104540 VO/2 Bioinformatik SS2025

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/cbio

104540 VO/2 Bioinformatik SS2025

PART I (Hubert Hackl)

I Transcriptional regulation

II Biological sequence analyses

III Gene expression analyses

PART II (Francesca Finotello)

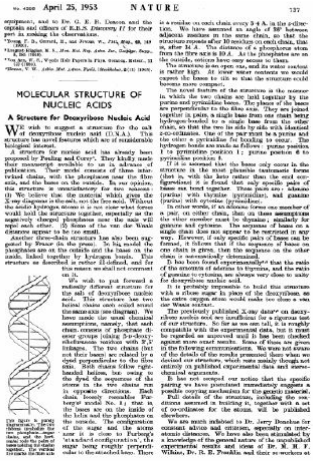
IV Functional and network analyses (Pathways, Enrichment)

V Single cell analyses (scRNAseq)

I Transcriptional regulation

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

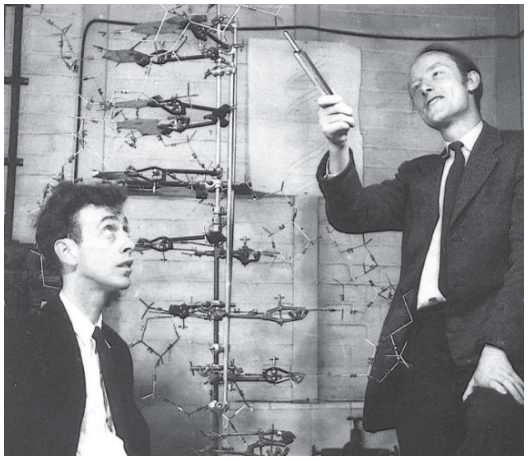
History



April 25, 1953 NATURE 737

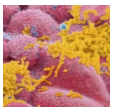
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

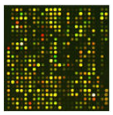
A Structure for Deoxyribose Nucleic Acid



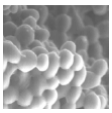
History

- **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

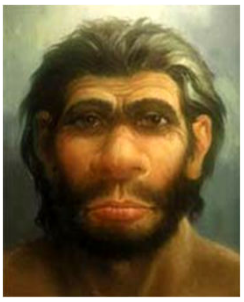
2000
- Draft version of the human genome (>10 years, >3 billion $, 20 labs)

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

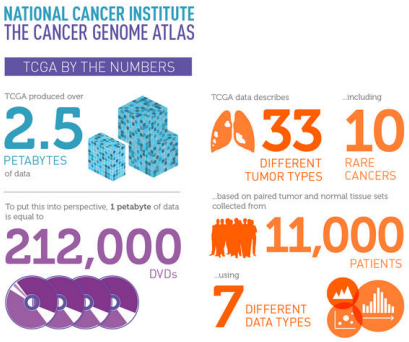
1000 Genomes Project (=> 100.000 genomes project)
- Study human genetic variation of >1.000 human genomes

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)

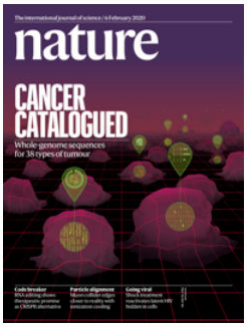
https://tcga-data.nci.nih.gov

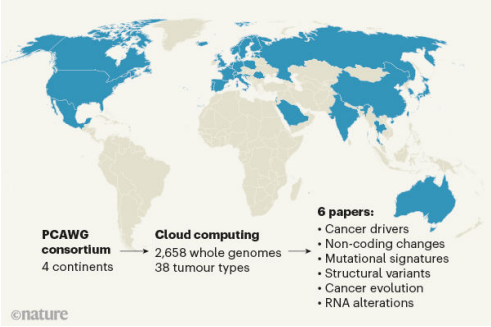


- Copy number
- Methylation
- Gene expression
- MicroRNA expression
- Somatic mutations
- Clinical data

Pan-Cancer Analysis of Whole Genomes Consortium

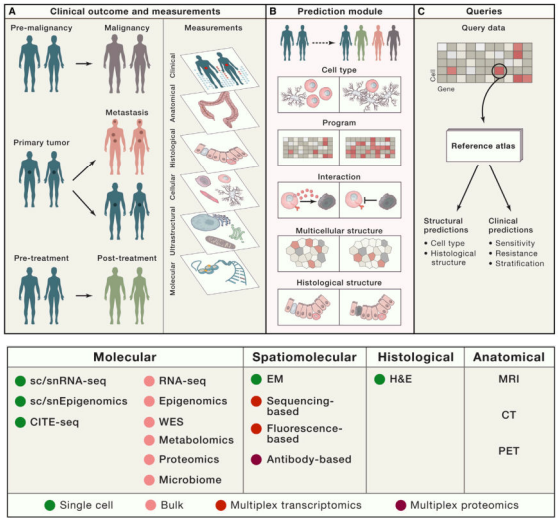
>2600 whole cancer genomes
38 tumor types
750 affiliations





Feb 2020

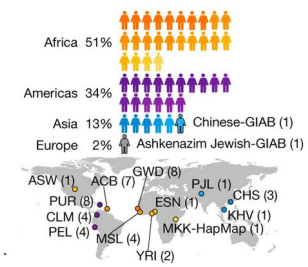
HTAN
HUMAN TUMOR ATLAS NETWORK

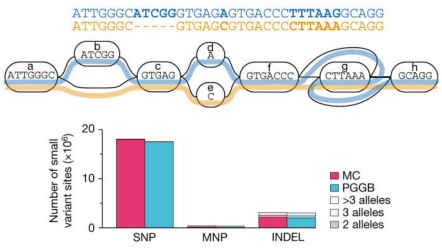


Johnson et al. Cell 2020

Human pangenome reference

- 47 phased, diploid assemblies from a cohort of genetically diverse individuals
- cover more than 99% of the expected sequence in each genome and are more than 99% accurate at the structural and base pair levels



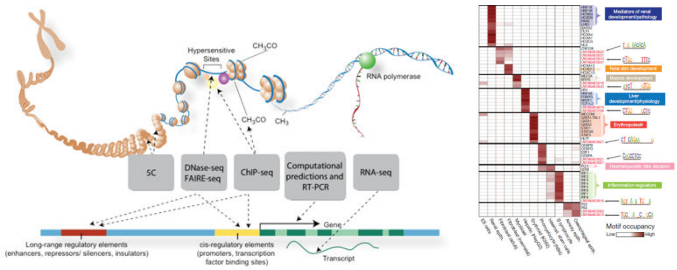


Lia et al. Nature 2023

ENCODE (Encyclopedia of DNA Elements)

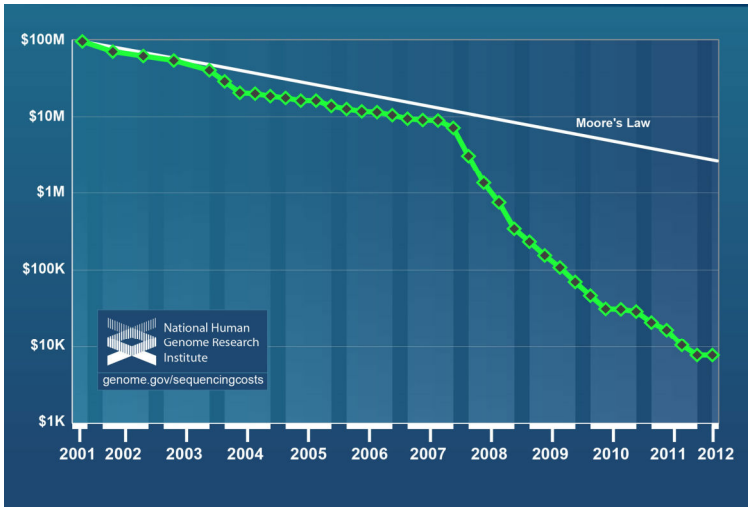
32 institutes
442 consortium members
1640 data sets
30 papers (Sept 2012)

http://www.nature.com/encode
http://genome.ucsc.edu/ENCODE/
http://www.genome.gov/10005107

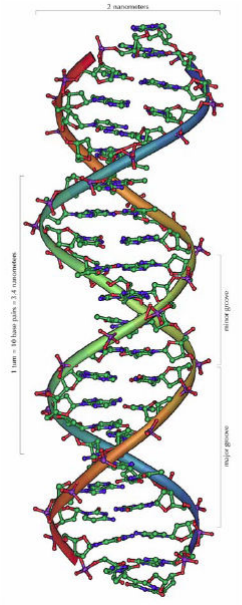


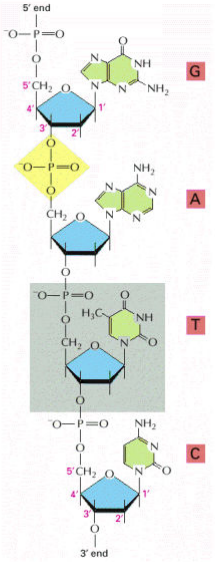
The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.

Cost per genome



DNA

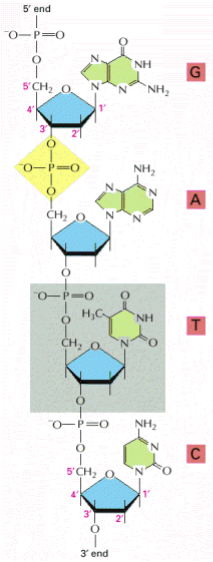


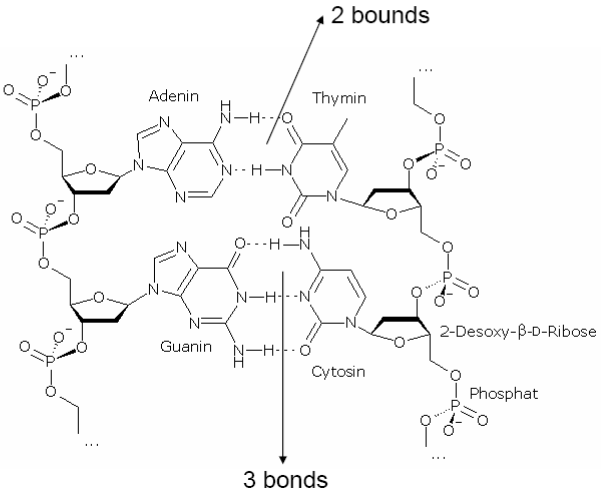


Guanine (binds to C)
Adenine (binds to T)
purine

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*

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

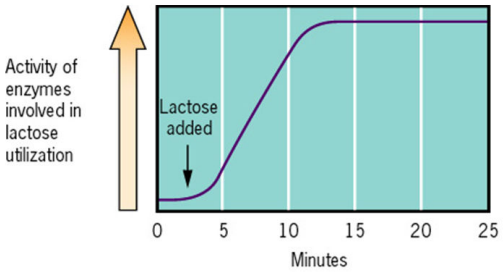
Gene regulation in prokaryotes

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

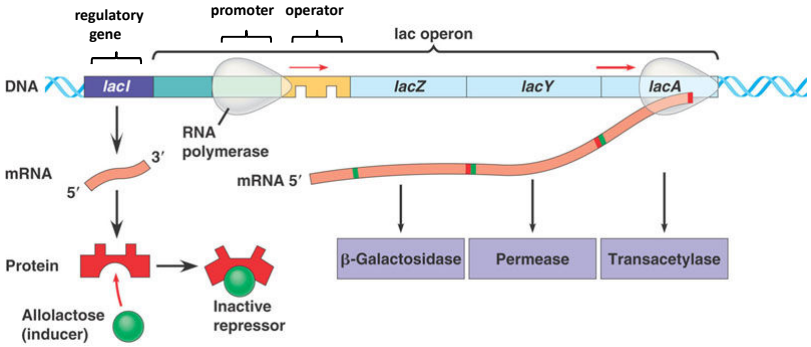
Example lack of glucose but abundance of lactose

- 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

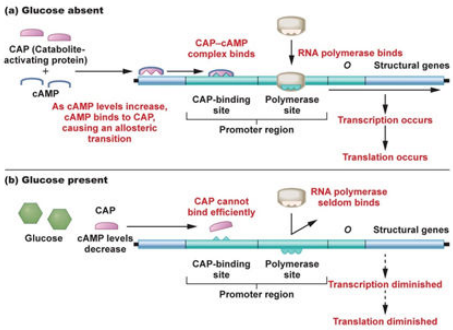
– If lactose is present repressor is inactivated by conformational changes => mRNA expression of structural genes

Prokaryotic transcriptional regulation

- Glucose and the lac operon

– Lactose is metabolised into glucose so what happens if glucose is present.

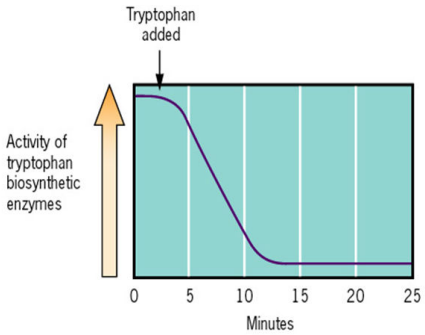
– Catabolite-activation protein (CAP): CAP must be present to make RNA polymerase binding efficiently



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

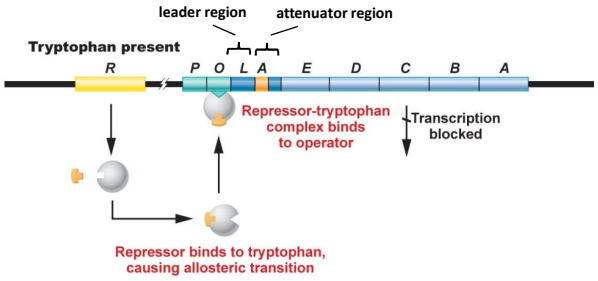
Response to environmental stimuli

- 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

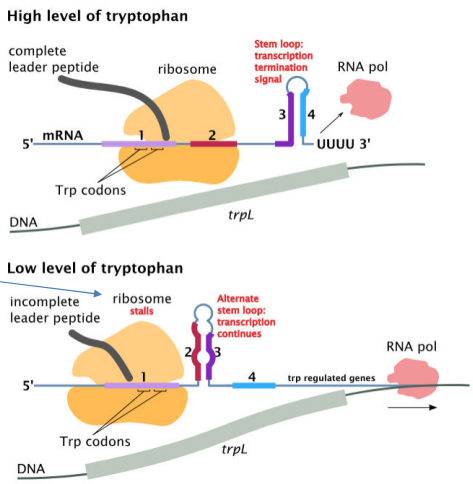


- If tryptophan is present the repressor-tryptophan complex binds to operator => no mRNA expression of structural genes.
- Translation and transcription are coupled (regulation by leader sequence and attenuation)

Attenuator mechanism

Translation is directly coupled to transcription

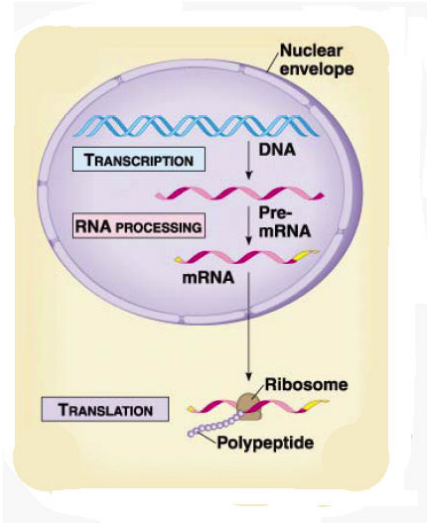
No Trp =>
takes time



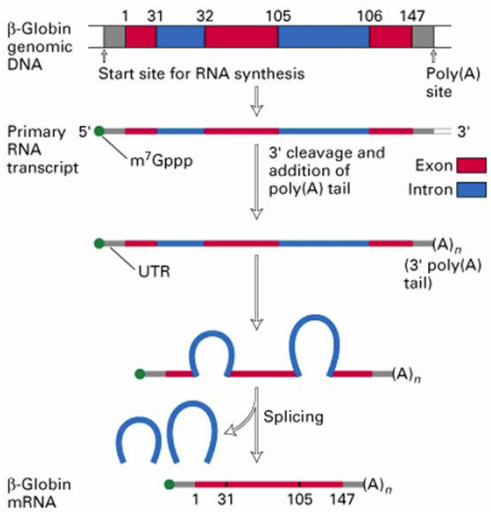
Gene regulation in eukaryotes

Gene expression in eukaryotes

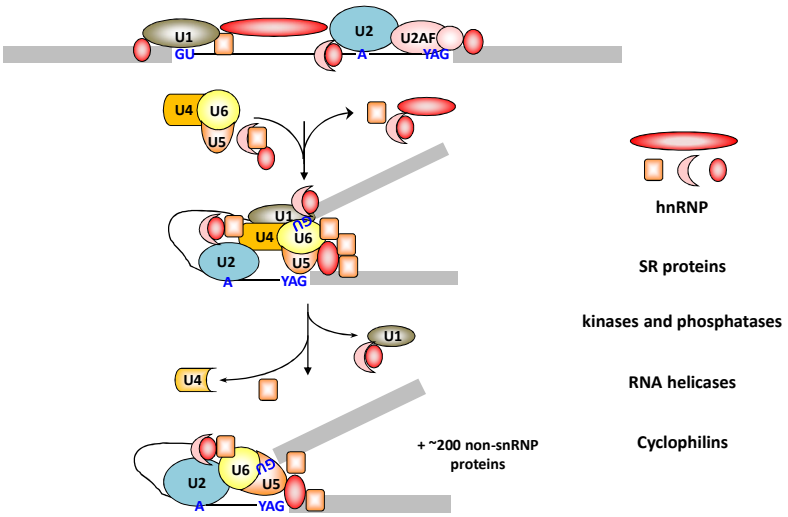
- Two cellular compartments:
  - Transcription in nucleus
  - Translation in cytoplasm
- RNA processing
  - 5´capping
  - RNA splicing
  - 3´polyadenylation



