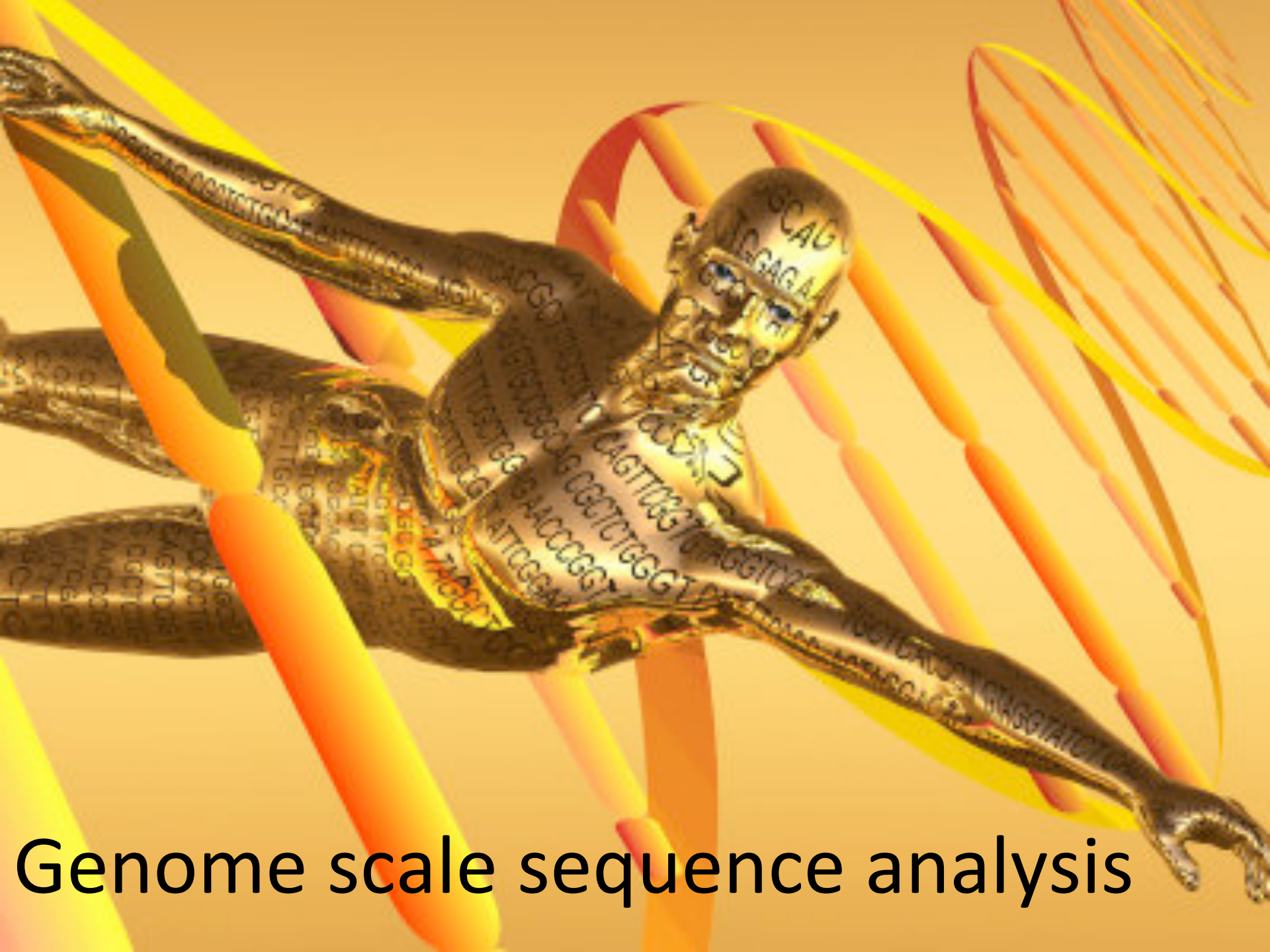
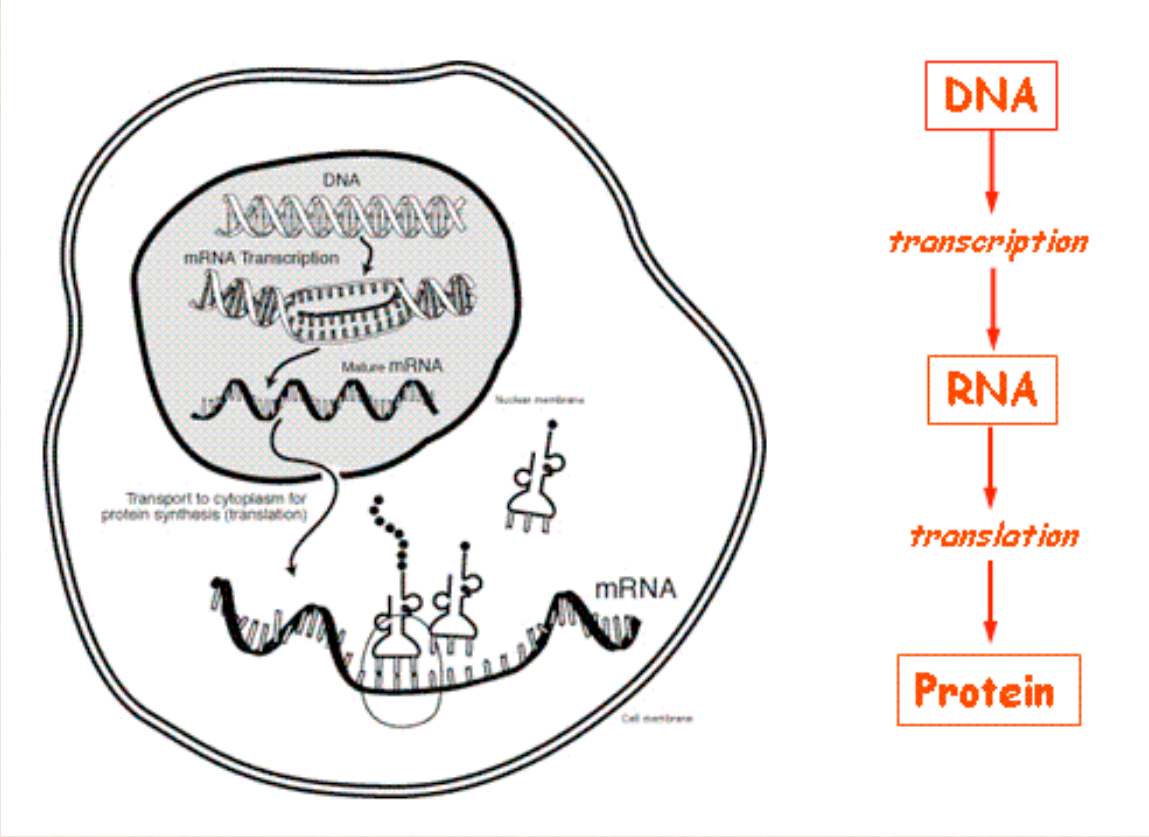
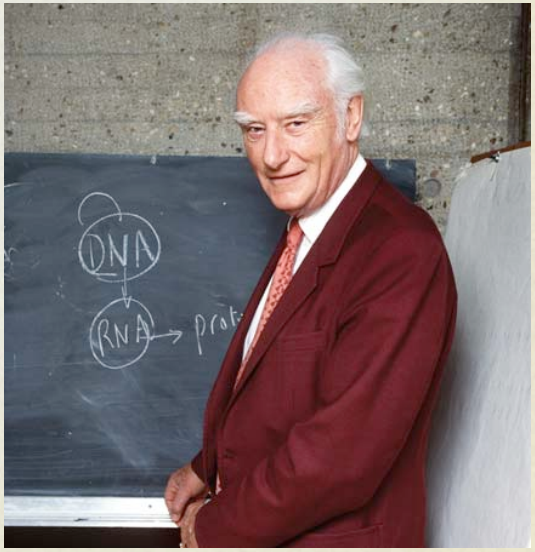

Genome scale sequence analysis

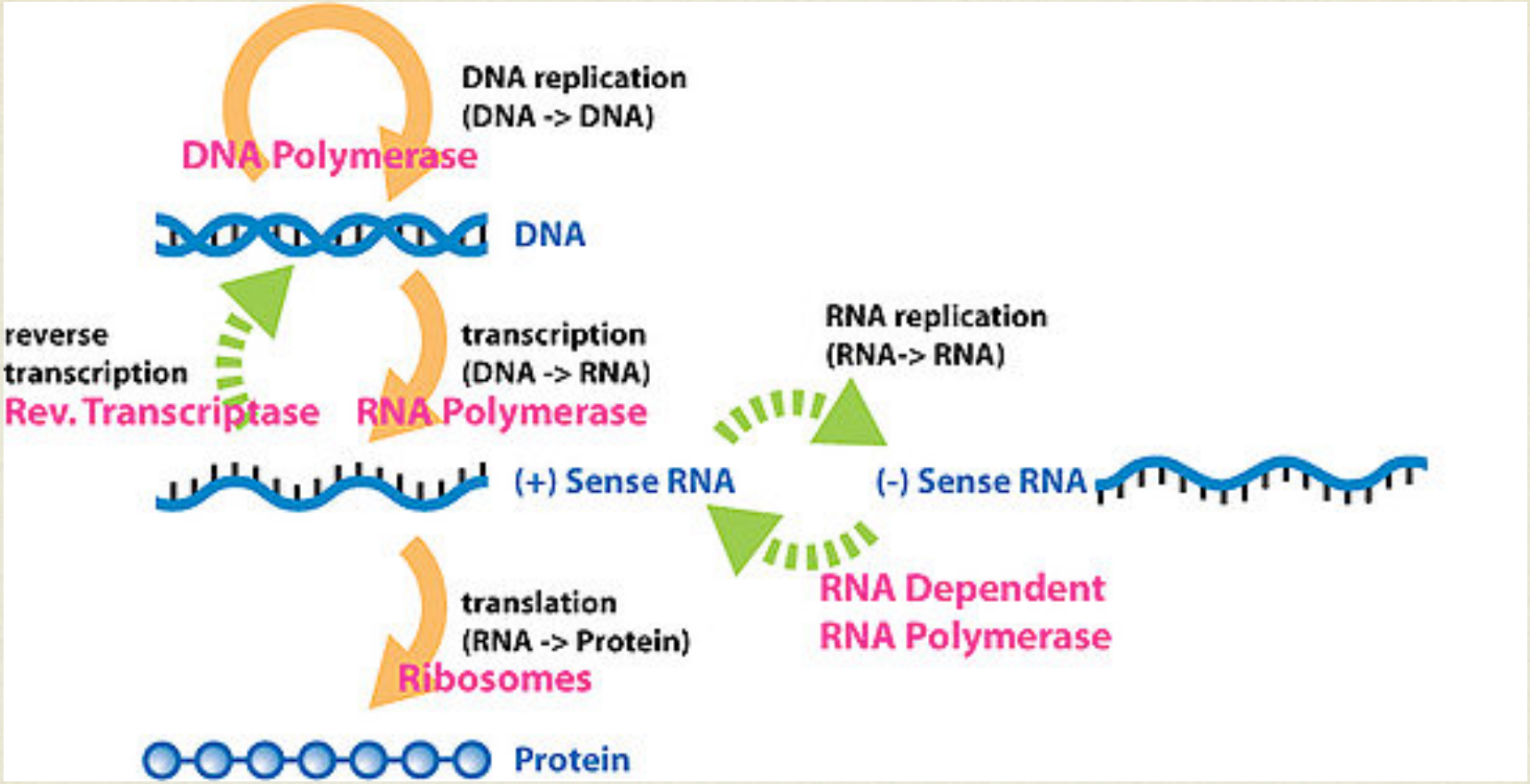
CENTRAL DOGMA OF MOLECULAR BIOLOGY





Classical definition

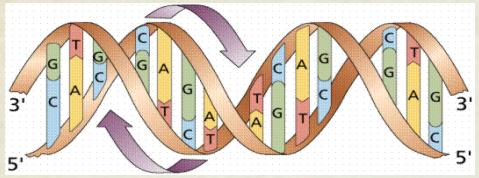
CENTRAL DOGMA OF MOLECULAR BIOLOGY



Modern definition

http://www.thefullwiki.org/

CENTRAL DOGMA OF MOLECULAR BIOLOGY



DNA

transcription

RNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

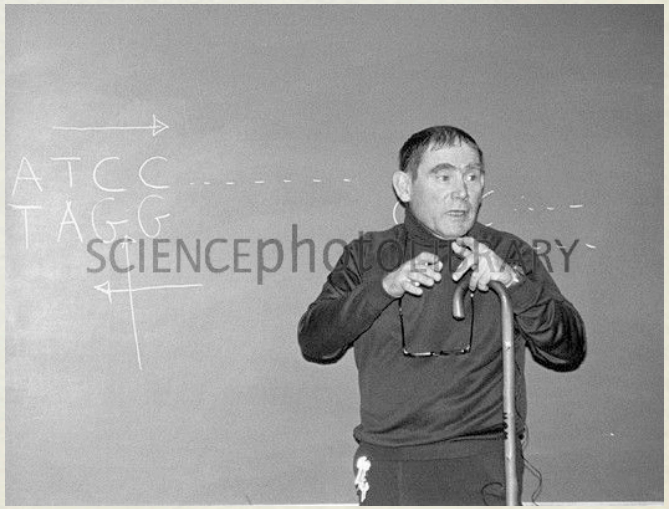
PEPTIDE

HISTORICAL DIGRESSION - DISCOVERY OF CODONS

Crick, Brenner et al. experiment

In the experiment, proflavin-induced mutations of the T4 bacteriophage gene, rIIB, were isolated. Proflavin causes mutations by inserting itself between DNA bases, typically resulting in insertion or deletion of a single base pair.

- Deletion of a single or two nucleotides changed the protein
- Deletion of three nucleotides didn't change the protein significantly
- CONCLUSION: each nucleotide triplet codes for a single amino acid



Crick FH, Barnett L, Brenner S, Watts-Tobin RJ (1961). "General nature of the genetic code for proteins". Nature 192: 1227–32.

HISTORICAL DIGRESSION - DISCOVERY OF CODONS

Let's consider the following sentence

THE SLY FOX AND THE SHY DOG

Let's remove one, two, or three letters after first "S"

THE SYF OXA NDT HES HYD OG

THE SFO XAN DTH ESH YDO G

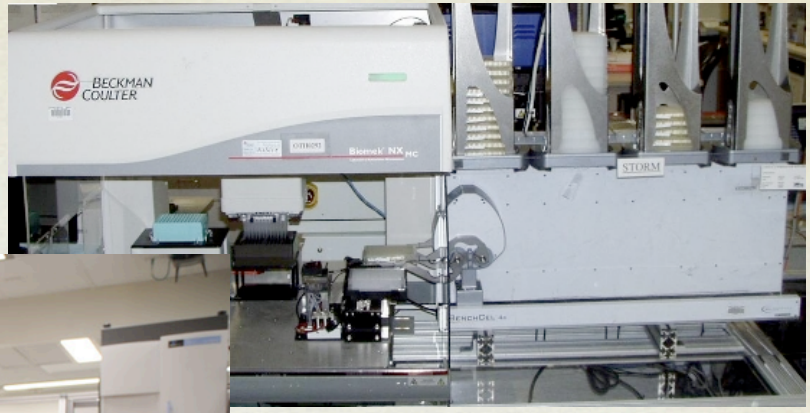
THE SOX AND THE SHY DOG

Which of the resulted sentences make sense?

