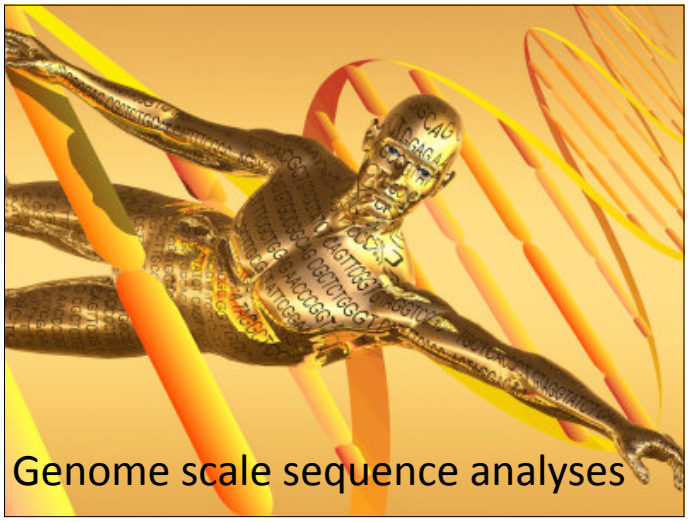
BIOINFORMATICS 1

or why biologists need computers

http://www.bioinformatics.uni-muenster.de/teaching/courses-2015/bioinf1/index.hbi

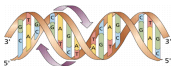
Prof. Dr. Wojciech Makałowski
Institute of Bioinformatics

1



2

CENTRAL DOGMA OF MOLECULAR BIOLOGY



DNA

transcription

RNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

3

CHALLENGE: HOW FROM THIS...

TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT

4

5

GENOME ANNOTATION

What are we looking for?

- protein coding genes
- RNA coding genes
- gene promoters
- repetitive elements

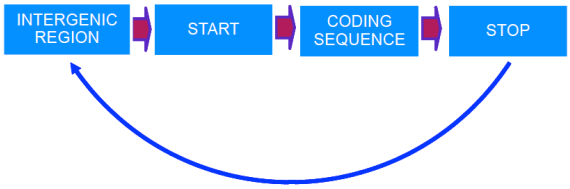
6

GENE IDENTIFICATION METHODS

- Molecular techniques
  - Very laborious
  - Time consuming
  - Expensive
  - Low rate of false positives
- Computational methods
  - Fast
  - Relatively low cost
  - High rate of false positives
  - Poor performance on less typical genes

