033002 S MM 4.2 Bioinformatik II VO

Hubert Hackl
Institute of Bioinformatics, Biocenter
Medical University of Innsbruck,
Innrain 80, CCB, 6020 Innsbruck, Austria
+43-512-9003-71403
hubert.hackl@i-med.ac.at
http://icbi.at

Organization

1. VO Hubert Hackl

Mo	02 Mar	[13:00-15:45]	M.01.470
Di	03 Mar	[13:00-14:45]	M.01.470
Mi	04 Mar	[13:00-14:45]	M.01.470
Do	05 Mar	[13:00-14:45]	M.01.470
Do	12 Mar	[13:00-15:45]	M.01.470
Fr	13 Mar	[13:00-15:45]	M.01.470

Schriftliche Prüfung (8 offene Fragen, zB.Hierarchisches Clustering)

2. UE Dietmar Rieder
   Computerraum C1 (FP301200), Protokolle
3. Unterlagen: https://icbi.at/courses/
4. Lernziele:
   - Schritte der genomweiten Genexpressionsanalyse
   - Methoden für die Interpretation von Gensignaturen
   - Biomolekulare Netzwerkanalyse
   - Identifizierung von regulatorischen Sequenzen
   - Online Tools/Softwarepakete und Genom-Browser

1. Introduction
   - Gene regulation
   - Genomics and genome analyses

2. Bioinformatics tools and methods
   - Regulatory sequences and motif discovery
   - Transcription factor binding sites

3. Technologies
   - Microarrays
   - Deep sequencing and applications (RNAseq)

4. Clustering
   - Unsupervized clustering (HCA,K-means, SOM)
   - Dimension reduction (PCA)
   - Supervized clustering (classification)

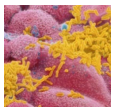
5. Gene ontology, Pathways, Enrichment analysis
   - Databases and tools
   - Gene set enrichment analysis

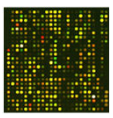
6. Biomolecular networks
   - Small world networks
   - Topology and parameter
   - Network motifs

Gene Regulation

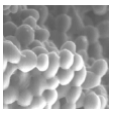
History

1995
- Two bacterial genomes decoded (TIGR)
  Mycoplasma genitalium (580.070 bp)
  Haemophilus influenza (1,830.137 bp, 1.740 genes)
- First DNA microarray studies published





1996
- *Saccharomyces cerevisiae* (bakers yeast) decoded (12,000.000 bp, 6.000 genes)



1998
- *Caenorhabditis elegans (*worm*)* genome decoded (97,000.000bp, 19.000 genes)

2000
- Genome of *Drosophila melanogaster* (fruit fly) (180,000.000bp, 14.000 genes)

Human genome project

2000
- Draft version of the human genome (>10 years, >3 billion $, 2500 scientists in 20 labs*)*

2003
- completed (high quality reference sequence) (3,000,000.000bp, 25.000 genes)

2007
- J Craig Venter genome sequence
- James Watson genome sequence (2 months, 454 sequencing, 1 million $)

2012
- > 150 eukaryotic genomes sequenced
- > 20 mammals
- > 10k of sequenced bacteria and viruses

Large scale genomics projects

1000 Genomes Project
- Study human genetic variation of >1.000 human genomes

UK10K
- Project to identify rare genetic events by studying the whole genome sequences of > 10.000 people (Wellcome Trust).

Genome10k
- Whole genome sequencing of 10.000 vertebrates

Human pangenome reference (Lia et al. Nature 2023)

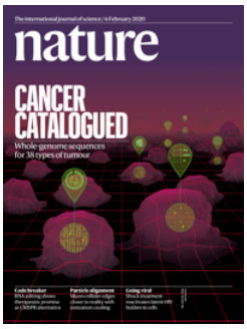
International Cancer Genome Consortium (ICGC) and
The Cancer Genome Atlas (TCGA)
- To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes.

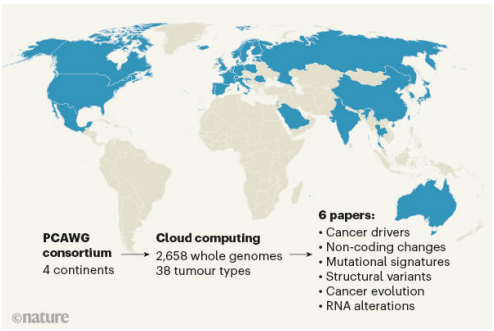
Human Tumor Atlas Network (HTAN)

Human Cell Atlas (Single cell and spatial analyses)

Pan-Cancer Analysis of Whole Genomes Consortium

>2600 whole cancer genomes
38 tumor types
750 affiliations

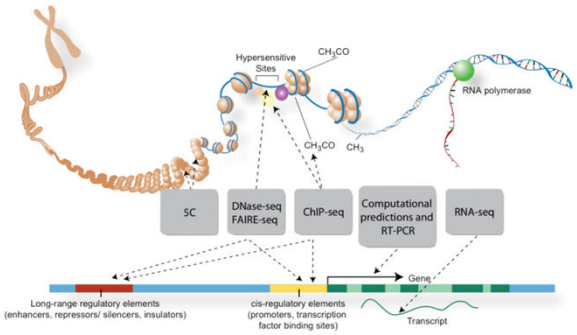


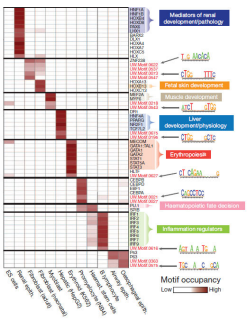


Feb 2020

ENCODE (Encyclopedia of DNA Elements)

32 institutes, 442 members, 1640 datasets, 30 publications (2012)





The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.

We have the genome sequence, so do we know everything?

No

The genome (transcriptome) is dynamic, the activity of the genes is changing over time and according to the environment or signals.

How is this regulated?