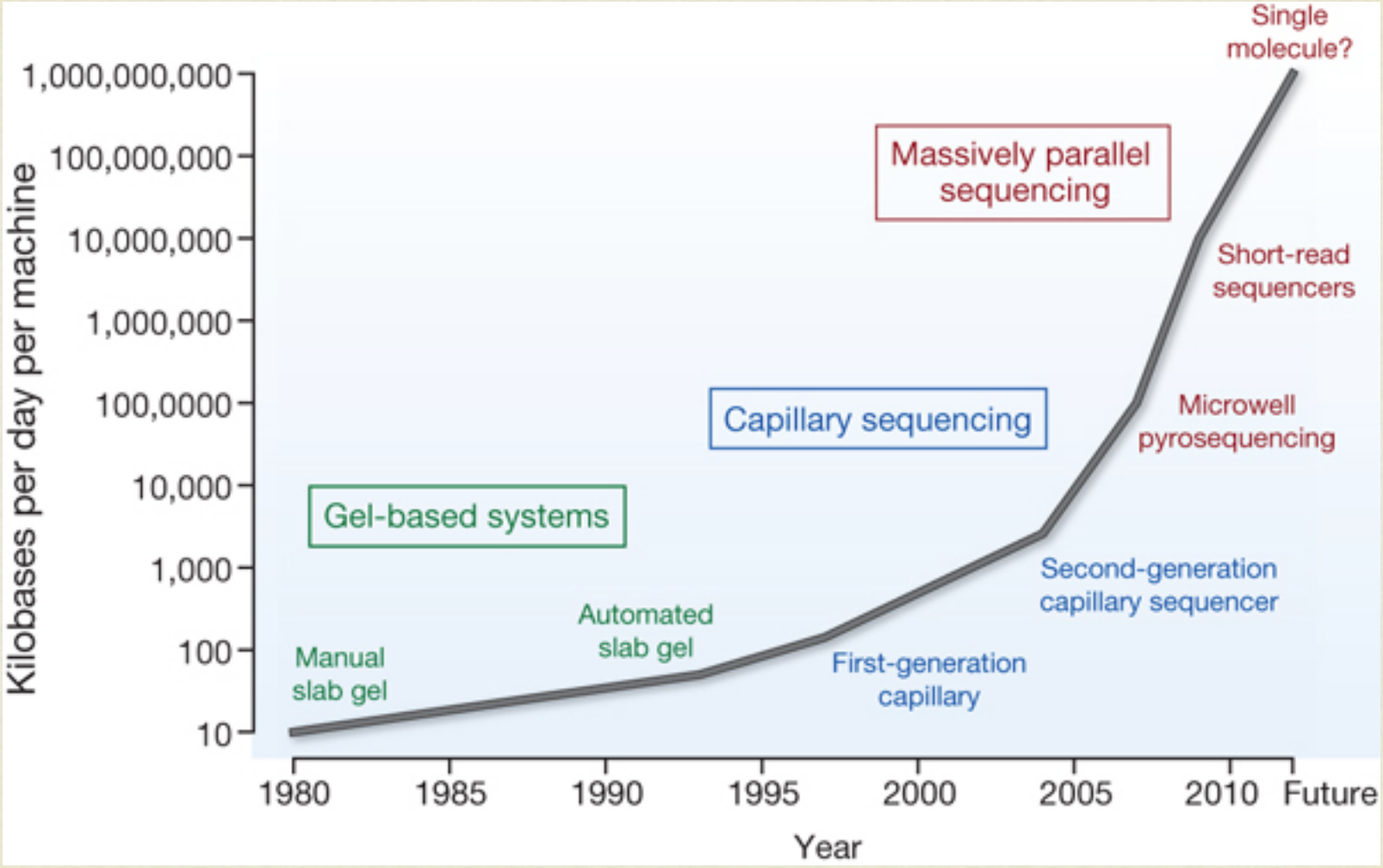
GETTING SEQUENCES

TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT

TECHNOLOGY MEETS BIOLOGY



IMPROVING TECHNOLOGY



MR Stratton *et al. Nature* **458**, 719-724 (2009)

EXISTING SEQUENCING TECHNOLOGIES (2012)

	Year of introduction	Amplifi-cation	Sequen-cing	Gb per run	Run time	Read length	Run price ($)
Sanger	1977	PCR	Sequencing by synthesis			1000 bp	
454	2004	Emulsion PCR	Pyrose-quencing	0.7	1 day	1000 bp	6,000
Illumina	2006	Bridge amplific.	Sequencing by synthesis	600	11 days	150 bp	23,000
SOLiD	2008	Emulsion PCR	Ligation-based	240	10 days	50 bp	5,000
Ion Torrents	2010	Emulsion PCR	Ion semi-conductor	10	2 hours	200 bp	1,000
PacBio	2011	none	Single molecule seq. by synthesis	0.04	40 min	3000 bp	100

http://blueseq.com/knowledgebank/

GENOME SEQUENCING PROJECTS

Organism	Complete	Draft assembly	In progress	Total
Prokaryotes	1117	966	595	2678
Archaea	100	5	48	153
Bacteria	1017	961	547	2525
Eukaryotes	36	319	294	649
Animas	6	137	106	249
Plants	5	33	80	118
Fungi	17	107	59	183
Protists	8	39	46	93

As of 16/02/2012

http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

GETTING SEQUENCES

TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT

READING ≠ UNDERSTANDING

Maturescentium autem uvae vehementiores nigrae — ideo vinum et his minus iucundum —, suaviores albae, quoniam e tralucido facilius accipitur aër. Recentes stomachum et spiritus inflatione alvum turbant. Itaque in febri damnantur, utique largiores; gravedinem enim capiti morbumque lethargum faciunt.

Gaius Plinius Secundus - *Naturalis Historiæ*

READING ≠ UNDERSTANDING

We shall best understand the probable course of natural selection by taking the case of a country undergoing some physical change. If the country were open were open on its borders, new forms would certainly immigrate, and this also would bla, bla bla become extinct inhabitants.

Charles Darwin - *The Origin of Species*

READING ≠ UNDERSTANDING

We shall best understand the probable course of [redacted] by taking the case of a country undergoing some physical change. If the country were open were open on its borders, new forms would certainly immigrate, and this also would bla, bla bla become extinct inhabitants.

Charles Darwin - *The Origin of Species*

CHALLENGE: HOW FROM THIS…

TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT

Infer this

TWO APPROACHES TO GENOME SEQUENCING

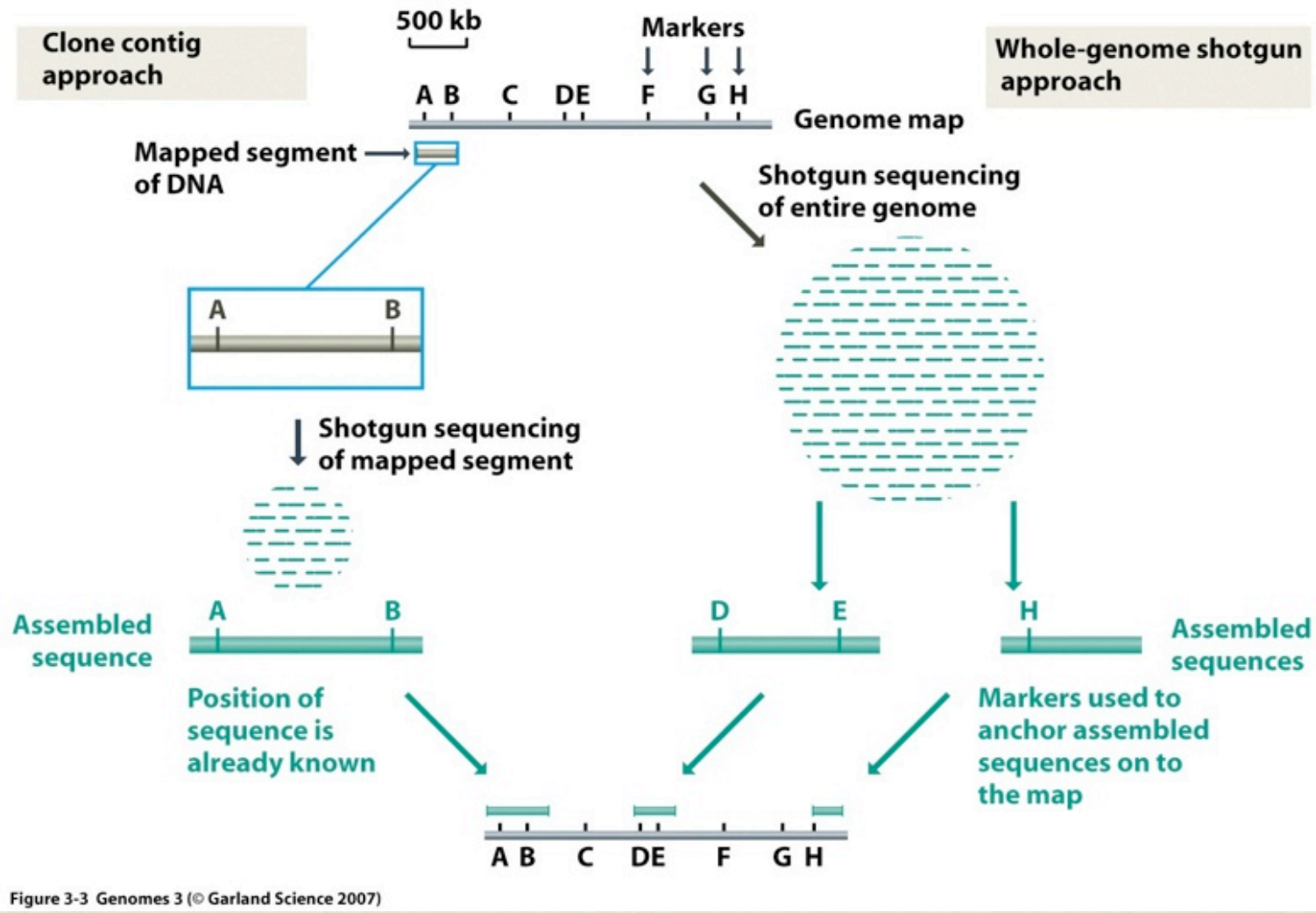
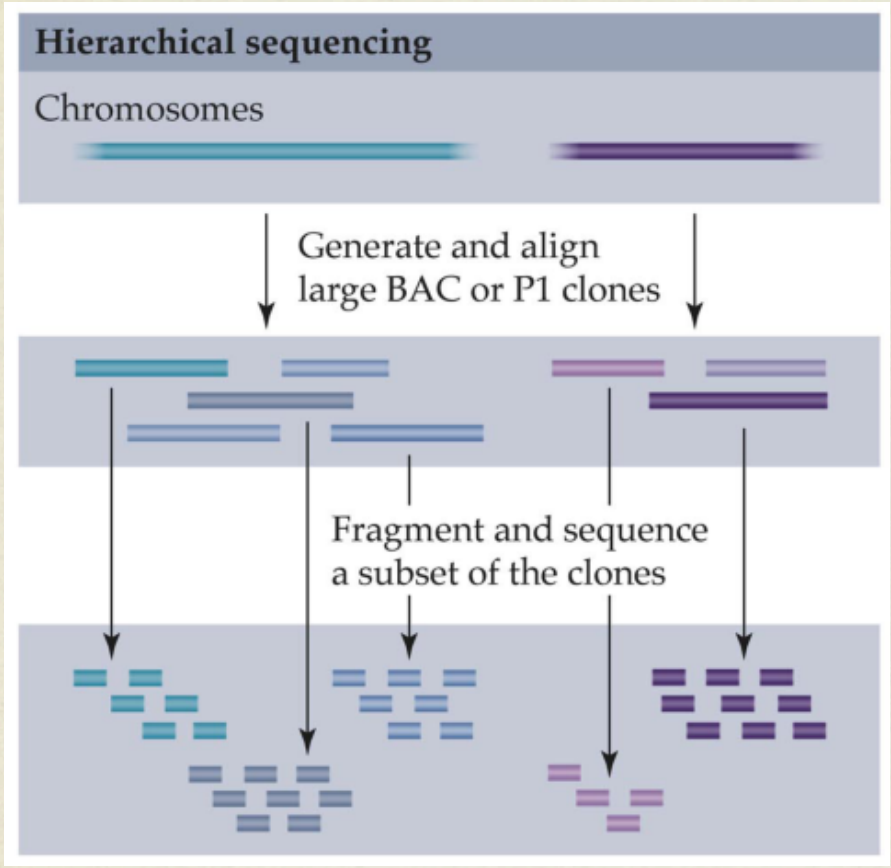
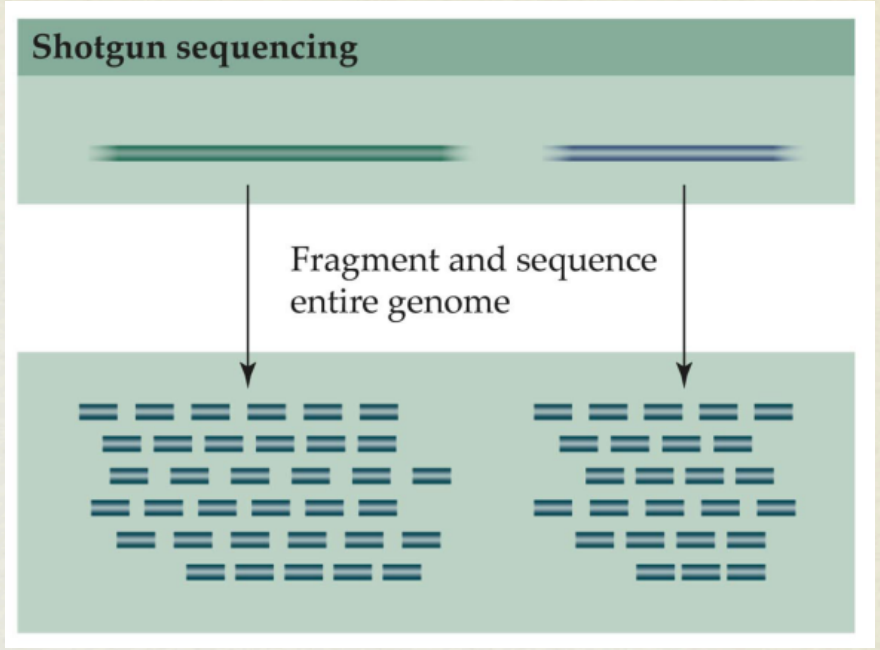


Figure 3-3 Genomes 3 (© Garland Science 2007)

TWO APPROACHES TO GENOME SEQUENCING



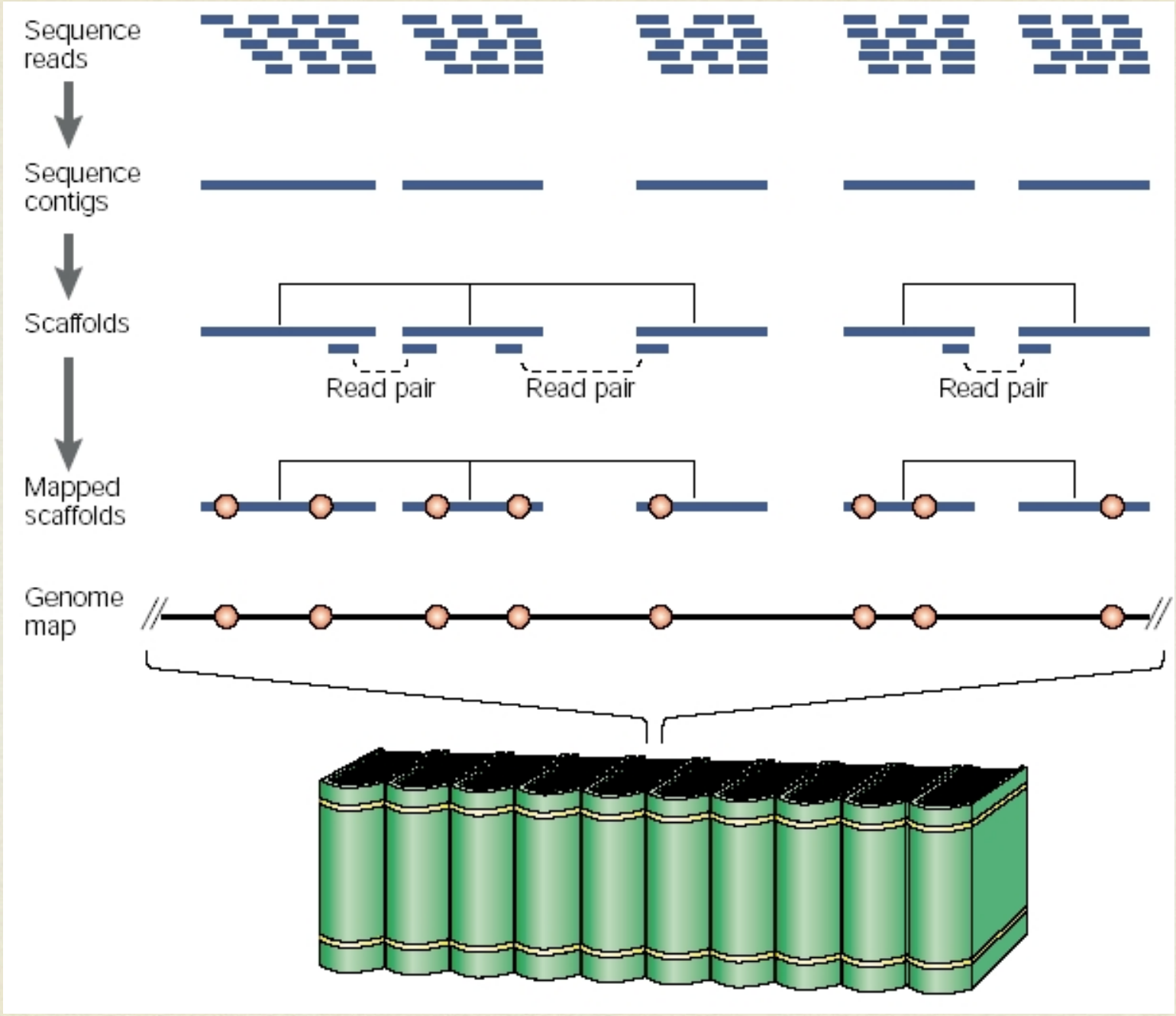


A PRIMER OF GENOME SCIENCE, Second Edition, Figure 2.7 (Part 2) © 2006 Sinauer Associates, Inc.