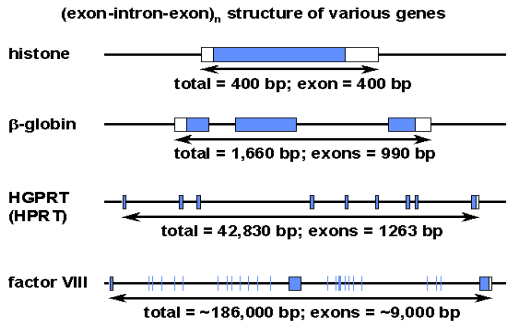
7

GENERAL MODEL OF A GENE



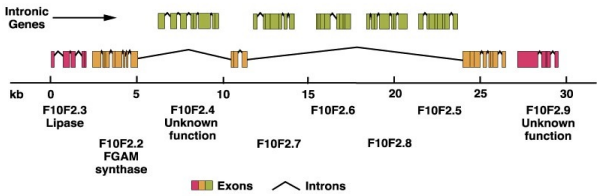
8

EUKARYOTIC GENE STRUCTURE



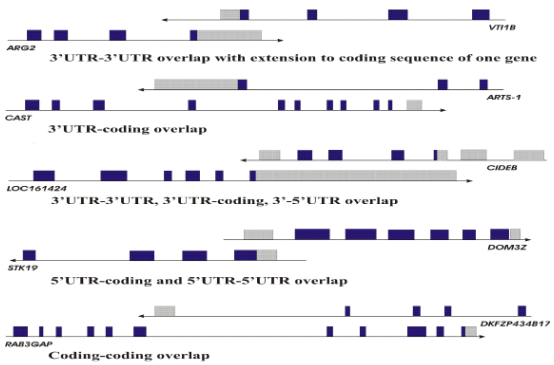
9

NESTED GENES



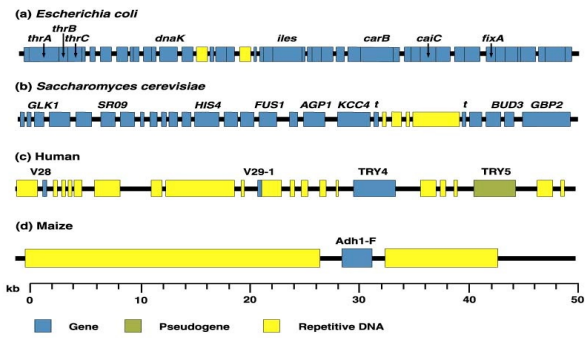
10

OVERLAPPING GENES



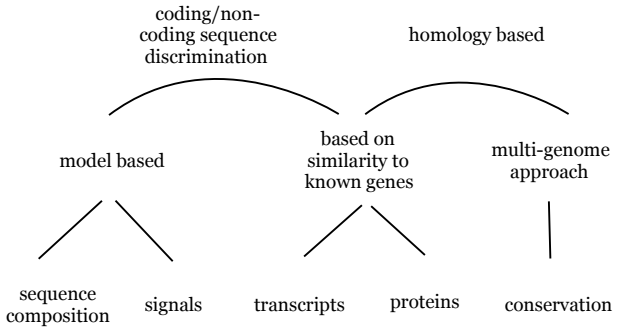
11

PSEUDOGENES AND REPETITIVE ELEMENTS



12

GENE FINDING METHODS



13

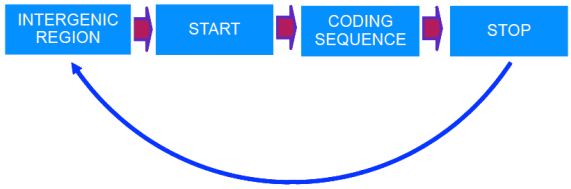
MODEL BASED METHODS

We take advantage of what we already learned about gene structures and features of coding sequences. Based on this knowledge we can build theoretical model, develop an algorithm to search for important features, train it on known data and use to search for coding sequences in anonymous genomic fragments.

However, we should remember that all models are wrong and only some are useful.

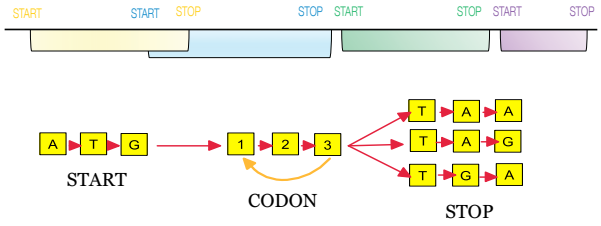
14

GENERAL MODEL OF A GENE



15

SEQUENCE CODING POTENTIAL



16

CODON USAGE

Codon preference in *E. coli* and *S. typhimurium* genes.[1]

	U			C			A			G			
U	UUU	Phe	19	UCU	Ser	9	UAU	Tyr	15	UGU	Cys	4	U
	UUC	Phe	17	UCC	Ser	10	UAC	Tyr	12	UGC	Cys	6	C
	UUA	Leu	11	UCA	Ser	6	UAA	STOP	2	UGA	STOP	0.8	A
	UUG	Leu	12	UCG	Ser	8	UAG	STOP	0.2	UGG	Trp	12	G
C	CUU	Leu	10	CCU	Pro	7	CAU	His	11	CGU	Arg	23	U
	CUC	Leu	10	CCC	Pro	5	CAC	His	10	CGC	Arg	23	C
	CUA	Leu	4	CCA	Pro	7	CAA	Gln	13	CGA	Arg	3	A
	CUG	Leu	55	CCG	Pro	15	CAG	Gln	31	CGG	Arg	5	G
A	AUU	Ile	27	ACU	Thr	9	AAU	Asn	17	AGU	Ser	7	U
	AUC	Ile	27	ACC	Thr	25	AAC	Asn	24	AGC	Ser	16	C
	AUA	Ile	4	ACA	Thr	6	AAA	Lys	36	AGA	Arg	2	A
	AUG	Met	26	ACG	Thr	15	AAG	Lys	12	AGG	Arg	1	G
G	GUU	Val	17	GCU	Ala	16	GAU	Asp	33	GGU	Gly	24	U
	GUC	Val	16	GCC	Ala	25	GAC	Asp	22	GGC	Gly	33	C
	GUA	Val	12	GCA	Ala	16	GAA	Glu	43	GGA	Gly	6	A
	GUG	Val	26	GCG	Ala	37	GAG	Glu	20	GGG	Gly	10	G

17

SEQUENCE FEATURES

We can check if sequence in particular ORF has some other features which could tell us if this is a putative coding sequence or the ORF is false positive. We can look at the sequence content and compare it with known coding sequence and non-coding sequence and check to which of these two the ORF sequence is more similar to.

18

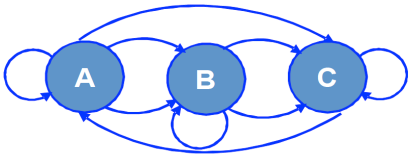
HIDDEN MARKOV MODELS

- HHM is a statistical model for an ordered sequence of symbols, acting as a stochastic state machine that generates a symbol each time a transition is made from one state to the next. Transitions between states are specified by transition probabilities. A Markov process is a process that moves from state to state depending on the previous n states.
- HHM has been previously used very successfully for speech recognition.
- In biology is used to produce multiple sequence alignments, in generating sequence profiles, to analyze sequence composition and patterns, to produce a protein structure prediction, and to locate genes.
- In gene identification HMM is a model of periodic patterns in a sequence, representing, for example, patterns found in the exons of a gene. HMM provides a measure of how close the data pattern in the sequence resemble the data used to train the model.

19

MARKOV CHAINS

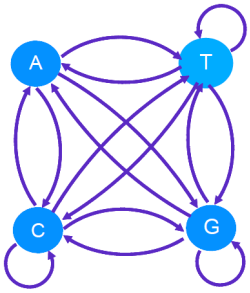
A Markov Chain is a non-deterministic system in which it is assumed that the probability of moving from one state to another doesn't vary with time. This means the current state and transition does not depend on what happened in the past. The Markov Chain is defined by probabilities for each occurring transition.



20

MARKOV CHAINS

In a sequence analysis we look at probabilities of transitions from one nucleotide to another. We can check, for example, if certain patterns of transition are more frequent in coding sequences than in non coding sequences.



21

ORDER OF MARKOV CHAINS

GCGCTAGCGCCGATCATCTACTCG

GCGCT**AG**CGCCGATCATCTACTCG
GCGCTA**GC**GCCGATCATCTACTCG — First order

GCGCT**TAG**CGCCGATCATCTACTCG
GCGCT**AGC**GCCGATCATCTACTCG — Second order

G**CGCTAG**CGCCGATCATCTACTCG
GC**GCTAGC**GCCGATCATCTACTCG — Fifth order

22

HOW FAR CAN WE GO?