- Gene regulation in prokaryotes
- Gene regulation in eukaryotes

Gene regulation in prokaryotes

Prokaryotic transcriptional regulation

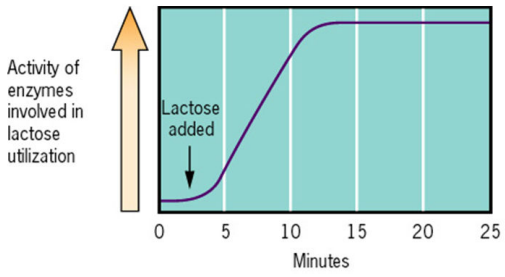
1. Lead to rapid increases and decreases in the expression of genes in response to environmental stimuli
   - Plasticity to respond to ever changing environment
2. Those that involve pre-programmed or cascades of gene expression
   - Set A → Set B → Set C……
   - Usually expressed in order

Response to environmental stimuli

- Gene expression (protein production) energetically expensive
- Extensive and sophisticated systems to regulate gene expression to conserve precious metabolic energy
- Transcriptional regulation has largest effect on phenotype

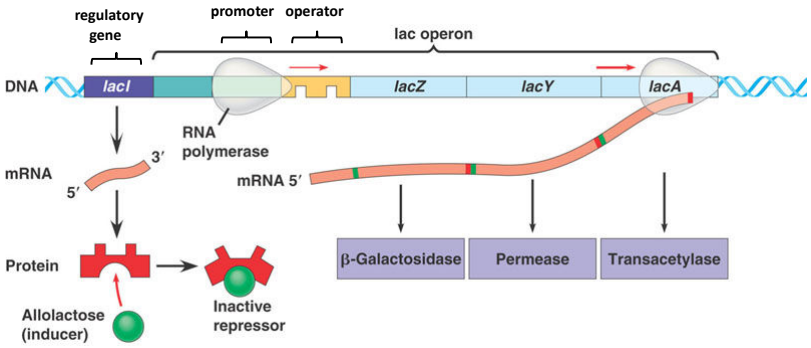
Example lack of glucose but abundance of lactose

- Turn on or induce expression of Lactose catabolism genes
- Induces transcription of gene for lactose utilization
- Catabolic (degradative) pathways often are inducible



Prokaryotic transcriptional regulation

lac operon as example for inducible system (*E. coli*)

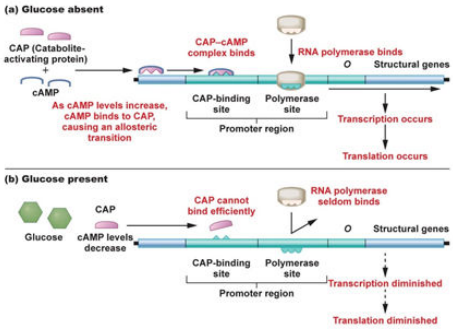


- If lactose is not present (resting state) repressor binding to promoter prevents binding of polymerase => **no** mRNA expression
- If lactose is present repressor is inactivated by conformational changes => mRNA expression of structural genes

Prokaryotic transcriptional regulation

Glucose and the lac operon

- Lactose is metabolised into glucose so what happens if glucose is present.
- Catabolite-activation protein (CAP): CAP must be present to make RNA polymerase binding efficiently

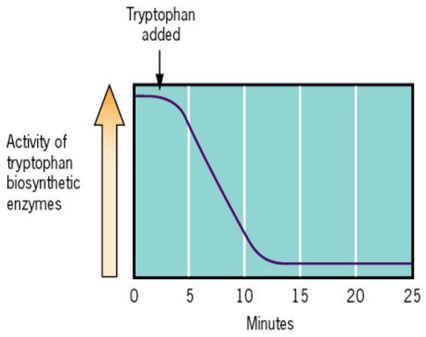


- In the presence of glucose the CAP is altered and prevents RNA polymerase binding to the promoter region and so prevents transcription.

Response to environmental stimuli

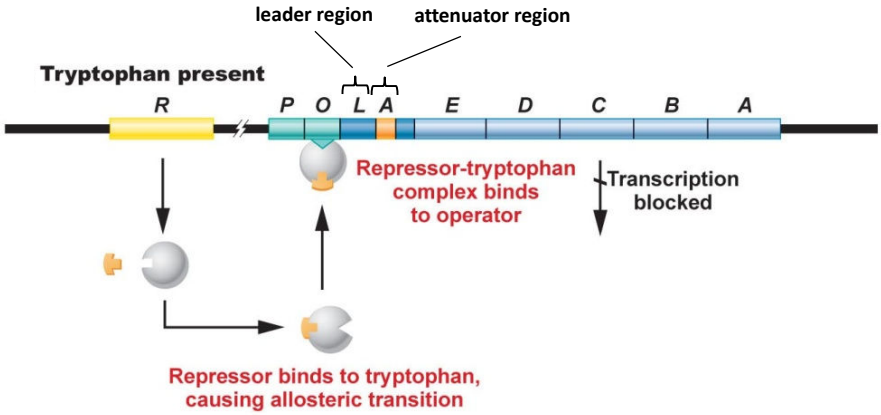
Example tryptophan (essential amino acid)

- *E.coli* can synthesize most molecules needed to growth (Amino acids, purines, pyrimidines, and vitamins)
- When Trp is present in the environment biosynthesis should be turned off
- Anabolic (biosynthetic) pathways often are repressible



