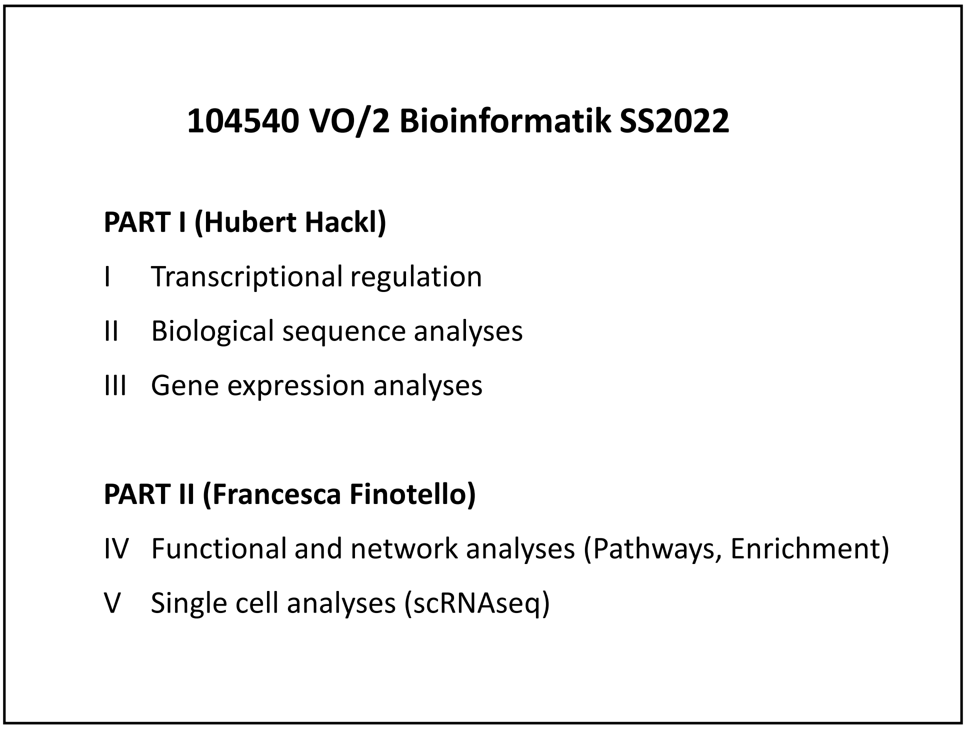


104540 VO/2 Bioinformatik SS2022

PART I (Hubert Hackl)

I Transcriptional regulation

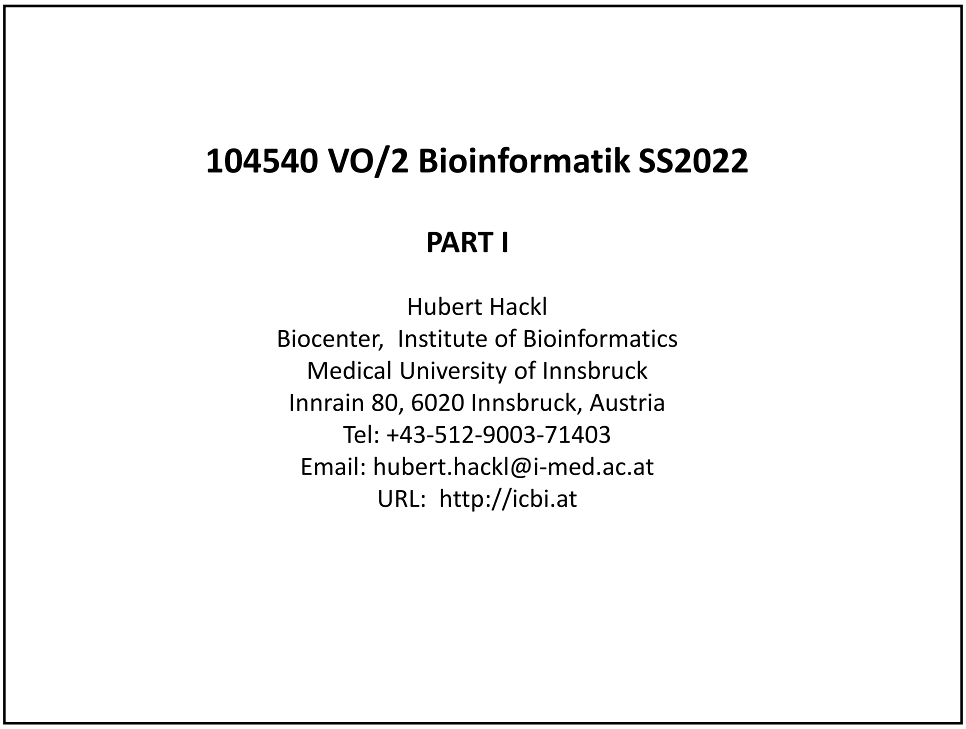
II Biological sequence analyses

III Gene expression analyses

PART II (Francesca Finotello)

IV Functional and network analyses (Pathways, Enrichment)

V Single cell analyses (scRNAseq)



104540 VO/2 Bioinformatik SS2022

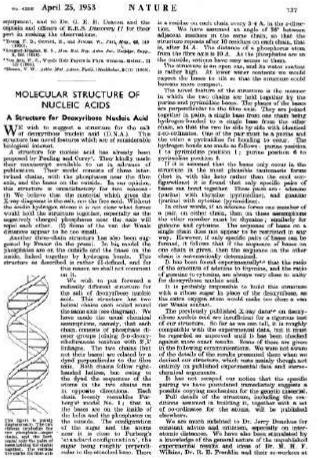
PART I

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

I Transcriptional regulation

- Introduction
- Gene Regulation
  - Prokaryotes
  - Eukaryotes
- Genome analysis
  - Hidden Markov Models

History



No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, 40, 149 (1920).

[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, 5, 285 (1949).

[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., 11 (3) (1950).

[4] Ekman, V. W., *Arkiv Mat. Astron. Fysik. (Stockholm)*, 2 (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

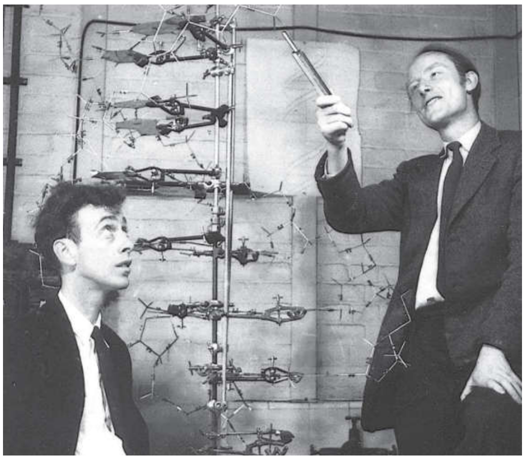
It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

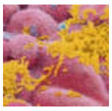
Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

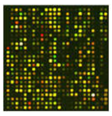
We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at



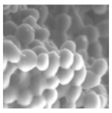
History

- **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 $, 20 labs*)*

2003

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

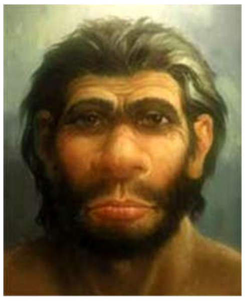
2007

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

2012

- >150 eukaryotic genomes sequenced
- > 20 mammals
- Hundreds of sequenced bacteria and viruses

Neandertal genome sequence



- Department of Evolutionary Genetics, Max-Planck Institute for Evolutionary Anthropology
- Draft sequence 2010 (Science) using 454 pyro-sequencing (Roche)
- Comparison with human and chimpanzee (e.g. speech-related gene FOXP2 with the same mutations as in human in contrast to chimp)
- Neanderthal admixture in modern human DNA?

Large scale genomics projects

1000 Genomes Project (=> 100.000 genomes project)

- Study human genetic variation of >1.000 human genomes

Genome10k

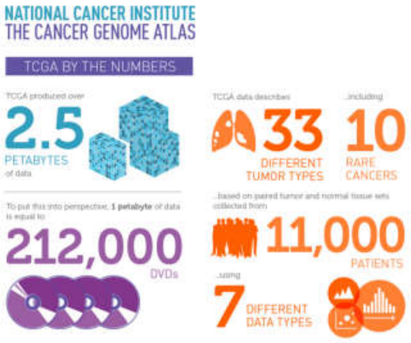
- whole genome sequencing of 10.000 vertebrates

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.

TCGA (The Cancer Genome Atlas)

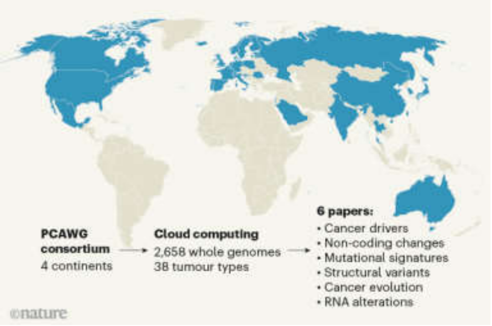
https://tcga-data.nci.nih.gov



- Copy number
- Methylation
- Gene expression
- MicroRNA expression
- Somatic mutations
- Clinical data

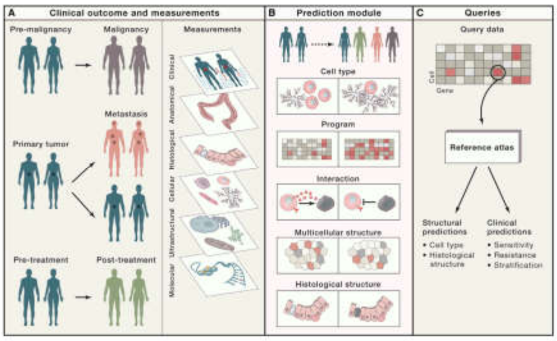
Pan-Cancer Analysis of Whole Genomes Consortium

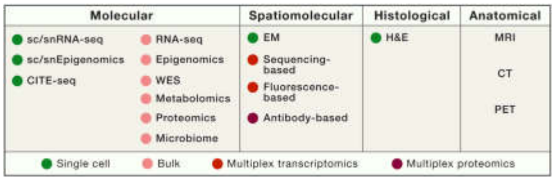
>2600 whole cancer genomes
38 tumor types
750 affiliations



Feb 2020

HTAN
HUMAN TUMOR ATLAS NETWORK





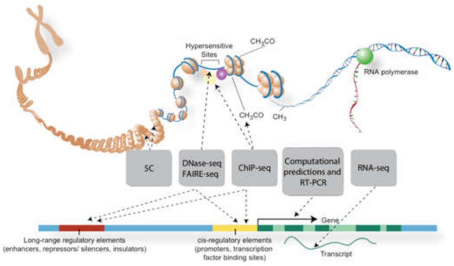
Molecular		Spatiomolecular	Histological	Anatomical
🟢 sc/snRNA-seq	🩷 RNA-seq	🟢 EM	🟢 H&E	MRI
🟢 sc/snEpigenomics	🩷 Epigenomics	🔴 Sequencing-based		
🟢 CITE-seq	🩷 WES	🔴 Fluorescence-based		CT
	🩷 Metabolomics			
	🩷 Proteomics	🟣 Antibody-based		PET
	🩷 Microbiome			

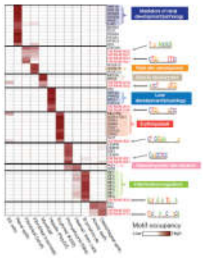
🟢 Single cell 🩷 Bulk 🔴 Multiplex transcriptomics 🟣 Multiplex proteomics

Johnson et al. Cell 2020

ENCODE (Encyclopedia of DNA Elements)

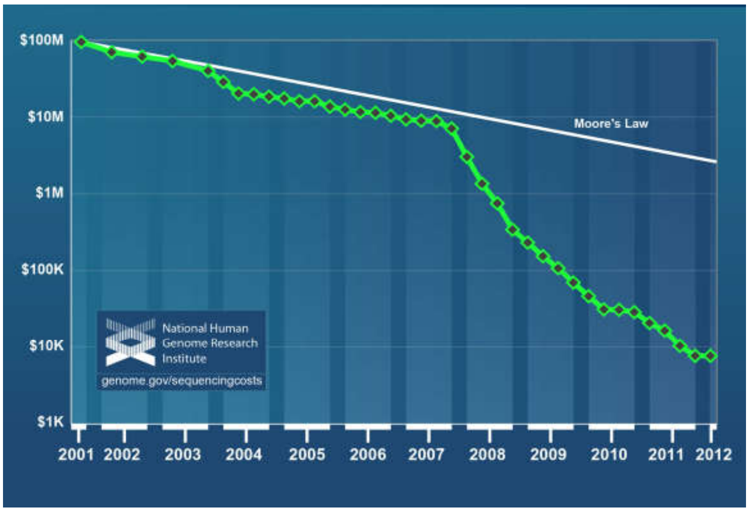
32 institutes	http://www.nature.com/encode
442 consortium members	http://genome.ucsc.edu/ENCODE/
1640 data sets	http://www.genome.gov/10005107
30 papers (Sept 2012)	


Hypersensitive Sites
CH_3CO
CH_3CO
CH_3
RNA polymerase
SC
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

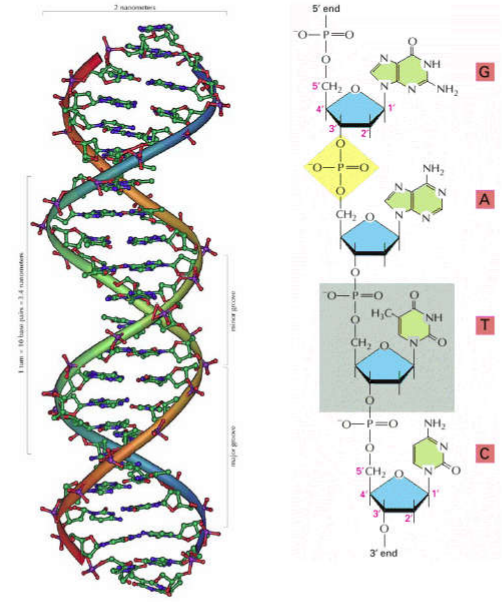


The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.

Cost per genome



DNA



Guanine (binds to C)

Adenine (binds to T)

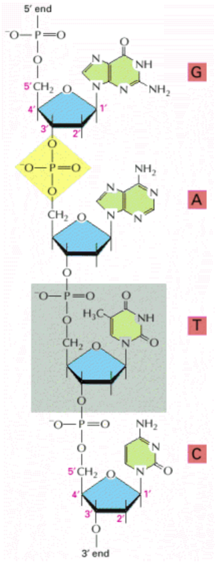
purine

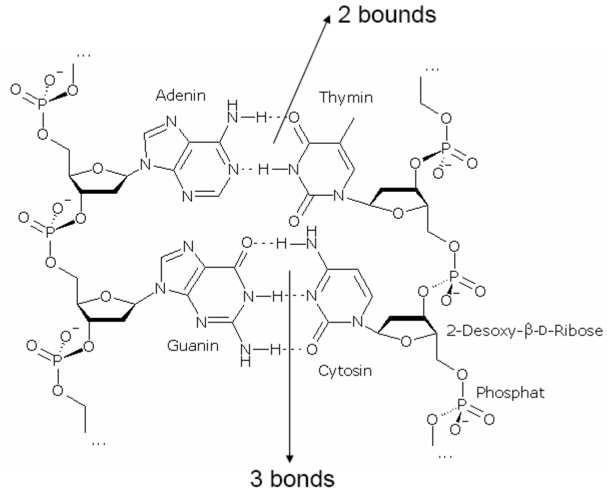
Thymine (binds to A)

Cytosine (binds to G)

pyrimidine

DNA

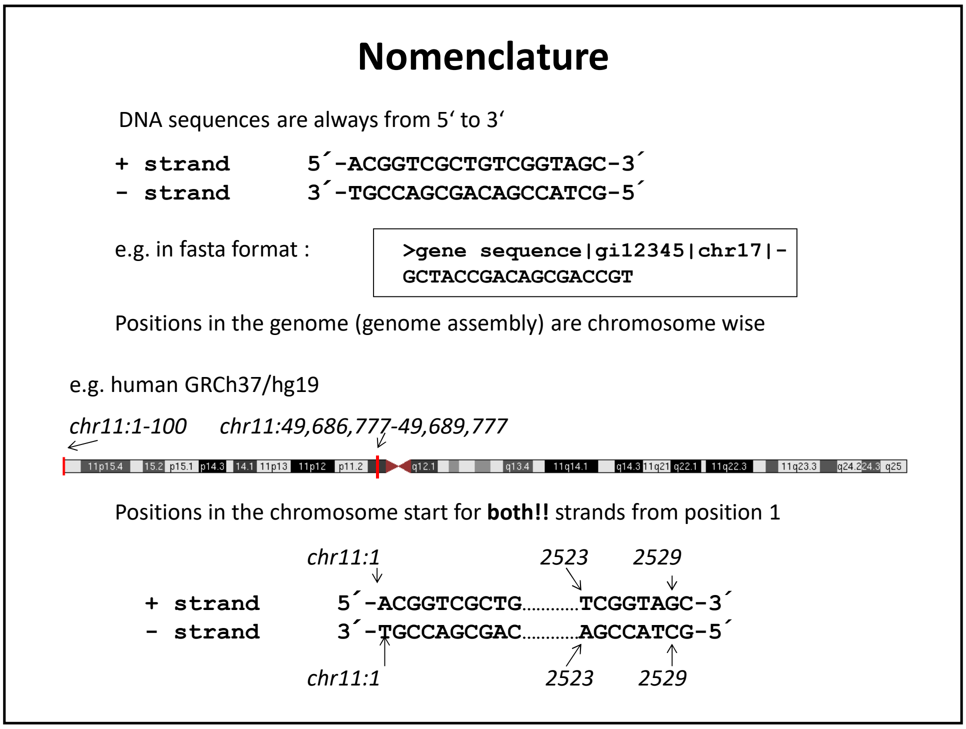




Nomenclature of nucleic acids

Base	Symbol	Occurrence
Adenin	A	DNA, RNA
Guanin	G	DNA, RNA
Cytosin	C	DNA, RNA
Thymin	T	DNA
Uracil	U	RNA

Symbol	Meaning	Description
R	A or G	pu**R**ine
Y	C or T	p**Y**rimidine
W	A or T	**W**eak hydrogen bonds
S	G or C	**S**trong hydrogen bonds
M	A or C	a**M**ino groups
K	G or T	**K**eto groups
H	A, C, or T (U)	not G, (**H** follows G)
B	G, C, or T (U)	not A, (**B** follows A)
V	G, A, or C	not T (U), (**V** follows U)
D	G, A, or T (U)	not C, (**D** follows C)
N	G, A, C or T (U)	a**N**y nucleotide



Nomenclature

DNA sequences are always from 5‘ to 3‘

```
+ strand      5´-ACGGTCGCTGTCGGTAGC-3´
- strand      3´-TGCCAGCGACAGCCATCG-5´
```

e.g. in fasta format :

```
>gene sequence|gi12345|chr17|-
GCTACCGACAGCGACCGT
```

Positions in the genome (genome assembly) are chromosome wise

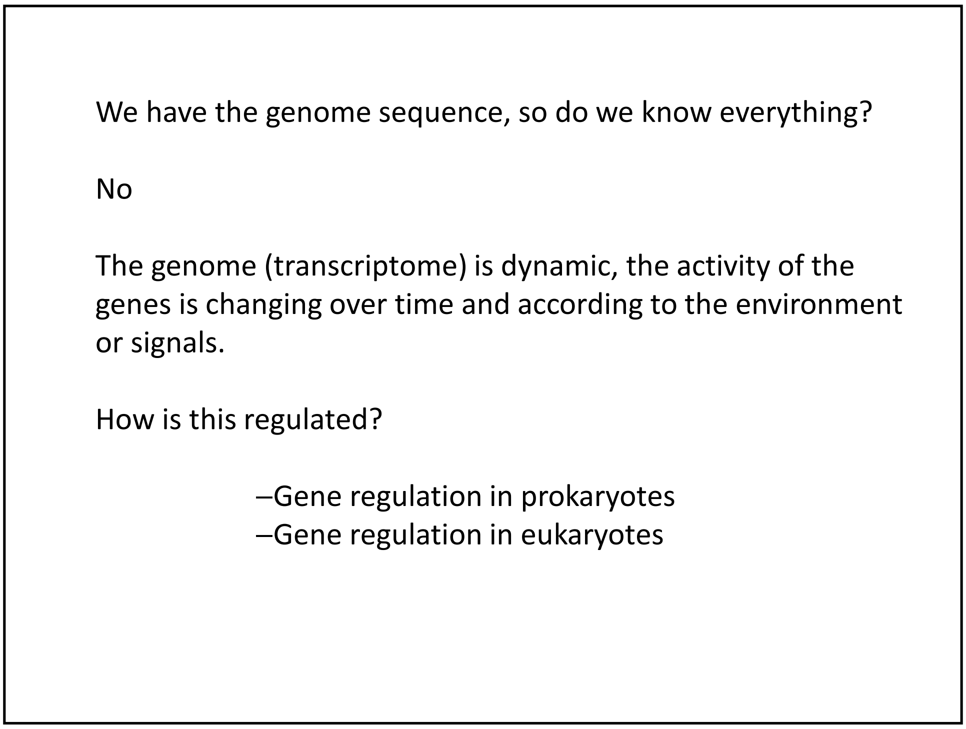
e.g. human GRCh37/hg19

chr11:1-100 *chr11:49,686,777-49,689,777*

11p15.4
15.2
p15.1
p14.3
14.1
11p13
11p12
p11.2
q12.1
q13.4
11q14.1
q14.3
11q21
q22.1
11q22.3
11q23.3
q24.2
24.3
q25

Positions in the chromosome start for **both!!** strands from position 1

```
                chr11:1                  2523      2529
+ strand      5´-ACGGTCGCTG............TCGGTAGC-3´
- strand      3´-TGCCAGCGAC............AGCCATCG-5´
                chr11:1                   2523      2529
```



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

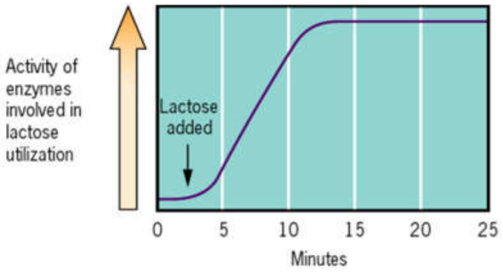
   – Set A → Set B → Set C……
   – Usually expressed in order

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

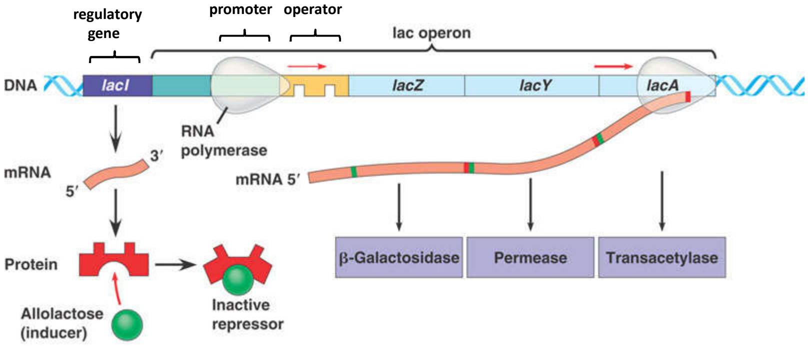
Example lack of glucose but abundance of lactose

- 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

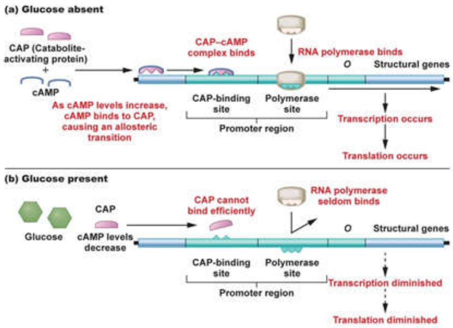
- *lac* operon as example for inducible system (*E. coli*)



  - 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

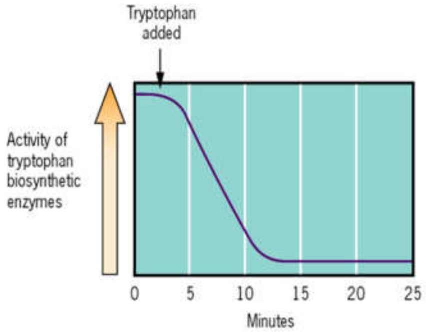
- Glucose and the lac operon
  - Lactose is metabolised into glucose so what happens if glucose is present.
  - Catabolite-activation protein (CAP): CAP must be present to make RNA polymerase binding efficiently



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

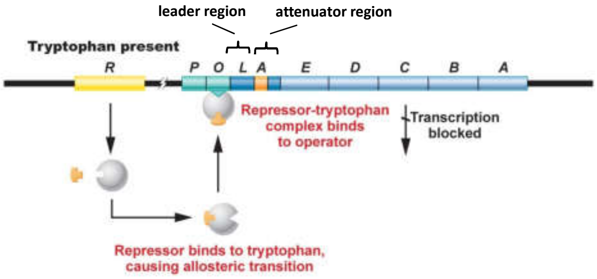
Response to environmental stimuli

- 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


leader region
attenuator region
Tryptophan present
R
P
O
L
A
E
D
C
B
A
Repressor-tryptophan complex binds to operator
Transcription blocked
Repressor binds to tryptophan, causing allosteric transition

- If tryptophan is present the repressor-tryptophan complex binds to operator => no mRNA expression of structural genes.
- Translation and transcription are coupled (regulation by leader sequence and attenuation)

Translational Control of Gene Expression

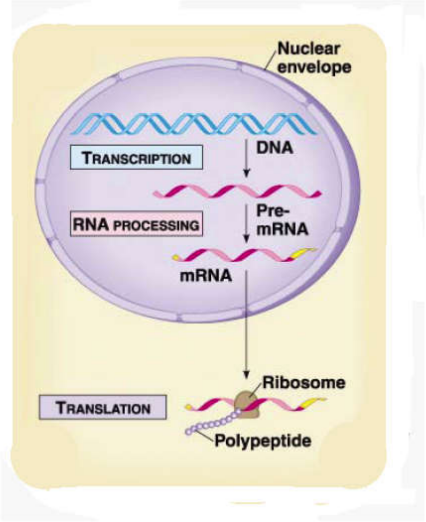
- Prokaryotes regulate at Transcription
- Translational control used for fine tuning
- Transcription, Translation, mRNA degradation are coupled
- Three general mechanisms

  1. Unequal efficiencies of translational initiation
  2. Altered efficiencies of ribosome movement
  3. Differential rates of mRNA degradation

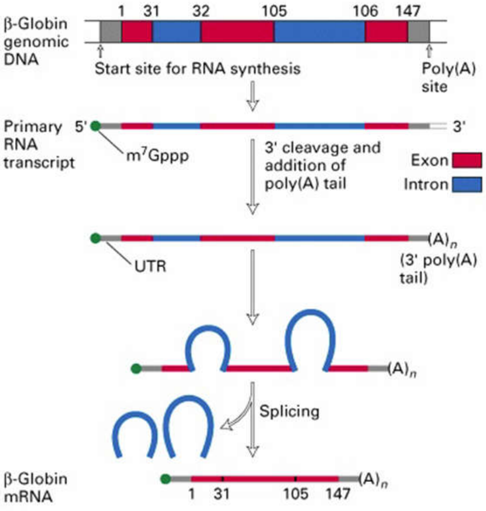
Gene regulation in eukaryotes

Gene expression in eukaryotes

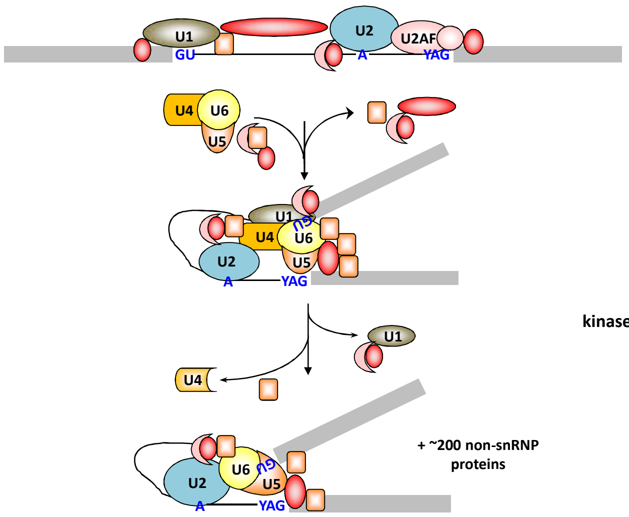
- Two cellular compartments:
  - Transcription in nucleus
  - Translation in cytoplasm
- RNA processing
  - 5´capping
  - RNA splicing
  - 3´polyadenylation



mRNA processing



Spliceosome assembly



+ ~200 non-snRNP proteins

hnRNP

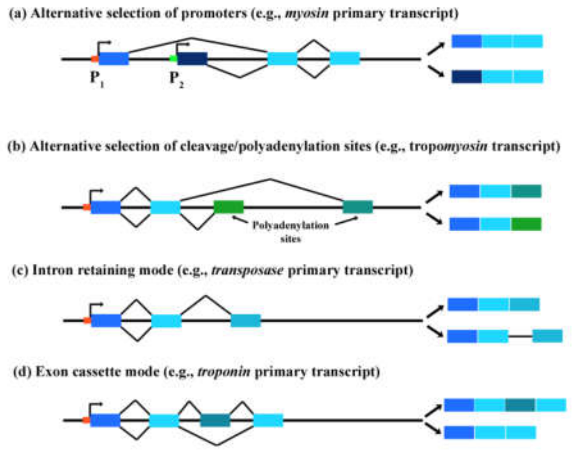
SR proteins

kinases and phosphatases

RNA helicases

Cyclophilins

Alternative splicing


(a) Alternative selection of promoters (e.g., *myosin* primary transcript)

