceding it (*G* /*C*).

Markov Models

Markov Models of Order 1

- probability of a nucleotide depends only on the preceding nucleotide

Model of coding DNA is based on

- probabilities of the 4 nucleotides at each codon position depending on the nucleotide occurring at the preceding codon position - *transition probabilities*

$\Rightarrow$ three 4 x 4 *transition matrices*: F^1, F^2, and F^3, corresponding to 3 codon positions (in Codon Prototype, only single matrix required)

Markov Models

$F^r(i, j)$ - probability of nucl. *i* in codon position *r* +1 given that nucl. *j* is at position *r*

i.e., prob. of T at codon pos^n 2, if A is at codon pos^n 1

These matrices are estimated from a sample set of genes of a given species.

$F^r(i, j)$ is estimated by the no. of times di-nucleotide *j, i* appears at codon position *r* over the total number of times the nucleotide *j* appears at codon position *r*

i.e., prob. of di-nucleotide AT with A at codon pos^n 1 divided by (AG+AT+AC+AA), the no. of times A occurs at codon pos^n 1

Markov Models

Table III: Probabilities of 4 nucleotides at different codon positions conditioned to the nucleotide in the preceding codon position

codon position 2	codon position 1			
	A	C	G	T
A	.36	.27	.35	.18
C	.21	.23	.24	.27
G	.19	.14	.23	.23
T	.24	.35	.19	.31

codon position 3	codon position 2			
	A	C	G	T
A	.16	.19	.15	.07
C	.28	.44	.41	.33
G	.40	.12	.27	.45
T	.16	.25	.17	.16

codon position 1	codon position 3			
	A	C	G	T
A	.22	.33	.24	.13
C	.21	.29	.27	.21
G	.44	.15	.37	.53
T	.13	.22	.12	.13

For a sequence AGGACG, the probability of occurrence of *G* at codon position 3

Markov Models

The probability of *S* = *AGGACG*, coding in frame 1

$$P^1(S) = f(\mathrm{A},1)F^1(\mathrm{G},\mathrm{A})F^2(\mathrm{G},\mathrm{G})F^3(\mathrm{A},\mathrm{G})F^1(\mathrm{C},\mathrm{A})F^2(\mathrm{G},\mathrm{C})$$

Random model of non-coding DNA

- probability of a nucleotide *i* does not depend on the preceding nucleotide *j*, then

$$F_0(i, j) = 0.25 \text{ , for all } i,j$$

Compute Log-likelihood ratio for each frame

***f*(A,1) - given by probability of nucleotide depending on its codon position in the Codon Prototype table**

Markov Models

Markov Models of Order 2 – prob. of a nucleotide at a given codon pos^{n} depends on the di-nucleotide preceding it, e.g., F^{2}(A, GC) – prob. of A following GC at codon pos^{n} 2, i.e., A at codon pos^{n} 3

Transition matrices - 4 x 16 (3 – each codon pos^{n})

Order of Markov Model

- indicates no. of preceding nucleotides on which the prob. of a given nucl. depends

Markov Models of higher order - capture more of the intrinsic features of coding DNA, but they also depend on more parameters.

Markov Models

Is there any relation between codon usage and Markov model of order 2?

- how many probabilities do we need to compute in each case?

Between hexamer usage and Markov model of order 5?

Markov Models

How many probabilities do we need to compute for codon usage and Markov model of order 2?

$\Rightarrow$ 64 & 208 (64x3 +16)

Between hexamer usage and Markov model of order 5?

$\Rightarrow$ 4096 & 13,312 (4096 x 3 + 1024)

Gene prediction programs

GENSCAN: Probabilistic model based on GHMM to identify complete gene structures in genomic DNA.

- different states of the model correspond to different functional units on a gene, e.g., promoter region, exon, intron, etc.

- uses a homogenous 5^{th} order Markov model for non-coding regions and 3-periodic (inhomogenous) 5^{th} order Markov model for coding regions

Signals are modeled by weight matrixes, weight arrays and maximal dependence decomposition techniques.

Trained on: Vertebrates, Maize, Arabidopsis, Homo sapiens, Accuracy lower for non-vertebrates