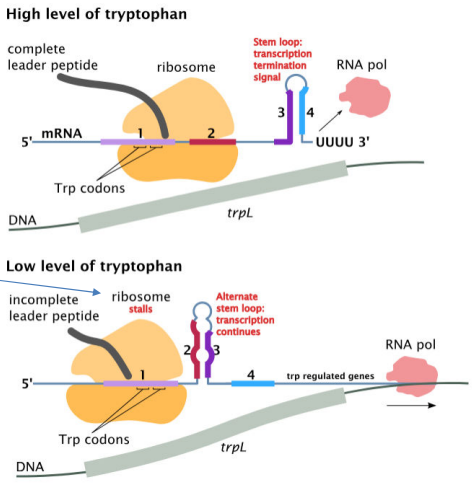
Prokaryotic transcriptional regulation

- *trp* operon as an example for a repressible system



Attenuator mechanism

Translation is directly coupled to transcription

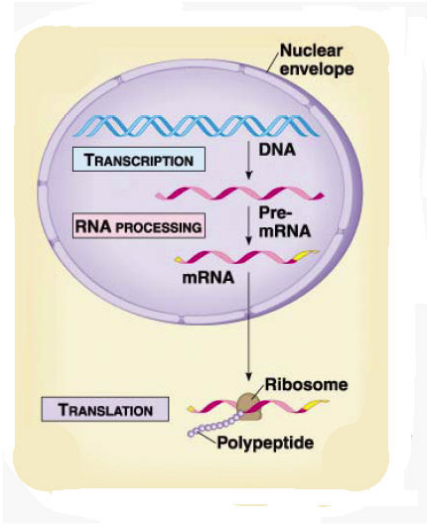


No Trp => takes time

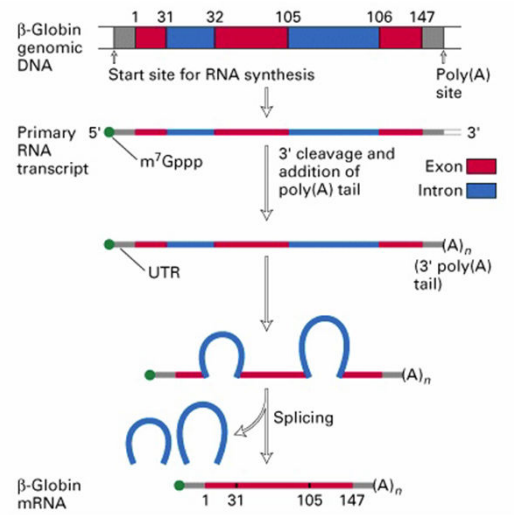
Gene regulation in eukaryotes

Gene expression in eukaryotes

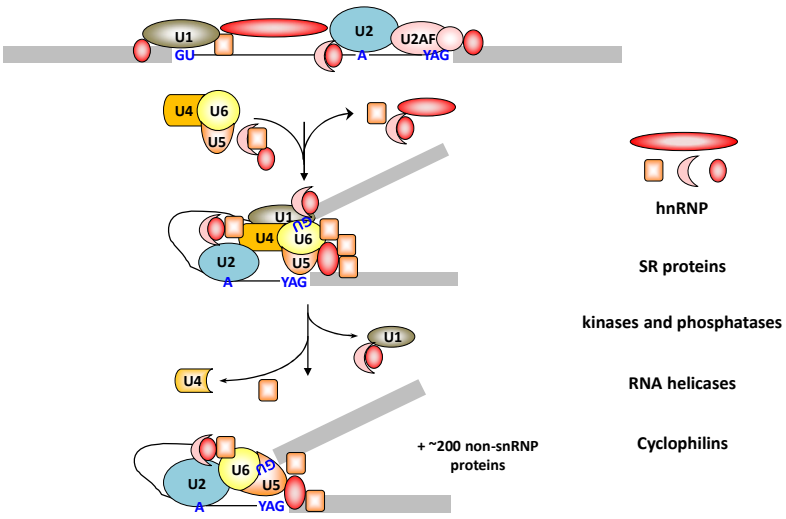
- Two cellular compartments:
  - Transcription in nucleus
  - Translation in cytoplasm
- RNA processing
  - 5´capping
  - RNA splicing
  - 3´polyadenylation



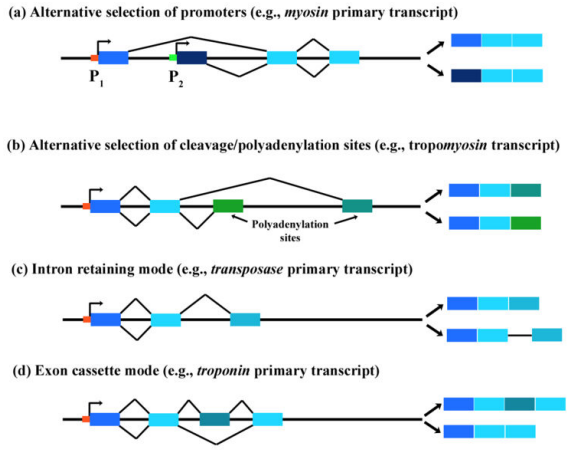
mRNA processing



Spliceosome assembly

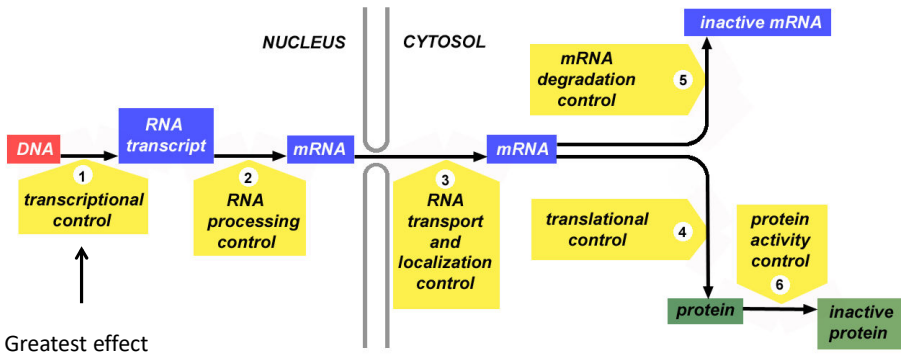


Alternative splicing

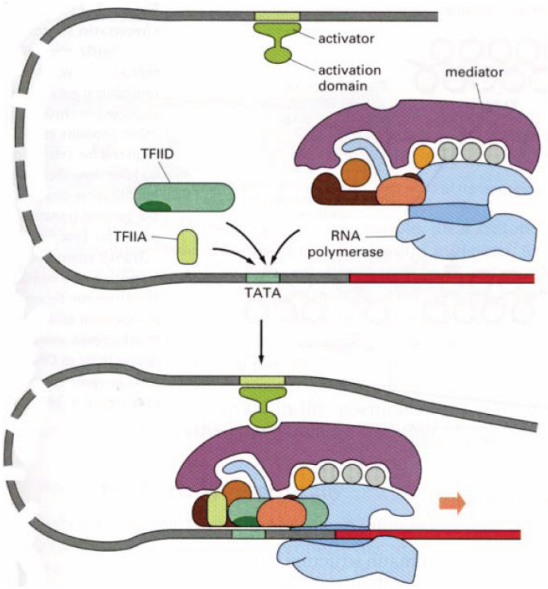


- Dependent on RNA/Spliceosome interaction
- Economizes on genetic information
- Create numerous related yet different proteins

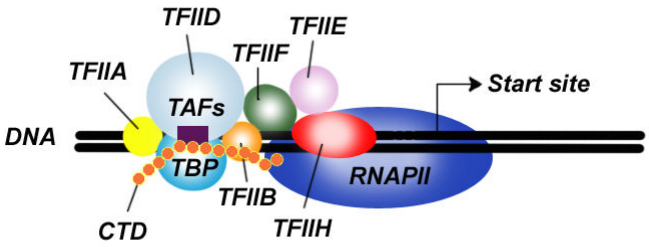
Different levels of regulation



Regulation of eukaryotic transcription



Basal transcription factors



Cis elements: sequences on DNA that affects the level of transcription.

Trans elements: DNA-binding proteins that change the level of transcription by basal transcription machinery.