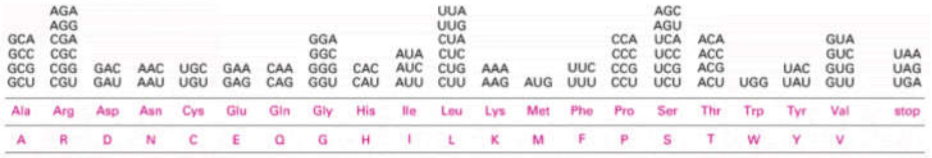
(b) Alternative selection of cleavage/polyadenylation sites (e.g., tropo*myosin* transcript)

(c) Intron retaining mode (e.g., *transposase* primary transcript)

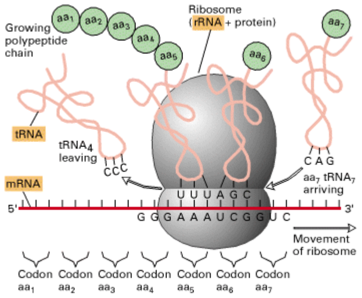
(d) Exon cassette mode (e.g., *troponin* primary transcript)

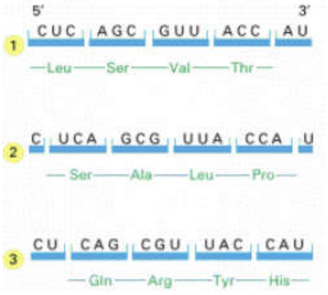
- Dependent on RNA/Spliceosome interaction
- Economizes on genetic information
- Create numerous related yet different proteins

Translation, genetic code and reading frames

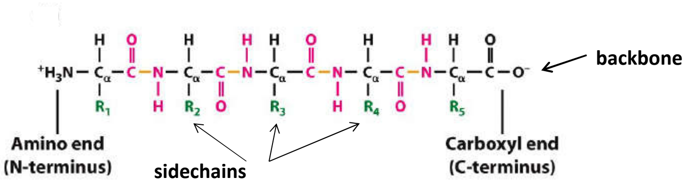


GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	





Peptid chain, amino acid sequence, proteins

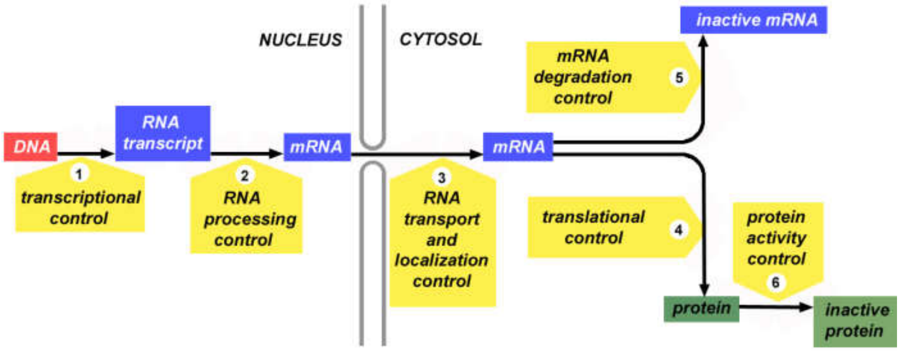


Protein sequences are always form N-terminal end to C-terminal end

E.g.. SCD sequence in fasta format

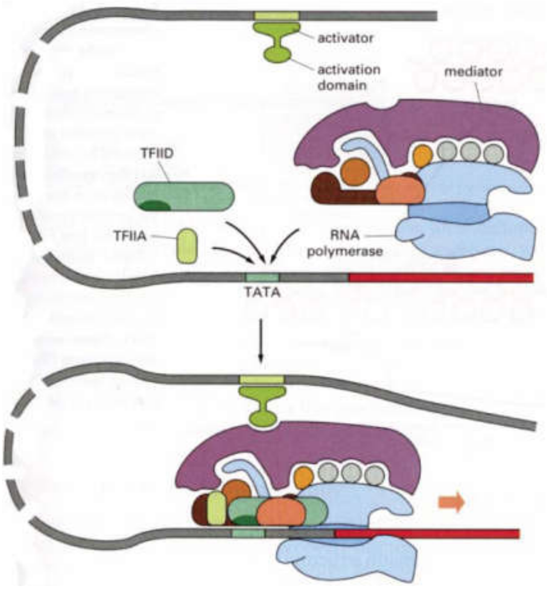
```
>gi|53759151|ref|NP_005054.3| acyl-CoA desaturase [Homo sapiens]
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVE
YVWRNIILMSLLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLII
ANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKL
VMFQRRYYKPGLLMMCFILPTLVPWYFWGETFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNI
SPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFTTFFIDCMAALGLAYDRKKVSKAAILARIKR
TGDGNYKSG
```

Different levels of regulation

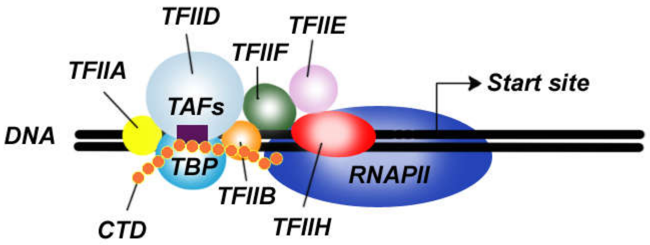


Transcriptional regulation has largest effect on phenotype!

Regulation of eukaryotic transcription



Basal transcription factors



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

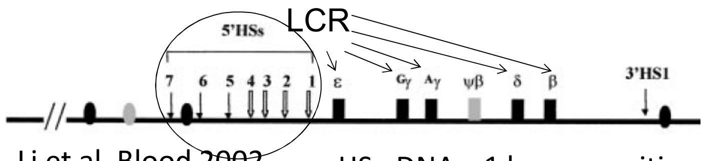
Trans elements: DNA-binding proteins that change the level of transcription by basal transcription machinery.

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)
- **LCR (locus control region)**
  Super-enhancer sequences in eukaryotic cells that control the expression of distant gene families (e.g. beta-globin)
- **Enhancers (distal regulation elements)**
  Eukaryotic DNA sequences that are necessary to activate gene transcription (p300, H3K4me1)
- **Insulators**
  Separates active from inactive chromatin domains and interferes with enhancer activity when placed between an enhancer and a promoter (CTCF)
- **Repressor/silencer**
  Negative regulators of gene expression (REST,SUZ12)

Locus Control Regions (LCR)

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

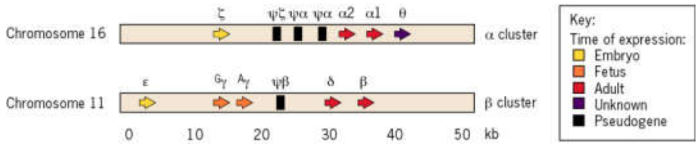


Li et al. Blood 2002

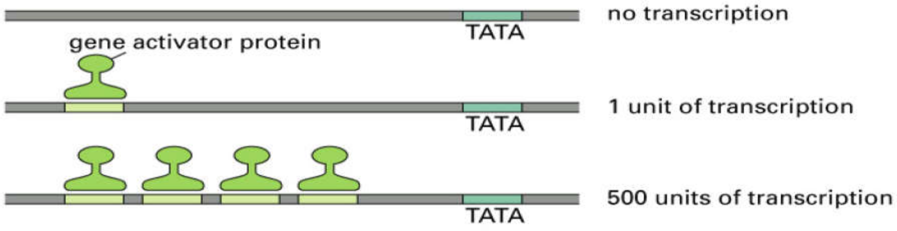
HS.. DNAse1 hypersensitive sites

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

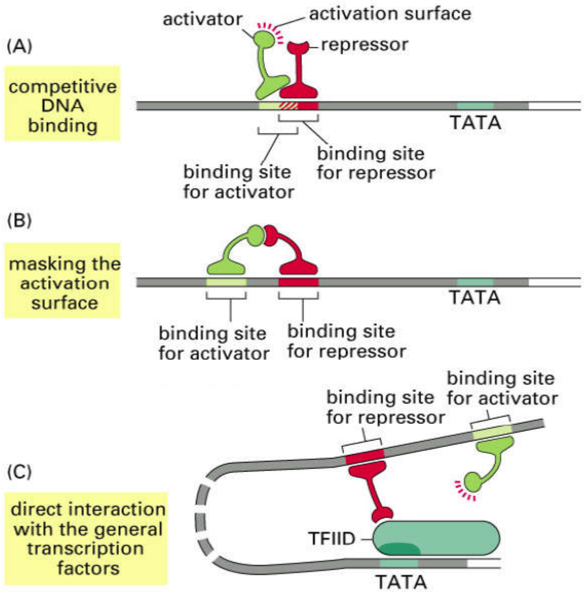
– Temporal regulation of hemoglobin (tetramer 2xα +2xβ)



Transcriptional synergy

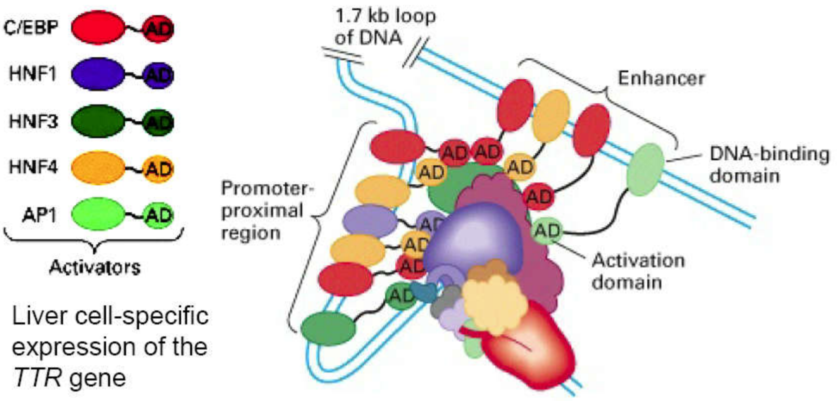


Eukaryotic gene repressors


(A)
competitive DNA binding
activator
activation surface
repressor
TATA
binding site for activator
binding site for repressor
(B)
masking the activation surface
TATA
binding site for activator
binding site for repressor
(C)
direct interaction with the general transcription factors
binding site for repressor
binding site for activator
TFIID
TATA

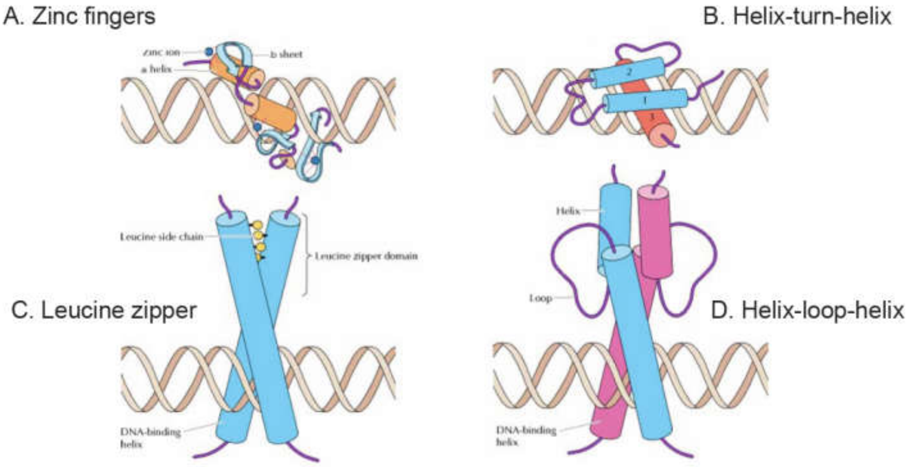
Transcription factor combinations

Most genes are regulated by multiple transcription factors



Liver cell-specific expression of the *TTR* gene

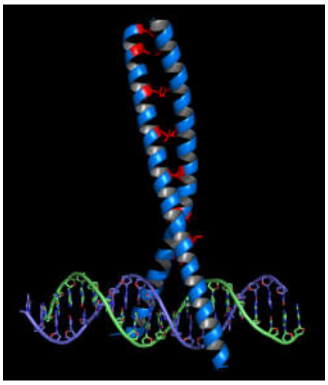
Classification of TF by DNA binding



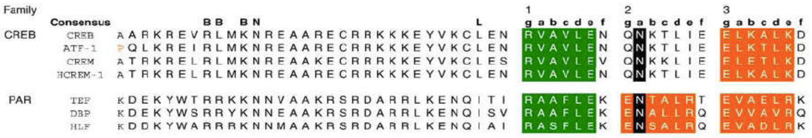
http://www.gene-regulation.com/pub/databases/transfac/cl.html

Transcription factor dimerization

Leucine zippers



- homo dimerization
- hetero dimerization



Family	Consensus		1	2	3
		B B B N L	g a b c d e f	g a b c d e f	g a b c d e f
CREB	CREB	A A R K R E V R L M K N R E A A R E C R R K K K E Y V K C L E N	R V A V L E N	Q N K T L I E	E L K A L K D
	ATF-1	P Q L K R E I R L M K N R E A A R E C R R K K K E Y V K C L E N	R V A V L E N	Q N K T L I E	E L K T L K D
	CREM	A T R K R E L R L M K N R E A A K E C R R R K K E Y V K C L E S	R V A V L E V	Q N K K L I E	E L E T L K D
	HCREM-1	A T R K R E L R L M K N R E A A R E C R R K K K E Y V K C L E N	R V A V L E N	Q N K T L I E	E L K A L K D
PAR	TEF	K D E K Y W T R R K K N N V A A K R S R D A R R L K E N Q I T I	R A A F L E K	E N T A L R T	E V A E L R K
	DBP	K D E K Y W S R R Y K N N E A A K R S R D A R R L K E N Q I S V	R A A F L E K	E N A L L R Q	E V V A V R Q
	HLF	K D D K Y W A R R R K N N M A A K R S R D A R R L K E N Q I A I	R A S F L E K	E N S A L R Q	E V A D L R K

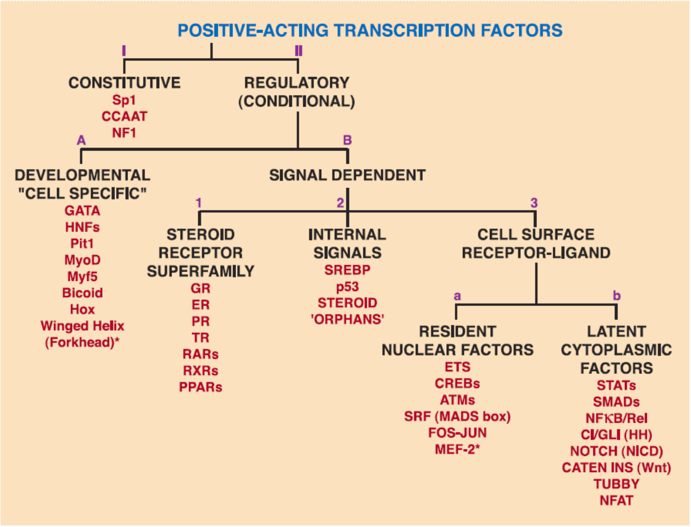
Signaling

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

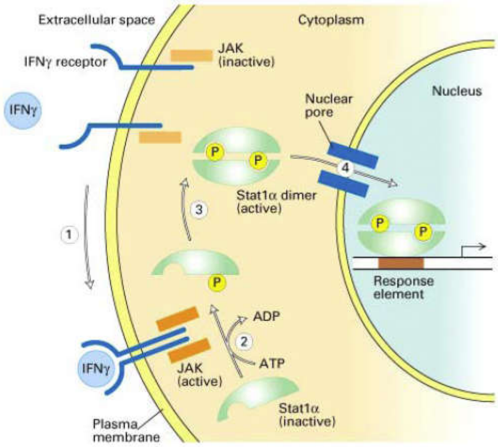
Classification of TF by function



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

Regulation by phosphorylation

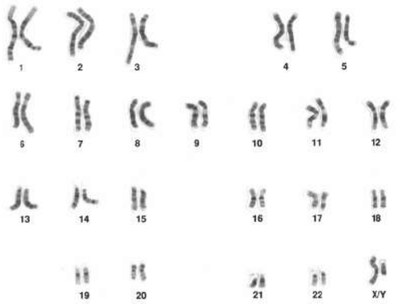
- Hormone activates kinase
- Kinase phosphorylates transcription factor
- Transcription factor is activated

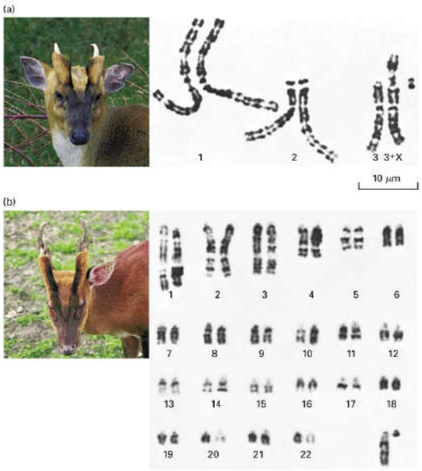


Principles of TF regulation

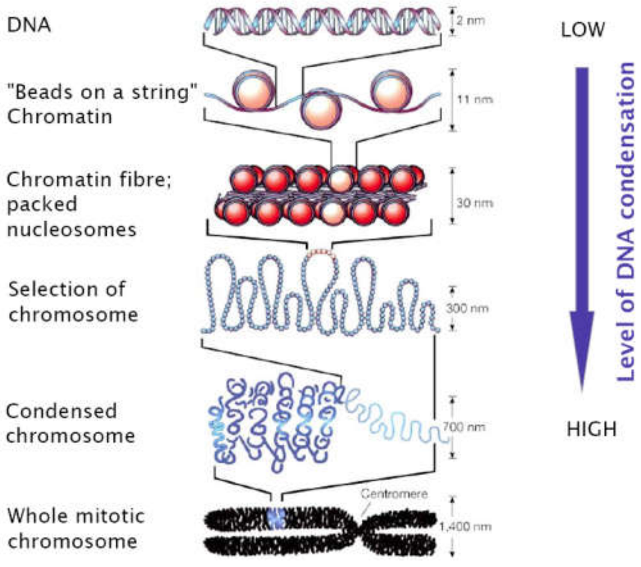
- 1 TF can target promoter of many genes
- >1 TF regulate expression of 1 gene (modules)
- Cascade of TF possible
- Positive feedback loop (autoregulation)
- Feed forward loop

Chromosomes





DNA packing



DNA

"Beads on a string" Chromatin

Chromatin fibre; packed nucleosomes

Selection of chromosome

Condensed chromosome

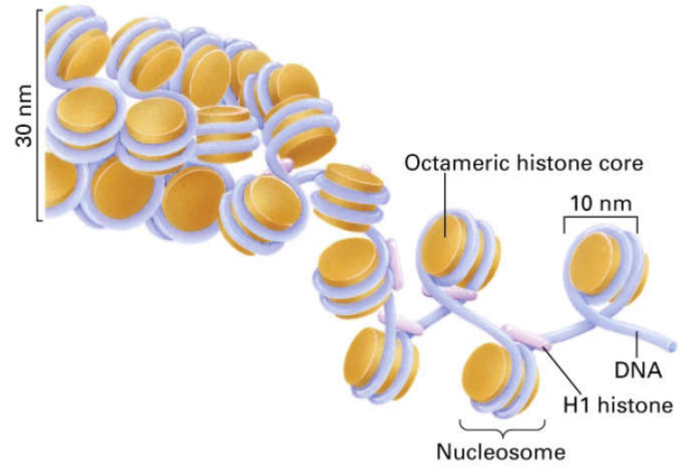
Whole mitotic chromosome

LOW

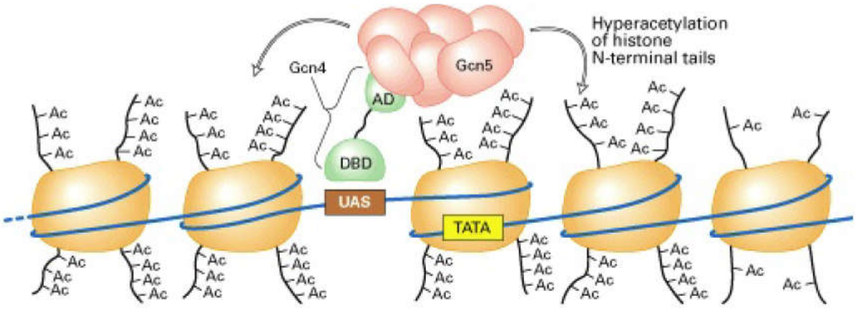
Level of DNA condensation

HIGH

The solenoid model of condensed chromatin

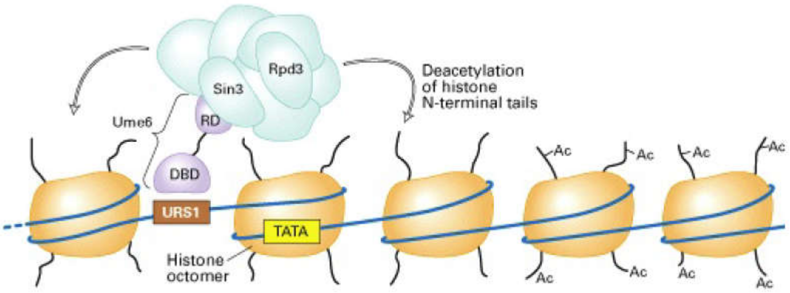


Activators: histone acetylation



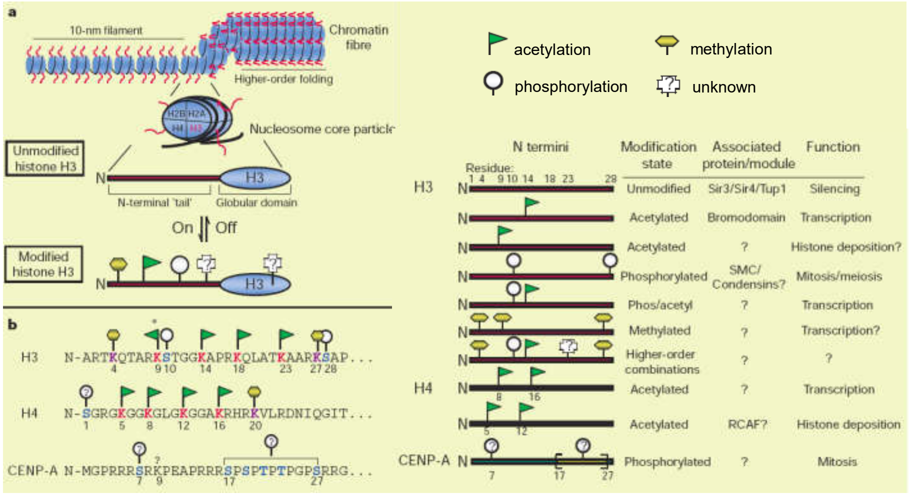
- Some activators recruit histone acetylase, which adds acetyl groups to histones
- Allows transcriptional machinery access to less condensed template DNA (euchromatin)

Repressors: histone deacetylation



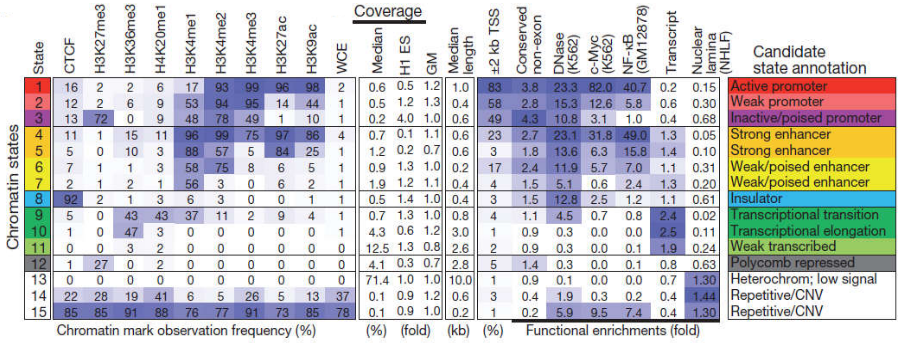
- Some repressors recruit histone deacetylase, which removes acetyl groups from histones
- Prevents transcriptional machinery access by condensing template DNA (heterochromatin)

Histone modification and histone code



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

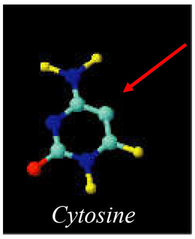
Chromatin states



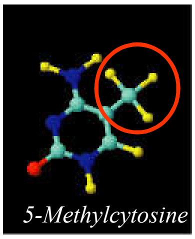
State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage Median	Coverage H1 ES	Coverage GM	Median length	±2 kb TSS	Conserved non-exon	DNase (K562)	c-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	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 (%)										(%)	(fold)		(kb)	(%)	Functional enrichments (fold)						

Ernst et al. Nature 2011.

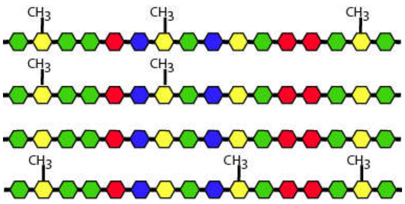
DNA methylation



Cytosine



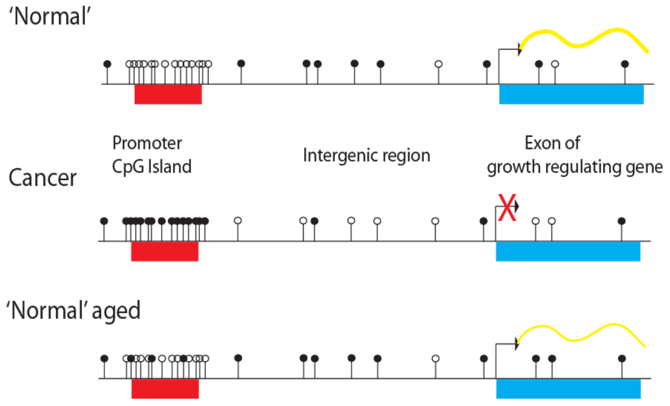
5-Methylcytosine



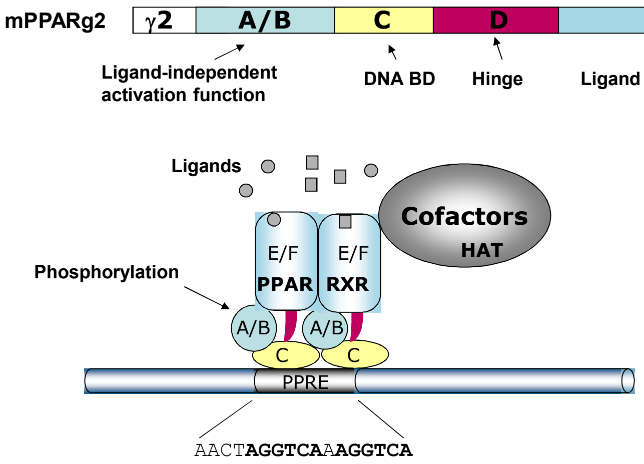
DNA methylation

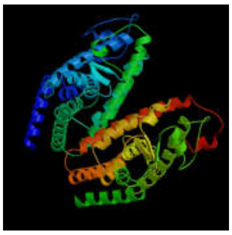
- Once differential expression patterns have been set up **epigenetic mechanisms** can ensure that differential expression patterns are stably inherited when cells divide
- Methylation does not alter base pairing
- 3% of cytosines in human DNA are methylated
- ~76% - 100% of cytosines in CpG islands are methylated
- DNA methyltransferases (DNMT1, DNMT3A, DNMT3b), for maintenance and *de novo* methylation of DNA
- CpG methylation is regulated tightly during development and is associated with gene silencing, X-inactivation, and allele specific

Aberrant methylation patterns



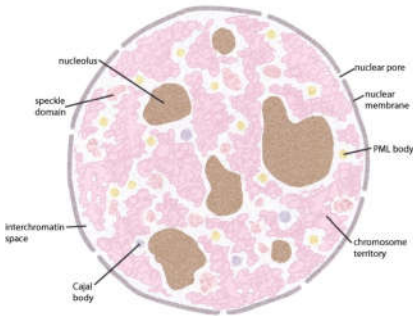
Nuclear receptors



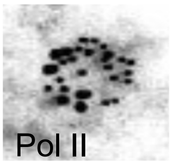


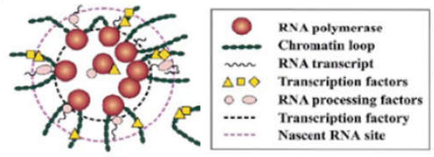
Functional compartmentalization of the nucleus

Compartments



Transcription factories

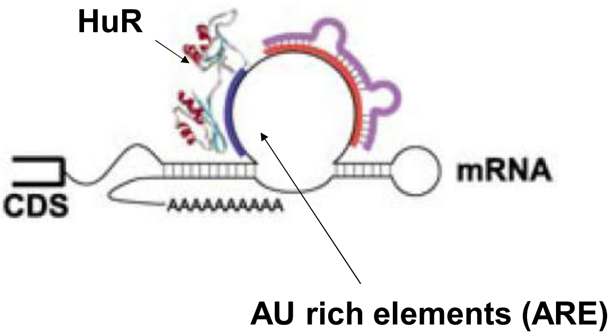



● RNA polymerase
Chromatin loop
RNA transcript
▲■◆ Transcription factors
○ ⬭ RNA processing factors
----- Transcription factory
----- Nascent RNA site

