Bioinformatics and Computerscience

Systems Biology

Data collection

1. ARRAY BASED

2. NEXT-GEN SEQUENCING
RNA-Seq analysis
ChIP-seq
Bulked segregant analysis

Network Inference

1. Sequence-based data analysis
MotifSuite
ModuleDigger
Crossed

2. Network reconstruction
Lemone
Distiller
Comodo
Bayesian network reconstruction

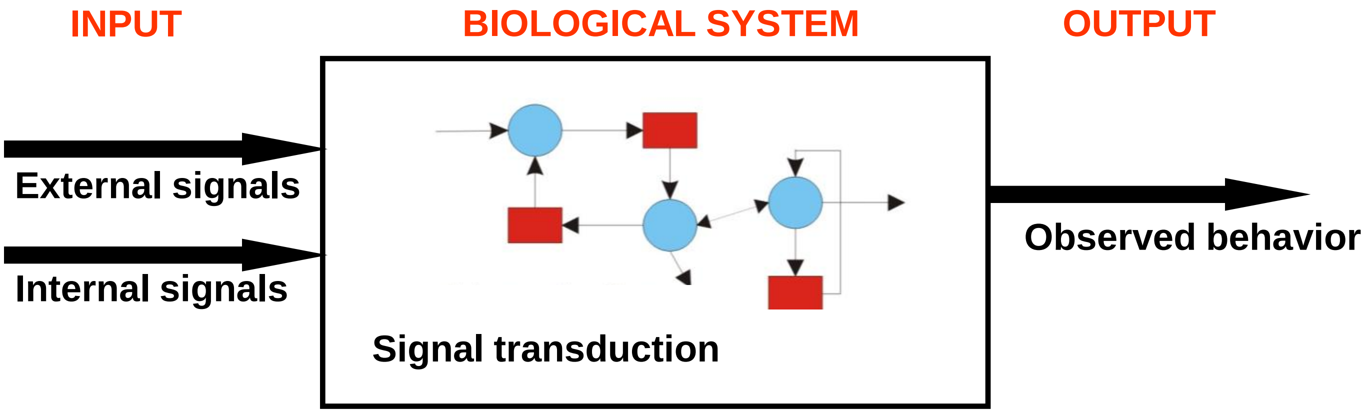
Network-based dataintegration

1. Network-based analysis of unstructured gene lists
2. Network-based gene prioritization
3. Network-based eQTL analysis
4. Network-based subtyping

- Development of methods that assist systems biologists
- Methods based on data-mining, statistics
- Unique in combining biologically relevant assumptions with rigorous statistical and datamining framework (pragmatic but not too much ad hoc)
- All tools have been validated on real biological cases

http://bioinformatics.psb.ugent.be/DBN/dbn/software

Systems Biology



INPUT

BIOLOGICAL SYSTEM

OUTPUT

External signals

Internal signals

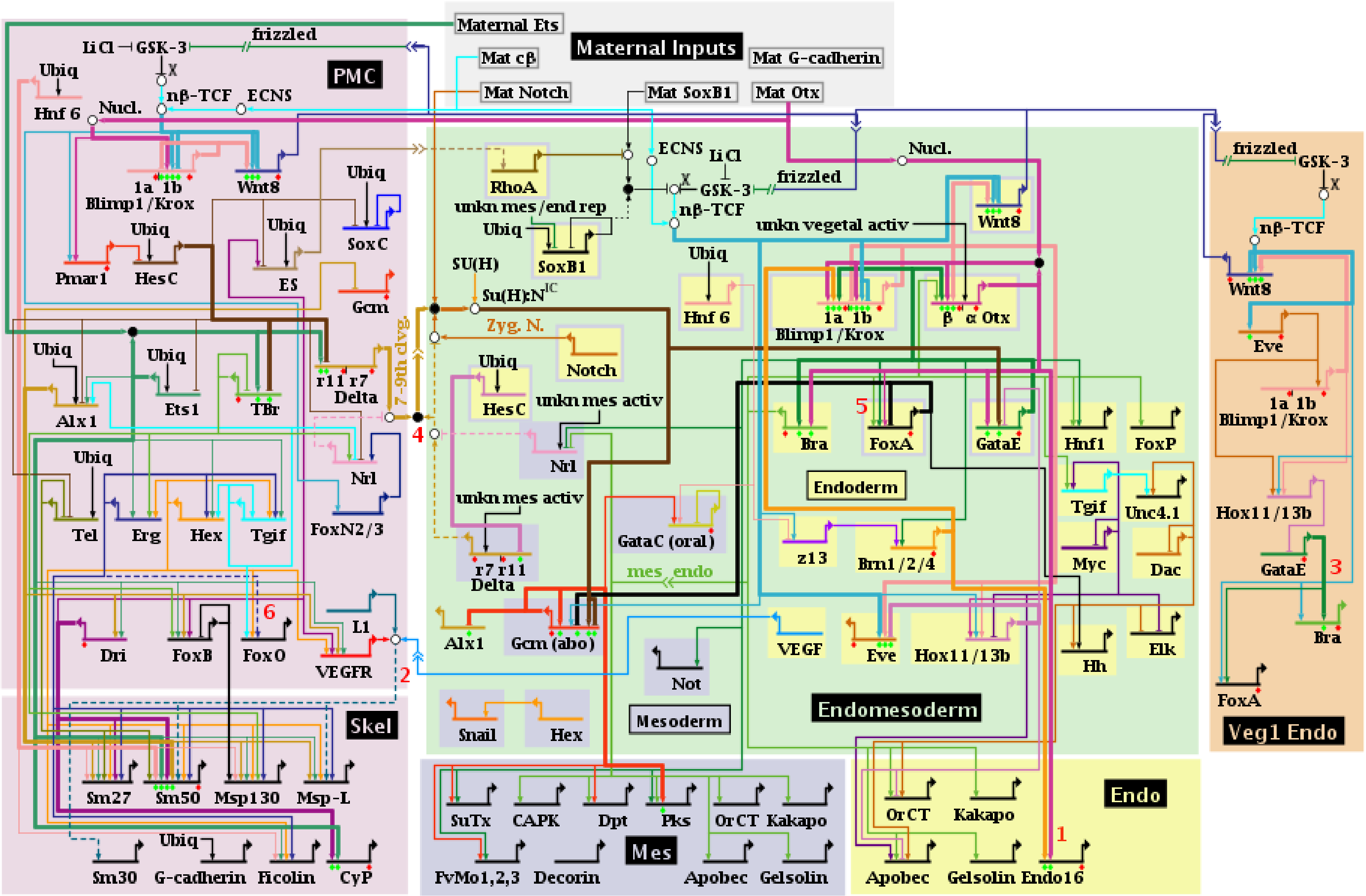
Signal transduction

Observed behavior

Systems Biology

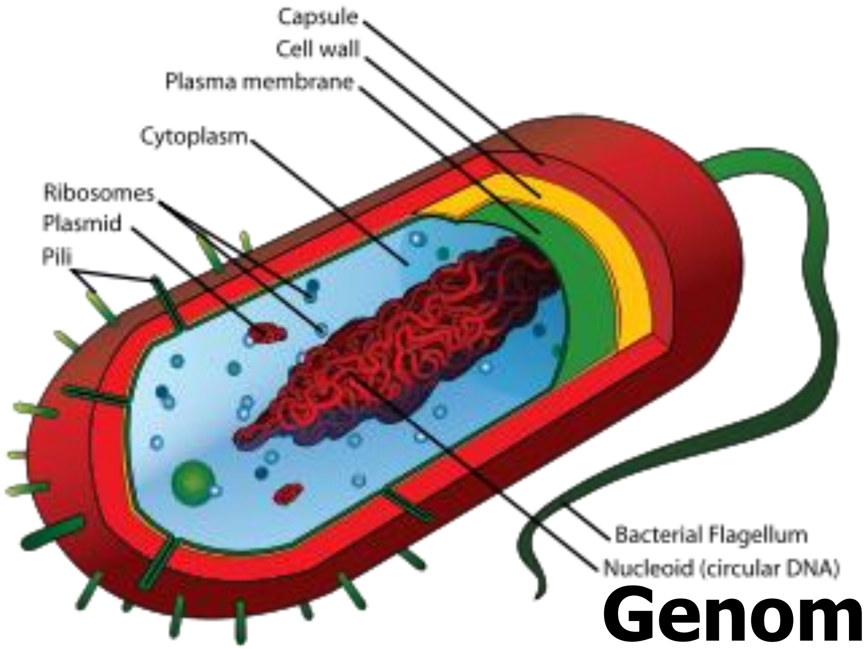
Endomesoderm Specification to 30 Hours

March 10, 2008



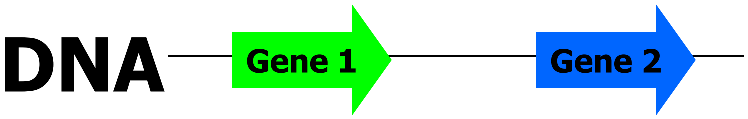
Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; χ = β-catenin source;
nβ-TCF = nuclearized b-β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Systems Biology