Cis-regulatory elements of transcription

- Promoter (proximal regulation elements)

  Region that is located immediately upstream of a protein-coding gene and binds to RNA polymerase II; where transcription is initiated; (TATA box) (H3K4me3)

- Enhancers (distal regulation elements)

  Eukaryotic DNA sequences that are necessary to activate gene transcription (p300, H3K4me1)

- LCR (locus control region)

  Super-enhancer sequences in eukaryotic cells that control the expression of distant gene families (e.g. beta-globin)

- Insulators

  Separates active from inactive chromatin domains and interferes with enhancer activity when placed between an enhancer and a promoter (CTCF)

Properties of enhancers and promoters

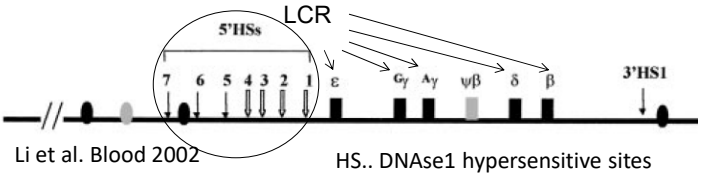
1. Act over large distances
   - >= kbp
2. Act independent of orientation
   - Normal or inverted
3. Effects are independent of position
   - Upstream, downstream, intronic

- Promoters
  - Immediately upstream
  - Function in only one orientation

30

Locus Control Regions (LCR)

- Example β-globin locus (5 genes in human)

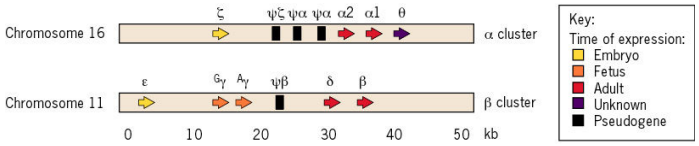


Li et al. Blood 2002

HS.. DNAse1 hypersensitive sites

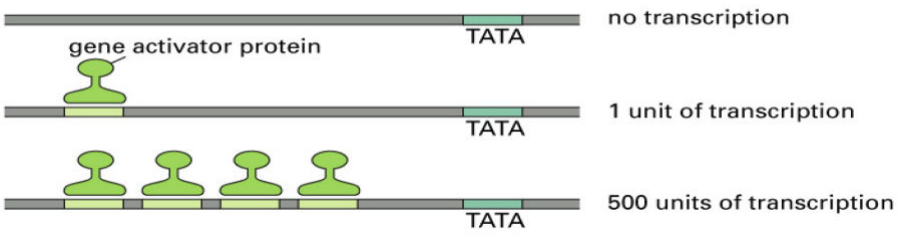
  - strong, transcription-enhancing activity
  - establishment and maintenance of an open chromatin domain

- **Temporal regulation** of hemoglobin (tetramer 2xα +2xβ)

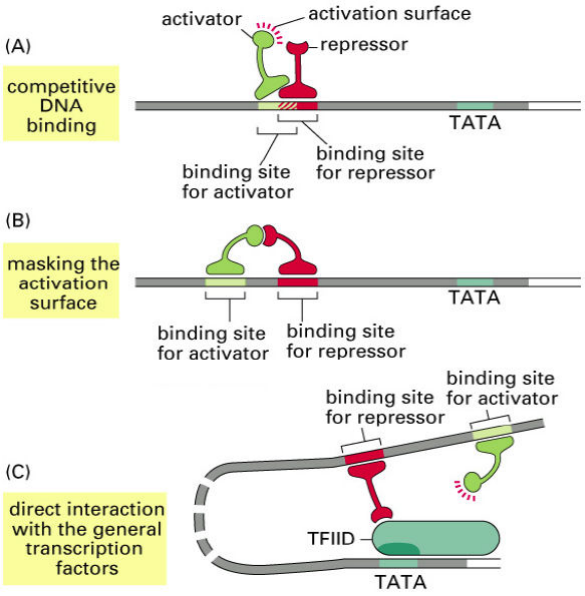


31

Transcriptional synergy

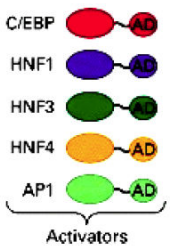


Eukaryotic gene repressors

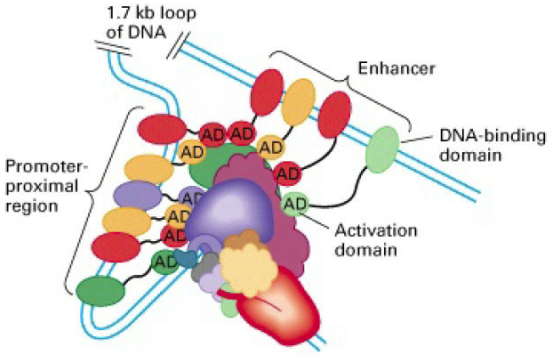


Transcription factor combinations

Most genes are regulated by multiple transcription factors



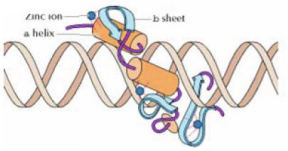
Liver cell-specific expression of the *TTR* gene