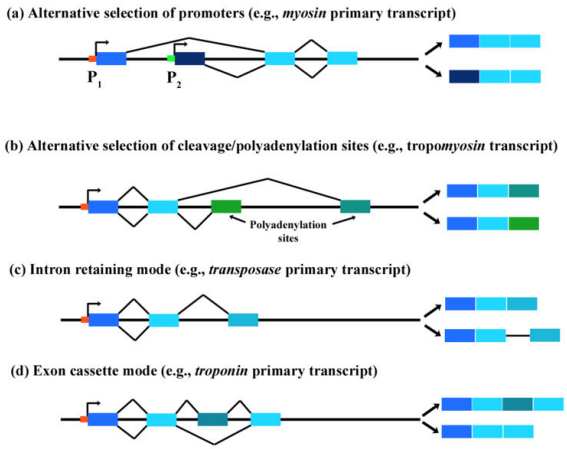
mRNA processing



Spliceosome assembly



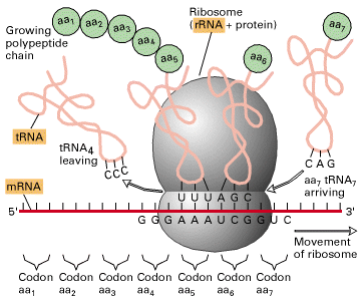
Alternative splicing

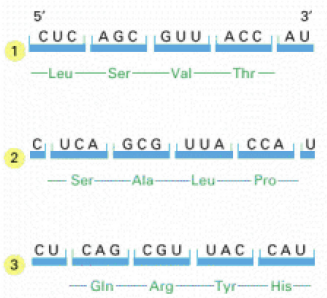


- Dependent on RNA/Spliceosome interaction
- Economizes on genetic information
- Create numerous related yet different proteins

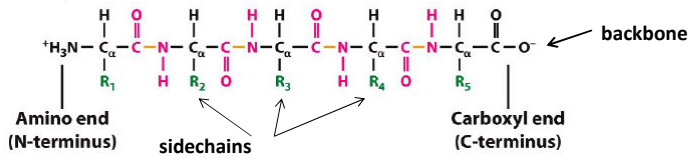
Translation, genetic code and reading frames

	AGA									UUA					AGC				
	AGG									UUG					AGU				
GCA	CGA						GGA			CUA				CCA	UCA	ACA			GUA	
GCC	CGC						GGC		AUA	CUC				CCC	UCC	ACC			GUC	UAA
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUG	AAA		UUC	CCG	UCG	ACG		UAC	GUG	UAG
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	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