- Order of our model will have influence on specificity and sensitivity of our program.
  - Too short sequences may not be specific enough and program may return a lot of false positives.
  - Long chains may be too specific and our program will not be sensitive enough returning false negatives.

23

ORDER OF MARKOV CHAINS

GCGCTAGCGCCGATCATCTACTCG

GCGCT**AG**CGCCGATCATCTACTCG
GCGCTA**GC**GCCGATCATCTACTCG — First order

GCGCT**TAG**CGCCGATCATCTACTCG
GCGCT**AGC**GCCGATCATCTACTCG — Second order

G**CGCTAG**CGCCGATCATCTACTCG
GC**GCTAGC**GCCGATCATCTACTCG — Fifth order

20 G
7 GA 7/20
1 GG 1/20
5 GT 5/20
7 GC 7/20

For non-coding sequence we assume that probability of each transition is equal. The more 'popular' in coding sequence transition, the higher probability the sequence is coding

24

Probability matrix

Number of probabilities in a DNA matrix of a given order can be calculated according to the following formula:

$$4^{k+1}$$

where 4 represents number of letters in the DNA alphabet and k stands for the order number.

Hence, first order Markov Model matrix consists of $4^2 = 16$ probabilities

p(A/A), p(A/T), p(A/C), p(A/G),

p(T/A), p(T/T), p(T/C), p(T/G),

p(C/A), p(C/T), p(C/C), p(C/G),

p(G/A), p(G/T), p(G/C), p(G/G)

GCG CTA GCG CCG ATC ATC TAC TCG

G CGC TAG CGC CGA TCA TCT ACT CG

GC GCT AGC GCC GAT CAT CTA CTC G

Frequencies of transitions may depend on in which codon position (1st, 2nd, or 3rd) is a given nucleotide (state)

Number of probabilities

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 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 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

$$4^{1+1} = 4^2 = 16$$

$$3\,(4^{1+1}) = 3 \times 4^2 = 48$$

CALCULATING CODING POTENTIAL OF A GIVEN SEQUENCE

To estimate if the sequence is coding we have to calculate probability that sequence is coding and probability the sequence is non-coding. Next we calculate logarithm from the ratio of these two probability values.

$$LP(S) = \log \frac{Pi\,(S)}{P_0\,(S)}$$

If the calculated value is > 0 the likelihood that the sequence is coding is higher than the sequence is not coding, if value is < 0 there is higher likelihood that sequence is not coding.

CODING VS. NON CODING SEQUENCE

A/A	C/A	G/A	T/A	coding
0.36	0.21	0.19	0.24	
A/A	**C/A**	**G/A**	**T/A**	**non coding**
0.25	0.25	0.25	0.25	

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

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 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 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

S=AGGACG

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

$P(S) = 0.27 \times 0.19 \times 0.27 \times 0.24 \times 0.21 \times 0.12 = 0.00008377$

$P(S) = 0.25 \times 0.25 \times 0.25 \times 0.25 \times 0.25 \times 0.25 = 0.0002441$

$LP(S) = \log(0.00008377/0.0002441) = -0.4644$

* in the case of first position in the analyzed sequence we put the frequency of a particular letter in the analyzed genome

CALCULATING LP

$$LP(S) = \log \frac{P_i(S)}{P_0(S)}$$

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 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 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

S=AGGACG

$$LP(S) = \log\frac{0.27}{0.25} + \log\frac{0.19}{0.25} + \log\frac{0.27}{0.25} + \log\frac{0.24}{0.25} + \log\frac{0.21}{0.25} + \log\frac{0.12}{0.25}$$

LP(S) = log 1.08 + log 0.76 + log 1.08 + log 0.96 + log 0.84 + log 0.48

LP(S) = 0.0334 + (-0.1191) +0.0334 + (-0.0177) + (-0.0757) + (-0.3187)

LP(S) = -0.4644

31

GLIMMER

- Gene finding program for prokaryotes (Saltzberg et. al, 1998)
- For prediction uses:
  - Start
  - Stop
  - Sequence composition
  - Interpolated Markov Models

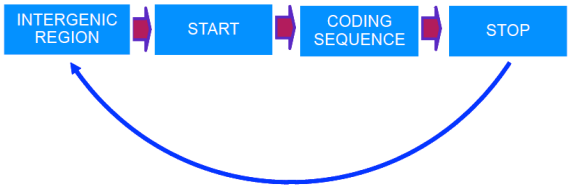
32

PROKARYOTIC VS. EUKARYOTIC GENES

Prokaryotes	Eukaryotes
small genomes	large genomes
high gene density	low gene density
no introns (or splicing)	introns (splicing)
no RNA processing	RNA processing
similar promoters	heterogeneous promoters
terminators important	terminators not important
overlapping genes	overlapping genes
	polyadenylation

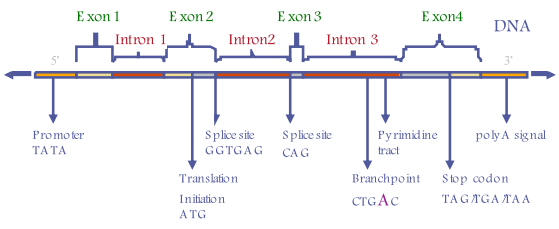
33

CODING REGIONS IN PROKARYOTES



34

EUKARYOTIC GENE STRUCTURE

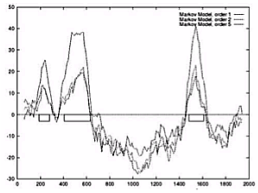


35

SEARCHING FOR CODING SEQUENCES USING MARKOV CHAINS

In this case we do not want check if given sequence fragment is coding or not but we rather want to identify coding fragments in a long sequence. In most cases this is done by calculating statistics in overlapping windows.