Timothy P. O'Brien et al.
Genome Res. 2003. 13: 1029-1041

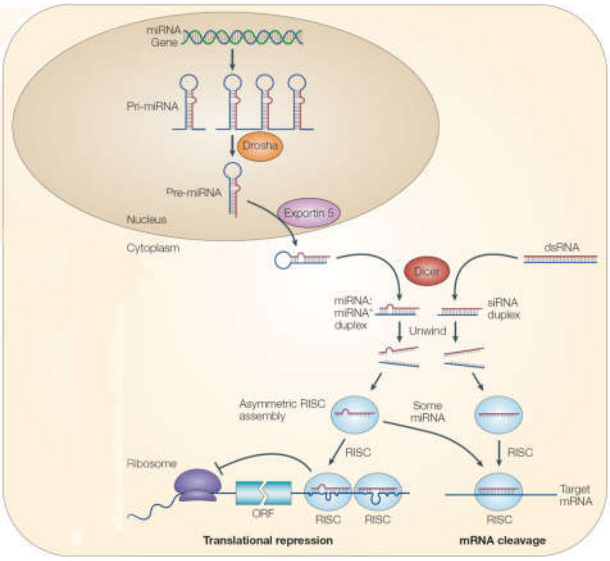
Iborra et al.
J Cell Sci 1996

RNA binding proteins for mRNA stability



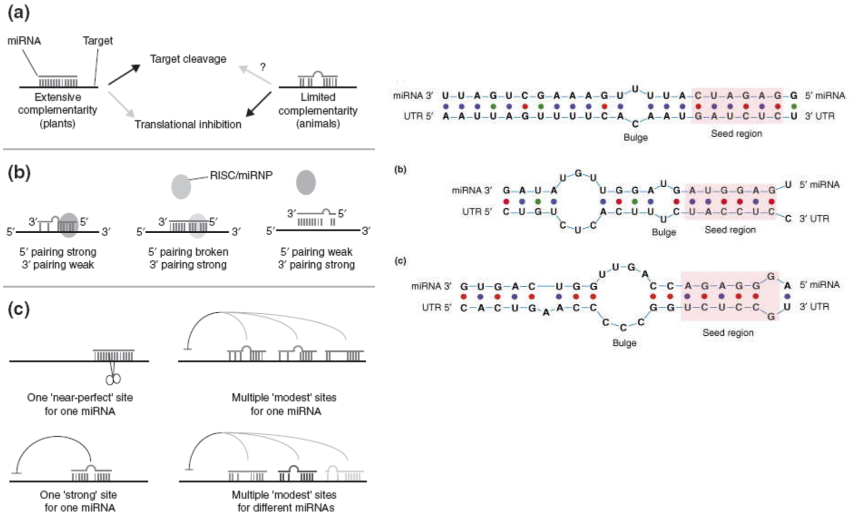
Cox-2	UAUUAAUUUAAUUAUUUAAUAAUAUUUAUAUUAAA
IL-1β	UAUUUAUUUAUUUAUUUGUUUGUUUGUUUUAUU
IL-2	UAUUUAUUUAAAUAUUUAAAUUUUAUAUUUAUU
IL-4	AUAUUUUAAUUUAUGAGUUUUUGAUAGCUUUAUUUUUUAAG
IL-8	UAUUUAUUAUUUAUGUAUUUAUUUAA
TNFα	AUUAUUUAUUAUUUAUUUAUUAUUUAUUUAUUUA

microRNA and siRNA

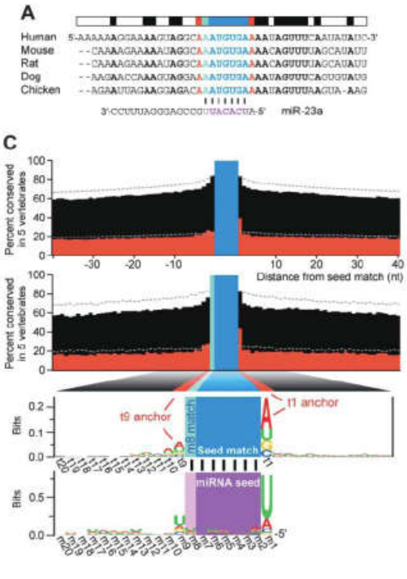


He L., Hannon GJ. Nature Reviews Genetics. 2004. 5:522-531

miRNA-mRNA targeting


(a)
miRNA
Target
Target cleavage
?
Extensive complementarity (plants)
Translational inhibition
Limited complementarity (animals)
(b)
RISC/miRNP
5′ pairing strong 3′ pairing weak
5′ pairing broken 3′ pairing strong
5′ pairing weak 3′ pairing strong
(c)
One 'near-perfect' site for one miRNA
Multiple 'modest' sites for one miRNA
One 'strong' site for one miRNA
Multiple 'modest' sites for different miRNAs
miRNA 3′ U-U-A-G-U-C-G-A-A-A-G-U-U-U-U-A-C-U-A-G-A-G-G 5′ miRNA
UTR 5′ A-A-U-U-A-G-U-U-U-C-A-C-A-A-U-G-A-U-C-U-C-U 3′ UTR
Bulge
Seed region
(b)
miRNA 3′ G-A-U-A-U-G-U-U-G-G-A-U-G-A-U-G-G-A-G-U 5′ miRNA
UTR 5′ C-U-G-U-C-U-C-A-C-U-U-U-C-U-A-C-C-U-C-C 3′ UTR
Bulge
Seed region
(c)
miRNA 3′ G-U-G-A-C-U-G-G-U-G-A-C-C-A-G-A-G-G-A 5′ miRNA
UTR 5′ C-A-C-U-G-A-A-C-C-C-C-C-G-G-U-C-U-C-C-G-U 3′ UTR
Bulge
Seed region

Conservation of microRNA target sequences



Genome analyses

Human Genome

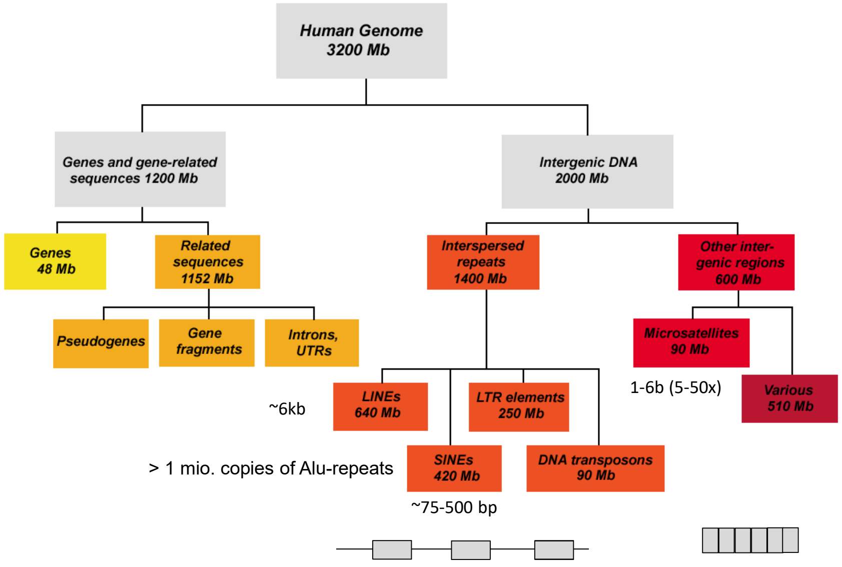
2.95 Gbases of 3.2 Gbases is euchromatin

- >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

Organization of the human genome



Human Genome 3200 Mb

Genes and gene-related sequences 1200 Mb

Intergenic DNA 2000 Mb

Genes 48 Mb

Related sequences 1152 Mb

Pseudogenes

Gene fragments

Introns, UTRs

Interspersed repeats 1400 Mb

Other inter-genic regions 600 Mb

Microsatellites 90 Mb

Various 510 Mb

LINEs 640 Mb

LTR elements 250 Mb

SINEs 420 Mb

DNA transposons 90 Mb

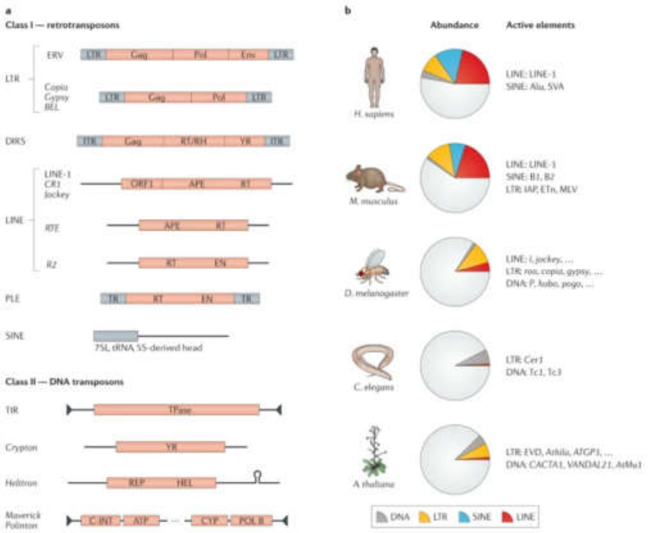
~6kb

> 1 mio. copies of Alu-repeats

~75-500 bp

1-6b (5-50x)

Transposons



Deniz et al. Nat Rev Genet. 2019

Bioinformatics challenges in genome analysis

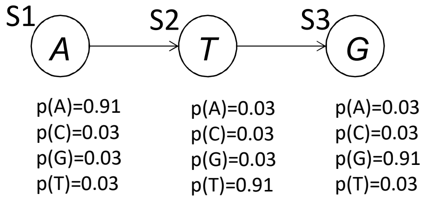
- Gene finding
- Start codon
- Exon-intron borders
- CpG-islands
- Repetitive sequences (Repeat Masker)
- Regulatory sequences

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



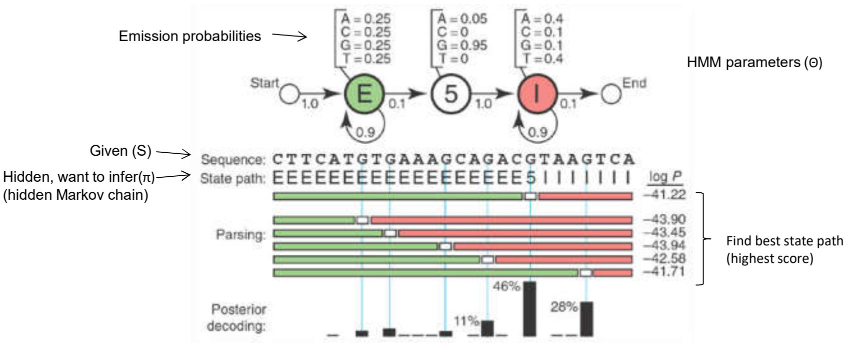
S1	S2	S3
p(A)=0.91	p(A)=0.03	p(A)=0.03
p(C)=0.03	p(C)=0.03	p(C)=0.03
p(G)=0.03	p(G)=0.03	p(G)=0.91
p(T)=0.03	p(T)=0.91	p(T)=0.03

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

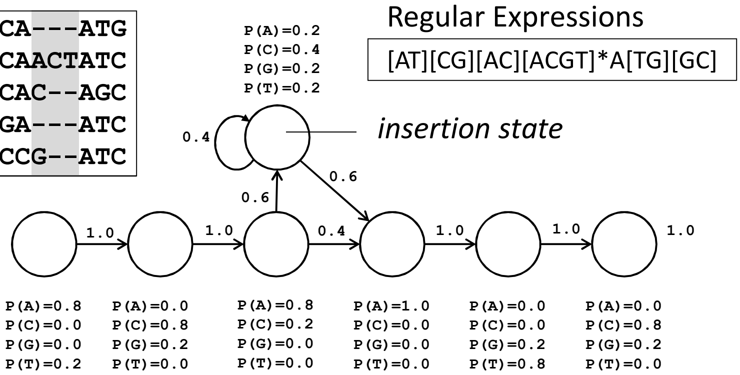
Profile Hidden Markov Model

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

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

Regular Expressions

[AT][CG][AC][ACGT]*A[TG][GC]



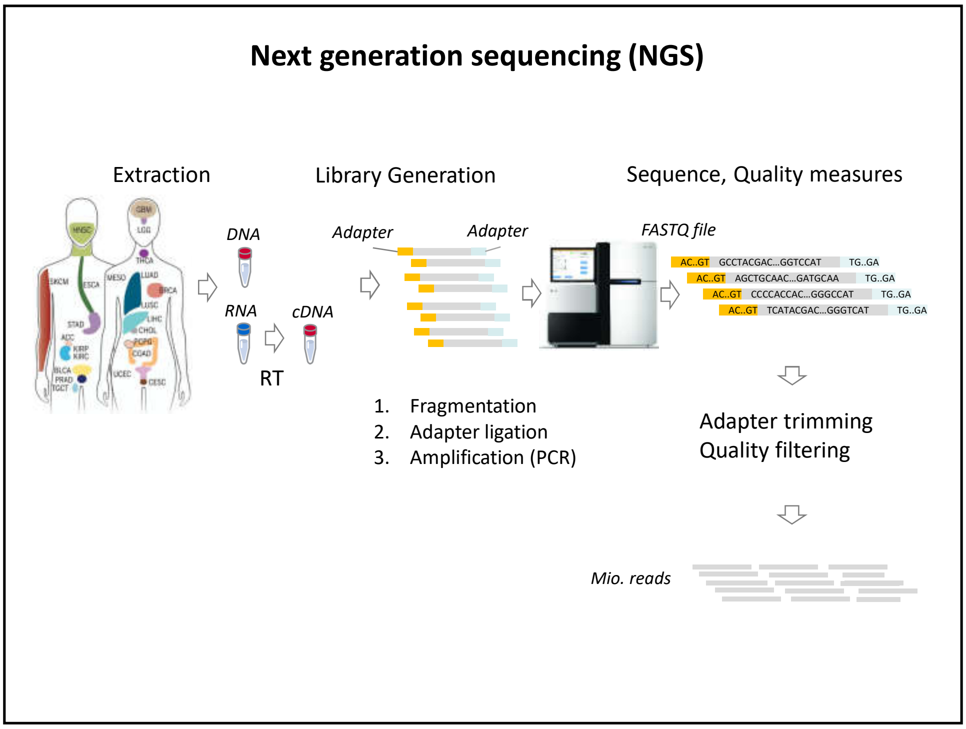
p(*ACACATC*)=0.8*1*0.8*1*0.8*0.6 *0.4*0.6*1*1*0.8*1*0.8=0.047

log-odds=log(p(S)/0.25^{L})=log(0.047/0.25^{7})

II Biological sequence analyses

- Mapping algorithms for NGS data
- Sequence alignment of 2 sequences
- Multiple sequence alignment
- Predictive models using protein sequences
- Regulatory sequences

Mapping algorithms for NGS data



Next generation sequencing (NGS)

Extraction

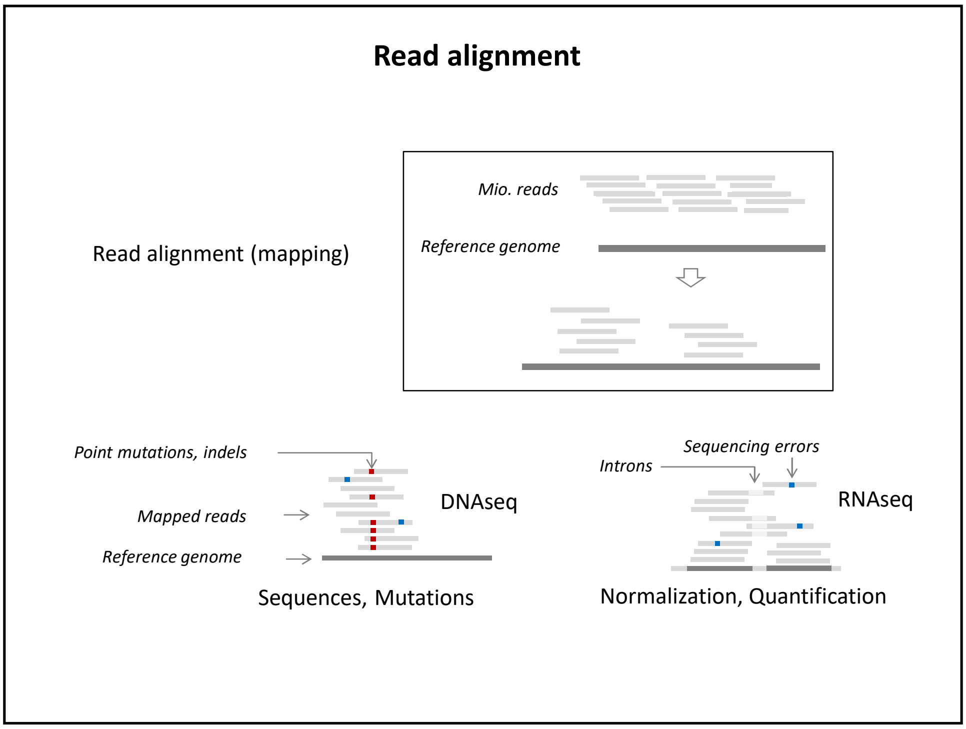
Library Generation

Sequence, Quality measures

AC..GT GCCTACGAC...GGTCCAT TG..GA
AC..GT AGCTGCAAC...GATGCAA TG..GA
AC..GT CCCCACCAC...GGGCCAT TG..GA
AC..GT TCATACGAC...GGGTCAT TG..GA

1. Fragmentation
2. Adapter ligation
3. Amplification (PCR)

Adapter trimming
Quality filtering



Read alignment

Read alignment (mapping)

Sequences, Mutations

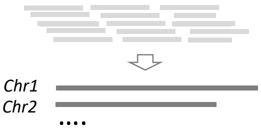
Normalization, Quantification

Exact string matching

Problem

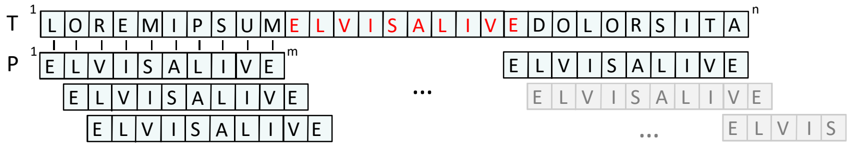
10 mio. short sequence reads (100 bp)

Reference genome (hg38) ($3*10^9$ bp)



⇨ String matching problem in text processing

1 Naïve approach


T 1 L O R E M I P S U M E L V I S A L I V E D O L O R S I T A n

P 1 E L V I S A L I V E m

E L V I S A L I V E

E L V I S A L I V E

...

E L V I S A L I V E

E L V I S A L I V E

...

E L V I S

$O[(n-m+1)*m]$

$s=10^7$ $m=10^2$ $n=3*10^9$ ⇨ $10^7*(3*10^9-99)*10^2$= max. $3*10^{18}$ comparisons

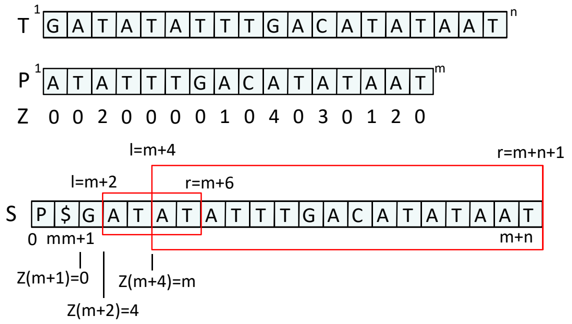
⇕

Desktop PC: 10^{12} floating point operations/s

Exact string matching algorithms

Z-box algorithm

Z(*k*)= longest substring starting at *k* which is also prefix of the string


T 1 G A T A T A T T T G A C A T A T A A T n

P 1 A T A T T T G A C A T A T A A T m

Z 0 0 2 0 0 0 0 1 0 4 0 3 0 1 2 0

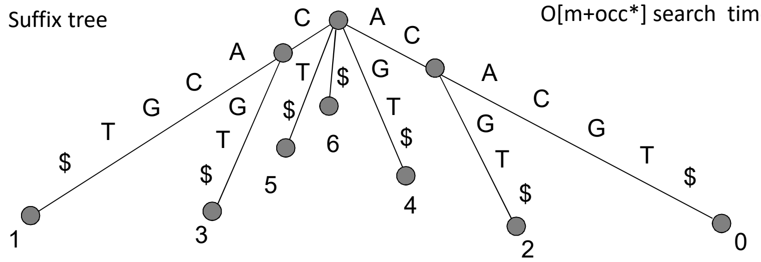
O[n+m]

- There are a number of improvements and other string matching algorithms such as *Boyer-Moore* or *Knutt-Morris-Pratt*

Suffix trees (ordered tree data structure)

Sequence ACACGT$

Suffix tree



O[m+occ*] search time

Suffix array

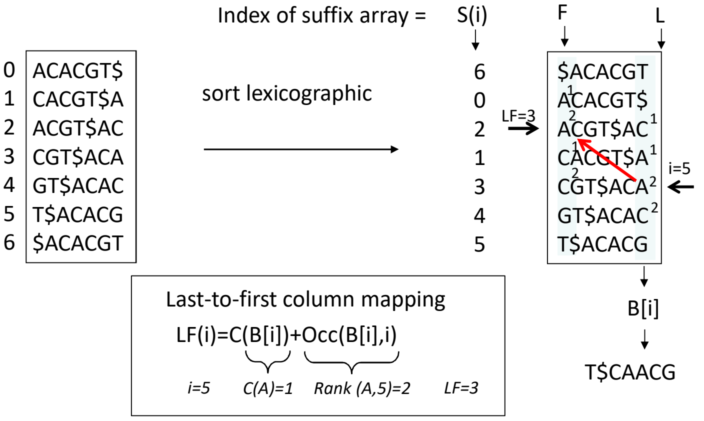
0	ACACGT$
1	CACGT$
2	ACGT$
3	CGT$
4	GT$
5	T$
6	$

O[m+log n+occ*] search time

*occ =number of occurences of P in T

Burrows-Wheeler transform

1. Append character (not part of alphabet)
2. Cyclic permutations
3. Sort lexicographic
4. Last column is Burrows-Wheeler transform (BWT, B[i])



Index of suffix array = S(i)

0	ACACGT$
1	CACGT$A
2	ACGT$AC
3	CGT$ACA
4	GT$ACAC
5	T$ACACG
6	$ACACGT

sort lexicographic

S(i)	F … L
6	$ACACGT
0	A[1]CACGT$
2	A[2]CGT$AC[1]
1	C[1]ACGT$A[1]
3	C[2]GT$ACA[2]
4	GT$ACAC[2]
5	T$ACACG

LF=3

i=5

B[i]

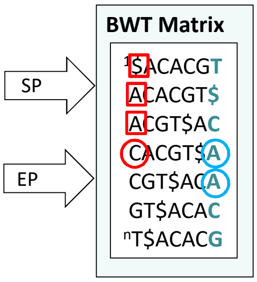
T$CAACG

Last-to-first column mapping

$$LF(i)=C(B[i])+Occ(B[i],i)$$

i=5 *C(A)=1* *Rank (A,5)=2* *LF=3*

Backward search algorithm (FM index)



C(c)

$	A	C	G	T
0	1	3	5	6

P= 1 | C | A | C | m

Occ(c,i) (=rank)

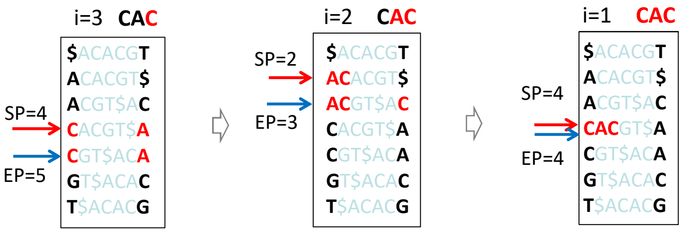
$	A	C	G	T
0	0	0	0	1
1	0	0	0	1
1	0	1	0	1
1	1	1	0	1
1	2	1	0	1
1	2	2	0	1
1	2	2	1	1

```
SP=1, EP=n
for i=m to 1 do
        SP=C(P[i])+Occ(P[i],SP-1)+1
        EP=C(P[i])+Occ(P[i],EP)
        if SP>EP then return Ø
end
return (SP,EP)
```

O[m]

FM …Full-text index in Minute space

Backward search algorithm for exact string matching

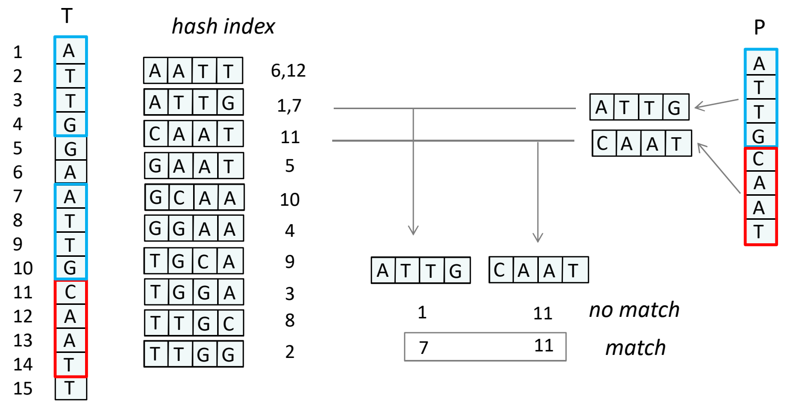


- FM-index can be also used for approximate string matching (k-mismatch search) by *backtracking.*
- BWT is compressible (run length encoding, move-to-front)
- In the original *Bowtie* implementation of the BWT-based FM-index for the human genome requires only 1.3 GB of memory.

Hash index based methods

Hashing

– Using *k*-mer seeds



hash index

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

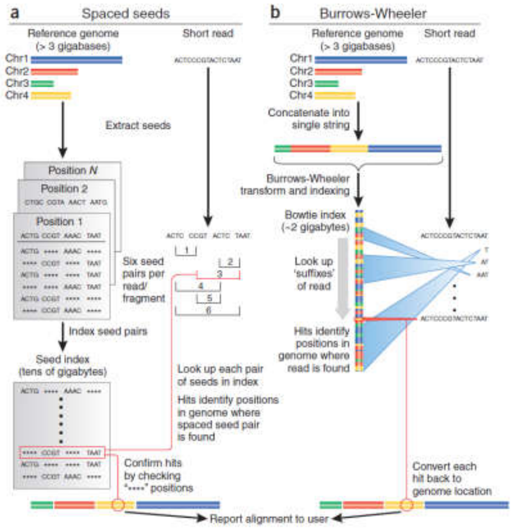
no match

match

– An extension step may account for errors or mismatches (spaced seeds)

Examples

Maq

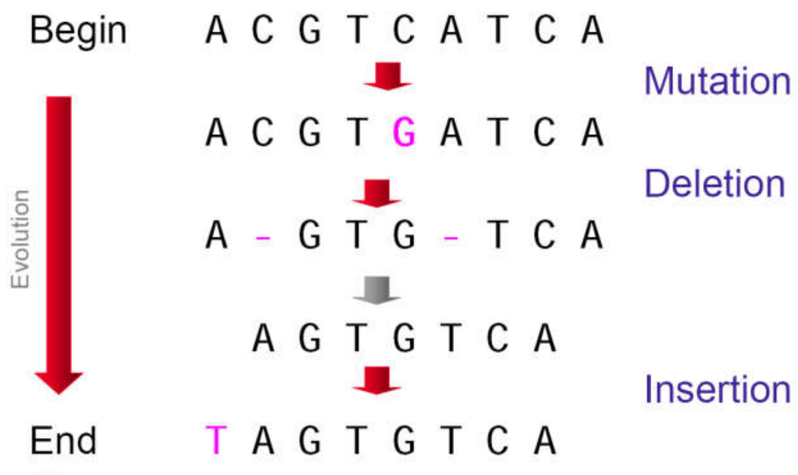


Tophat

Trapnell C, Salzberg S. Nature Biotech. 2009

Sequence alignment of 2 sequences

Genomes change over time



Begin A C G T C A T C A

Mutation

A C G T G A T C A

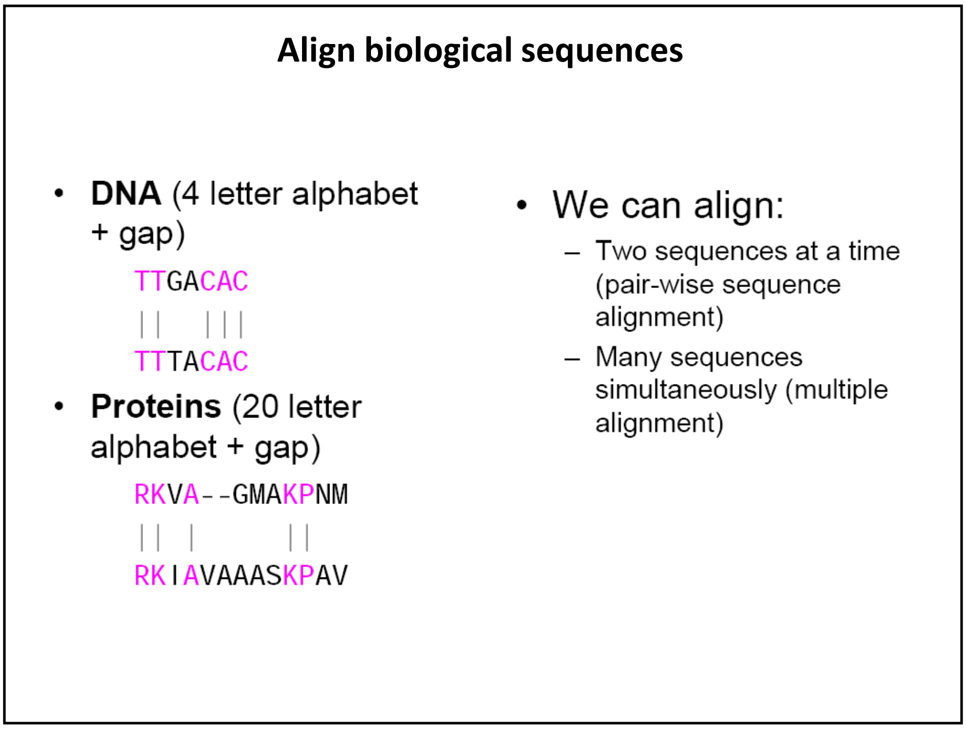
Deletion

A - G T G - T C A

A G T G T C A

Insertion

End T A G T G T C A



Align biological sequences

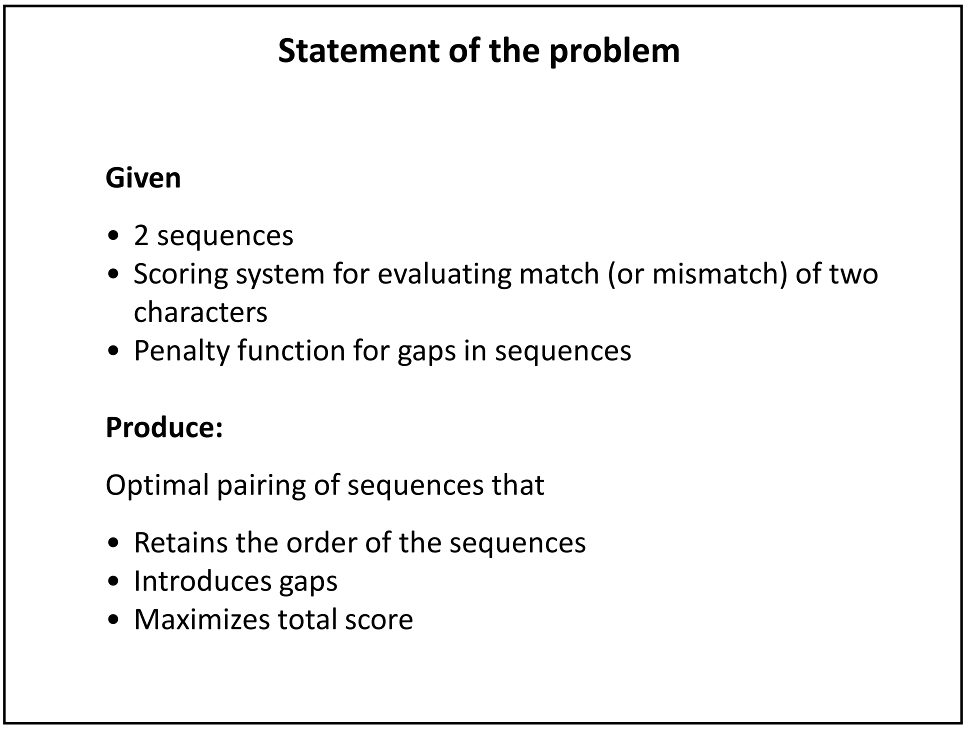
- **DNA** (4 letter alphabet + gap)

```
TTGACAC
||  |||
TTTACAC
```

