Chapter 8 DNA Analysis

8.1 Genome Anatomy

8.2 Gene Finding

1. Hidden Markov models
2. Neural networks
3. Homology Search
4. Promoter prediction
5. EST Cluster

Performance of Gene Prediction Methods

8.3 Alternative Splicing and Nucleosomes

Alternative splicing

Nucleosomes

8.4 Comparative Genomes

8.5 Genomic Individuality

Sequence repeats

SNPs

8. DNA Analysis
Introduction

- **Genome Sequencing: whole genome**
  - DNA, RNA, ncRNA, iRNA, miRNA
  - Proteins and proteins-RNA complexes
  - Repetitive and redundant sequences: transposons and consequences
  - Exons and Introns: Alternative splicing
  - SNPs: advantages and draw backs
  - Genome variability: N° genes, placement, duplications and rearrangement comparisons

↓

Whole genome comparison and genome mapping

8. DNA Analysis
Introduction

- Bioinformatics Challenge

DNA information detection: Inter/intra specie comparison

↓

Conditions/diseases relation

↓

Predictions

↓ ↓

Experimental designs

individual response to treatments

drugs/medication quantities

long term effects

8. DNA Analysis
8.1 Genome Anatomy

Prokaryotes: circular DNA

Eukaryotes: chromosomes forming nucleosomes enclosed into nucleus

Sequenced genomes

Prokaryotes

Hemophilus influenzae and E.coli (58% match in two genomes)

1977 Fred Sanger Bacteriophage (11 genes)

1981 Anderson et al. Human mitochondrion (16 568 bp, 13 proteins, 2 ribosomals RNAs and 22 tRNAs)

1986 Plant chloroplast organelle (120-200 kbp)

Eukaryotes

1992 Oliver et al. S. cerevisiae (315 kbp, 182 genes)

1995 H.influenzae (1,83 Mbp, 1743 genes)

8. DNA Analysis
8.1 Genome Anatomy

GOBASE: The Organelle Genome Database release 23

Based on GenBank releases 42.000 new mitochondrial sequences and 39.000 new chloroplast sequence . Tuesday 19th May 2009

Resources for Genomics, Molecular Biology and Evolutionary Research

OGMP: The Organelle Genome Megasequencing Program

Interested in the evolution of mitochondria and plastids and their genomes

FMGP: Fungal Mitochondrial Genome Project

Protist genome sequencing projects

8. DNA Analysis
8.1 Genome Anatomy

Organism	Group	Genome (Mbp)	Genes	kb containing one gene
Methanococcus jannaschii 1996	archaea	1.66	1,682	0.99
Escherichia coli 1997	bacteria	4.6	4,288	1.07
Hemophilus influenzae 1995	bacteria	1.83	1,743	1.05
Mycoplasma pneumoniae 1996	bacteria	0.82	676	1.21
Bacillus subtilis 1997	bacteria	4.2	4,098	1.02
Aquifex aeolicus 1998	bacteria	1.55	1,512 1.03	
Synecgicystus sp. 1996	bacteria	3.57	3,168 1.13	
Arabidopsis thaliana	plant	125	25,000	5.0
Caenorhabditis elegans	worm	100	18,424	5.43
Drosophila melanogaster	fruit fly	180	13,601	13.23
Saccharomyces cerevisiae	budding yeast	13.5	6,241	2.16
Homo sapiens	human	2900	> 30,000	96.67 → 1 gene/80kbp

Clustered with high gene density, GC content, SINE and low LINE

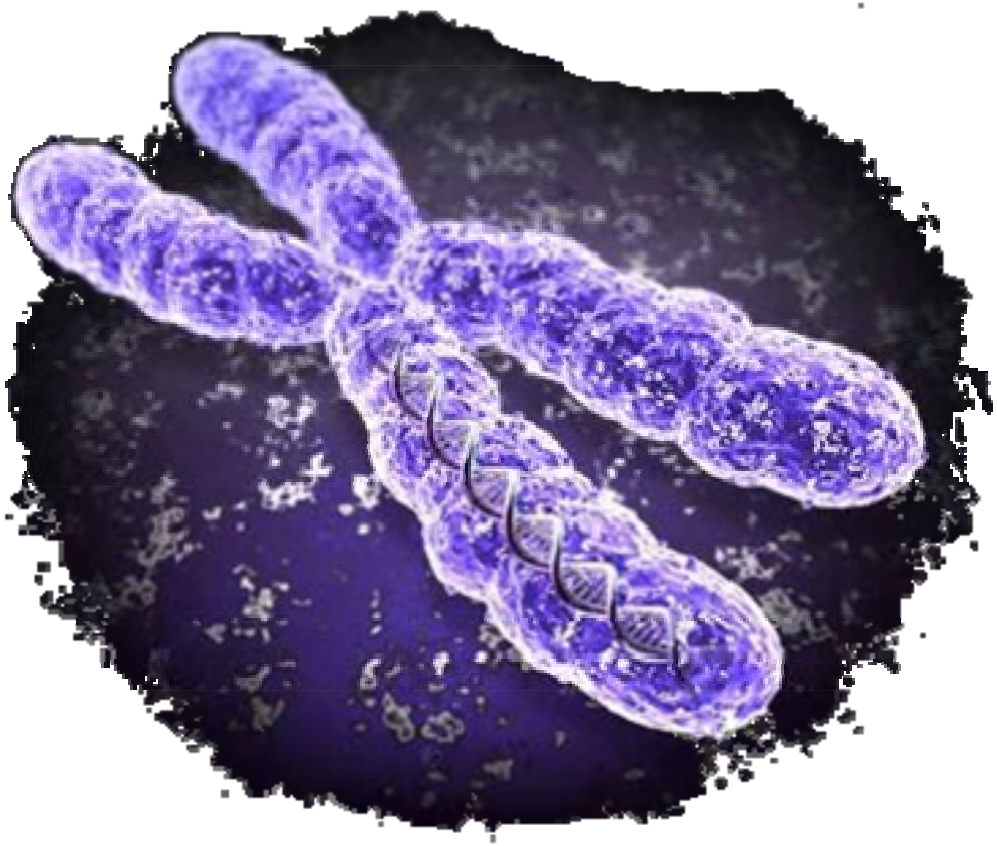
Eukaryotic cells In chromosomes

Heterochromatin as tightly packed form of DNA and not transcribed: centromeres and telomeres

Euchromatin as lightly packed form of DNA and actively transcribed: rich in gene concentration

Pseudogenes as gene copies which have lost their protein-coding ability or are otherwise no longer expressed in the cell: duplication and no transcription, housekeeping genes (ribosomal proteins)

www.pseudogene.org



http://www.tiricosuave.com/images/chromosome.jpg

8. DNA Analysis
8.2 Gene Finding

- Each organism has gene codon preferences and splite junctions: each genome specifies its own gene finding model (HMM and NN)

  1. Whole genome sequencing
  2. ORF identification: start (AUG ^{met}) and stop codon + reading frame controlled up and downstream with three starting positions
  3. ORF checked by homology gene search (known gene), codon specific usage and statistics (pairwise codon frequency), GC content (bias in the 3rd position)
  4. TESTCODE and CODONFREQUENCY
  5. Difficulties in finding ORFs due to introns

     Eukaryotes

     Promoters identified

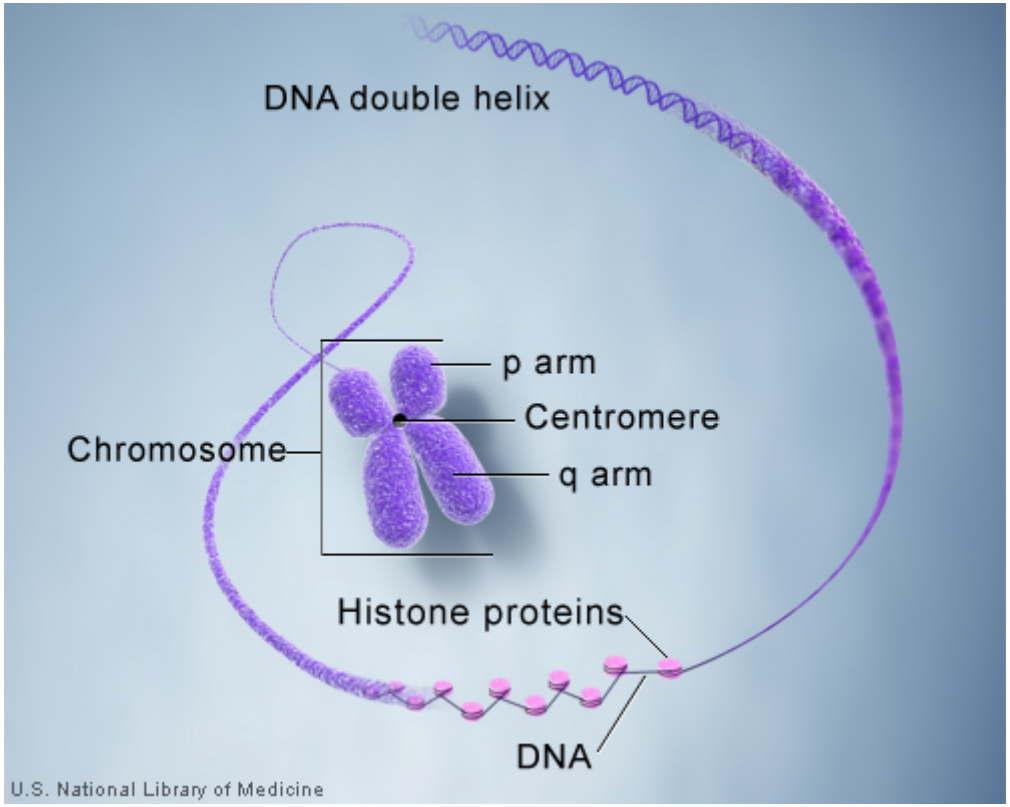
     Introns determined and removed

     mRNA sequences translated (1st start codon- 1st stop codon)