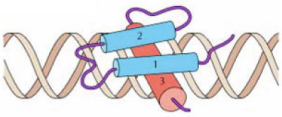


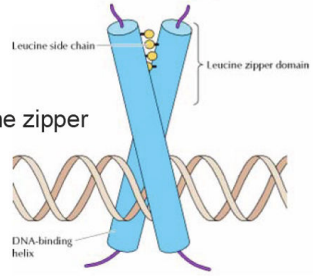
Classification of TF by DNA binding

A. Zinc fingers

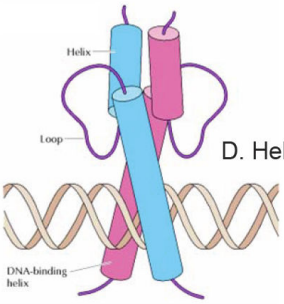
B. Helix-turn-helix







C. Leucine zipper

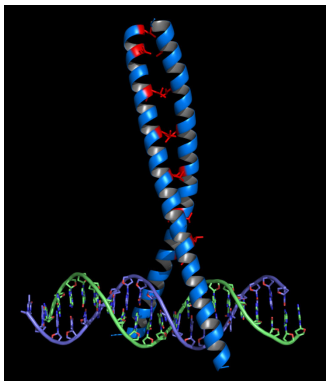


D. Helix-loop-helix

http://www.gene-regulation.com/pub/databases/transfac/cl.html

Transcription factor dimerization

Leucine zippers



- homo dimerization
- hetero dimerization

every 7th position
Leucine

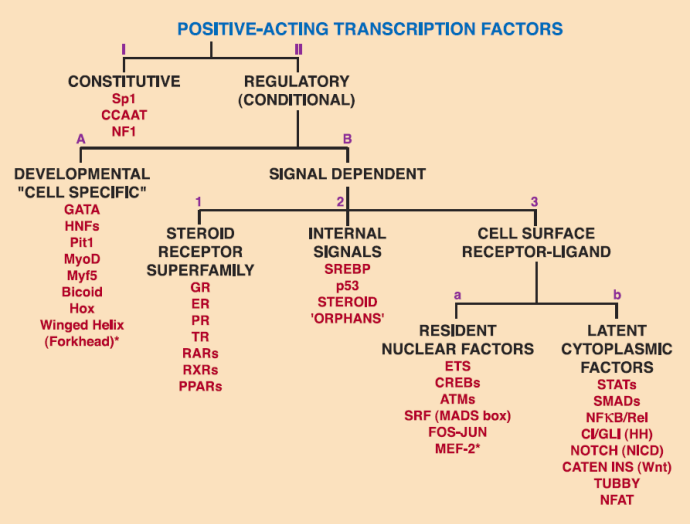
Signaling

Induction of transcription by environmental factors are less common in eukaryotes

Intercellular communication mediated by hormones

- Steroid Hormones
  - cholesterol derivatives
  - Easy pass through cell membrane
  - Ex. Estrogen, progesterone, testosterone, glucocorticoids, ecdysone
- Peptide Hormones
  - Peptides
  - Don't pass through membrane
  - Ex. Insulin, growth hormone, prolactin
- Other non-hormone proteins
  - Nerve growth factor
  - Epidermal growth factor

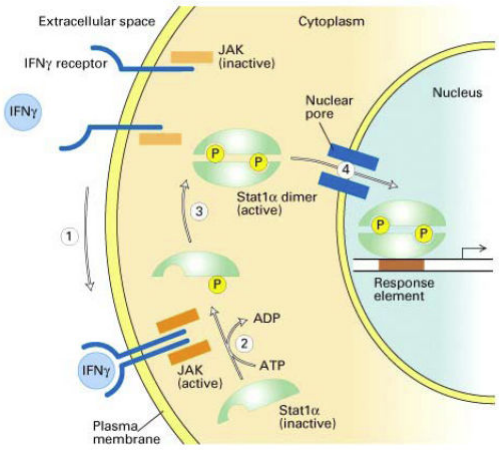
Classification of TF by function



Brivanlou AH, Darnell Jr JE. Science. 295: 813-818 (2002)

Regulation by phosphorylation

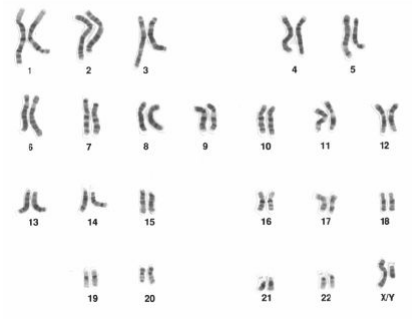
- Hormone activates kinase
- Kinase phosphorylates transcription factor
- Transcription factor is activated

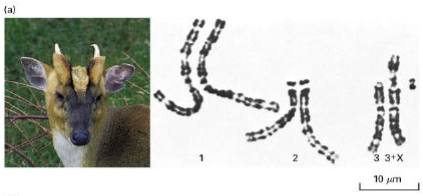


Principles of TF regulation

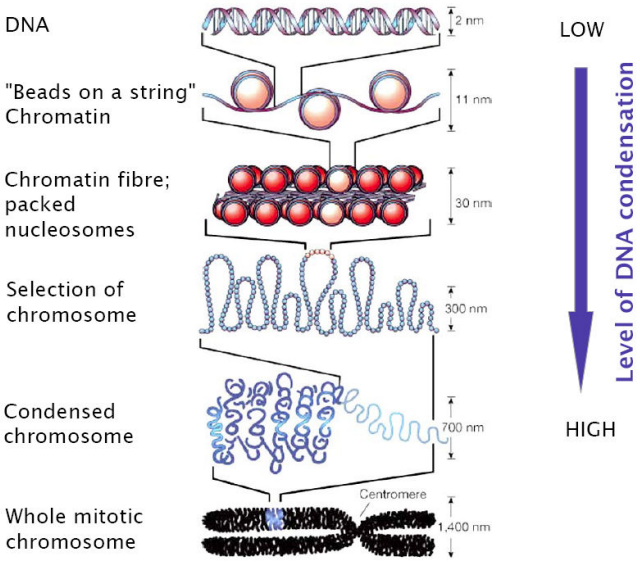
- 1 TF can target promoter of many genes
- >1 TF regulate expression of 1 gene (modules)
- Cascade of TF possible
- Positive feedback loop
- Feed forward loop