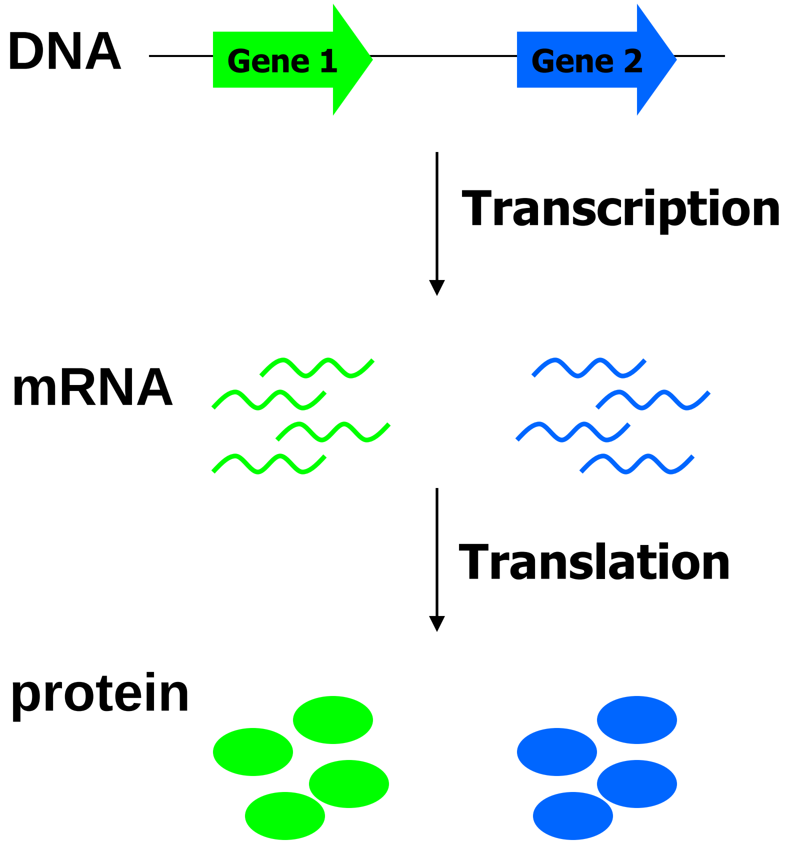
Genome

Systems Biology



AGCACTGTCCACTGCATGGTGAGGATGGGGGTGAGCTCCCT
TTGTGGCTAGGTGCTTAAACGTCTATCGGACGCTCA**GTGAA**
GGGCTATTGCGGGGTAACAGGAAACCCTGAGGTGCT
GATAGGTCAAAGATGGAGAAGGCGTAGGGATATGTG
TTGGCAGAGGAACCGAAGAATACCAGGCCATTTCCG
AAGCCACTCATTTTCCTCGAATTCCTTCTTTATGCCTTC
AGTCTCTATTGACCGTAAATTTGGTTGTTGTCTCCCAGCTGT
TTATTTCTGTAACAGATCTTGGAGGCTGCGGTCTGGATCCCT
CGCCAAGAACCAGATCCAGGAGAAAACGTGCTCAACGTGCA
GCTCTGCTCCTACTGATTATAGCCCCACAGATGACATCGCTC
CATAGTCACACCAAGTCTCCTGTGGGAGTCTTGCTCCTCGTT
CTCAGTGTCTGTTACAGCTCGGTATTTTAGTGTCAGGACGTC
GGCTCCCAGCCCGCATCTCCGCTCAGCAATGCCATTATCTTC
TCAGCCAAGTCCTAGAAATGGGTTG**GCTTCCCATTTGCAA**
AAACATCGCTCCATAGTCACACCAAGTCTCCTGTGGG
AGTCTTGCTCCTCGTTCTCAGTGTCTGTTACAGCTCG
GTATTTTAGTGTCAGGACGTCGGCTCCCAGCCCGCAT
CTCCGCTCAGCAATGCCATTATCTTCTCAGCCAAGTCCT
AGAAATGGGTTGGCTTCCCATTTGCAAAAACATCGCTCCATA
GTCACACCAAGTCTCCTGTGGG

Systems Biology

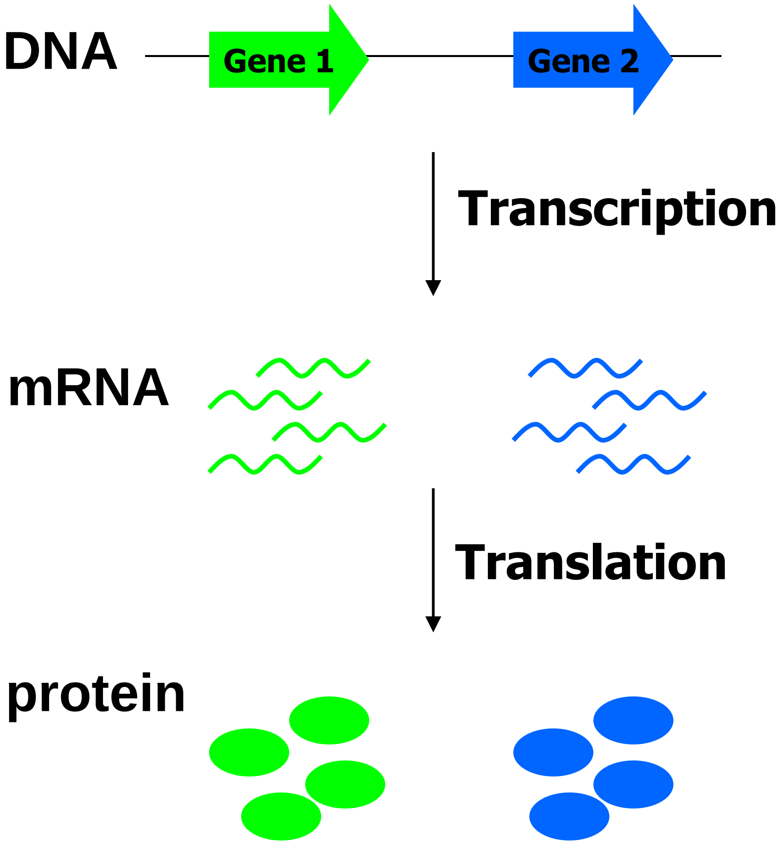


CACUUCCCGAUAACGCCCCAUUGUCCUUUGGGACU
CCACGACUAUCCAGUUUCUACCUCUUCCGCAUCCCU
AUACACAACCGUCUCCUUGGCUUCUUAUGGUCCGG
UAAAGGCUUCGGUGAGUAAAAGGAGCUUAAGGAAG
AAAUA

ASDVAHHPHPHAGKTALHKFTRVSLSPSNRPLLFGQT
ALHKRTFLS

XVAAAMLLRSCPVLSQGPTGLLGKVAKTYQFIFSIGRC
PILATQGPTCS

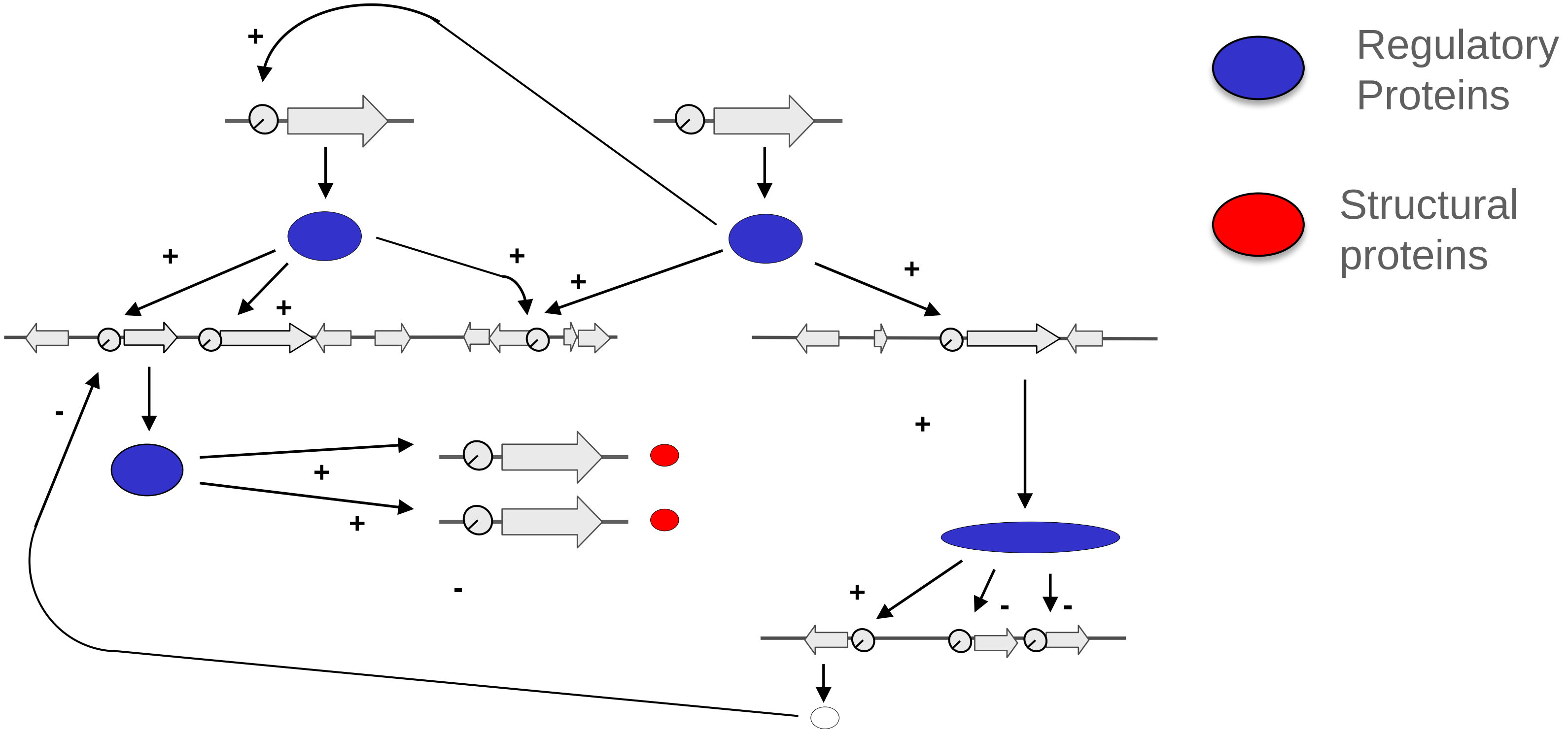
Systems Biology



Regulatory Proteins

Structural proteins

Systems Biology



Static network is encoded in the genome

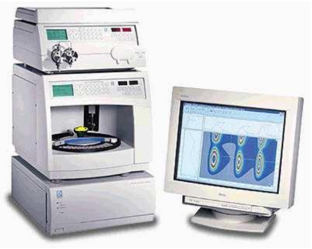
Not all proteins are made at all times (network is condition-dependent)