- **Proteins** (20 letter alphabet + gap)

```
RKVA--GMAKPNM
|| |     ||
RKIAVAAASKPAV
```

- We can align:
  - Two sequences at a time (pair-wise sequence alignment)
  - Many sequences simultaneously (multiple alignment)



Statement of the problem

Given

- 2 sequences
- Scoring system for evaluating match (or mismatch) of two characters
- Penalty function for gaps in sequences

Produce:

Optimal pairing of sequences that

- Retains the order of the sequences
- Introduces gaps
- Maximizes total score

Enumeration of all possible alignments

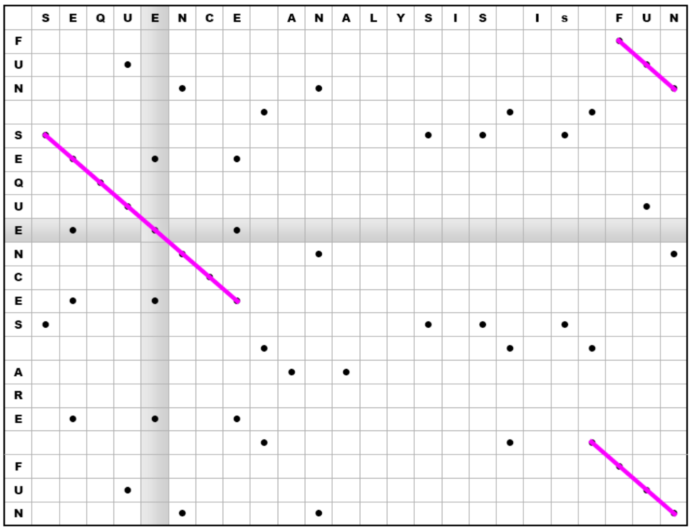
- Number of possible alignments of 2 sequences with length n and m

$$\begin{bmatrix} n+m \\ m \end{bmatrix} = \frac{(m+n)!}{(m!)^2} \approx \frac{2^{m+n}}{\sqrt{\pi \cdot m}}$$

- For 2 sequences of length n

n	enumeration
10	184,756
20	1.40E+11
100	9.00E+58

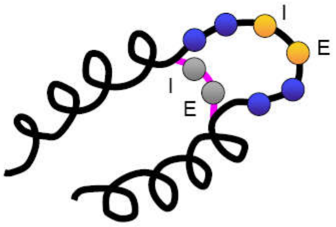
Dot matrix



	S	E	Q	U	E	N	C	E		A	N	A	L	Y	S	I	S		I	s		F	U	N
F																						●		
U				●																			●	
N						●					●													●
									●									●			●			
S	●														●		●			●				
E		●			●			●																
Q			●																					
U				●																			●	
E		●			●			●																
N						●					●													●
C							●																	
E		●			●			●																
S	●														●		●			●				
									●									●			●			
A										●		●												
R																								
E		●			●			●																
									●									●			●			
F																						●		
U				●																			●	
N						●					●													●

Biology of gaps

```
AGKLAVRSTMIESTRVILTWRKW
AGKLAVRS--IE--RVILTWRKW
vs.
AGKLAVRSTMIEST--RVILTWRKW
AGKLAVRS------IERVILTWRKW
vs.
Many others…
```

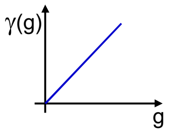


Gap penalties

We expect to penalize gaps - the standard cost associated with a gap of length g:

- Linear gap penalty function

γ (g) = -g*d



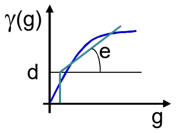
- Convex gap penalty function (more realistic)

Affine score:

γ (g) = - d – (g-1)*e

gap open penalty

gap extend penalty



Distance scoring (DNA sequnces)

- **Hamming distance:**
  Number of letters in which sequences differ (not valid if the sequences have different length)

s	AAT	AGCAA	AGCACACA
t	TAA	ACATA	A-CACACTA
HD(s,t)	2	3	2

- **Levenshtein distance:**
  w(a,a)=0
  w(a,b)=1 for a≠b
  w(-,a)=w(b,-)=1

deletion insertion

s	AGCACAC-A
t	A-CACACTA
d(s,t)	2

For two sequences, the distance is unique, but the optimal alignment (the one with minimal cost or distance) is not unique

Substitutions matrices (protein sequences)

- Unrelated or random model assumes that letter a occurs independently with some frequency *qa.*

$$P(x,y|R) = \prod q_{xi} \prod q_{xj}$$

- The alternative match model of aligned pairs of residues occurs with a joint probability *pab.*

$$P(x,y|M) = \prod p_{xi\,yi}$$

- Odds ratio

$$\frac{P(x,y|M)}{P(x,y|R)} = \frac{\prod p_{xi\,yi}}{\prod q_{xi} \prod q_{yj}} = \prod \frac{p_{xi\,yi}}{q_{xi}\, q_{yj}}$$

Substitution matrices

- Log-odds ratio (*score matrix* or *substitution matrix*)

  $S = \Sigma s(xi,yi)$ where $s(a,b) = \log \frac{p_{ab}}{q_a\, q_b}$ for aligned pair(a,b)

  s>0 … more likely than random, *s*<0 … less likely than random

- Physical properties of amino acids (e.g. hydrophob vs. hydrophil) are the reason that there are differences in the substitution scores

- Manually align protein structures (or, more risky, sequences)

- Look for frequency of amino acid substitutions at structurally nearly constant sites.

PAM matrices

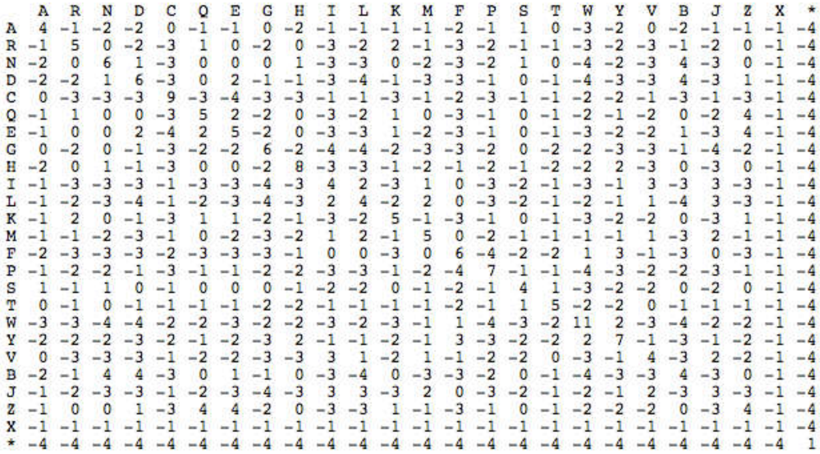
- Margaret Dayhoff, 1978

- Point Accepted Mutation (PAM)
  - Look at patterns of substitutions in related proteins
  - The new side chain must function the same way as the old one (“acceptance”)
  - On average, 1 PAM corresponds to 1 amino acid change per 100 residues
  - 1 PAM ~ 1% divergence
  - Extrapolate to predict patterns at longer distances

BLOSUM matrices

- Henikoff and Henikoff, 1992
- Blocks Substitution Matrix (BLOSUM n)
  - Look only for differences in conserved, ungapped regions of a protein family
  - More sensitive to structural or functional substitutions
  - Contribution of sequences > n% identical weighted to 1

BLOSUM62



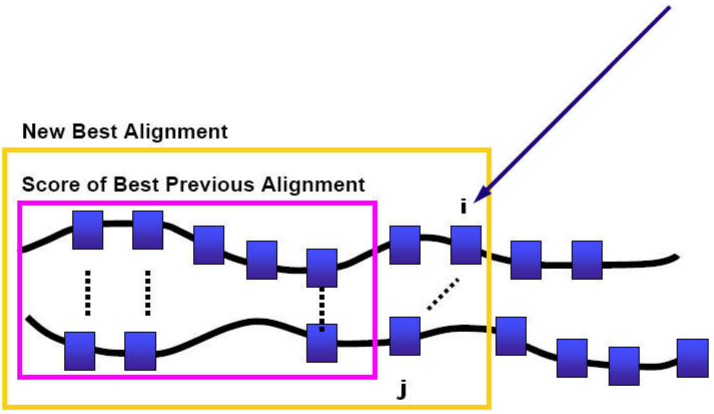
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	J	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	-1	-1	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	-2	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	4	-3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	-3	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-1	-3	-1	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	-2	4	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	-3	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-4	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	-3	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	3	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	-3	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	2	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	0	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-3	-1	-1	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	-2	0	-1	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	-1	-1	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-2	-2	-1	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-1	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	2	-2	-1	-4
B	-2	-1	4	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	-3	0	-1	-4
J	-1	-2	-3	-3	-1	-2	-3	-4	-3	3	3	-3	2	0	-3	-2	-1	-2	-1	2	-3	3	-3	-1	-4
Z	-1	0	0	1	-3	4	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-2	-2	-2	0	-3	4	-1	-4
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Summary of substitutions matrices

- Triple-PAM strategy (Altschul, 1991)
  - PAM 40 short alignments, highly similar
  - PAM 120
  - PAM 250 longer, weaker local alignments
- BLOSUM (Henikoff, 1993)
  - BLOSUM 90 short alignments, highly similar
  - BLOSUM 62 most effective in detecting known members of a protein family (Standard in BLAST)
  - BLOSUM 30 longer, weaker local alignments
- No single matrix is the complete answer for all sequence comparisons

Dynamic programing for sequence alignment

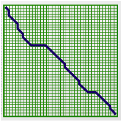
New Best Alignment = Previous Best + Local Best



Sequence alignment

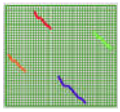
- Global alignment

  Needleman-Wunsch algorithm



- Local alignment

  Smith-Waterman algorithm



Mike Waterman

Temple Smith

Global alignment: Needleman-Wunsch algorithm

- Construct a matrix F(i,j) where i is index from sequence 1 and j is the index from sequence 2
- Starting with F(0,0)=0

substitution matrix

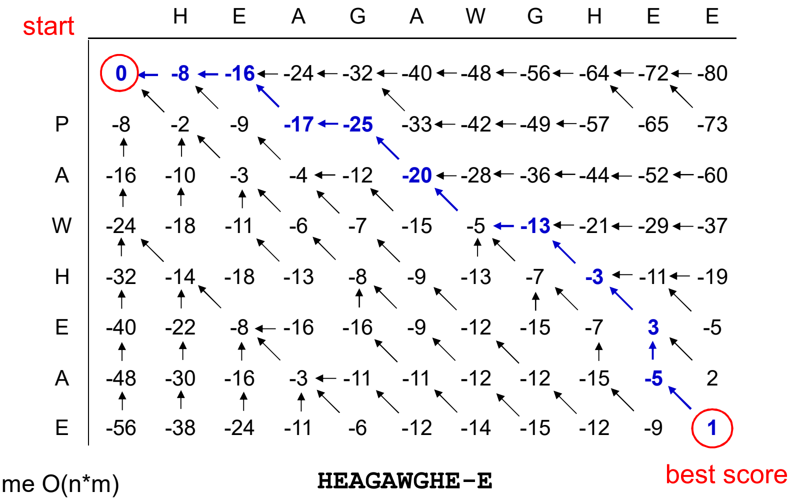
$$F(i,j) = \max \begin{cases} F(i-1,j-1)+s(x_i,y_j) \\ F(i-1,j)-d \\ F(i,j-1)-d \end{cases}$$

gap penalty

$$\begin{tikzcd}
F(i-1,j-1) \arrow[dr, "s(x_i{,}y_j)"'] & F(i,j-1) \arrow[d, "-d"] \\
F(i-1,j) \arrow[r, "-d"'] & F(i,j)
\end{tikzcd}$$

Global sequence alignment

Example with S=BLOSUM50 and d=8



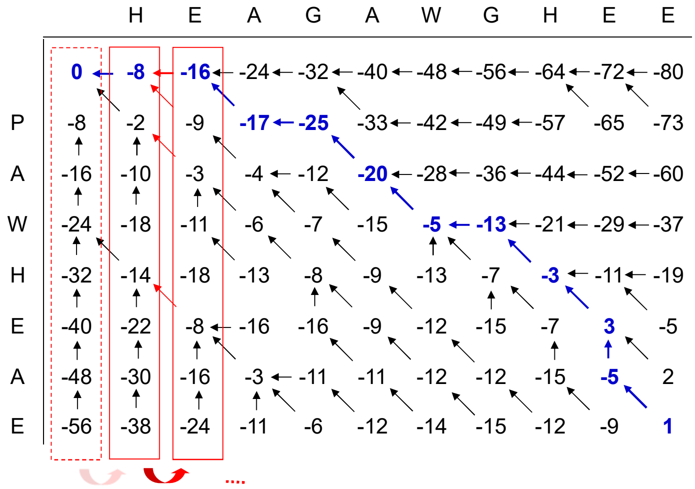
		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

Time O(n*m)
Space O(n*m)

```
HEAGAWGHE-E
--P-AW-HEAE
```

Linear space alignment

- Do calculate the score for column j only column j-1 is needed



		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

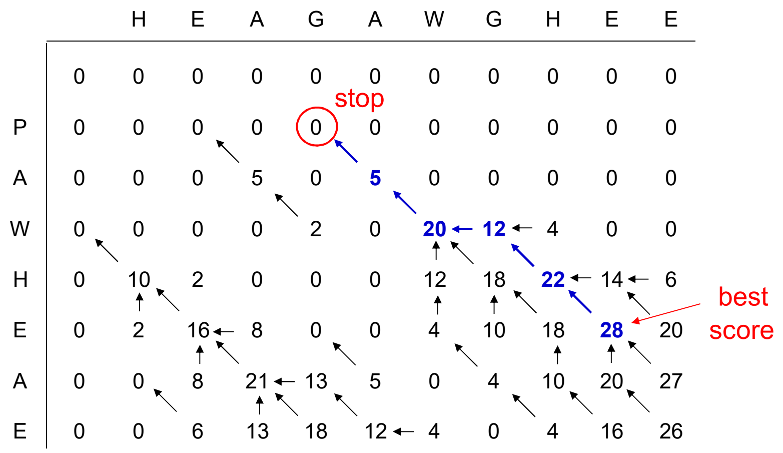
Local alignment: Smith-Waterman algorithm

- Look for best alignments between subsequences
- E.g. two proteins sharing a common domain
- Algorithm is similar to global alignment

$$F(0,j) = F(i,0)=0$$

$$F(i,j) = \max \begin{cases} 0 \\ F(i-1,j-1)+s(x_i,y_j) \\ F(i-1,j)-d \\ F(i,j-1)-d \end{cases}$$

Local alignment



	H	E	A	G	A	W	G	H	E	E
0	0	0	0	0	0	0	0	0	0	0
P 0	0	0	0	0	0	0	0	0	0	0
A 0	0	0	5	0	5	0	0	0	0	0
W 0	0	0	0	2	0	20	12	4	0	0
H 0	10	2	0	0	0	12	18	22	14	6
E 0	2	16	8	0	0	4	10	18	28	20
A 0	0	8	21	13	5	0	4	10	20	27
E 0	0	6	13	18	12	4	0	4	16	26

```
AWGHE
AW-HE
```

Database search

- Database:
  AIKWQPRSTW…
  IKMQRHIKW…
  HDLFWHLWH…
  ……………………

- Query:
  RGIKW

- Output: sequences *similar* to query

How to answer the query

We could just scan the whole database

- But:
  - Query must be very fast
  - Most sequences will be completely unrelated to query
  - Individual alignment needs not be perfect. Can finetune

- Exploit nature of the problem
  - If you're going to reject any match with idperc < 90%, then why bother even looking at sequences which don't have a fairly long stretch of matching a.a. in a row.

W-mer indexing

- Preprocessing:

  For every W-mer (e.g., W=3) store every location in the database where it occurs (can use hashing if W is large)

- Query:
  - Generate W-mers and look them up in the database.
  - Process the results

- Running time benefit:
  - For W=3, if the sequences are “random”, then roughly one W-mer in 23^3 will match, i.e., one in a ten thousand
  - We hit only a small fraction of all sequences

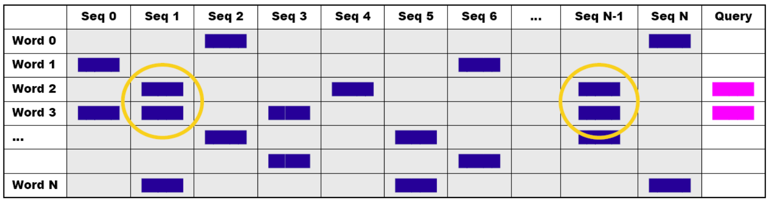
FASTA

- Use hash table of short words of the database (DB) sequence and query sequence (2-6 chars)
- For words in query sequence, find similar words in DB using (fast) hash table lookup, and compute

  R = position(query) – position (DB).

  Areas of long match will show same R for many words.

- Score matching segments based on content of these matches. Extend the good matches empirically.



	Seq 0	Seq 1	Seq 2	Seq 3	Seq 4	Seq 5	Seq 6	…	Seq N-1	Seq N	Query
Word 0			■							■	
Word 1	■						■				
Word 2		■			■				■		■
Word 3	■	■		■■					■		■
…			■			■			■		
				■■			■				
Word N		■				■				■	

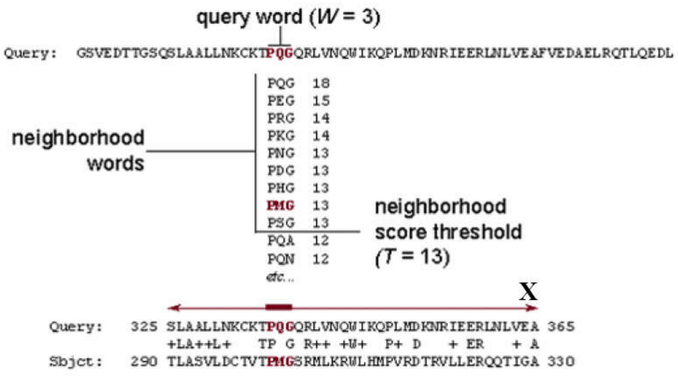
BLAST

- Finds inexact, ungapped “seeds” using a hashing technique (like FASTA) and then extends the seed to maximum length possible.
- Based on strong statistical/significance framework “What is a significantly high score of two segments of length N and M?”
- Most commonly used for fast searches and alignments. New versions now do gapped segments.

Stephen Altschul

Samuel Karlin

High-scoring segment pairs



```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

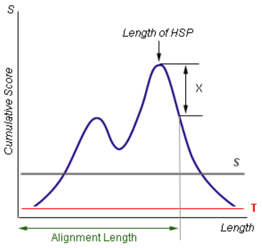
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

High-scoring segment pairs

- Receive query
  - Split query into overlapping words of length W
  - Find neighborhood words for each word until threshold T
  - Look into the table where these neighbor words occur: seeds
  - Extend seeds until score drops off under X



- Evaluate statistical significance of score
- Report scores and alignments

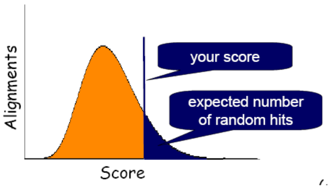
Significance of scores

The number of unrelated matches with score greater than S is approximately Poisson distributed with mean

$$\boxed{E(S)=Kmne^{-\lambda S}}$$

where λ is a scaling factor m and n are the length of the sequences

The probability that there is a match of score greater than S follows a extreme value distribution:



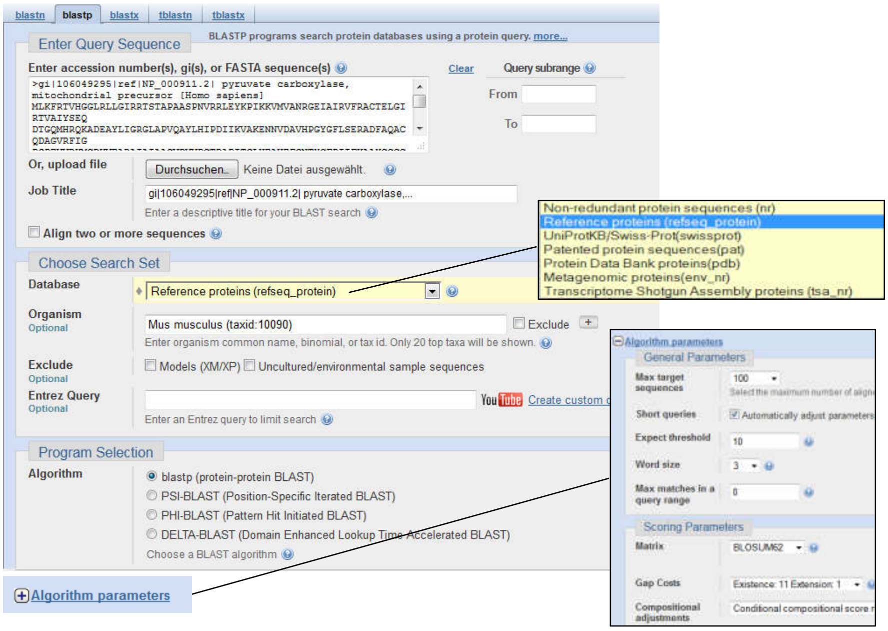
$$\boxed{P(x>S)=1-e^{-E(S)}}$$

Karlin S, Altschul S. *Proc Natl Acad Sci* (1990)

NCBI Blast

Program	*Query sequence*	*Subject sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide six-frame translation	Protein
TBLASTN	Protein	Nucleotide six-frame translation
TBLASTX	Nucleotide six-frame translation	Nucleotide six-frame translation

NCBI Blast Example



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>gi|106049295|ref|NP_000911.2| pyruvate carboxylase,
mitochondrial precursor [Homo sapiens]
MLKFRTVHGGLRLLGIRRTSTAPAASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELGI
RTVAIYSEQ
DTGQMHRQKADEAYLIGRGLAPVQAYLHIPDIIKVAKENNVDAVHPGYGFLSERADFAQAC
QDAGVRFIG
```

Query subrange

From

To

Or, upload file: Durchsuchen... Keine Datei ausgewählt.

Job Title: gi|106049295|ref|NP_000911.2| pyruvate carboxylase,...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Reference proteins (refseq_protein)

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- UniProtKB/Swiss-Prot(swissprot)
- Patented protein sequences(pat)
- Protein Data Bank proteins(pdb)
- Metagenomic proteins(env_nr)
- Transcriptome Shotgun Assembly proteins (tsa_nr)

Organism (Optional): Mus musculus (taxid:10090) □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional):

Enter an Entrez query to limit search

Program Selection

Algorithm:

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Algorithm parameters

General Parameters

Max target sequences: 100

Select the maximum number of align

Short queries: ☑ Automatically adjust parameters

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

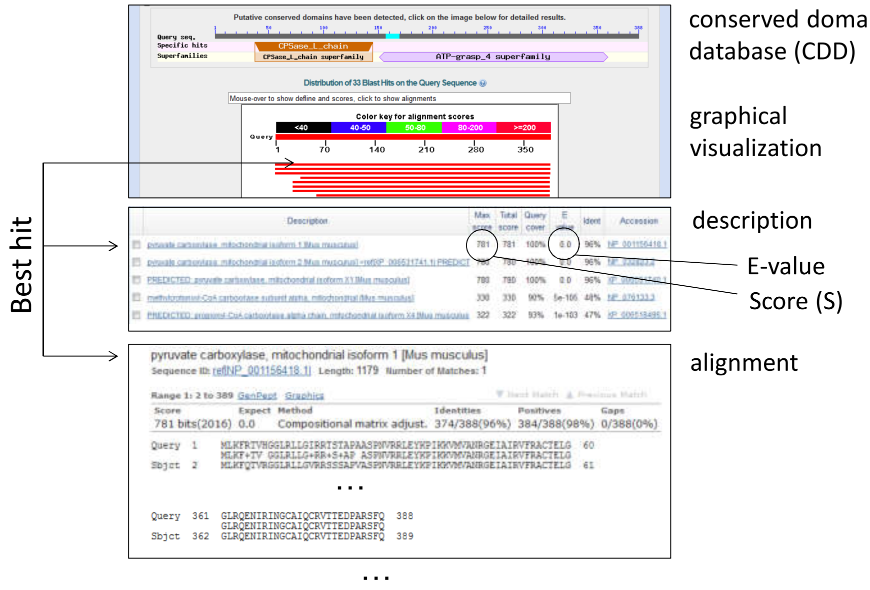
Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score r

Blast Results



Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 33 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments

Color key for alignment scores

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	pyruvate carboxylase, mitochondrial isoform 1 [Mus musculus]	781	781	100%	0.0	96%	NP_001156418.1
□	pyruvate carboxylase, mitochondrial isoform 2 [Mus musculus] >ref|XP_006531741.1| PREDICT	780	780	100%	0.0	96%	NP_032823.2
□	PREDICTED: pyruvate carboxylase, mitochondrial isoform X1 [Mus musculus]	780	780	100%	0.0	96%	XP_006531740.1
□	methylcrotonoyl-CoA carboxylase subunit alpha, mitochondrial [Mus musculus]	330	330	90%	5e-106	48%	NP_076133.3
□	PREDICTED: propionyl-CoA carboxylase alpha chain, mitochondrial isoform X4 [Mus musculus	322	322	93%	1e-103	47%	XP_006518495.1

pyruvate carboxylase, mitochondrial isoform 1 [Mus musculus]

Sequence ID: ref|NP_001156418.1| Length: 1179 Number of Matches: 1

Range 1: 2 to 389 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
781 bits(2016)	0.0	Compositional matrix adjust.	374/388(96%)	384/388(98%)	0/388(0%)

```
Query  1    MLKFRTVHGGLRLLGIRRTSTAPAASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG  60
            MLKF+TV GGLRLLG+RR+S+AP ASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG
Sbjct  2    MLKFQTVRGGLRLLGVRRSSSAPVASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG  61

                              . . .

Query  361  GLRQENIRINGCAIQCRVTTEDPARSFQ  388
            GLRQENIRINGCAIQCRVTTEDPARSFQ
Sbjct  362  GLRQENIRINGCAIQCRVTTEDPARSFQ  389
```

. . .

Best hit

conserved domain database (CDD)

graphical visualization

description

E-value

Score (S)

alignment

Multiple sequence alignment

Multiple sequence alignment

Often simple extension of pairwise alignment:

- **Given:**
  - Set of sequences
  - Match matrix
  - Gap penalties
- **Find:**
  - Alignment of sequences such that optimal score is achieved.