Computer models for introns recognition must be constructed

8.2 Gene Finding

1. Hidden Markov models
   GLIMMER; GENEZILLA
1. Neural networks
   GRAIL;GeneParser;NetGene
1. Homology Search
2. Promoter prediction
3. EST Cluster



Performance of Gene Prediction Methods

Each genome requires a model
trained to its specific characteristics

8.2.1 HMM

Hidden Markov Models

- Attempt to detect :
  - Coding regions boundaries: Start and stop codons
  - Transcription initiation and termination sites
  - PolyA sites
  - Splite sites
  - Protein binding sites (Transcription factors, TATA-box, Topoisomerase I and II)
  - Ribosomal binding sites
  - Branch points

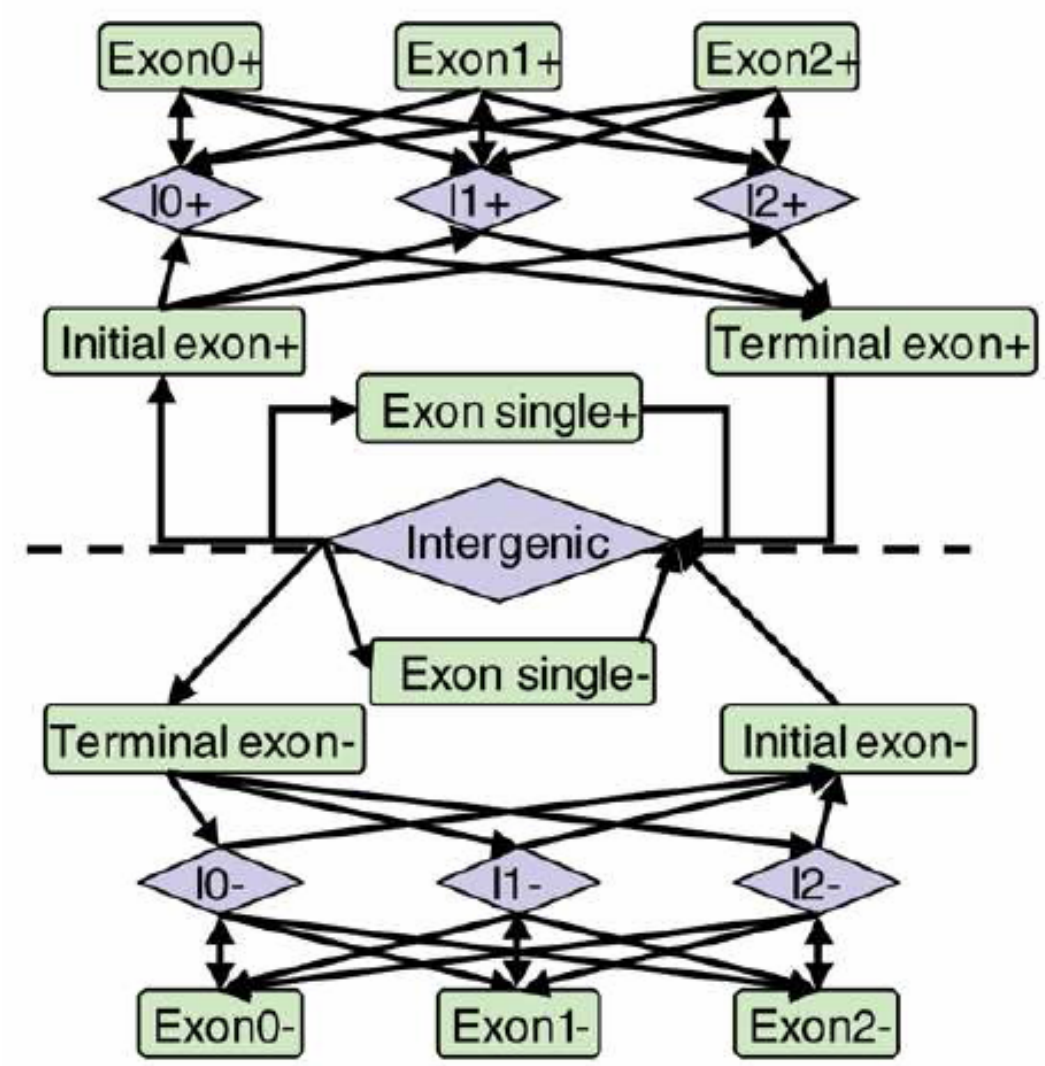
- Gene Mark, Mark.hmm, GLIMMER, GRAIL, GenScan /GenomeScan, Genie

  Constructed hierarchically through region modules

  Exon module: initial, internal and terminal exon

  Intron modules

  Intergenic modules



HMM

State-transition diagram

Each state implemented as a separate submodel such as weight array matrix or an IMM-Interpolated Markov Model

GLIMMER: Interpolated MM Long known patterns search

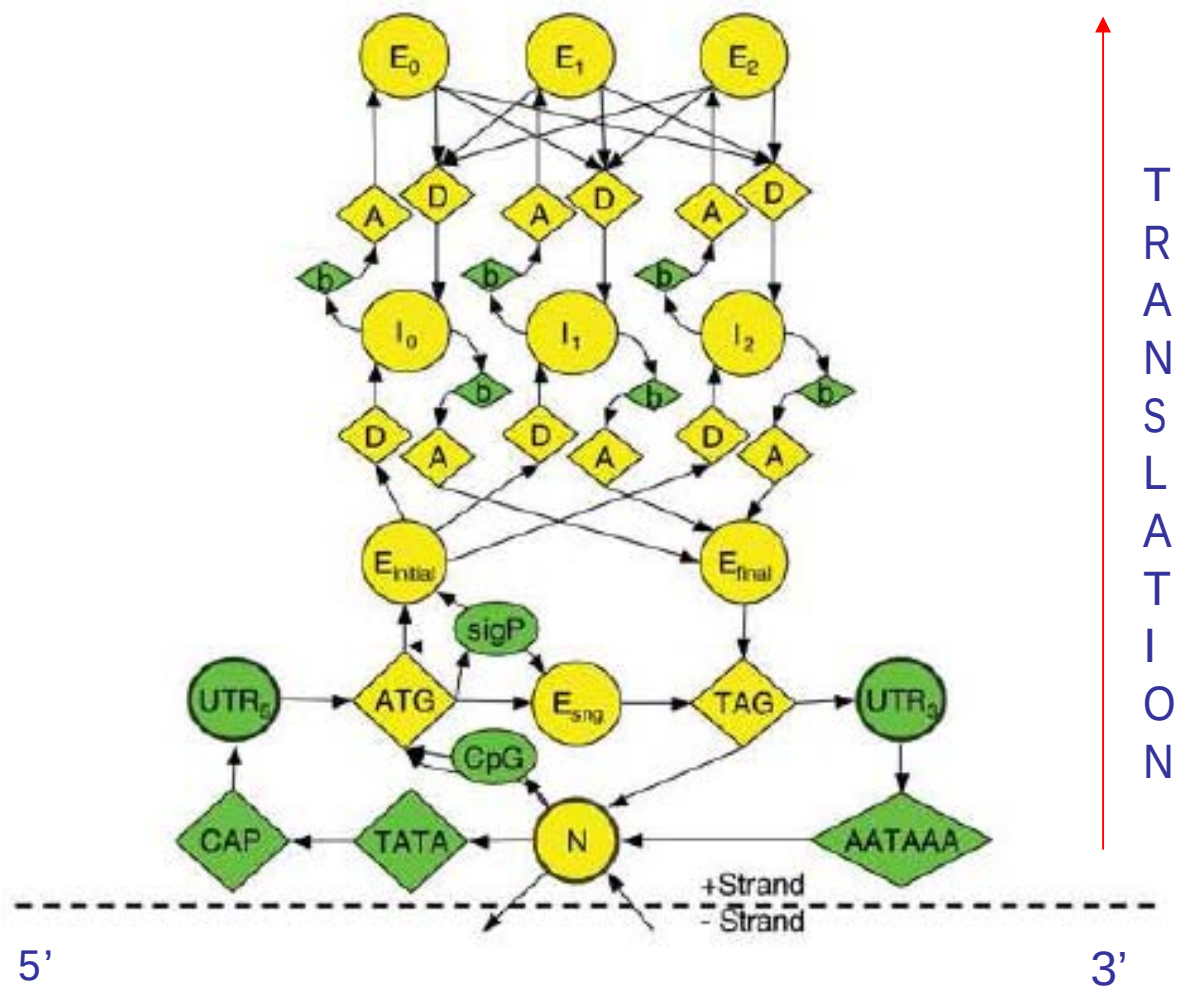
Pattern recognition

+

Probabilistic modeling

Long frequent pattern modeled by higher order and higher probability than short ones