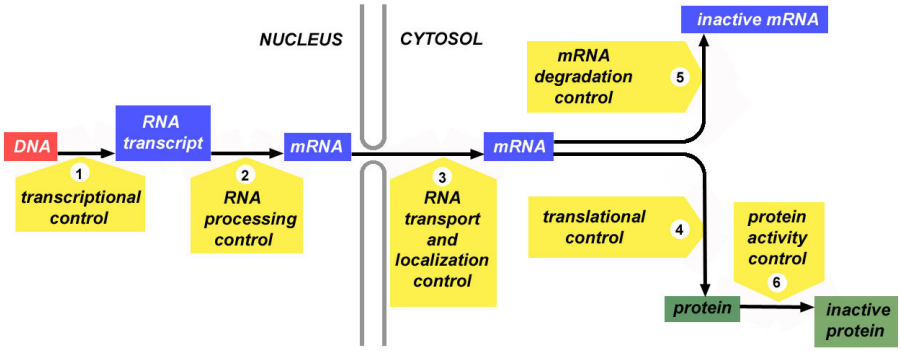


Protein sequences are always form N-terminal end to C-terminal end

E.g.. SCD sequence in fasta format

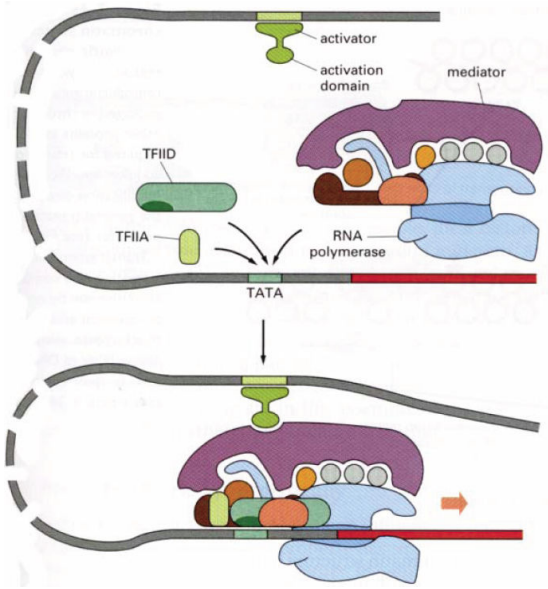
```
>gi|53759151|ref|NP_005054.3| acyl-CoA desaturase [Homo sapiens]
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVE
YVWRNIILMSLLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLII
ANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKL
VMFQRRYYKPGLLMMCFILPTLVPWYFWGETFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNI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

Different levels of regulation

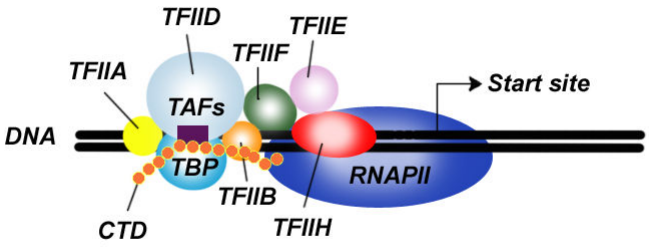


Transcriptional regulation has largest effect on phenotype!

Regulation of eukaryotic transcription



Basal transcription factors



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

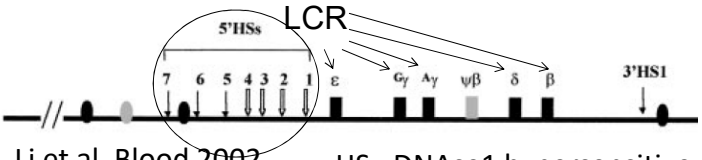
Trans elements: DNA-binding proteins that change the level of transcription by basal transcription machinery.

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)

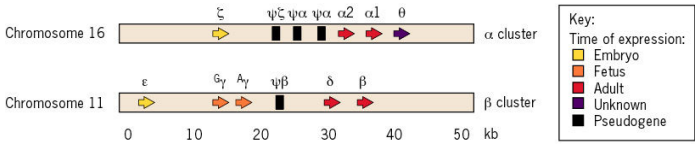


Li et al. Blood 2002

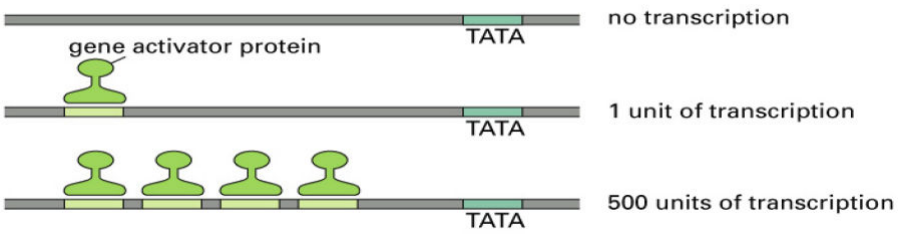
HS.. DNAse1 hypersensitive sites

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

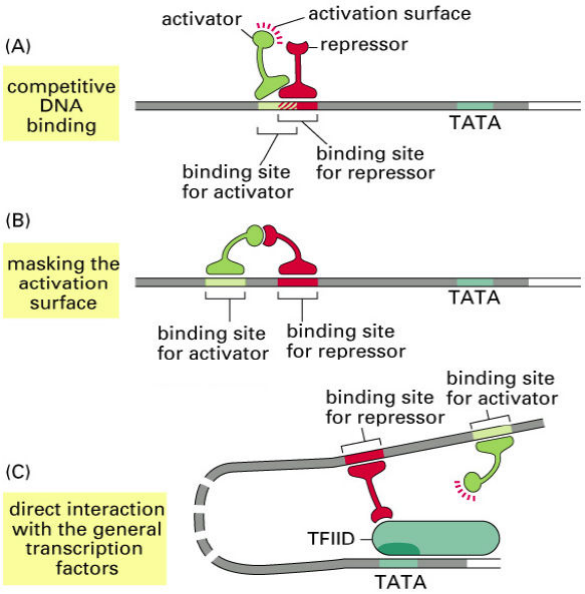
- Temporal regulation of hemoglobin (tetramer 2xα +2xβ)



Transcriptional synergy

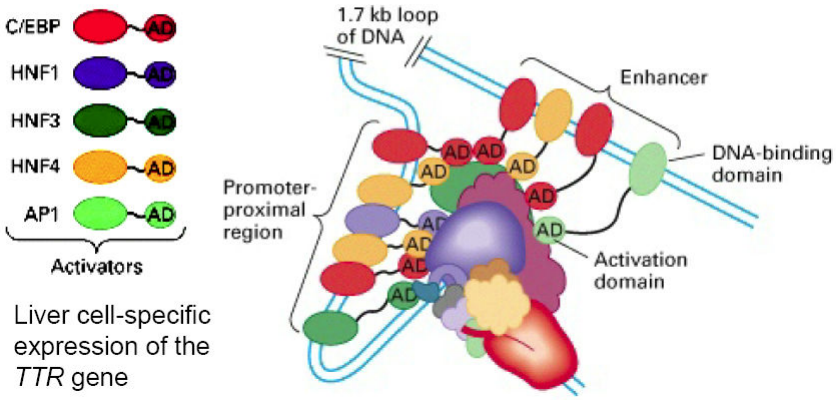


Eukaryotic gene repressors



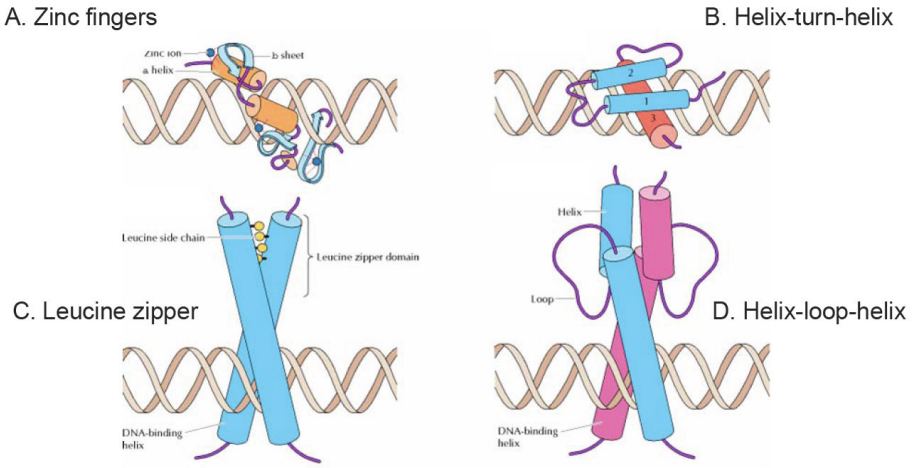
Transcription factor combinations

Most genes are regulated by multiple transcription factors



Liver cell-specific expression of the *TTR* gene

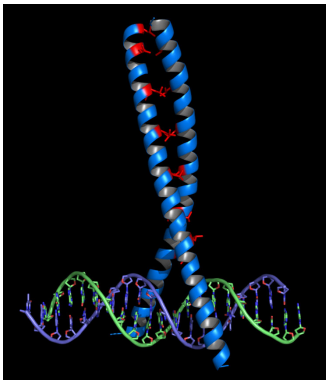
Classification of TF by DNA binding



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

Transcription factor dimerization

Leucine zippers



- homo dimerization
- hetero dimerization

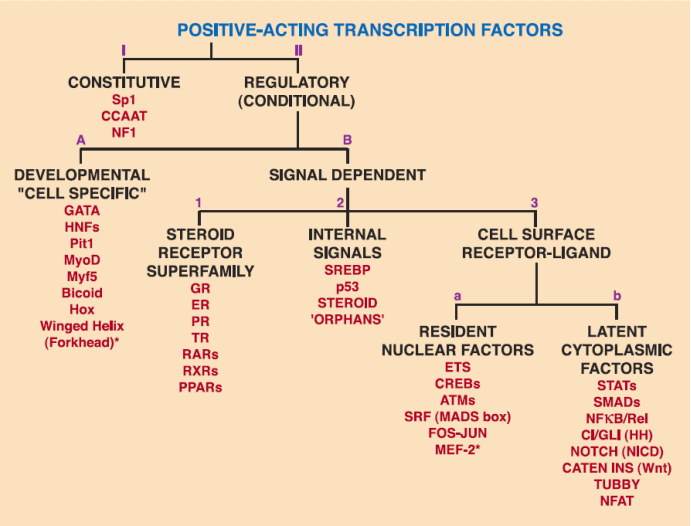
Signaling

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

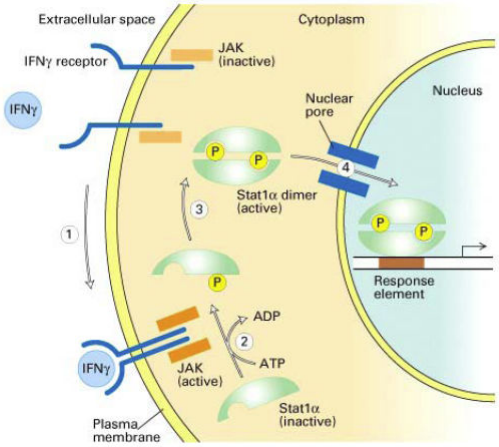
Classification of TF by function



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

Regulation by phosphorylation

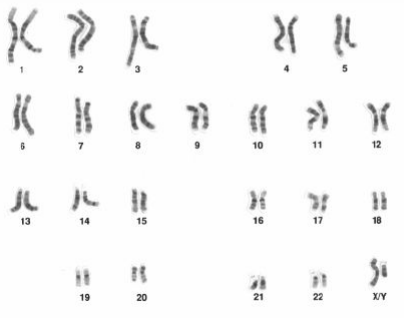
- Hormone activates kinase
- Kinase phosphorylates transcription factor
- Transcription factor is activated

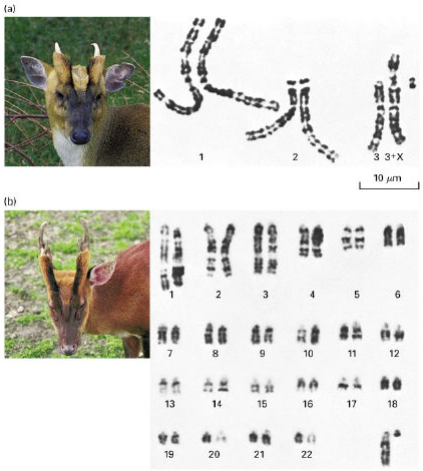


Principles of TF regulation

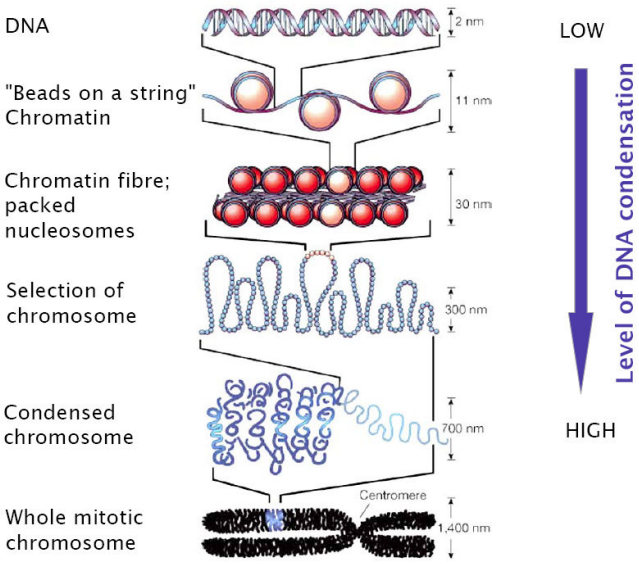
- 1 TF can target promoter of many genes
- >1 TF regulate expression of 1 gene (modules)
- Cascade of TF possible
- Positive feedback loop (autoregulation)
- Feed forward loop

Chromosomes

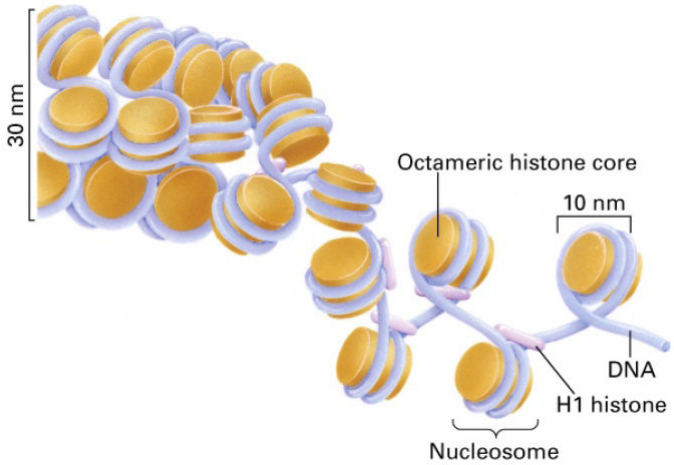




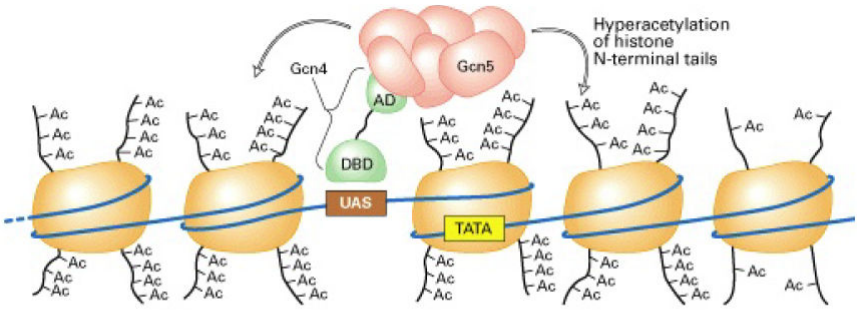
DNA packing



The solenoid model of condensed chromatin

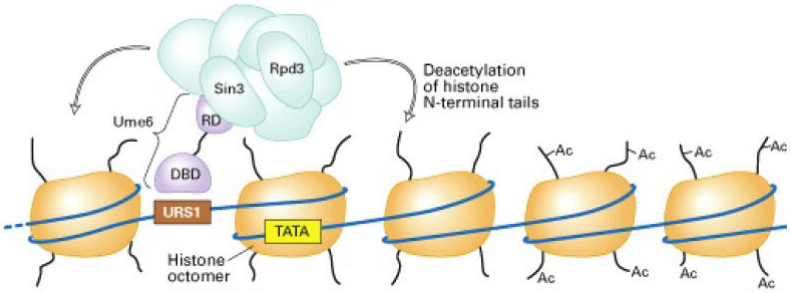


Activators: histone acetylation



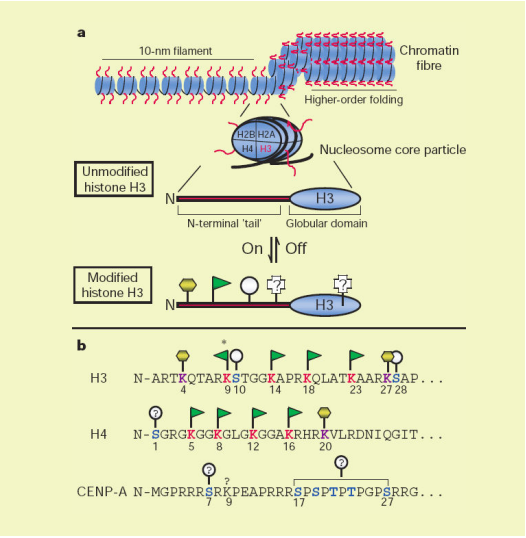
- Some activators recruit histone acetylase, which adds acetyl groups to histones
- Allows transcriptional machinery access to less condensed template DNA (euchromatin)

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



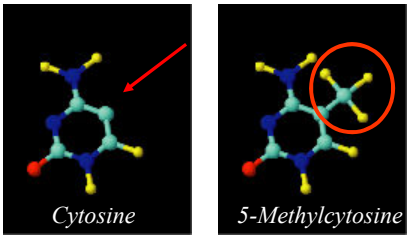
Strahl BD, Allis CD. Nature 2000. 403:41-45

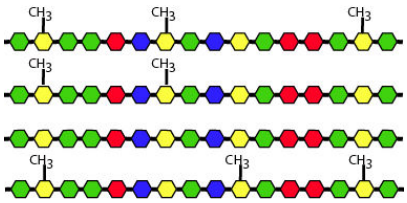
Chromatin states

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