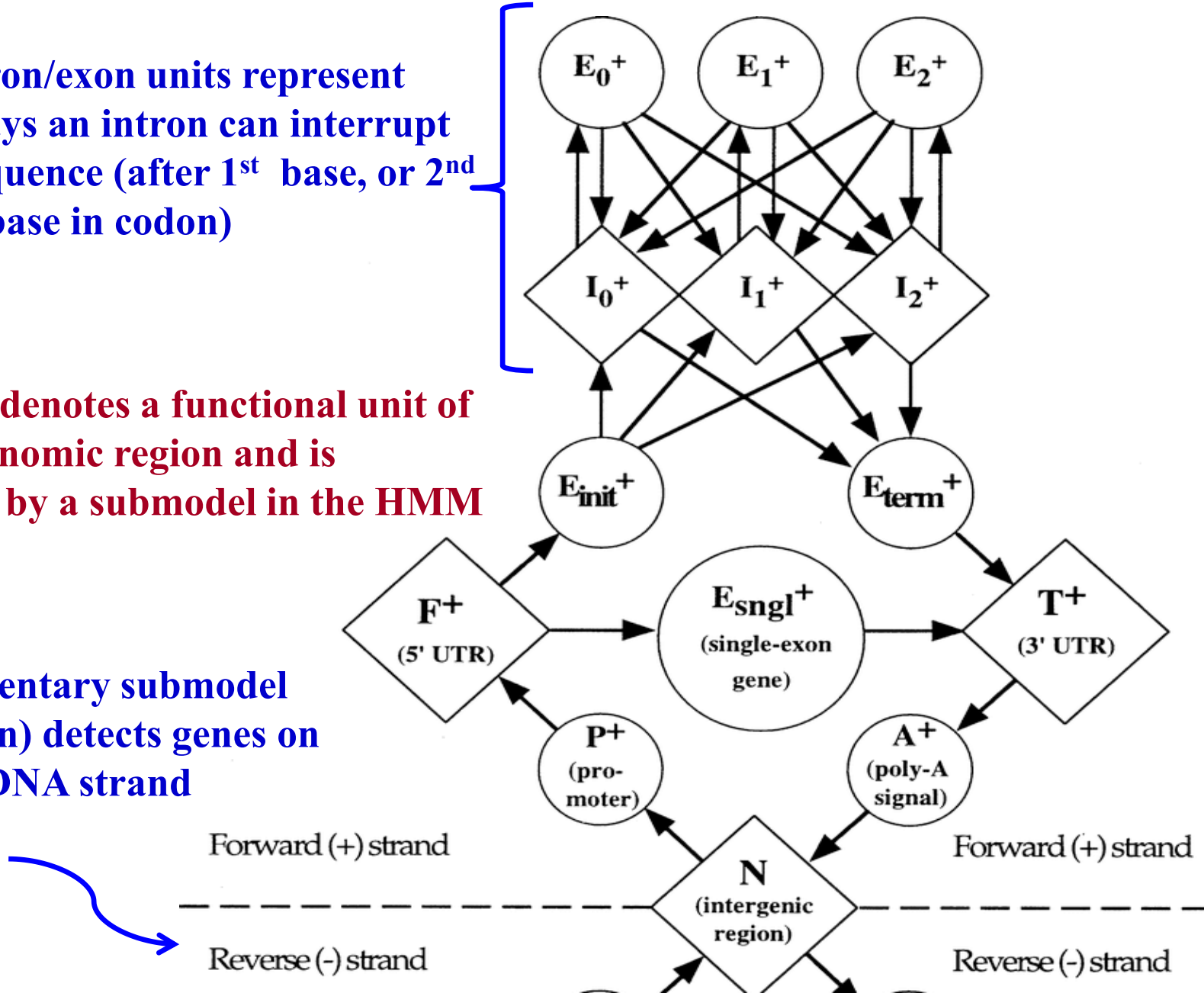
http://genes.mit.edu/GENSCAN.html

GENSCAN HMM for Eukaryotic Gene Finding

Pairs of intron/exon units represent different ways an intron can interrupt a coding sequence (after 1st base, or 2nd base or 3rd base in codon)

Each shape denotes a functional unit of a gene or genomic region and is represented by a submodel in the HMM

Complementary submodel (not shown) detects genes on opposite DNA strand



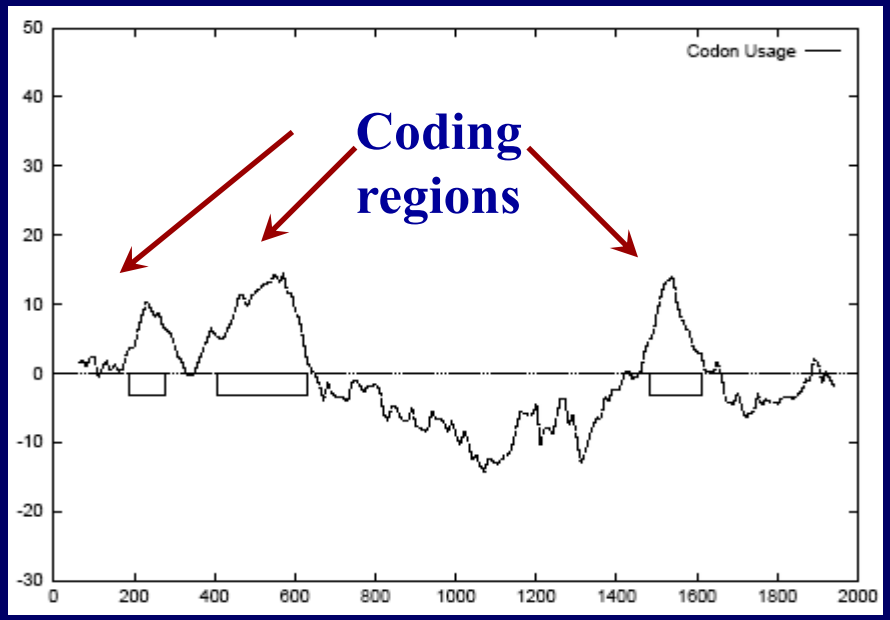
Finding Coding Regions

To locate the (usually small) coding regions within large genomic sequences:

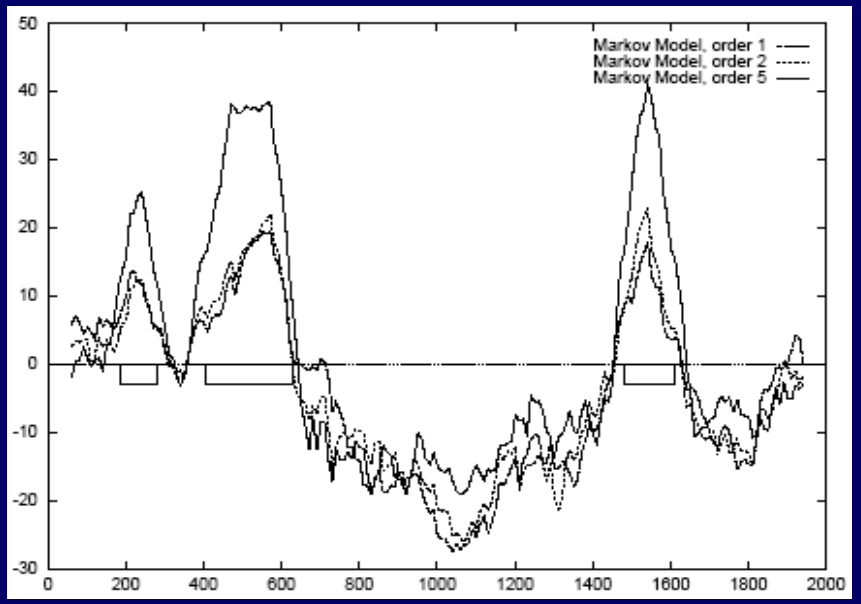
- compute the value of a coding statistic in successive (usually overlapping) sliding windows
- record the value of the statistic for each window

This generates a profile along the sequence in which peaks correspond to coding regions and valleys to non-coding ones.

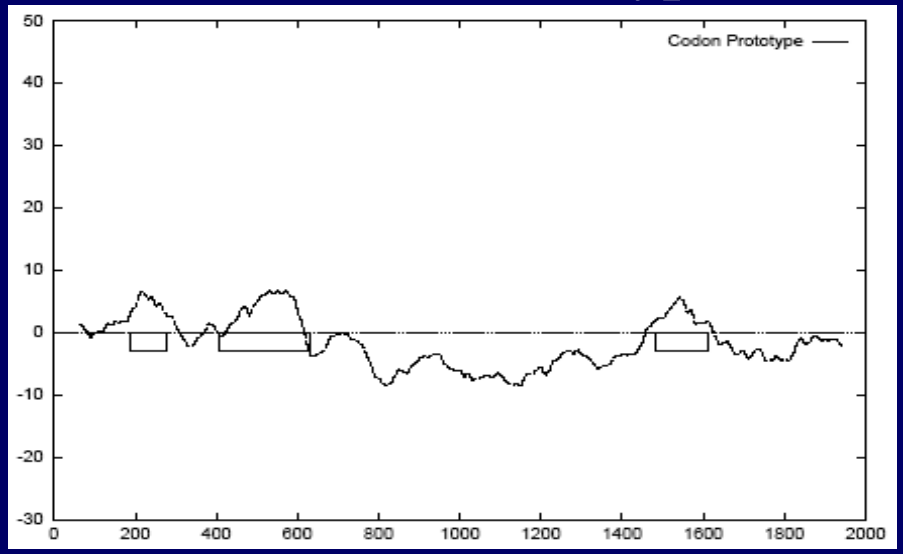
Codon Usage



Markov Models



Codon Prototype



Model dependent Coding Statistics along 2000 bp human β-globin gene, computed on an sliding window of length 120 and step 10.

Measures Independent of a Model of Coding DNA

Measures Independent of a Model of Coding DNA

Methods reviewed so far rely on

- a probabilistic model of coding DNA under which coding likelihood of DNA sequences is computed

- non-biased sample of coding DNA required as for most species such a sample may not exist, or likely to biased towards highly expressed genes

Coding measures not depending on an "*a priori*" model of coding DNA would, therefore, be very useful

Measures Independent of a Coding Model

Underlying assumption - coding DNA is less "random" or "homogeneous" than non-coding DNA

Deviation from randomness can be measured independent of a reference model, and the resulting score correlated with coding function.

Since there is no reference model, these scores do not have a direct probabilistic meaning.

Measures Independent of a Coding Model

One such measure is based on the usage of codons - measure the degree of homogeneity in codon usage between the three frames of the sequence.