GIANT PUZZLE

SEQUENCE ASSEMBLY

- A fundamental goal of DNA sequencing has been to generate large, continuous regions of DNA sequence
- Capillary sequencing reads ~600-800bp in length
  - Overlap based assembly algorithms (phrap, phusion, arachne)
  - Compute all overlaps of reads and then resolve the overlaps to generate the assembly
- Volume and read length of data from next-gen sequencing machines meant that the read-centric overlap approaches were not feasible
  - already in 1980's Pevzner et al. introduced an alternative assembly framework based on de Bruijn graph
  - Based on a idea of a graph with fixed-length subsequences (k-mers)
  - Key is that not storing read sequences – just k-mer abundance information in a graph structure



Green, E.D. (2001) NRG 2, 573-583

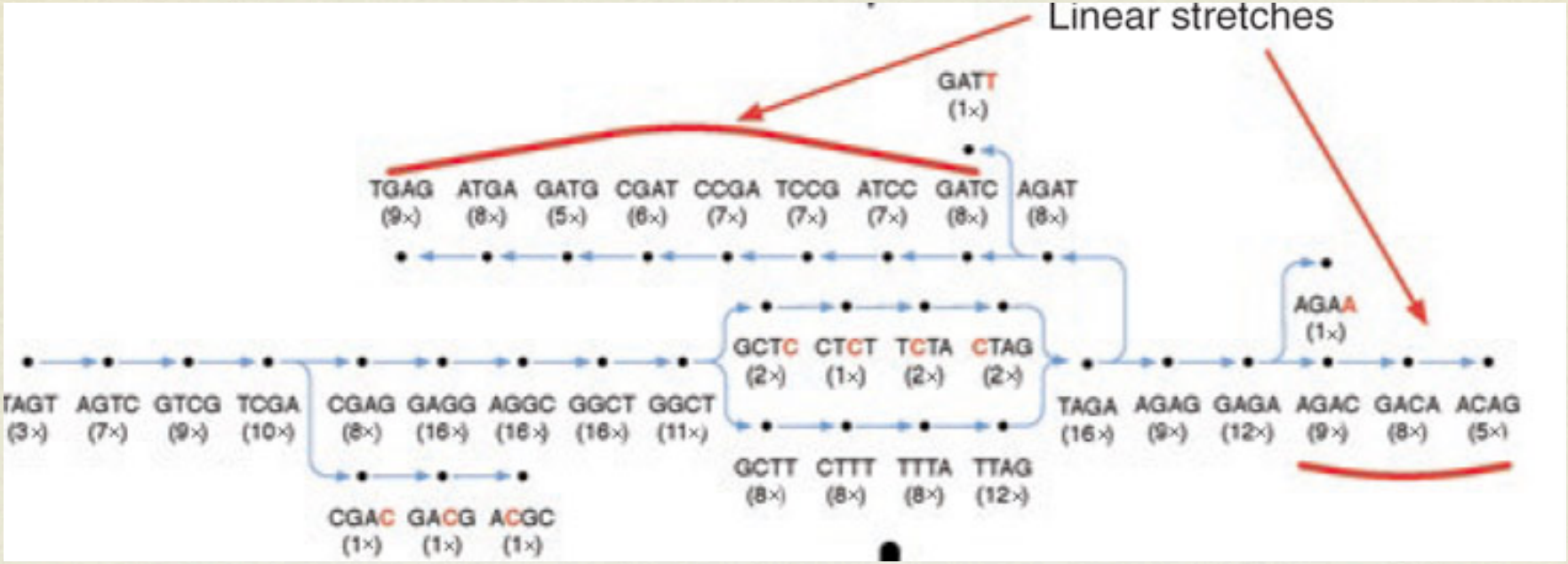
De Bruijn GRAPH CONSTRUCTION

Genome is sampled with random sequencing of for example 7 bp reads.

Note errors in the reads are represented in red

Flicek & Birney (2009) Nat Meth, 6: S6-S12.

De Bruijn GRAPH CONSTRUCTION

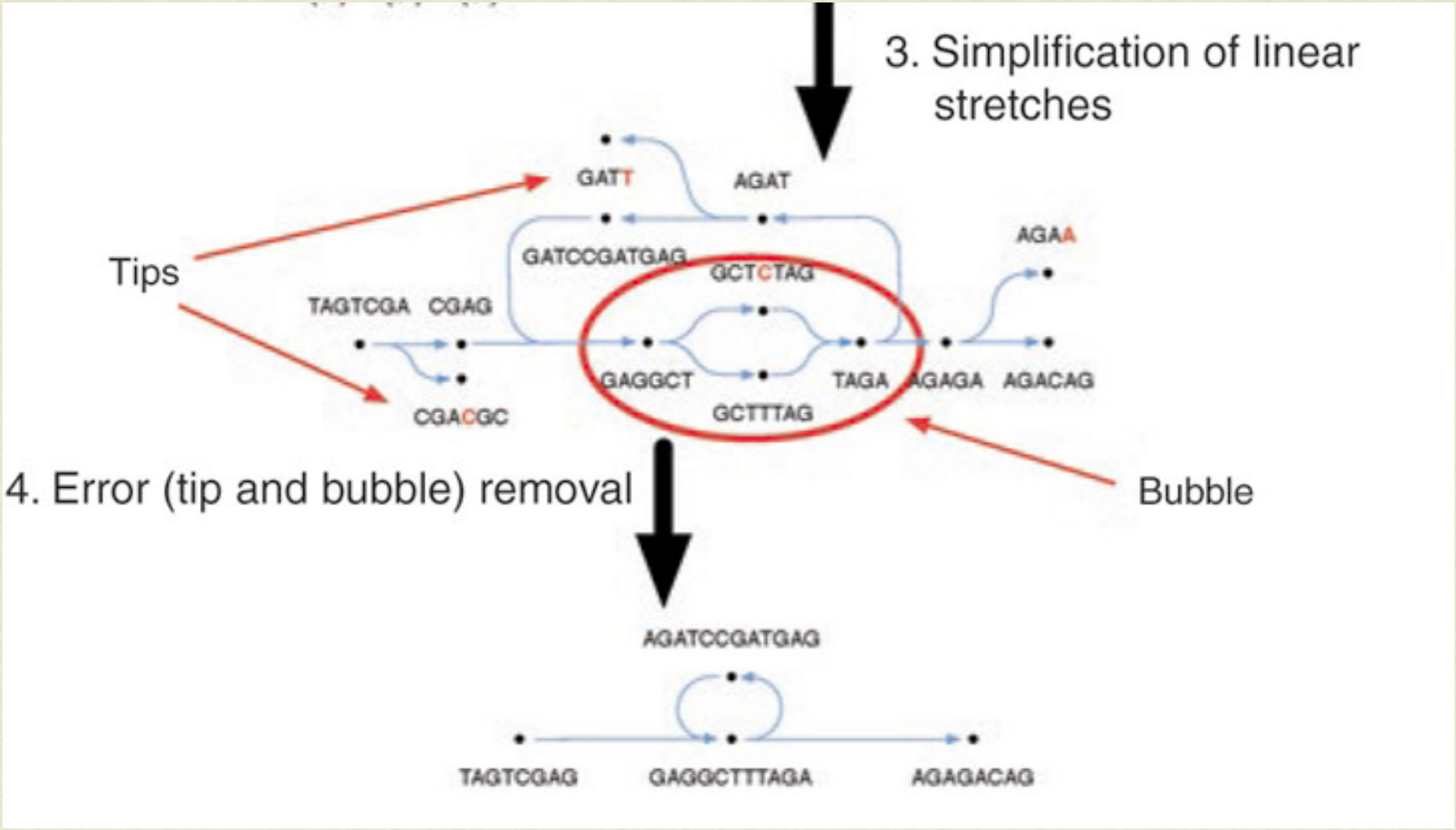


The k-mers in the reads (4-mers in this example) are collected into nodes and the coverage at each node is recorded (numbers at nodes).

Features:

- continuous linear stretches within the graph
- Sequencing errors are low frequency tips in the graph

Flicek & Birney (2009) Nat Meth, 6: S6-S12.



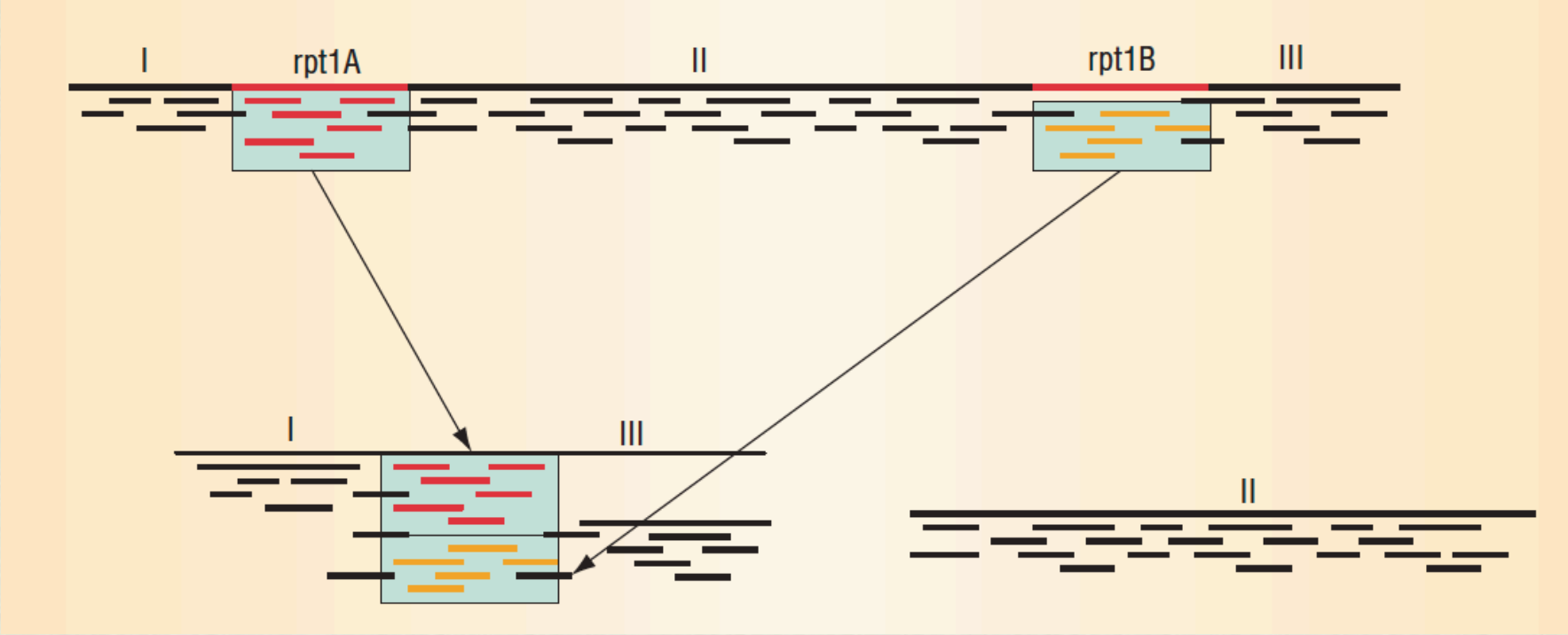
Graph is simplified to combine nodes that are associated with the continuous linear stretches into single, larger nodes of various k-mer sizes.

Error correction removes the tips and bubbles that result from sequencing errors.

Final graph structure that accurately and completely describes in the original genome sequence

Flicek & Birney (2009) Nat Meth, 6: S6-S12.

REPEATS PROBLEM

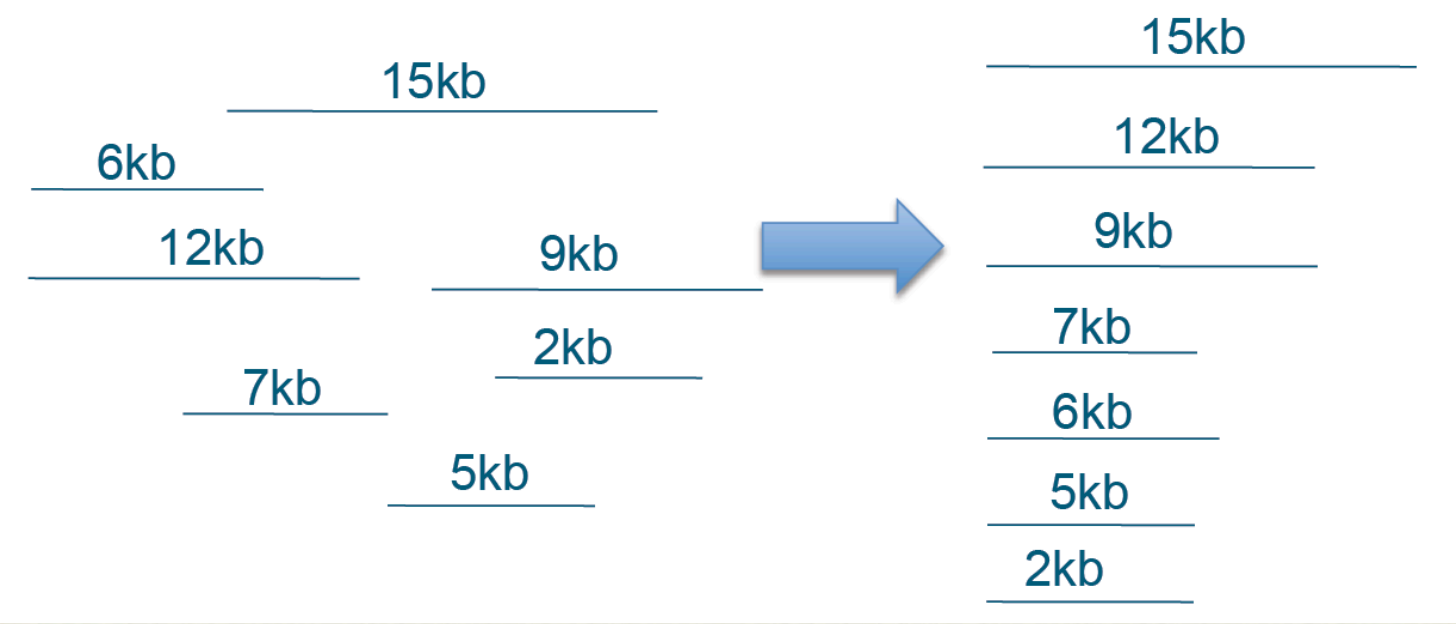


Very similar sequences may lead to false assembly, especially if the repeated region is longer than average reads length, e.g. recent tandem duplications or recent transpositions of mobile elements.

NEXT-GEN ASSEMBLERS

- First de Bruijn based assembler was Newbler developed by 454 Life Scinces
  - Adapted to handle main source of error in 454 data – indels in homopolymer tracts
- Many de Bruijn assemblers subsequently developed
  - SHARCGS, VCAKE, VELVET, EULER-SR, EDENA, ABySS and ALLPATHS, SOAP
  - Most can use mate-pair information
- Slightly different approach to transcriptome assembly
  - It has to allow many discontinuous graphs representing single transcript, including paralogs and alternatively spliced ones.
  - SOAP-Trans, Trinity

ASSEMBLY EVALUATION - N50

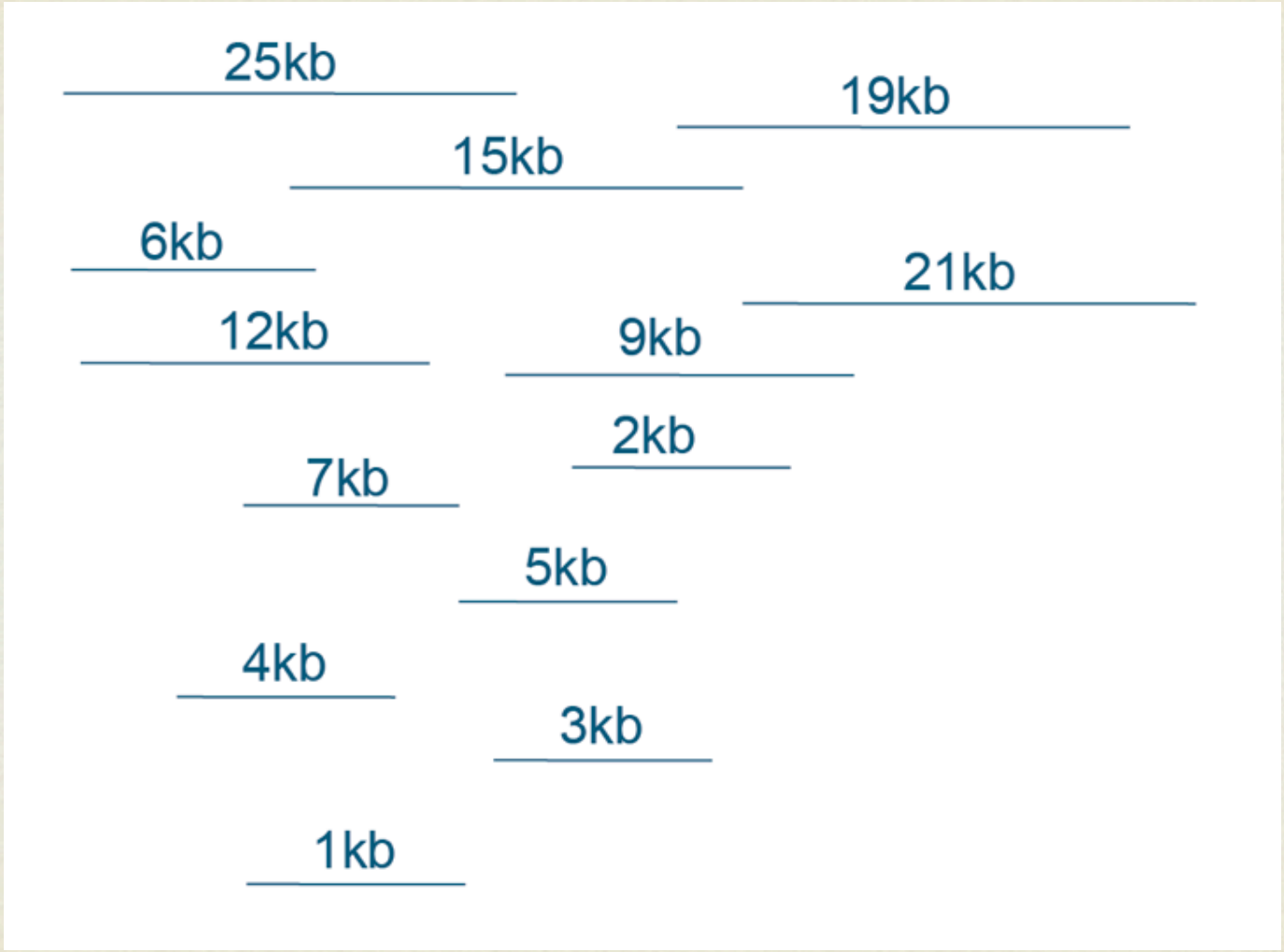


If one orders the set of contigs produced by the assembler by size, then N50 is the size of the contig such that 50% of the total bases are in contigs of equal or greater size.

