Computer genomics lectures

- Internet-resources for DNA sequences search and analysis
- General description of bioinformatics
- Genes and genomes databases
- Genomic DNA: superposition of many codes
- Repeats in DNA, formal description and analysis
- Consensus, position weight matrices, sequence alignment methods
- Machine learning methods overview: hidden markov models, neural networks, genetic algorithm
- Methods of eukaryotic gene structure recognition
- Statistical approaches used for different recognition methods accuracy comparison
- Investigation of qualitative and quantitative characteristics of transcriptome

Internet-resources for DNA sequences search and analysis

Major bioinformatics databases

GenBank EMBL DDBJ	http://www.pubmed.com http://www.ebi.ac.uk/embl.html http://www.ddbj.nig.ac.jp	Nucleotide and Protein Sequences
PubMed	http://www.pubmed.com	Bibliographic database
EnsEMBL UCSC Genome Browser	http://www.ensembl.org http://genome.ucsc.edu	Genes and genomes analysis and annotation. Genes and genomes search and visualization tool
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq	Non-redundant sequence database of genomes, transcripts and proteins
UniGene STACK GeneCards GenAtlas GeneOntology TIGR Gene Indices	http://www.ncbi.nlm.nih.gov/UniGene http://www.sanbi.ac.za/Dbases.html http://www.genecards.org http://www.dsi.univ-paris5.fr/genatlas http://www.geneontology.org http://www.tigr.org/tdb/tgi.shtml	Gene Database
SWISSPROT	http://www.expasy.ch	Protein Database
EPD	http://www.epd-isb-sib.ch	Eukaryotic Promoter Database