Goals of multiple sequence alignment

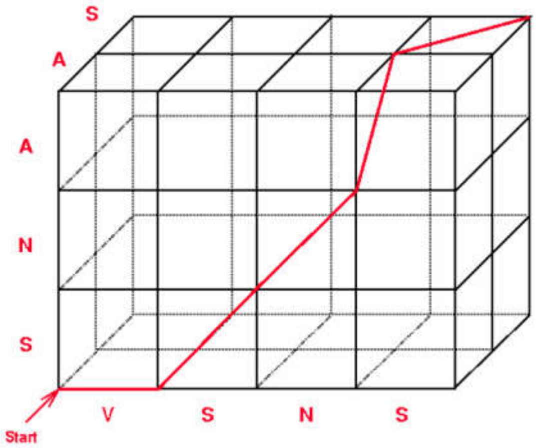
- Determine Consensus Sequences
  - Prosite, eMOTIF
  - ClustalW, MACAW, Pileup, T-Coffee
- Building Gene Families
  - Blocks, Prints, ProDom, pFAM, DOMO, eBLOCKs
- Develop Relationships & Phylogenies
  - Clusters
  - Relationships
  - Evolutionary Models
  - Phylip, GrowTree, MACAW, PAUP
- Model Protein Structures for Threading and Fold Prediction
  - Profiles, Templates, HSSP, FSSP
  - Hidden Markov Models, pFAM, SAM
  - Network Models, Neural Nets, Belief Nets
  - Statistical Models, Generalized Linear Models

Exhaustive search using dynamic programming

Why not just use same technique as for pairwise alignment?

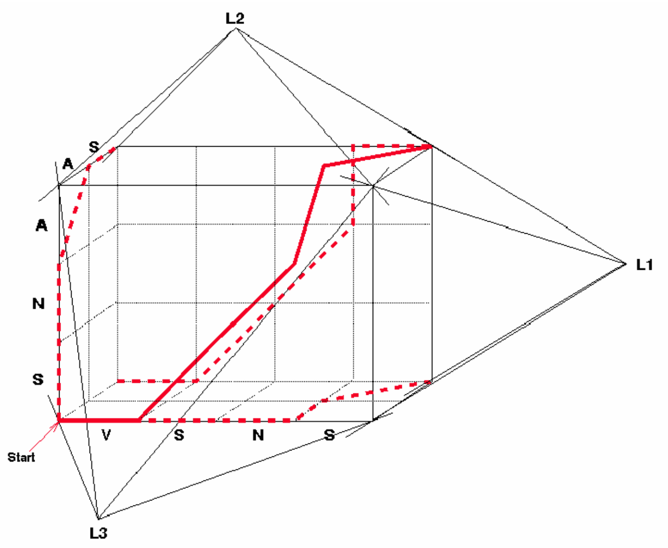
- Instead of 2-dimensional SCORE matrix, use N dimensional. Fill from one corner to diagonal corner in N dimensions.
- Complexity increases with number of sequences O(MN), so only N < 10 and lengths (M)~ 200 can be accommodated.

Dynamic Programming



```
V S N _ S
_ S N A _
_ _ _ A S
```

Dynamic Programming



MSA Algorithm

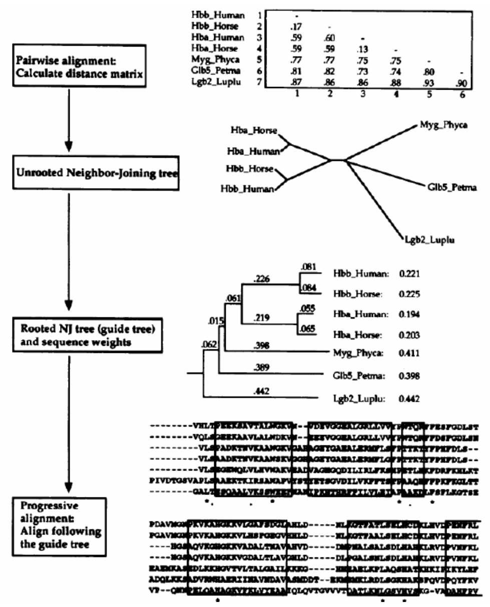
Based on dynamic programming concept:

1. Compute optimal pairwise alignments to get upperbound on any pair of alignments. (MA can't do any better than sum of optimal pairwise alignments.)

2. Create heuristic multiple alignment in ad hoc fashion to create lowerbound on MA score (e.g. align all sequences to the first).

3. Search N-dimensional scoring matrix (as in pairwise case) for optimal path, where S[i,j,k…] is the best score including ith element of sequence 1, jth of sequence 2, kth of sequence 3, etc…

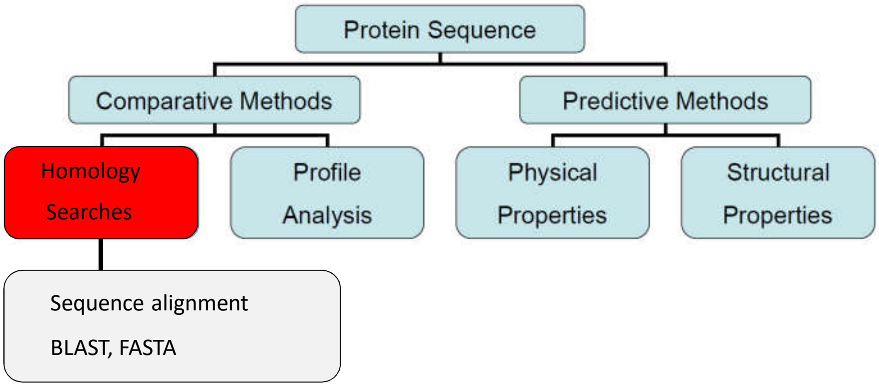
Progressive tree alignment (ClustalW)



Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

Predictive methods using protein sequences

Homology searches



Protein Sequence

Comparative Methods

Predictive Methods

Homology Searches

Profile Analysis

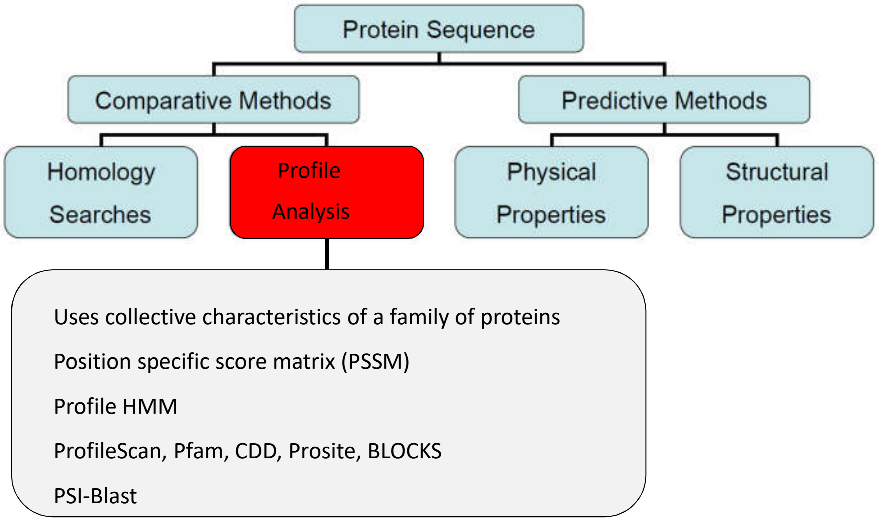
Physical Properties

Structural Properties

Sequence alignment

BLAST, FASTA

Profile Analysis



Protein Sequence

Comparative Methods

Predictive Methods

Homology Searches

Profile Analysis

Physical Properties

Structural Properties

Uses collective characteristics of a family of proteins

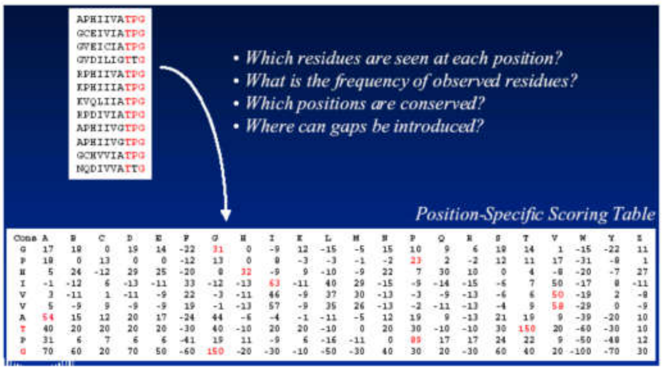
Position specific score matrix (PSSM)

Profile HMM

ProfileScan, Pfam, CDD, Prosite, BLOCKS

PSI-Blast

Profile Construction



APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG

- *Which residues are seen at each position?*
- *What is the frequency of observed residues?*
- *Which positions are conserved?*
- *Where can gaps be introduced?*

Position-Specific Scoring Table

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-5	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	-41	19	11	-9	6	-16	-11	0	85	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

$$\mathrm{PSSM}(p,a) = \sum_{b=1}^{20} f(p,b)*s(a,b)$$

f(p,b) = frequency of amino acid b in position p
s(a,b) is the score of (a,b) (from, e.g., BLOSUM or PAM)

PSI-BLAST

- Position-Specific Iterated BLAST search
- Used to identify distantly related sequences that are possibly missed during a standard BLAST search
- Easy-to-use version of a profile-based search
  - Perform BLAST search against protein database
  - Use results to calculate a position-specific scoring matrix
  - PSSM replaces query for next round of searches
  - May be iterated until no new significant alignments are found

Altschul et al., Nucleic Acids Res. 25: 3389-3402, 1997

Profile Hidden Markov Model

- Allows position dependent gap penalties
- Can be obtained from a multiple alignment (DNA or Protein)
- Can be used for searching a database for other members of the family

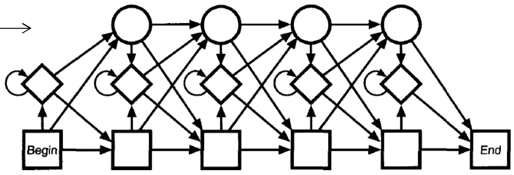
Insert states

```
G G W W R G d y . g g k k q L W F P S N Y V
I G W L N G y n e t t g e r G D F P G T Y V
P N W W E G q l . . n n r r G I F P S N Y V
D E W W Q A r r . . d e q i G I V P S K - -
G E W W K A q s . . t g q e G F I P F N F V
G D W W L A r s . . s g q t G Y I P S N Y V
G D W W D A e l . . k g r r G K V P S N Y L
- D W W E A r s l s s g h r G Y V P S N Y V
```

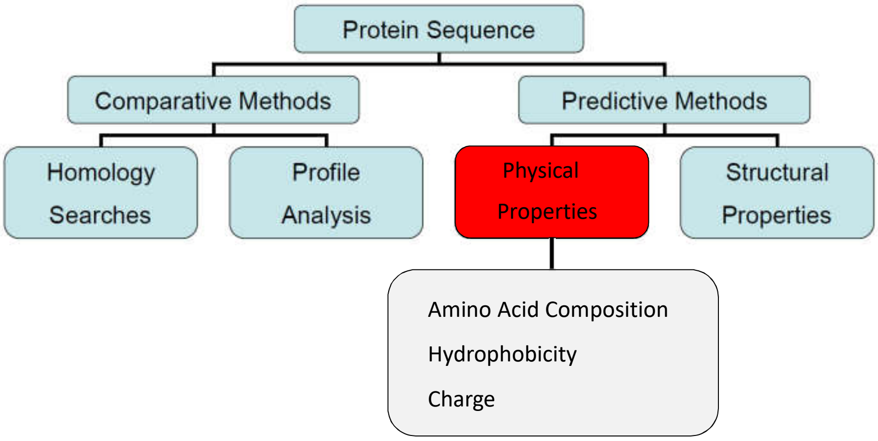
Delete (silent, null) states

Insert states to model highly variable regions in the alignemnt

Main states (gray)



Protein Sequence Analysis



Protein Sequence

Comparative Methods

Predictive Methods

Homology Searches

Profile Analysis

Physical Properties

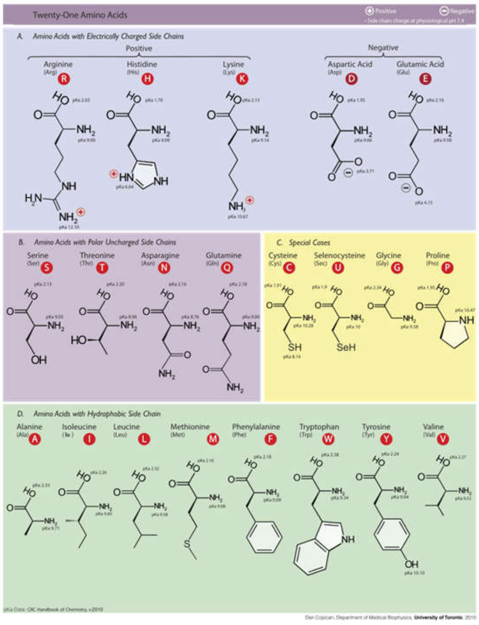
Structural Properties

Amino Acid Composition

Hydrophobicity

Charge

Amino Acids



pKa Data: CRC Handbook of Chemistry, v.2010

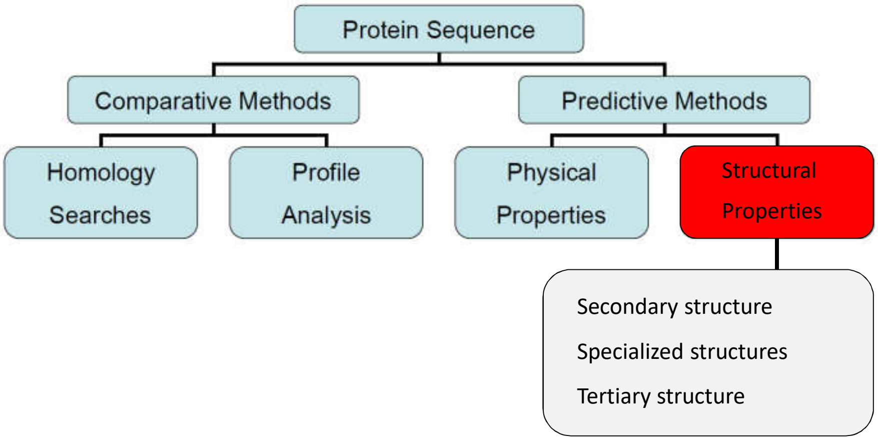
Dan Cojocari, Department of Medical Biophysics, **University of Toronto**, 2010

ProtParam

- **Computes physicochemical parameters**
  - Molecular weight
  - Theoretical pI
  - Amino acid composition
  - Extinction coefficient

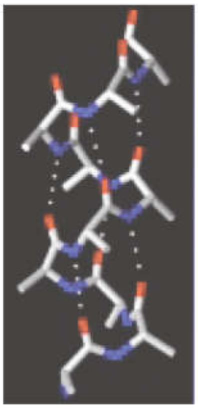
http://web.expasy.org/protparam

Protein Sequence Analysis



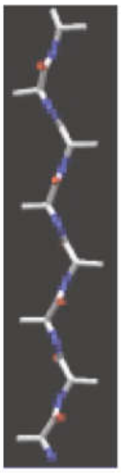
Alpha-helix

- Corkscrew
- Main chain forms backbone, side chains project out
- Hydrogen bonds between CO group at n and NH group at n+4
- Helix-formers: Ala, Glu, Leu, Met
- Helix-breaker: Pro



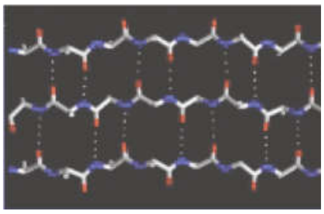
Beta-strand

- Extended structure (“pleated”)
- Peptide bonds point in opposite directions
- Side chains point in opposite directions
- No hydrogen bonding within strand

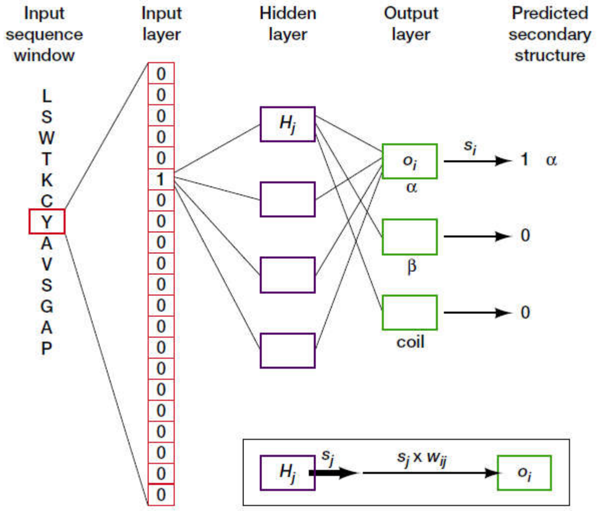


Beta-sheet

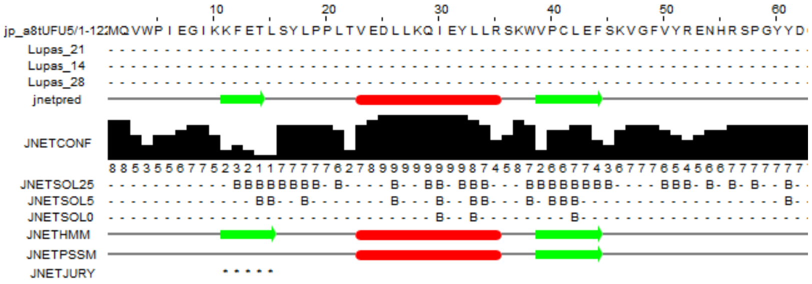
- Stabilization through hydrogen bonding
- Parallel or antiparallel
- Variant: beta-turn



Neuronal network for secondary structure prediction



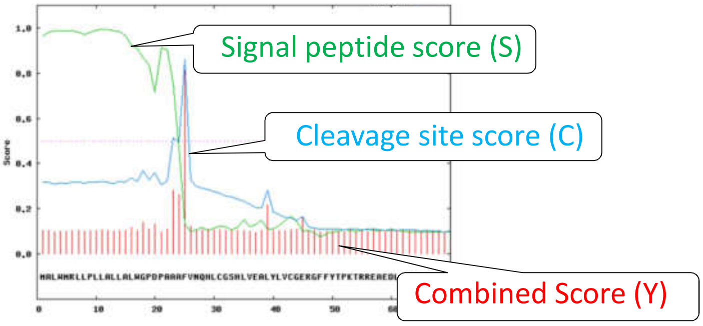
Protein secondary structure prediction (Jpred)



```
                               10         20         30         40         50         60
jp_a8tUFU5/1-12:MQ V W P I E G I K K F E T L S Y L P P L T V E D L L K Q I E Y L L R S K W V P C L E F S K V G F V Y R E N H R S P G Y Y D
    Lupas_21    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
    Lupas_14    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
    Lupas_28    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
    jnetpred

   JNETCONF
                8 8 5 3 5 5 6 7 7 5 2 3 2 1 1 7 7 7 7 7 6 2 7 8 9 9 9 9 9 9 9 9 8 7 4 5 8 7 2 6 6 7 7 4 3 6 7 7 7 6 5 4 5 6 6 7 7 7 7 7 7 7
  JNETSOL25     - - - - - - - - - - - B B B B B B B B - B - - - - B - - B B - B B B - - - B B B B B B B B - - - - B B B - B - B - B - - B - -
   JNETSOL5     - - - - - - - - - - - - - B B - - B - - - - - - - B - - - B - - B B - - - B - B B B - - - - - - - - - - - - - - - - - B - -
   JNETSOL0     - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - B - - B - - - - - - - - B - - - - - - - - - - - - - - - - - - -
   JNETHMM
  JNETPSSM
   JNETJURY                         * * * * *
```

SignalP

- Neural network trained based on phylogeny
  - Gram-negative prokaryotic
  - Gram-positive prokaryotic
  - Eukaryotic
- Predicts secretory signal peptides
- http://www.cbs.dtu.dk/services/SignalP/



Signal peptide score (S)

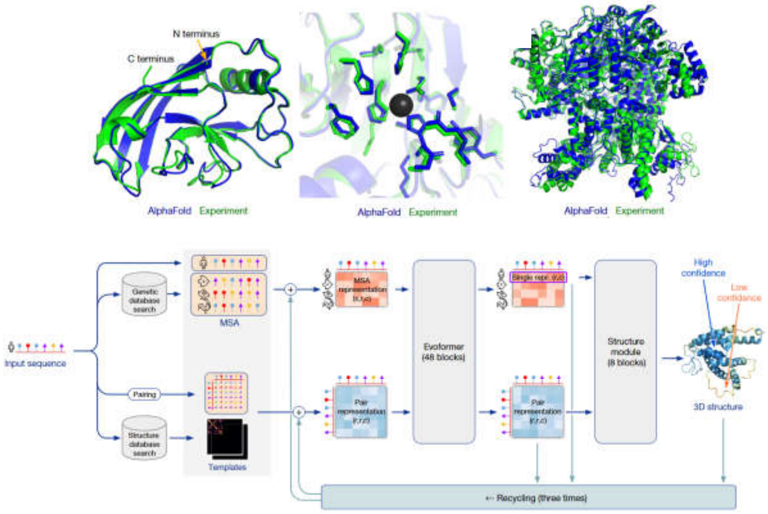
Cleavage site score (C)

Combined Score (Y)

PredictProtein

- Multi-step predictive algorithm (Rost et al., 1994)
  - Protein sequence queried against SWISS-PROT
  - MaxHom used to generate iterative, profile-based multiple sequence alignment (Sander and Schneider,1991)
  - Multiple alignment fed into neural network (PHDsec)
- Accuracy: Average > 70%, Best-case > 90%
- http://www.predictprotein.org/

Protein folding from sequence (AlphaFold2)



Jumper et al. Nature 2021

Regulatory sequences

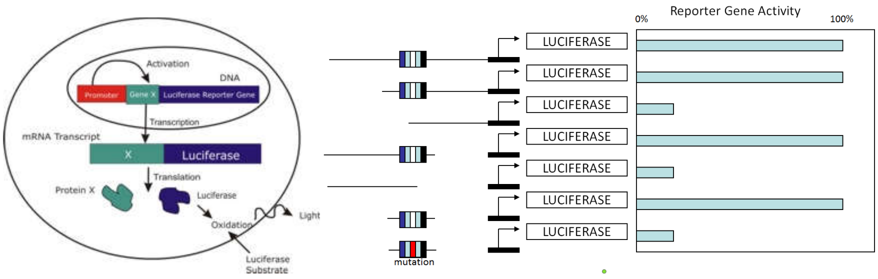
- Transcription factor binding sites
    - Experimental methods
    - Computational methods
        - Matrix based methods
        - Motif discovery
- MicroRNA target prediction

Transcription factor binding sites

Experimental methods

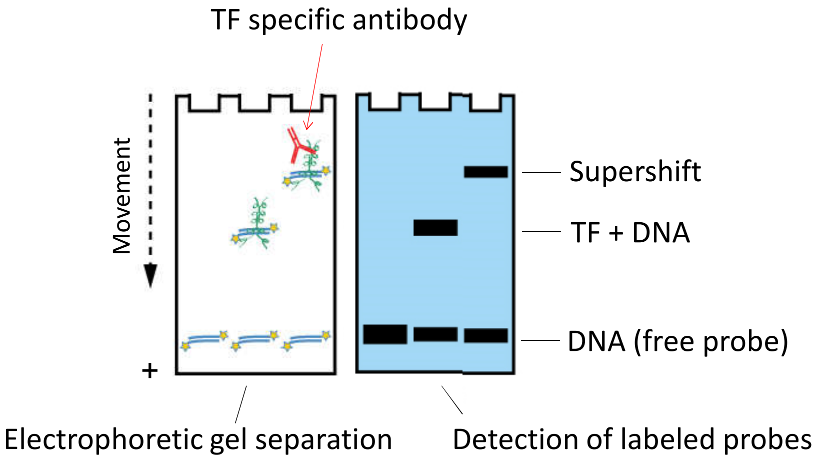
- Reporter gene assays (luciferase)
- Electro mobility shift assays (EMSA)
- DNase I and Exonulease Footprinting
- SELEX
- Chromatin immuno precipitation (ChIP)

Luciferase reporter assays



- Identify functional regulatory region within a sequence and delineate specific TFBS through mutagenesis
- Evidence that TF binding has an effect on transcription (not only binding to DNA)

Electromobility/Gel Shift Assays



TF specific antibody

Movement

+

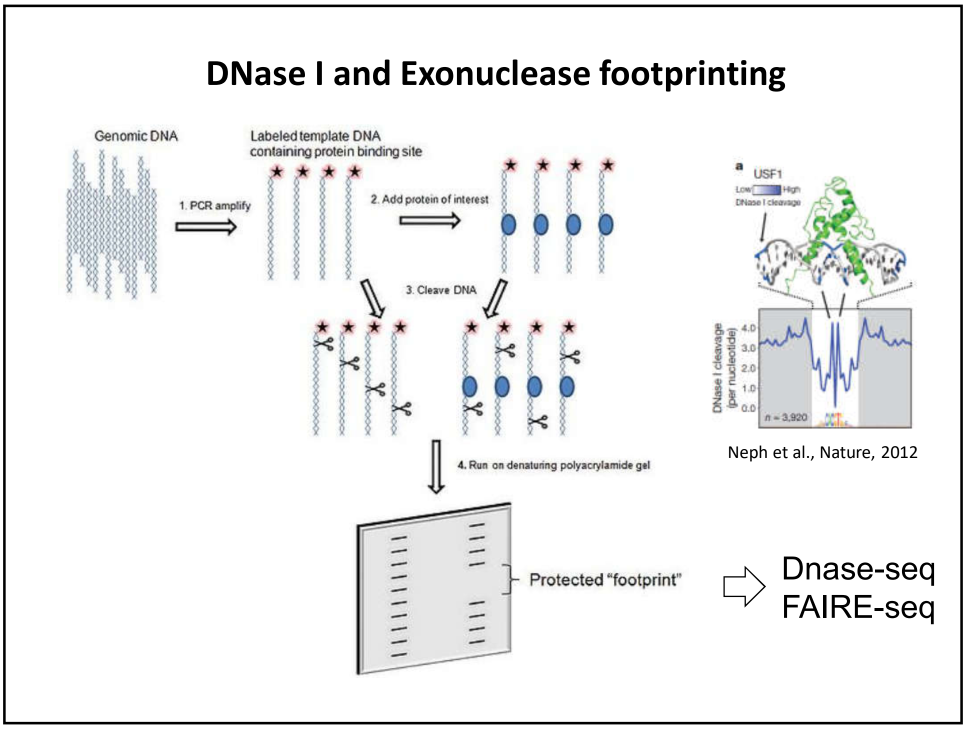
Supershift

TF + DNA

DNA (free probe)

Electrophoretic gel separation

Detection of labeled probes

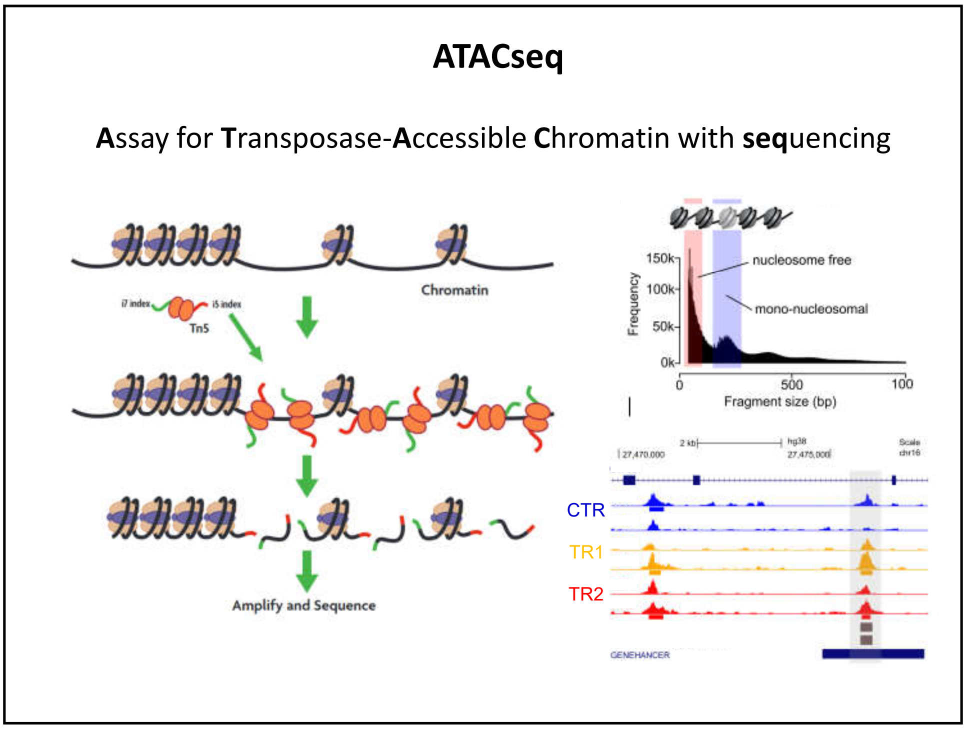


DNase I and Exonuclease footprinting

Genomic DNA
1. PCR amplify
Labeled template DNA containing protein binding site
2. Add protein of interest
3. Cleave DNA
4. Run on denaturing polyacrylamide gel
Protected "footprint"
a
USF1
Low
High
DNase I cleavage
DNase I cleavage (per nucleotide)
4.0
3.0
2.0
1.0
0.0
n = 3,920