15+12+9+7+6+5+2 = 56

56/2 = 28 -> N50 is 9kb (15+12 = 27 is less than 50%)

EXERCISE: CALCULATE N50



CHALLENGE: HOW FROM THIS…

TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT

Infer this

GENOME ANNOTATOTION

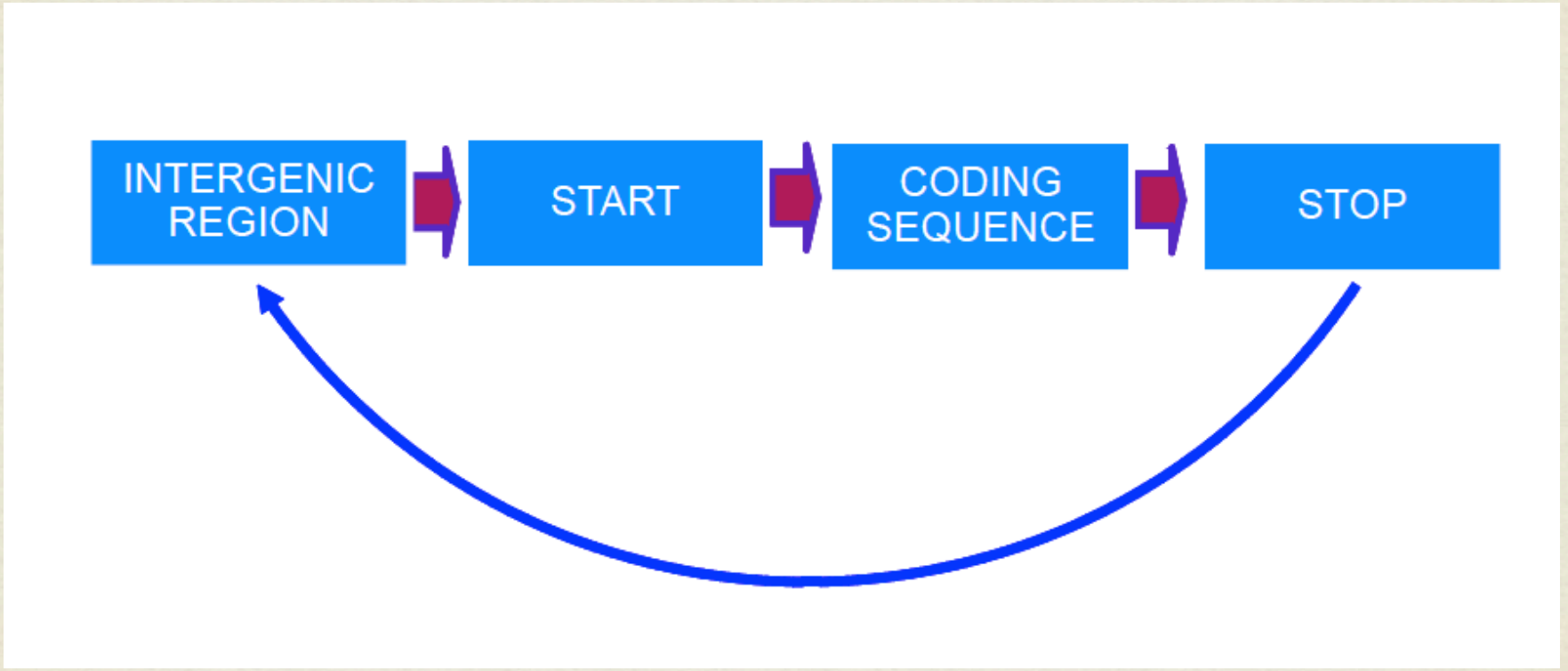
What are we looking for?