State of the network can be measured through gene/protein expression

Systems Biology

computer cluster

cell culture robot



HPLC

PCR

microarray platform

MALDI-TOF mass spectrometer

DNA sequencers

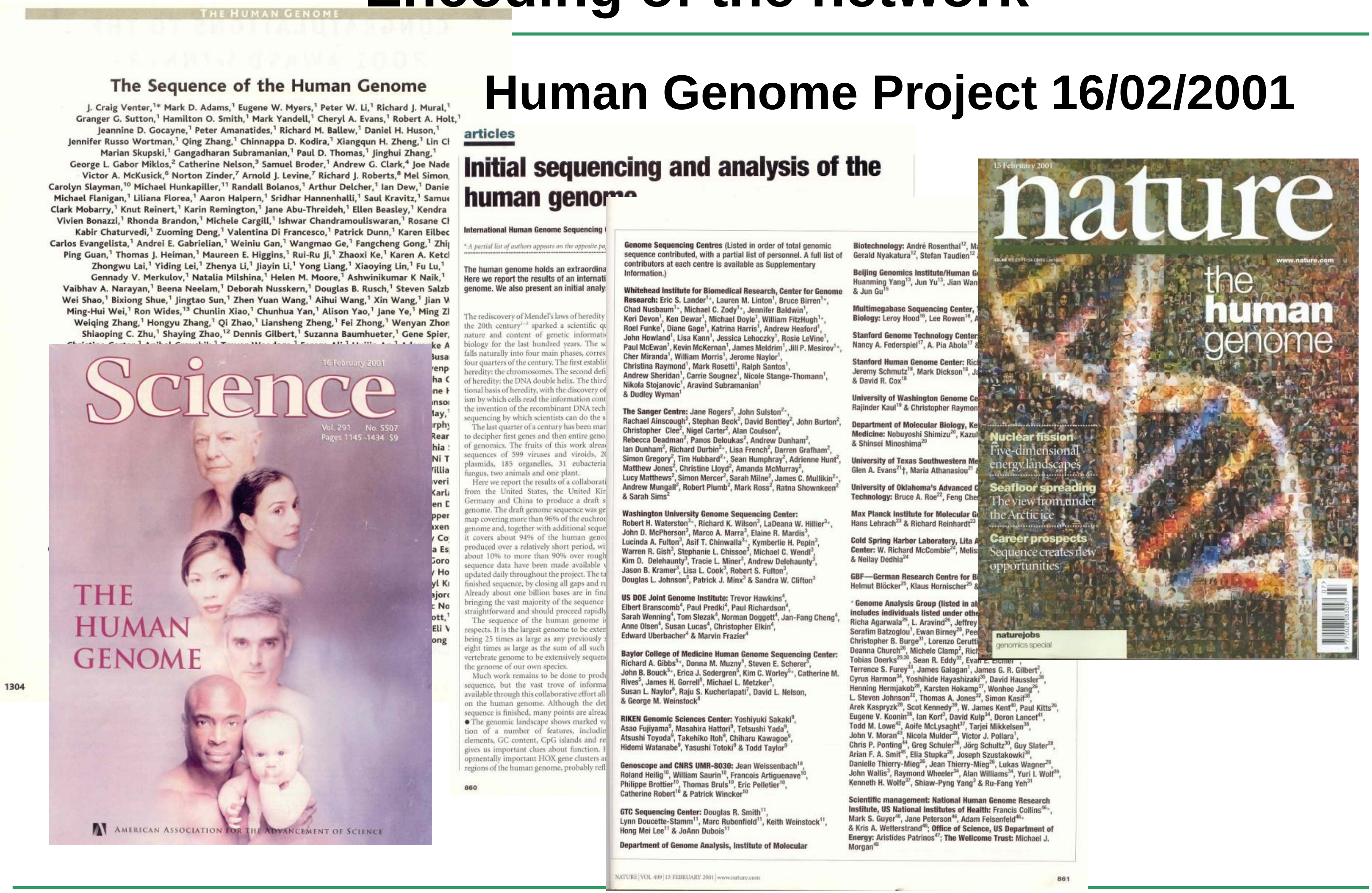
Encoding of the network

Genome

AGCACTGTCCACTGCATGGTGAGGATGGGGGTGAGCTCCCT
TTGTGGCTAGGTGCTTAAACGTCTATCGGACGCTCA**GTGAA**
GGGCTATTGCGGGGTAACAGGAAACCCTGAGGTGCT
GATAGGTCAAAGATGGAGAAGGCGTAGGGATATGTG
TTGGCAGAGGAACCGAAGAATACCAGGCCATTTCCG
AAGCCACTCATTTTCCTCGAATTCCTTCTTTATGCCTTC
AGTCTCTATTGACCGTAAATTTGGTTGTTGTCTCCCAGCTGT
TTATTTCTGTAACAGATCTTGGAGGCTGCGGTCTGGATCCCT
CGCCAAGAACCAGATCCAGGAGAAAACGTGCTCAACGTGCA
GCTCTGCTCCTACTGATTATAGCCCCACAGATGACATCGCTC
CATAGTCACACCAAGTCTCCTGTGGGAGTCTTGCTCCTCGTT
CTCAGTGTCTGTTACAGCTCGGTATTTTAGTGTCAGGACGTC
GGCTCCCAGCCCGCATCTCCGCTCAGCAATGCCATTATCTTC
TCAGCCAAGTCCTAGAAATGGGTTG**GCTTCCCATTTGCAA**
AAACATCGCTCCATAGTCACACCAAGTCTCCTGTGGG
AGTCTTGCTCCTCGTTCTCAGTGTCTGTTACAGCTCG
GTATTTTAGTGTCAGGACGTCGGCTCCCAGCCCGCAT
CTCCGCTCAGCAATGCCATTATCTTCTCAGCCAAGTCCT
AGAAATGGGTTGGCTTCCCATTTGCAAAAACATCGCTCCATA
GTCACACCAAGTCTCCTGTGGG

Encoding of the network

Human Genome Project 16/02/2001



THE HUMAN GENOME

The Sequence of the Human Genome

J. Craig Venter,[1*] Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Cl
Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1]
George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nade
Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,
Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Danie
Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samue
Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra
Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Cl
Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbec
Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhi
Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketc
Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1]
Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1]
Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzb
Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian W
Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zl
Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhon
Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,

1304

16 February 2001

Science

Vol. 291 No. 5507
Pages 1145–1434 $9

THE
HUMAN
GENOME

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing

* A partial list of authors appears on the opposite pa

The human genome holds an extraordina
Here we report the results of an internati
genome. We also present an initial analy

The rediscovery of Mendel's laws of heredity
the 20th century[1–3] sparked a scientific qu
nature and content of genetic informatio
biology for the last hundred years. The sc
falls naturally into four main phases, corres
four quarters of the century. The first establis
heredity: the chromosomes. The second defi
of heredity: the DNA double helix. The third
tional basis of heredity, with the discovery of
ism by which cells read the information cont
the invention of the recombinant DNA tech
sequencing by which scientists can do the s

The last quarter of a century has been mar
to decipher first genes and then entire geno
of genomics. The fruits of this work alrea
sequences of 599 viruses and viroids, 20
plasmids, 185 organelles, 31 eubacteria
fungus, two animals and one plant.

Here we report the results of a collaborati
from the United States, the United Kin
Germany and China to produce a draft s
genome. The draft genome sequence was ge
map covering more than 96% of the euchror
genome and, together with additional seque
it covers about 94% of the human geno
produced over a relatively short period, wi
about 10% to more than 90% over rough
sequence data have been made available v
updated daily throughout the project. The ta
finished sequence, by closing all gaps and re
Already about one billion bases are in fina
bringing the vast majority of the sequence
straightforward and should proceed rapidly

The sequence of the human genome is
respects. It is the largest genome to be exten
being 25 times as large as any previously s
eight times as large as the sum of all such
vertebrate genome to be extensively sequen
the genome of our own species.

Much work remains to be done to produ
sequence, but the vast trove of informa
available through this collaborative effort all
on the human genome. Although the det
sequence is finished, many points are alrea
● The genomic landscape shows marked va
tion of a number of features, includin
elements, GC content, CpG islands and re
gives us important clues about function. I
opmentally important HOX gene clusters ar
regions of the human genome, probably refl

860

Genome Sequencing Centres (Listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors at each centre is available as Supplementary Information.)

Whitehead Institute for Biomedical Research, Center for Genome Research: Eric S. Lander[1*], Lauren M. Linton[1], Bruce Birren[1*], Chad Nusbaum[1*], Michael C. Zody[1*], Jennifer Baldwin[1], Keri Devon[1], Ken Dewar[1], Michael Doyle[1], William FitzHugh[1*], Roel Funke[1], Diane Gage[1], Katrina Harris[1], Andrew Heaford[1], John Howland[1], Lisa Kann[1], Jessica Lehoczky[1], Rosie LeVine[1], Paul McEwan[1], Kevin McKernan[1], James Meldrim[1], Jill P. Mesirov[1*], Cher Miranda[1], William Morris[1], Jerome Naylor[1], Christina Raymond[1], Mark Rosetti[1], Ralph Santos[1], Andrew Sheridan[1], Carrie Sougnez[1], Nicole Stange-Thomann[1], Nikola Stojanovic[1], Aravind Subramanian[1] & Dudley Wyman[1]

The Sanger Centre: Jane Rogers[2], John Sulston[2*], Rachael Ainscough[2], Stephan Beck[2], David Bentley[2], John Burton[2], Christopher Clee[2], Nigel Carter[2], Alan Coulson[2], Rebecca Deadman[2], Panos Deloukas[2], Andrew Dunham[2], Ian Dunham[2], Richard Durbin[2*], Lisa French[2], Darren Grafham[2], Simon Gregory[2], Tim Hubbard[2*], Sean Humphray[2], Adrienne Hunt[2], Matthew Jones[2], Christine Lloyd[2], Amanda McMurray[2], Lucy Matthews[2], Simon Mercer[2], Sarah Milne[2], James C. Mullikin[2*], Andrew Mungall[2], Robert Plumb[2], Mark Ross[2], Ratna Shownkeen[2] & Sarah Sims[2]

Washington University Genome Sequencing Center: Robert H. Waterston[3*], Richard K. Wilson[3], LaDeana W. Hillier[3*], John D. McPherson[3], Marco A. Marra[3], Elaine R. Mardis[3], Lucinda A. Fulton[3], Asif T. Chinwalla[3*], Kymberlie H. Pepin[3], Warren R. Gish[3], Stephanie L. Chissoe[3], Michael C. Wendl[3], Kim D. Delehaunty[3], Tracie L. Miner[3], Andrew Delehaunty[3], Jason B. Kramer[3], Lisa L. Cook[3], Robert S. Fulton[3], Douglas L. Johnson[3], Patrick J. Minx[3] & Sandra W. Clifton[3]

US DOE Joint Genome Institute: Trevor Hawkins[4], Elbert Branscomb[4], Paul Predki[4], Paul Richardson[4], Sarah Wenning[4], Tom Slezak[4], Norman Doggett[4], Jan-Fang Cheng[4], Anne Olsen[4], Susan Lucas[4], Christopher Elkin[4], Edward Uberbacher[4] & Marvin Frazier[4]

Baylor College of Medicine Human Genome Sequencing Center: Richard A. Gibbs[5*], Donna M. Muzny[5], Steven E. Scherer[5], John B. Bouck[5*], Erica J. Sodergren[5], Kim C. Worley[5*], Catherine M. Rives[5], James H. Gorrell[5], Michael L. Metzker[5], Susan L. Naylor[6], Raju S. Kucherlapati[7], David L. Nelson, & George M. Weinstock[8]

RIKEN Genomic Sciences Center: Yoshiyuki Sakaki[9], Asao Fujiyama[9], Masahira Hattori[9], Tetsushi Yada[9], Atsushi Toyoda[9], Takehiko Itoh[9], Chiharu Kawagoe[9], Hidemi Watanabe[9], Yasushi Totoki[9] & Todd Taylor[9]

Genoscope and CNRS UMR-8030: Jean Weissenbach[10], Roland Heilig[10], William Saurin[10], Francois Artiguenave[10], Philippe Brottier[10], Thomas Bruls[10], Eric Pelletier[10], Catherine Robert[10] & Patrick Wincker[10]