Probability combination in the final model



TRANSLATION

HMM

State-transition diagram

GENEZILLA

TRANSLATION

8.2.2 Neural Networks

Artificial Neural Network - Model : supervised learning neural networks approximate a function from training data

Training data consists of *n* input vectors and the corresponding outputs

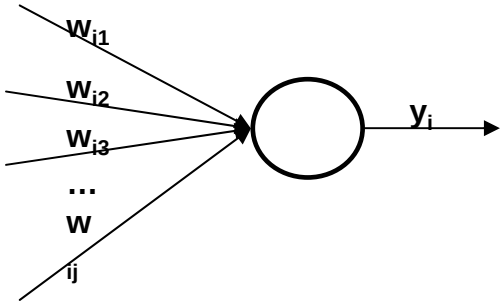
class label, continous value

$\{(x_p, d_p) | 1 \le p \le n\}$

Model

–units (input, output, hidden), weighted connections

–parameters of the neural network → weights



$$net_i = \sum_j w_{ij} \cdot y_j$$

$$y_i = f(net_i)$$

$$f(z) = \frac{1}{1+e^{-z}}$$

$$net_i(t) = \sum_j w_{ij} \cdot y_j(t-1)$$

$$y_i(t) = f(net_i(t))$$

Artificial Neural Networks

GRAIL→ NN based system for gene finding in **Coding/non coding regions**

Identifies polyA sites and promoters →constructs protein sequences

With inputs as

- score of 6-mers in candidate region
- score of 6-mers in flanking regions
- Markov Model score
- Flanking region GC composition
- Candidate region GC composition
- Score for slicing acceptor site
- Score for splicing donator site
- Length of region

Scores are log-likelihood scores of simple probabilistic methods

Goal: to construct sequences by identifying coding/non coding regions

Artificial Neural Networks

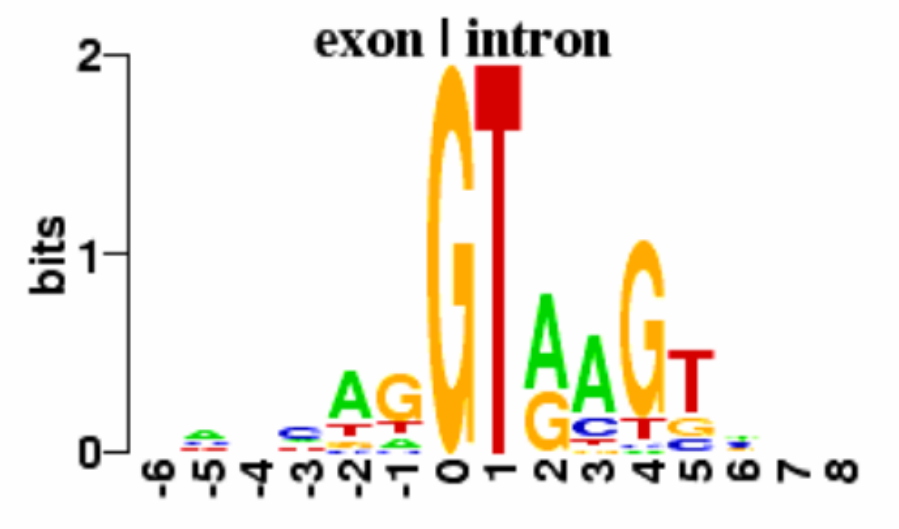
GeneParser → **splice site** recognition system by alignment of

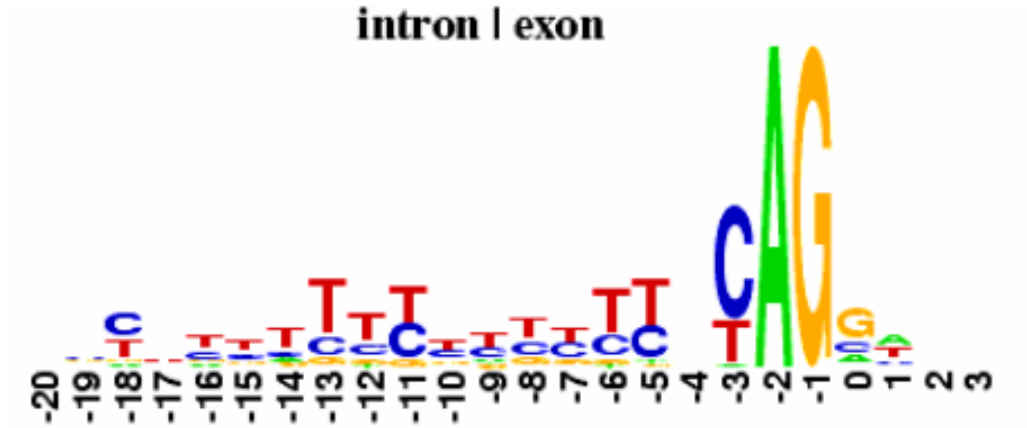
- Exons and introns tarts and ends
- Splice site indicators weighted by NN because the alignment scores are combined into one

Goal: log-likelihood score

NetGene → combination of **splice sites** prediction and **coding-non coding** regions into NN

- Three networks are combined with an input window of 15, 41 and 301 bp
- First and second network are donator and acceptor
- Third network as global network





ANN

Exon-intron and Intron-exon boundaries

→ Specific patterns

Size of the letter gives its frequency at its position

WebLogo

8.2.3 Homology search

Translation of all ORFs to amino acid sequences followed by alignment comparison method

- BLAST, WU-BLAST to known sequences
  - Match with low e-value (p-value)→ gene found
- BLASTX, FASTAX including the translation
- TBLASTN, TFASTX including the translation of both query and known sequence (database entry)

Local alignment methods may work even when intron-exon boundaries are not recognized

- When correctly translated exon→ corresponding exon found in an amino acid sequence already known

8.2.4 Promoter prediction

Promoter: 5' end of genes and containing the starting regions

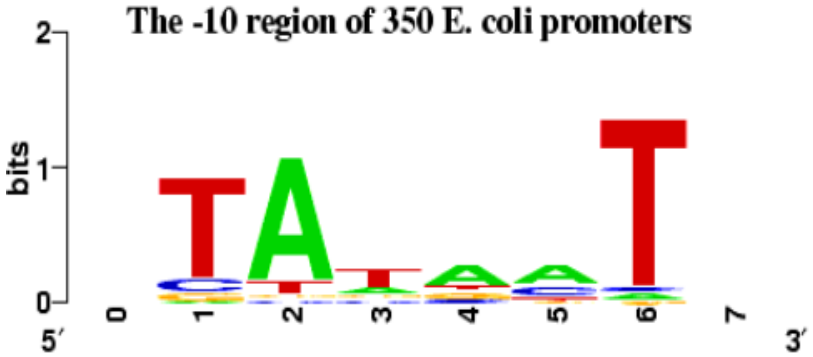
8.2.4.1 Prokaryotes: E.coli

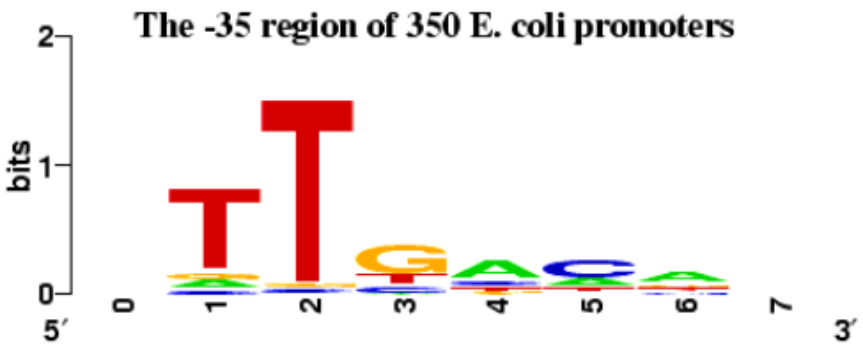
Alignment: promoter sequences aligned by using the transcription start site as anchor point

Pattern recognition: **TATAAT** Pribnow box and **TTGACA** pattern

AT rich region at +1 and -35 position

New promoters by Scoring matrix building (PSI-BLAST)