Neph et al., Nature, 2012

Dnase-seq
FAIRE-seq

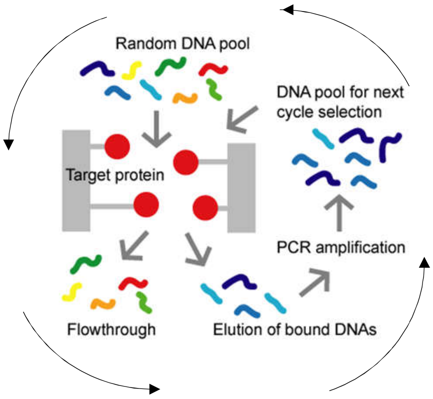


ATACseq

Assay for **T**ransposase-**A**ccessible **C**hromatin with **seq**uencing

SELEX

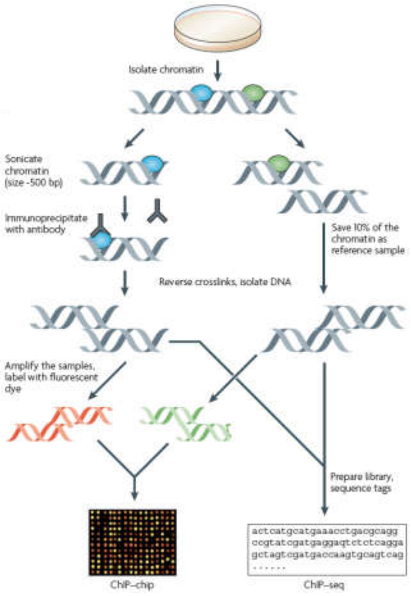
Systematic evolution of ligands by exponential enrichment



several cycles

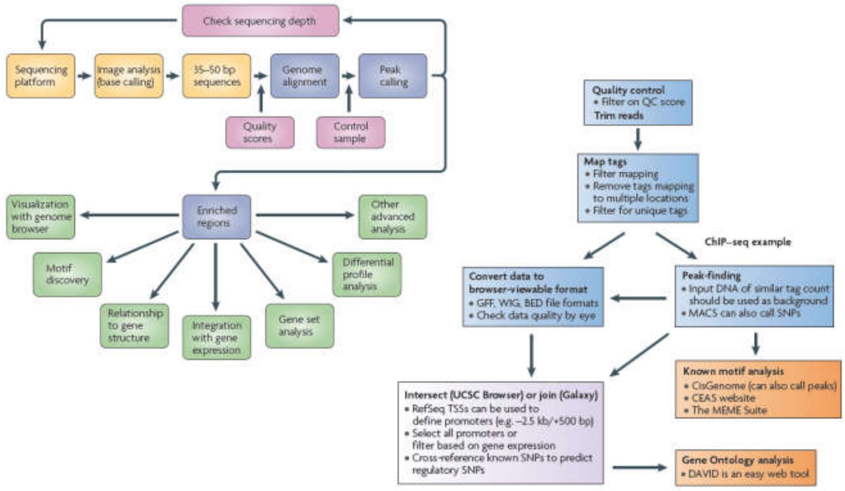
Most position weight matrices (PWMs) in the databases are derived by SELEX

ChIP procedure


Isolate chromatin
Sonicate chromatin (size ~500 bp)
Immunoprecipitate with antibody
Save 10% of the chromatin as reference sample
Reverse crosslinks, isolate DNA
Amplify the samples, label with fluorescent dye
Prepare library, sequence tags
actcatgcatgaaacctgacgcagg
ccgtatcgatgaggagtctctcagga
gctagtcgatgaccaagtgcagtcag
......
ChIP–chip
ChIP–seq

Farnham, Nature Rev Genetics, 2009

ChIP-seq analysis



Check sequencing depth

Sequencing platform

Image analysis (base calling)

35–50 bp sequences

Genome alignment

Peak calling

Quality scores

Control sample

Enriched regions

Visualization with genome browser

Other advanced analysis

Motif discovery

Differential profile analysis

Relationship to gene structure

Integration with gene expression

Gene set analysis

Quality control
- Filter on QC score

Trim reads

Map tags
- Filter mapping
- Remove tags mapping to multiple locations
- Filter for unique tags

ChIP–seq example

Convert data to browser-viewable format
- GFF, WIG, BED file formats
- Check data quality by eye

Peak-finding
- Input DNA of similar tag count should be used as background
- MACS can also call SNPs

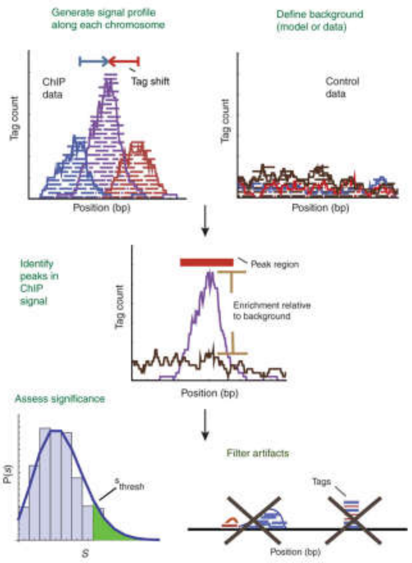
Known motif analysis
- CisGenome (can also call peaks)
- CEAS website
- The MEME Suite

Intersect (UCSC Browser) or join (Galaxy)
- RefSeq TSSs can be used to define promoters (e.g. –2.5 kb/+500 bp)
- Select all promoters or filter based on gene expression
- Cross-reference known SNPs to predict regulatory SNPs

Gene Ontology analysis
- DAVID is an easy web tool

Hawkins et al., Nature Rev Genetics, 2010

ChIP-seq (Peak calling)

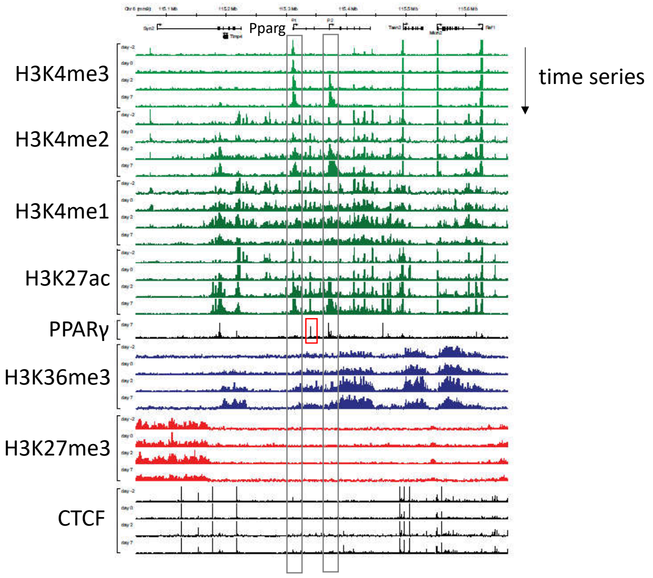


Tools:

- CisGenome
- ERANGE
- FindPeaks
- F-Seq
- GLITR
- MACS
- PeakSeq
- QuEST
- SICER
- SiSSRs
- Spp
- Useq

Pepke, Nature Methods, 2009

Chromatin state and TF localization



Mikkelsen et al., Cell, 2010

Computational methods

- Problem: sequences are short (e.g. 6-10 bp) and degenerated, many false positives
- Matrix based methods (knowledge about TF)
  Position weight matrix (PWM), HMM
- Motif discovery
  Word counting, EM
- MicroRNA target prediction

Experimental verified binding sites

Gene	Organism	5'-3' Sequence	Ref
CYP4A6/P450 IV	rabbit	AACT AGGGCA A AGTTGA	[1]
CYP4A1/P450 IV	rat	AACT AGGGTA A AGTTCA	[2]
L-fatty acid binding protein	rat	ATAT AGGCCA T AGGTCA*	[3]
3-hydroxy-3-methyl-glutaryl-CoA-synthase	rat	AACT GGGCCA A AGGTCT*	[4]
Enoyl-CoA-hydratase	rat	ATGT AGGTAA T AGTTCA*	[1]
Malic enzyme	rat	TTCT GGGTCA A AGTTGA	[5]
Phosphoenolpyruvate carboxikinase	rat	AACT GGGATA A AGGTCT	[6]
Phosphoenolpyruvate carboxikinase)	rat	CCCA CGGCCA A AGGTCA*	[6]

. . . .

Uncoupling protein 1	mouse	AGTG TGGTCA A GGGTGA*	[12]
Apolopoprotein C-III	human	GCGC TGGGCA A AGGTCA*	[1]
Acyl-CoA oxidase	human	TAGA AGGTCA G CTGTCA	[13]
Lipoprotein lipase	human	GTCT GCCCTT T CCCCCT*	[14]
Muscle type carnitine palmitoyltransferase I	human	CCTT TTCCCT A CATTTG	[15]
Consensus		AWCT AGGNCA A AGGTCA	[16]

Position frequency matrix

- Position frequency matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	10	8	4	3	11	0	1	1	2	19	15	17	2	0	0	0	16
C	3	4	11	5	1	1	2	6	15	0	1	4	1	1	2	17	2
G	3	2	4	2	7	20	19	6	1	1	2	1	17	15	1	4	1
T	6	8	3	12	3	1	0	7	4	2	4	0	2	6	19	1	3

- Position weight matrix (PWM),
  position specific scoring matrix (PSSM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	0.86	0.54	-0.46	-0.87	1.00	-1.32	-2.46	-2.32	-1.46	1.79	1.45	1.63	-1.46	-1.32	-1.32	-1.32	1.54
C	-0.87	-0.46	1.00	-0.14	-2.46	-2.46	-1.46	0.26	1.45	-1.32	-2.46	-0.46	-2.46	-2.46	-1.46	1.63	-1.46
G	-0.87	-1.46	-0.46	-1.46	0.35	1.86	1.79	0.26	-2.46	-2.46	-1.46	-2.46	1.63	1.45	-2.46	-0.46	-2.46
T	0.13	0.54	-0.87	1.13	-0.87	-2.46	-1.32	0.49	-0.46	-1.46	-0.46	-1.32	-1.46	0.13	1.79	-2.46	-0.87

Position weight matrix (PWM)

Probability of base b at position i

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \epsilon \{A,C;G,T\}} s(b')}$$

N ... number of sites
s(b) ... pseudo counts
$F_{b,i}$ ... frequency of base b in position i

PWM

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

p(b) ... background probability of base b

Evaluation of sequences

$$S = \sum_{i=1}^{w} W_{b,i}$$

w ... width of PWM
b ... nucleotide in position i
S ... PWM score of a sequence

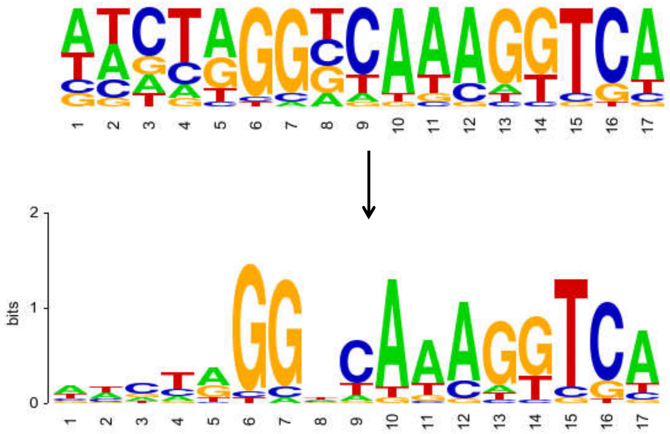
	1	2	3	4	5	6
A	1.00	-1.32	-2.46	-2.32	-1.46	1.79
C	-2.46	-2.46	-1.46	0.26	1.45	-1.32
G	0.35	1.86	1.79	0.26	-2.46	-2.46
T	-0.87	-2.46	-1.32	0.49	-0.46	-1.46

…ACGT$\boxed{\text{AGGTCA}}$TAGAGTA.. S=1+1.86+1.79+0.49+1.45+1.79=8.38

…ACGTAGG$\boxed{\text{TCATAG}}$AGTA.. S=-0.87-2.46-2.46+0.49-1.46-2.46=-9.22

Optimized similarity score to minimize false predictions

From Frequency to Sequence Logo



Information content in position i

$$\boxed{D_i = 2 + \sum_b p(b,i)\log_2 p(b,i) - e(n)}$$

e(n) … correction factor if only few samples n
D_i … information content at position i
b … base A,C,G, or, T

All bases with equal probabilities at position i

$D_i=2+4*0.25*\log_2 0.25=0$ bits

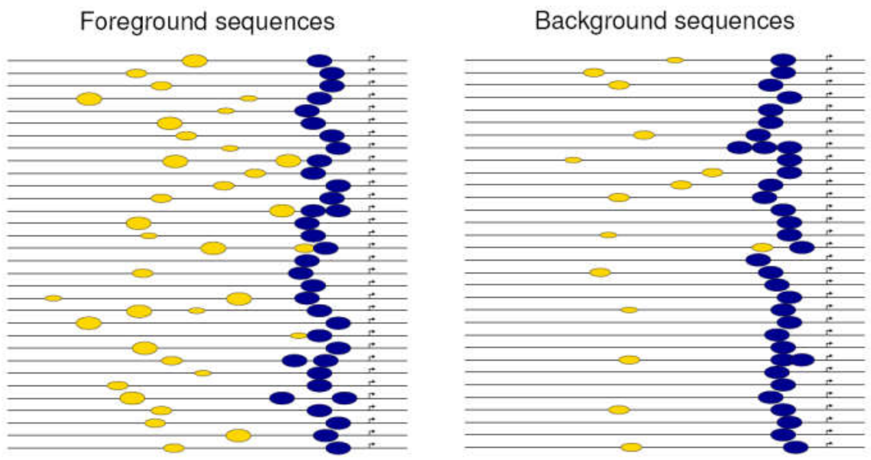
Only one base is present at position i

$D_i=2+1*\log_2 1+3*0.001*\log_2 0.001=1.97$ bits

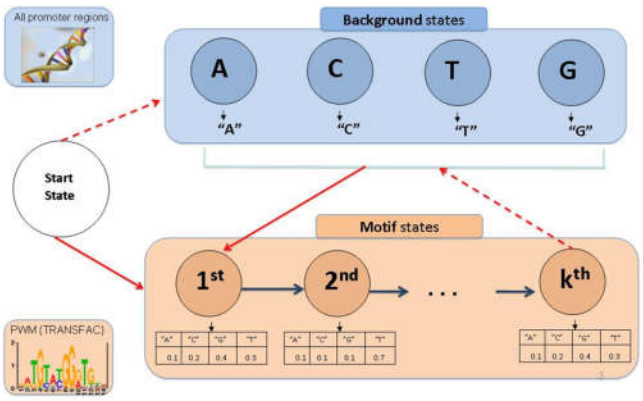
↑

from pseudocounts ($\log_2 0$ is not defined!!)

Using a set of background sequences



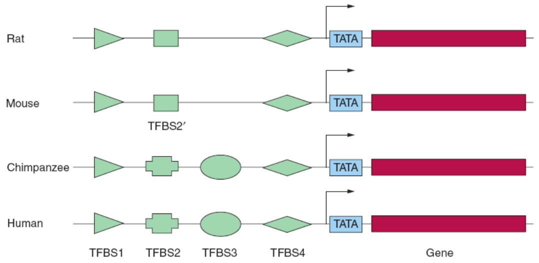
Profile hidden markov models (HMM)



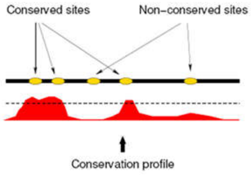
Levkovitz et al. PLoS One. 2010

Phylogenetic footprinting

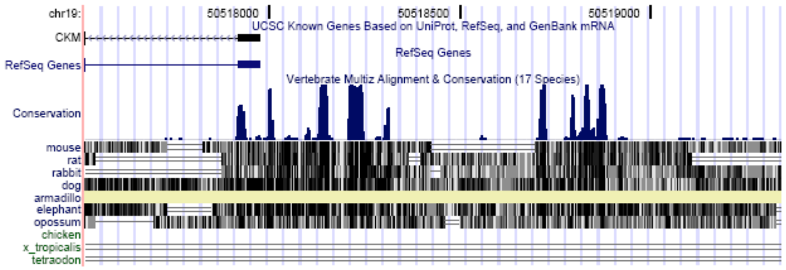
- Functional regulatory sites are conserved between species

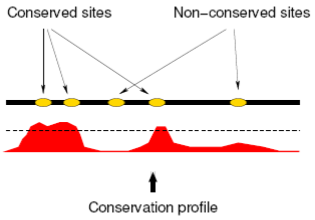


- Multiz alignment of UCSC genome browser

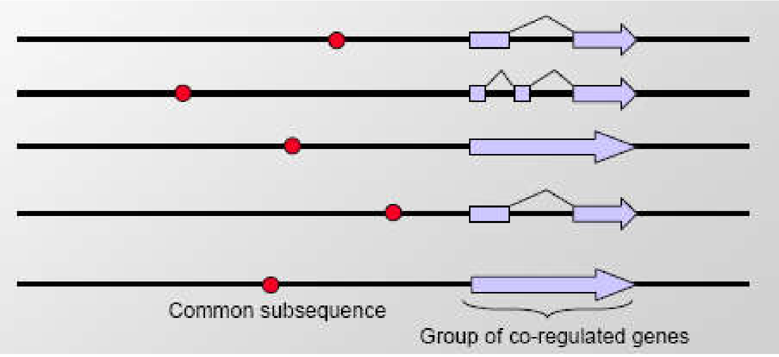


Phylogenetic footprinting





Motif discovery



Word counting

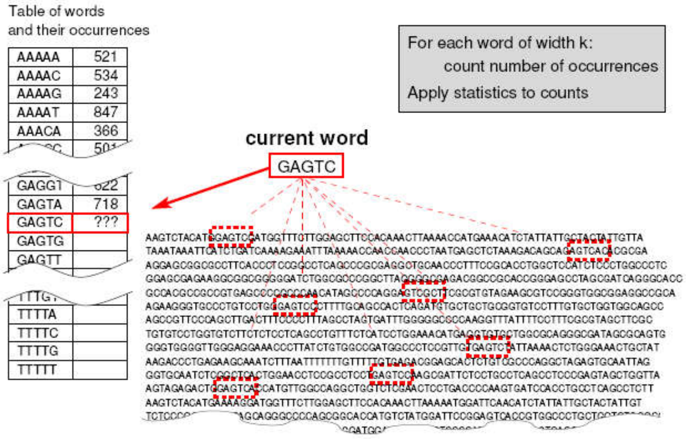


Table of words and their occurrences

AAAAA	521
AAAAC	534
AAAAG	243
AAAAT	847
AAACA	366
AAACC	501
GAGGT	622
GAGTA	718
GAGTC	???
GAGTG	
GAGTT	
TTTGT	
TTTTA	
TTTTC	
TTTTG	
TTTTT	

For each word of width k:
count number of occurrences
Apply statistics to counts

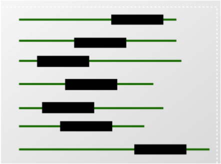
current word

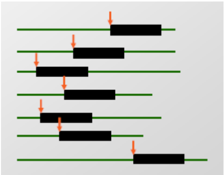
GAGTC

```
AAGTCTACATGGAGTCGATGGTTTCTTGGAGCTTCCACAAACTTAAAACCATGAAACATCTATTATTGCTACTATTGTTA
TAAATAAATTCATCTGATCAAAAGAAATTTAAAAACCAACCAACCCTAATGAGCTCTAAAGACAGCAGAGTCACACGCGA
AGGAGCGGCGCCTTCACCCTCCGGCCTCAGCCCGCGAGGCTGCAACCCTTTCCGCACCTGGCTCCATCTCCCTGGCCCTC
GGAGCGAGAAGGCGGCGGGGGATCTGGCGCCCGGCTTAGGGGCGAGACGGCCGCACCGGGAGCCTAGCGATCAGGGCACC
GCCACGCCGCCGTGAGCCCCGCCCCAACATAGCCCCAGGAGTCGCTTCGCGTGTAGAAGCGTCCGGGTGGCGGAGGCCGCA
AGAAGGGTGCCCTGTCCTGGGAGTCCCTTTTGCAGCCACTCAGATGTGCTGCTGCGGTGTCCTTTGTGCTGGTGGCAGCC
AGCCGTTCCCAGCTTGACTTTCCCCTTTAGCCTAGTGATTTGGGGGCGCCAAGGTTTATTTTCCTTTCGCGTAGCTTCGC
TGTGTCCTGGTGTCTTCTCTCCTCAGCCTGTTTCTCATCCTGGAAACATGAGGTGTGCTGGCGCAGGGCGATAGCGCAGTG
GGGTGGGGTTGGGAGGAAACCCTTATCTGTGGCCGATGGCCCTCCGTTGTGAGTCTATTAAAACTCTGGGAAACTGCTAT
AAGACCCTGAGAAGCAAATCTTTAATTTTTTTGTTTTTGTGAGACGGAGCACTCTGTCGCCCAGGCTAGAGTGCAATTAG
GGTGCAATCTCGGCTCACTGGAACCTCCGCCTCCTGAGTCCAAGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGTTA
AGTAGAGACTGGAGTCACCATGTTGGCCAGGCTGGTCTCGAACTCCTGACCCCAAGTGATCCACCTGCCTCAGCCTCTT
AAGTCTACATGAAAAGGATGGTTTCTTGGAGCTTCCACAAACTTAAAAATGGATTCAACATCTATTATTGCTACTATTGT
TCTCCC      AGCAGGGCCCCAGCGGCACCATGTCTATGGATTCCGGAGTCACCGTGGCCCTGCT
                           CGATGG
```

Expectation maximum

- Problem: Don't know what the motif looks like or where the starting positions are





→ Use expectation maximum (EM)

- EM is a family of algorithms for learning probabilistic models in problems that involve *hidden state*
- In our problem, the hidden state is where the motif starts in each training sequence

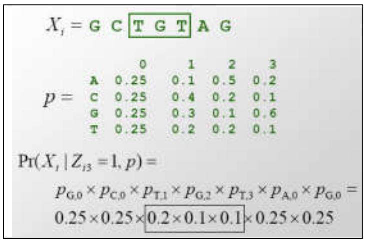
Basic EM-approach

p

A motif is represented by a matrix of probabilities: P_{ck} represents the probability of character *c* in column *k*

Z

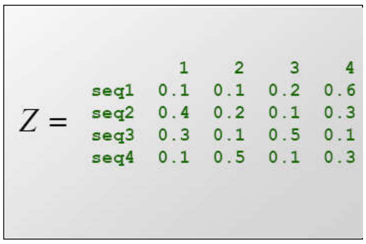
The element Z_{ij} of the matrix Z represents the probability that the motif starts in position *j* in sequence *i.*



$X_i = \texttt{G C} \boxed{\texttt{T G T}} \texttt{A G}$

$$p = \begin{array}{ccccc} & 0 & 1 & 2 & 3 \\ \texttt{A} & 0.25 & 0.1 & 0.5 & 0.2 \\ \texttt{C} & 0.25 & 0.4 & 0.2 & 0.1 \\ \texttt{G} & 0.25 & 0.3 & 0.1 & 0.6 \\ \texttt{T} & 0.25 & 0.2 & 0.2 & 0.1 \end{array}$$

$$\Pr(X_i \mid Z_{i3} = 1, p) =$$
$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =$$
$$0.25 \times 0.25 \times \boxed{0.2 \times 0.1 \times 0.1} \times 0.25 \times 0.25$$

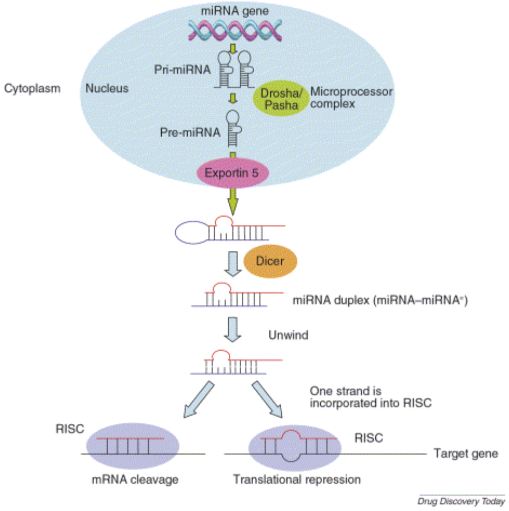


$$Z = \begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ \texttt{seq1} & 0.1 & 0.1 & 0.2 & 0.6 \\ \texttt{seq2} & 0.4 & 0.2 & 0.1 & 0.3 \\ \texttt{seq3} & 0.3 & 0.1 & 0.5 & 0.1 \\ \texttt{seq4} & 0.1 & 0.5 & 0.1 & 0.3 \end{array}$$

– The basic EM approach has been enhanced by MEME (ChIP-MEME)

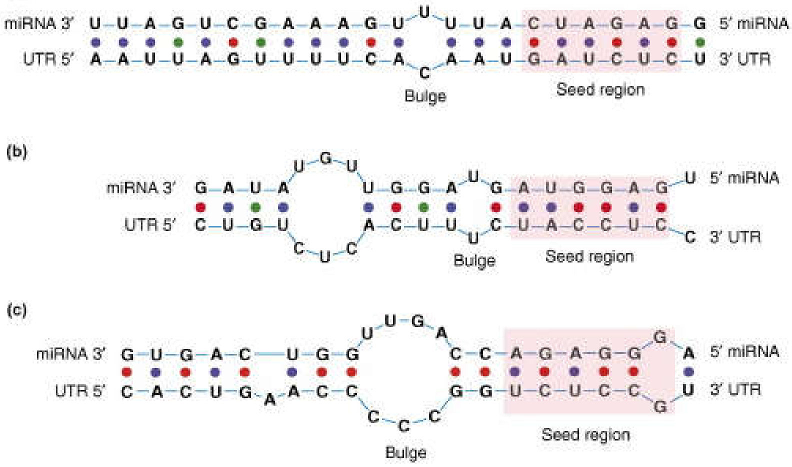
MicroRNA target prediction

microRNA biogenesis



Drug Discovery Today

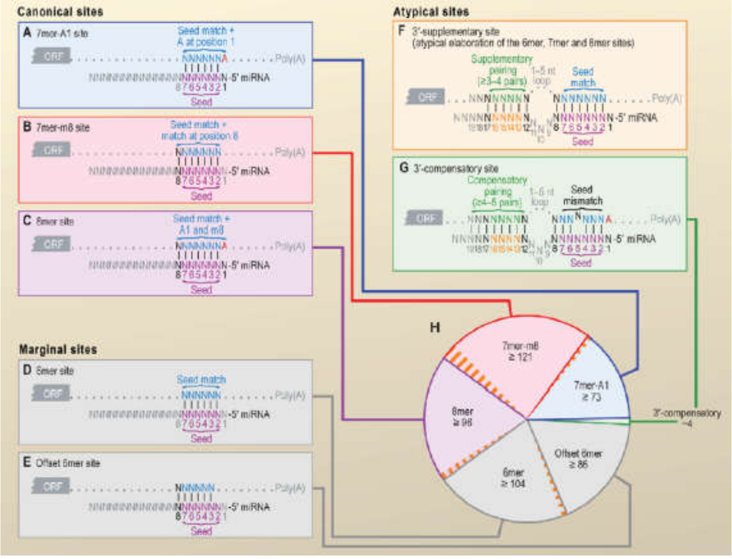
microRNA/mRNA pairing


miRNA 3'
U-U-A-G-U-C-G-A-A-A-G-U U U-U-A-C-U-A-G-A-G-G
5' miRNA
UTR 5'
A-A-U-U-A-G-U-U-U-U-C-A C A-A-U-G-A-U-C-U-C-U
3' UTR
Bulge
Seed region
(b)
miRNA 3'
G-A-U-A U-G-U U-G-G-A U G-A-U-G-G-A-G-U
5' miRNA
UTR 5'
C-U-G-U C-U-C A-C-U-U U C-U-A-C-C-U-C-C
3' UTR
Bulge
Seed region
(c)
miRNA 3'
G-U-G-A-C-U-G-G U-U-G-A C-C-A-G-A-G-G G A
5' miRNA
UTR 5'
C-A-C-U-G A A-C-C C-C-C G-G-U-C-U-C-C G U
3' UTR
Bulge
Seed region

Principles of microRNA target prediction

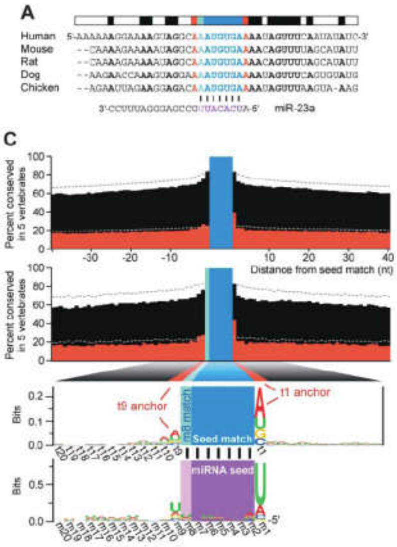
1. Sequence complementarity
2. Conservation
3. Thermodynamics
4. Site accessibility
5. UTR Context
6. Anticorrelation of expression profiles

Sequence complementarity



Bartel, Cell,2009

Conservation



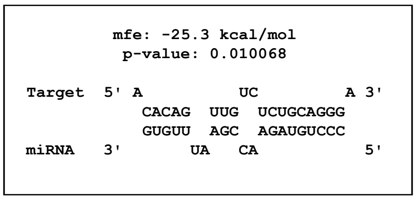
Lewis BP et al., Cell, 2003

Thermodynamics

1. Minimum free energy

⇨ e

Mfold (Zuker et al.)
RNAfold (Hofacker et al.)