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

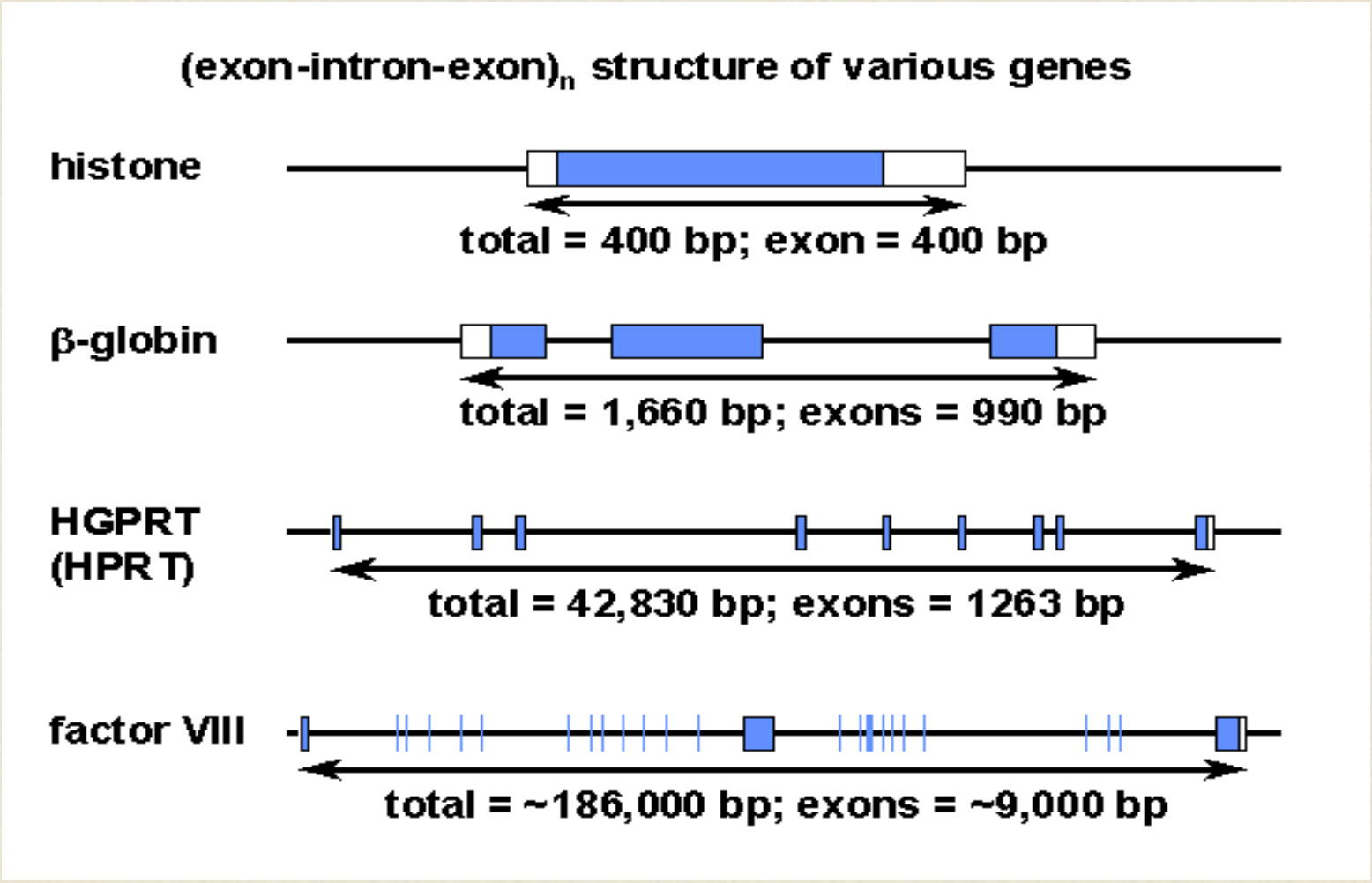
GENE IDENTIFICATION METHODS

- 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

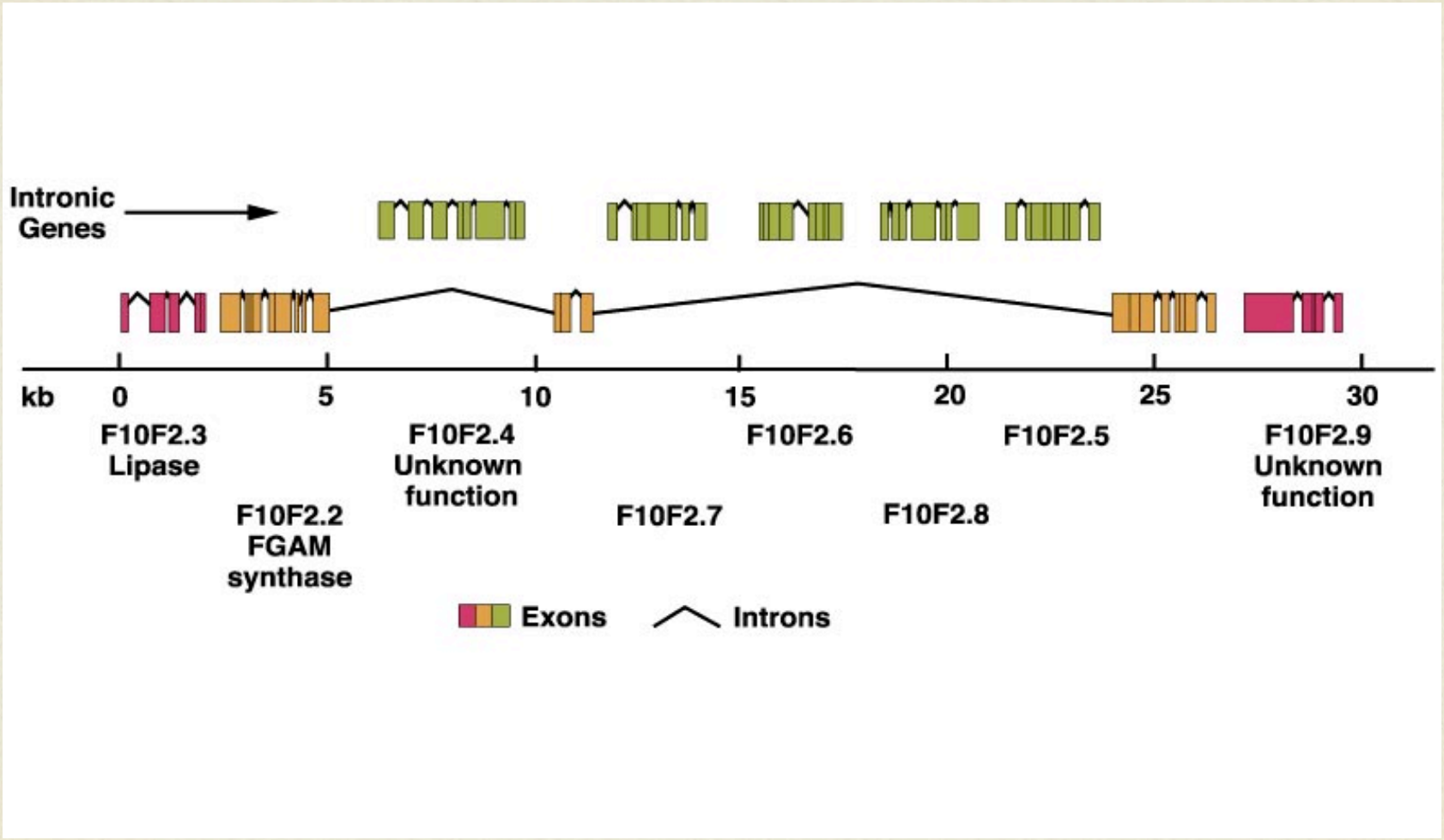
GENERAL MODEL OF A GENE



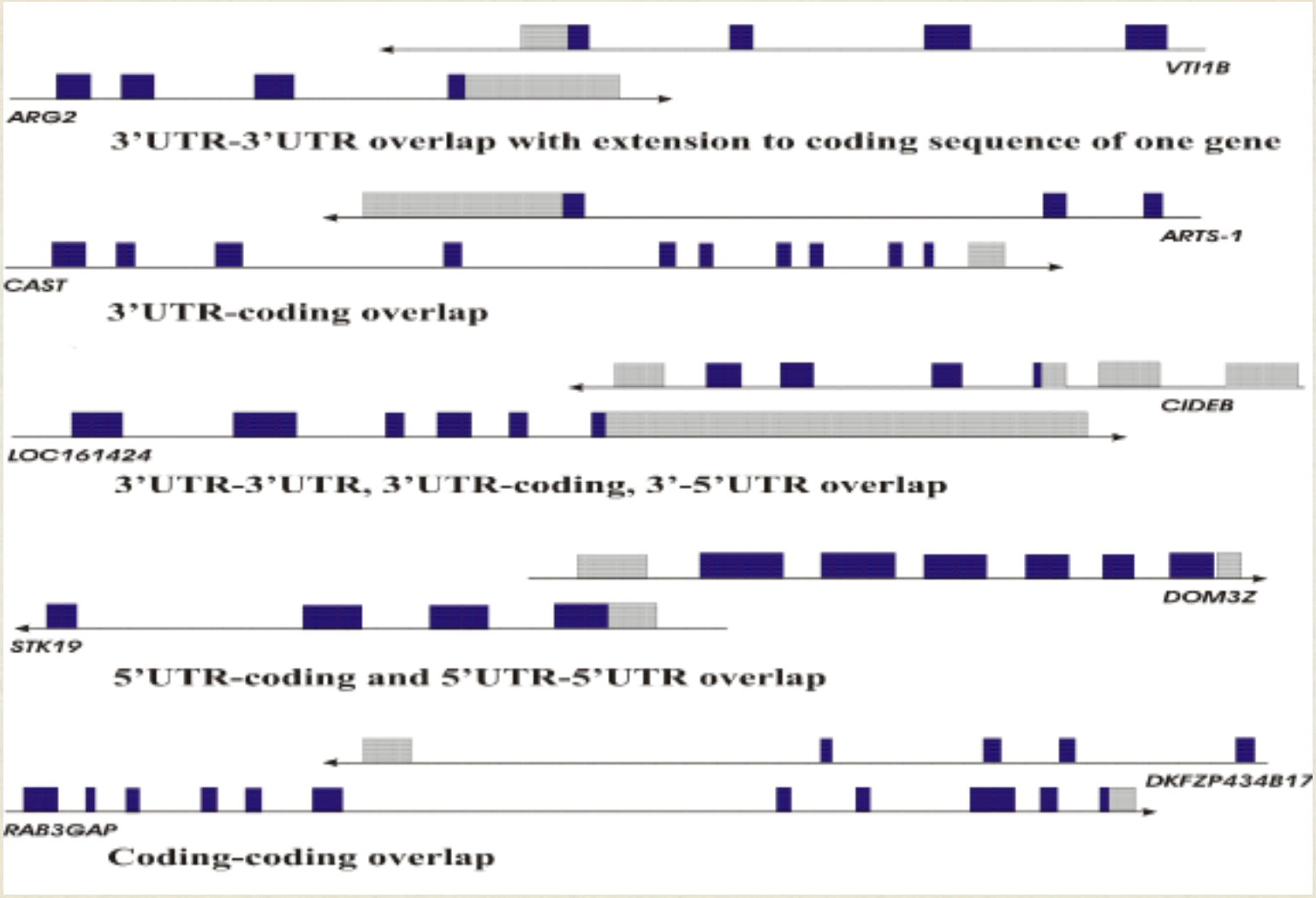
EUKARYOTIC GENE STRUCTURE



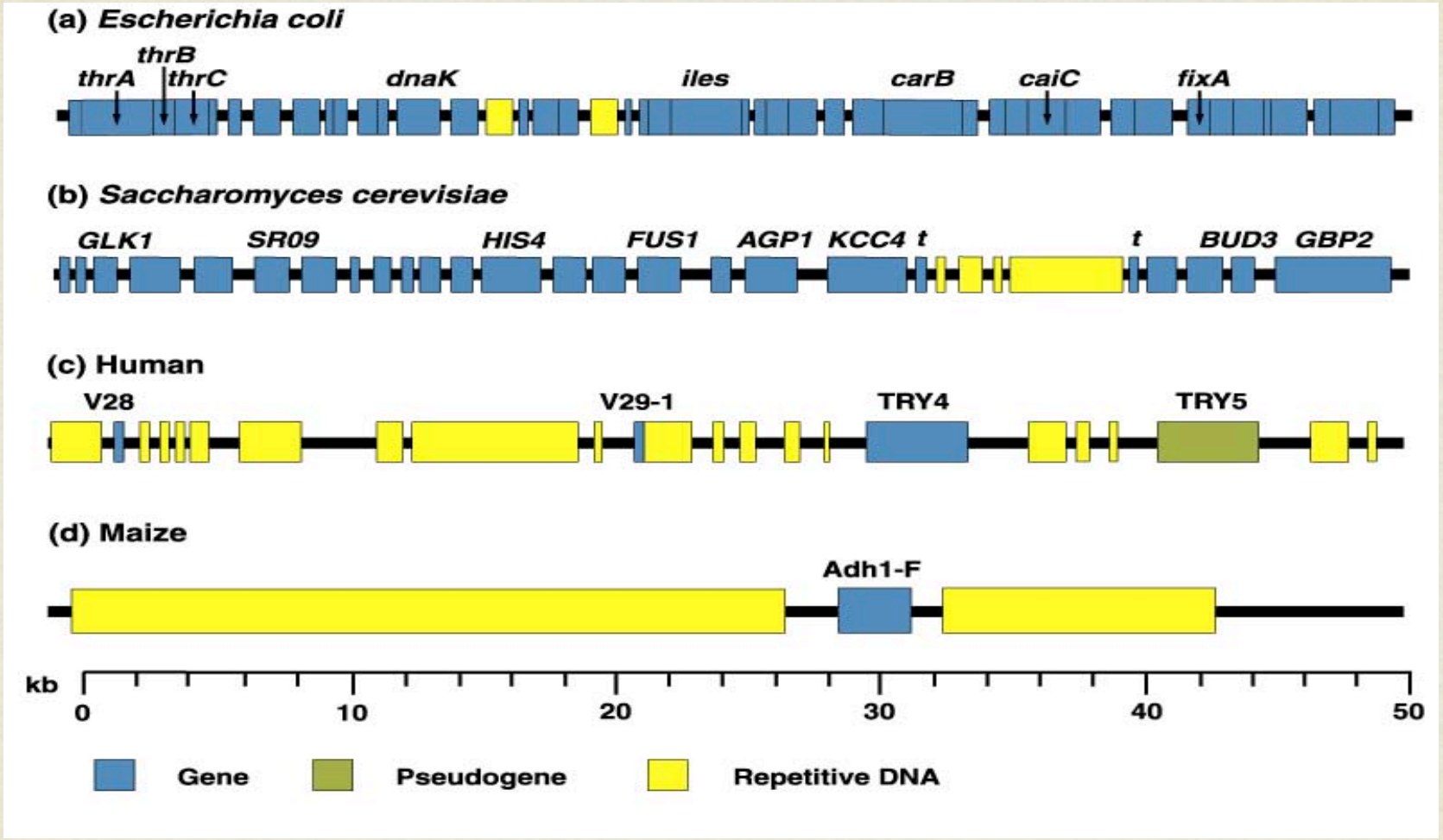
NESTED GENES



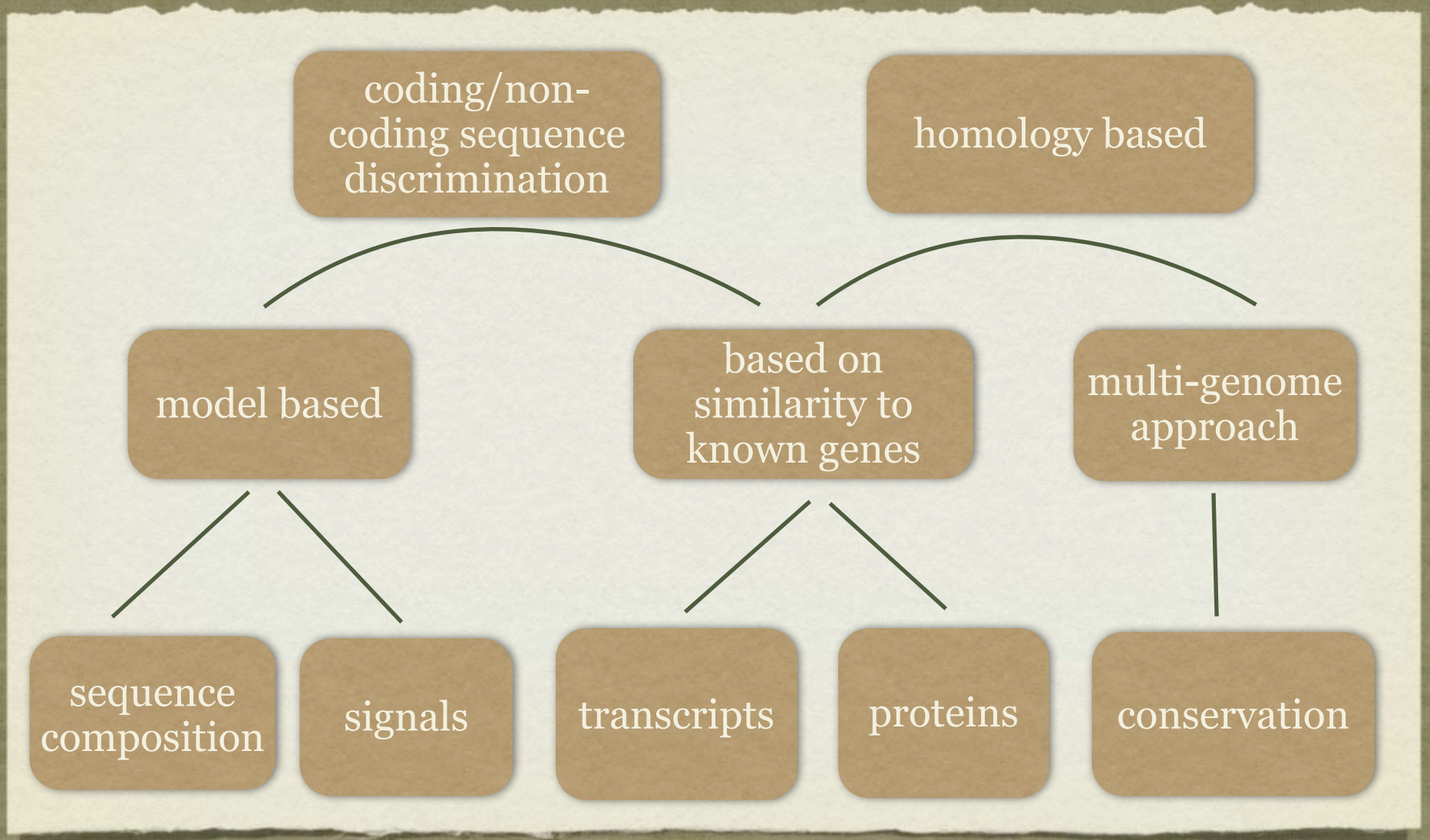
OVERLAPPING GENES



PSEUDOGENES AND REPETITIVE ELEMENTS



GENE FINDING METHODS

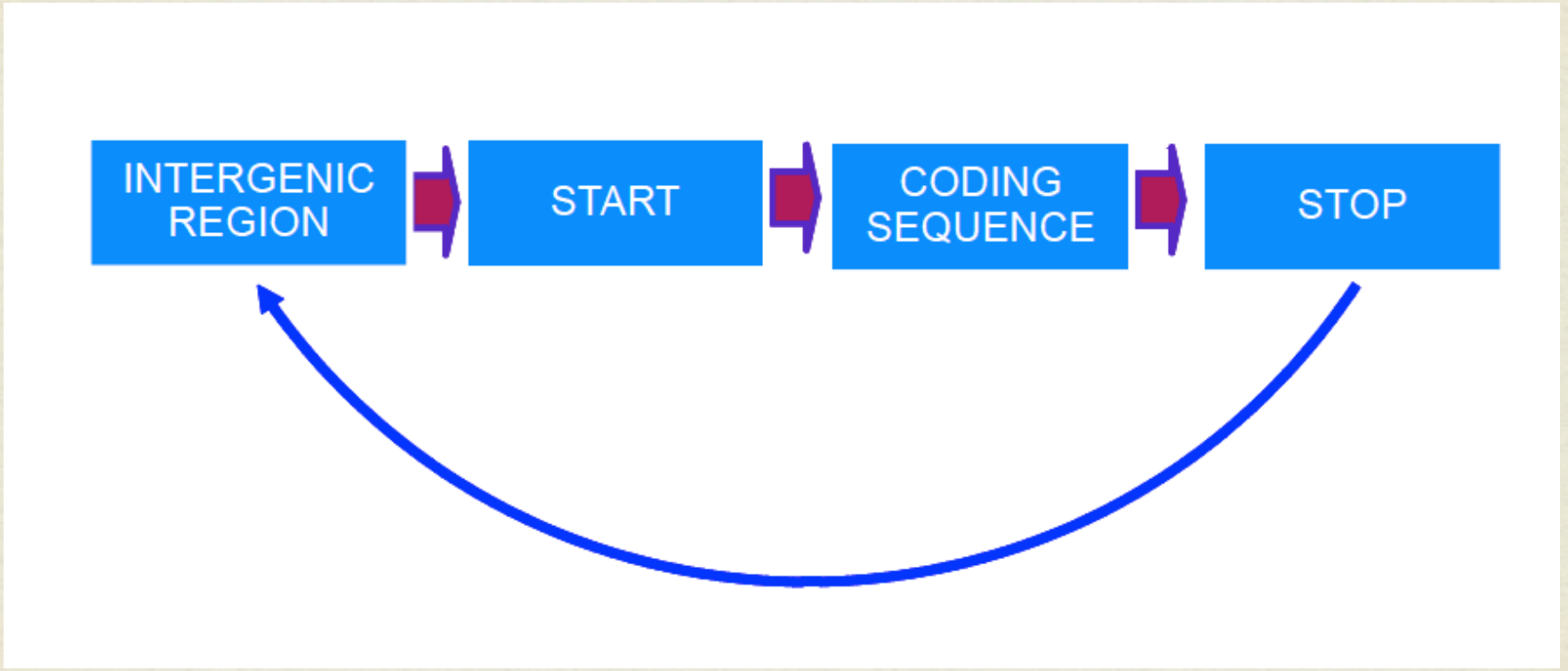


MODEL BASED METHODS

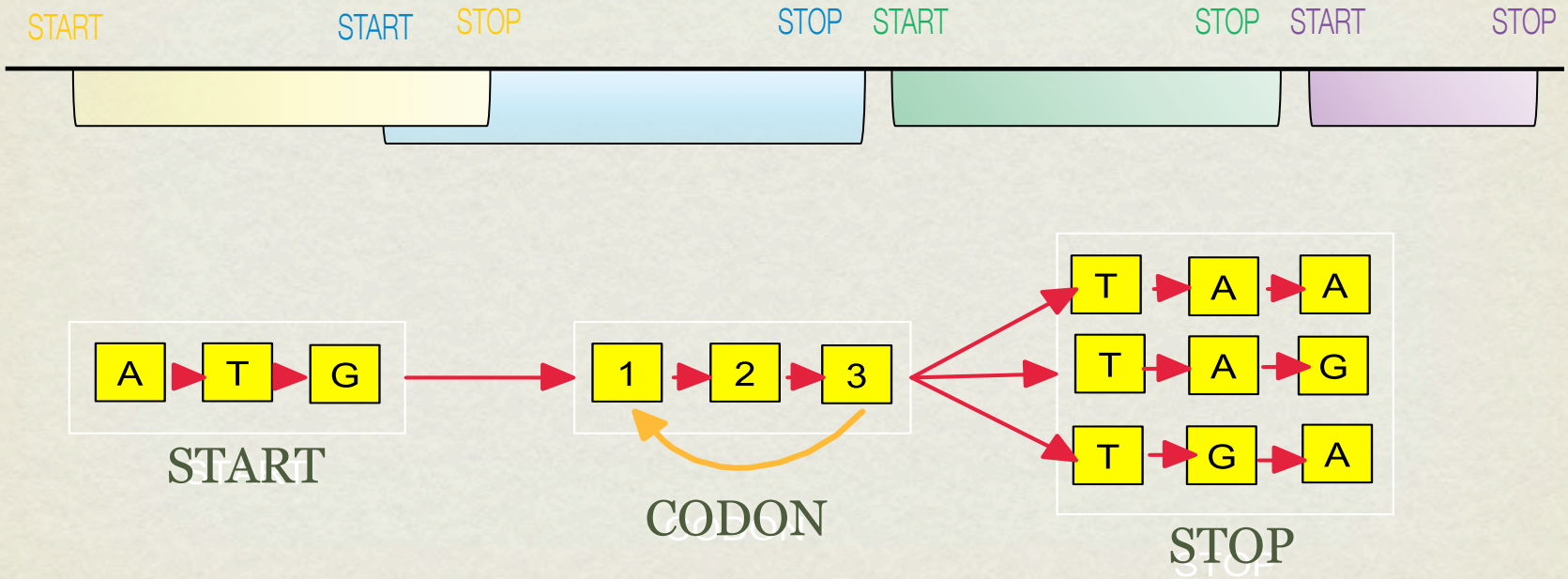
We take advantage of what we already learned about gene structures and features of coding sequences. Based on this knowledge we can build theoretical model, develop an algorithm to search for important features, train it on known data and use to search for coding sequences in anonymous genomic fragments.

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

GENERAL MODEL OF A GENE



SEQUENCE CODING POTENTIAL



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**

SEQUENCE FEATURES

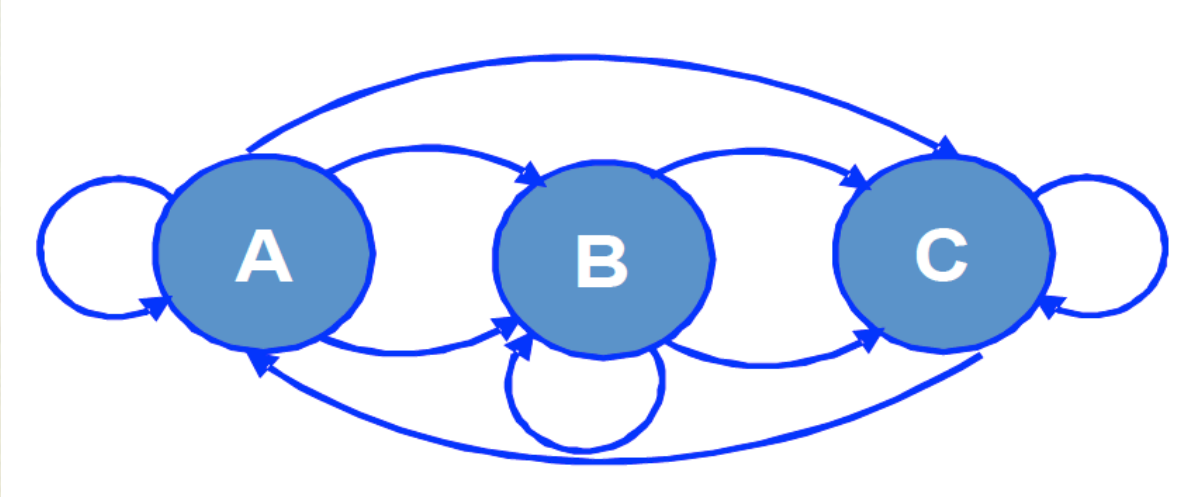
We can check if sequence in particular ORF has some other features which could tell us if this is a putative coding sequence or the ORF is false positive. We can look at the sequence content and compare it with known coding sequence and non-coding sequence and check to which of these two the ORF sequence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

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.

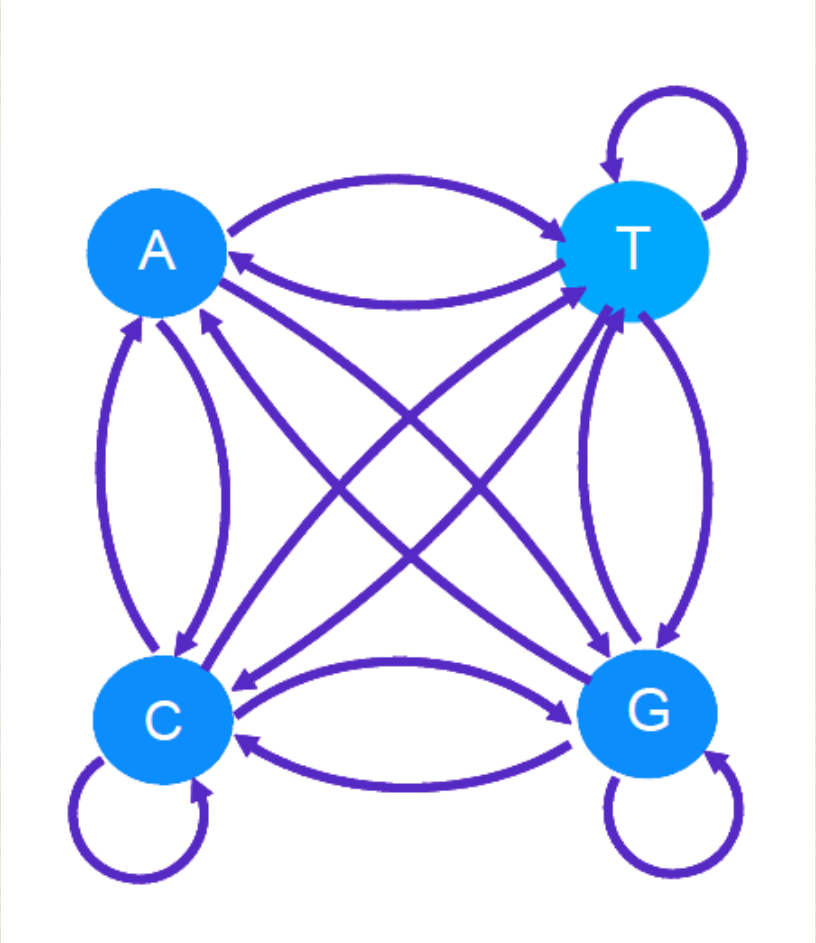
MARKOV CHAINS

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.



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.



ORDER OF MARKOV CHAINS

GCGCTAGCGCCGATCATCTACTCG

GCGCTAGCGCCGATCATCTACTCG
GCGCTAGCGCCGATCATCTACTCG

First order

GCGCTAGCGCCGATCATCTACTCG
GCGCTAGCGCCGATCATCTACTCG

Second order

GCGCTAGCGCCGATCATCTACTCG
GCGCTAGCGCCGATCATCTACTCG

Fifth order

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.