RNA polymerase Promoter

8.2.4 Promoter prediction

Neural Networks: by using a local code

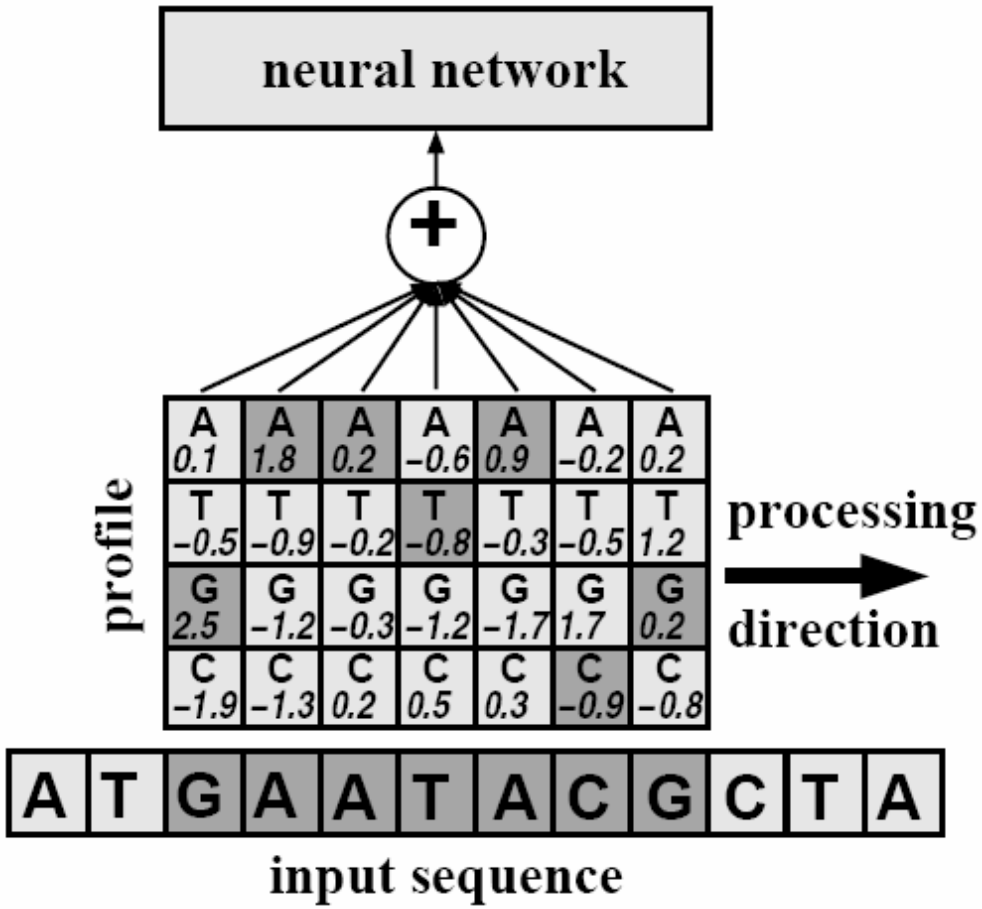
1. Each nucleotide coded by one vector with 4 components

A= (1,0,0,0); T= (0,1,0,0);
G= (0,0,1,0); C = (0,0,0,1)

2. Window over the sequence is used where the inputs weight of the NN are equivalent to a scoring matrix

Neural network ingoing weights

HMM: alignment coded into the MM or trained into short promoter sequences by an EM model



8.2.4 Promoter prediction

Promoter: 5' end of genes and containing the starting regions

8.2.4.2 Eukaryotes : Profiles search by RNApolyII binding sites upstream

Short conserved patterns contained within long regions

TF binding sites with position given respect to the transcription start site:TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH

TATA box consensus sequence **TATA**[**A**,**T**] {**C**}[**G**,**A**]

GC box

Remodel the nucleosome structure by acetylation and deacetylation of histones (DNA accessibility)

Transcription control by RNApolyII by Phosphorilation

Cell cycle genes

Pentamers for cell cycle genes: **ACGCGT** for late G_1 phase and **CCCTT** for early G_1 phase

Co-regulated genes by microarray expression profiles and common binding sites

8.2.4 Promoter prediction

Prediction methods

NN: NNPP and PROMOTER 2.0

Profiles: weight matrices to identify promoter sites (PromoterScan, TFsearch, TESS, MatInspector, ConsInspector…)

LDA (Linear Discriminant Functions): TATA-box score as discriminant (TSSD, TSSW)

Quadratic discriminant analysis: sequence length, different and overlapping windows as discriminants (CorePromoter)

Multiple pattern: binding sites are clustered (FastIM)

Eukaryotic Promoter DB (EPD):

ftp://ftp.epd.unil.ch/pub/databases/epd/views

Splice site and gene recognition:

http://linkage.rockefeller.edu/wli/gene/programs.html

http://hto-13.usc.edu/software/procrustes/index.html

http://cmgm.standford.edu/classes/genefinding

http://www1.imim.es/courses/SeqAnalysis/GeneIdentification/Evaluation.html

8.2.5 ESTs Clusters

From mRNA

↓

cDNA

↓

Cloned into cDNAs libraries

↓

Sequenced at each end to obtain ESTs

↓

ESTs compared to one another and to genomic sequences

↓

When building a cluster of overlapping sequences: Possible Gene found

GrailEXP: ESTs data searches for predicted genes confirmation

8.2.5 Performance of Gene Prediction Methods

- TP: positive correctly predicted
- FN: positive incorrectly predicted
- FP: negative incorrectly predicted
- TN: negative correctly predicted

true	predicted		total
	+1	-1	
+1	TP	FN	TP + FN
-1	FP	TN	FP + TN
total	TP + FP	FN + TN	N

Confusion matrix Accuracy = TP + TN / N

Specificity = TN/TN + FP

Sensitivity = TP/TP + FN

Balanced E = ½(Sp + Se)

8.2.5 Performance of Gene Prediction Methods

Method	Sensitivity	Specificity	Matthews
GenParser	0.69-0.75	0.68-0.78	0.66-0.69
GeneID	0.65-0.67	0.74-0.78	0.66-0.67
Grail	0.48-0.65	0.86-0.87	0.61-0.72

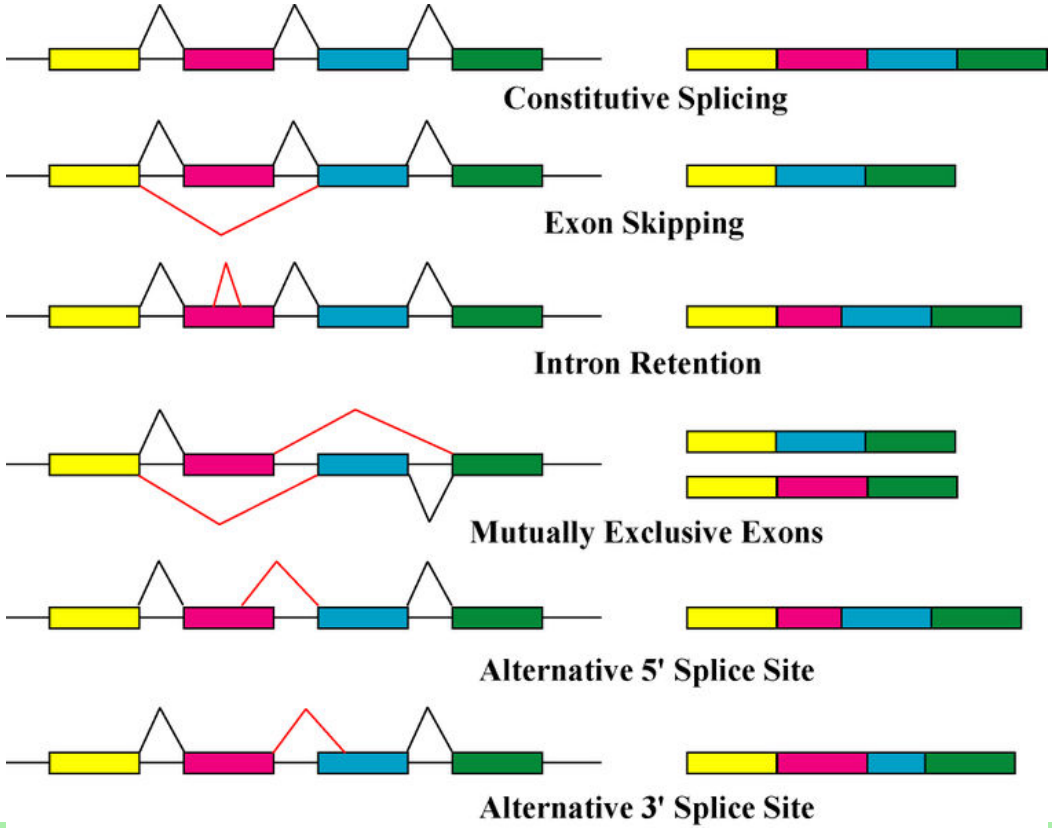
Finding the nucleotides ends of exons → Higher prediction performance