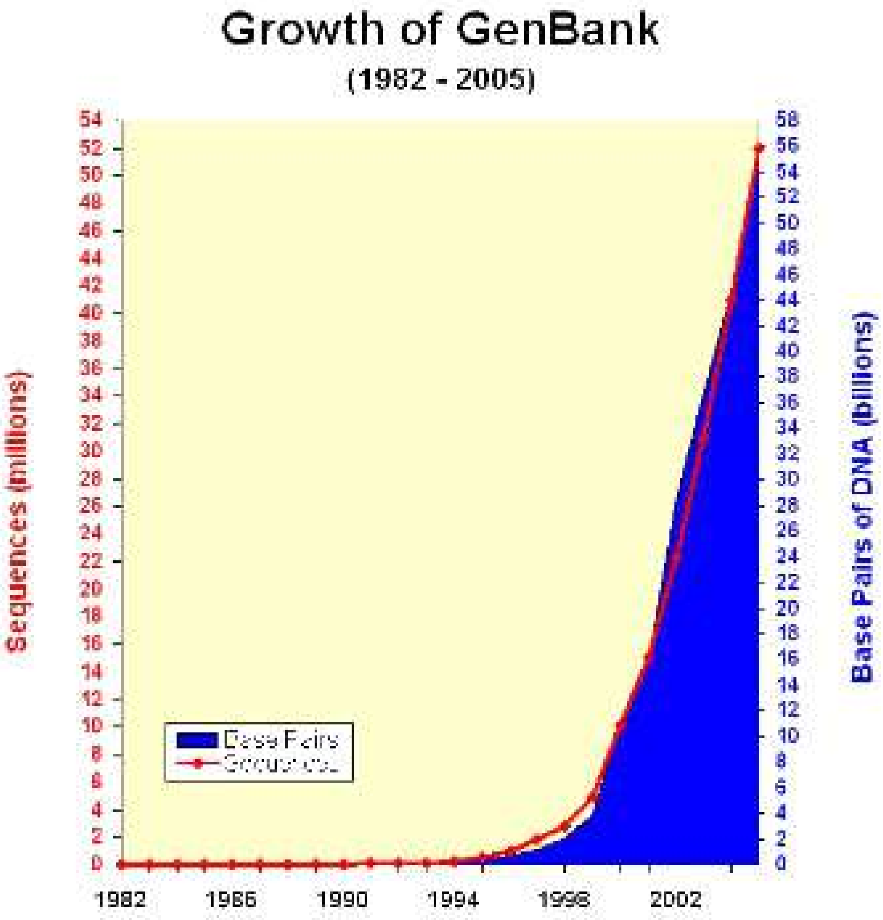
General description of bioinformatics

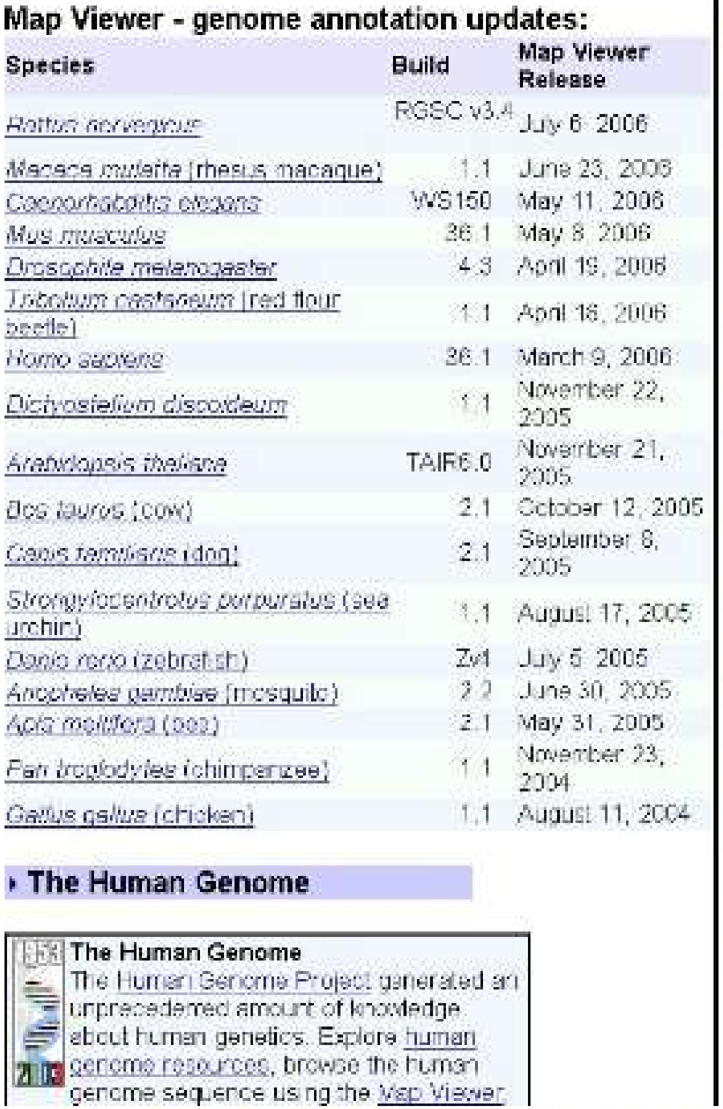
Basic directions of bioinformatics development

- Homology search, multiple alignment
- Statistical analysis of genetic texts, genome segmentation
- Recognition of coding sequences and open reading frames
- Recognition of DNA functional sites
- Analysis of RNA secondary structure
- Analysis of protein sequences, protein secondary structure prediction, recognition of functional sites in proteins
- Phylogenetic analysis
- DNA-chips, DNA microarray: expression analysis
- Database surfing: manipulation with a large amount of data

Genes and genomes databases

Genome annotation progress



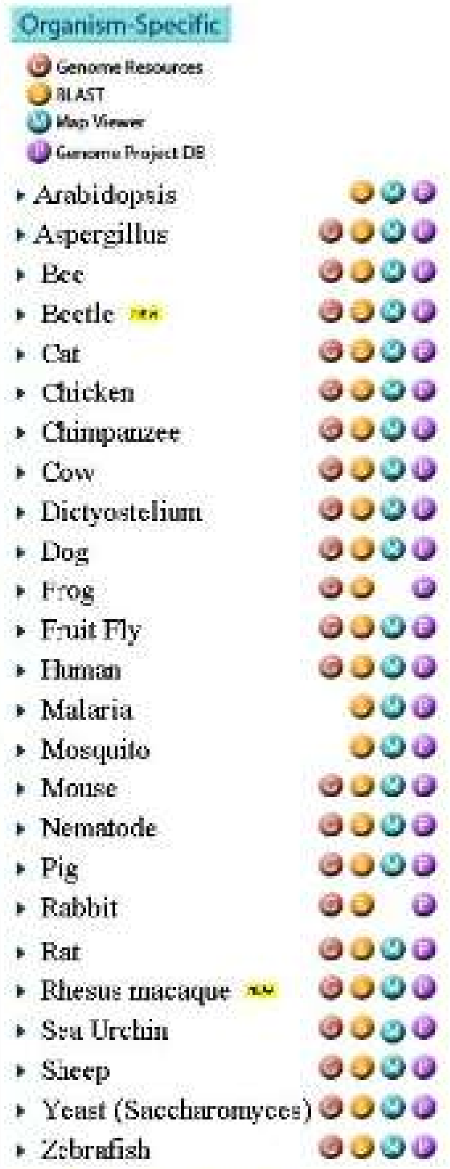


Map Viewer - genome annotation updates:

Species	Build	Map Viewer Release
Rattus norvegicus	RGSC v3.4	July 6, 2006
Macaca mulatta (rhesus macaque)	1.1	June 23, 2006
Caenorhabditis elegans	WS150	May 11, 2006
Mus musculus	36.1	May 8, 2006
Drosophila melanogaster	4.3	April 19, 2006
Tribolium castaneum (red flour beetle)	1.1	April 18, 2006
Homo sapiens	36.1	March 9, 2006
Dictyostelium discoideum	1.1	November 22, 2005
Arabidopsis thaliana	TAIR6.0	November 21, 2005
Bos taurus (cow)	2.1	October 12, 2005
Canis familiaris (dog)	2.1	September 8, 2005
Strongylocentrotus purpuratus (sea urchin)	1.1	August 17, 2005
Danio rerio (zebrafish)	Zv4	July 5, 2005
Anopheles gambiae (mosquito)	2.2	June 30, 2005
Apis mellifera (bee)	2.1	May 31, 2005
Pan troglodytes (chimpanzee)	1.1	November 23, 2004
Gallus gallus (chicken)	1.1	August 11, 2004

The Human Genome

The Human Genome
The Human Genome Project generated an unprecedented amount of knowledge about human genetics. Explore human genome resources, browse the human genome sequence using the Map Viewer,



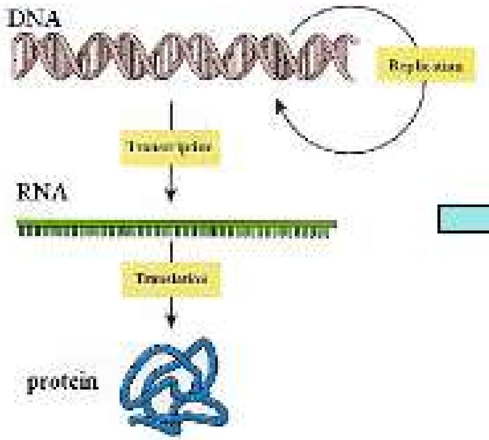
Organism-Specific

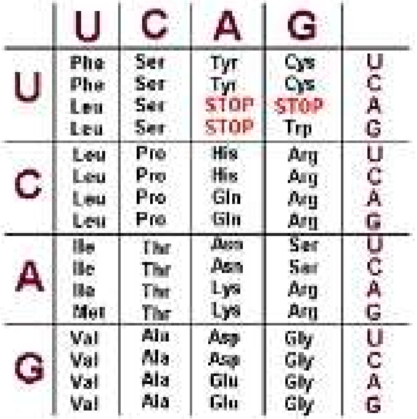
- ● Genome Resources
- ● BLAST
- ● Map Viewer
- ● Genome Project DB

Arabidopsis		●	●	●
Aspergillus	●	●	●	●
Bee	●	●	●	●
Beetle NEW	●	●	●	●
Cat	●	●	●	●
Chicken	●	●	●	●
Chimpanzee	●	●	●	●
Cow	●	●	●	●
Dictyostelium	●	●	●	●
Dog	●	●	●	●
Frog	●	●		●
Fruit Fly	●	●	●	●
Human	●	●	●	●
Malaria		●	●	●
Mosquito		●	●	●
Mouse	●	●	●	●
Nematode	●	●	●	●
Pig	●	●	●	●
Rabbit	●	●		●
Rat	●	●	●	●
Rhesus macaque NEW	●	●	●	●
Sea Urchin	●	●	●	●
Sheep	●	●	●	●
Yeast (Saccharomyces)	●	●	●	●
Zebrafish	●	●	●	●

http://www.ncbi.nih.gov/Genbank/genbankstats.html http://www.ncbi.nlm.nih.gov/Genomes/