	Chromatin mark observation frequency (%)										(%)	(fold)		(kb)	(%)	Functional enrichments (fold)						
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

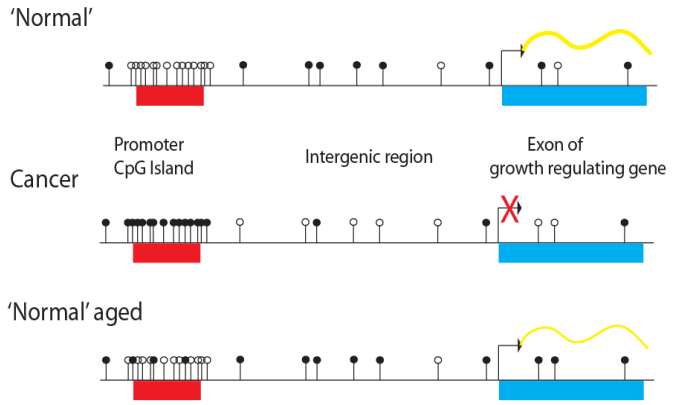
Ernst et al. Nature 2011.

DNA methylation

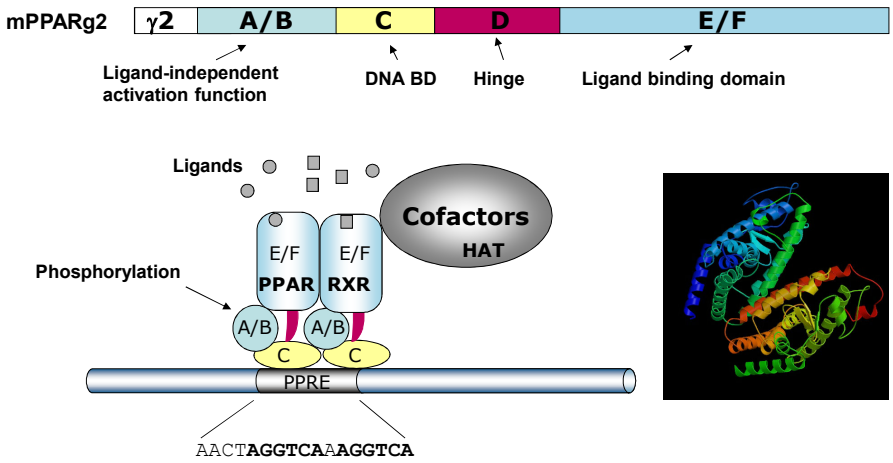




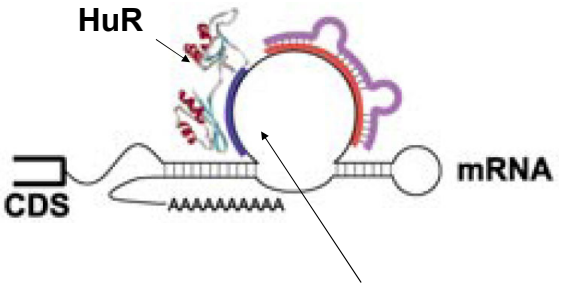
Aberrant methylation patterns



Nuclear receptors



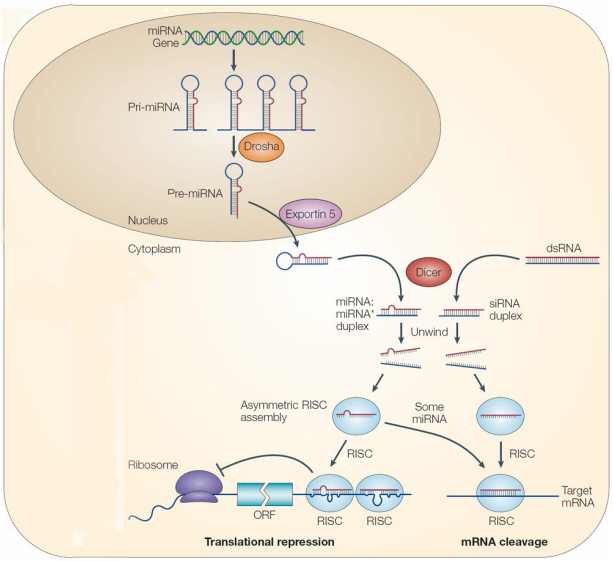
RNA binding proteins for mRNA stability



AU rich elements (ARE)

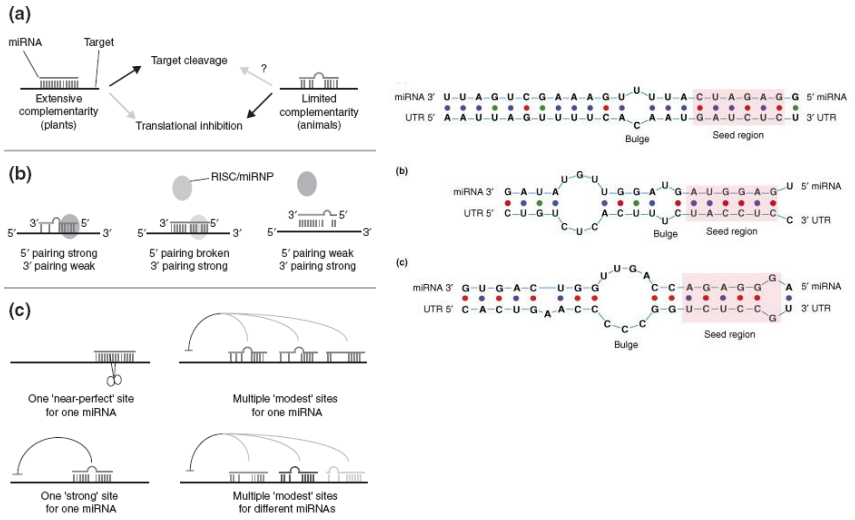
Cox-2	UAUUAAUUUAAUUAUUUAAUAAUAUUUAUAUUAAA
IL-1β	UAUUUAUUUAUUUAUUUGUUUGUUUGUUUUAUU
IL-2	UAUUUAUUUAAAUAUUUAAAUUUUAUAUUUAUU
IL-4	AUAUUUUAAUUUAUGAGUUUUUGAUAGCUUUAUUUUUUAAG
IL-8	UAUUUAUUAUUUAUGUAUUUAUUUAA
TNFα	AUUAUUUAUUAUUUAUUUAUUAUUUAUUUAUUUA

microRNA and siRNA

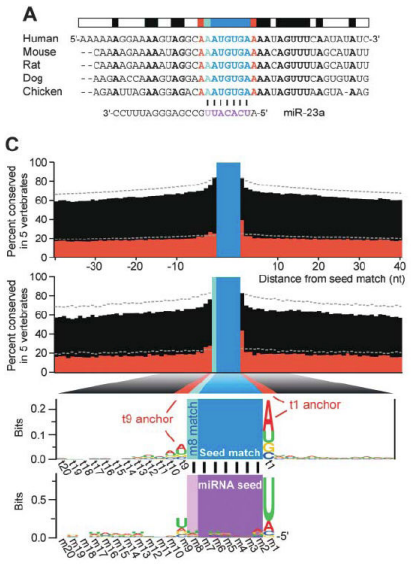


He L., Hannon GJ. Nature Reviews Genetics. 2004. 5:522-531

miRNA-mRNA targeting



Conservation of microRNA target sequences



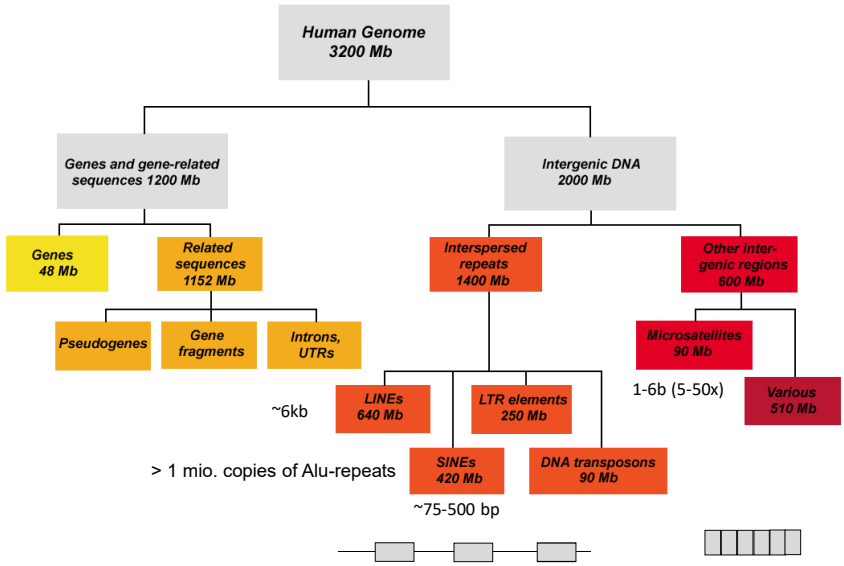
Genome analyses

Human Genome

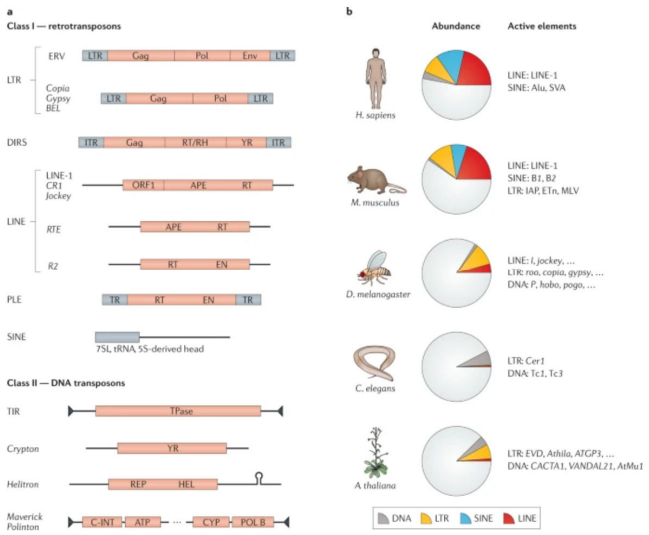
2.95 Gbases of 3.2 Gbases is euchromatin
- >90% of euchromatin sequenced
- ~1% of sequence encodes protein sequences

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



Transposons



Deniz et al. Nat Rev Genet. 2019

Bioinformatics challenges in genome analysis

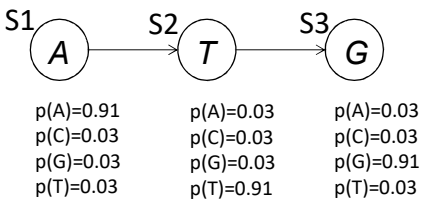
- Gene finding
- Start codon
- Exon-intron borders
- CpG-islands
- Repetitive sequences (Repeat Masker)
- Regulatory sequences

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

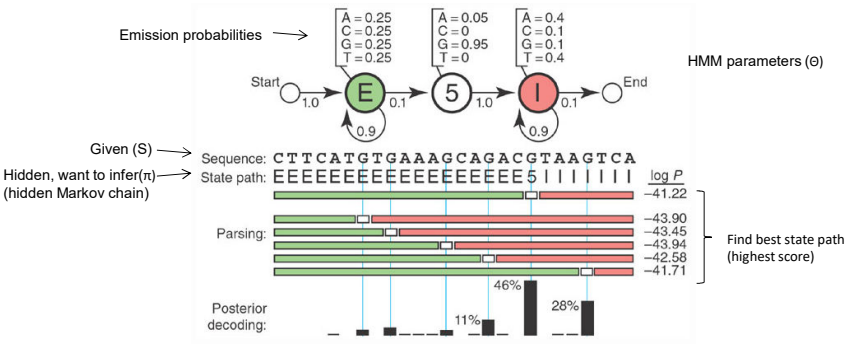


Markov chain 0th order
p(*ATG*)=0.91^3=0.752

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

Hidden Markov Model (HMM)

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



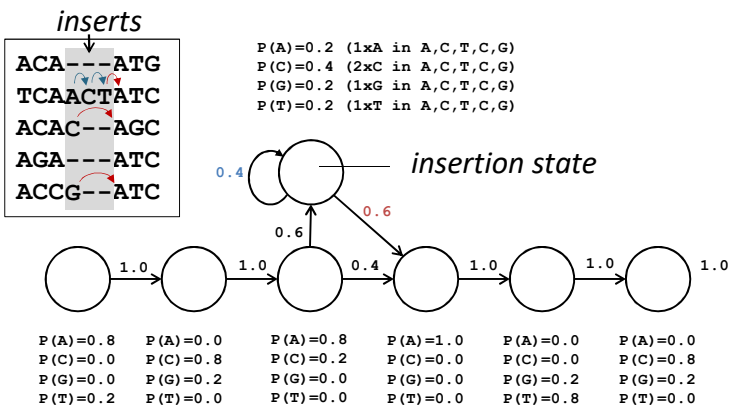
G T A A G T C A

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

Eddy SR, Nat Biotech 2004

Profile Hidden Markov Model

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



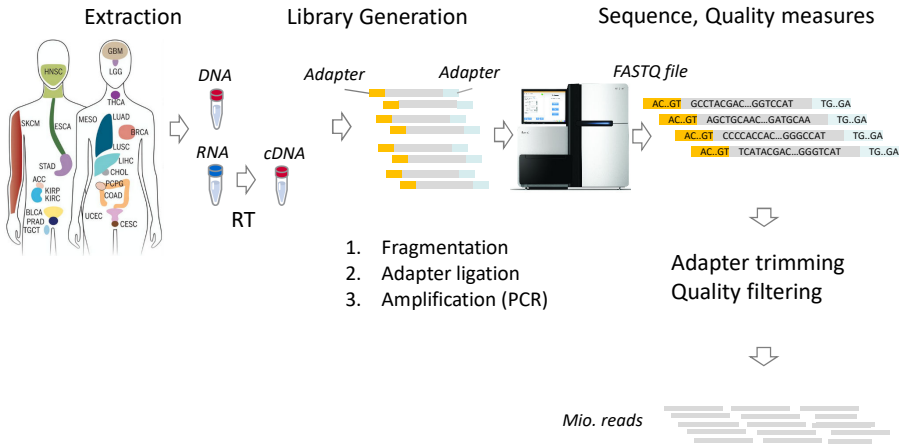
p(*ACACATC*)=0.8*1*0.8*1*0.8*0.6 *0.4*0.6*1*1*0.8*1*0.8=0.047
log-odds=log(p(S)/0.25^L)=log(0.047/0.25^7)

II Biological sequence analyses

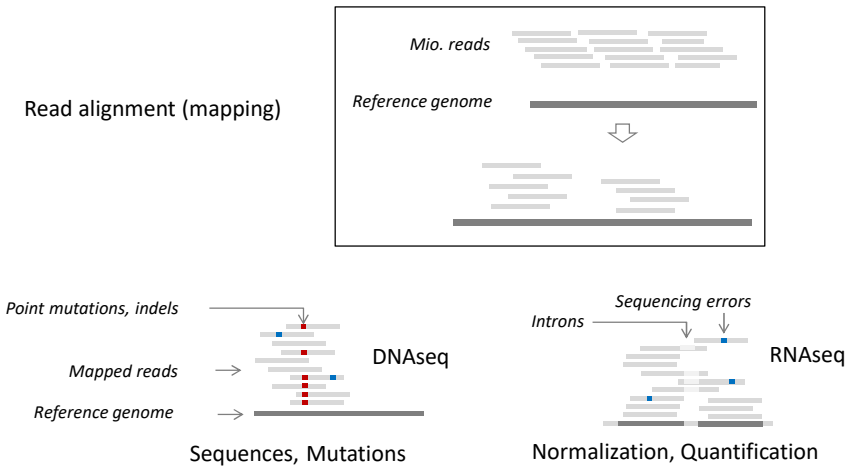
- Mapping algorithms for NGS data
- Sequence alignment of 2 sequences
- Multiple sequence alignment
- Predictive models using protein sequences
- Regulatory sequences

Mapping algorithms for NGS data

Next generation sequencing (NGS)



Read alignment



Exact string matching

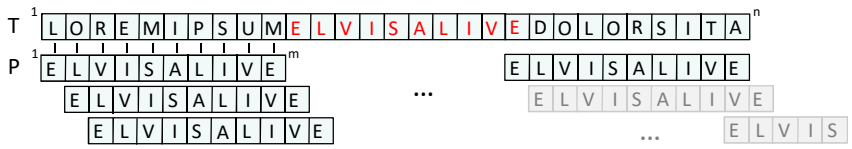
Problem

10 mio. short sequence reads (100 bp)

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

⇨ String matching problem in text processing

1 Naïve approach



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

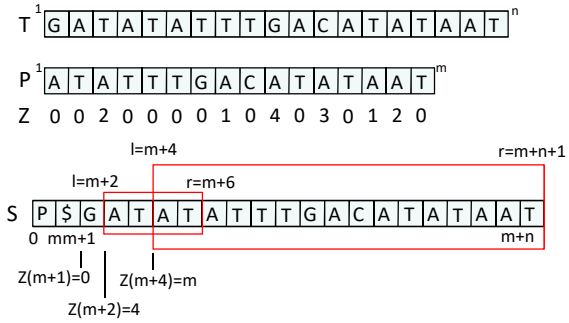
⇔

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



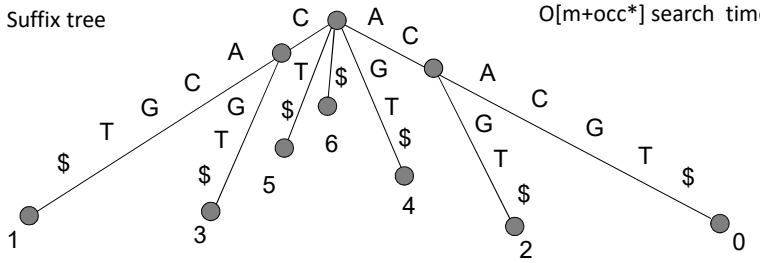
O[n+m]

- 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])

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

sort lexicographic →

Index of suffix array = S(i)

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

LF=3 → ACGT$AC

i=5 ← CGT$ACA

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)

SP

EP

BWT Matrix