Method	Sensitivity	Specificity	Matthews
Grail	0.79	0.92	0.83
FGENEH	0.93	0.93	0.85
MZEF	0.95	0.95	0.89

8.3.1 Alternative Splicing

60% of human genes are alternatively spliced

- one pre-mRNA produces more different mature mRNA→ more different proteins
- Tissue-specific, developmentally-regulated, responsitive to physical condition



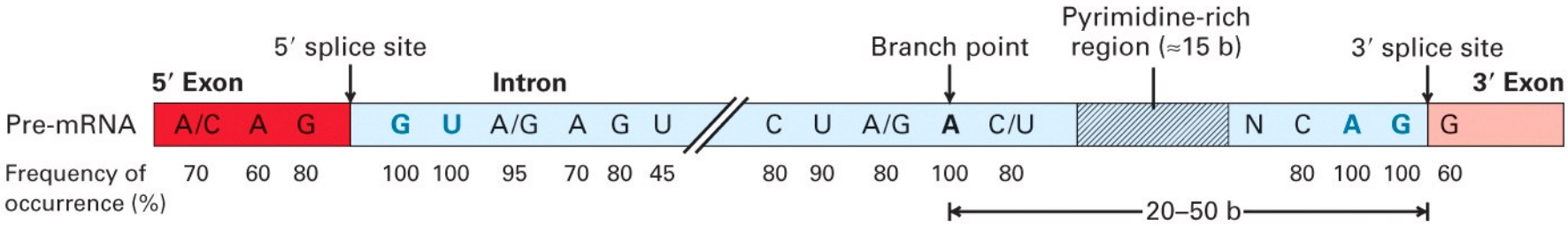
Nucleotides sequences signals needed for splicing:

5‘ end splice site

3‘ end splice site

branch point

8.3.1 Alternative Splicing



Splicing signals with consensus sequences

AS major machinery : **Spliceosome**

- 5 U-rich snRNAs (small nuclear RNAs) (U1, U2, U3, U4, U5)→ splicing signals Recognition
- snRNAs are associated with 6 to 10 proteins in snRNPs (small nuclear ribonucleoprotein particles)
- Video

8.3.1 Alternative Splicing

AS processes: hight complexity and hard to identify the splicing signals and operator sequences involved

Pattern recognition methods are used for the analysis of AS regulatory regions

Long Short-Term Memory (LSTM) model allows information in the activations to be retained over a long period of time steps

Recurrent neural network which is able to

deal with sequences of different lengths

recognise complex, global patterns in sequences by storing informations computed from the past inputs

input

- sequence
- position information
- length information

8.3.1 Alternative Splicing

Compared to feed-forward neural networks **recurrent neural networks** contain cycles

Recurrent neural network the information is stored in two distinct ways

Activations of the units → short-term memory

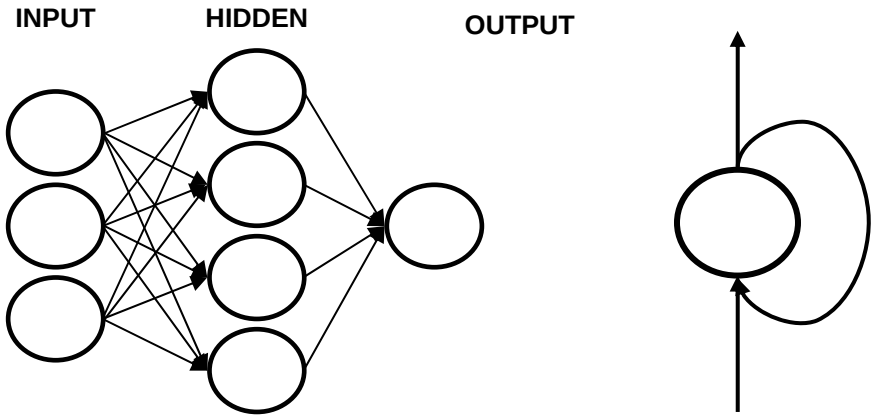
Weights → long-term memory

(weights are modified based on experience)

LSTM uses a linear activation function

for the memory cells (cells with feedback)

$$f(z) = z$$



Activation of the memory cell at the time *t* depends on the activation of the memory cell at the time *t-1*

$$y_t = f(w \cdot y_{t-1})$$

How changes in y_{t-1} influences the value of y_t can be expressed by the derivative

$$\frac{\partial y_t}{\partial y_{t-1}} = w \cdot f'(w \cdot y_{t-1})$$

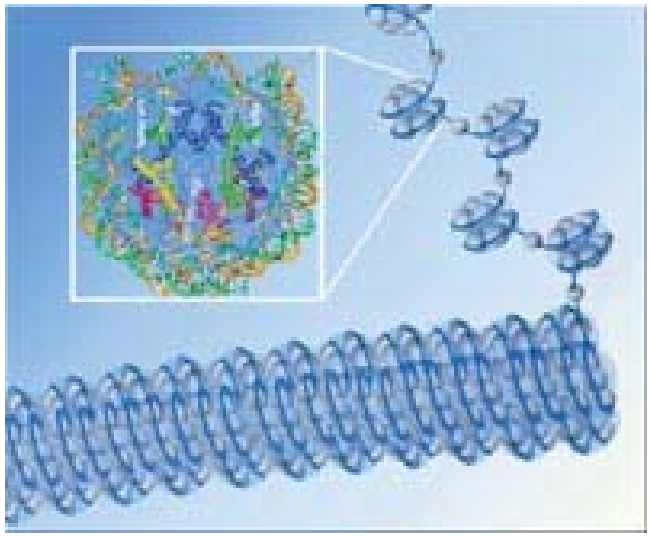
In order to achieve $\frac{\partial y_t}{\partial y_{t-1}} = 1$, *f(z)* is set to *z*

Then $y_t = w \cdot y_{t-1}$ and $\frac{\partial y_t}{\partial y_{t-1}} = w \Rightarrow w = 1$

8.3.2 Nucleosomes

Eukaryotic DNA wrapped around histone-protein complexes→ chromosomes

Gene expression regulation → accessibility to promoter sites



High density:

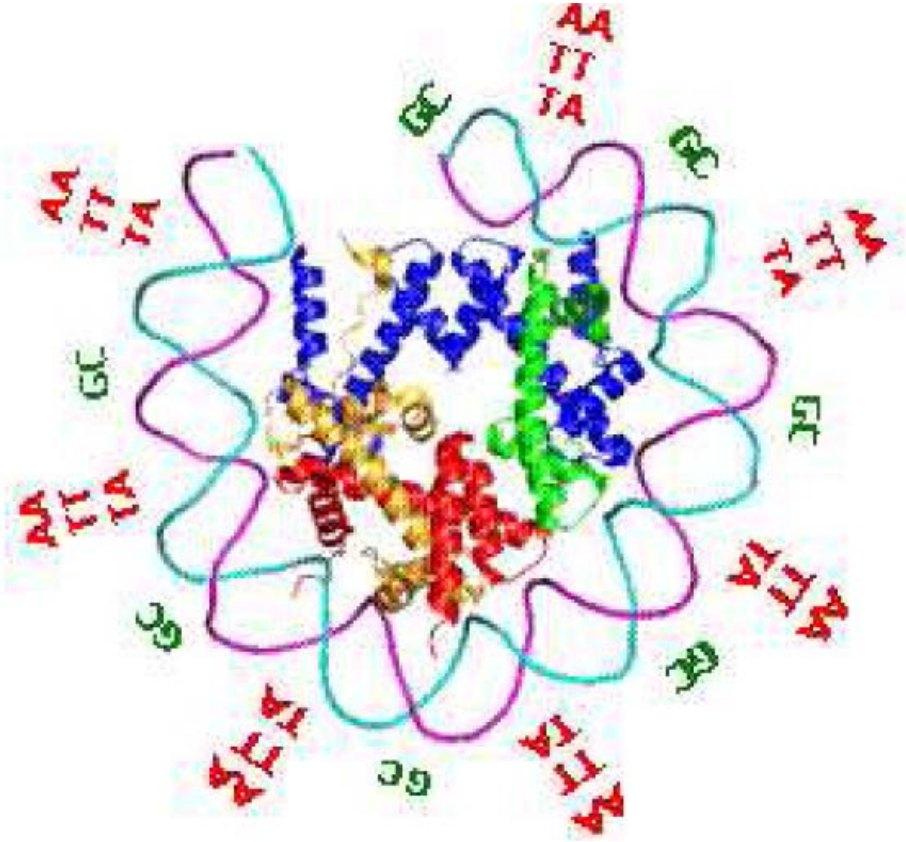
- Centromeres

Low density:

- TF binding sites
- Transcription initiation sites
- Ribosomal RNA and tRNA coding sites

http://www.lbl.gov/Publications/Currents/Archive/view-assets/Apr-30-2004/nucleosomes_chromatin.jpg

Segal et al. Nature, 2006*A "genomic code for nucleosome positioning* Markov model for nucleosome position"