Organization into chromosomes

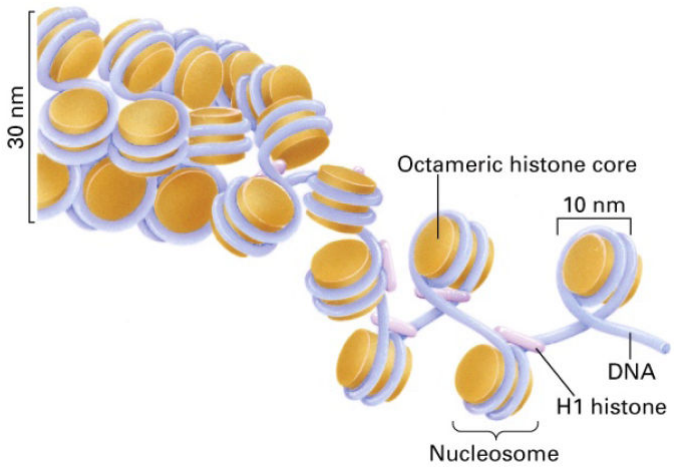




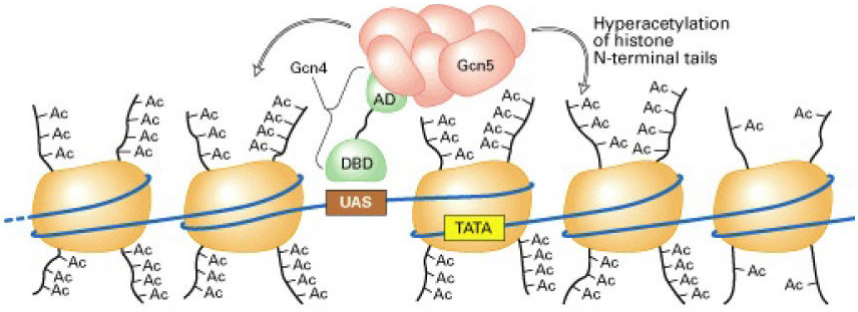
DNA packing



The solenoid model of condensed chromatin

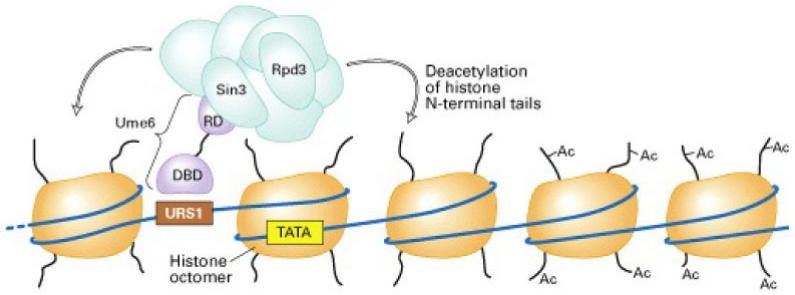


Activators: histone acetylation



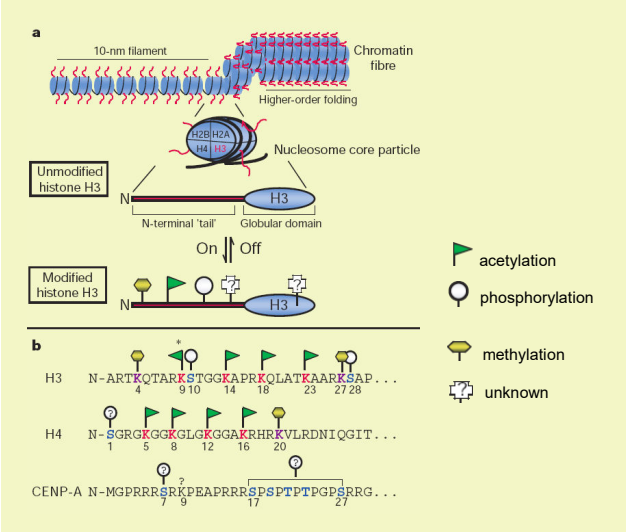
- Some activators recruit histone acetylase, which adds acetyl groups to histones
- Allows transcriptional machinery access to less condensed template DNA (euchromatin)

Repressors: histone deacetylation



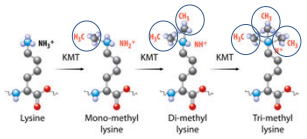
- Some repressors recruit histone deacetylase, which removes acetyl groups from histones
- Prevents transcriptional machinery access by condensing template DNA (heterochromatin)

Histone modification and histone code



Strahl BD, Allis CD. Nature 2000. 403:41-45

Chromatin states



State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage: Median (%)	Coverage: H1 ES (fold)	Coverage: GM (fold)	Median length (kb)	±2 kb TSS (%)	Conserved non-exon	DNase (K562)	c-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV

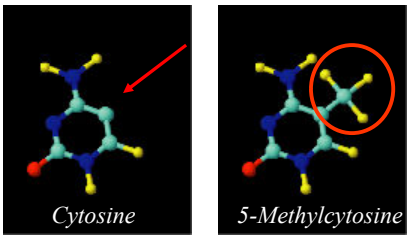
Chromatin states

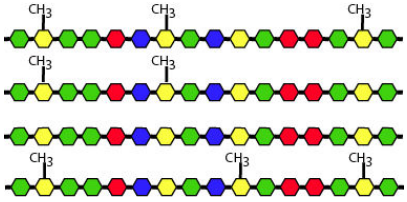
Chromatin mark observation frequency (%)

Functional enrichments (fold)

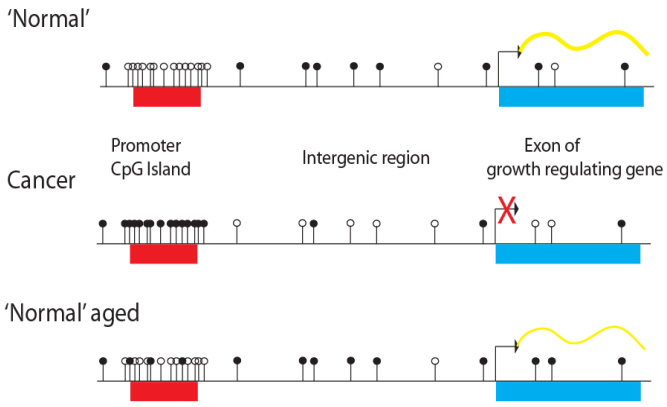
Ernst et al. Nature 2011.

DNA methylation

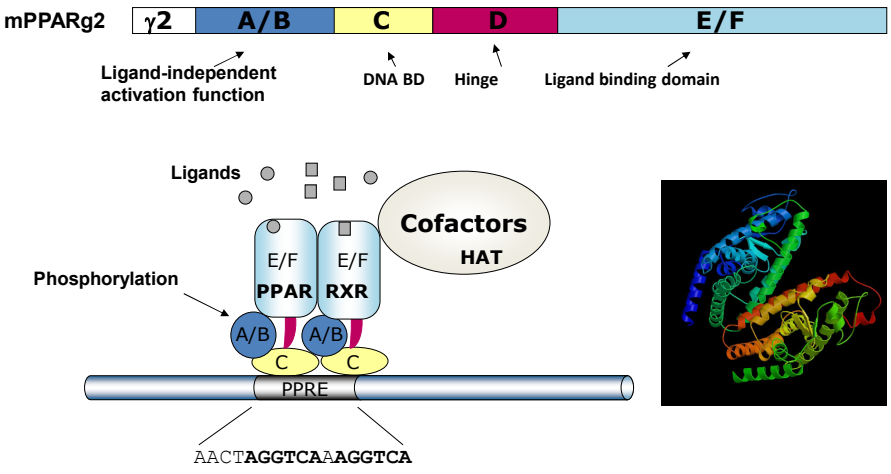




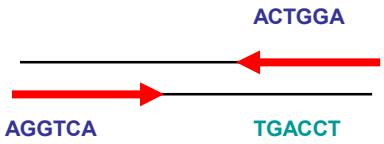
Aberrant methylation patterns



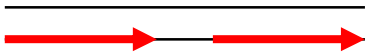
Nuclear receptors (=transcription factor+receptor)



Nuclear receptor response elements



PALINDROME
(INVERTED REPEAT)



DIRECT REPEAT

AGGTCAn...nAGGTCA

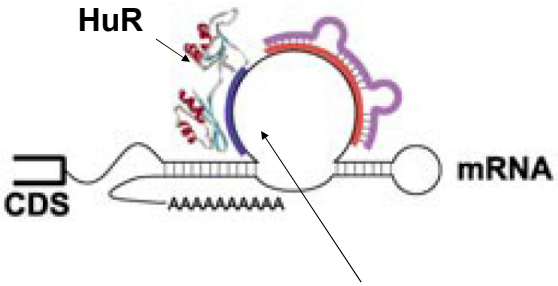
# of nucleotides

e.g.