Assumption - if the sequence is coding, codon usage will be markedly different in the coding frame compared to other two frames and will exhibit inhomogeneity in codon usage between frames,

If the sequence is not coding, codon usage will essentially be same in all three frames and codon usage will be homogeneous between frames

Position Asymmetry

Another measure is based on asymmetry in the base composition between codon positions

- measures how asymmetric is the distribution of nucleotides at the three codon positions in the sequence, i.e.,

- calculate asymmetry independently for each nucleotide and then combine the values into a single score.

Position Asymmetry

Average frequency of nucleotide *b* at the three codon positions is:

$$f_S(b) = \sum_{r=1}^{3} (f_S(b,r))/3$$

$f_s(b,r)$ – relative frequency of nucleotide *b* at codon position *r* in sequence *S*

Asymmetry in the distribution of nucleotide *b* is defined as the variance of this frequency

$$\text{asym}(b) = \sum_{i=1}^{3} (f_S(b,i) - f_S(b))^2$$

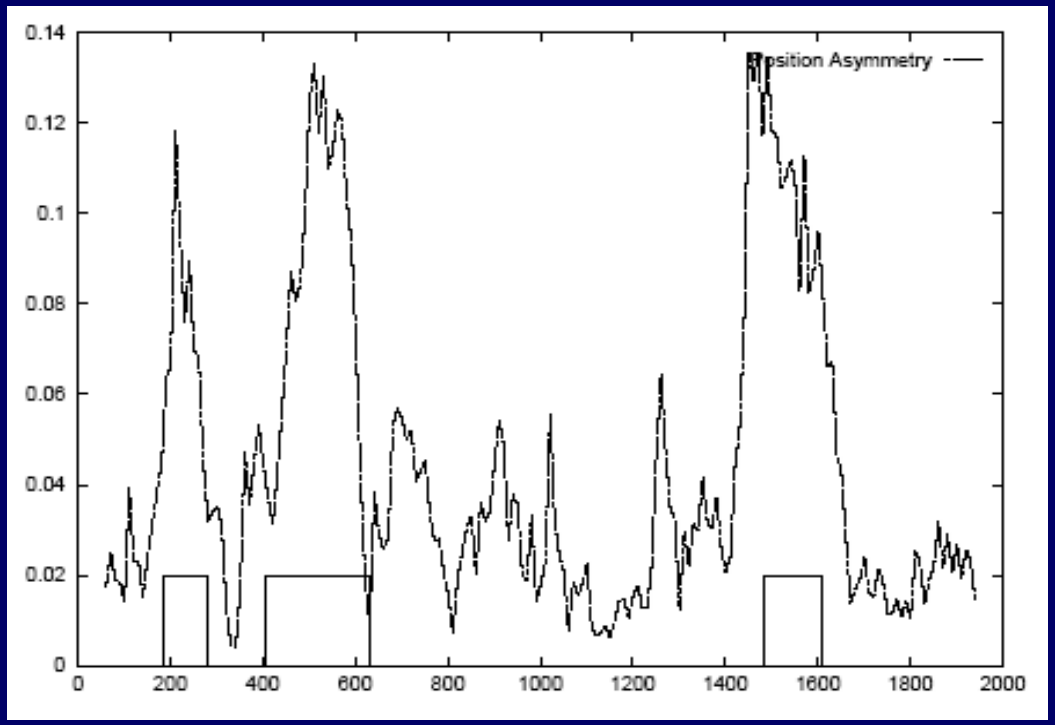
summed over the whole sequence.

Position Asymmetry

Position asymmetry of the sequence, *PA*(*S*) is

$$PA(S) = \mathrm{asym}(A) + \mathrm{asym}(C) + \mathrm{asym}(G) + \mathrm{asym}(T)$$

***asym*(*b*) – is independent of the frame in which codons are defined ⇒ only one value of asymmetry needs to be computed**



Computed along the 2000 bp human β-globin gene sequence, in a sliding window of length 120 and step 10

Fickett & Tung, TESTCODE

Methods Based on Periodic Correlations Between Nucleotide Positions

Given a DNA sequence, compute how many times nucleotide *i* is followed by nucleotide *j* at a distance of *k* nucleotides, $N_{ij}(k)$

For instance, if the sequence is

$$S = AGGACGGGATCA,$$

then $N_{GA}(1) = 2$, $N_{AT}(0) = 1$, $N_{GG}(0) = 3$, $N_{AA}(7) = 2$, and so on.