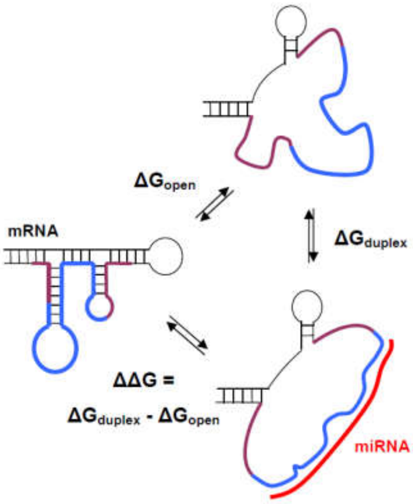
```
         mfe: -25.3 kcal/mol
          p-value: 0.010068

Target  5' A          UC          A 3'
            CACAG  UUG  UCUGCAGGG
            GUGUU  AGC  AGAUGUCCC
miRNA   3'      UA   CA           5'
```

2. Account for different sequence length

3. Extreme value distribution of MFE

Rehmsmeier M et al. RNA (2004)

Site accessibility



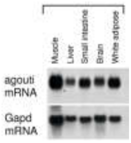
Leitner A, 2009

III Gene expression analyses

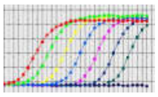
– Microarrays
– RNA sequencing
– Gene expression profiling
– Clustering and classification
– Gene ontology

Gene expression analyes

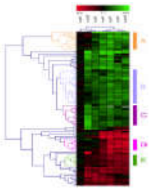
- Northern bloting
  - semi-quantitative
  - few genes



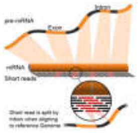
- Real time RT-PCR (qPCR)
  - medium throughput
  - 96/384 per run



- Microarray analysis
  - high throughput
  - 10.000-500.000 elements per chip

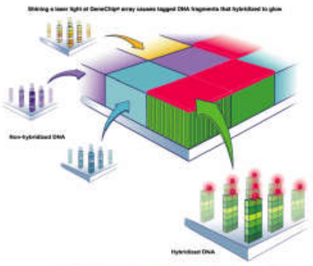


- RNA seq
  - high throughput
  - deep sequencing (short reads 25bp)

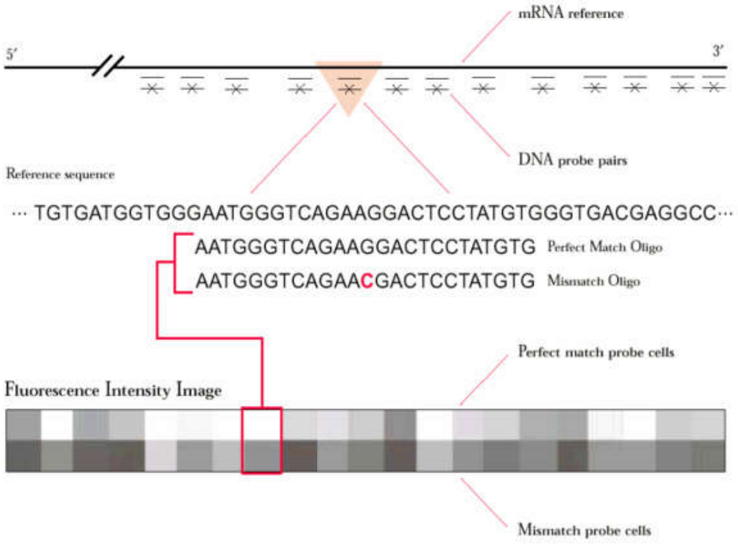


pre-mRNA
Exon
Intron
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

One color microarrays (Affymetrix)


Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

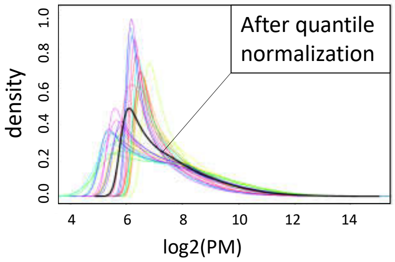
Affymetrix chips


mRNA reference
5′
3′
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG Mismatch Oligo
Fluorescence Intensity Image
Perfect match probe cells
Mismatch probe cells

Processing of Affymetrix chips

Robust Microarray Averaging (R/Bioconductor pkg. RMA)

- Background modeling (PM vs. MM)
- Quantile normalization across all arrays



- Probe summarization (median polish)
- Log2-transformation (log2-intensities)

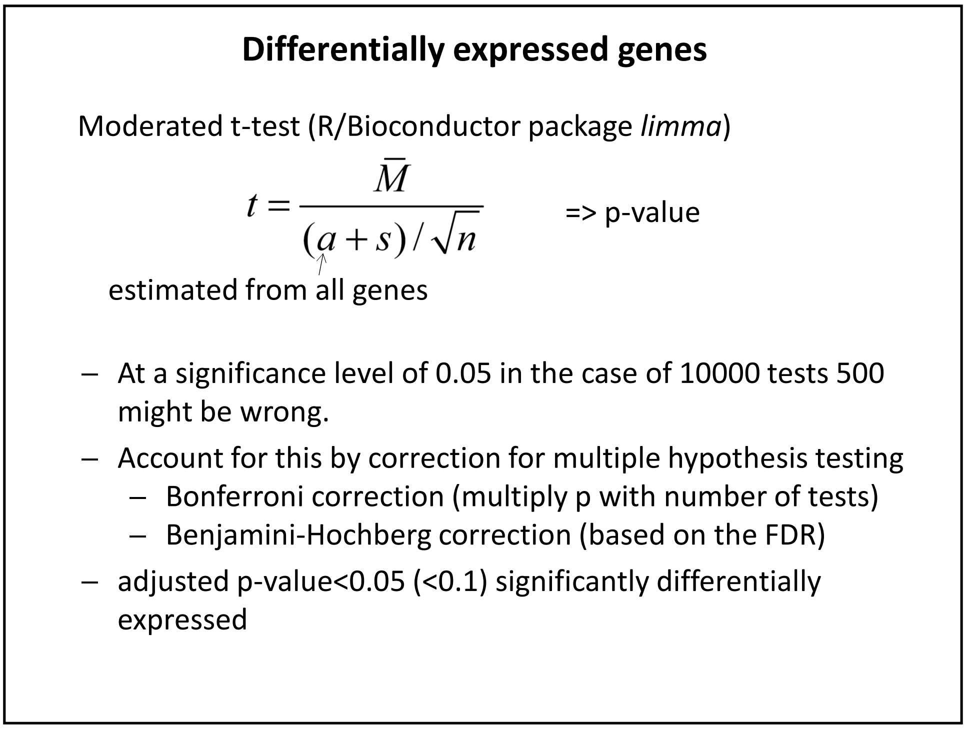
Differentially expressed genes

test

16134 probesets

ID	GENE	KO1	KO2	KO3	WT1	WT2	WT3	logFC	AveExpr	t	P.Value	adj.P.Val
10386473	Srebf1	5.72	5.58	6.06	4.91	4.88	5.09	0.83	5.33	7.66	3.7E-09	4.6E-05
10463355	Scd2	6.63	6.26	6.92	5.13	4.77	5.01	1.64	5.59	7.52	5.6E-09	4.6E-05
10548105	Ccnd2	5.56	5.48	5.49	5.05	5.11	5.02	0.45	5.23	5.21	7.3E-06	3.9E-02
10587284	Elovl5	5.81	5.67	5.97	5.05	5.06	5.35	0.66	5.44	4.87	2.1E-05	8.4E-02
10540122	Slc6a6	7.27	7.16	7.35	6.75	6.81	6.71	0.50	7.04	4.80	2.6E-05	8.5E-02
10605437	Pls3	5.50	5.63	5.41	4.88	4.93	4.87	0.62	5.20	4.63	4.3E-05	9.7E-02
10543791	Podxl	7.30	7.03	7.08	6.31	6.52	6.33	0.75	6.59	4.61	4.6E-05	9.7E-02
10356084	Irs1	8.30	8.76	7.61	6.62	7.33	7.19	1.18	7.60	4.57	5.2E-05	9.7E-02
10346164	Sdpr	5.68	5.37	5.43	5.00	5.03	4.95	0.50	5.17	4.54	5.7E-05	9.7E-02
10387625	Chrnb1	6.31	6.08	6.06	5.73	5.59	5.81	0.44	6.01	4.52	6.0E-05	9.7E-02
10407390	Ptbp1	4.84	5.26	5.07	4.22	3.98	4.64	0.77	4.88	4.43	8.0E-05	1.1E-01
10507539	Elovl1	5.08	4.58	4.89	4.33	4.34	4.55	0.44	4.61	4.40	8.7E-05	1.1E-01
10585988	Myo9a	4.05	4.00	4.01	3.50	3.64	3.79	0.38	3.93	4.39	9.1E-05	1.1E-01
10371959	Elk3	5.94	5.85	5.78	5.28	5.44	5.46	0.47	5.66	4.38	9.3E-05	1.1E-01

condition KO vs. condition WT



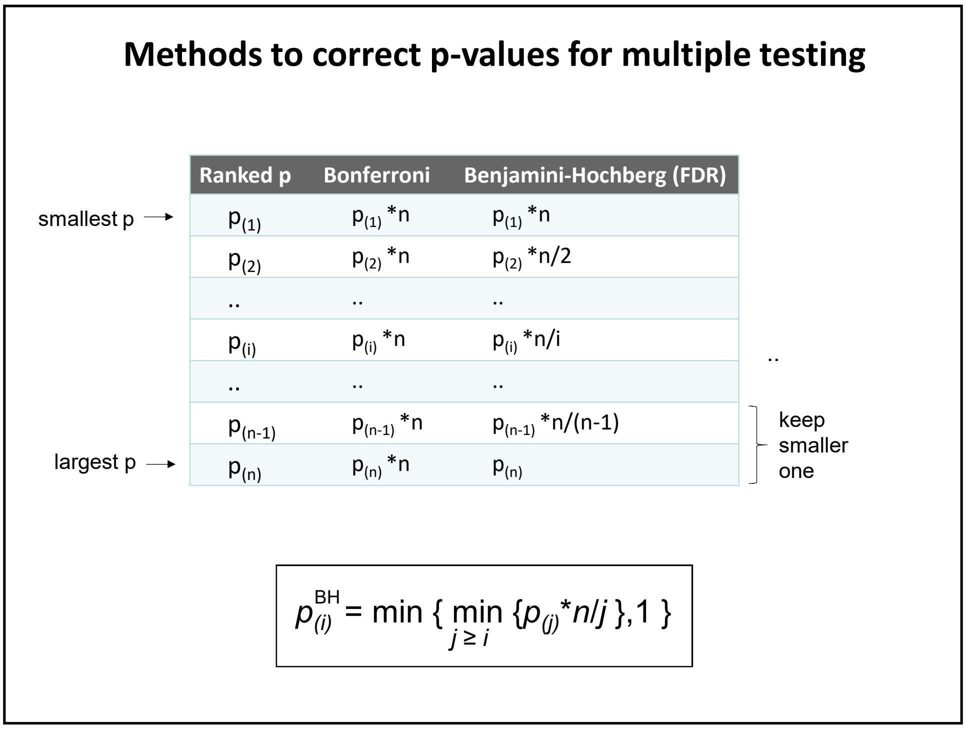
Differentially expressed genes

Moderated t-test (R/Bioconductor package *limma*)

$$t = \frac{\bar{M}}{(\underset{\uparrow}{a} + s)/\sqrt{n}} \qquad \Rightarrow \text{p-value}$$

estimated from all genes

- At a significance level of 0.05 in the case of 10000 tests 500 might be wrong.
- Account for this by correction for multiple hypothesis testing
  - Bonferroni correction (multiply p with number of tests)
  - Benjamini-Hochberg correction (based on the FDR)
- adjusted p-value<0.05 (<0.1) significantly differentially expressed

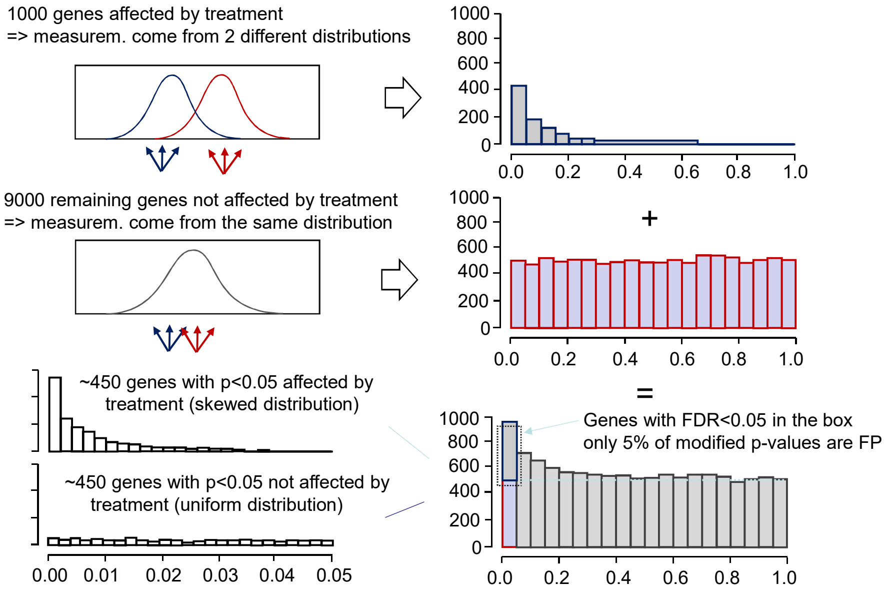


Methods to correct p-values for multiple testing

	Ranked p	Bonferroni	Benjamini-Hochberg (FDR)	
smallest p →	$p_{(1)}$	$p_{(1)}$ *n	$p_{(1)}$ *n	
	$p_{(2)}$	$p_{(2)}$ *n	$p_{(2)}$ *n/2	
	..	..	..	
	$p_{(i)}$	$p_{(i)}$ *n	$p_{(i)}$ *n/i	..
	..	..	..	
	$p_{(n-1)}$	$p_{(n-1)}$ *n	$p_{(n-1)}$ *n/(n-1)	keep smaller one
largest p →	$p_{(n)}$	$p_{(n)}$ *n	$p_{(n)}$	

$$\boxed{p^{BH}_{(i)} = \min\{\min_{j \geq i}\{p_{(j)}*n/j\},1\}}$$

P-value distribution



1000 genes affected by treatment
=> measurem. come from 2 different distributions

9000 remaining genes not affected by treatment
=> measurem. come from the same distribution

~450 genes with p<0.05 affected by treatment (skewed distribution)

~450 genes with p<0.05 not affected by treatment (uniform distribution)

Genes with FDR<0.05 in the box
only 5% of modified p-values are FP

Josh Starmer (StatQuest)

Deep (next generation) sequencing technologies

- Sanger (Thermo Fisher Scientific) — 1st gen.

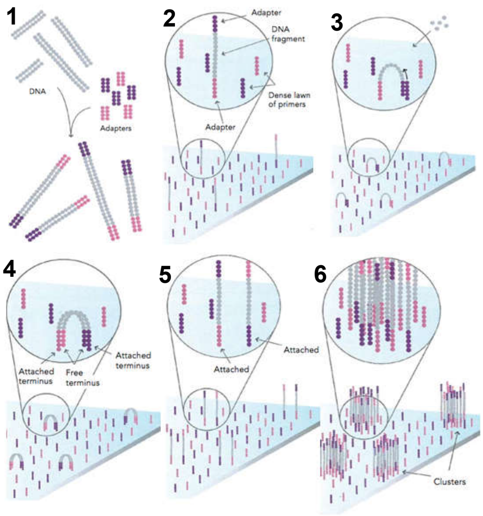
- 454 (Roche)
- Solexa (Illumina)
- Solid (Thermo Fisher Scientific)
- Ion Torrent (Thermo Fisher Scientific)

2nd gen.
(ampl)

- HeliScope (Helicos)
- Pacific Biosciences SMRT
- Oxford Nanopore Sequencing (MinION)

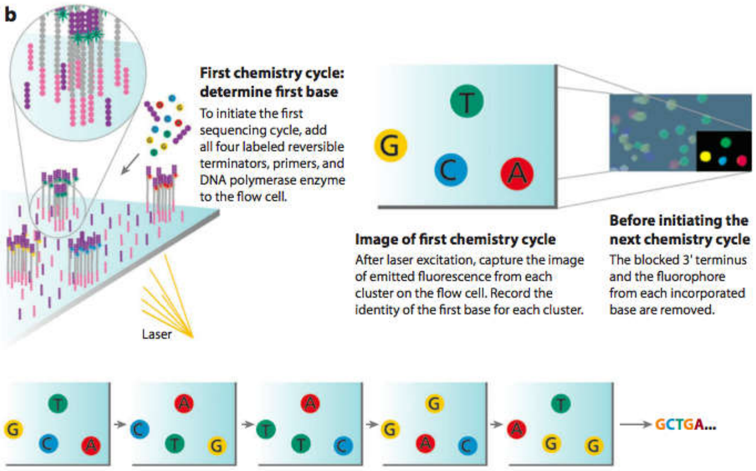
3rd gen.
(no ampl)

Solexa (Illumina)



1. Prepare genomic DNA sample
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature double stranded DNA
6. Complete amplification

Solexa (Illumina)



First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

Image of first chemistry cycle

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle

The blocked 3' terminus and the fluorophore from each incorporated base are removed.

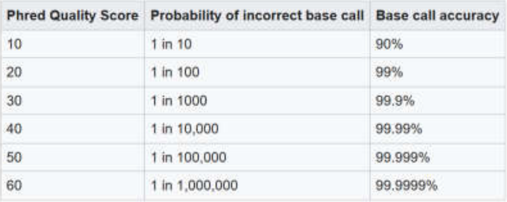
Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

Base calling (Phred score)

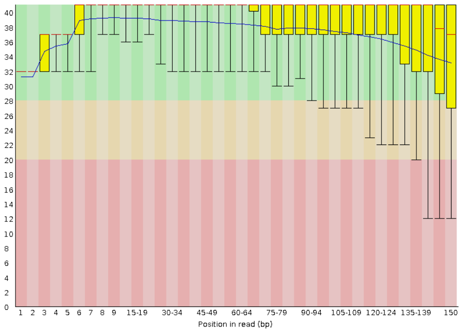
Base-calling error probabilities: P

Q=-10* log P



Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality of Sequencing (FASTQC)



Base calling (FastQ format)