GTC Sequencing Center: Douglas R. Smith[11], Lynn Doucette-Stamm[11], Marc Rubenfield[11], Keith Weinstock[11], Hong Mei Lee[11] & JoAnn Dubois[11]

Department of Genome Analysis, Institute of Molecular

Biotechnology: André Rosenthal[12], Ma
Gerald Nyakatura[12], Stefan Taudien[12]

Beijing Genomics Institute/Human G
Huanming Yang[13], Jun Yu[13], Jian Wan
& Jun Gu[15]

Multimegabase Sequencing Center,
Biology: Leroy Hood[16], Lee Rowen[16], A

Stanford Genome Technology Center
Nancy A. Federspiel[17], A. Pia Abola[17] &

Stanford Human Genome Center: Ric
Jeremy Schmutz[18], Mark Dickson[18], Ja
& David R. Cox[18]

University of Washington Genome Ce
Rajinder Kaul[19] & Christopher Raymon

Department of Molecular Biology, Ke
Medicine: Nobuyoshi Shimizu[20], Kazul
& Shinsei Minoshima[20]

University of Texas Southwestern Me
Glen A. Evans[21]†, Maria Athanasiou[21] &

University of Oklahoma's Advanced C
Technology: Bruce A. Roe[22], Feng Che

Max Planck Institute for Molecular G
Hans Lehrach[23] & Richard Reinhardt[23]

Cold Spring Harbor Laboratory, Lita A
Center: W. Richard McCombie[24], Melis
& Neilay Dedhia[24]

GBF—German Research Centre for Bi
Helmut Blöcker[25], Klaus Hornischer[25] &

* **Genome Analysis Group (listed in al**
includes individuals listed under othe
Richa Agarwala[26], L. Aravind[26], Jeffrey
Serafim Batzoglou[1], Ewan Birney[28], Pee
Christopher B. Burge[31], Lorenzo Cerutti
Deanna Church[26], Michele Clamp[2], Rich
Tobias Doerks[29,30], Sean R. Eddy[32], Evan E. Eichler
Terrence S. Furey[33], James Galagan[1], James G. R. Gilbert[2],
Cyrus Harmon[34], Yoshihide Hayashizaki[35], David Haussler[36],
Henning Hermjakob[28], Karsten Hokamp[37], Wonhee Jang[26],
L. Steven Johnson[32], Thomas A. Jones[32], Simon Kasif[38],
Arek Kaspryzk[28], Scot Kennedy[39], W. James Kent[40], Paul Kitts[26],
Eugene V. Koonin[26], Ian Korf[3], David Kulp[34], Doron Lancet[41],
Todd M. Lowe[42], Aoife McLysaght[37], Tarjei Mikkelsen[38],
John V. Moran[43], Nicola Mulder[28], Victor J. Pollara[1],
Chris P. Ponting[44], Greg Schuler[26], Jörg Schultz[30], Guy Slater[28],
Arian F. A. Smit[45], Elia Stupka[28], Joseph Szustakowki[38],
Danielle Thierry-Mieg[26], Jean Thierry-Mieg[26], Lukas Wagner[26],
John Wallis[3], Raymond Wheeler[34], Alan Williams[34], Yuri I. Wolf[26],
Kenneth H. Wolfe[37], Shiaw-Pyng Yang[3] & Ru-Fang Yeh[31]

Scientific management: National Human Genome Research Institute, US National Institutes of Health: Francis Collins[46*], Mark S. Guyer[46], Jane Peterson[46], Adam Felsenfeld[46*] & Kris A. Wetterstrand[46]; **Office of Science, US Department of Energy:** Aristides Patrinos[47]; **The Wellcome Trust:** Michael J. Morgan[48]

NATURE | VOL 409 | 15 FEBRUARY 2001 | www.nature.com

861

15 February 2001

nature

www.nature.com

the human genome

Nuclear fission
Five-dimensional energy landscapes

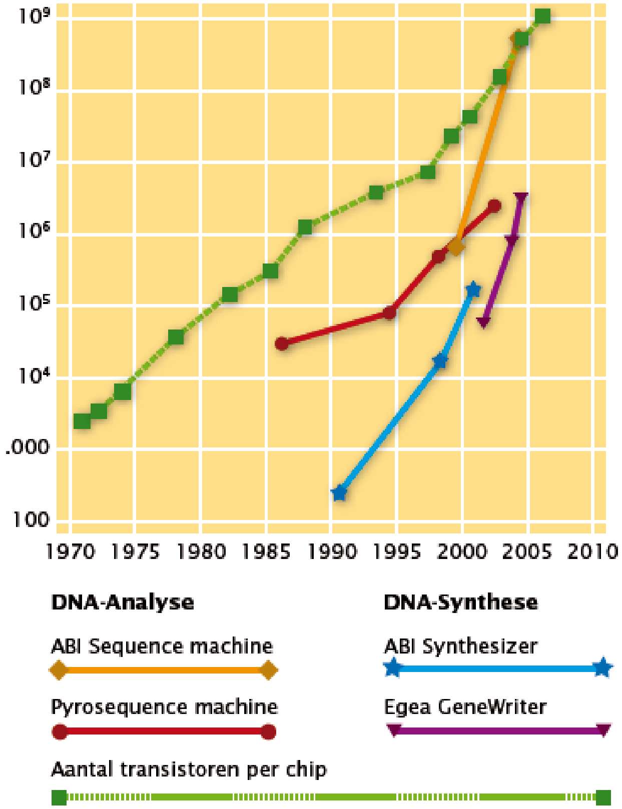
Seafloor spreading
The view from under the Arctic ice

Career prospects
Sequence creates new opportunities

naturejobs
genomics special

Sequencing human genome: 13 years/ 3 miljard dollar

Encoding of the network



De wet van Moore (exponentiële toename van de snelheid) is ook van toepassing op de analyse en synthese van DNA

Next generation sequencing follows Moore Law

454
Solid: 50 Gb/run;
Helicos
Illumina: 25 Gb/run; 75 bp reads

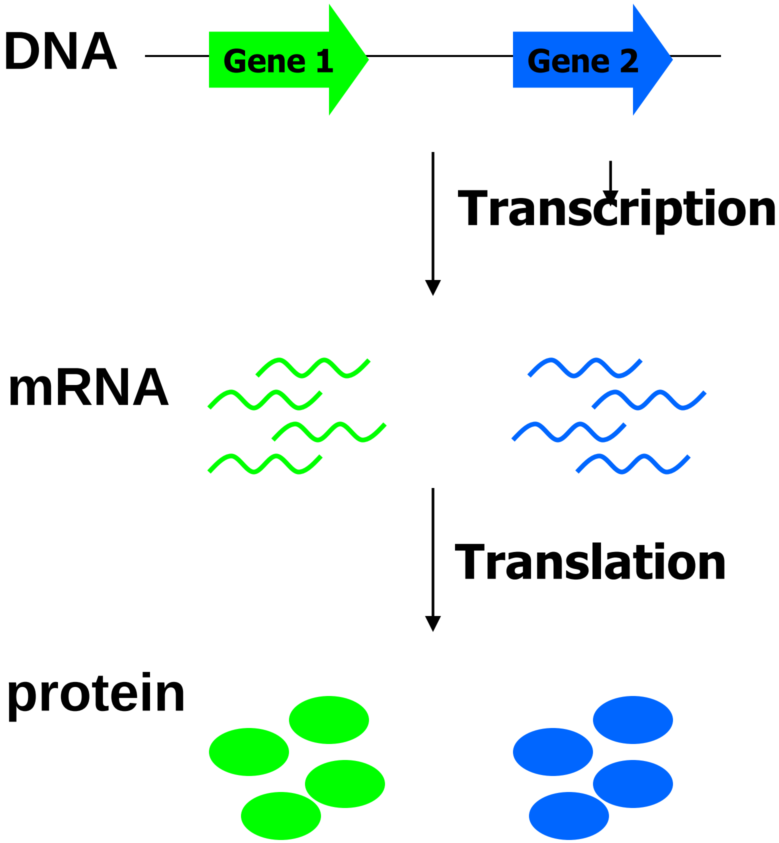
Encoding of the network

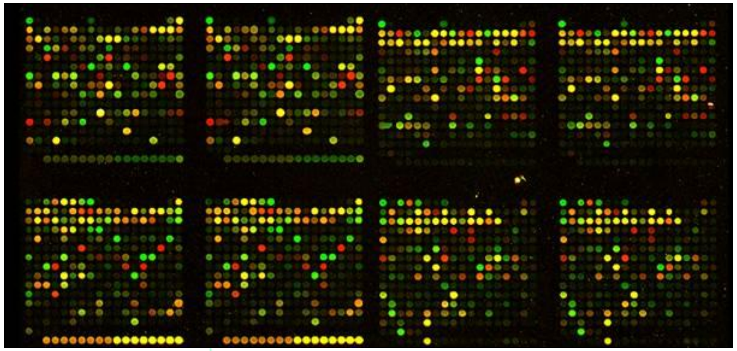
Next generation sequencing technology

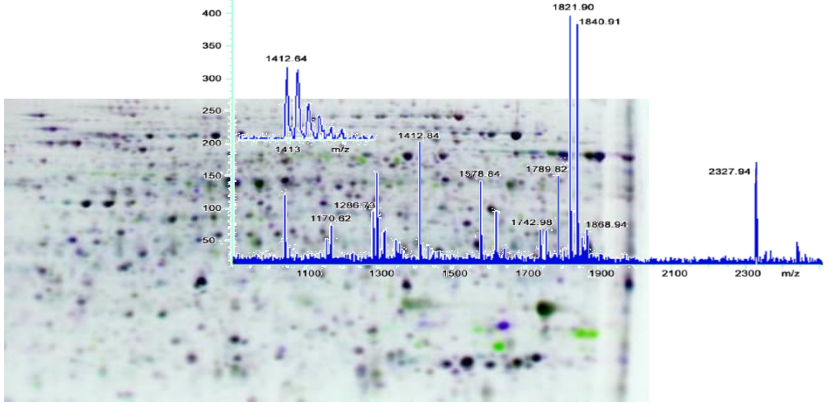
- **Sequencing human genome: 13 years/ 3 miljard dollar**
- **Genome Watson (454 techn): 20 persons/2 months. Totaal 1.000.000 dollar**
- **Now: 1000 dollar human genome**