```
1$ACACGT
ACACGT$
ACGT$AC
CACGT$A
CGT$ACA
GT$ACAC
nT$ACACG
```

C(c)

$	A	C	G	T
0	1	3	5	6

P= 1 C A C m

Occ(c,i) (=rank)

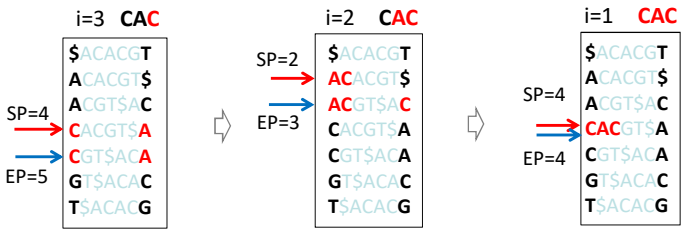
$	A	C	G	T
0	0	0	0	1
1	0	0	0	1
1	0	1	0	1
1	1	1	0	1
1	2	1	0	1
1	2	2	0	1
1	2	2	1	1

```
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

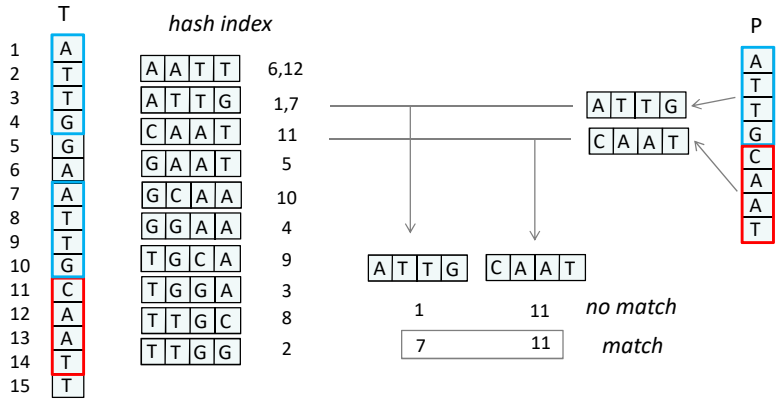


- FM-index can be also used for approximate string matching (k-mismatch search) by *backtracking.*
- BWT is compressible (run length encoding, move-to-front)
- In the original *Bowtie* implementation of the BWT-based FM-index for the human genome requires only 1.3 GB of memory.

Hash index based methods

Hashing

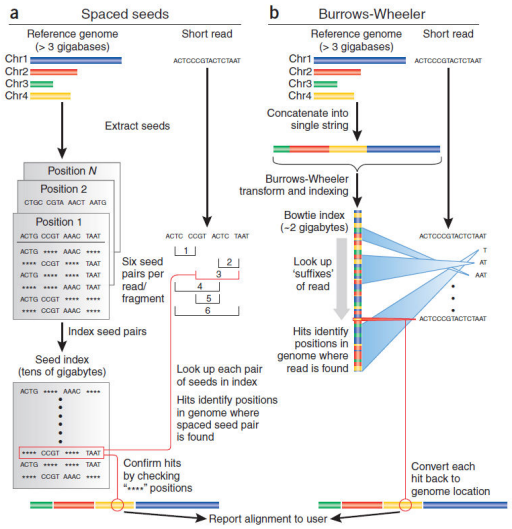
– Using *k*-mer seeds



– 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

Align biological sequences

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

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

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

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

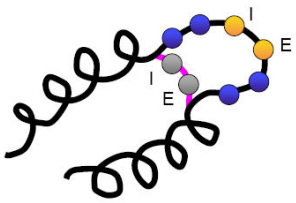
- We can align:
  - Two sequences at a time (pair-wise sequence alignment)
  - Many sequences simultaneously (multiple alignment)

Number of all possible alignments for length n and m $\binom{n+m}{m}$

=> 2 sequences with length 100 = 9 x 10^{58}

Biology of gaps

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

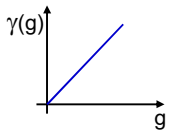


Gap penalties

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

- Linear gap penalty function

  $\gamma(g) = -g*d$

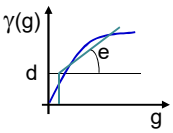


- Convex gap penalty function (more realistic)

  Affine score:

  $\gamma(g) = -d - (g-1)*e$

  gap open penalty — gap extend penalty



Distance scoring (DNA sequences)

- **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 q_a.

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

- The alternative match model of aligned pairs of residues occurs with a joint probability p_{ab}.

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

- Odds ratio

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

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

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

Substitution matrices

- Blocks Substitution Matrix (BLOSUM n)
  - Henikoff and Henikoff, 1992
  - Conserved, ungapped regions of a protein family
  - BLOSUM 90 short alignments, highly similar
  - BLOSUM 62 standard, members of protein family
  - BLOSUM 30 longer, weaker local alignments

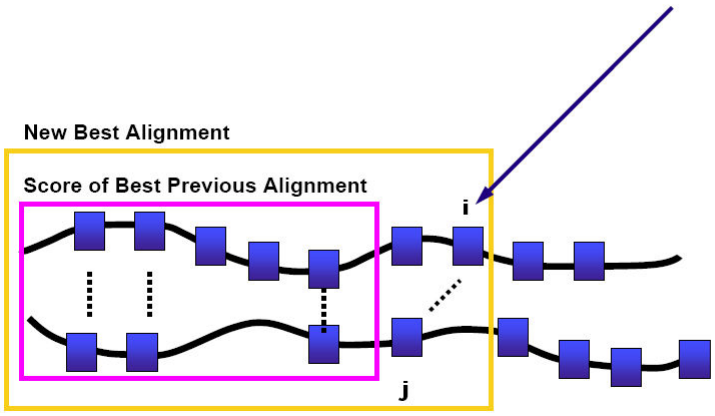
- Point Accepted Mutation (PAM n)
  - Margaret Dayhoff, 1978
  - Substitutions in related proteins
  - PAM 1 ~ 1 amino acid change per 100 residues
  - PAM 40 short alignments, highly similar
  - PAM 120
  - PAM 250 longer, weaker local alignments

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

Dynamic programing for sequence alignment

New Best Alignment = Previous Best + Local Best



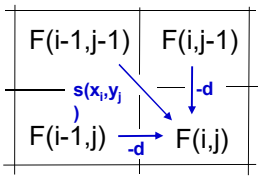
Global alignment: Needleman-Wunsch algorithm

- 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



Global sequence alignment

Example with S=BLOSUM50 and d=8

start

		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

best score

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

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

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

stop

best score

AWGHE
AW-HE

Database search

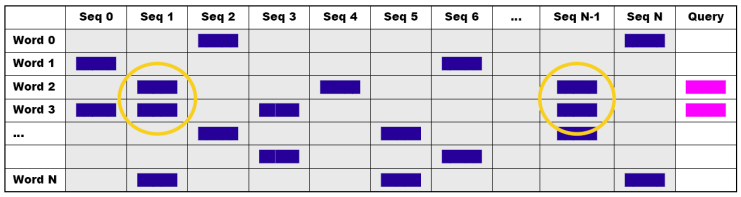
- Database:
  AIKWQPRSTW...
  IKMQRHIKW...
  HDLFWHLWH...
  .....
- Query:
  RGIKW
- Output: sequences *similar* to query

W-mer indexing (hashing)

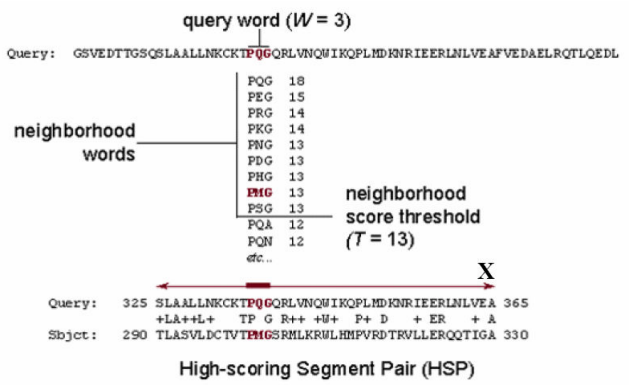
- Preprocessing:
  For every W-mer (e.g., W=3) store every location in the database where it occurs
- Query:
  Generate W-mers and look them up in the database.

FASTA

R = position(query) – position (DB).



Basic Local Alignment Search Tool (BLAST)



– 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

Significance of scores

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

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

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:

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

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

NCBI Blast

Program	*Query sequence*	*Subject sequence*
Nucleotide BLAST	Nucleotide	Nucleotide
Protein BLAST	Protein	Protein
blastx	Nucleotide six-frame translation	Protein
tblastn	Protein	Nucleotide six-frame translation

NCBI Blast Example

blastn | blastp | blastx | tblastn | tblastx

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

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) — Clear — 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 — YouTube — Create custom

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

Algorithm parameters

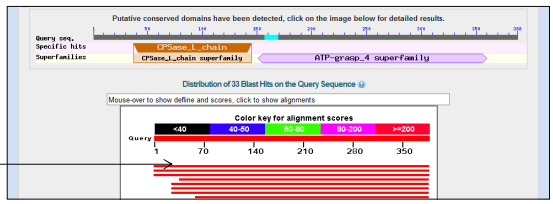
General Parameters

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 m

Blast Results



conserved domain database (CDD)

graphical visualization

Best hit

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_032802.2
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-105	48%	NP_076133.3
PREDICTED: propionyl-CoA carboxylase alpha chain, mitochondrial isoform X4 [Mus musculus]	322	322	93%	1e-103	47%	XP_006518496.1

description

E-value

Score (S)

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%)

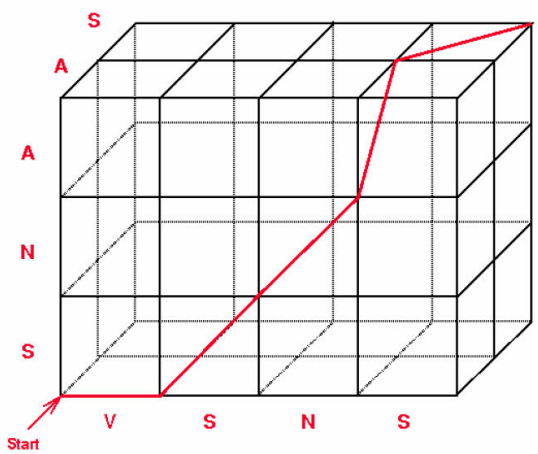
```
Query  1    MLKFRTVHGGLRLLGIRRTSTAPAASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG  60
            MLKF+TV GGLRLLG+RR+S+AP ASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG
Sbjct  2    MLKFQTVRGGLRLLGVRRSSSAPVASPNVRRLEYKPIKKVMVANRGEIAIRVFRACTELG  61
                                  ...
Query  361  GLRQENIRINGCAIQCRVTTEDPARSFQ  388
            GLRQENIRINGCAIQCRVTTEDPARSFQ
Sbjct  362  GLRQENIRINGCAIQCRVTTEDPARSFQ  389
```

alignment

...