AGTACGATATTAGCGGCAATCGTATGACTACGTCTTGCTACGTCTTCTCTCGTCTGCTCTAG



This example shows a profile for a sequence analyzed using a 120-bp window and a 10-bp step.

36

CODON USAGE

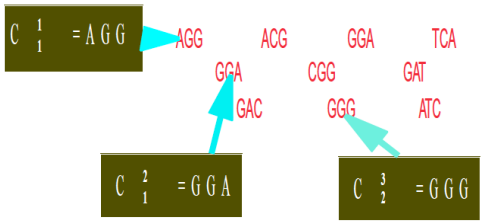
Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38

37

CODON USAGE

DNA sequence can be divided into non-overlapping codons in three reading frames

C = C1C2...Cm

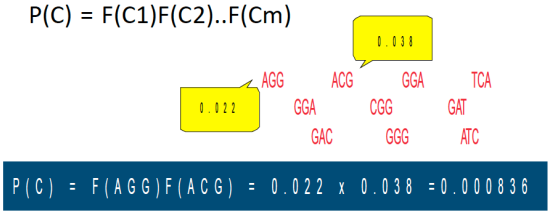


38

PROBABILITY THAT SEQUENCE IS CODING

Probability that sequence is coding is equal probability that sequence of codons is coding. Assuming independence between adjacent codons the probability that sequence is coding will be equal to the product of codon frequencies.

P(C) = F(C1)F(C2)..F(Cm)



39

PROBABILITY THAT SEQUENCE IS NON-CODING

If the sequence is non-coding the codon frequency will be random and each codon will be equally probable. In this case frequency for each codon will be 0.0156. This is because we have 64 codons and each of them is equally likely.

Therefore probability that the sequence is non-coding will be:

P(C) = F(AGG)F(AGC) = 0.0156 x 0.0156 = 0.000244

40

LOG-LIKELIHOOD RATIO

To determine if the sequence is more likely coding or non-coding we can use log-likelihood ratio:

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

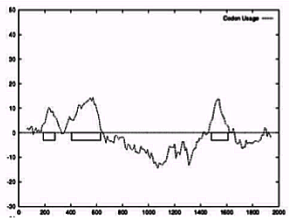
In our example (AGGAGC)

$$LP(S) = \log \frac{P^i(S)}{P_o(S)} = \log \frac{0.000836}{0.000234} = \log 3.57 = 0.55$$

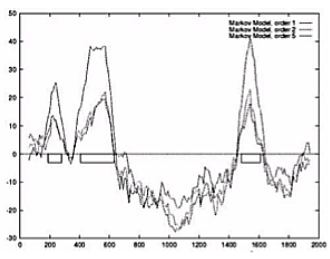
Conclusion: AGGAGC string is more likely to be protein coding

41

Codon usage

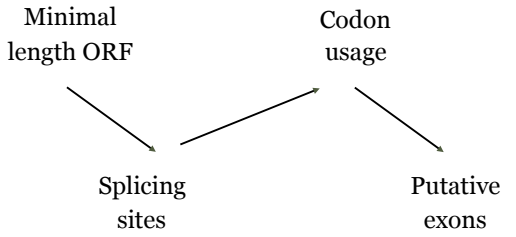


Markov models



42

RULE BASED METHODS



43

GENE IDENTIFICATION PROGRAMS

- The first generation of programs was designed to identify approximate locations of coding regions in genomic DNA (e.g. GRAIL). These methods could not accurately predict precise exon location.
- The second generation (e.g. MZEF, SORFIND, and Xpound) combined splice signals and coding region identification but did not attempt to assemble predicted exons into complete genes.
- Third generation (GeneID, GeneParser, GenLang,FGENES) predicted entire gene structures but their performance was rather poor. One of problems was the assumption that the input sequence contains complete genes.
- Fourth generation of programs is represented by GENSCAN or TWINSCAN. With improved accuracy and less restricted requirements (e.g. allow partial genes) these programs are considered to be the best and are widely used in large-scale genomes analysis.

44

CLASSES OF GENE PREDICTION METHODS

Sequence similarity based

- BLAST can be used for aligning ESTs or proteins to the genomic sequence
- PROCRUSTES and GenWise use global alignment of homologous protein to genomic sequence
- The biggest limitation to this type of approaches:
  - only about half of genes being discovered have significant similarity to genes in the database
  - genes with very limited expression may never be discovered

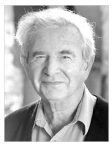
Model based

- Limitations of these approaches:
  - Newly sequenced genomes very often lack large enough samples of known genes to estimate model parameters
  - Need to be retrained as the number of available genes is growing
  - Genes of less typical structure or having rare signals may not be discovered

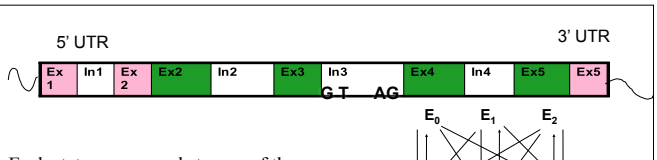
45

GENSCAN

- Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.
- Search for general and specific compositional properties of distinct functional units in eukaryotic genes
- General fifth-order Markov model of coding regions
- Analyzes both DNA strands
- Sequences may contain multiple and/or partial genes
- http://genes.mit.edu/GENSCAN