Methods Based on Periodic Correlations Between Nucleotide Positions

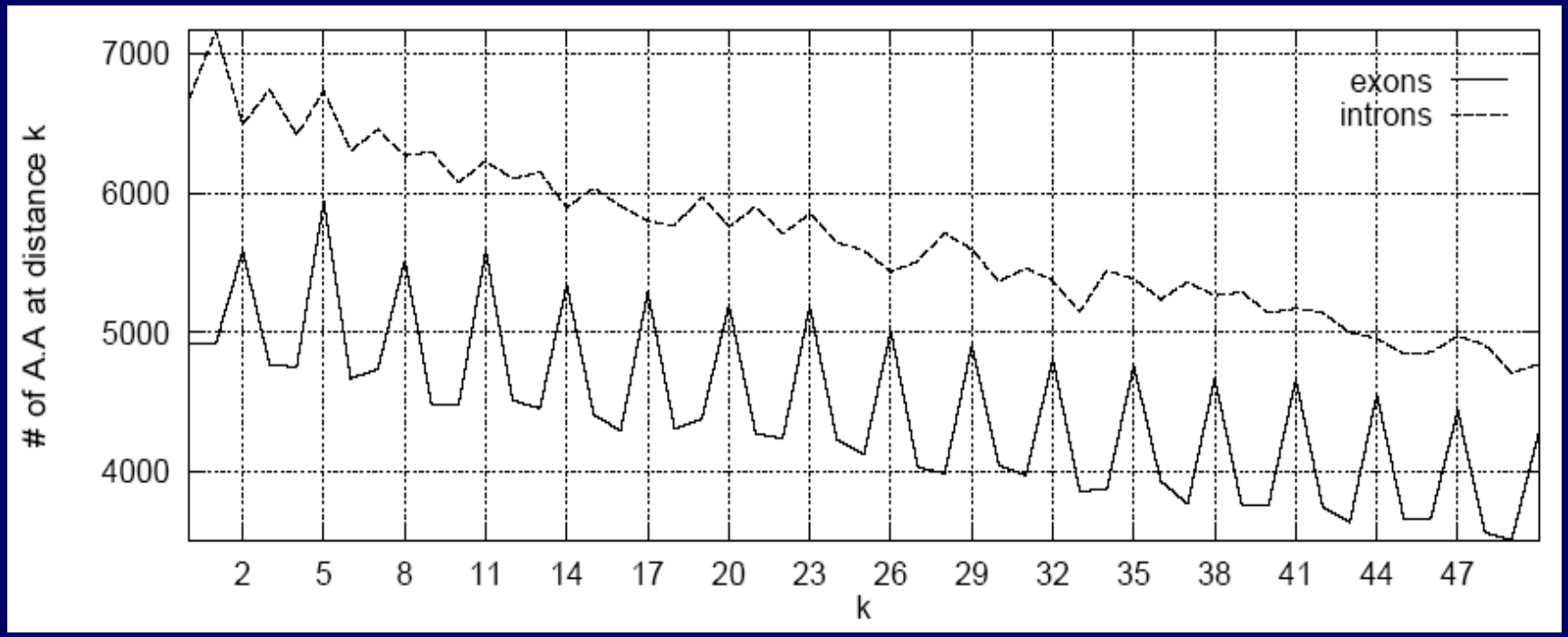
On computing absolute frequency of pair *A…A* with *k* nucleotides between them in the first 200 base pairs of sequences in test sets of human exons and introns, it was observed that

$\Rightarrow$ a clear periodic pattern arises from a set of exons

$\Rightarrow$ nucleotide *A* is more likely to be found at a distance *k* = 2, 5, 8, … from another *A*, than at other distances

Note: nucleotide pairs at distances *k* = 2, 5, 8, … are at the same codon position.

Periodic Structure in DNA sequence



First 200 bases in a set of 1761 human exons & 1753 human introns

A clear period-3 pattern observed in coding regions, which is absent in non-coding regions - reflects correlation between nucleotide positions along coding sequences

Methods Based on Periodic Correlation

A number of coding statistics have been devised based on measuring the periodic structure of DNA sequences:

- Periodic Asymmetry Index
- Average Mutual Information
- Fourier Spectrum

Fourier Series

Mathematically, a Fourier series decomposes a *periodic* function into a sum of simple oscillating functions, namely, *sines* and *cosines*.

Fourier series for *f* on the interval [-π, π]:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos(nx) + b_n \sin(nx) \right]$$

Using Euler's formula, $e^{inx} = \cos(nx) + i\sin(nx)$:

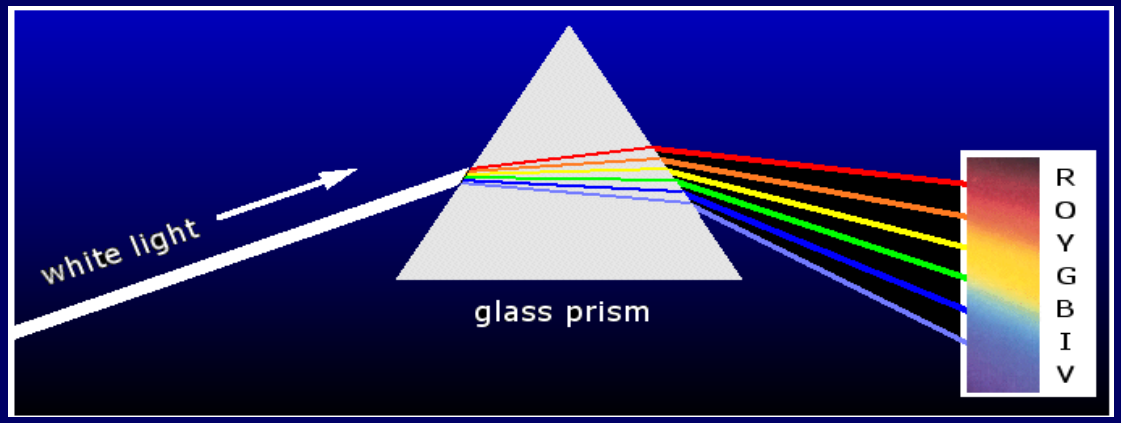
$$f(x) = \sum_{-\infty}^{\infty} c_n e^{inx} \qquad c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx$$

the Fourier coefficients are then given by $a_n = c_n + c_{-n}$, $b_n = i(c_n - c_{-n})$

Fourier Transform

Goal of Fourier transform method is given a function in time signal to be able to compute the frequency components of the signal.