To find a nt in position i depends ONLY on the nt in position i-1

The nucleosome wrapping composed by 147 bp

Model:

$$p(s) = p_1(s_1) \prod_{i=2}^{147} p_i(s_i \mid s_{i-1})$$

10 bp frequency AA, TT, TA with alternates GC

↓

Expected bpN° pro turn around the nucleosome (HMM basis)= 10

8. DNA Analysis
8.4 Comparative Genomics

Genes:

Homologous: with a common ancestor and sharing the same function

Orthologous: with a common ancestor but evolved through Speciation,,

The specie diverged into two species (gene replacement experiment)

Paralogous: with a common ancestor and evolved after duplication with possible new function acquisition (mutation)

Gene duplication: Pseudogenes

New function (mutations on one copy are not penalized)

Comparisons and clustering

Proteins comparisons

Genome comparisons (also on the basis of protein sequences, ESTs + homolog searches and clustering)

8. DNA Analysis
8.4 Comparative Genomics

Synteny: local gene order conservation

Conserved between close related species

Evolutionary trees: Rearrangements and synteny to estimate the evolutionary distance between species

↓

Computed distance by elementary rearrangement steps to transfer the genome from one specie into another

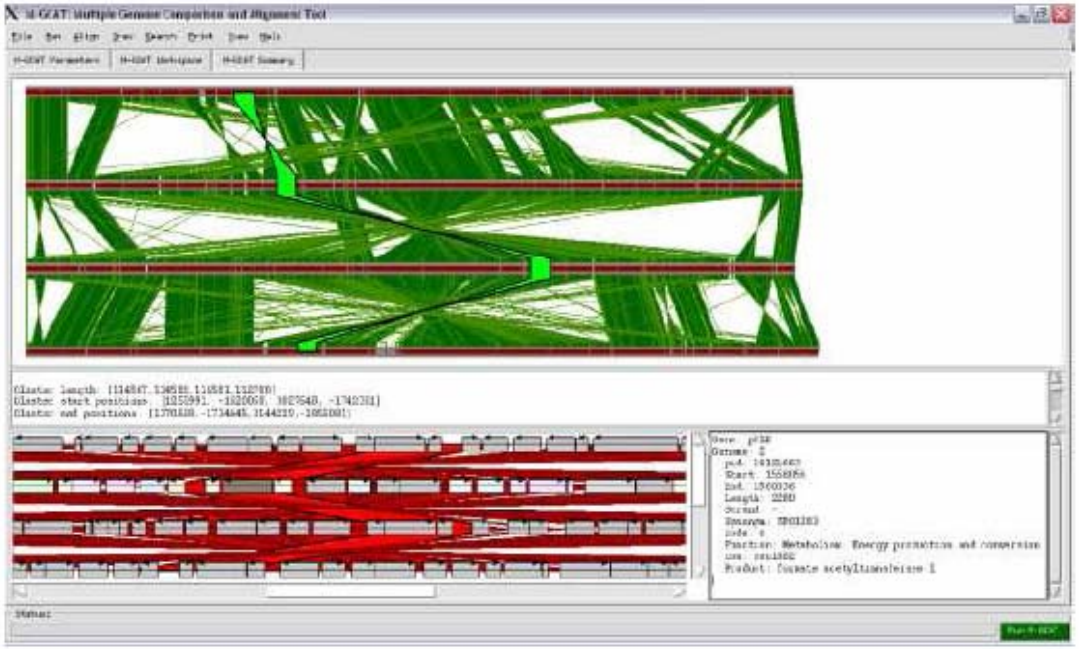
HGT: Horizontal Gene Transfer,, genomic material from one specie is included directly into the genome of another specie (mitochondria or chloroplast)

Detection by a deviation of base frequencies in a region of a genome

LGT: Lateral Gene Transfer

8. DNA Analysis
8.4 Comparative Genomics

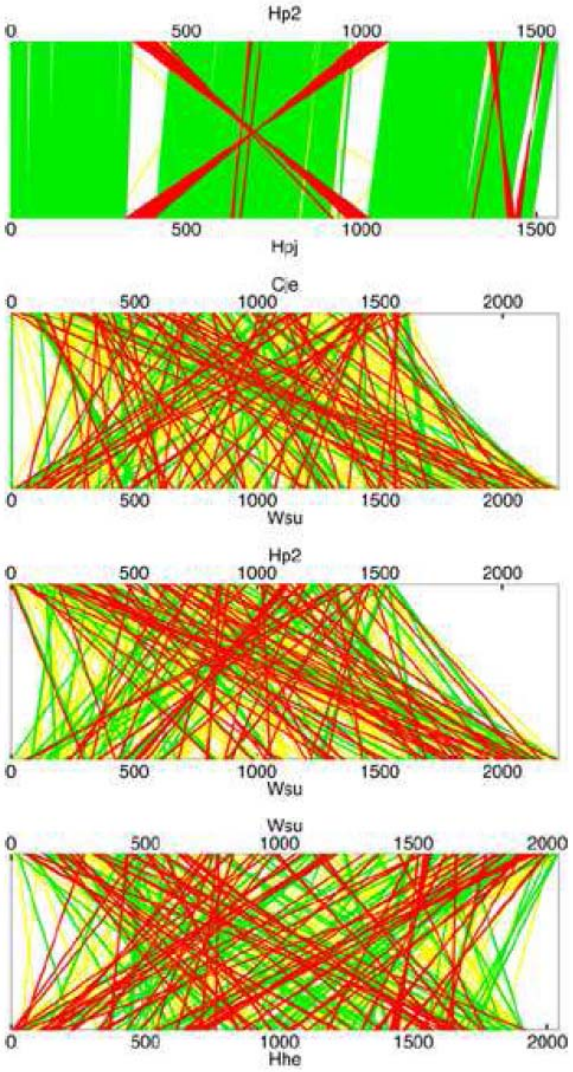
Study of the relationship of **genome structure** and function across different biological species or strains (wikipedia): Gene locations, duplications, sequence repeats (location and length), single mutations (promoter)



M-GCAT Multiple Genome Comparison and Alignment Tool (alignment frameworks among closely related bacterial species)

Maps as Genome comparison results between species

8. DNA Analysis
8.4 Comparative Genomics



M-GCAT Multiple Genome Comparison and Alignment Tool

Genome comparison of campylobacterales

Within one specie the genomic order is conserved but not between species

HGT: Horizontal Gene Transfer

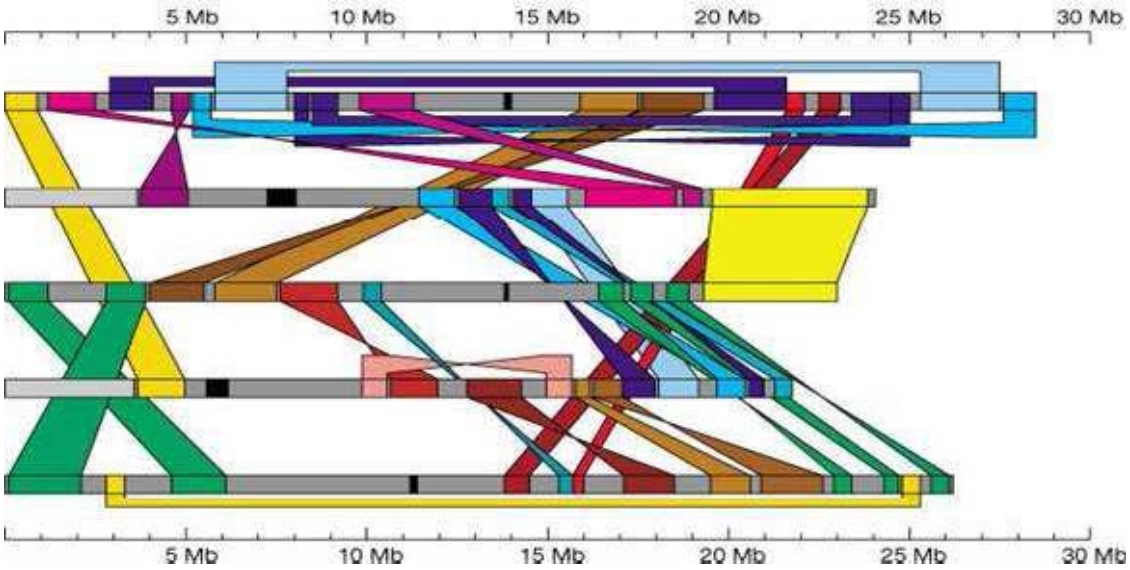
Adaptation, mutations and rearrangements

Wolinella and Bdellovibrio genome comparison

8. DNA Analysis
8.4 Comparative Genomics

One genome compared to itself or its chromosomes one another

Segmental duplication between the chromosomes of Arabidopsis



Sequenced regions cover 115.4 Mbp over the 125 Mbp of the entire genome)