State of the network

Functional data: transcriptome, proteome, metabolome

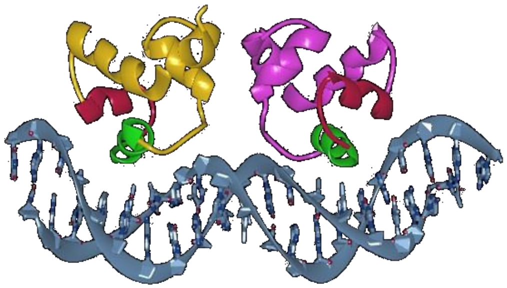


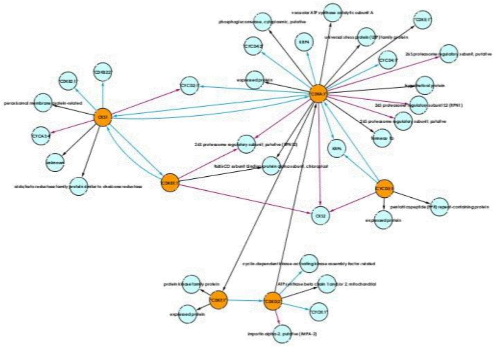


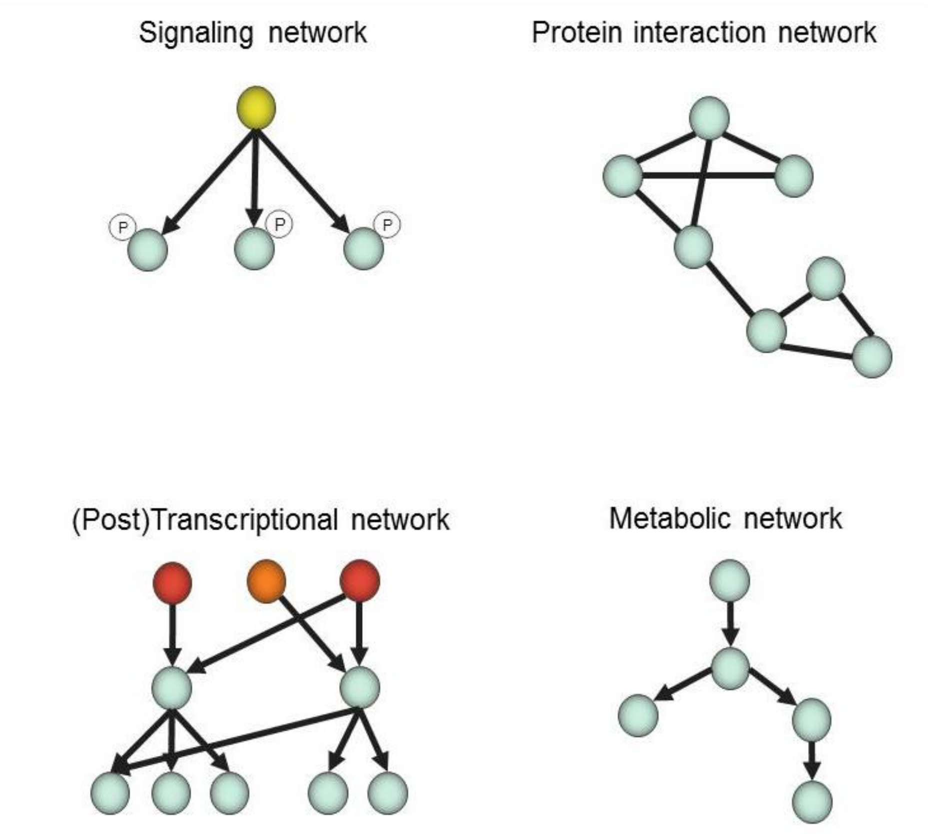


Interactions of the network

Physical data:

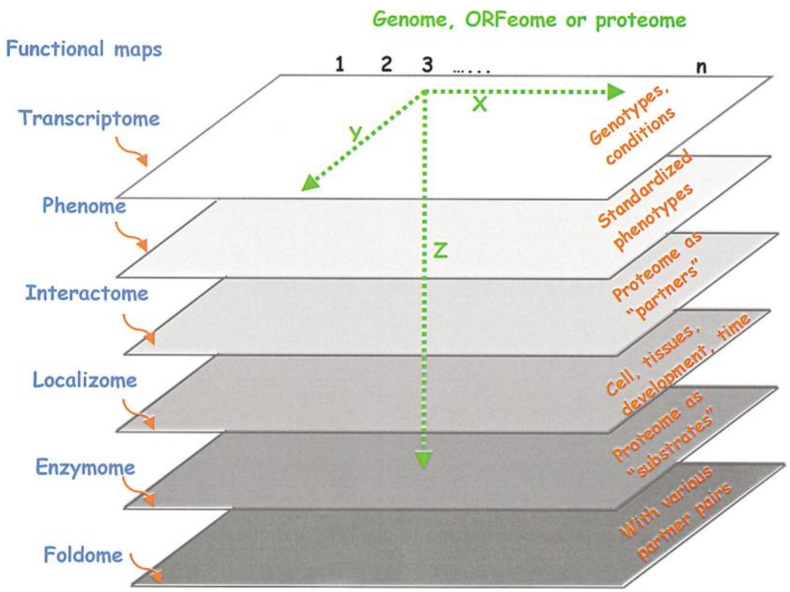




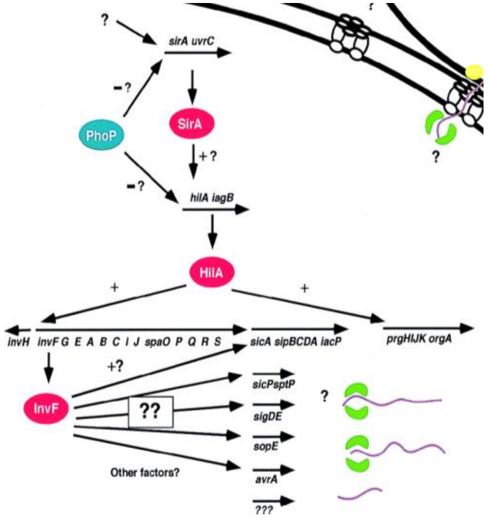


PHYSICAL INTERACTION NETWORKS

Inferring the network



High throughput data



Mechanistic insight in the biological system at molecular biological level (holistic insight)

Inferring the network

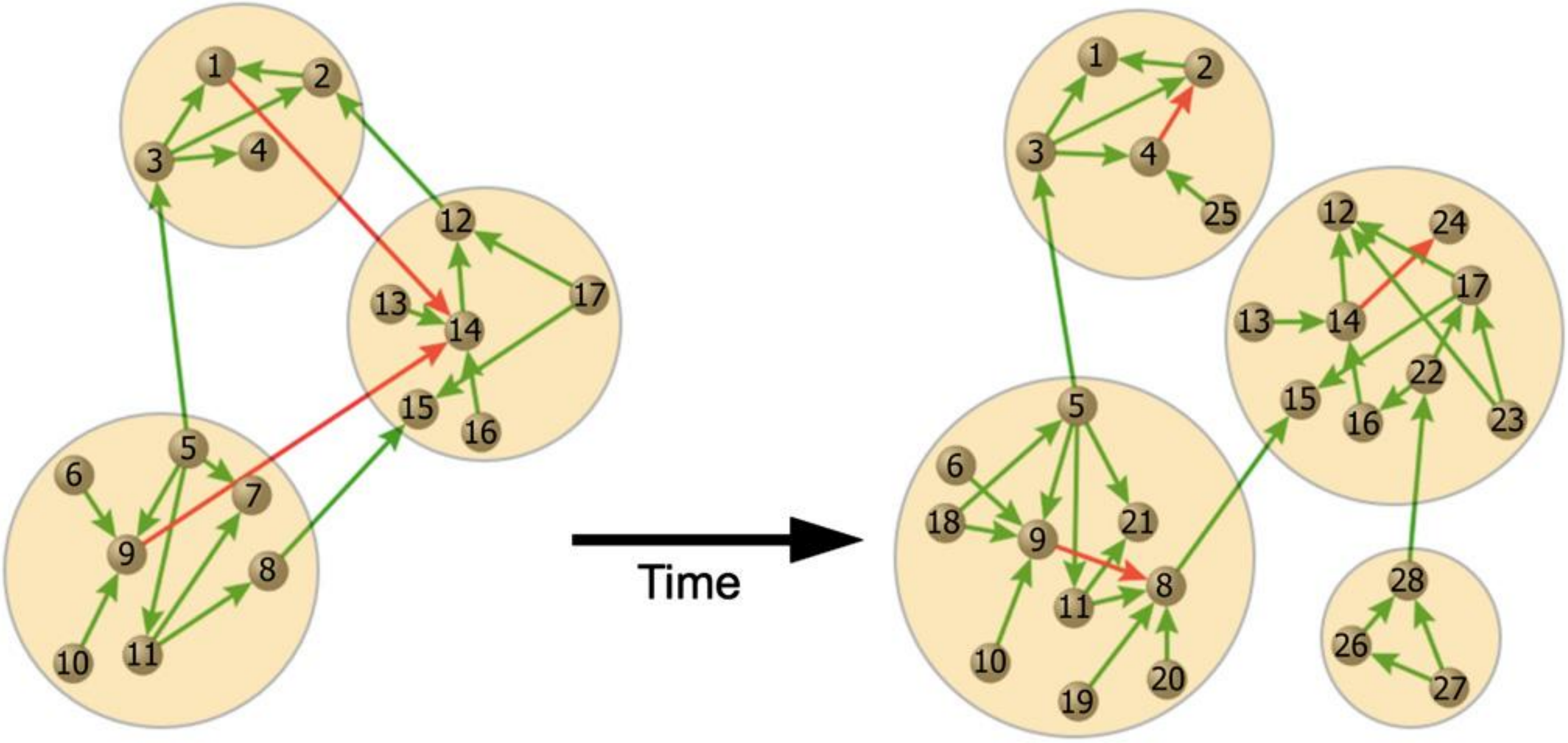
- Omics data are noisy
- Omics data are incomplete