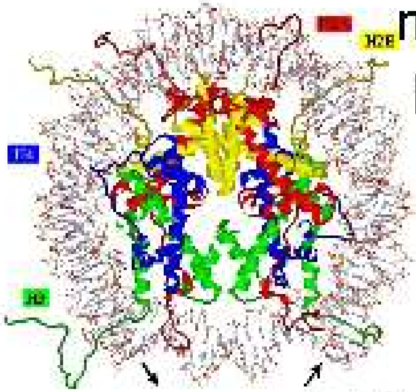
Genomic DNA: superposition of many codes





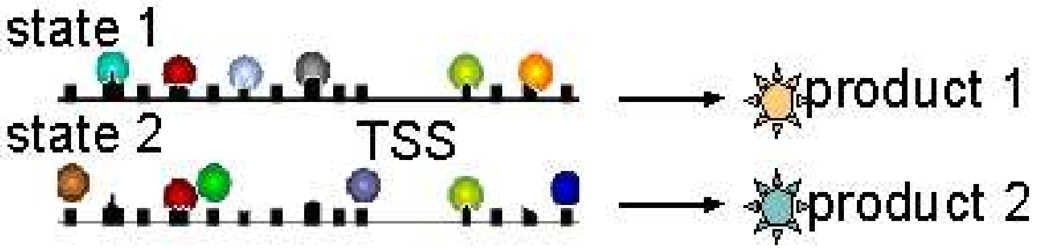
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

= **genomic DNA codes?**



nucleosome positioning

transcription regulation



state 1

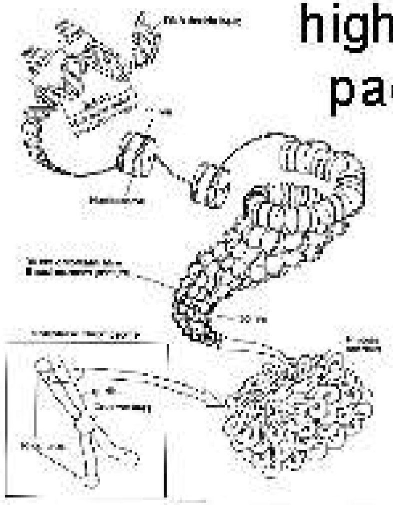
state 2

TSS

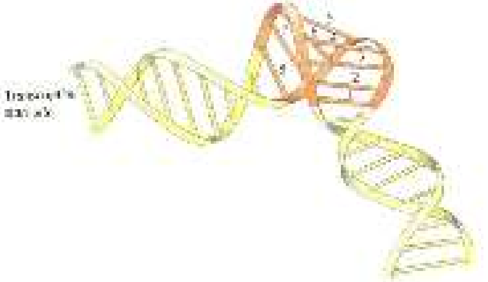
product 1

product 2

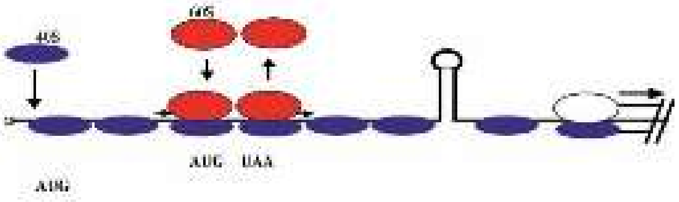
higher order packaging



DNA confomation



translation regulation



etc...