AGGTCAnnnnnAGGTCA

DR-5

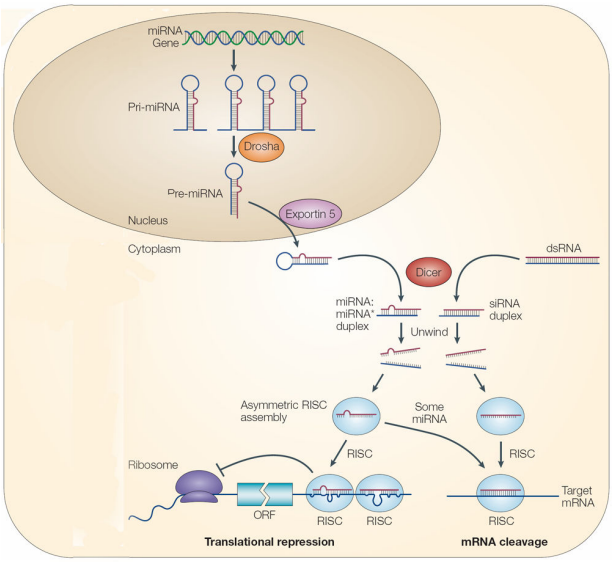
RNA binding proteins for mRNA stability



AU rich elements (ARE)

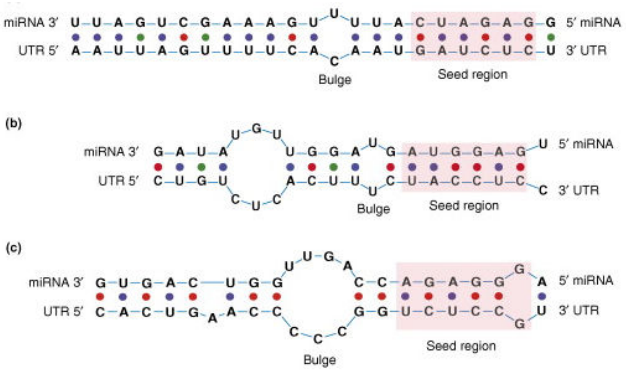
Cox-2	UAUUAAUUUAAUUAUUUAAUAAUAUUUAUAUUAAA
IL-1β	UAUUUAUUUAUUUAUUUGUUUGUUUGUUUUAUU
IL-2	UAUUUAUUUAAAUAUUUAAAUUUUAUAUUUAUU
IL-4	AUAUUUUAAUUUAUGAGUUUUUGAUAGCUUUAUUUUUUAAG
IL-8	UAUUUAUUAUUUAUGUAUUUAUUUAA
TNFα	AUUAUUUAUUAUUUAUUUAUUAUUUAUUUAUUUA

microRNA and siRNA (RNAi)

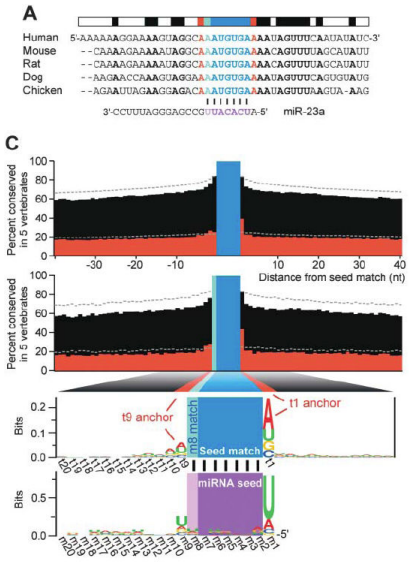


He L., Hannon GJ. Nature Reviews Genetics. 2004. 5:522-531

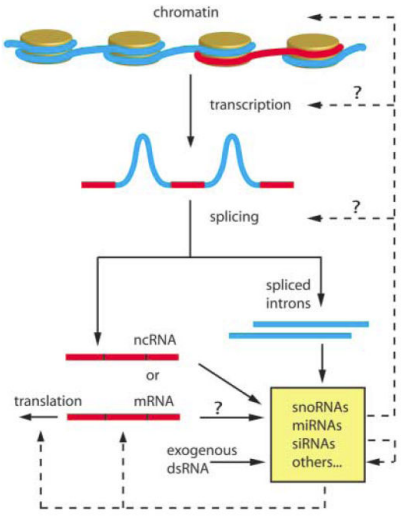
miRNA-mRNA targeting



Conservation of microRNA target sequences



Function of none coding RNA (ncRNA)



Genome analyses

Human Genome

2.95 Gbases of 3.2 Gbases is euchromatin

- >90% of euchromatin sequenced
- ~1% of sequence encodes protein sequences

23,000 genes

- Small # considering:
  - Yeast - 6,000 genes
  - *Drosophila* - 13,000 genes
  - *C. elegans* - 19,000 genes
  - *A. thaliana* - 26,000 genes

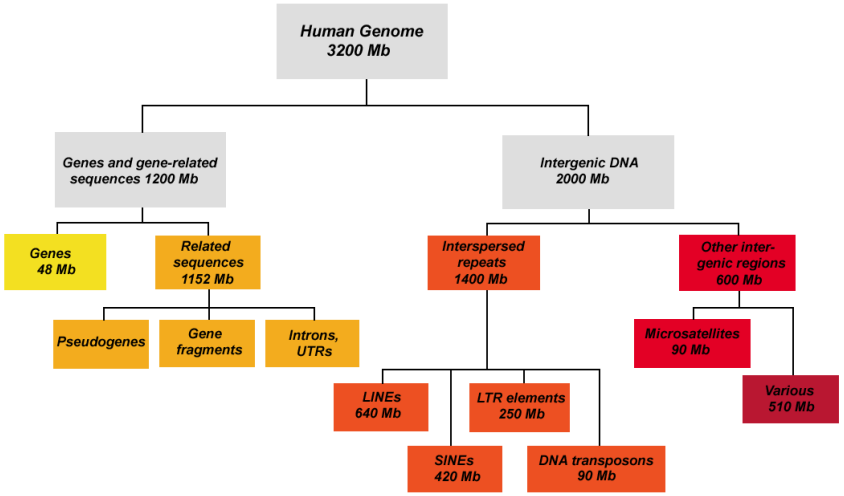
Genes

- Exons 1.1 %
- Introns ~24%
- Intergenic 74%

- Average gene size – ~70 kb
- Average # of introns - 10
- Only 94 of 1,278 protein families are specific to vertebrates