ORDER OF MARKOV CHAINS

GCGCTAGCGCCGATCATCTACTCG

GCGCTAGCGCCGATCATCTACTCG
GCGCTAGCGCCGATCATCTACTCG

First order

GCGCTAGCGCCGATCATCTACTCG
GCGCTAGCGCCGATCATCTACTCG

Second order

GCGCTAGCGCCGATCATCTACTCG
GCGCTAGCGCCGATCATCTACTCG

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

Probability matrix

4^{K+1}

first order Markov Model - matrix of 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)

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

$4^{2+1} = 4^3 = 64$

$4^{3+1} = 4^4 = 256$

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

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

$$3\,(4^{1+1}) = 3 \times 4^2 = 16$$

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.

$$LP(S) = \log \frac{P^{i}(S)}{P_{0}(S)}$$

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

MARKOV MODELS - PROBABILITIES

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

S=AGGACG

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

$$P_1(S) = f(A,1)F_1(G,A)F_2(G,G)F_3(A,G)F_1(C,A)F_2(G,C)$$

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

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

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

CALCULATING LP

$$LP(S) = \log \frac{P^{i}(S)}{P_{0}(S)}$$

S=AGGACG

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

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

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

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

$$LP(S) = -0.4644$$

GLIMMER

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

PROKARYOTIC VS. EUKARYOTIC GENES

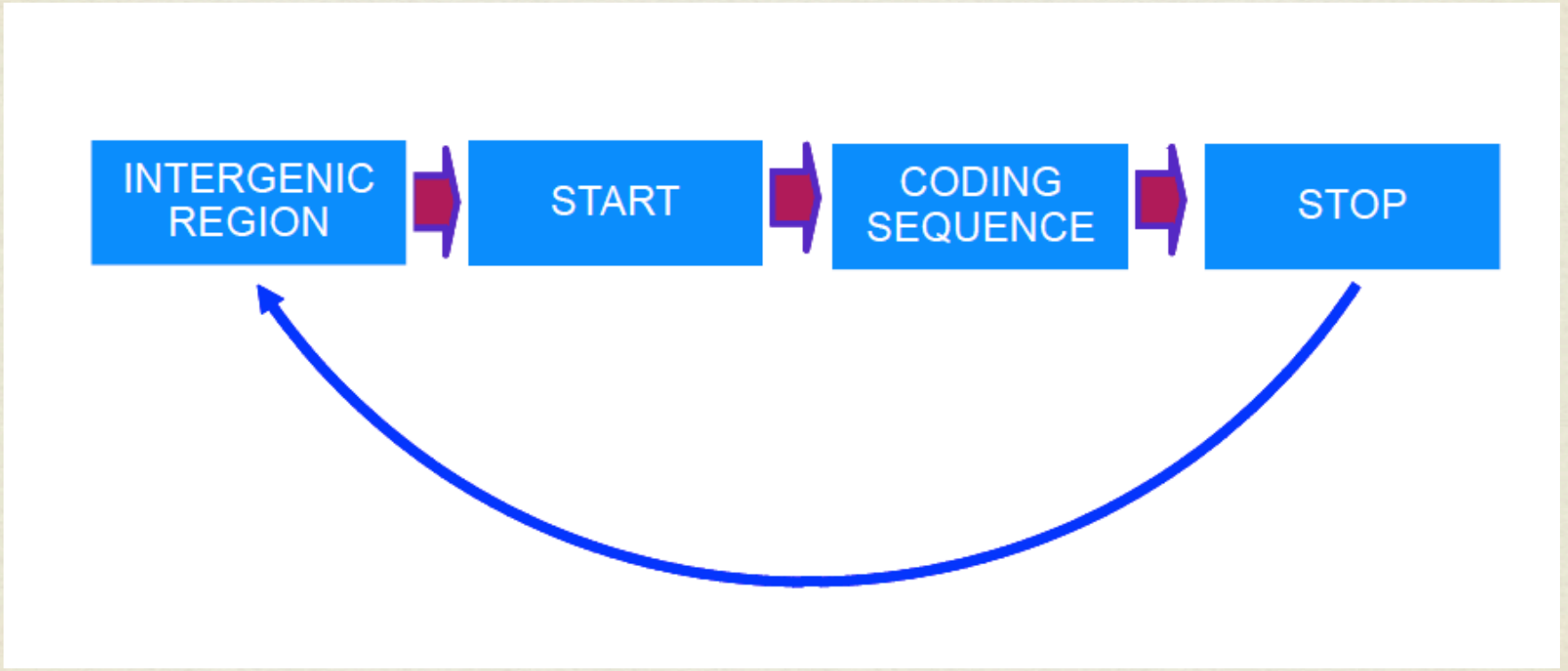
Prokaryotes

- small genomes
- high gene density
- no introns (or splicing)
- no RNA processing
- similar promoters
- terminators important
- overlapping genes

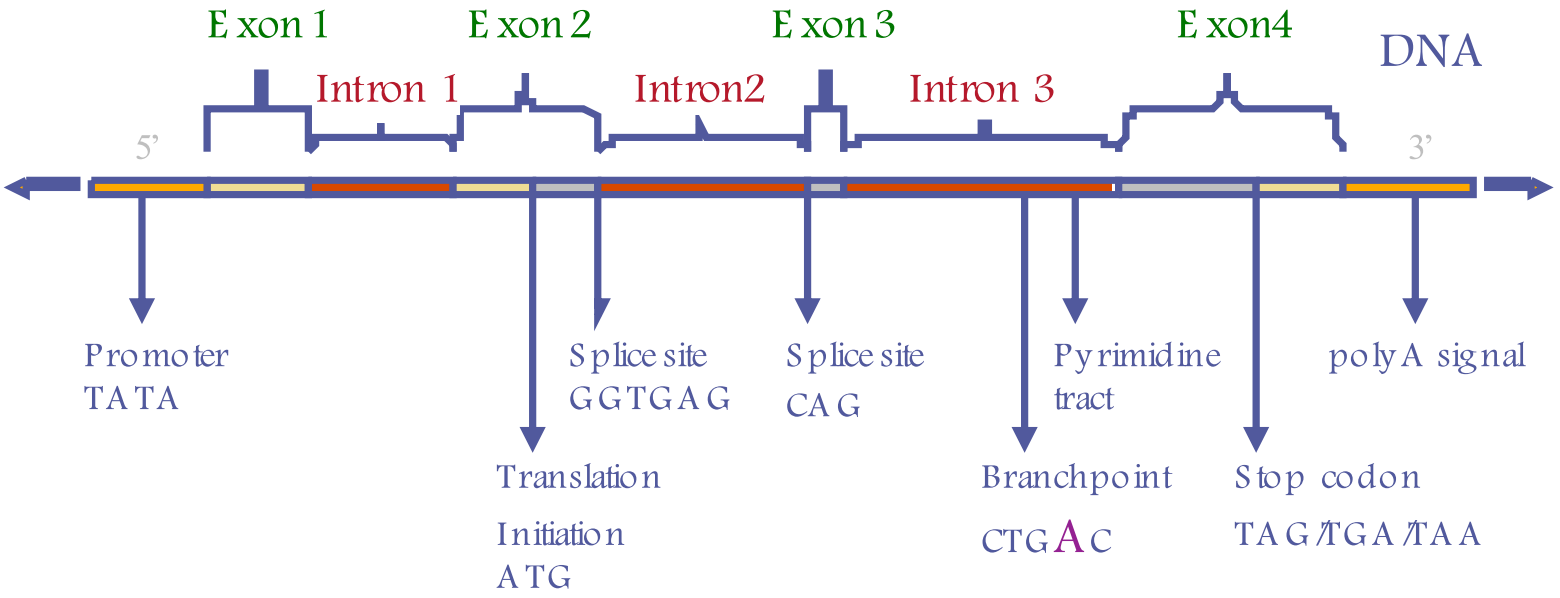
Eukaryotes

- large genomes
- low gene density
- introns (splicing)
- RNA processing
- heterogeneous promoters
- terminators not important
- overlapping genes
- polyadenylation

CODING REGIONS IN PROKARYOTES



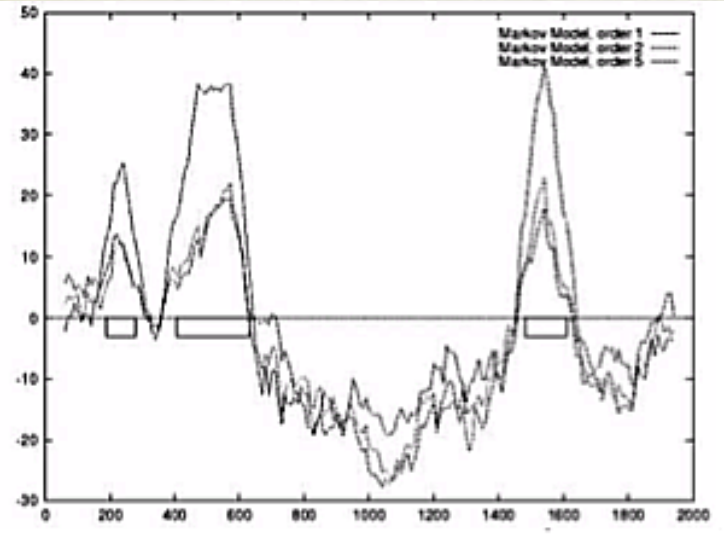
EUKARYOTIC GENE STRUCTURE



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.

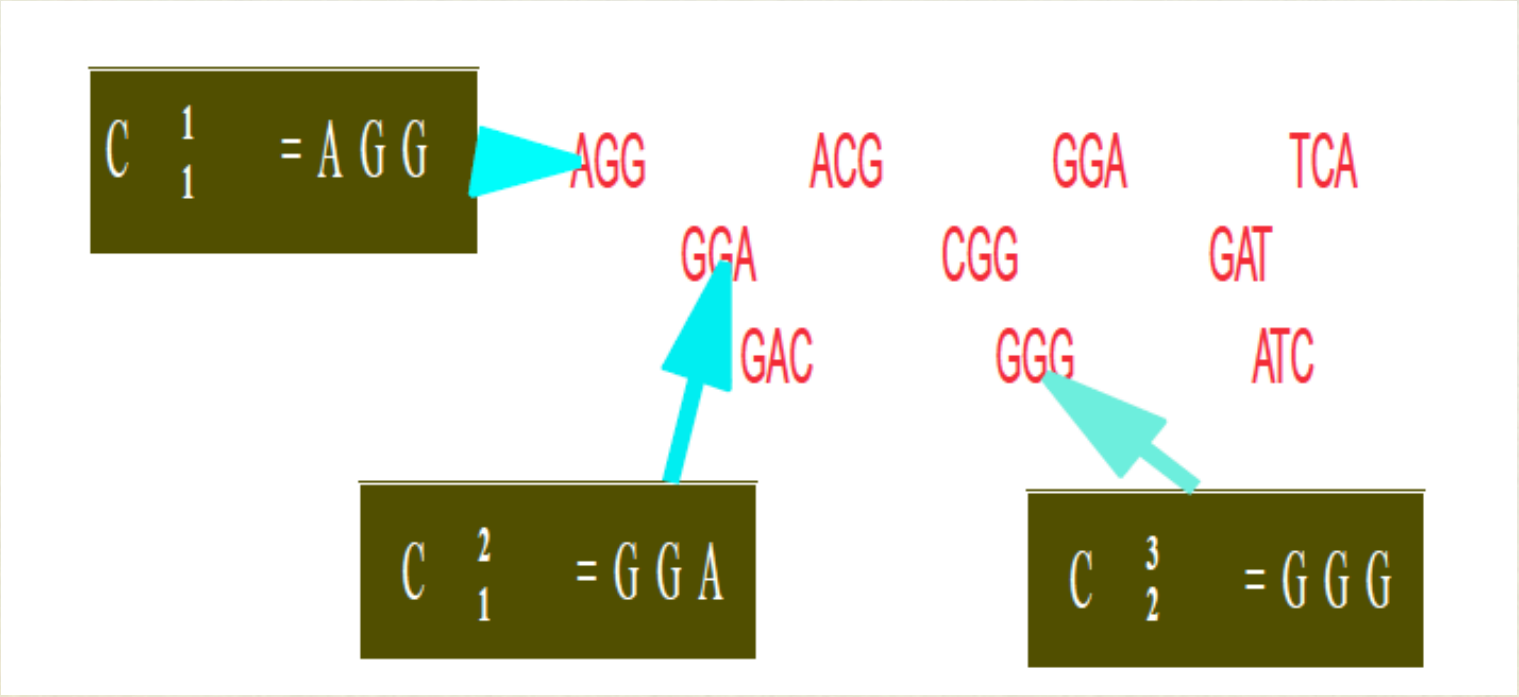
CODON USAGE

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

CODON USAGE

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

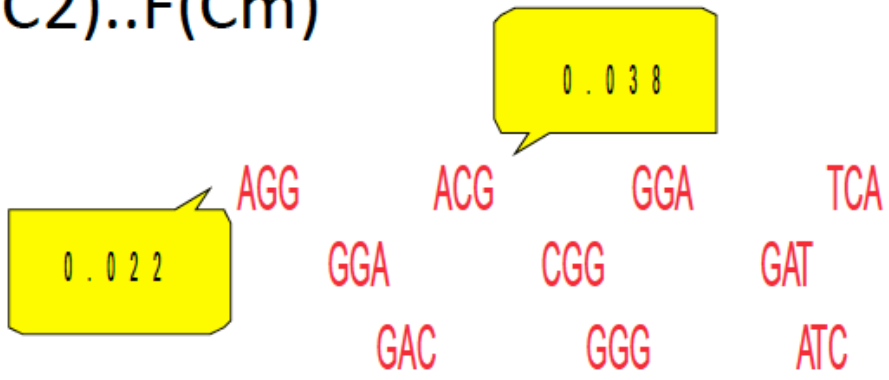
C = C1C2…Cm



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.

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



$$P(C) = F(AGG)F(ACG) = 0.022 \times 0.038 = 0.000836$$

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 possible.