46



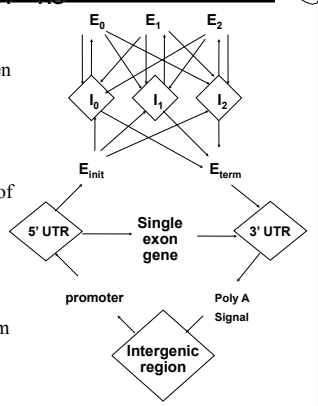
Each state corresponds to one of the seven categories with which all nucleotides are ultimately labeled- promoter, 5'UTR, exon, intron, 3'UTR, PolyA, intergenic

Three components:

Transition model – specifies probability of moving from any one state to another

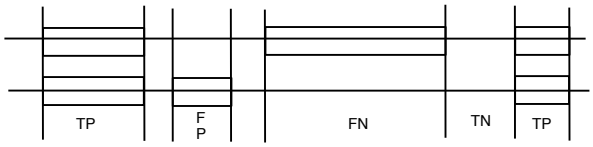
Duration model – specifies the probablity of staying in a given state

State specific sequence models – specifies the probability of any given nucleotide sequence being generated from any given state



47

Evaluation statistics



TP - true positive
FP - false positive
FN - false negative
TN - true negative

Sensitivity Fraction of actual coding regions that are correctly predicted as coding, ranging from 0 to 1
Sn = TP/(TP+FN)

Specificity Fraction of the prediction that is actually correct, ranging from 0 to 1
Sp = TP/(TP+FP)

Correlation Combined measure of sensitivity and specificity, ranging from -1 (always wrong) to +1 (always right)

$$CC = \frac{TP \times TN + FP \times FN}{\sqrt{(PP)(PN)(AP)(AN)}}$$

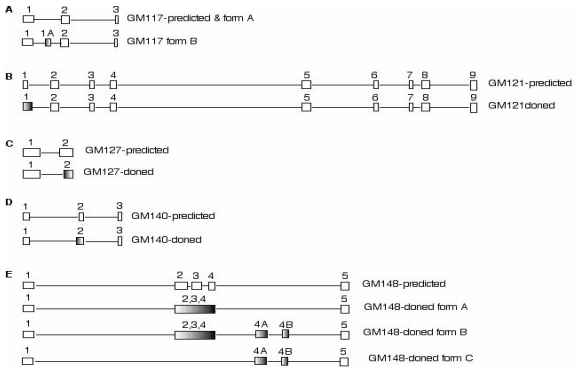
48

PREDICTION PROGRAMS PERFORMANCE

37 genes were tested, 16 of them (43%) were confirmed. At the exon level 159 exons were predicted and 58 (36%) were found to be real.

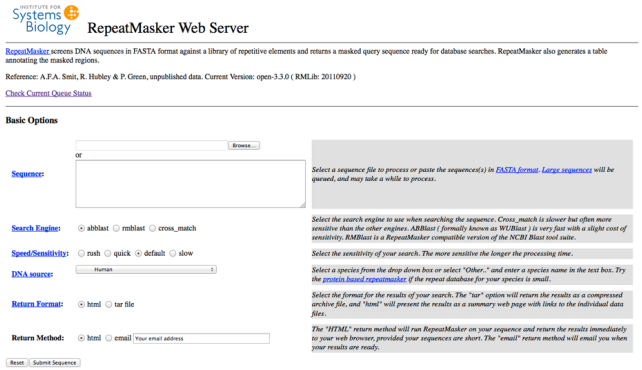
	predicted exons	specificity	sensitivity
MZEF	34	**0.51**	0.56
GRAIL	11	0.48	0.19
GENSCAN	52	0.46	**0.91**
FGENES	45	0.37	0.75

PROBLEMS RELATED TO GENE PREDICTION - GENE STRUCTURE AND ALTERNATIVE SPLICING



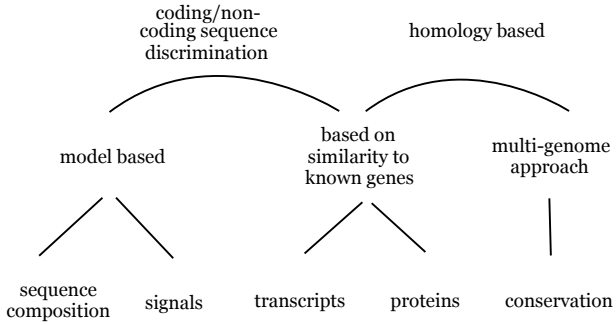
I. Makalowska et al. Gene 284: 203-213

RepeatMasker



http://www.repeatmasker.org/

GENE FINDING METHODS



GENE FINDING STRATEGIES

- Search for conserved regions
- Presence of ORF
- Codon usage
- Splice sites
- Polyadenylation signal
- Similarity search
- Presence of regulatory elements

WHY IS PROMOTER PREDICTION DIFFICULT?