25 498 genes
11 000 protein families

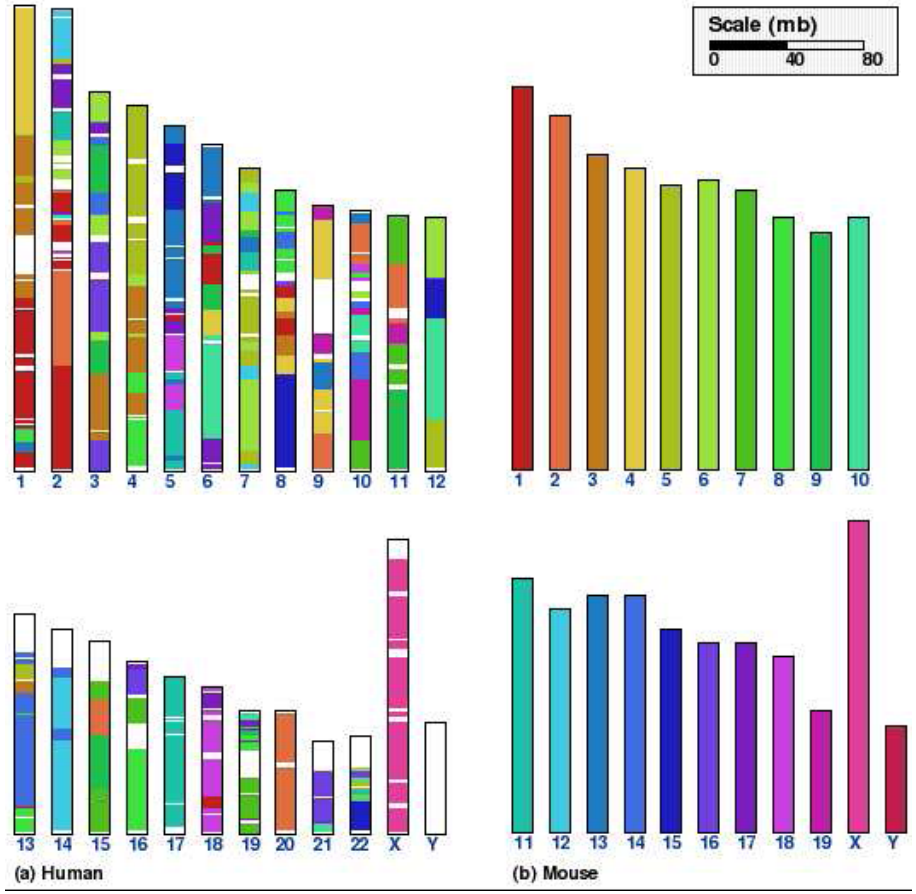
Evolution: whole genome duplication + gene loss + extensive local gene duplications

Dynamic genome enriched by LGT (Lateral Gene Transfer)

©Arabidopsis Genome Initiative

8. DNA Analysis
8.4 Comparative Genomics

Analysis within the genome : Interdependent or related genes are clustered and inherited as blocks

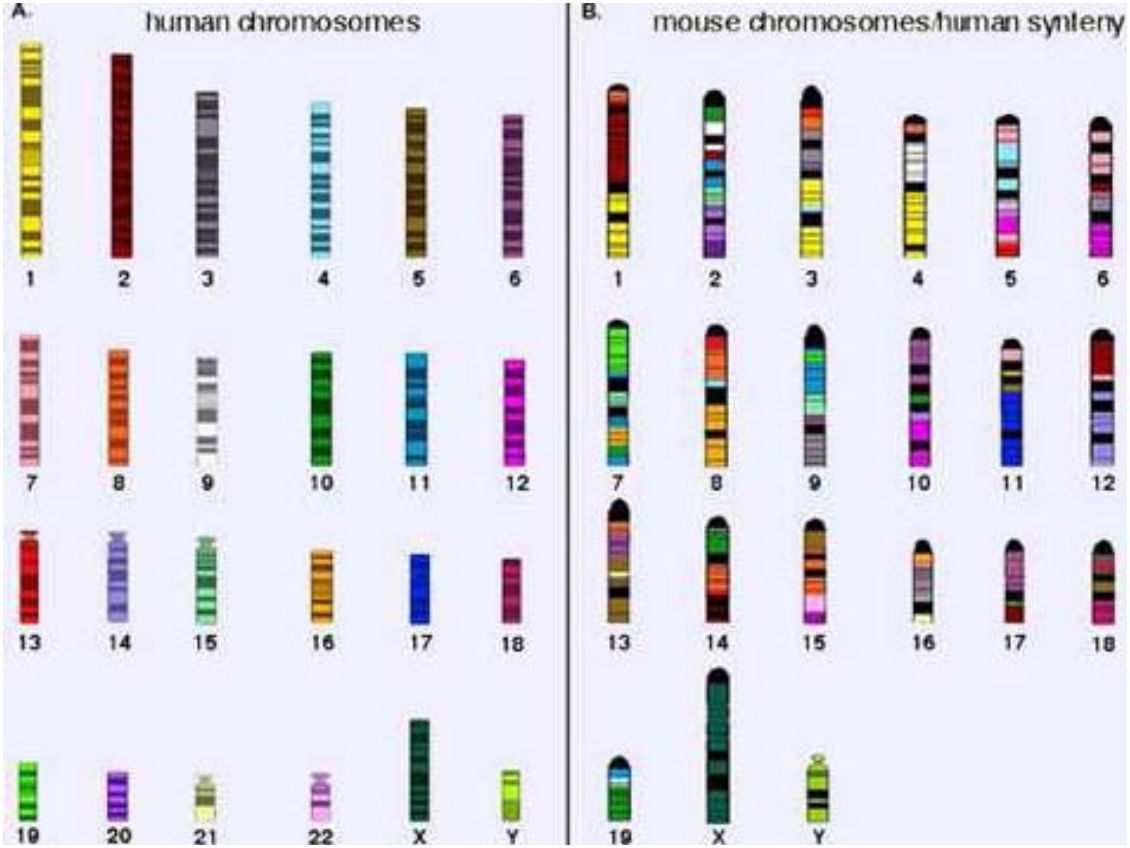


Mouse and Human genomic similarities

Mouse chromosomes mapped the human chromosomes

FISH and Giemsa staining as techniques to identify corresponding chromosomes location in different species

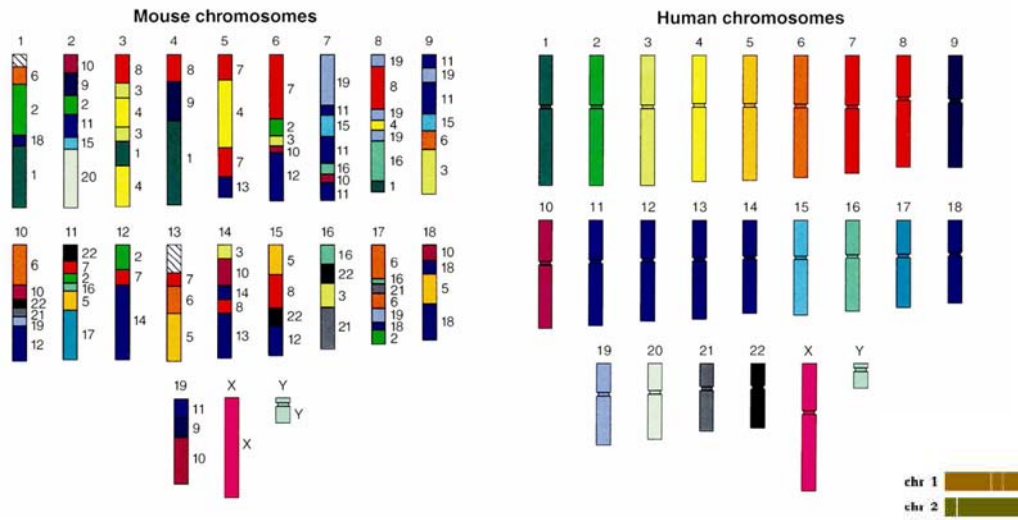
Color TO Chromosome number



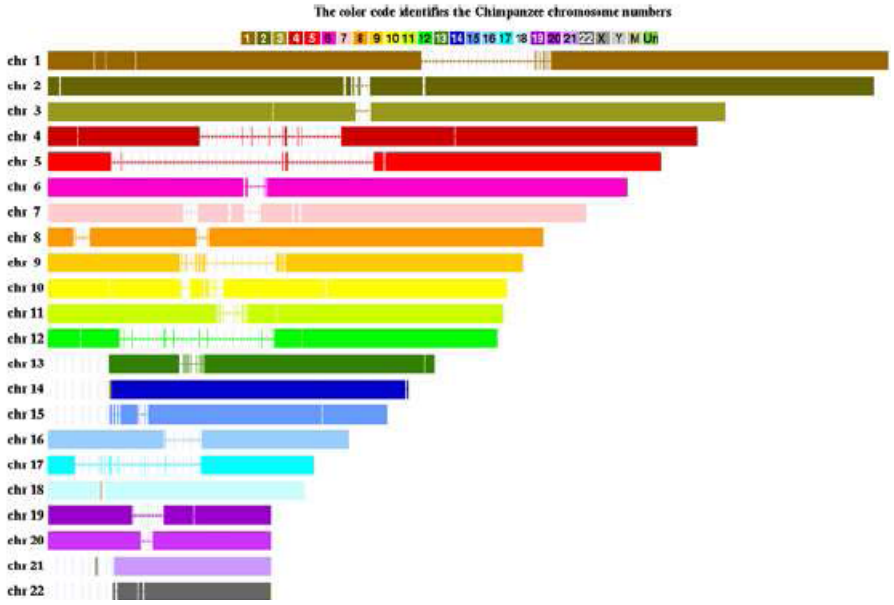
Syntony: gene local order

Human chromosomes mapped to the mouse C.