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

$$P(C) = F(AGG)F(AGC) = 0.0156 \times 0.0156 =$$

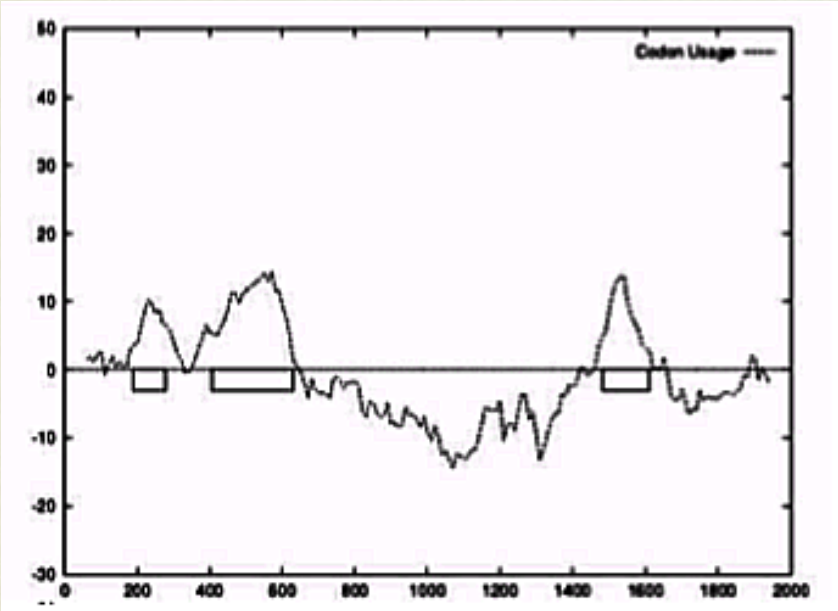
0.000244

LOG-LIKELIHOOD RATIO

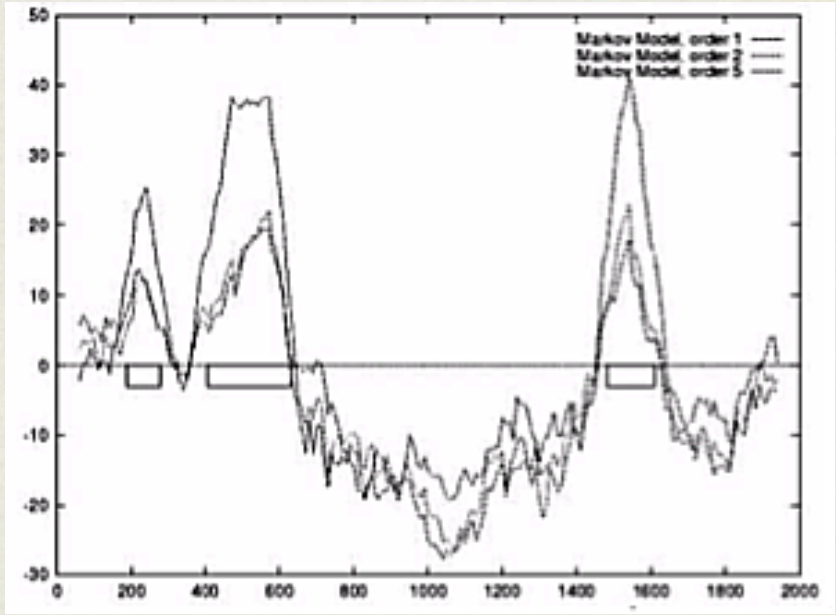
$$LP(S) = \log \frac{P_i(S)}{P_0(S)}$$

$$LP(S) = \log 1.4102 + \log 2.4358 = 0.1493 + 0.3866 = 0.53 > 0$$

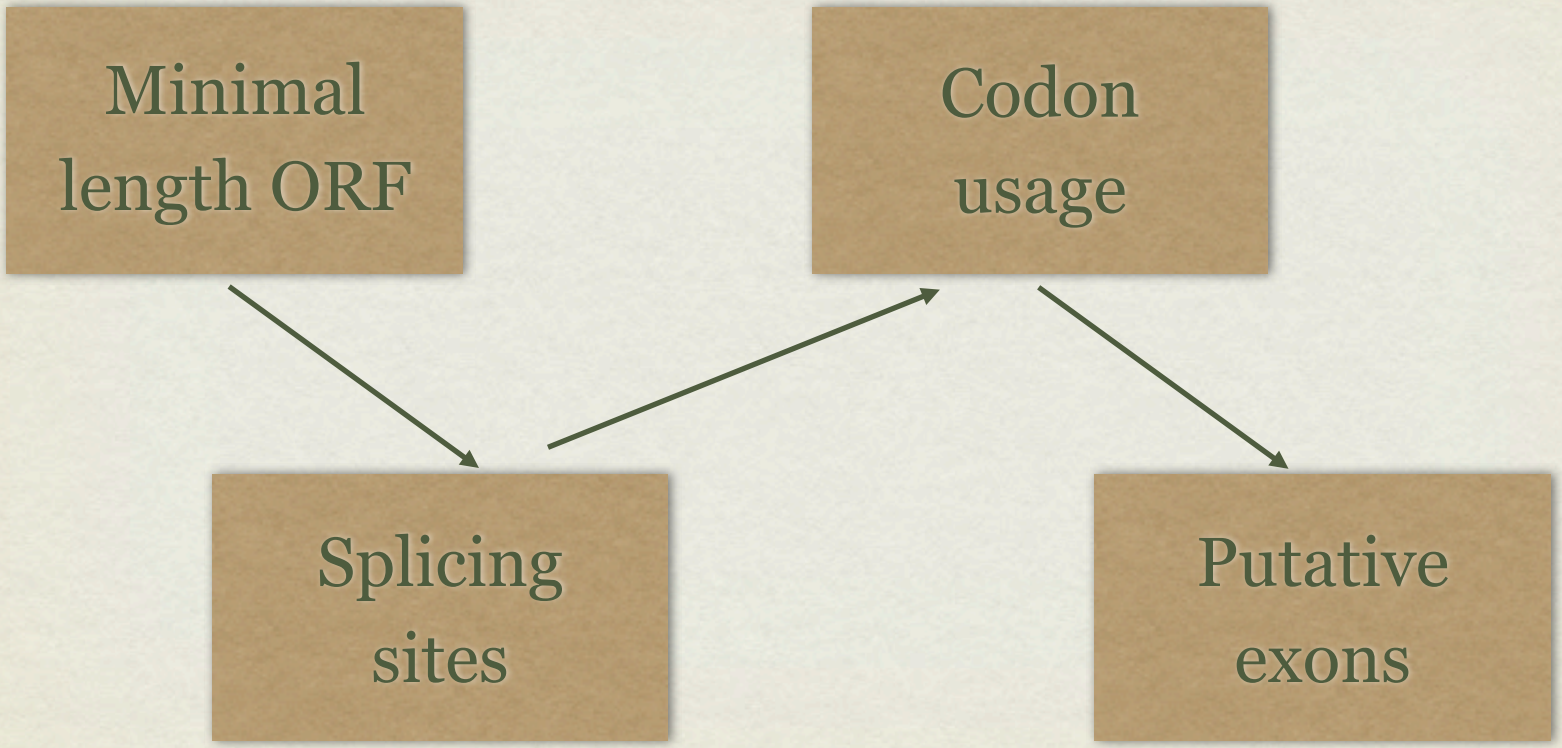
Codon usage



Markov models



RULE BASED METHODS



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.

CLASSES OF GENE PREDICTION METHODS

Sequence similarity based

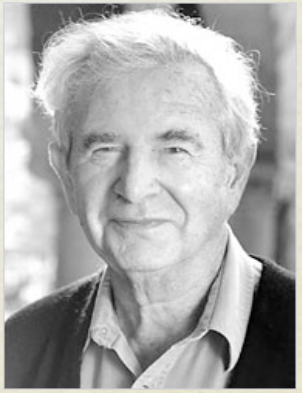
- BLAST can be used for aligning ESTs or proteins to the genomic sequence
- PROCRUSTES and GenWise use global alignment of homologous protein to genomic sequence
- The biggest limitation to this type of approaches:
  - only about half of genes being discovered have significant similarity to genes in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

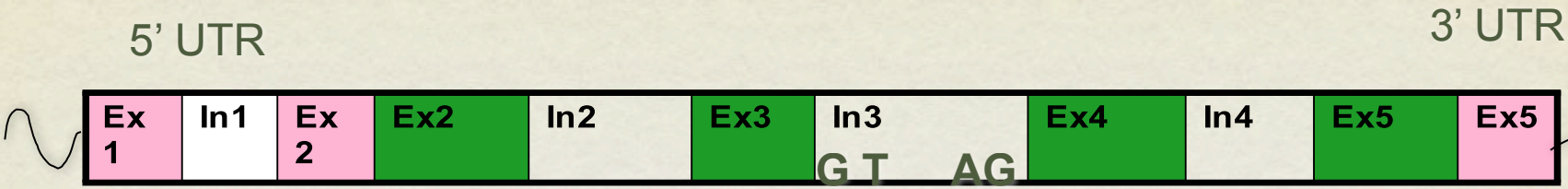
Model based

- 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

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





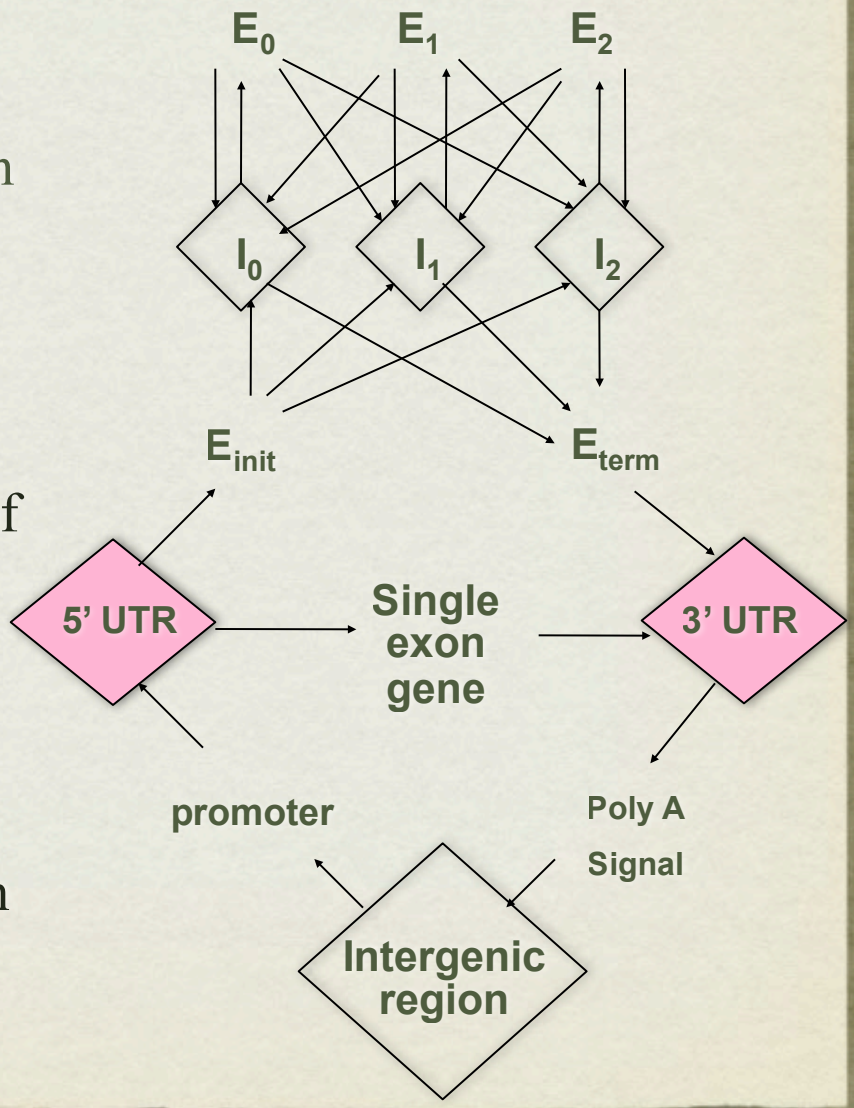
Each state corresponds to one of the seven categories with which all nucleotides are ultimately labeled- promoter, 5'UTR, exon, intron, 3'UTR, PolyA, intergenic

Three components:

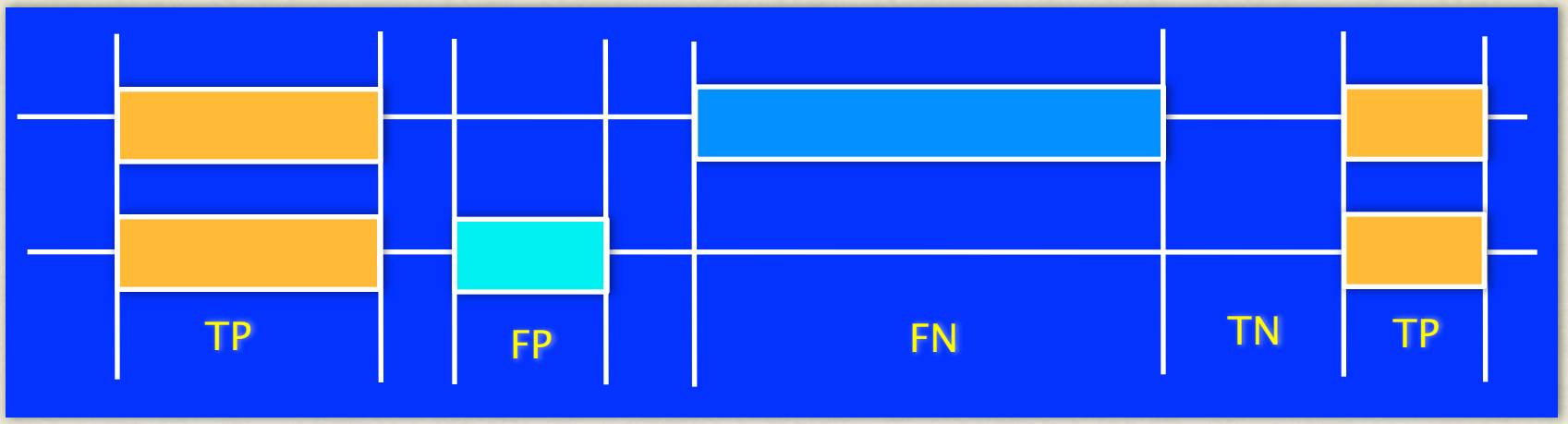
Transition model – specifies probability of moving from any one state to another

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



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 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)

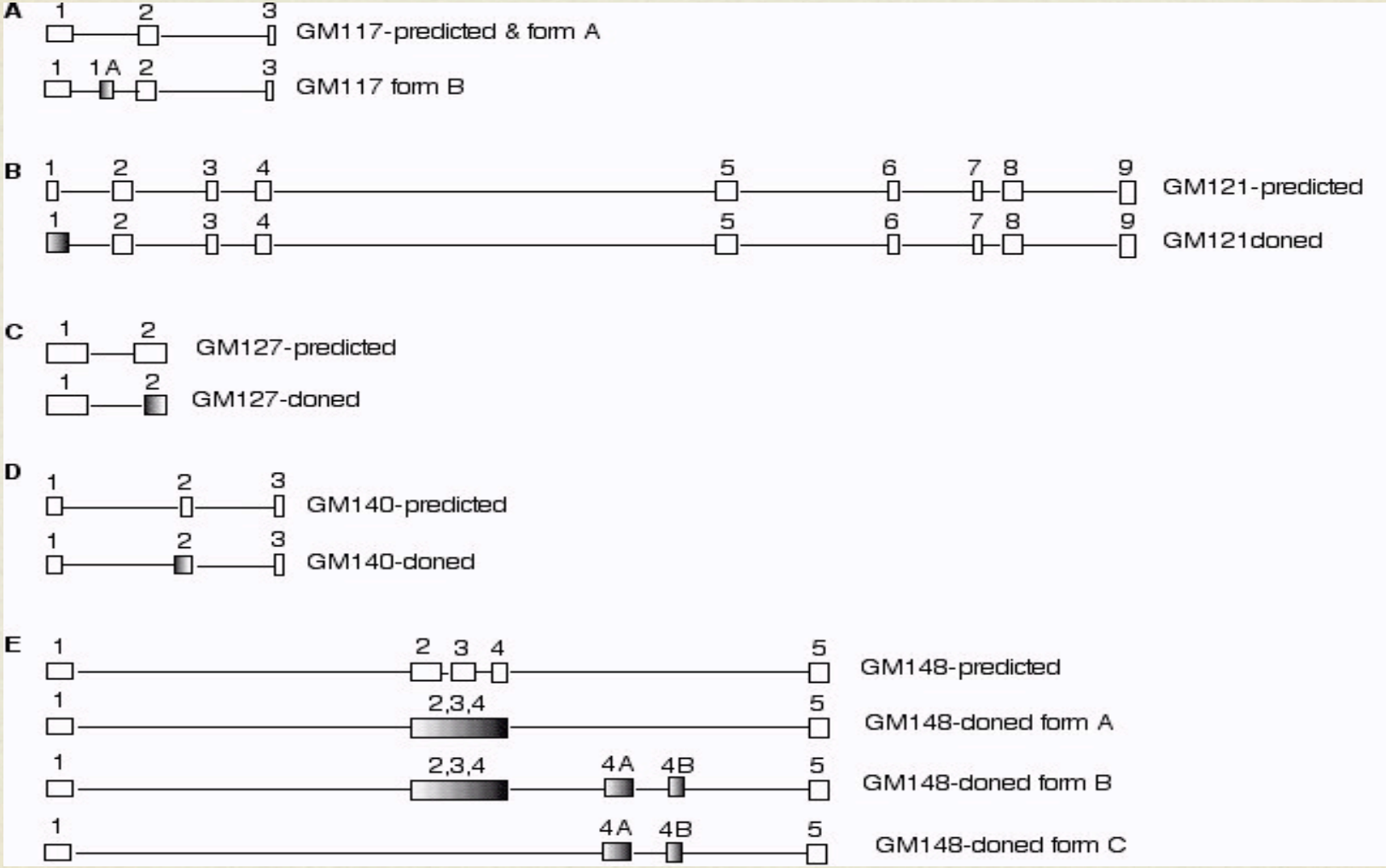
$$CC = \frac{TP \times TN + FP \times FN}{\sqrt{(PP)(PN)(AP)(AN)}}$$

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.

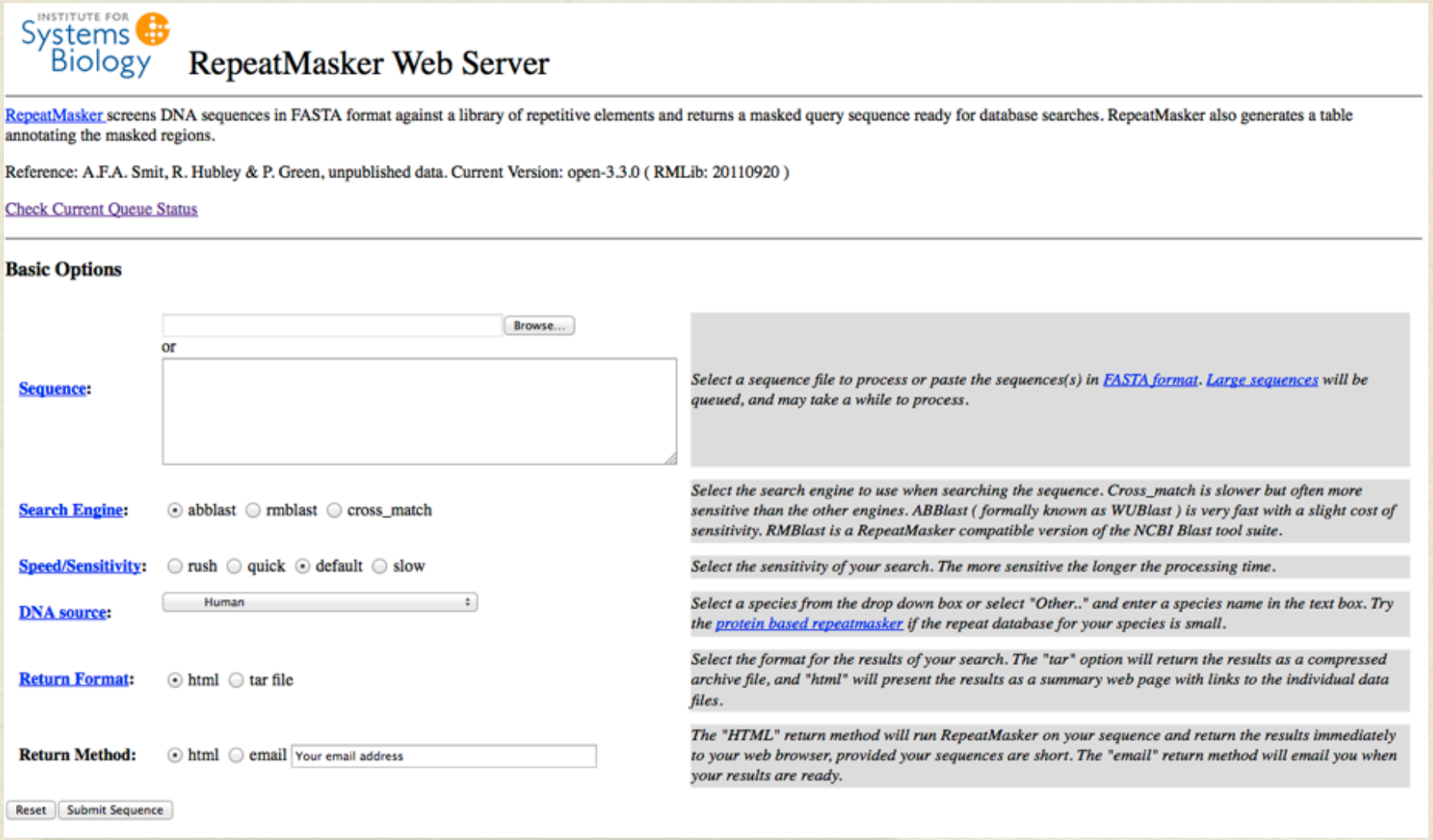
	predicted exons	specificity	sensitivity
MZEF	34	**0.51**	0.56
GRAIL	11	0.48	0.19
GENSCAN	52	0.46	**0.91**
FGENES	45	0.37	0.75