Definition: <fastq> := <block>+

<block> := @<seqname> \n <seq> \n \+<seqname>? \n <qual> \n

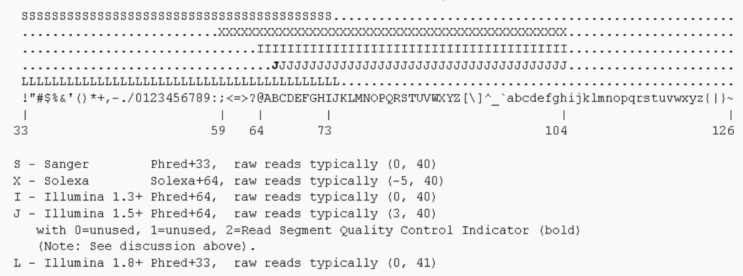
<seqname> := [A-Za-z0-9_.:-]+

<seq> := [A-Za-z\n\.~]+

<qual> := [!-~\n]+

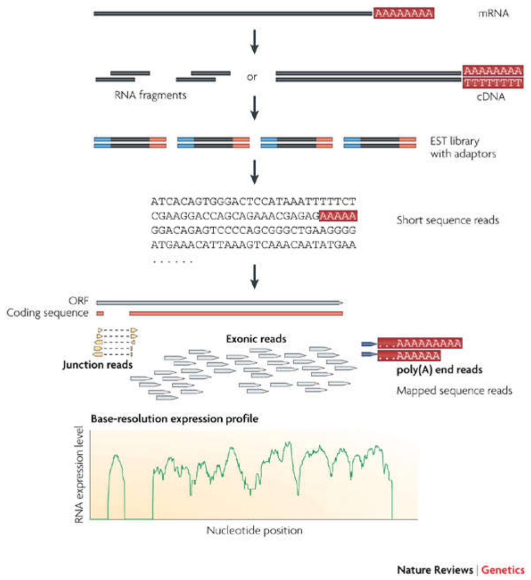
```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

Quality scores are encoded in ASCII



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Transcriptome sequencing (RNAseq)



Wang et al., Nature Rev Gen, 2009

Analysis steps

0. Image analysis and base calling (Phred quality score)

=> FastQ files (sequence and corresponding quality levels)

1. Trimming adaptors and low quality reads (FastQC, Trimmomatic)
2. Read mapping (Spliced alignment) (STAR)

=> SAM/BAM files

3. Transcriptome reconstruction (reference transcriptome, GTF file)

4. Expression quantification (transcript isoforms) (featureCounts)

=> raw count matrix

5. Differential expression analysis (negative-binomial test) (DESeq, edgeR)

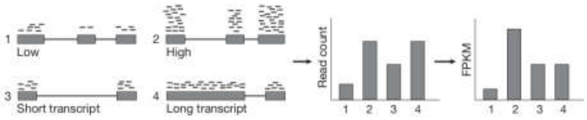
=> List of genes with log2FC, p-value, FDR, average expression

6. Normalization

Normalization

Within-samples

- Reads per kilobase per million reads (RPKM)
- Fragments per kilobase per million (FPKM) for paired-end seq.



- TPM (transcripts per million) (preferable)

Between-samples

- Quantile normalization (upper quantile normalization)
- TMM (trimmed mean of M values) (edgeR)
- Relative log expression (RLE) (DESeq2)

RPKM (FPKM)

GENE	S1	S2	S3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Tens(Mio)	3.5	4.5	10.6

1. Divide by millions of reads

RPM

A (2kb)	2.86	2.61	2.83
B (4kb)	5.71	5.43	5.66
C (1kb)	1.43	1.96	1.42
D (10kb)	0.00	0.00	0.09

2. Divide by gene length in kb

RPKM

A (2kb)	1.43	1.30	1.42
B (3kb)	1.43	1.36	1.42
C (1kb)	1.43	1.96	1.42
D (10kb)	0.00	0.00	0.01

TPM

GENE	S1	S2	S3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

1. Divide by gene length in kb

RPK

A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

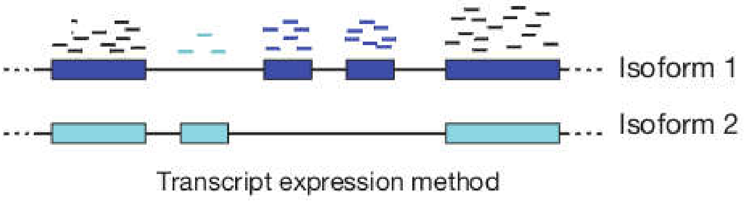
Tens(Mio)	1.5	2.025	4.51

2. Divide by millions of RPK

TPM

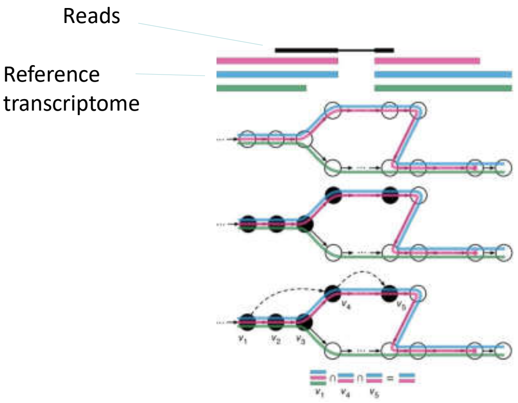
A (2kb)	3.33	2.96	3.326
B (3kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

Isoform quantification



- Uncertainy in assigning reads to isoforms
- Paired-end sequencing
- Spliced alignment
- Alternative splicing (statistical significant?)

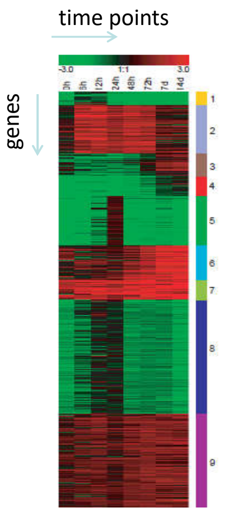
RNA seq quantification using pseudoalignment (kallisto)

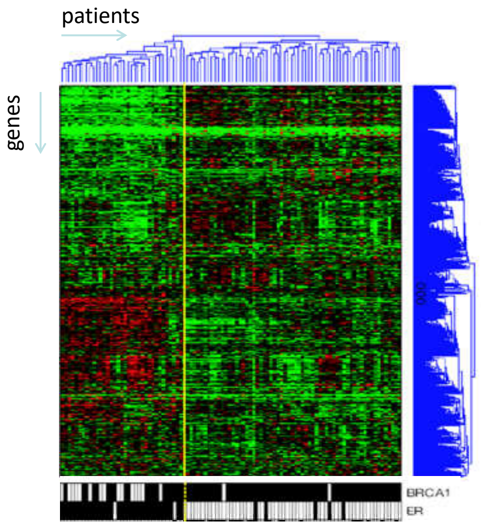


Transcriptome de Bruijn Graph (T-DBG) where nodes (v1, v2, v3, ...) are *k*-mers

Bray et al. Nature Biotechnology 2016

Gene expression profiling



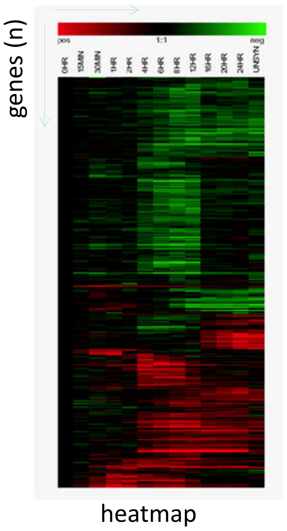


cell development

cancer

Representation of gene expression

conditions (patients) (m)



heatmap

n x *m* matrix with *n* genes and *m* samples

- Representation as heatmap (e.g. *red* upregulated genes, green down regulated genes, *black* no change)

For experiments in reference design:

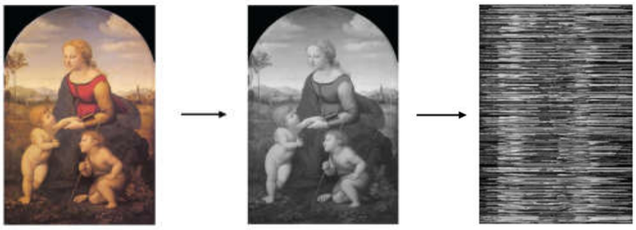
- log2-fold change (log2FC, log2(A/B), log2 ratio)

For patient samples and no reference:

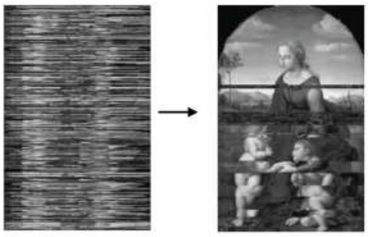
- Mean (median) centered log2-levels for each gene
  log2-intensities for one-color arrays
  log2-RPKM for RNAseq
- z-score of log2-levels
  Z= (X-m)/s   m…mean,
  s…standard deviation

Organize data

grayscale random order slices



clustering algorithms (hierarchical clustering)



Sherlock G, Kishan M, Narisamhan S

Clustering

- Unsupervized clustering
  - Hierarchichal Clustering
  - K-Means Clustering
  - Principal Component Analysis (PCA)

- Supervized clustering (Classification)
  - Support vector machines (SVM)
  - Logistic regression
  - Cross validation

Clustering

- Agglomerative

  Bottom up approach, whereby single expression profiles are successively joined to form nodes.

- Divisive

  Top down approach, each cluster is successively split in the same fashion, until each cluster consists of one single profile.

Similarity (distance) between expression profiles

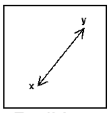
- Pearson correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$-1 \leq r \leq 1$

- Euclidian distance

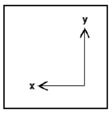
$$d_E = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



Euclidean

- Manhattan distance

$$d_M = (\sum_{i=1}^{n} |x_i - y_i|)$$

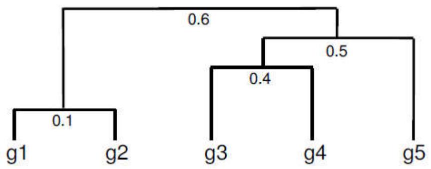


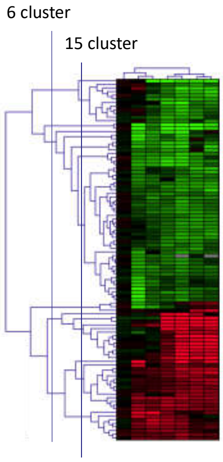
Manhattan

Hierarchical clustering

- Agglomerative (bottom up), unsupervized
- Cluster genes or samples (or both= biclustering)
- Distances are encoded in dendogram (tree)
- Cut tree to get clusters
- Pearson correlation (usually used)
- Computational intensive (correlation matrix)

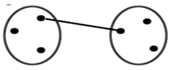
1. Identify clusters (items) with closest distance
2. Join to new clusters
3. Compute distance between clusters (items) (see linkage)
4. Return to step 1



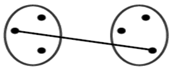


Linkage

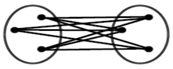
- Single-linkage clustering
  Minimal distance



- Complete-linkage clustering
  Maximal distance



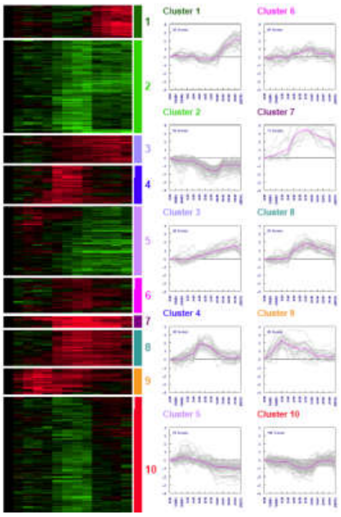
- Average-linkage clustering
  Calculated using average distance (UPGMA)
  Average from distances not! expression values



K-means

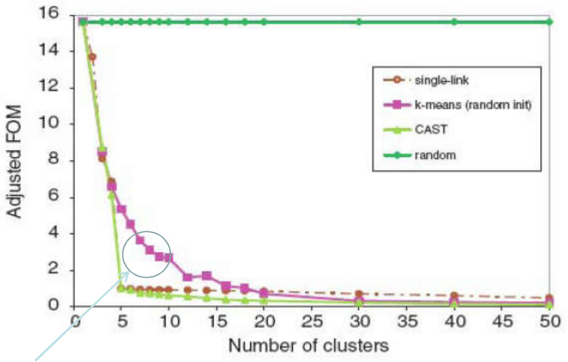
- partition *n* genes into *k* clusters, where *k* has to be predetermined
- k-means clustering minimizes the variability within and maximize between clusters
- Moderate memory and time consumption

1. Generate random points (“cluster centers”) in n dimensions (results are depending on these seeds).
2. Compute distance of each data point to each of the cluster centers.
3. Assign each data point to the closest cluster center.
4. Compute new cluster center position as average of points assigned.
5. Loop to (2), stop when cluster centers do not move very much.

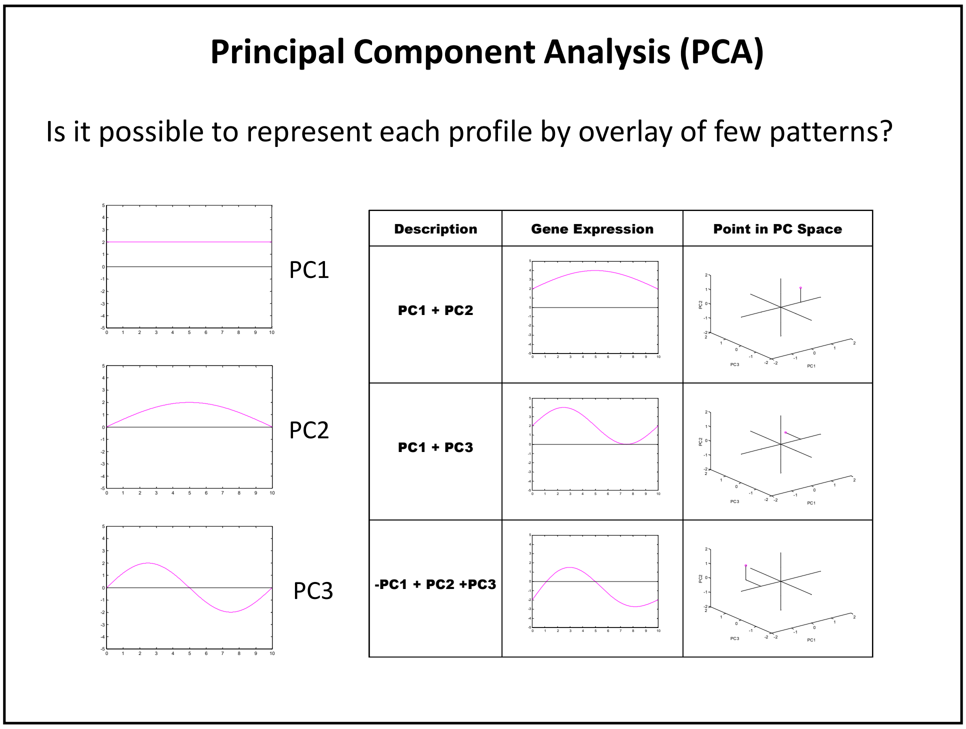


How to choose k

Figure of Merit (FOM)



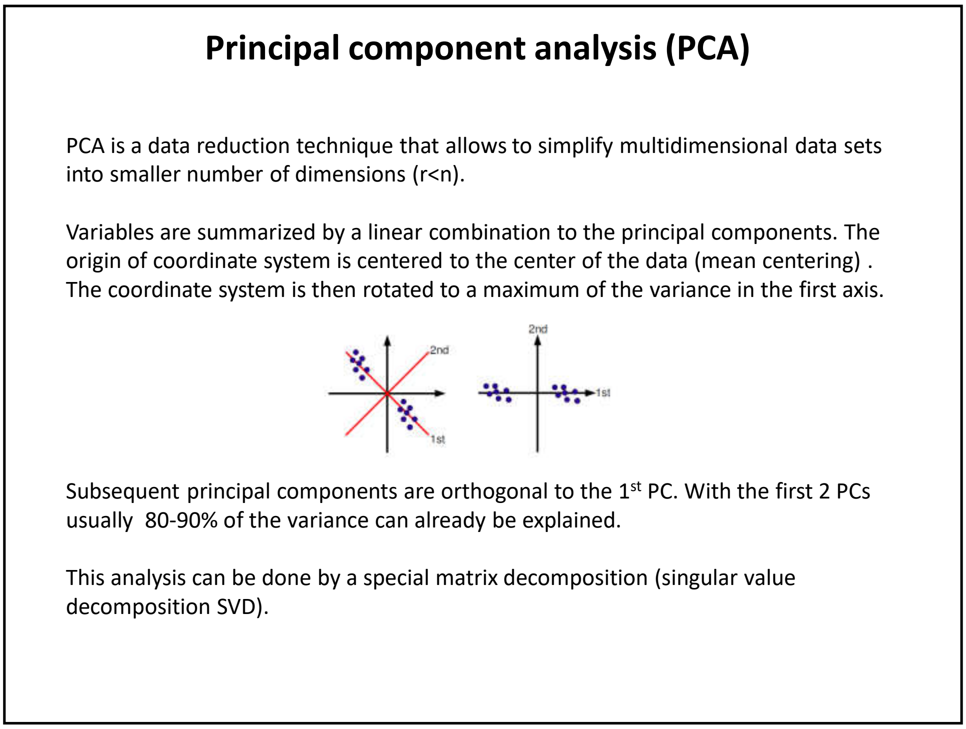
choose k here (e.g. k=8)



Principal Component Analysis (PCA)

Is it possible to represent each profile by overlay of few patterns?

Description	Gene Expression	Point in PC Space
PC1 + PC2		
PC1 + PC3		
-PC1 + PC2 +PC3		



Principal component analysis (PCA)

PCA is a data reduction technique that allows to simplify multidimensional data sets into smaller number of dimensions (r<n).

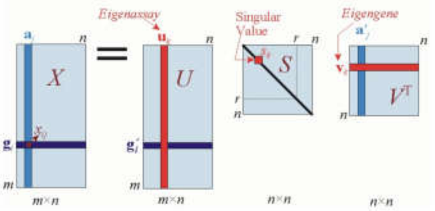
Variables are summarized by a linear combination to the principal components. The origin of coordinate system is centered to the center of the data (mean centering) . The coordinate system is then rotated to a maximum of the variance in the first axis.

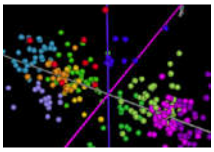
Subsequent principal components are orthogonal to the 1st PC. With the first 2 PCs usually 80-90% of the variance can already be explained.

This analysis can be done by a special matrix decomposition (singular value decomposition SVD).

Singular value decomposition (SVD)

$X = USV^T$ *with* $UU^T = V^TV = VV^T = I$



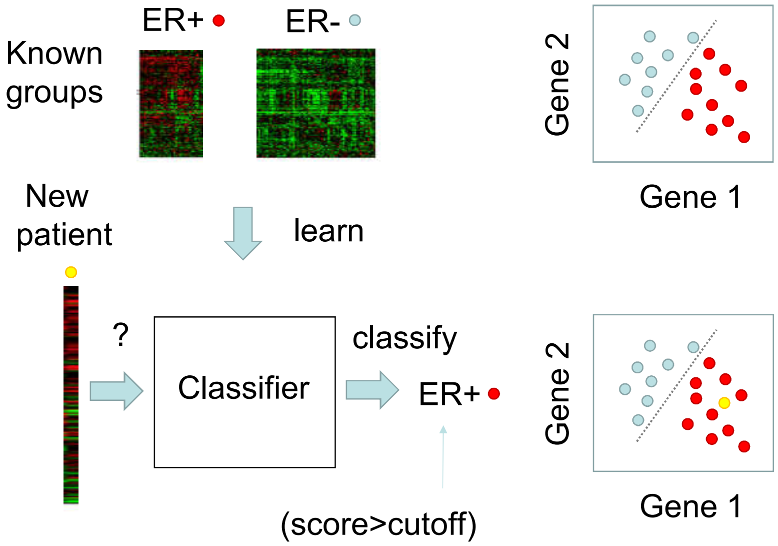


For mean centered data the Covariance matrix *C* can be calculated by XX^T. *U* are eigenvectors of XX^T and the eigenvalues are in the diagonal of *S* defined by the characteristic equation $|C - \lambda I| = 0$.

Transformation of the input vectors into the principal component space can be described by $Y = XU$ where the projection of sample *i* along the axis is defined by the *j*-th PC:

$$y_{ij} = \sum_{t=1}^{m} x_{it} u_{tj}$$

Classification



Known groups

ER+ ●

ER- ○

Gene 2

Gene 1

New patient

learn

?

Classifier

classify

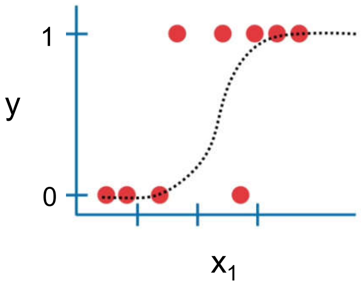
ER+ ●

(score>cutoff)

Gene 2

Gene 1

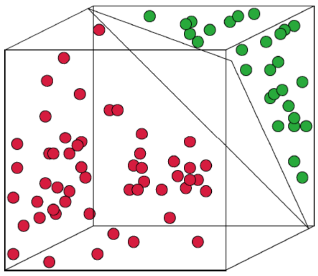
Logistic regression

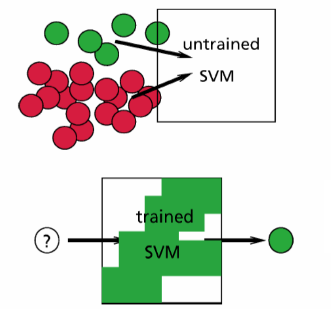


$\ln (P/(1-P))=b_0+b_1*x_1+b_2*x_2+\dots$

- Binary outcome (y)
- With logit transformation analog to linear regression

Support vector machines (SVM)

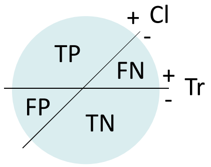




A SVM tries to find an optimal hyperplane that separates all training samples correctly and maximizes the margin (maximizes the distance between it and the nearest data point of each class). If this is not possible in the input space (for example in 2 dimensions) a hyperplane can be found in the higher dimensional feature space (e.g. 3D-space)

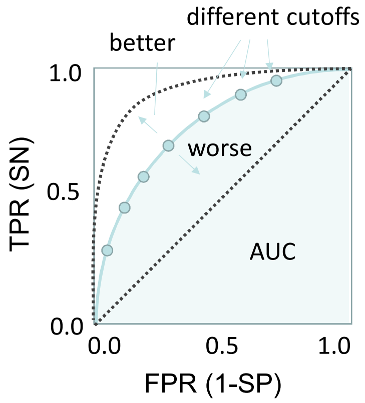
Receiver operator characteristics (ROC)

		Truely	
		ER+	ER-
Classified (> cutoff)	ER+	TP	FP
	ER-	FN	TN



Sensitivity
SN=TP/(TP+FN)

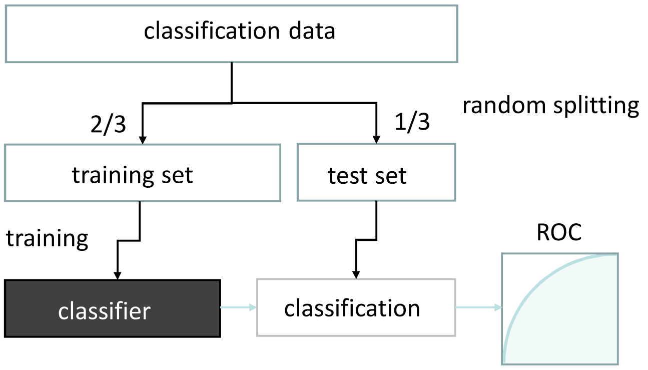
Specificty
SN=TN/(TN+FP)



Area under curve (AUC)
AUC=1.0 optimal
AUC=0.5 random

Holdback cross validation

To avoid overfitting data should be splitted into training and test set



classification data

random splitting

2/3

1/3

training set

test set

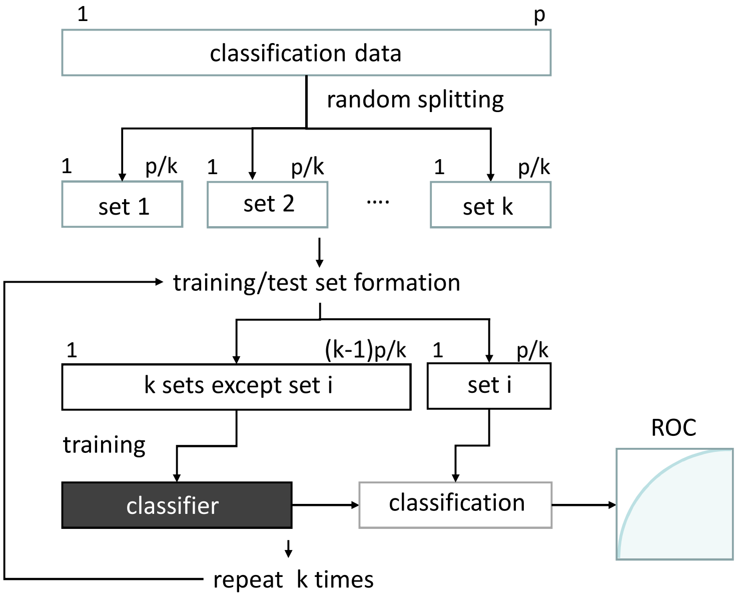
training

ROC

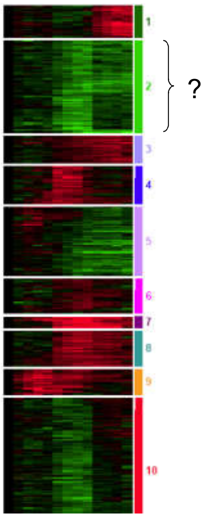
classifier

classification

K-fold cross validation



Biological meaning of the gene sets



- Guilt-by-association
- Regulation by the same transcription factor
- Gene ontology terms
- Over representation analysis
- Pathways

Gene Ontology

Gene Ontology (GO)

The Gene Ontology project (http://geneontology.org) provides a **controlled vocabulary** to describe gene and gene product attributes in any organism.

The three organizing principles (categories) of GO are

- cellular component

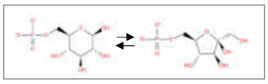
mitochondrium

- biological process

cell cycle

- molecular function

isomerase activity



What’s in a GO term?

- **Term**

  transcription initiation

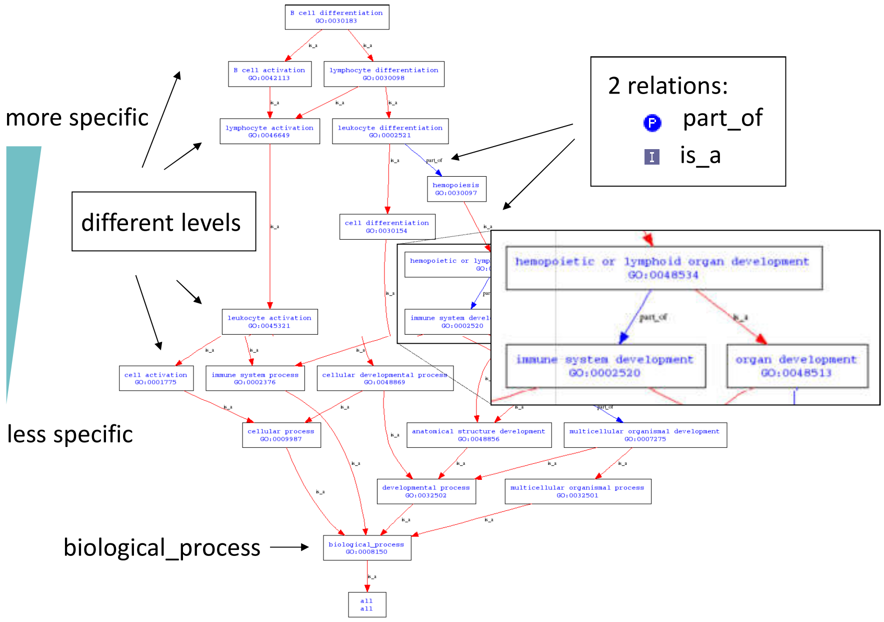
- **ID**

  GO:0006352

- **Definition**

  Processes involved in starting transcription, where transcription is the synthesis of RNA by RNA polymerases using a DNA template.

Parent /child relation in directed acyclic graph (DAG)


B cell differentiation GO:0030183
B cell activation GO:0042113
lymphocyte differentiation GO:0030098
lymphocyte activation GO:0046649
leukocyte differentiation GO:0002521
hemopoiesis GO:0030097
cell differentiation GO:0030154
hemopoietic or lymphoid organ development GO:0048534
immune system development GO:0002520
organ development GO:0048513
leukocyte activation GO:0045321
cell activation GO:0001775
immune system process GO:0002376
cellular developmental process GO:0048869
cellular process GO:0009987
anatomical structure development GO:0048856
multicellular organismal development GO:0007275
developmental process GO:0032502
multicellular organismal process GO:0032501
biological_process GO:0008150
all all
is_a
part_of

more specific

different levels

less specific

biological_process

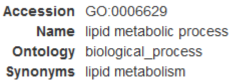
2 relations:

- P part_of
- I is_a

Gene Ontology Browser (Amigo2)

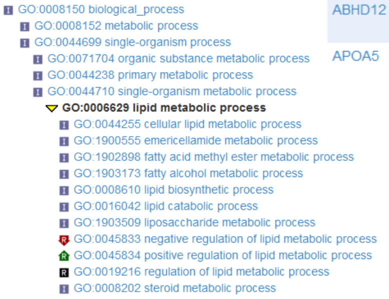
http://amigo2.geneontology.org (http://geneontology.org/)

Term information



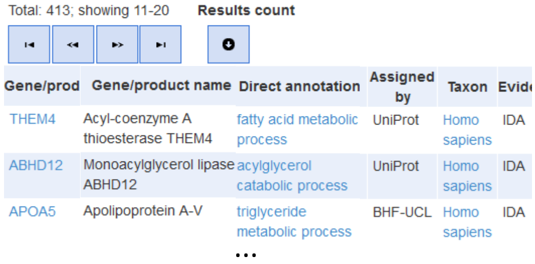
Accession	GO:0006629
Name	lipid metabolic process
Ontology	biological_process
Synonyms	lipid metabolism

Inferred tree view



- GO:0008150 biological_process
  - GO:0008152 metabolic process
  - GO:0044699 single-organism process
    - GO:0071704 organic substance metabolic process
    - GO:0044238 primary metabolic process
    - GO:0044710 single-organism metabolic process
      - **GO:0006629 lipid metabolic process**
        - GO:0044255 cellular lipid metabolic process
        - GO:1900555 emericellamide metabolic process
        - GO:1902898 fatty acid methyl ester metabolic process
        - GO:1903173 fatty alcohol metabolic process
        - GO:0008610 lipid biosynthetic process
        - GO:0016042 lipid catabolic process
        - GO:1903509 liposaccharide metabolic process
        - GO:0045833 negative regulation of lipid metabolic process
        - GO:0045834 positive regulation of lipid metabolic process
        - GO:0019216 regulation of lipid metabolic process
        - GO:0008202 steroid metabolic process

Annotation



Total: 413; showing 11-20 **Results count**

Gene/prod	Gene/product name	Direct annotation	Assigned by	Taxon	Evid
THEM4	Acyl-coenzyme A thioesterase THEM4	fatty acid metabolic process	UniProt	Homo sapiens	IDA
ABHD12	Monoacylglycerol lipase ABHD12	acylglycerol catabolic process	UniProt	Homo sapiens	IDA
APOA5	Apolipoprotein A-V	triglyceride metabolic process	BHF-UCL	Homo sapiens	IDA

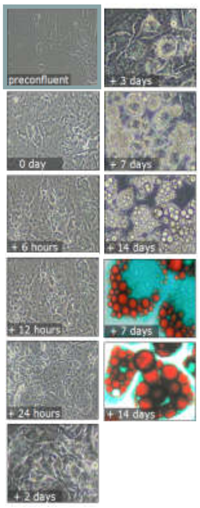
…

Evidence code for GO annotations

ISS	Inferred from Sequence Similarity
IEP	Inferred from Expression Pattern
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IPI	Inferred from Physical Interaction
IDA	Inferred from Direct Assay
RCA	Inferred from Reviewed Computational Analysis
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IC	Inferred by Curator
ND	No biological Data available

Case study: fat cell differentiation

Reference (proliferating cells)

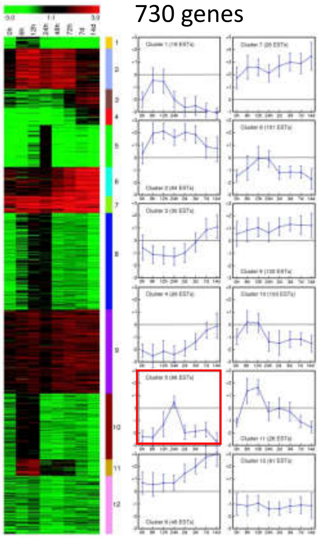


Microarray analysis

⇨

Time series
0h-ref
+6h-ref
+12h-ref
…
+14d-ref

730 genes



Heatmap (k-means clustering)

Biological function of genes in clusters

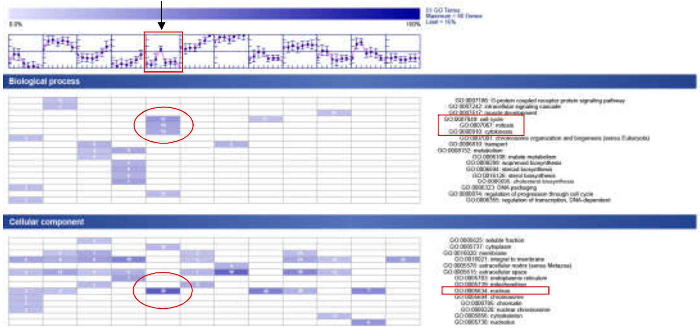
⇨

GO Analysis

Hackl H, Burkard TR et al. Genome Biol. 2005

GO terms for gene sets

cluster 5

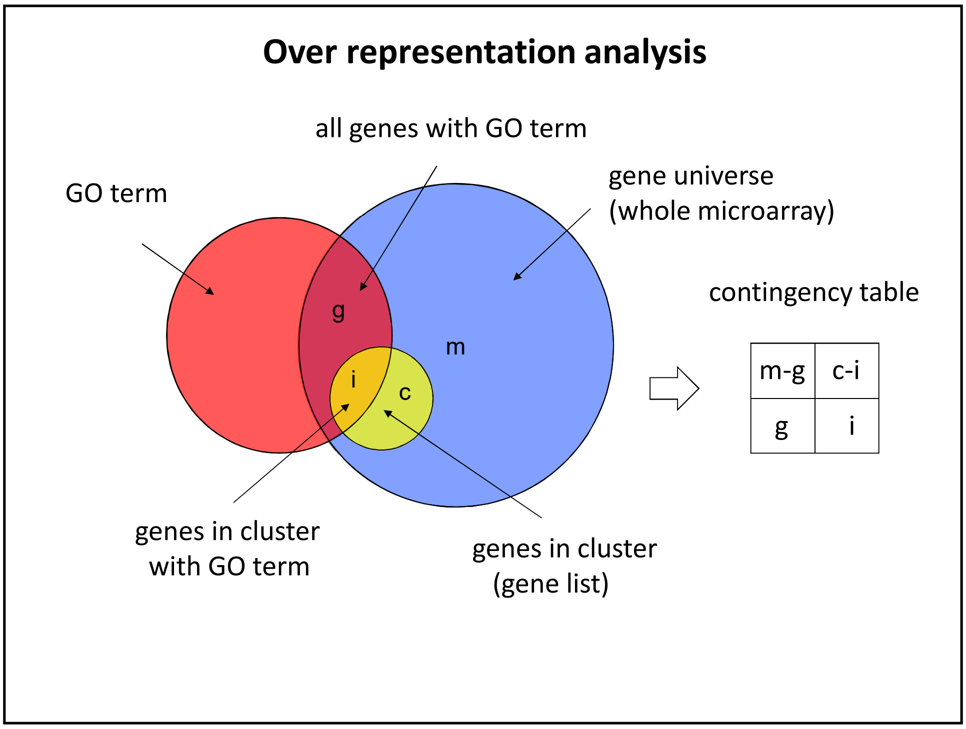


cell cycle (17)
mitosis (14) ⟶
cytokineses (13)

nucleus (30) ⟶

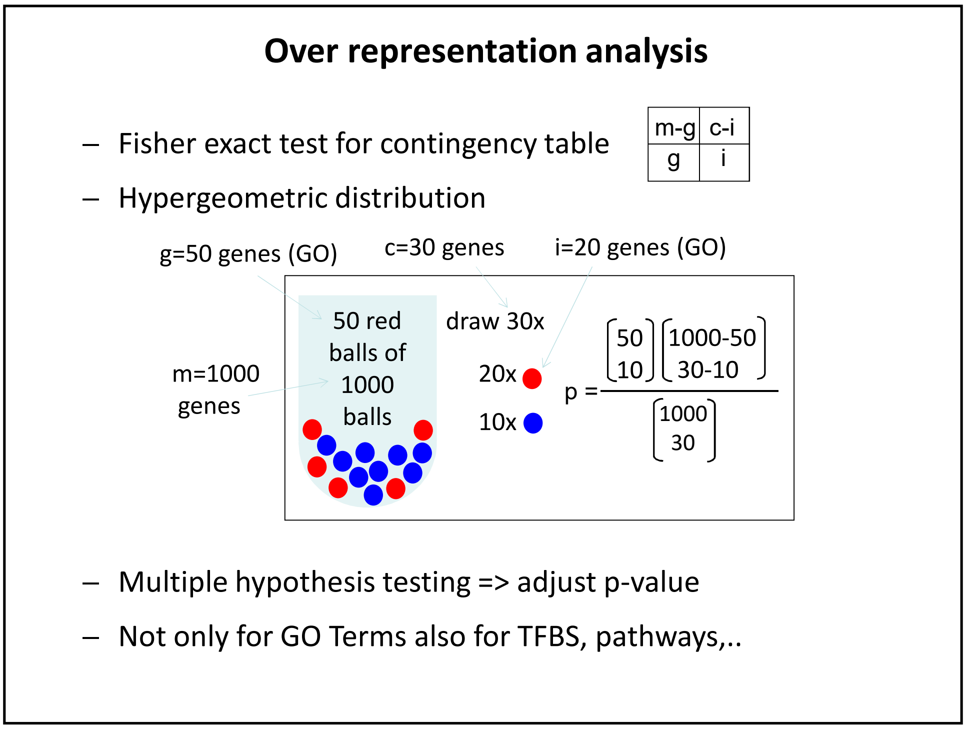
– 3T3-L1 cell line undergoes ≥ 1 cell cycle before terminal adipocyte differentiation around 1 day after induction (clonal expansion)

Are results just by chance?
⇨ Over representation analysis



Over representation analysis

m-g	c-i
g	i



Over representation analysis

- Fisher exact test for contingency table

m-g	c-i
g	i

- Hypergeometric distribution

$$p = \frac{\binom{50}{10}\binom{1000-50}{30-10}}{\binom{1000}{30}}$$

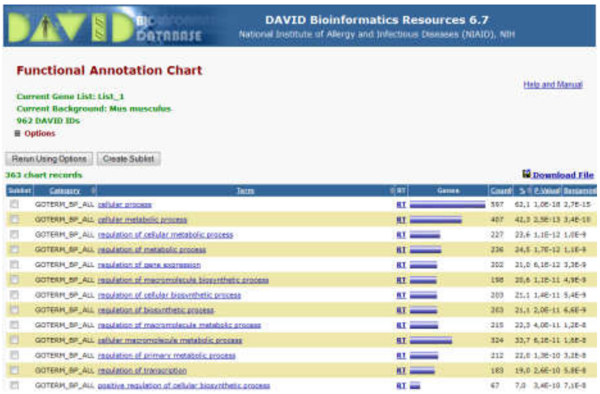
- Multiple hypothesis testing => adjust p-value
- Not only for GO Terms also for TFBS, pathways,..

DAVID

- Database for Annotation, Visualization and Integrated Discovery
- https://david.ncifcrf.gov
- Functional annotation tool (over representation analysis)

1019 mouse
gene symbols

Dnajb1
Wnt11
Sorbs3
D230025D16Rik
Sfxn3
Hspa5
Golga3
Hgs
Npc1
Mta2
Cnn2
Spg20
Zpr1
…



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Mus musculus
962 DAVID IDs
▪ Options

Rerun Using Options | Create Sublist

363 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_BP_ALL	cellular process	RT		597	62,1	1,0E-18	2,7E-15
□	GOTERM_BP_ALL	cellular metabolic process	RT		407	42,3	2,5E-13	3,4E-10
□	GOTERM_BP_ALL	regulation of cellular metabolic process	RT		227	23,6	1,1E-12	1,0E-9
□	GOTERM_BP_ALL	regulation of metabolic process	RT		236	24,5	1,7E-12	1,1E-9
□	GOTERM_BP_ALL	regulation of gene expression	RT		202	21,0	6,1E-12	3,3E-9
□	GOTERM_BP_ALL	regulation of macromolecule biosynthetic process	RT		198	20,6	1,1E-11	4,9E-9
□	GOTERM_BP_ALL	regulation of cellular biosynthetic process	RT		203	21,1	1,4E-11	5,4E-9
□	GOTERM_BP_ALL	regulation of biosynthetic process	RT		203	21,1	2,0E-11	6,6E-9
□	GOTERM_BP_ALL	regulation of macromolecule metabolic process	RT		215	22,3	4,0E-11	1,2E-8
□	GOTERM_BP_ALL	cellular macromolecule metabolic process	RT		324	33,7	6,1E-11	1,8E-8
□	GOTERM_BP_ALL	regulation of primary metabolic process	RT		212	22,0	1,3E-10	3,2E-8
□	GOTERM_BP_ALL	regulation of transcription	RT		183	19,0	2,6E-10	5,8E-8
□	GOTERM_BP_ALL	positive regulation of cellular biosynthetic process	RT		67	7,0	3,4E-10	7,1E-8

…