- Not a one single type of core promoter
- Promoter needs additional regulatory elements
- Transcription may be activated or repressed by many regulatory proteins
- Transcriptional activators and repressors act very specifically both in terms of the cell type and point in the cell cycle
- Not all regulatory factors have been characterized

PROKARYOTIC PROMOTER PREDICTION

- Most bacterial promoters contain:
  - The Pribnow box, at about -10bp from the start codon there is consensus sequence: 5'-TATAAT-3'
  - The -35 sequence, centered about -35bp from the start codon there is consensus sequence: 5'-TTGACA

NNNTTGACANNNNNNNNNNNNNNNNNTATAATNNNNNATGcccccc

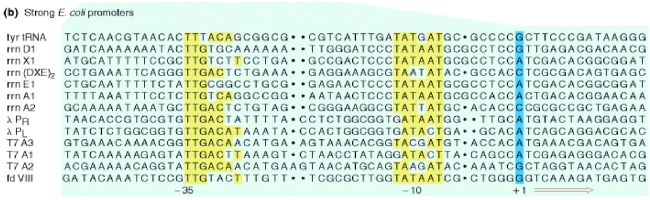
-35 region

-10 region

RNA start site

55

E.COLI PROMOTERS



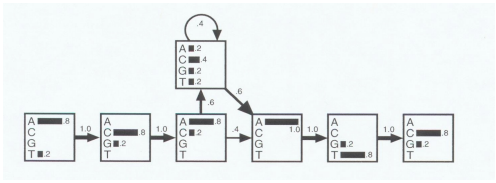
Promoters sequences can vary tremendously.

RNA polymerase in eukaryotes recognizes hundreds of different promoters

56

MARKOV MODELING - AGAIN

A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C



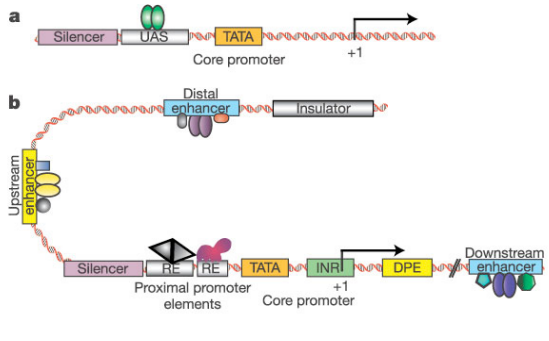
57

EUKARYOTIC PROMOTERS

- Three types of RNA polymerase (I, II, III), each binding to various kinds of promoters
- Polymerase II transcribes genes coding for proteins
- Core Promoter - most have TATA box that is centered around position -25 and has the consensus sequence: 5'-TATAAAA-3'
- Several promoters have a CAAT box around -90 with the consensus sequence: 5'-GGCCAATCT-3'
- promoters for "housekeeping" genes contain multiple copies of a GC-rich element that includes the sequence 5'-GGGCGG-3'
- Proximal Promoter Regions - transcription factor binding regions within ~200 bp of the Core Promoter
- Enhancers - transcription factor binding regions that can act to regulate transcription from the core promoter even from many kilobases away from the core promoter

58

EUKARYOTIC PROMOTERS



59

CISTER : CIS-ELEMENT CLUSTER FINDER

- Detects cis-elements clusters by using Hidden Markov Model
- For each element uses separate matrix with frequencies of each nucleotide in each position; user can input matrix for elements not included in the basic option
- User can specify:
  - distance between neighboring cis-elements within a cluster
  - number of cis-elements in the cluster
  - distance between clusters
  - half-width of the sliding window

60

EXAMPLE OF MATRIX

```
NA   AML-1a
XX
DE   runt-factor AML-1
XX
BF   T02256; AML1a; Species: human, Homo
sapiens.
XX
P0      A      C      G      T
01      5      1      2     49      T
02      2      2     52      1      G
03      4     14      1     38      T
04      0      0     57      0      G
05      1      0     55      1      G
06      1      4      0     52      T
```

Sequences of experimentally identified elements are aligned and frequencies in each position are calculated

61

EXAMPLE OF MATRIX

Sequences of experimentally identified elements are aligned and frequencies in each position are calculated

```
P1 P2 P3 P4 P5 P6
TGTGGT
TGCGGT
TGTGGT
AGTGGT
TGTGGC
```

```
NA   AML-1a
XX
DE   runt-factor AML-1
XX
BF   T02256; AML1a; Species: human, Homo sapiens.
XX
P0      A      C      G      T
01      1      0      0      4      T
02      0      0      5      0      G
03      0      1      0      4      T
04      0      0      5      0      G
05      0      0      5      0      G
06      0      1      0      4      T
```

62

HTTP://ZLAB.BU.EDU/~MFRITH/CISTER.SHTML

Cister : Cis-element Cluster Finder

Instructions

Paste a DNA sequence into the box or enter a GenBank identifier:

OR upload a DNA sequence from a file: Browse...

(Optional) Set subsequence From: To:

Choose a bunch of cis-elements:

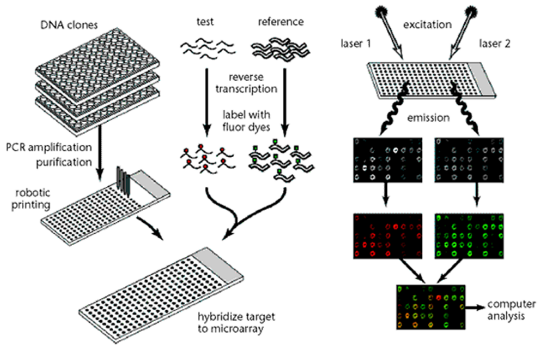
TATA Sp1 CRE ERE NF-1 E2F Mef-2 Myf
CCAAT AP-1 Ets Myc GATA LSF SRF Tef

AND / OR enter your own cis-elements:
(Get cis-element matrices from TRANSFAC - free registration required)

AND / OR upload cis-elements from a file: Browse...