Repeats in DNA, formal description and analysis

Repeats: basic types

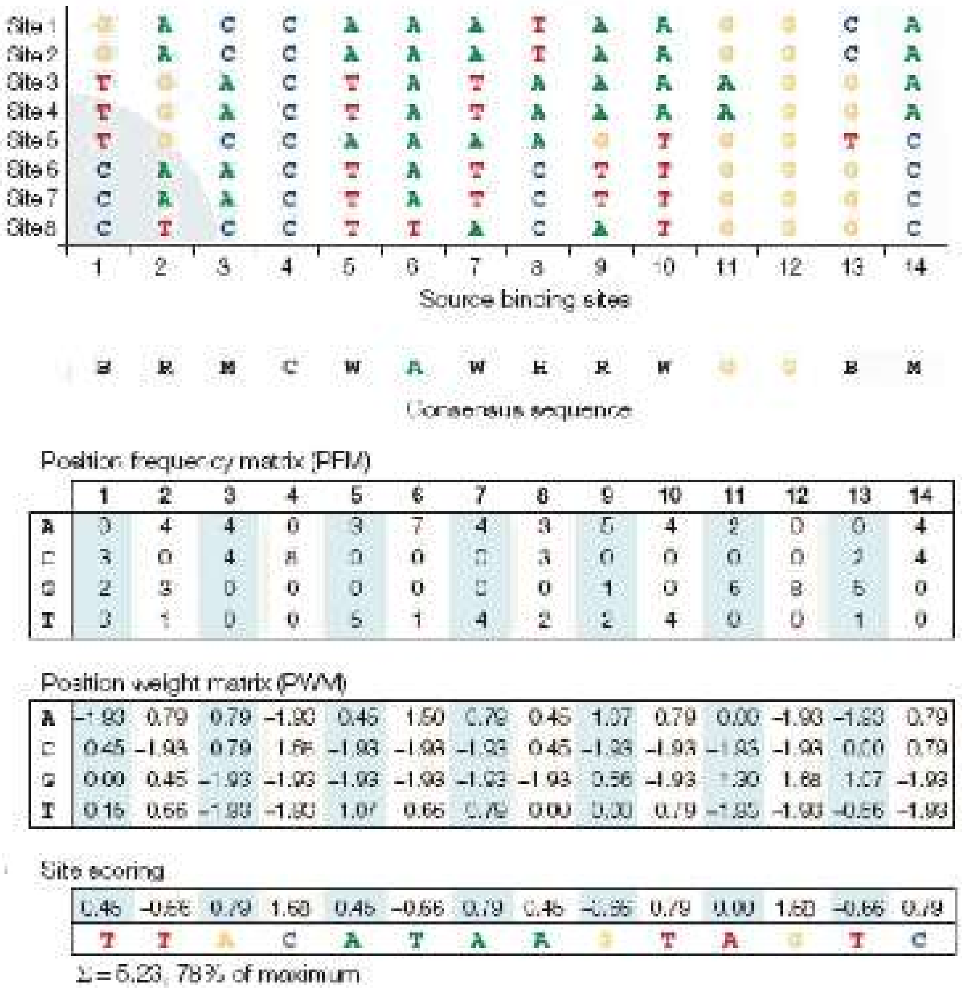
Type	Example	Direction	Complementarity
Direct	AGCTTT AGCTTT TCGAAA TCGAAA	Forward	No
Invert	AGCTTT AAAGCT TCGAAA TTTCGA	Reverse	Yes
Symmetric	AGCTTT TTTCGA TCGAAA AAAGCT	Reverse	No
Direct complementary	AGCTTT TCGAAA TCGAAA AGCTTT	Forward	Yes
Palindrome	AAGCCGAA TTCGGCTT	Reverse	No
Complementary palindrome	AAGCGCTT TTCGCGAA	Reverse	No

Repeats: possible mutual positioning

Type	Example
Dispersed	...AGTTC......AGTTC...
Tandem	... AGTTCAGTTC...
Overlapped	...AGTTCAGTTCAGTTC...

Consensus, position weight matrices, sequence alignment methods

Position weight matrix (PWM) model



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Position weight matrix (PWM)

A	–1.93	0.79	0.79	–1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	–1.93	–1.93	0.79
C	0.45	–1.93	0.79	1.68	–1.93	–1.93	–1.93	0.45	–1.93	–1.93	–1.93	–1.93	0.00	0.79
G	0.00	0.45	–1.93	–1.93	–1.93	–1.93	–1.93	–1.93	–0.66	–1.93	1.30	1.68	1.07	–1.93
T	0.45	–0.66	–1.93	–1.93	1.07	–0.66	0.79	0.00	0.00	0.79	–1.93	–1.93	–0.66	–1.93

Site scoring

0.45	–0.66	0.79	1.68	0.45	–0.66	0.79	0.45	–0.66	0.79	0.00	1.68	–0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum

A set of aligned binding sites

Consensus model

Position frequency matrix (PFM): the count of observed nucleotides at each position