It would give information on how signals of different frequencies are represented in a signal.



Effect of white light on a prism is like a Fourier transform, all the different frequencies of light can be observed.

Fourier Spectrum

Periodic correlations in DNA sequences can be examined by means of Fourier analysis

The partial spectrum of a DNA sequence *S* of length *l* corresponding to nucleotide *b* is

$$S_\alpha(f) = \frac{1}{N}\sum_{n=0}^{N-1} x_\alpha[n] e^{-2\pi f n i}$$

$x_\alpha[n] = 1$ if $S_j = \alpha$; 0 otherwise;

f is the discrete frequency, $f = m/N$, with m = 1, 2, …, $N/2$

In the analysis of a DNA sequence, we have a signal in space, which we want to convert in frequency domain.

Fourier Transform of DNA Sequences

Convert the DNA sequence into four nucleotide subsequences:

$$x_A[n],\ x_T[n],\ x_G[n],\ x_C[n]$$

Let $x_\alpha[n] = 1$ if character α is present at the n^{th} position of the DNA sequence; $x_\alpha[n] = 0$ otherwise;

$x_\alpha[n]$ is an indicator sequence for the presence or absence of character α in the DNA sequence.

Fourier Transform of DNA Sequences

For example, $x_\alpha[n]$ components for DNA sequence:

ACTGCTAGCAAT

Σ	A	C	T	G	C	T	A	G	C	A	A	T
$x_A\ n$	1	0	0	0	0	0	1	0	0	1	1	0
$x_T\ n$	0	0	1	0	0	1	0	0	0	0	0	1
$x_C\ n$	0	1	0	0	1	0	0	0	1	0	0	0
$x_G\ n$	0	0	0	1	0	0	0	1	0	0	0	0

Fourier transform

$$S_\alpha(f) = \frac{1}{N}\sum_{n=0}^{N-1} x_\alpha[n] e^{-2\pi f n i}$$

for $0 \le f \le 0.5$, $f = m/N$, $m = 1, 2, \ldots, N/2$

Fourier Transform of DNA Sequences

From the Fourier product spectrum

$S(f) = \sum_{\alpha} |S_{\alpha}(f)|^2, \qquad \alpha = \{A, T, G, C\}$

If a period **P** repeat exists in the DNA sequence, S(f) would show a peak at ***f* = 1/P**.

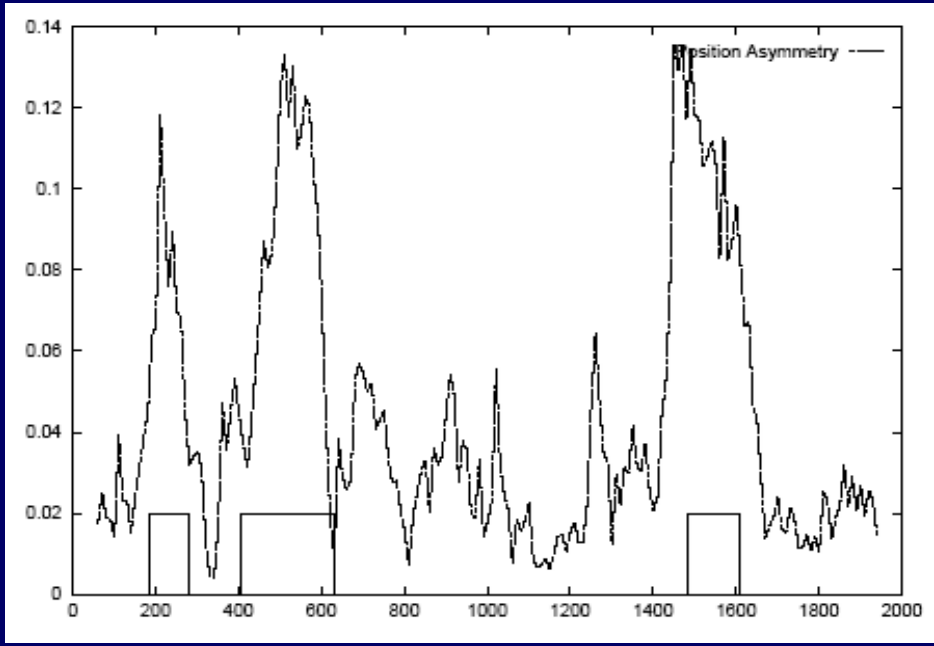
Total Fourier Spectrum of the DNA sequence is the sum of the four partial Spectra:

$$S(f) = \sum_{b \in \{A,C,G,T\}} S_b(f)$$

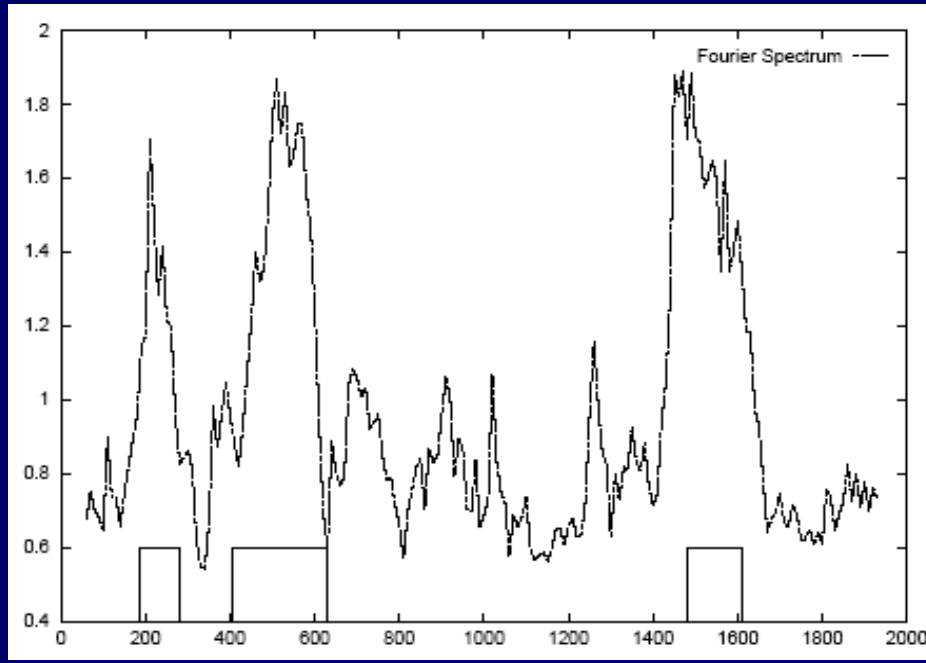
DNA **coding regions** reveal characteristic periodicity of **3** as a **distinct peak** at frequency ***f* = 1/3**.

No such "peak" is seen in non-coding sequences

Position Asymmetry



Fourier Spectrum



Model independent Coding Statistics along 2000 bp human β-globin gene, computed on an sliding window of length 120 and step 10.

Model-independent Measures

Model-independent coding measures are useful when no previous coding sequences are known for a given genome.

However, the signal they produce is weaker than that produced by model-dependent measures,

⇒ usually longer sequences required to obtain discrimination.

⇒ limits their utility mostly to prokaryotic genomes where genes are continuous ORFs.

		exon sequence			intron sequence		
		coding frame	non coding frames		frame 1	frame 2	frame 3
Codon Usage		24.06	-16.13	-3.16	-14.36	-23.74	-19.67
Hexamer Usage		27.62	-11.64	-6.51	-20.90	-27.56	-22.07
		39.98	-14.58	-8.46	-26.73	-27.81	-25.87
Codon Preference		15.97	-1.32	7.24	-7.96	-12.70	-14.93
Amino Acid Usage		8.17	-14.87	-10.17	-6.15	-10.69	-4.57
Codon Prototype		9.87	-11.23	-10.30	-11.45	-17.44	-14.49
Markov Model	order 1	29.92	-2.69	-3.31	-35.44	-42.40	-41.73
	order 2	34.73	-18.26	-7.77	-29.61	-41.76	-40.05
	order 5	72.69	-21.38	13.56	-37.63	-30.99	-36.40
Position Asymmetry		0.0957			0.0211		
Periodic Asymmetry Index		1.159			1.009		
Average Mutual Information		0.00681			0.000344		
Fourier Spectrum		2.278			0.892		

Values of different coding statistics in the 223 bp long second coding exon of the human β-globin gene, and in a 223 bp long sequence from the middle of the second intron of the same gene

Complications in Gene Prediction

Problem of gene identification is complicated in case of eukaryotes by the vast variation that is found in the structure of genes.

On an average, a vertebrate gene is ~ 30Kb long. Of this, coding region is only about 1Kb.

Coding region typically consists of 6 exons, each about 150bp long.

These are average statistics

Complications in Gene Prediction

Huge variations from the average observed

Biggest human gene, dystrophin ~ 2.5Mb long.