PROBLEMS RELATED TO GENE PREDICTION - GENE STRUCTURE AND ALTERNATIVE SPLICING



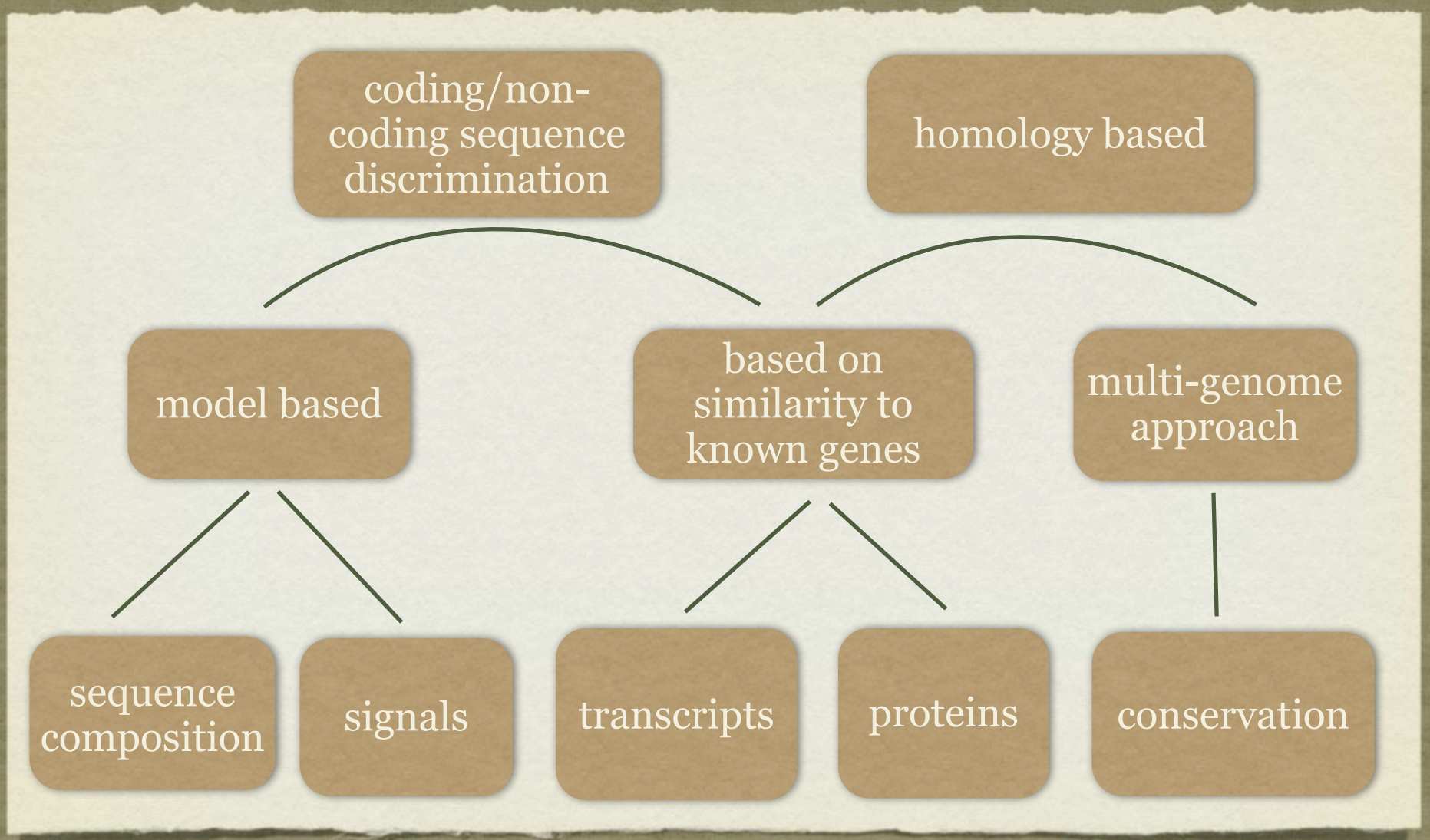
I. Makalowska et al. Gene 284: 203-213

RepeatMasker



http://www.repeatmasker.org/

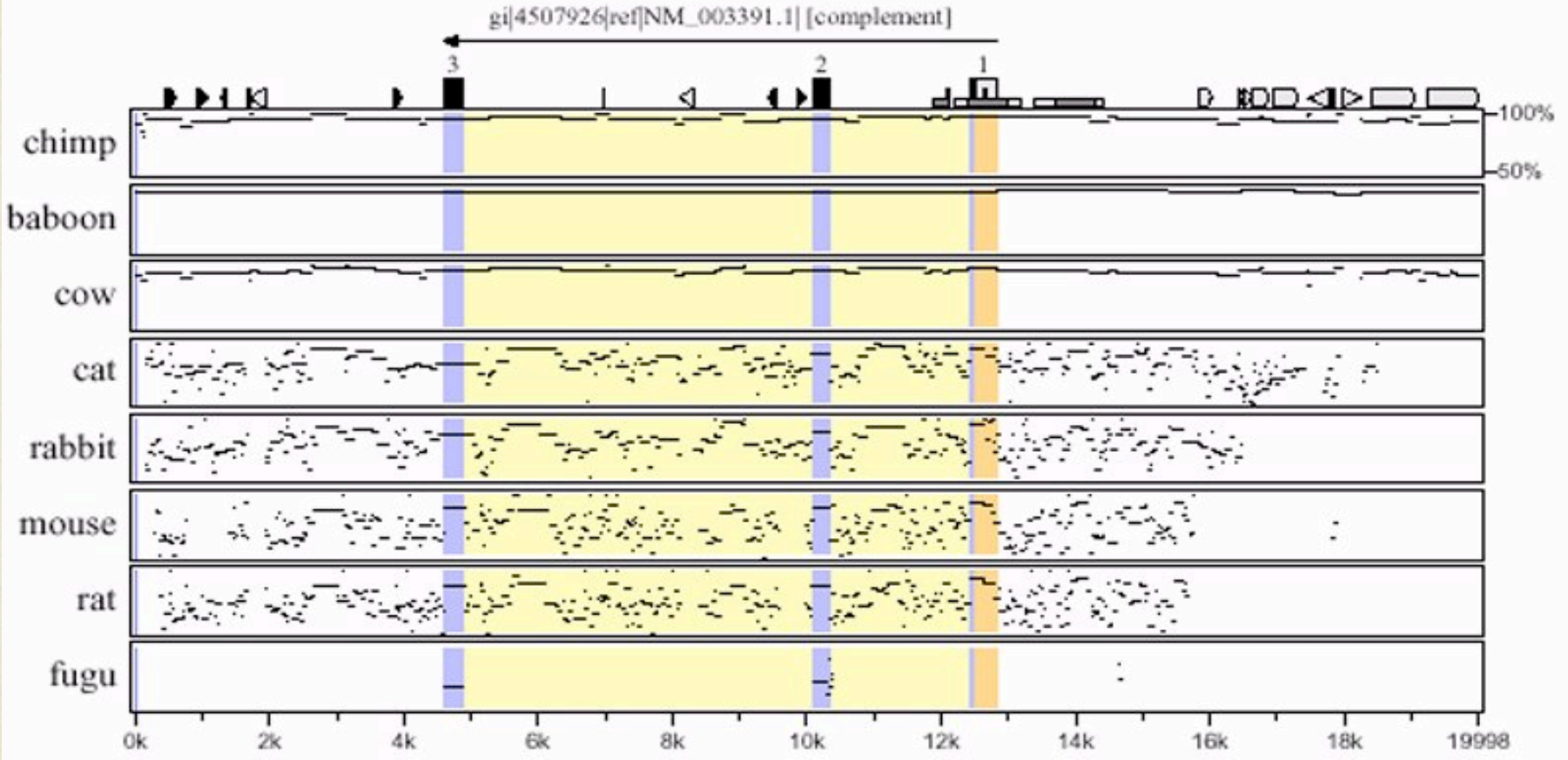
GENE FINDING METHODS



PipMaker http://pipmaker.bx.psu.edu/pipmaker/

- Computes alignments of similar regions in two or more DNA sequences
- Resulting alignments are summarized with a "percent identity plot"
- As an output PipMaker generates PDF or PostScript document
- MultiPipMaker can be requested to compute true multiple alignment and return a nucleotide level view of the results

MULTIPIPMAKER



GENE FINDING STRATEGIES

- Search for conserved regions
- Presence of ORF
- Codon usage
- Splice sites
- Polyadenylation signal
- Similiratity 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

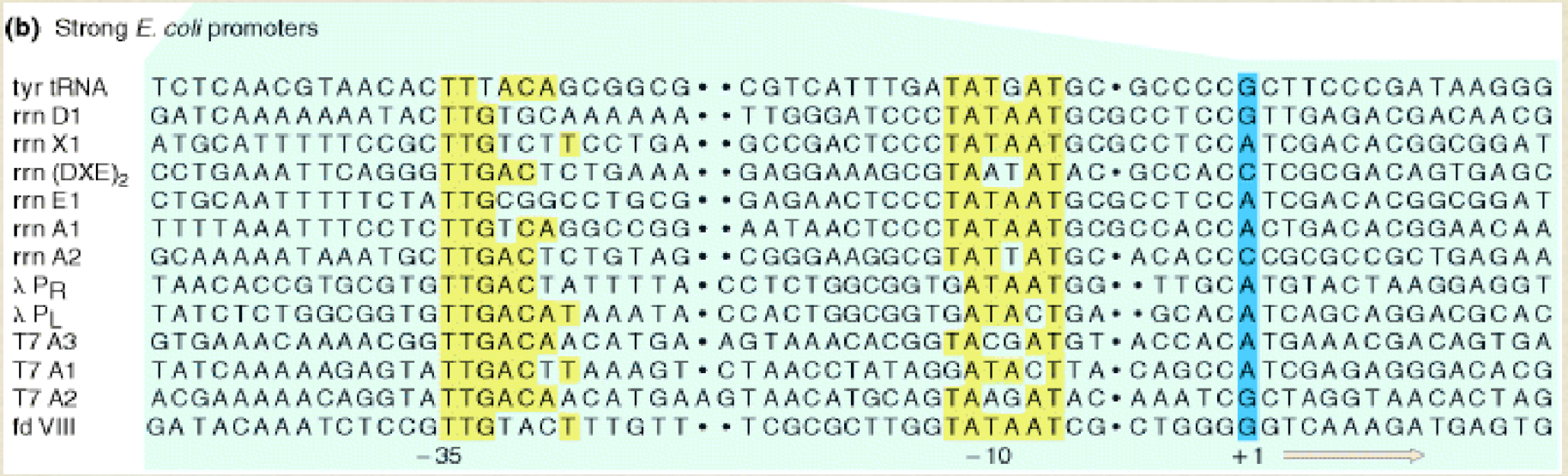
NNNTTGACANNNNNNNNNNNNNNNNNTATAATNNNNNNATGcccccc

-35 region

-10 region

RNA start site

E.COLI PROMOTERS

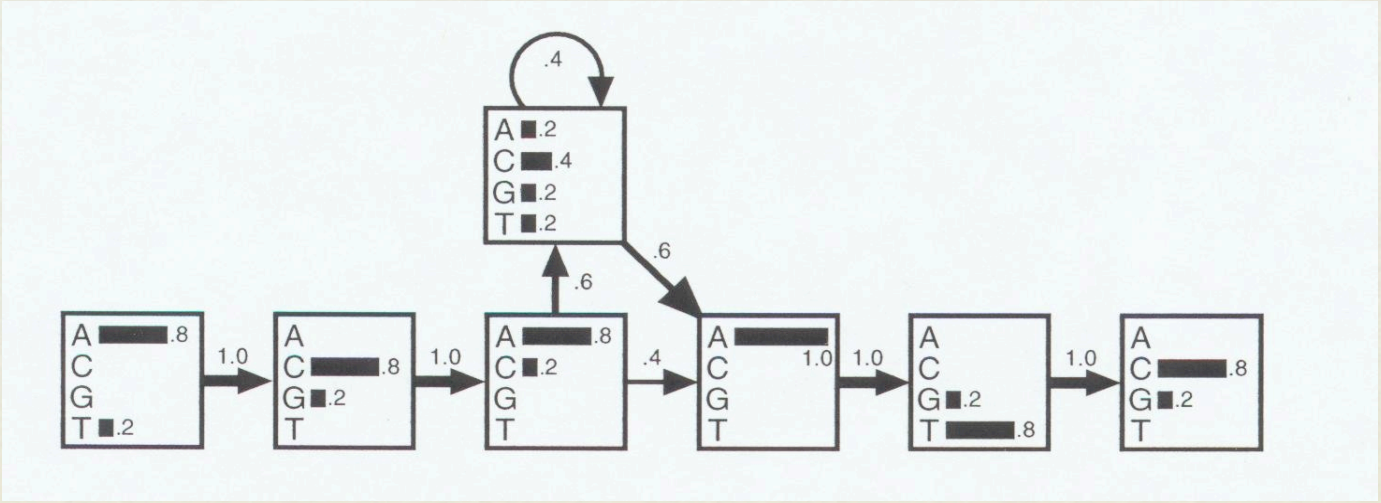


Promoters sequences can vary tremendously.

RNA polymerase in eukaryotes recognizes hundreds of different promoters

MARKOV MODELING - AGAIN

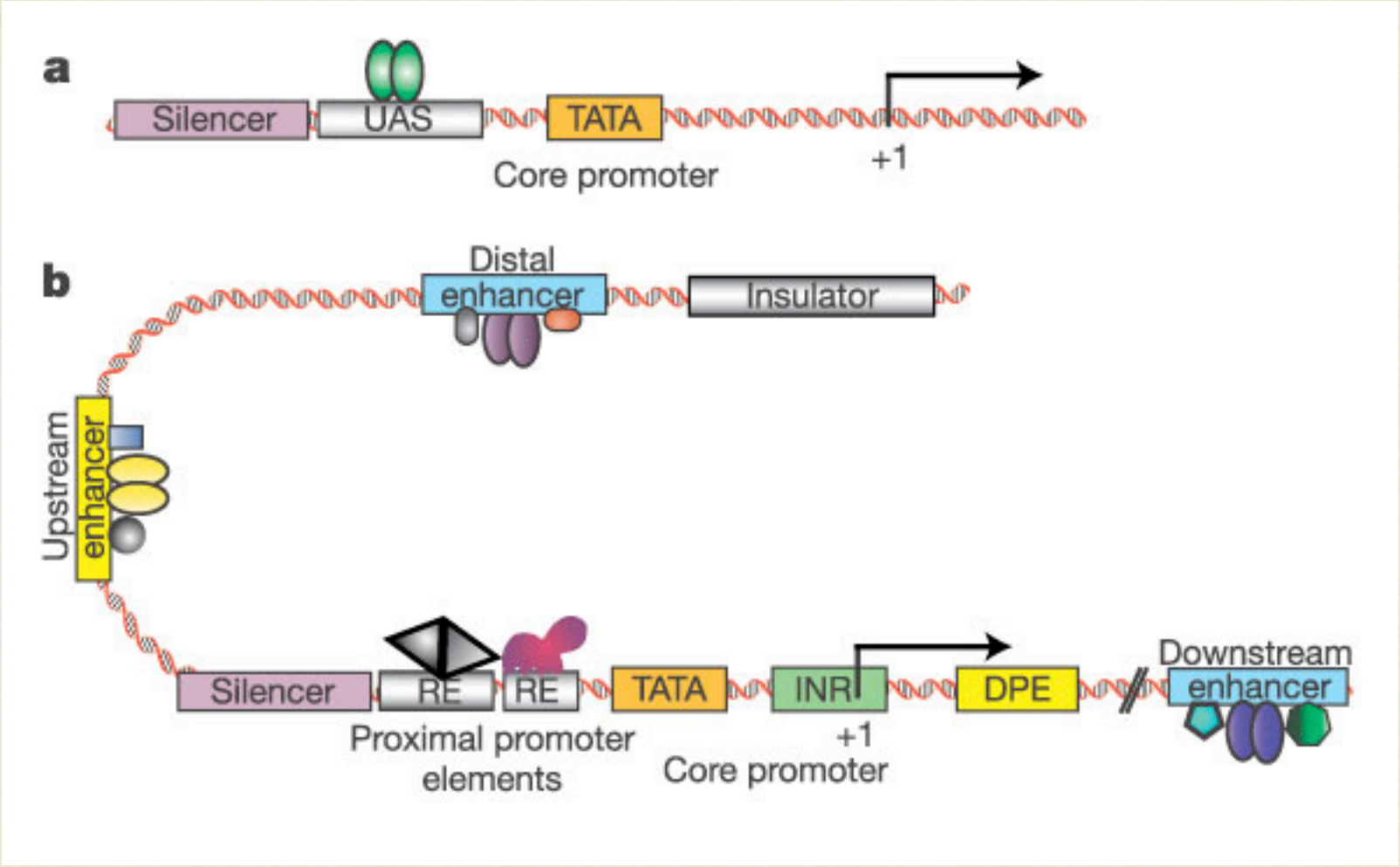
A C A – – – A T G
T C A A C T A T C
A C A C – – A G C
A G A – – – A T C
A C C G – – A T C



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

EUKARYOTIC PROMOTERS



Nature **424**, 147-151

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

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

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
```

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...