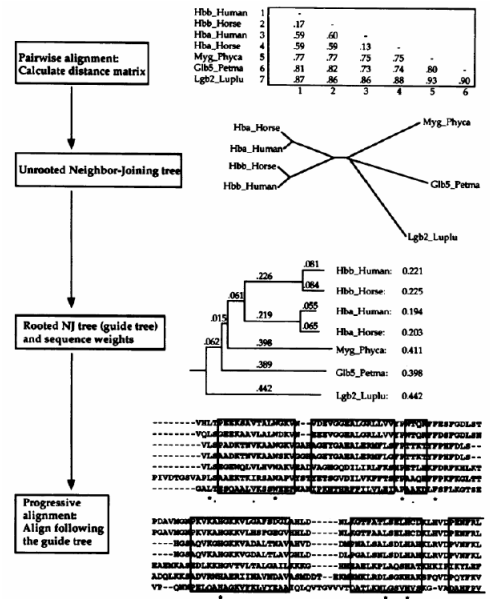
Multiple sequence alignment

Dynamic Programming



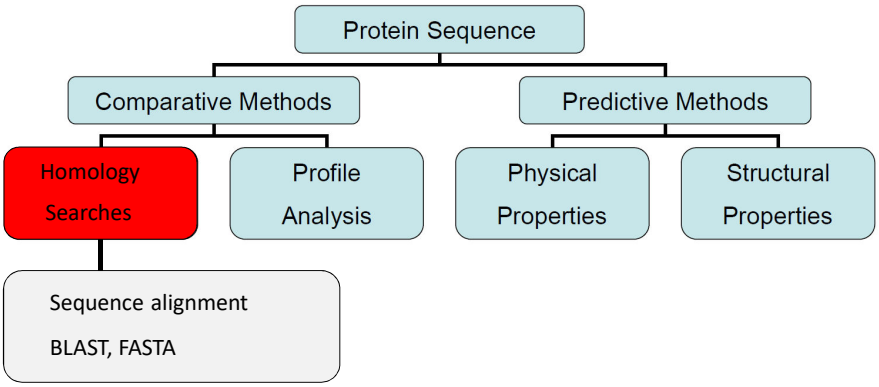
V S N _ S
_ S N A _
_ _ _ A S

Progressive tree alignment (ClustalW)

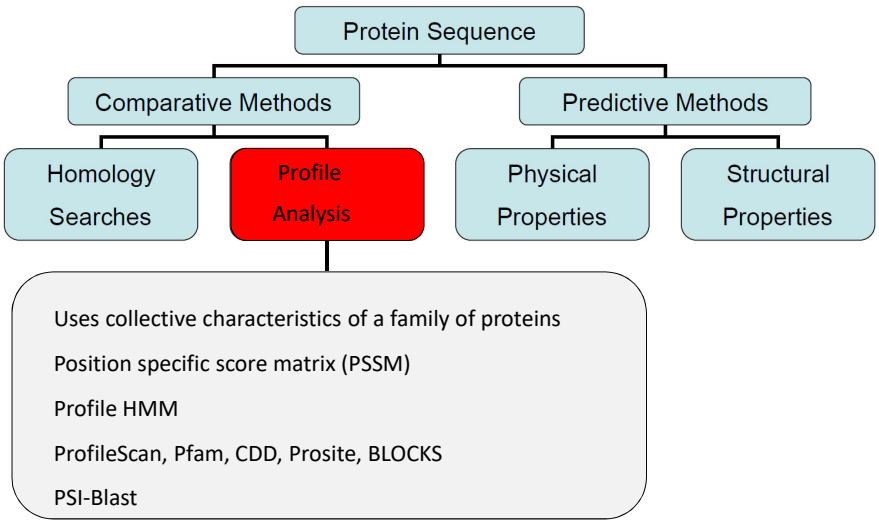


Predictive methods using protein sequences

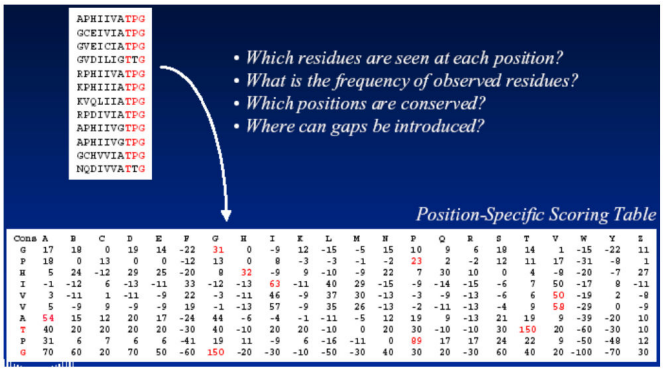
Homology searches



Profile Analysis



Profile Construction

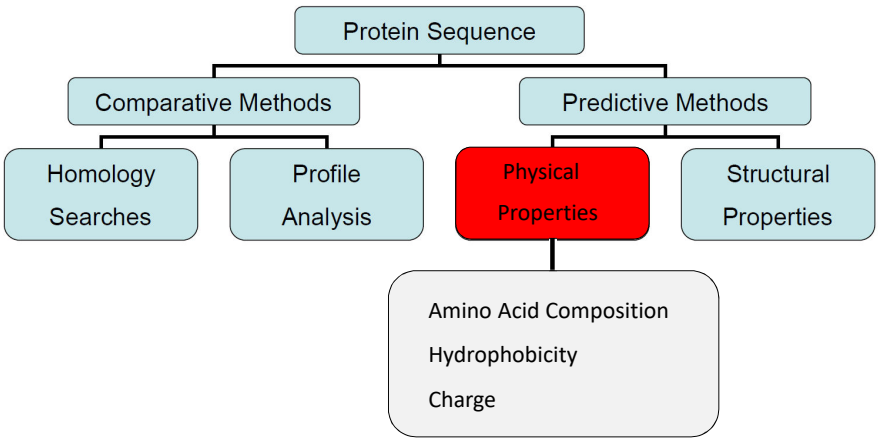


Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-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	89	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

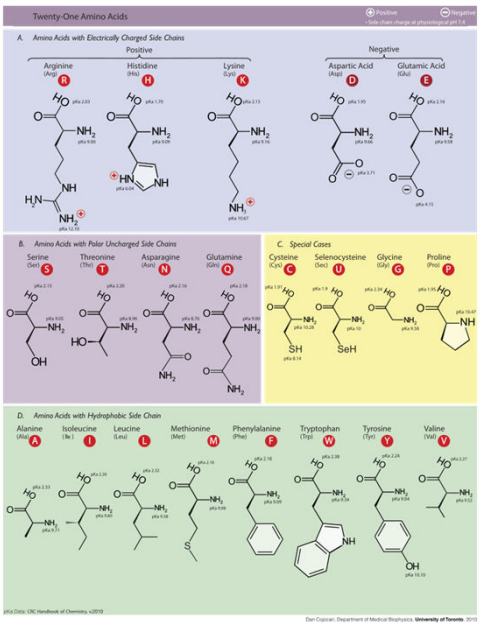
$$\text{PSSM}(p,a) = \sum_{b=1}^{20} f(p,b)*s(a,b)$$

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)

Protein Sequence Analysis



Amino Acids

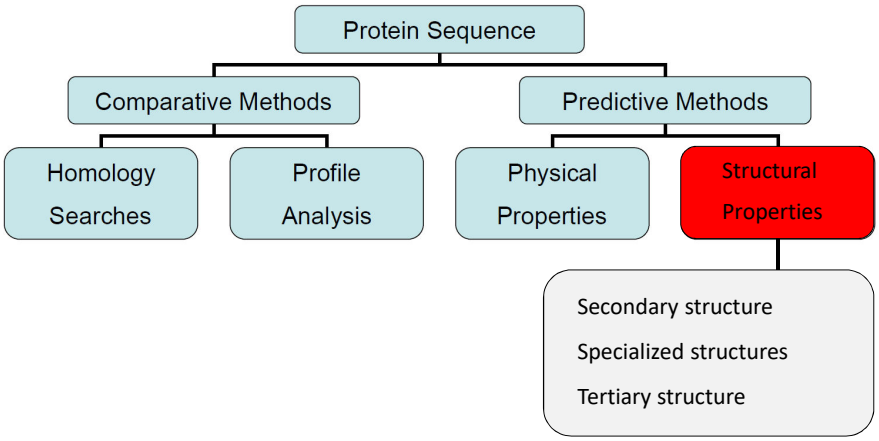


ProtParam

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

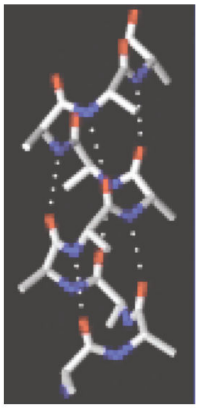
http://web.expasy.org/protparam

Protein Sequence Analysis



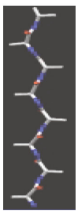
Alpha-helix

- 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



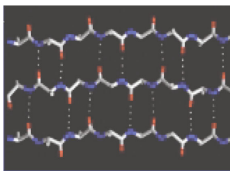
Beta-strand

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

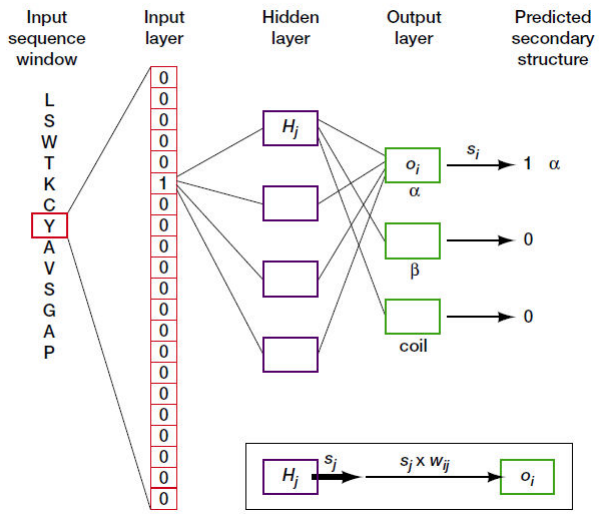


Beta-sheet

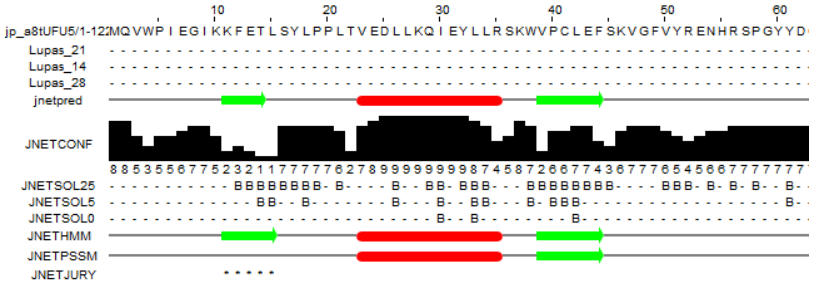
- Stabilization through hydrogen bonding
- Parallel or antiparallel
- Variant: beta-turn



Neuronal network for secondary structure prediction

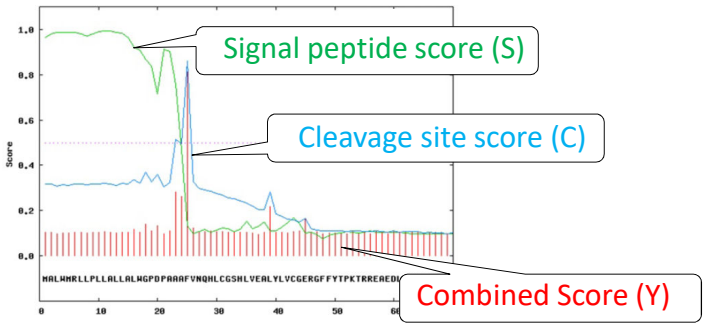


Protein secondary structure prediction (Jpred)



SignalP

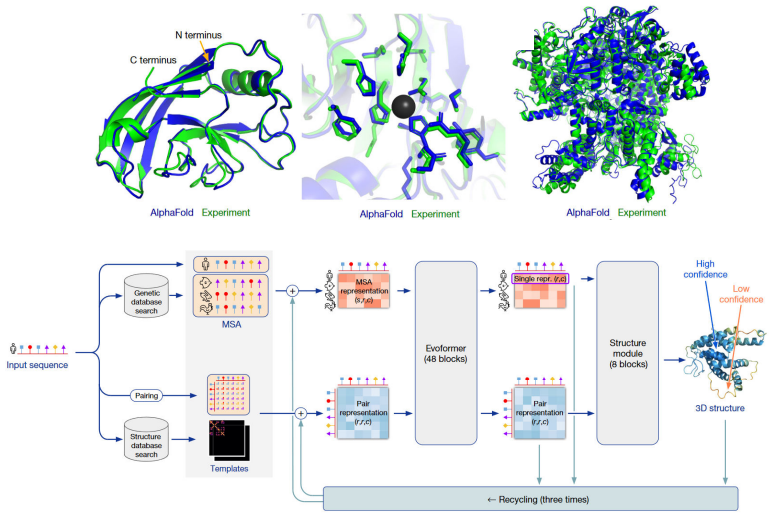
- Neural network trained based on phylogeny
  - Gram-negative prokaryotic
  - Gram-positive prokaryotic
  - Eukaryotic
- Predicts secretory signal peptides
- http://www.cbs.dtu.dk/services/SignalP/



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

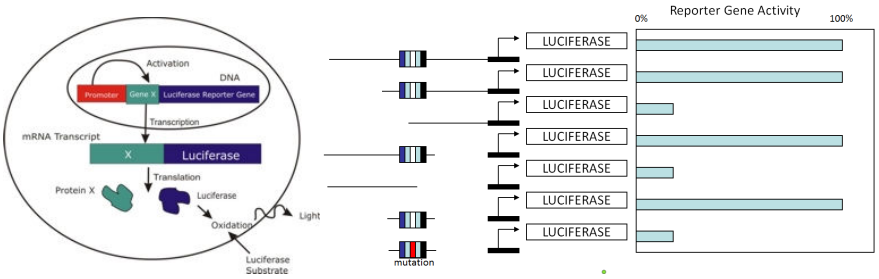
- Transcription factor binding sites
  - Experimental methods
  - Computational methods
    - Matrix based methods
    - Motif discovery
- MicroRNA target prediction

Transcription factor binding sites

Experimental methods

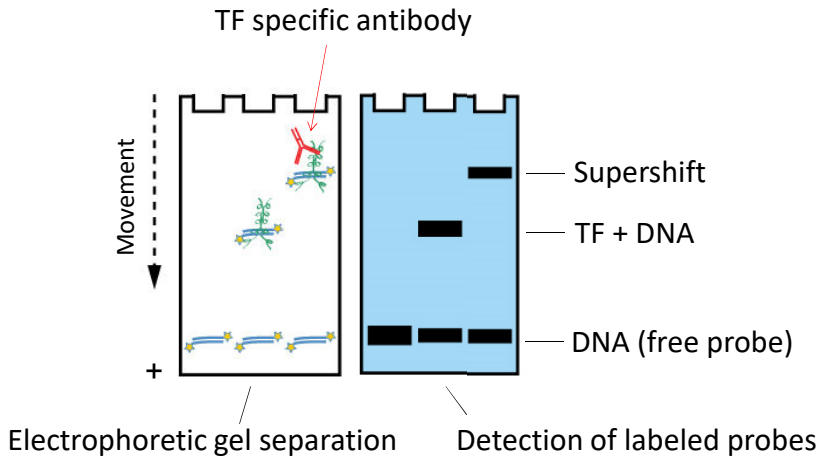
- Reporter gene assays (luciferase)
- Electro mobility shift assays (EMSA)
- DNase I and Exonulease Footprinting
- SELEX
- Chromatin immuno precipitation (ChIP)

Luciferase reporter assays

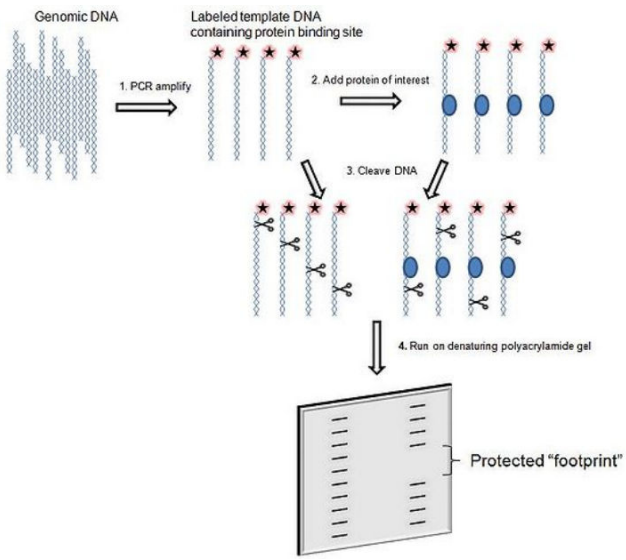


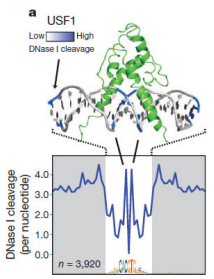
- Identify functional regulatory region within a sequence and delineate specific TFBS through mutagenesis
- Evidence that TF binding has an effect on transcription (not only binding to DNA)

Electromobility/Gel Shift Assays



DNase I and Exonuclease footprinting



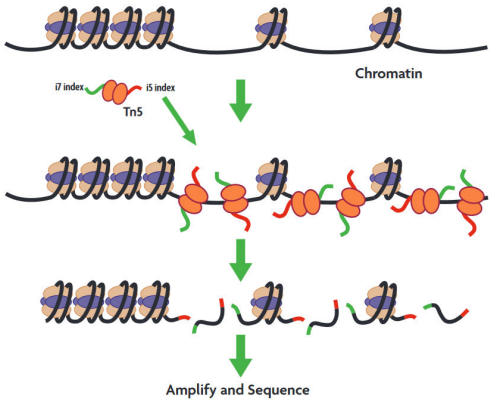


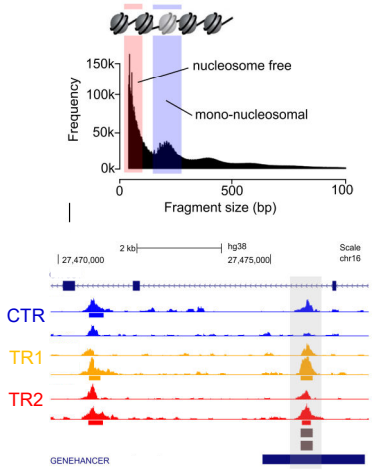
Neph et al., Nature, 2012

⇨ Dnase-seq
FAIRE-seq

ATACseq

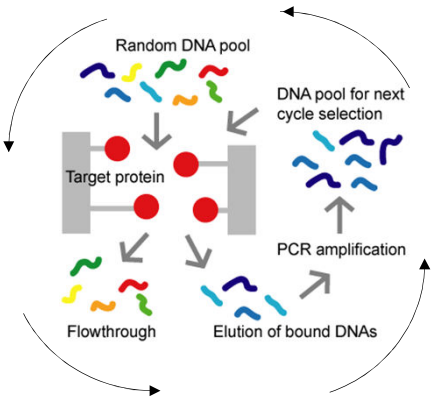
Assay for **T**ransposase-**A**ccessible **C**hromatin with **seq**uencing