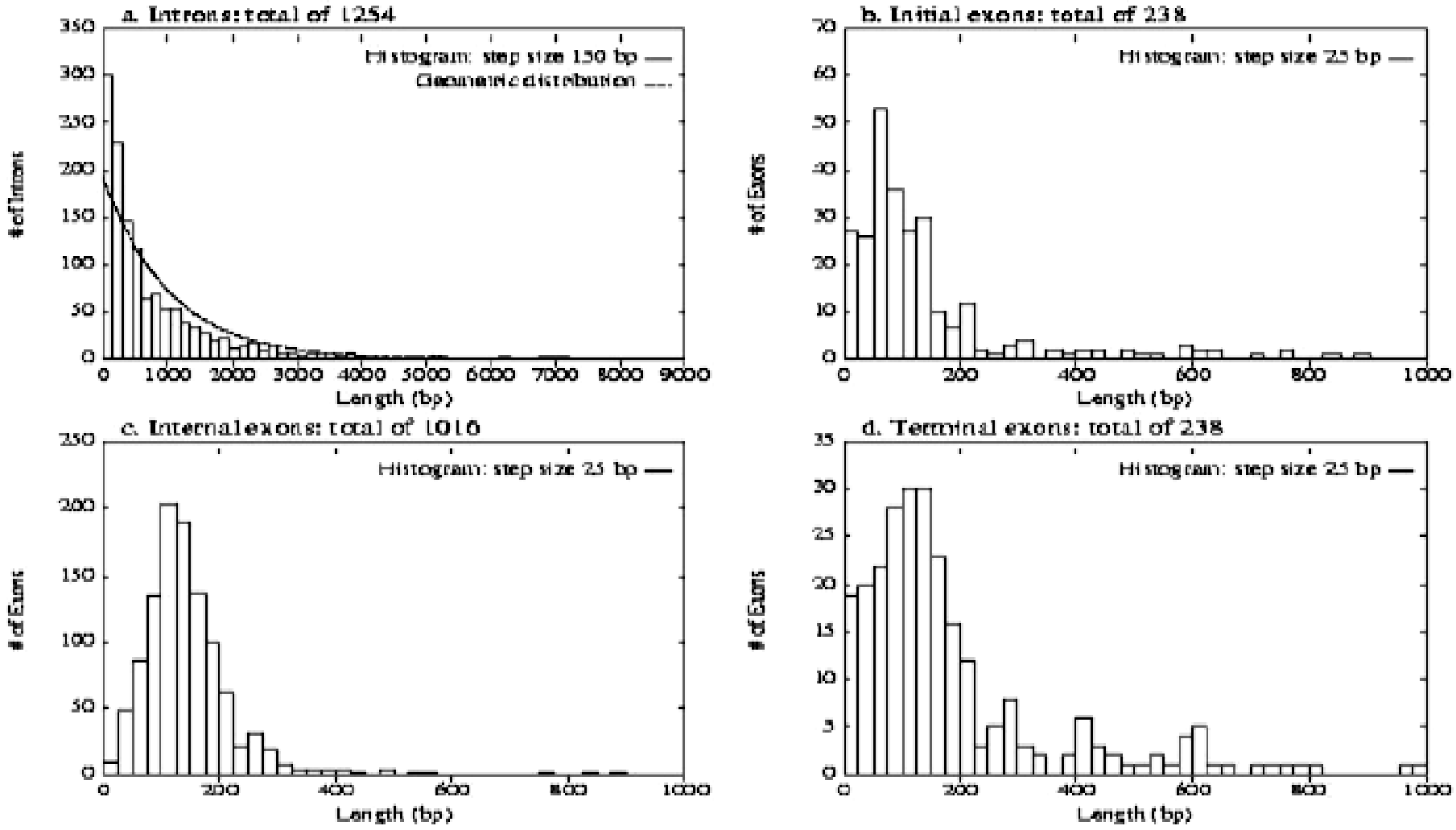
Human blood coagulation factor gene ~ 186Kb, has 26 exons, sizes varying from 69 - 3106 bp, its 25 introns range in size from 207 - 32,400 bp. Intron 22 has two transcripts, one in the same orientation and another in the reverse.

An average 5' UTR is 750bp long, but it can be longer & span several exons (e.g., in MAGE family)

On an average, 3' UTR is about 450bp long, but for e.g., in case of the gene for Kallman's syndrome, the length exceeds 4Kb.

Couple all this with alternative splicing

Length distribution of human exons and introns



A large variation in the size of genes and exons observed in the eukaryotic genome – 238 multi-exon genes analysis shown

Some facts about human genes

- Comprise about 3% of the genome
- Average gene length: ~ 8,000 bp
- Average no. of exons/gene: 5 - 6
- Average exon length: ~ 200 bp
- Average intron length: ~ 2,000 bp
- ~ 8% genes have a single exon

Some exons can be as small as 1 or 3 bp

Complications in Gene Prediction

In higher eukaryotes gene finding is difficult:

- **Multiple ORFs** need to be combined to obtain a spliced coding region.
- **Alternative splicing** is not uncommon,
- **Variation** in gene structure - exons can be very short, and introns can be very long.

Recent research has shown that some long non-coding RNAs have features similar to protein-coding genes.

References

DNA Composition, Codon Usage and Exon Prediction, Roderic Guigo, Chapter published in "Genetic Databases", M.J. Bishop ed., Academic Press, 1999

For an up-to-date list of references by Wentian Li
www.nslij-genetics.org/wli/