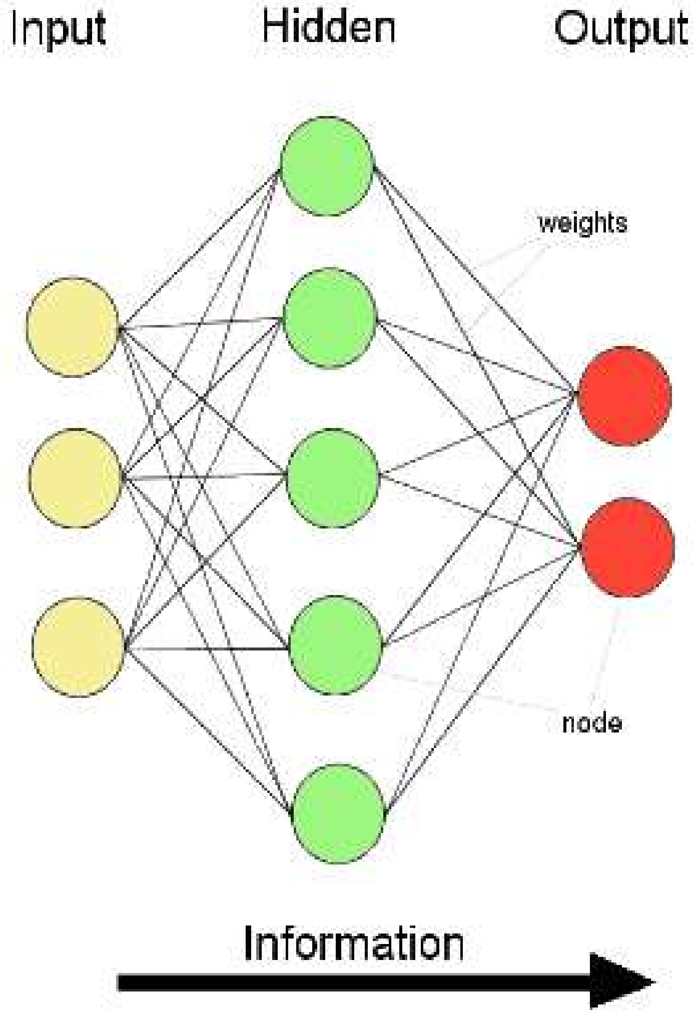
PFM is converted to a position weight matrix (PWM) using a special formula

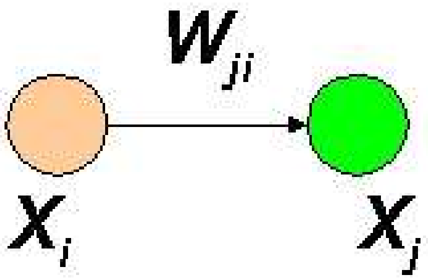
Using a PWM model, a score for any DNA sequence can be calculated by summation over all positions

- *Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004, 276-287.*

Machine learning methods overview: neural networks, genetic algorithm

Artificial neural network: oriented multigrapf of artificial neurons with weighted connections





Network parameter

weight of connection between the neurons $x_i$and x_j.