- Integrate different data to obtain higher precision and coverage

- Reconstruct network
  - Different datasources
  - Different Molecular layers

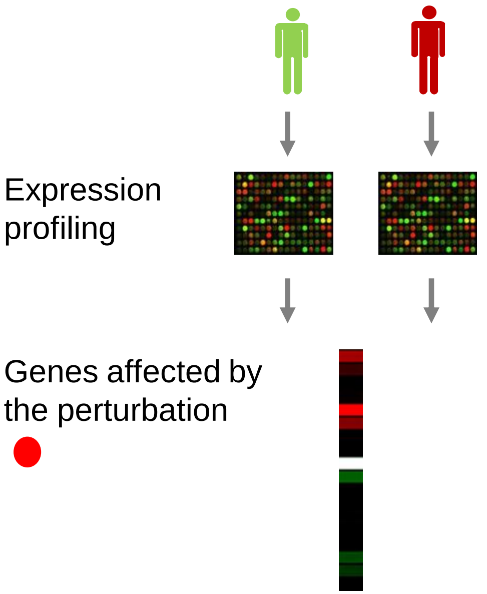
Fundamental knowledge

Evolution: comparing network between species or over time



Using the network to interpret data

In house data

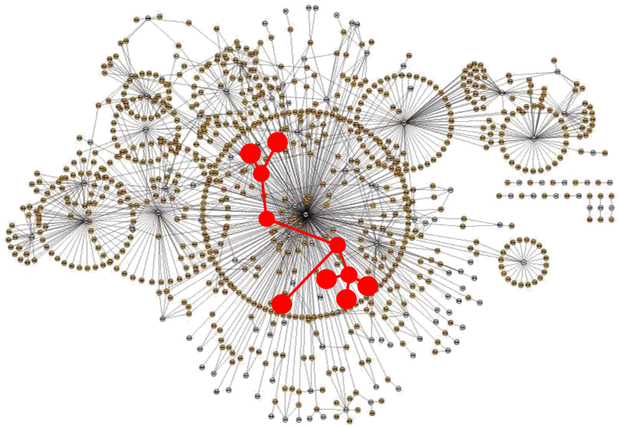


Expression profiling

Genes affected by the perturbation

●

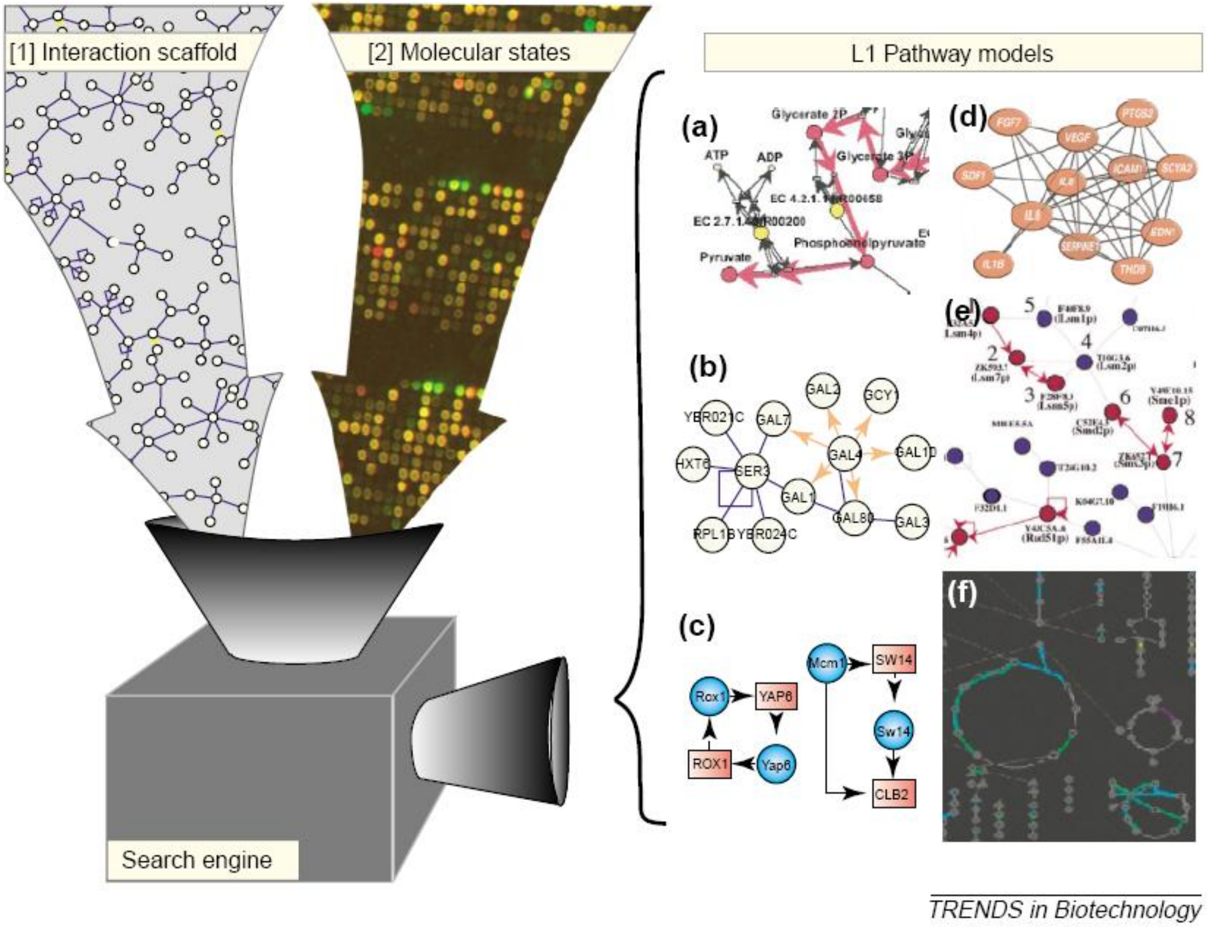
Physical interaction network



Infer the **hidden paths** between genomic variations and expression alteration

Network-based analysis of unstructured gene lists

Bioinformatics and datamining



Ideker and Lauffenburger, TiB 2003

Bioinformatics and datamining

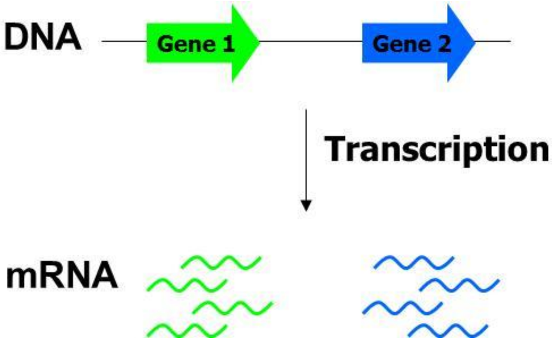
What does it require applying a computer science framework to bioinformatics?

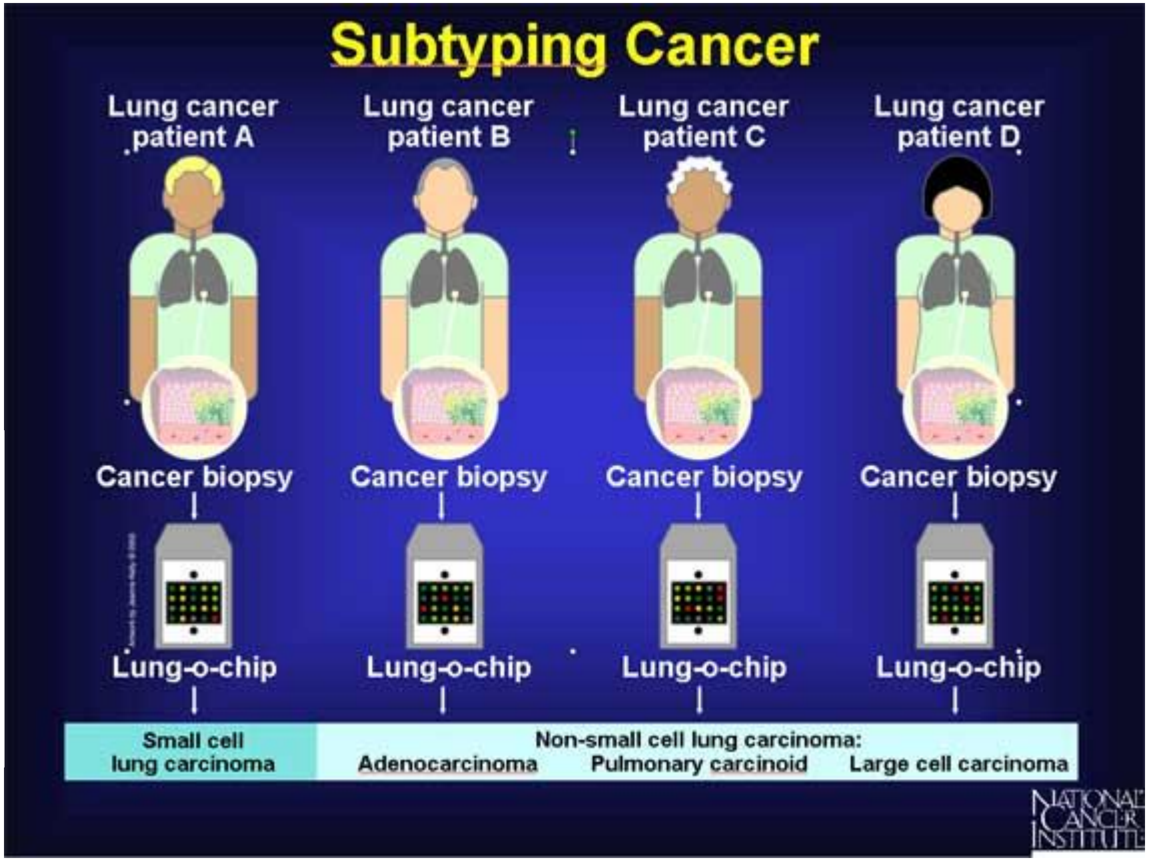
Fast solution

- The wet lab scientist rules
- Competition is fierce
  - Often the high impact papers are the conceptual ones
  - You tackle a research problem for the FIRST time
- Biological message is more imprtant than the method used to analyse the data
  - Code is sloppy , undocumented
  - Unsustainable code /tool development

Problems are complex

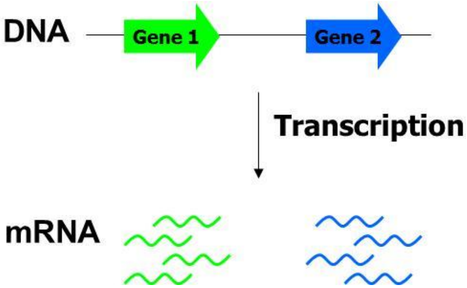
Cancer subtyping
& biomarker
identification