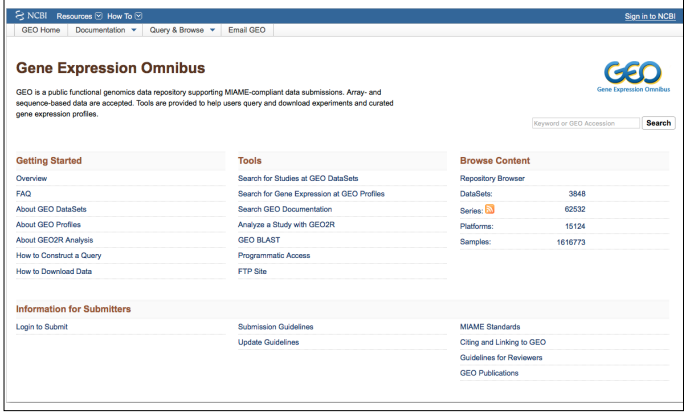
63

GENE EXPRESSION ANALYSIS - MICROARRAY



64

GENE EXPRESSION ANALYSIS - NCBI REPOSITORY



65

"Next Generation" Sequencers



		Total read length	Read length	Run time (day)	Run cost
Roche	GS FLX+	0.7Gb	700 bp	0.9	$15,000
Illumina	HiSeq2500	600Gb	100bp	12	$20,000
AB	SOLiD	500Gb	50 bp	7	$10,000
Heliscope	Helicos	37Gb	32 bp	8	?
Pacific Bioscience	PacBio-RS	50Mb	>10 kb	30min	$1,000

※Human Genome= 3Gb

66

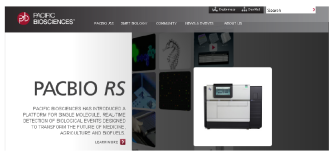
For personal use… (~1Gb sequencing)

Roch "GS Junior"

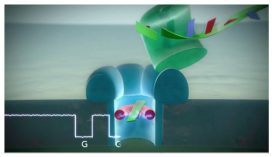
Illumina "MiSeq"

"IonTorrent (PGM)"

For single molecule sequencing (real time; long base)



BacBio



Nanopore

14

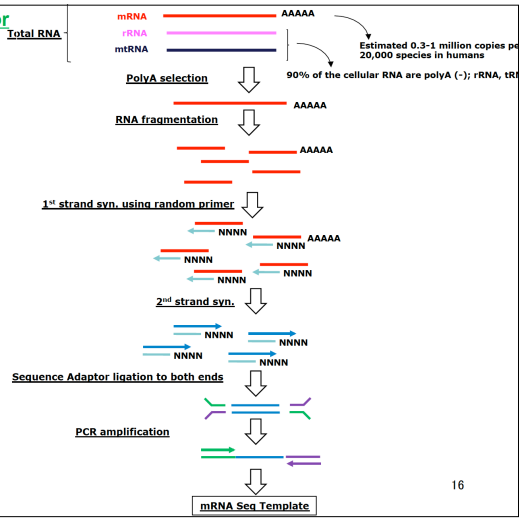
67

USE OF RNASeq

Task	Sequencing approach	Technique
Measuring of gene expression	36 bp single-end reads (tags counting)	RNASeq, small RNASeq
Measuring of gene interactions	36 bp single-end reads (tags counting)	RIP Seq CLIP Seq
Genome annotation	100 nt paired-end reads (sequence determinations)	RNASeq
Alternative splicing analysis	100 nt paired-end reads (sequence determinations)	RNASeq

68

Template Prep. for RNA Seq

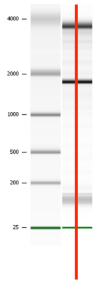


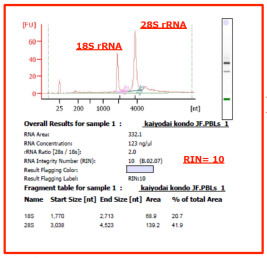
16

69

BioAnalyzer is essential for sample preparation

BioAnalyzer (Agilent):
Electrophoresis on microchip





RNA integrity number
(RIN value) for qualifying the RN

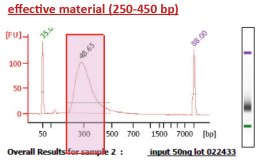
Dissection

17

70

Advantages in using BioAnalyzer (I)

effective material (250-450 bp)



Overall Results for sample 2 : input 50ng lot 022433

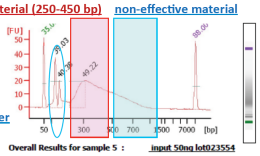
Number of peaks found: 1

Peak table for sample 2 : input 50ng lot 022433

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	285	48.18	256.0	

effective material (250-450 bp) non-effective material

Primer dimer



Overall Results for sample 5 : input 50ng lot023554

Number of peaks found: 3

Peak table for sample 5 : input 50ng lot023554

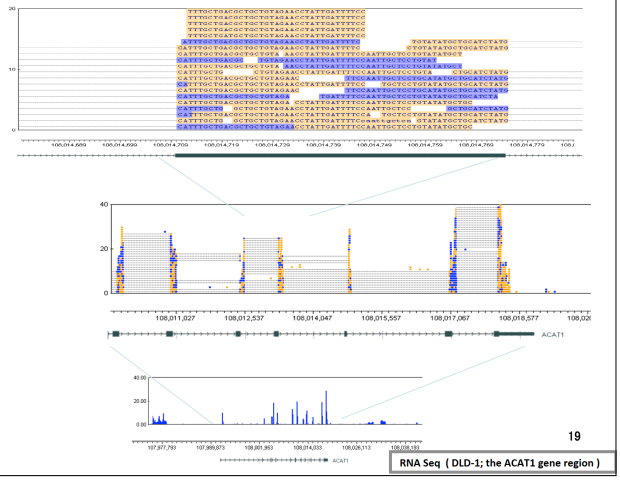
Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	125	0.84	82.7	
3	149	3.40	34.5	
4	307	12.36	61.0	