$$w_{ij} \leftarrow w_{ij} + \Delta w_{ij}$$

Network learning

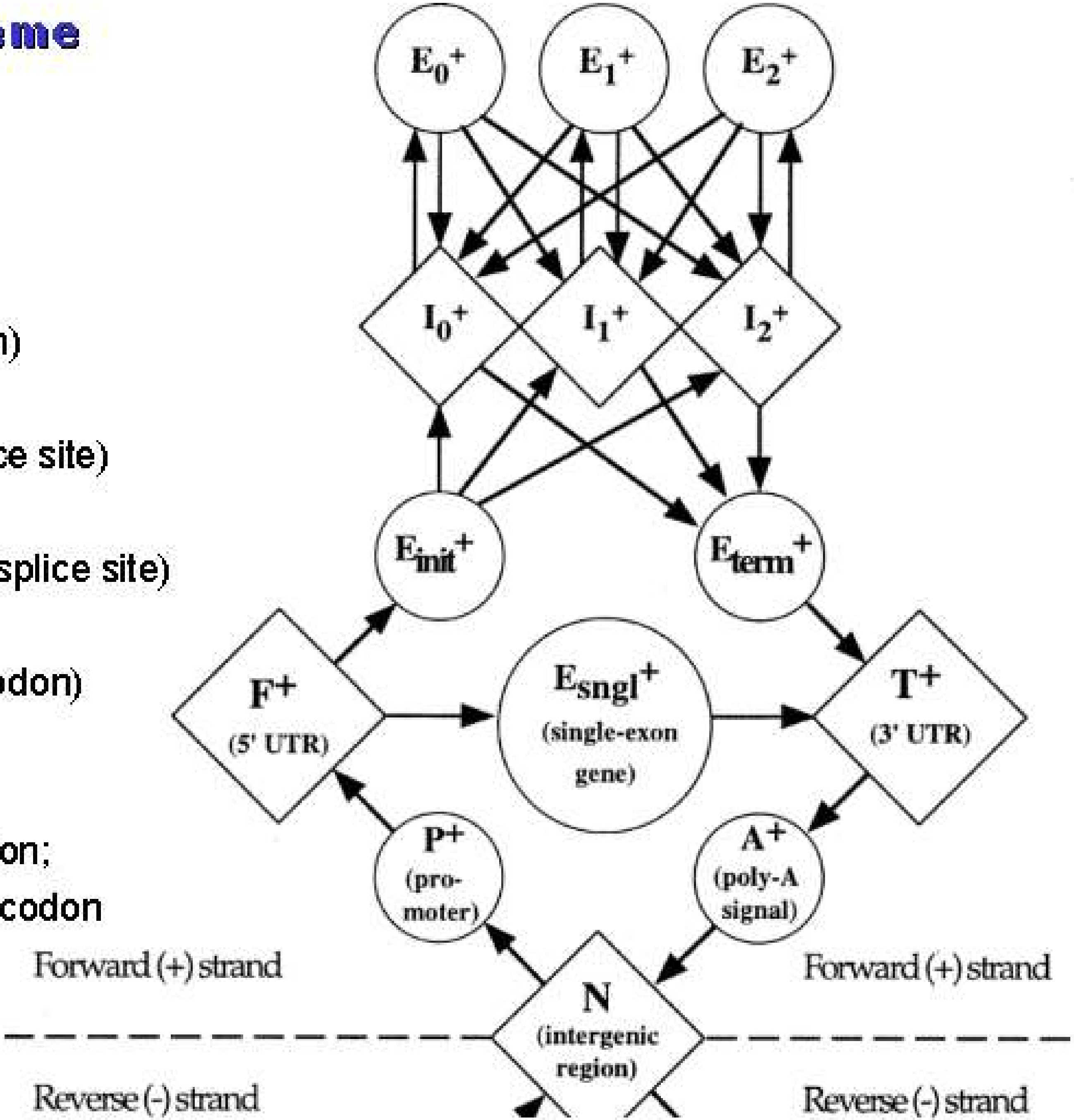
Weight modification according to the learning paradigm (supervised,unsupervised, Hebbian, reinforcement...)