SELEX

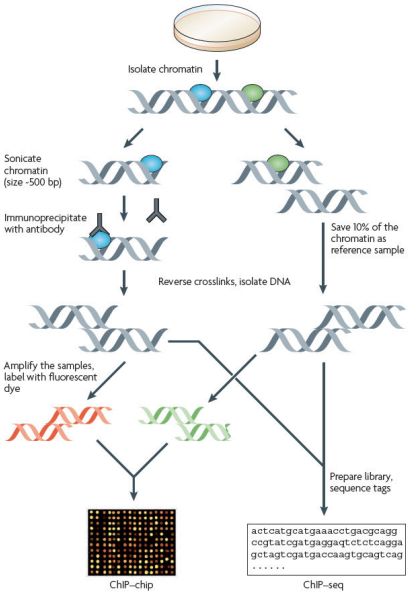
Systematic evolution of ligands by exponential enrichment



several cycles

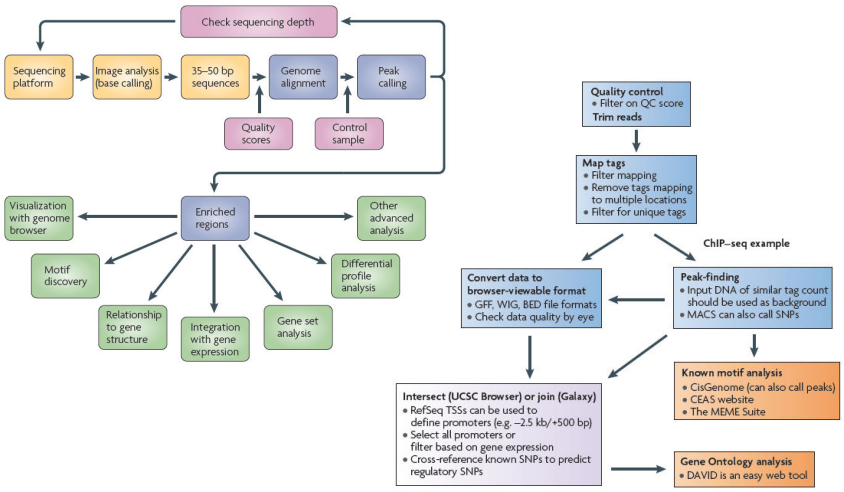
Most position weight matrices (PWMs) in the databases are derived by SELEX

ChIP procedure



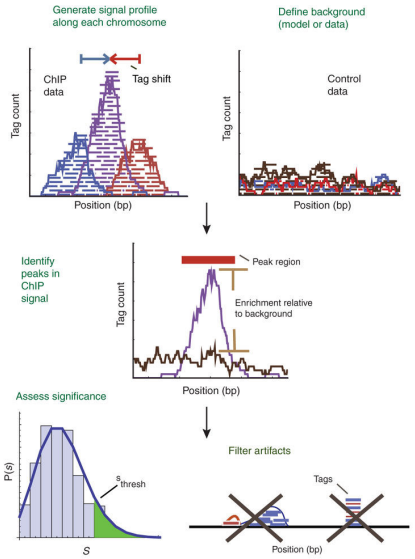
Farnham, Nature Rev Genetics, 2009

ChIP-seq analysis



Hawkins et al., Nature Rev Genetics, 2010

ChIP-seq (Peak calling)

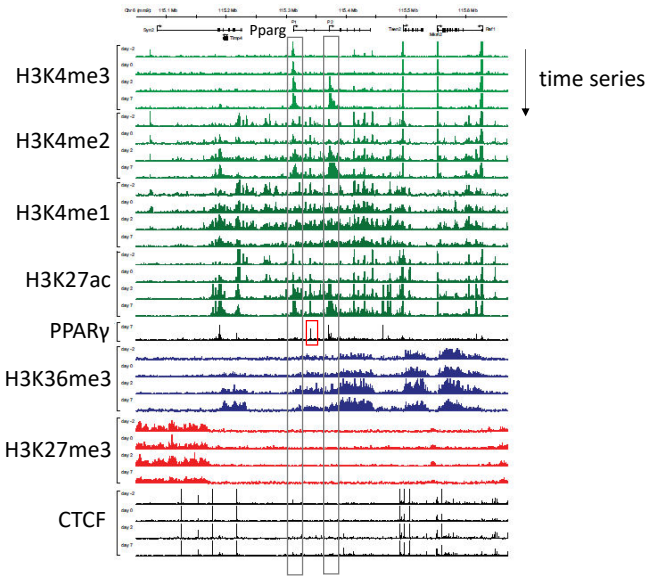


Tools:

- 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

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

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

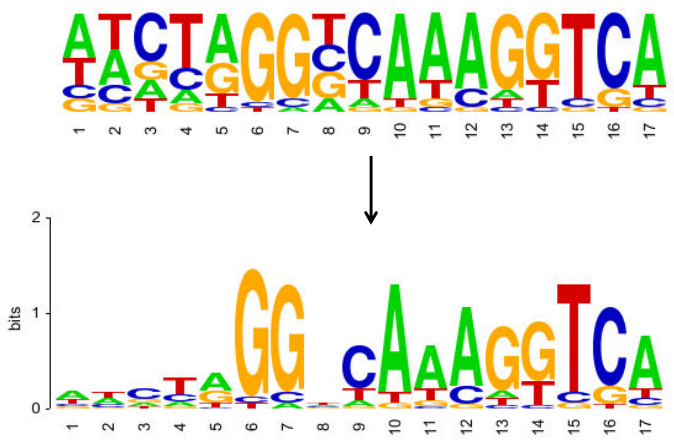
	1	2	3	4	5	6
A	1.00	-1.32	-2.46	-2.32	-1.46	1.79
C	-2.46	-2.46	-1.46	0.26	1.45	-1.32
G	0.35	1.86	1.79	0.26	-2.46	-2.46
T	-0.87	-2.46	-1.32	0.49	-0.46	-1.46

…ACGTAGGTCATAGAGTA.. S=1+1.86+1.79+0.49+1.45+1.79=8.38

…ACGTAGGTCATAGAGTA.. 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

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

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

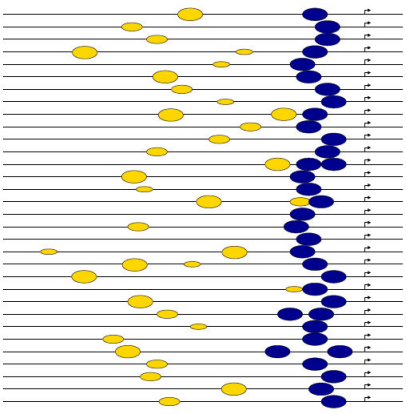
Only one base is present at position i

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

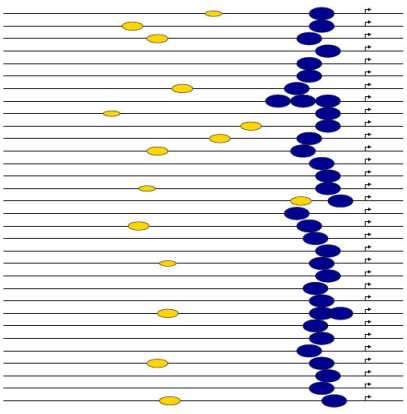
from pseudocounts ($\log_2 0$ is not defined!!)

Using a set of background sequences

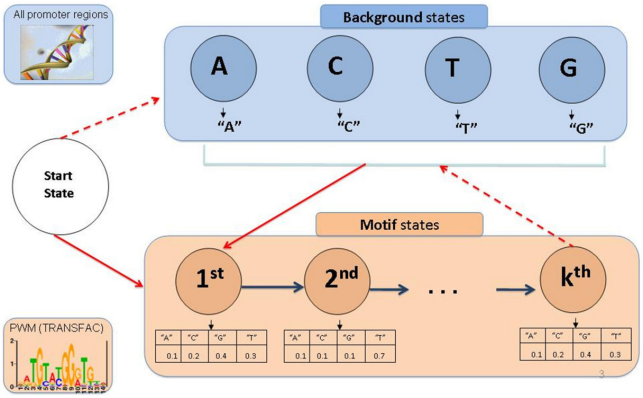
Foreground sequences



Background sequences



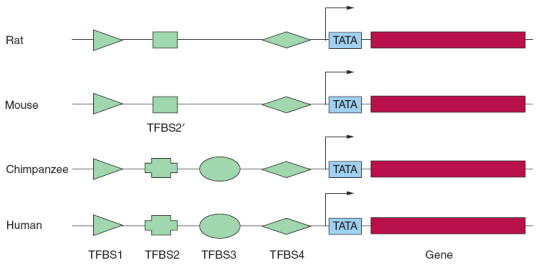
Profile hidden markov models (HMM)



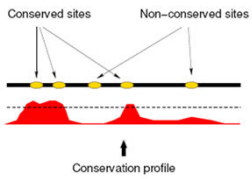
Levkovitz et al. PLoS One. 2010

Phylogenetic footprinting

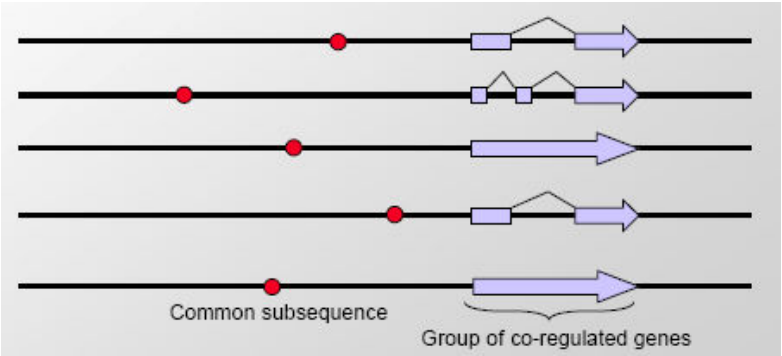
- Functional regulatory sites are conserved between species



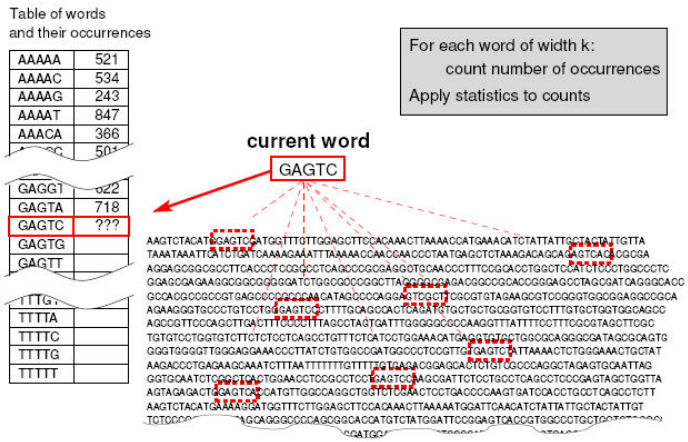
- Multiz alignment of UCSC genome browser



Motif discovery



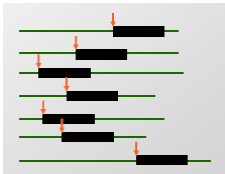
Word counting



Compared with background sequences (Fischer exact test => p-value)

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*

X_i = G C [T G T] A G

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

$$\Pr(X_i \mid Z_{i3} = 1, p) =$$

$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =$$

$$0.25 \times 0.25 \times \boxed{0.2 \times 0.1 \times 0.1} \times 0.25 \times 0.25$$

Z

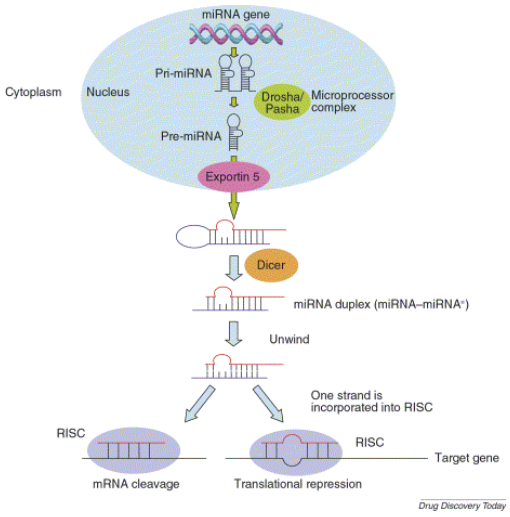
The element Z_{ij} of the matrix Z represents the probability that the motif starts in position *j* in sequence *i.*

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

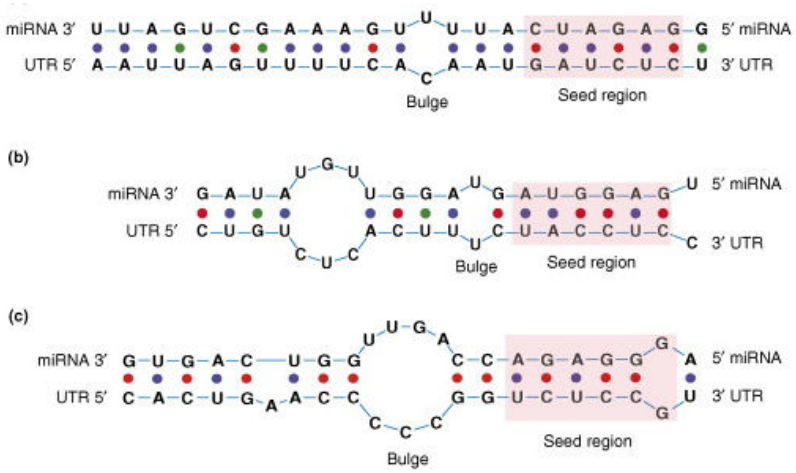
– The basic EM approach has been enhanced by MEME (ChIP-MEME)

MicroRNA target prediction

microRNA biogenesis



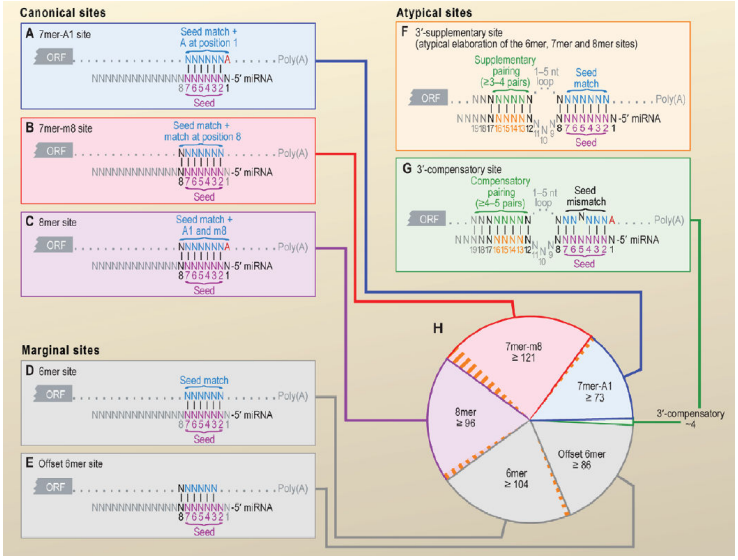
microRNA/mRNA pairing



Principles of microRNA target prediction

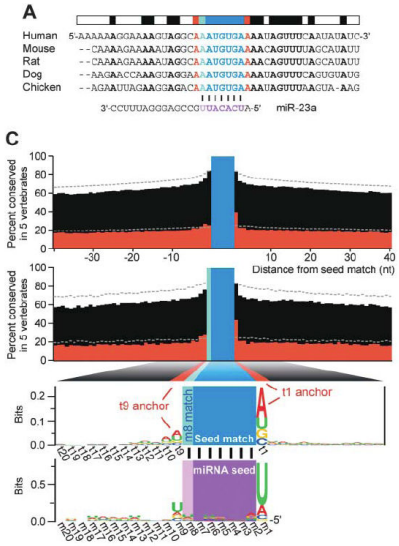
1. Sequence complementarity
2. Conservation
3. Thermodynamics
4. Site accessibility
5. UTR Context
6. Anticorrelation of expression profiles