To measure effective template amount

18

71

Examples of NGS data (RNA Seq on Genome Studio Viewer)



19

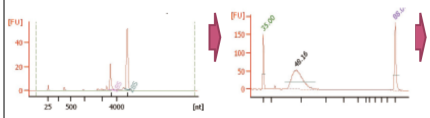
RNA Seq (DLD-1; the ACAT1 gene region)

72

Transcriptome Analysis of Tomatoes (young/old leaves)

Routine scheme for the RNA Seq

(1 μg total RNA) (450ng library) (0.2ng library)



36-base single-end read: 1 lane

De novo assemble (AbySS)

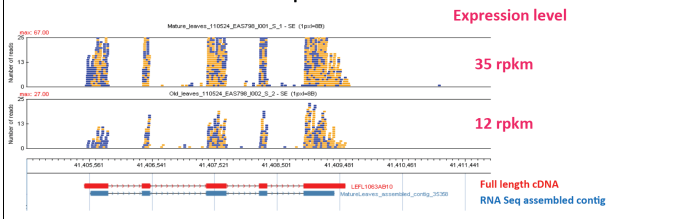
Sequence Summary

Tissue	# reads (36bp)	# Assembled contigs 500bp< / 1k < / 1.5k<	%Matched with cDNA 500bp< / 1k < / 1.5k<	%Matched with tBLASTX < 1e-50 500bp< / 1k < / 1.5k<
mature leaves	29,923,071	7,165/ 2,304/834	4,648/1,456/467	6,866/ 2,280/828
old leaves	28,711,676	6,118/1,890/653	4,001/1,199/361	5,869/1,871/649

24

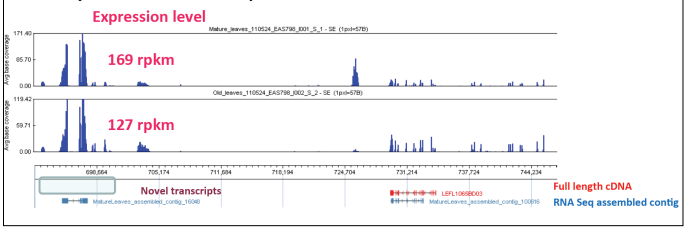
73

Once assembled to full-length cDNA, the tag count could be also used as an indicator of its expression levels



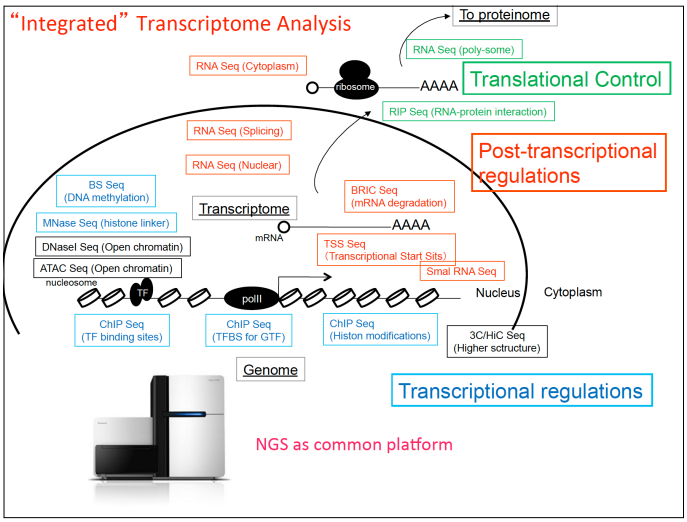
(rpkm: read per million tags per kb mRNA)

Discovery of the novel transcripts



74

"Integrated" Transcriptome Analysis



75

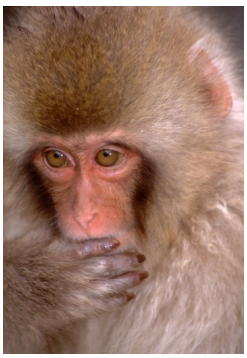
Bioinformatic tools for RNA Seq

Purpose	Software	URL
Mapping	BWA	http://bio-bwa.sourceforge.net/
	Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
	TopHat2	http://tophat.cbcb.umd.edu/
Expression	Cufflinks	http://cufflinks.cbcb.umd.edu/
	Cuffdiff	Same as above
	DEseq	http://bioconductor.org/packages/release/bioc/html/DESeq.html
Fusion genes	TopHat-fusion	http://tophat.cbcb.umd.edu/fusion_index.html
	deFuse	http://compbio.bccrc.ca/software/defuse/
	SOAPfuse	http://soap.genomics.org.cn/soapfuse.html
Assemble	Trans-Abyss	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
	Trinity	http://trinityrnaseq.sourceforge.net/
Viewer	UCSC Genome Browser	http://genome.ucsc.edu/cgi-bin/hgGateway
	IGV	https://www.broadinstitute.org/igv/home

76

BIOINFORMATICS CREED

- Remember about biology
- Do not trust the data
- Use comparative approach
- Use statistics
- Know the limits
- Remember about biology!!!



77