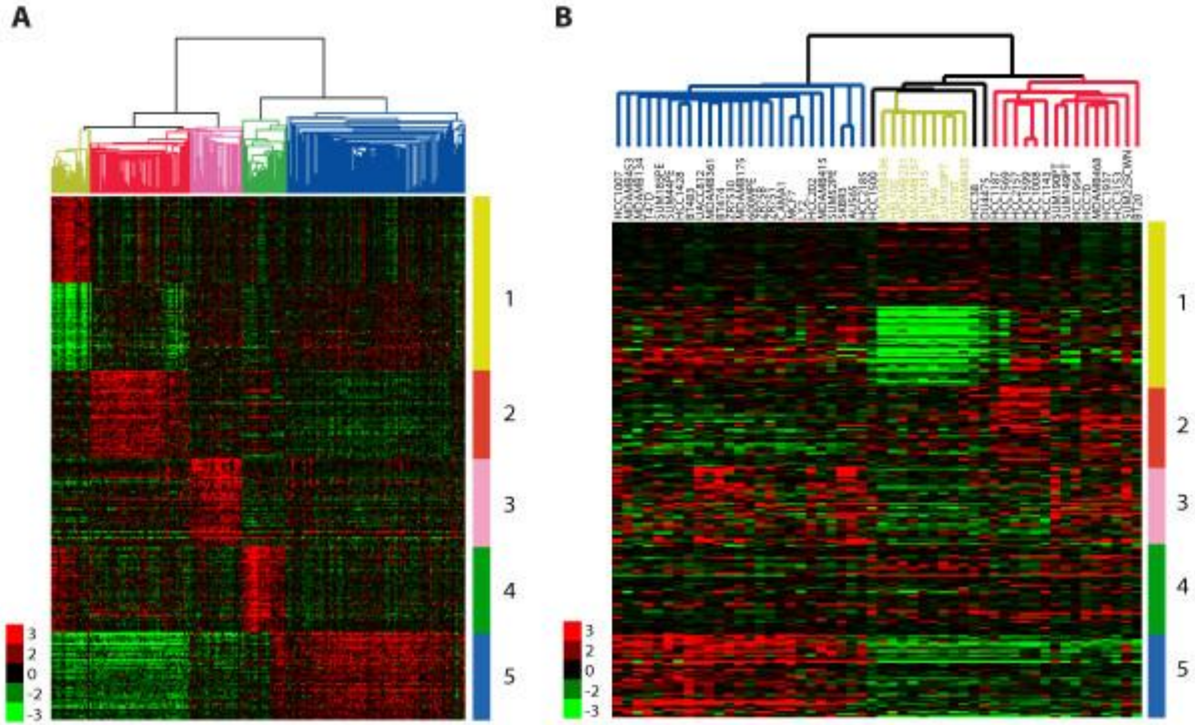




Classification problem

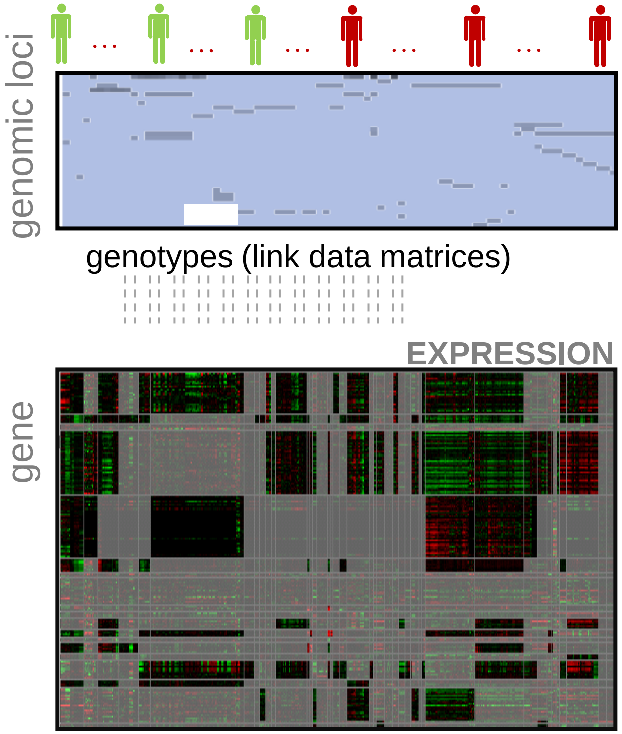
Problems are complex





(bi)clustering problem

Problems are complex

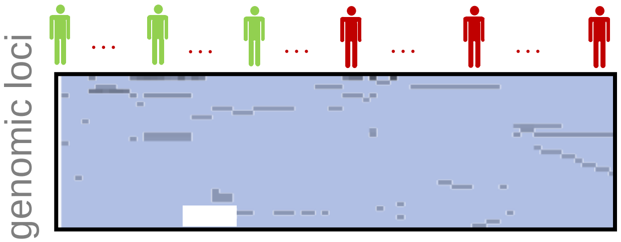


Dataintegration problem

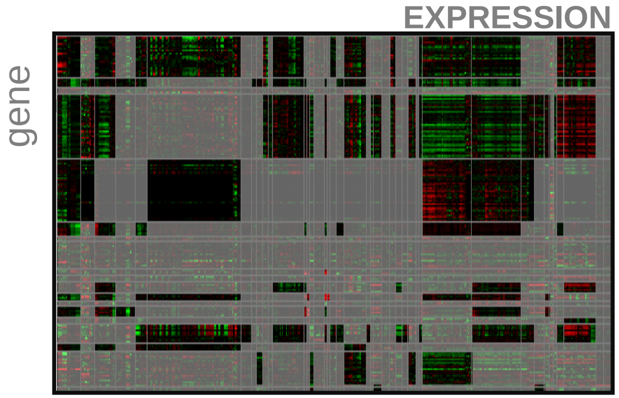
Problems are complex

Cancer subtyping

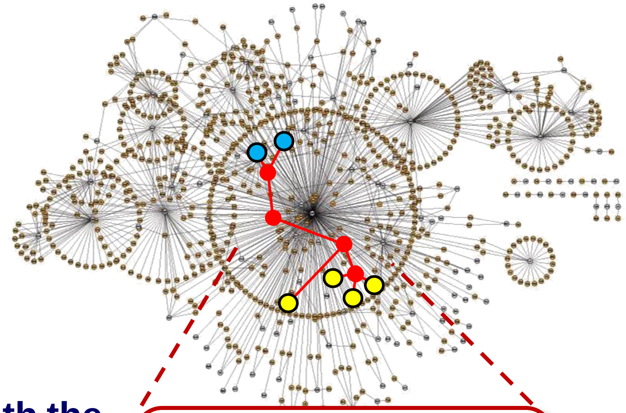
Preprocess the data



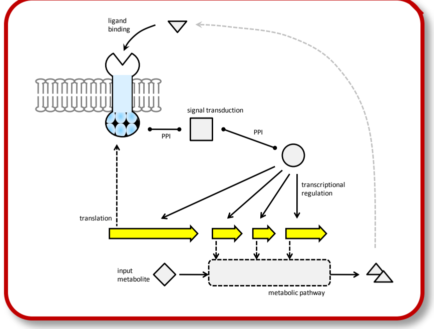
Associate genotypes to phenotype



Infer the interaction network from the omics data



Integrate all data with the interaction network



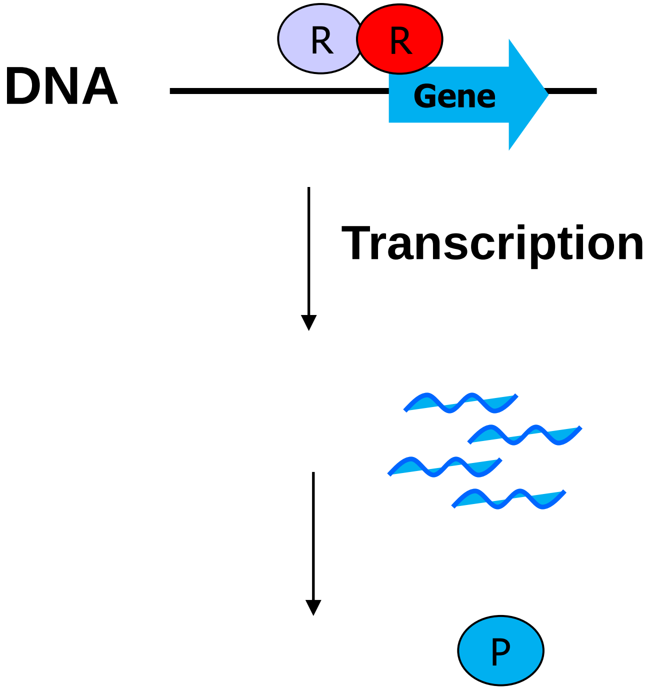
Dataintegration problem

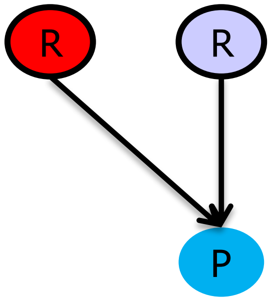
Bioinformatics and datamining

- Problems need a fast solution
- Problems are increasingly complex and can not be solved by one particular datamining tool (generic knowledge needed)
- Datamining in bioinformatics requires a quite thorough understanding of biology
- Problems are underdetermined and ill defined