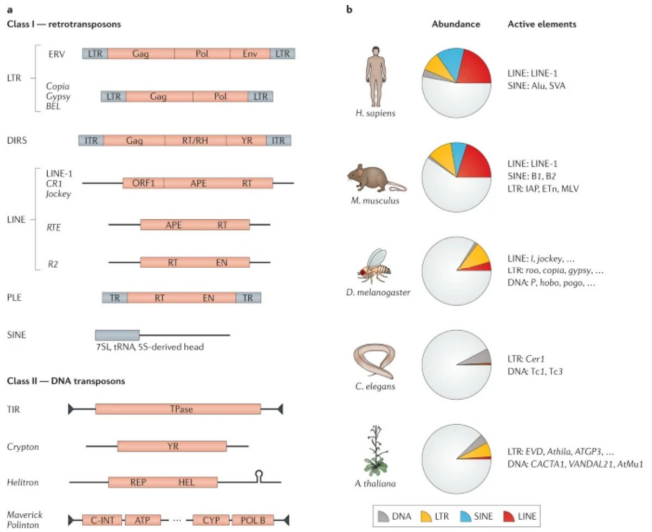
Genes involved in basic biochemical processes seem to evolve just once and have stayed fixed from bacteria to yeast to mammals

Organization of the human genome



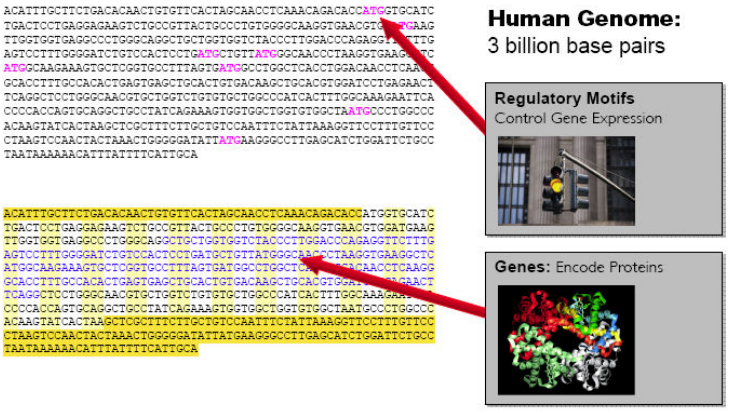
E.g. > 1 million copies of Alu-repeats

Transposons



Deniz et al. Nat Rev Genet. 2019

Extracting signal from noise



Bioinformatics challenges in genome analysis

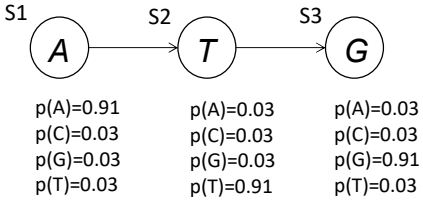
- Gene finding
- Start codon
- Exon-intron borders
- CpG-islands
- Repetitive sequences (Repeat Masker)
- Regulatory sequences

Solution: **Hidden Markov Models (HMM)**

Markov chains

Markov chains: a sequence of events that occur one after another. The main restriction on a Markov chain is that the probability assigned to an event at any location in the chain can depend on only a fixed number of previous events.

Scoring sequences (e.g. start codon *ATG*)
3 states (S1, S2, S3), p(A)=p(C)=p(G)=p(T)=0.25

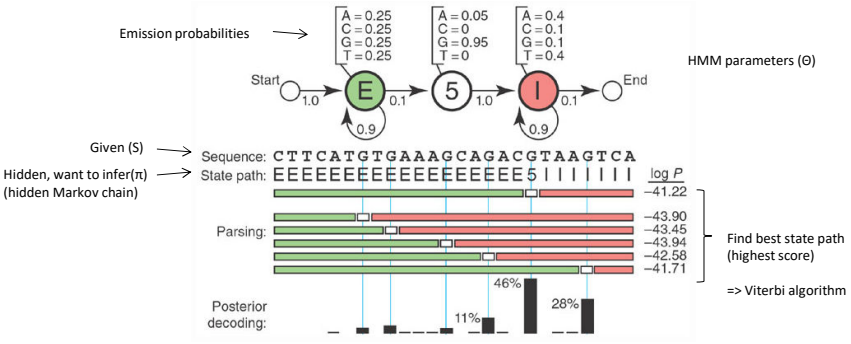


Markov chain 0th order
p(*ATG*)=0.91^3=0.752

Markov chain 1th order
p(*ATG*)=p(*A*)*p(*T*|*A*)*p(*G*|*T*)

Hidden Markov Model (HMM)

- Example exon-intron border
- 3 states: exon(E), 5'SS (5), intron (I)



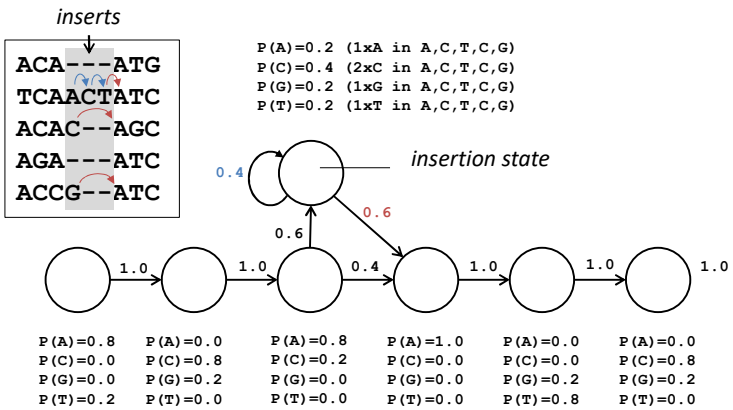
G T A A G T C A

$\log P(S,\pi|HMM,\Theta)=\log(1*0.25^{18}*0.9^{17}*0.1*0.95*1.0*0.4*0.9*0.4*0.9*0.4*0.9*0.1*0.9*0.4*0.9*0.1*0.9*0.4*0.1)$

Eddy SR, Nat Biotech 2004

Profile Hidden Markov Model

- For multiple alignments (e.g. DNA sequences)



p(*ACACATC*)=0.8*1*0.8*1*0.8*0.6 *0.4*0.6*1*1*0.8*1*0.8=0.047
log-odds=log(p(S)/0.25^L)=log(0.047/0.25^7)