8. DNA Analysis
8.4 Comparative Genomics



Mouse and Human genomic similarities and gene clusters

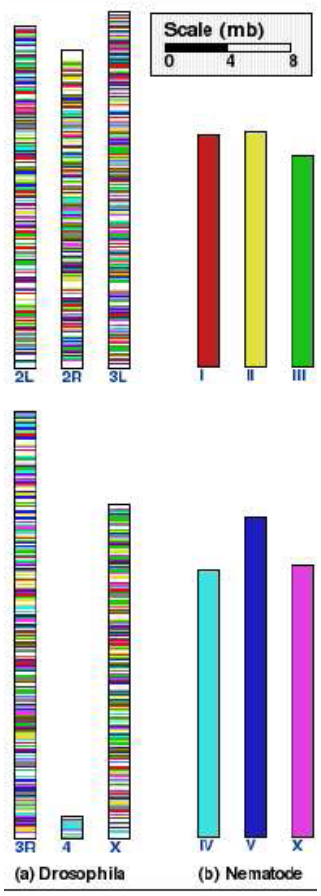


Chimpance and Human gene clusters

Color identifies the Chimpance chromosome numbers

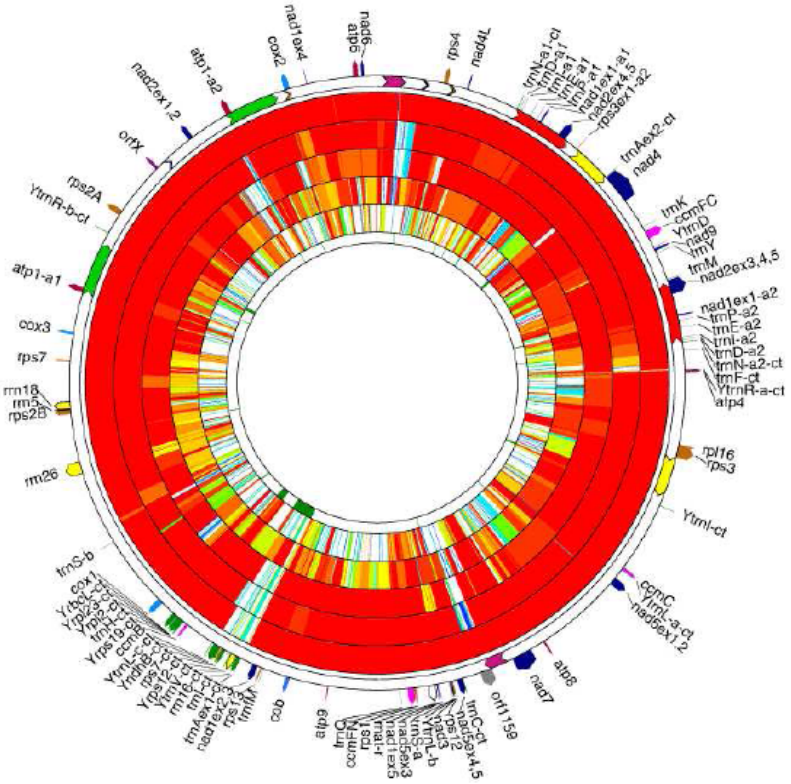
8. DNA Analysis
8.4 Comparative Genomics

Worm (nematode) chromosomes mapped to the fruit fly chromosomes



Color: nematode chromosome number

mtDNA genome comparisons within the *Zea mays* family (maize)



outermost circle: mtDNA Zea family

Innermost circle: Sorghum biocolor mtDNA

8. DNA Analysis
8.5 Genomic Individuality

8.5.1 Sequence Repeats

Sequence Repeats: Tandem repeats along side the whole genome

Junction of the same sequence

Special base pair distribution: mass per volume or buoyant density

Measurement methods able to separate DNA fragments depending on different densities

↓

Satellite DNA

Satellites	**Mini-satellites**	**Microsatellites**
Repeats of one thousand to several thousands bp in tandem region up to 100 million bases long	VNTRs repeats of 15bp in regions from 100 kb up to 1000 kb Varying in size (individuality identification)	SSRs or STR, short repeats of 2-6 bp in regions from 100 up to 100 bp Length inherited (evolutionary studies and gene markers /TTAGGG)

8.5.1 Sequence Repeats

Transposable elements: DNA regions which can jump from one location on the chromosome to another leaving prints as repetitive sequence

Organism	% transposable
H. sapiens - human	35
Z. mays - maize	50
D. melanogaster - fruit fly	15
A. thaliana - plant	2
C. elegans - nematode	1.8
S. cerevisiae - budding yeast	3.1

10% of SINE (Short Interspersed Nuclear Elements) ALu 1.2 mill copies
Implicated in several inherited human diseases; Human population genetics and evolution of primates

14.6 LINE (Long Interspersed Nuclear Elements)

↓

Reverse transcriptase and RNA based transposition (more dominant)

Long terminal repeat retrotransposons,

Long terminal repeats retroposons

Long terminal repeat retrovirus-like

OR

DNA based dynamics of transposition (200 000 copies in the human genome)

Hybrids in MITES (Miniature Inverted repeats Transposable Elements)

8.5.2 SNPs

Single Nucleotide Polymorphism: DNA variation

One single nucleotide differs from species in at least 1% population

Alleles C and T: gtagCccc gtagTccc

Placed in Exons: amino acid change in the protein

Placed in Introns: regulatory effects e.g. splicing alteration, transcription factor affinity influcence, ...

Polymorphism	*Schizophrenia* *n = 279*	*Control* *n = 255*	χ^2	*df*	*P*
g.-888G>C					
Genotype			2.99	2	0.22
GG	243 (87.1%)	209 (81.9%)			
GC	34 (12.2%)	42 (16.5%)			
CC	2 (0.7%)	4 (1.6%)			
Allele			3.16	1	0.08
G	520 (93.2%)	460 (90.2%)			
C	38 (6.8%)	50 (9.8%)			

Diseases and human sensitivity

Pathogens responses

Drug treatment & individual medicine

- Pharmacogenomics
- Pharmacogenetics

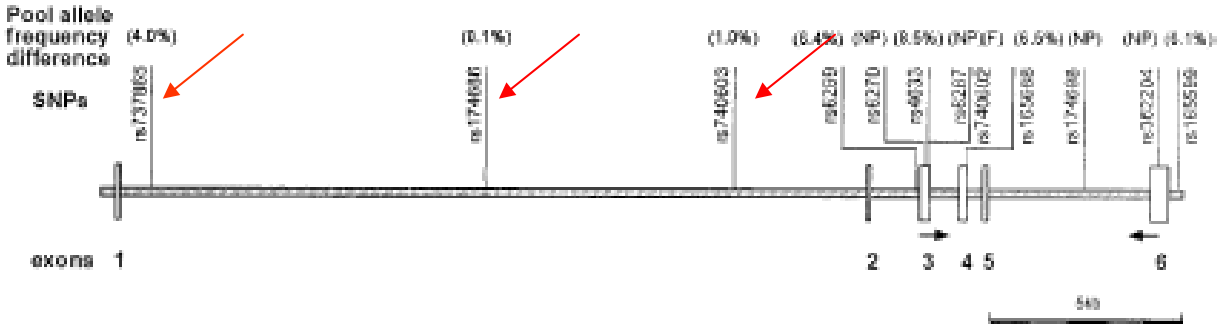
8. DNA Analysis

8.5 Genomic Individuality

SNPs detected by: specific enzymes (restrictases) and Microarray technique

COMT (catechol-O-Metyltransferase, dopamine metabolism) haplotype SNPs and Schizophrenia

3SNPs rs6270, rs6267, rs165688 → non synonymous changes



Lactase SNP and Milk metabolism

C>T Intron of gene MCM6 13 region →SI001784U/rs4988235

Allele C	C	C	5' - atacagataagataatgtag C ccctggcctcaaaggaactc - 3'
Allele T	T	T	5' - atacagataagataatgtag T ccctggcctcaaaggaactc - 3'

8. DNA Analysis
8.5 Genomic Individuality

SNPs are inherited within the blocks contained in the chromosomes during genome material replication: Haplotypes blocks

Two haplotypes in each human chromosome

One allele per chromosome

SNP database

www.ncbi.nlm.nih.gov/SNP

SNP consortium

http://snp.cshl.org

International HapMap Project (haplotype map project)

www.hapmap.org