Learned network with the fixed weights presents the knowledge about the world

Methods of eukaryotic gene structure recognition

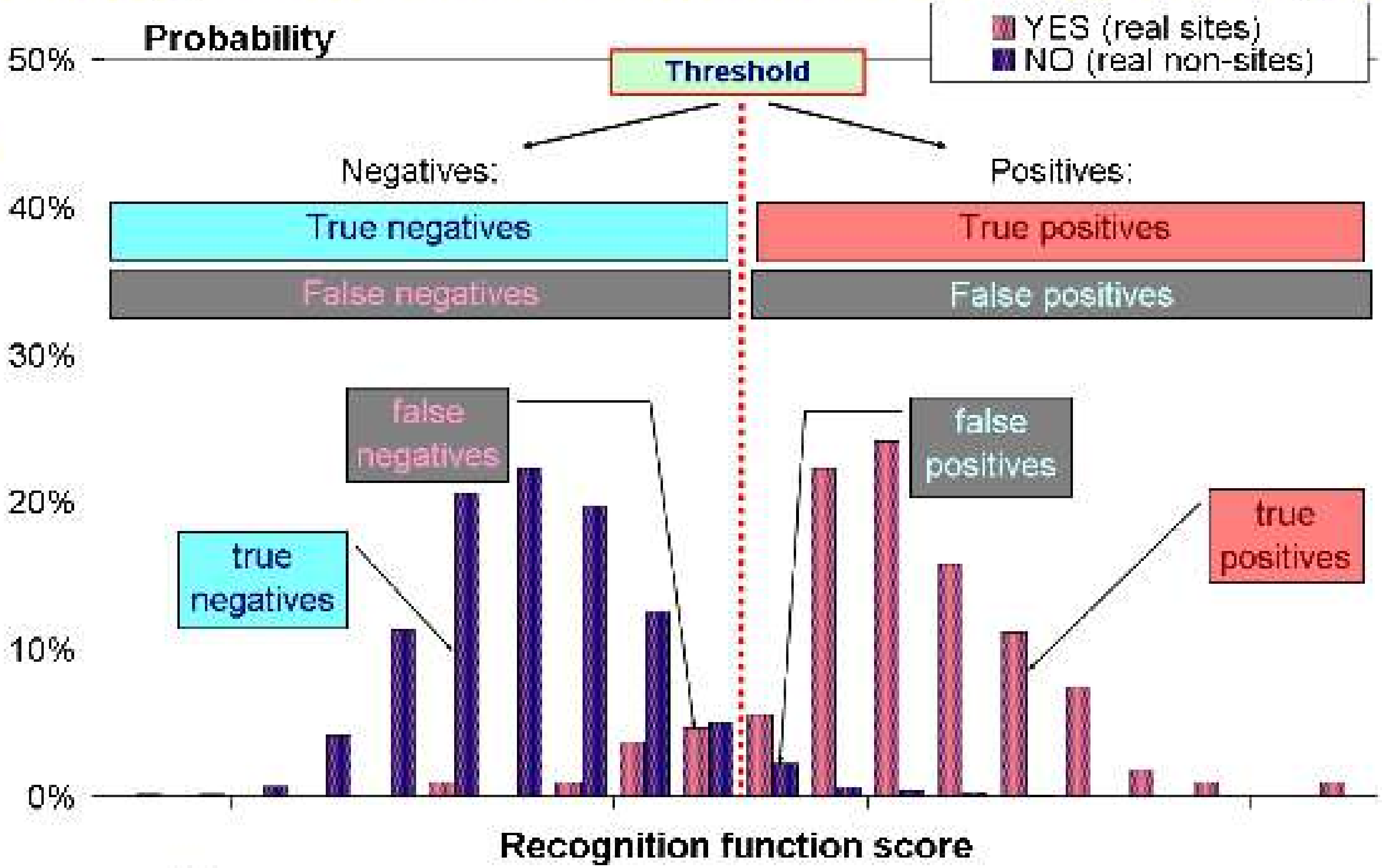
Program GenScan scheme

N	intergenic region
P	promoter
F	5'-untranslated region
E_{sngl}	single exon (intronless) (translation start –> stop codon)
E_{init}	initial exon (translation start –> donor splice site)
E_k	phase k internal exon (acceptor splice site –> donor splice site)
E_{term}	terminal exon (acceptor splice site –> stop codon)
I_k	phase k intron: 0 – between codons; 1 – after the first base of a codon; 2 – after the second base of a codon



Statistical approaches used for different recognition methods accuracy comparison

Predictions



Contingency table

		Prediction	
		Non-sites	Sites
Reality	Sites	FN, false negatives Not predicted real sites	TP, true positives Correctly predicted real sites
	Non-sites	TN, true negatives Correctly predicted non-sites	FP, false positives Real non-sites predicted as sites

Investigation of qualitative and quantitative characteristics of transcriptome

Comparision of methods for transcripts detection and abundance estimation

Method	Relative/absolute measurability, compatibility	Genes amount	Sensitivity & dynamic range	High throughput capacity
Direct mRNA detection via hybridization of transcripts with ssDNA or RNA probes				
Nothern blot hybridization	+	(1-5) x (5- 20)	- -	- -
Ribonuclease protection	+ +	(10-15) x (5-20)	+	- -
Detection of cDNA made by reverse transcription from mRNA				
Quantitative RT-PCA, Real time RT-PCA	+ +	tens	+ + (!)	+ +
Differential display	+	tens	+ + (!)	+ +
Oligonucleotide/cDNA microarrays	+ + (!)	thousands	+ + +	+ + +
Computational analysis of cDNA reads, «*in silico* hybridization» of transcripts				
SAGE	+ + +	thousands	+ + +	+ + +
MPSS	+ + + +	thousands	+ + + +	+ + + +
EST	+	thousands	+ + + (!)	+ + +

(!) Caution about possibility for nonlinearly distorted transcripts abundance estimations