Sequence complementarity



Bartel, Cell,2009

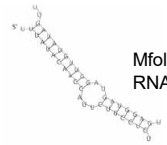
Conservation



Lewis BP et al., Cell, 2003

Thermodynamics

1. Minimum free energy



⇨ e

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

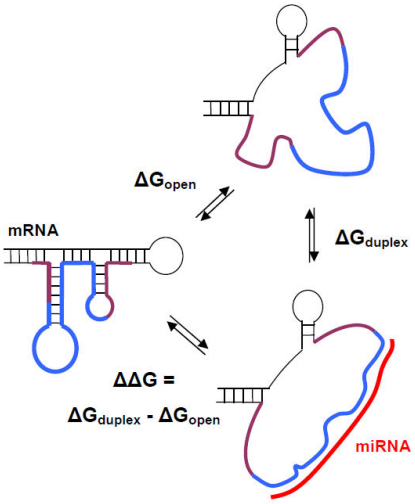
```
          mfe: -25.3 kcal/mol
           p-value: 0.010068

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



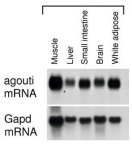
Leitner A, 2009

III Gene expression analyses

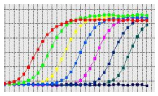
- Microarrays
- RNA sequencing
- Gene expression profiling
- Clustering and classification
- Gene ontology

Gene expression analyes

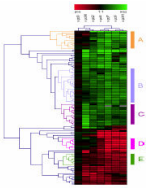
- Northern bloting
  - semi-quantitative
  - few genes



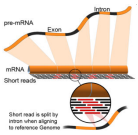
- Real time RT-PCR (qPCR)
  - medium throughput
  - 96/384 per run



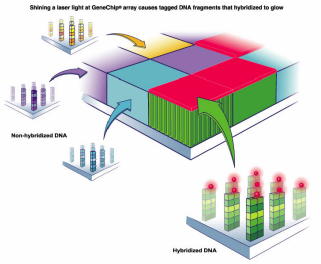
- Microarray analysis
  - high throughput
  - 10.000-500.000 elements per chip

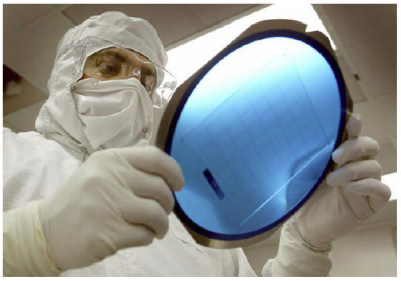


- RNA seq
  - high throughput
  - deep sequencing (short reads 25bp)

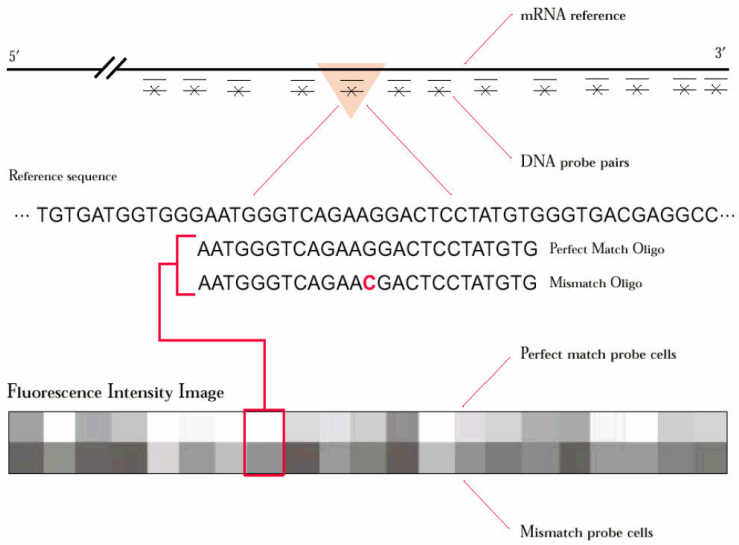


One color microarrays (Affymetrix)





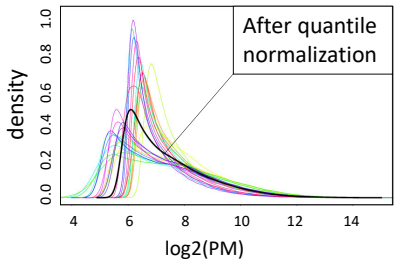
Affymetrix chips



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

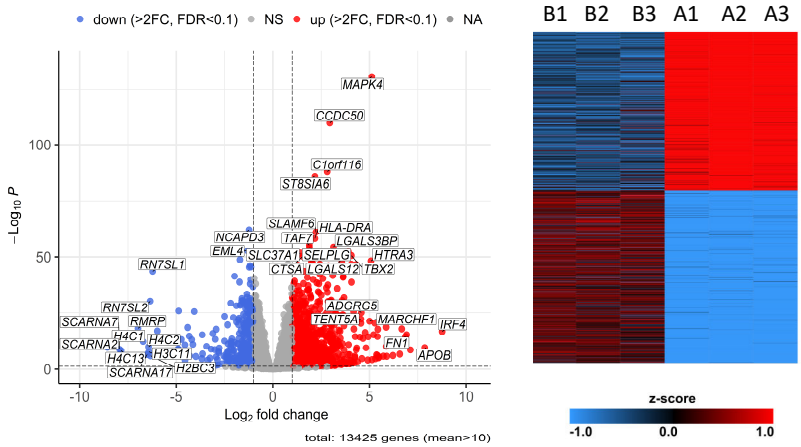
16134 probesets

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

Condition A vs. B



Differentially expressed genes

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

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}}$$ => 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

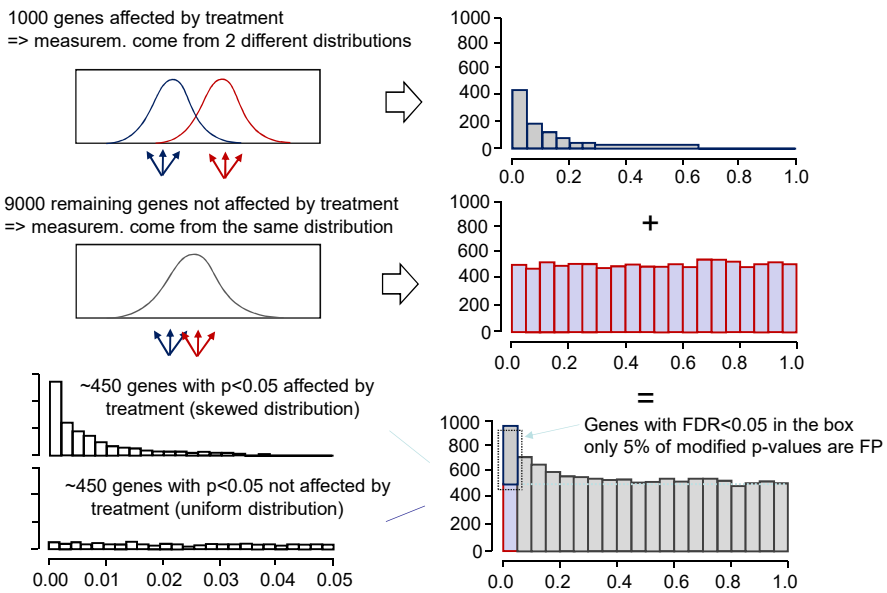
	Ranked p	Bonferroni	Benjamini-Hochberg (FDR)
smallest p →	$p_{(1)}$	$p_{(1)}*n$	$p_{(1)}*n$
	$p_{(2)}$	$p_{(2)}*n$	$p_{(2)}*n/2$
	..	..	..
	$p_{(i)}$	$p_{(i)}*n$	$p_{(i)}*n/i$
	..	..	..
	$p_{(n-1)}$	$p_{(n-1)}*n$	$p_{(n-1)}*n/(n-1)$
largest p →	$p_{(n)}$	$p_{(n)}*n$	$p_{(n)}$

..

keep smaller one

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

P-value distribution



Josh Starmer (StatQuest)

Deep (next generation) sequencing technologies

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

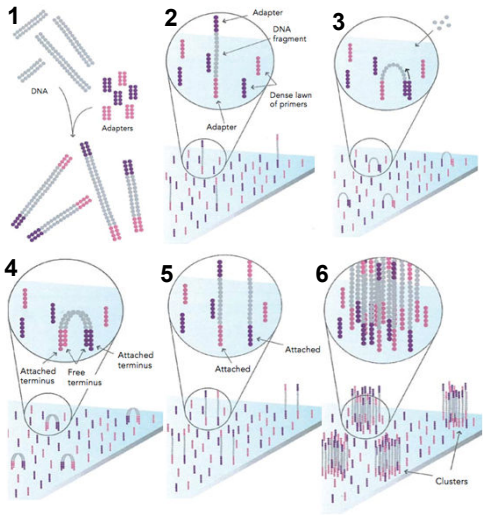
- 454 (Roche)
- Solexa (Illumina)
- Solid (Thermo Fisher Scientific)
- Ion Torrent (Thermo Fisher Scientific)

— 2nd gen. (ampl)

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

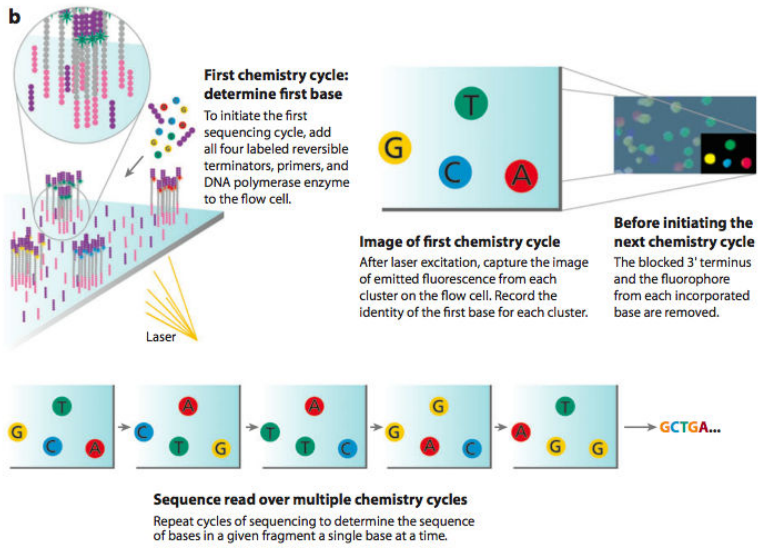
— 3rd gen. (no ampl)

Solexa (Illumina)

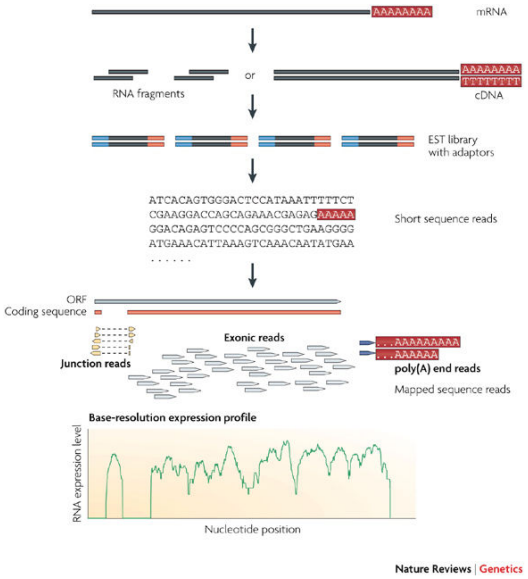


1. Prepare genomic DNA sample
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature double stranded DNA
6. Complete amplification

Solexa (Illumina)



Transcriptome sequencing (RNAseq)



Wang et al., Nature Rev Gen, 2009

Phred Quality Score

Q=-10* log P

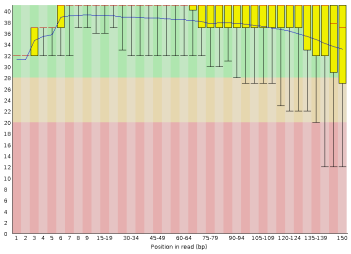
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

fastq format

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

← Q in ASCII

Quality of Sequencing (FASTQC)



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) (DESeq2, edgeR)

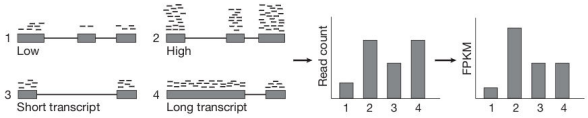
=> List of genes with log2FC, p-value, FDR, average expression

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

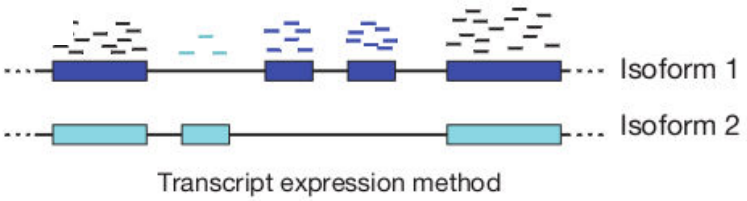
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1
Tens(Mio)	1.5	2.025	4.51

2. Divide by millions of RPK

TPM

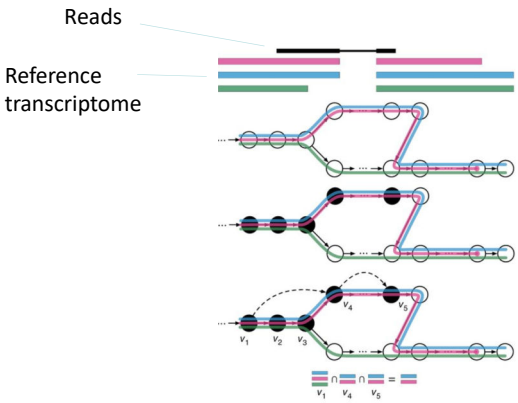
A (2kb)	3.33	2.96	3.326
B (3kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

Isoform quantification



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

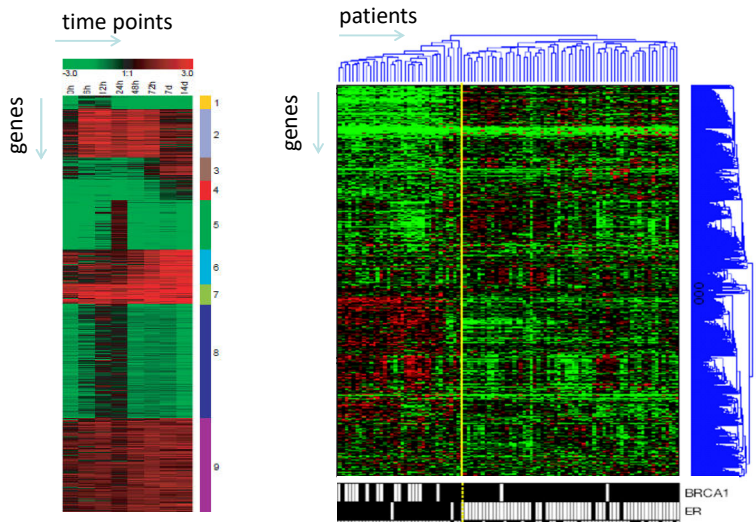
RNA seq quantification using pseudoalignment (kallisto)



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

Bray et al. Nature Biotechnology 2016

Gene expression profiling



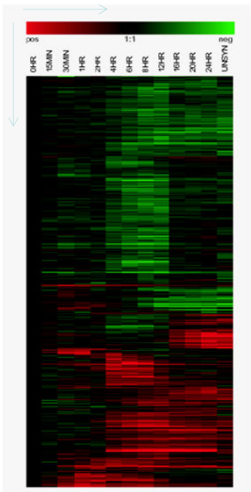
cell development

cancer

Representation of gene expression

conditions (patients) (m)

genes (n)



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:

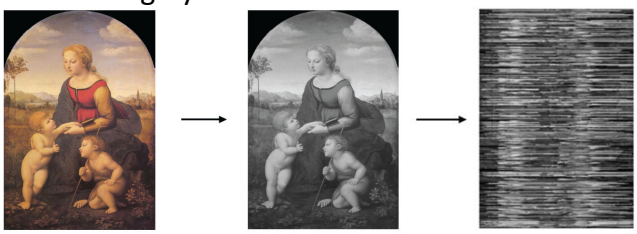
- log2-fold change (log2FC, log2(A/B), log2 ratio)

For patient samples and no reference:

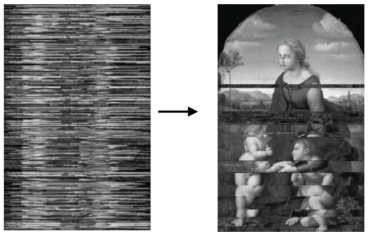
- Mean (median) centered log2-levels for each gene
  log2-intensities for one-color arrays
  log2-RPKM for RNAseq
- z-score of log2-levels
  Z= (X-m)/s m…mean,
  s…standard deviation

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

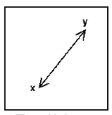
- Pearson correlation

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

$-1 \leq r \leq 1$

- Euclidian distance

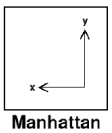
$$d_E = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



Euclidean

- Manhattan distance

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

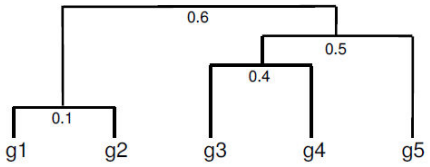


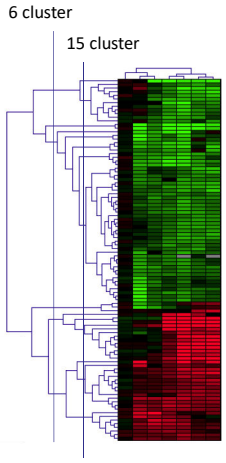
Manhattan

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)

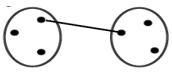
1. Identify clusters (items) with closest distance
2. Join to new clusters
3. Compute distance between clusters (items) (see linkage)
4. Return to step 1



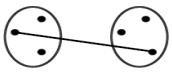


Linkage

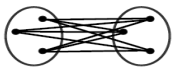
– Single-linkage clustering
  Minimal distance



– Complete-linkage clustering
  Maximal distance



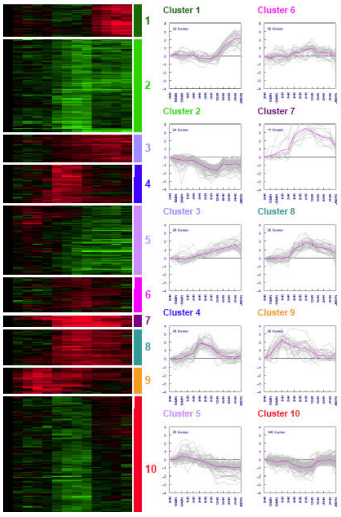
– Average-linkage clustering
  Calculated using average distance (UPGMA)
  Average from distances not! expression values



K-means

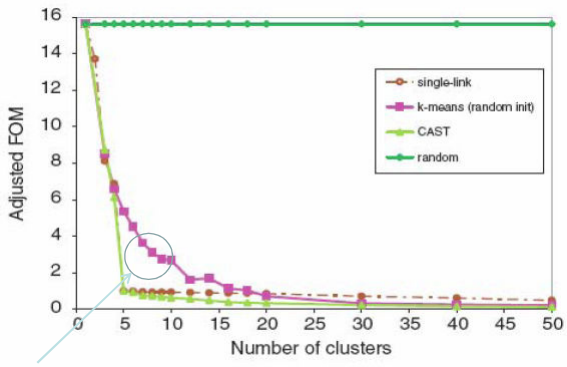
- partition *n* genes into *k* clusters, where *k* has to be predetermined
- k-means clustering minimizes the variability within and maximize between clusters
- Moderate memory and time consumption

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

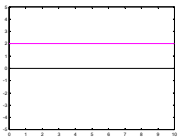
Figure of Merit (FOM)



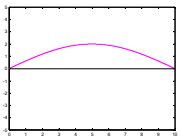
choose k here (e.g. k=8)

Principal Component Analysis (PCA)

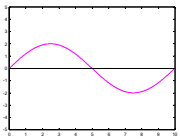
Is it possible to represent each profile by overlay of few patterns?



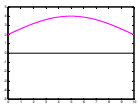
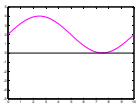
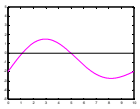
PC1



PC2



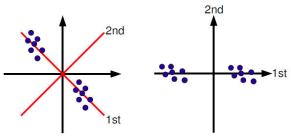
PC3

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.

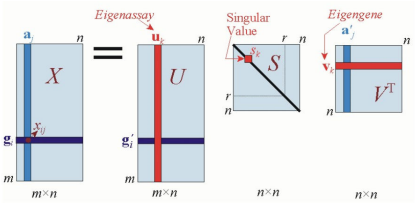


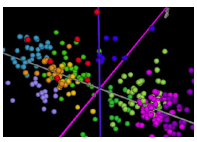
Subsequent principal components are orthogonal to the 1st PC. With the first 2 PCs usually 80-90% of the variance can already be explained.

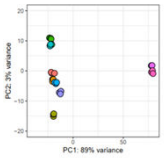
This analysis can be done by a special matrix decomposition (singular value decomposition SVD).

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:

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

Other dimension reduction methods

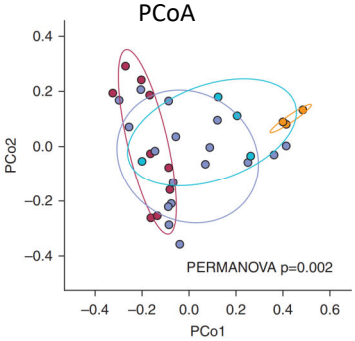
Metric multidimensional scaling (MDS)
Principal Coordinate Analysis (PCoA)

Bray-Curtis dissimilarity index
in microbiome analyses

Non-linear transformation
Keeps local and global structures

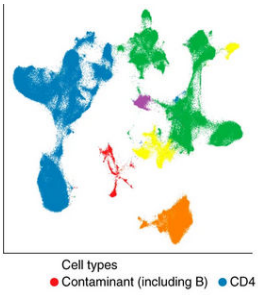
Single cell analysis

PCoA

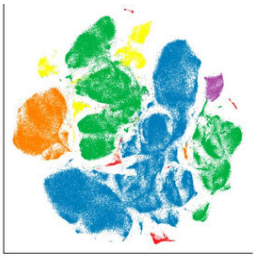


Moosbruger-Martinz et al. J Invest Derm 2020

UMAP

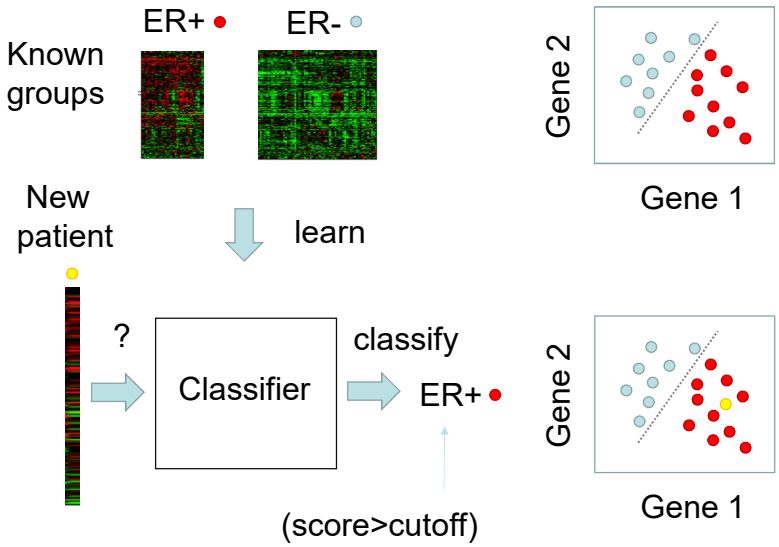


t-SNE

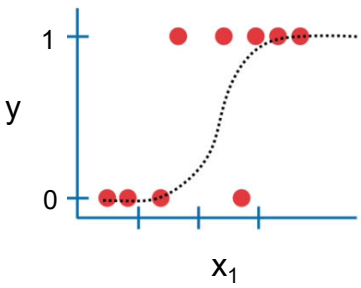


Becht E et al. Nat Biotechnol 2018

Classification



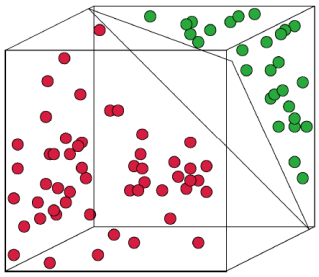
Logistic regression

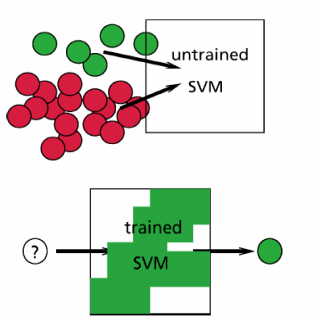


$\ln(P/(1-P)) = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots$

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

Support vector machines (SVM)



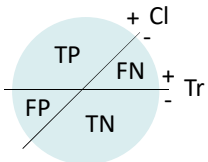


A SVM tries to find an optimal hyperplane that separates all training samples correctly and maximizes the margin (maximizes the distance between it and the nearest data point of each class). If this is not possible in the input space (for example in 2 dimensions) a hyperplane can be found in the higher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

Receiver operator characteristics (ROC)

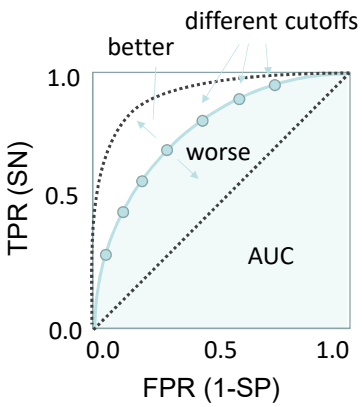
Truely

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



Sensitivity
SN=TP/(TP+FN)

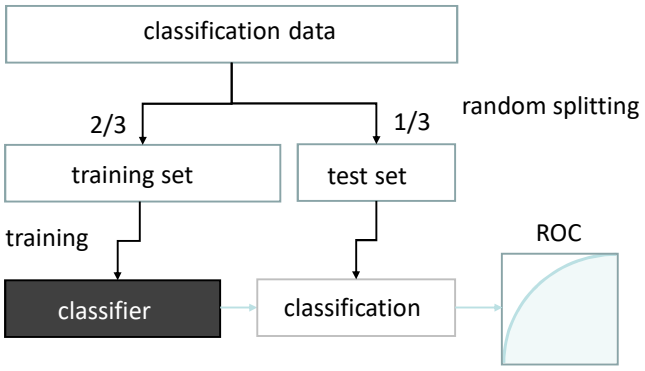
Specificty
SP=TN/(TN+FP)



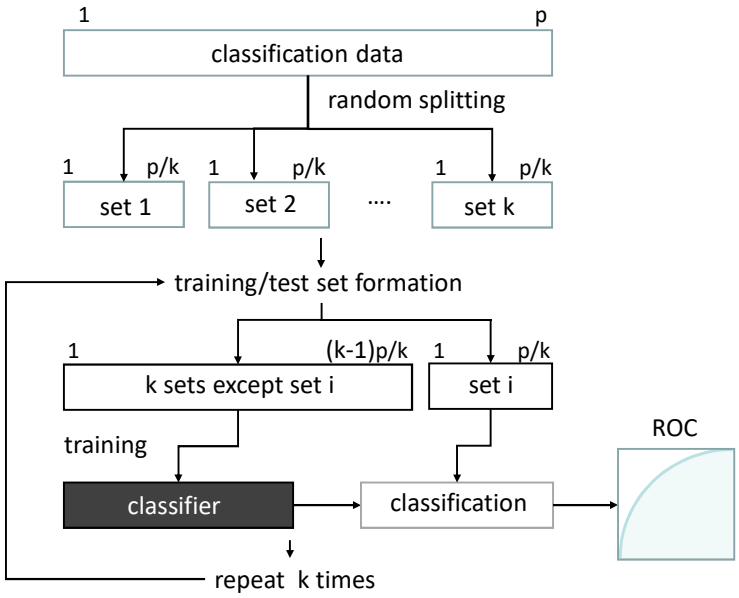
Area under curve (AUC)
AUC=1.0 optimal
AUC=0.5 random

Holdback cross validation

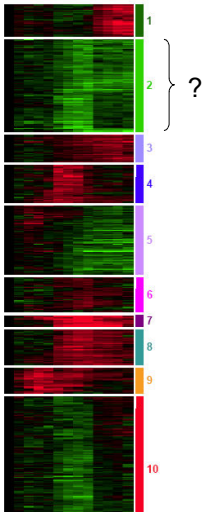
To avoid overfitting data should be splitted into training and test set



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

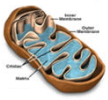
Gene Ontology (GO)

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

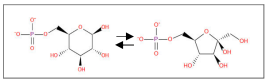


- biological process

cell cycle

- molecular function

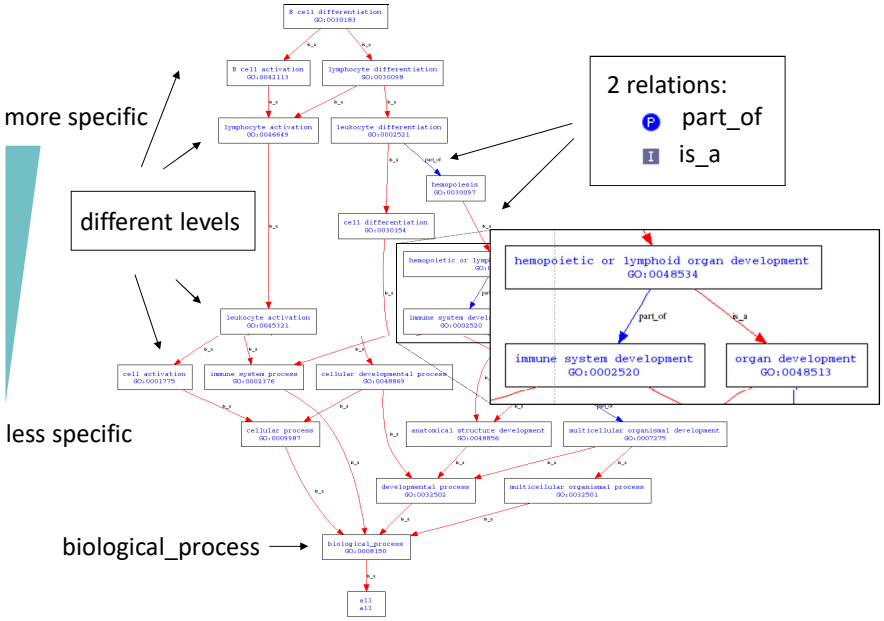
isomerase activity



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)



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

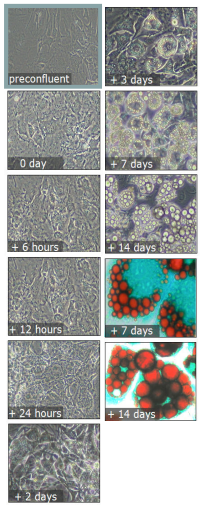
...

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)

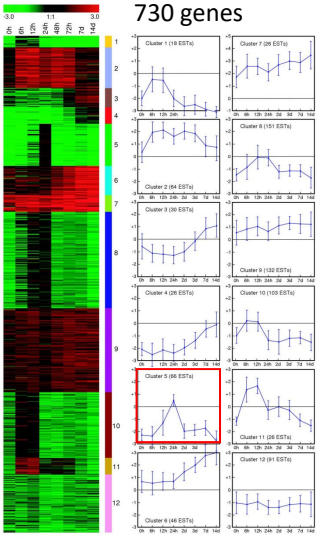


Microarray analysis

⇨

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

730 genes



Heatmap (k-means clustering)

Biological function of genes in clusters

⇨

GO Analysis

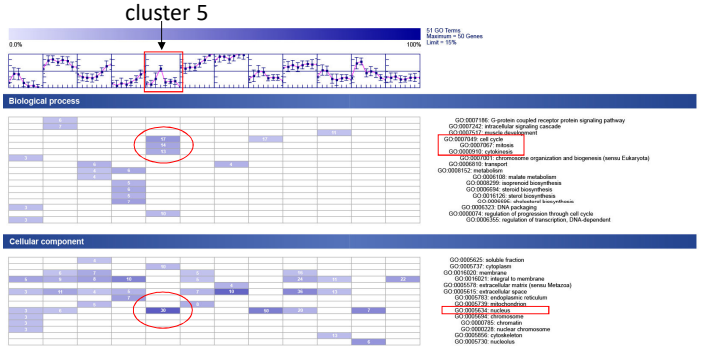
Hackl H, Burkard TR et al. Genome Biol. 2005

GO terms for gene sets

cluster 5

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

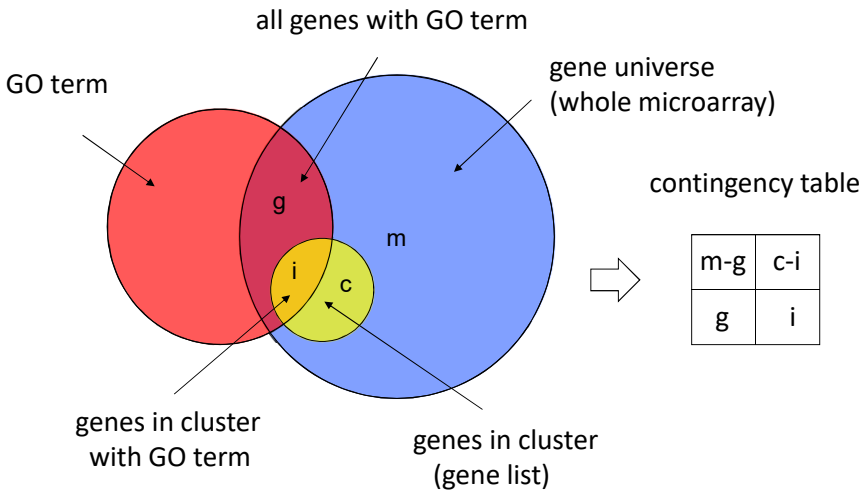
nucleus (30)



– 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

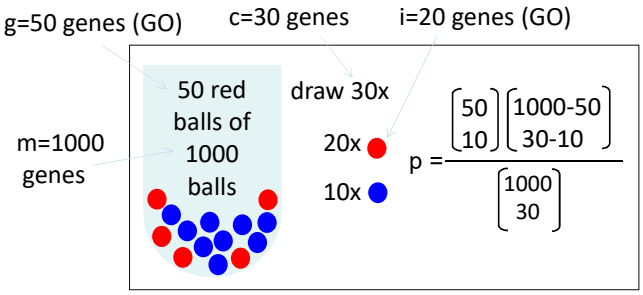


Over representation analysis

- Fisher exact test for contingency table

m-g	c-i
g	i

- Hypergeometric distribution



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

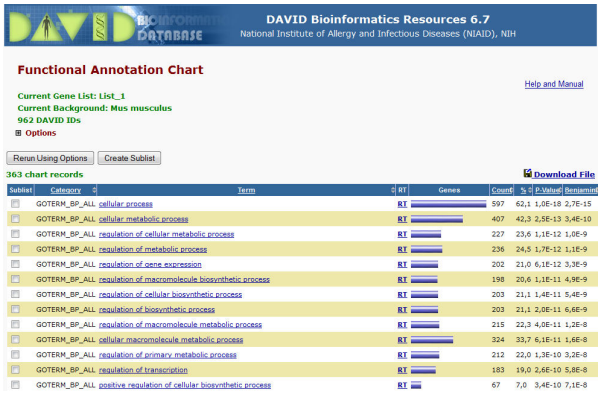
- Multiple hypothesis testing => adjust p-value
- Not only for GO Terms also for TFBS, pathways,..

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
...



...