Network inference

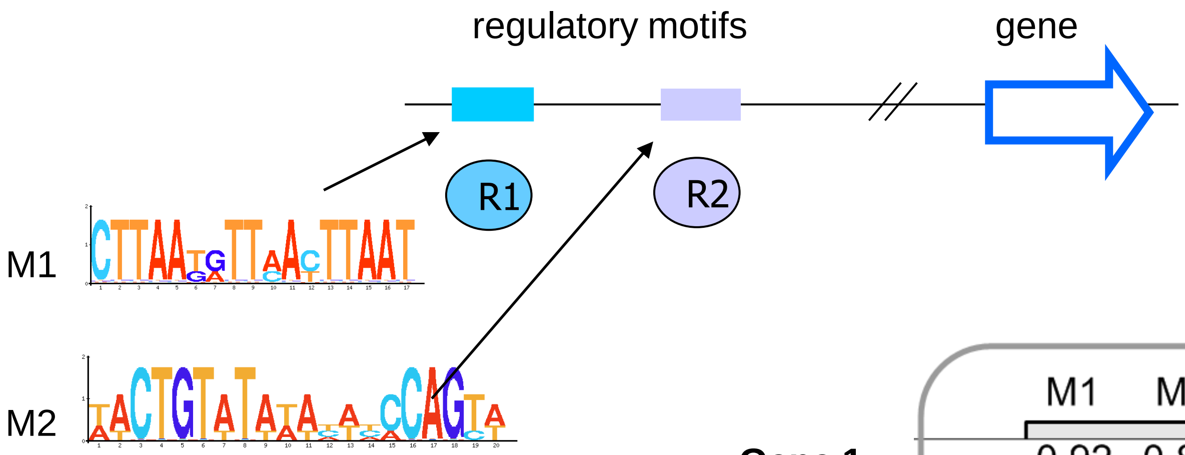
Transcriptional network inference



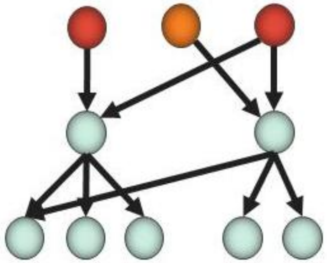


Transcriptional network

Network inference

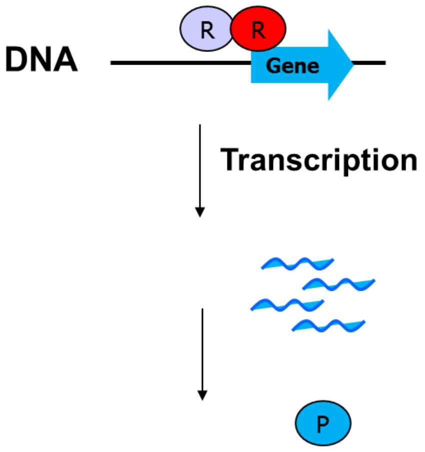


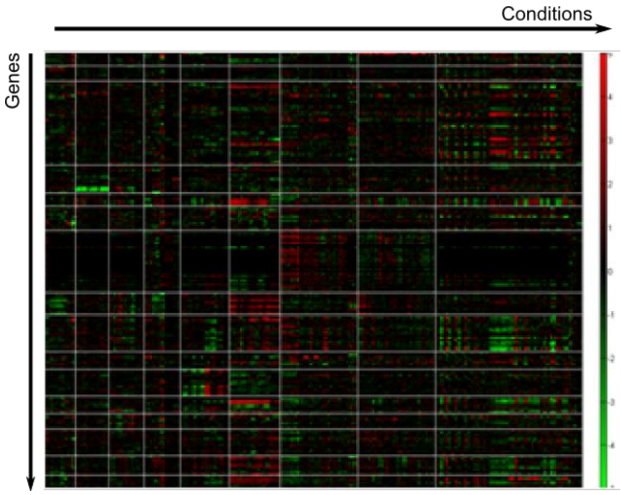
(Post)Transcriptional network

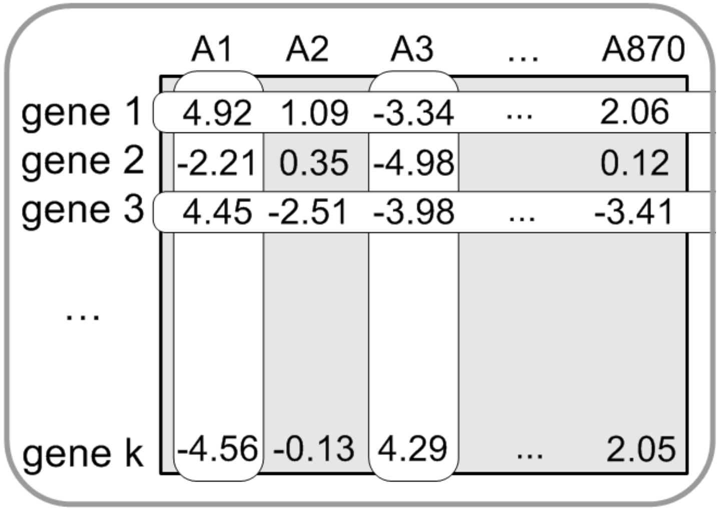


	M1	M2	M3	…	M67
Gene 1	0.92	0.86	0.99	...	0.86
Gene 2	0.91	0.95	0.38	...	0.72
	0.99	0.76	0.99	...	0.59
Gene n	0.33	0.98	0.99	...	0.65

Network inference

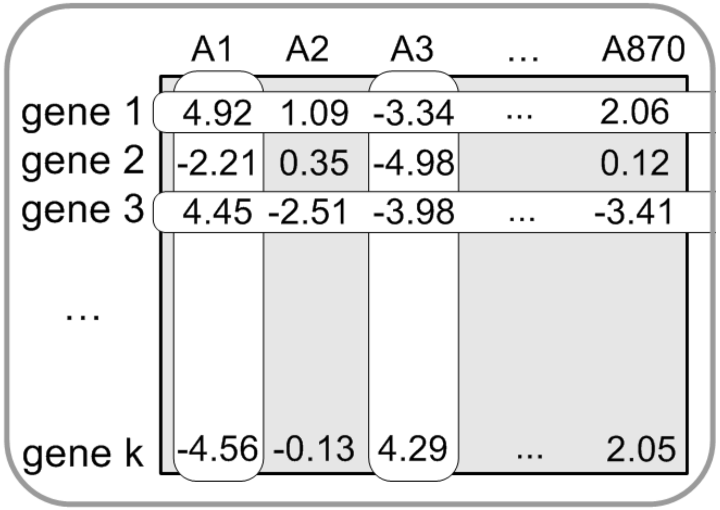




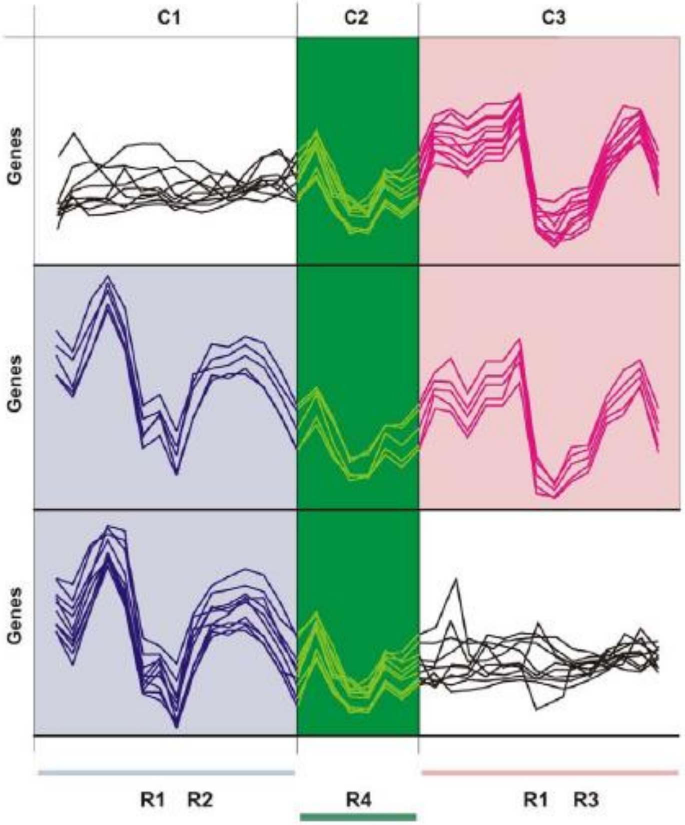


	A1	A2	A3	…	A870
gene 1	4.92	1.09	-3.34	…	2.06
gene 2	-2.21	0.35	-4.98		0.12
gene 3	4.45	-2.51	-3.98	…	-3.41
…					
gene k	-4.56	-0.13	4.29	…	2.05

Network inference



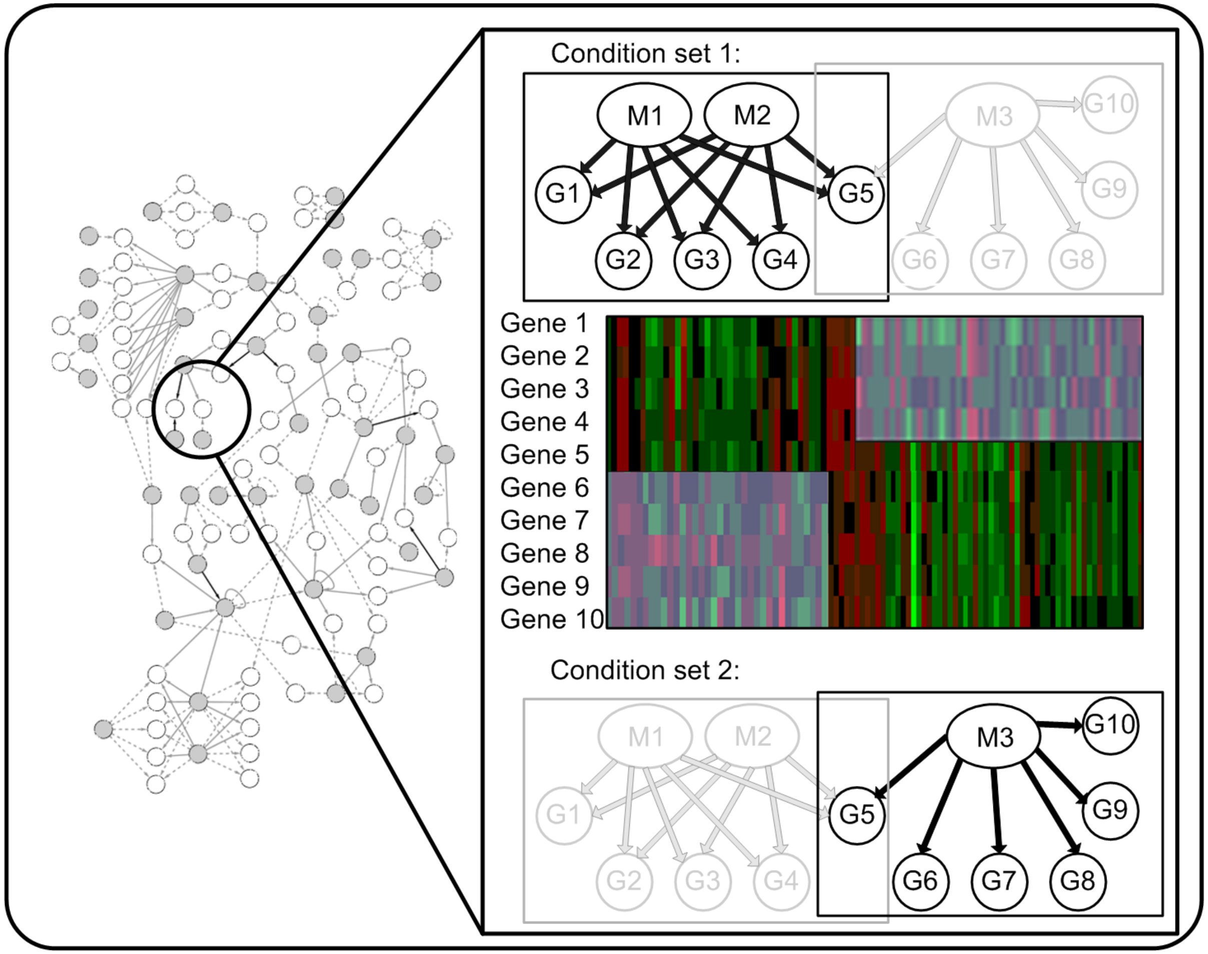
	A1	A2	A3	…	A870
gene 1	4.92	1.09	-3.34	…	2.06
gene 2	-2.21	0.35	-4.98		0.12
gene 3	4.45	-2.51	-3.98	…	-3.41
…					
gene k	-4.56	-0.13	4.29	…	2.05



Guilt by association

Coexpressed target genes are coregulated

Network inference



Network inference

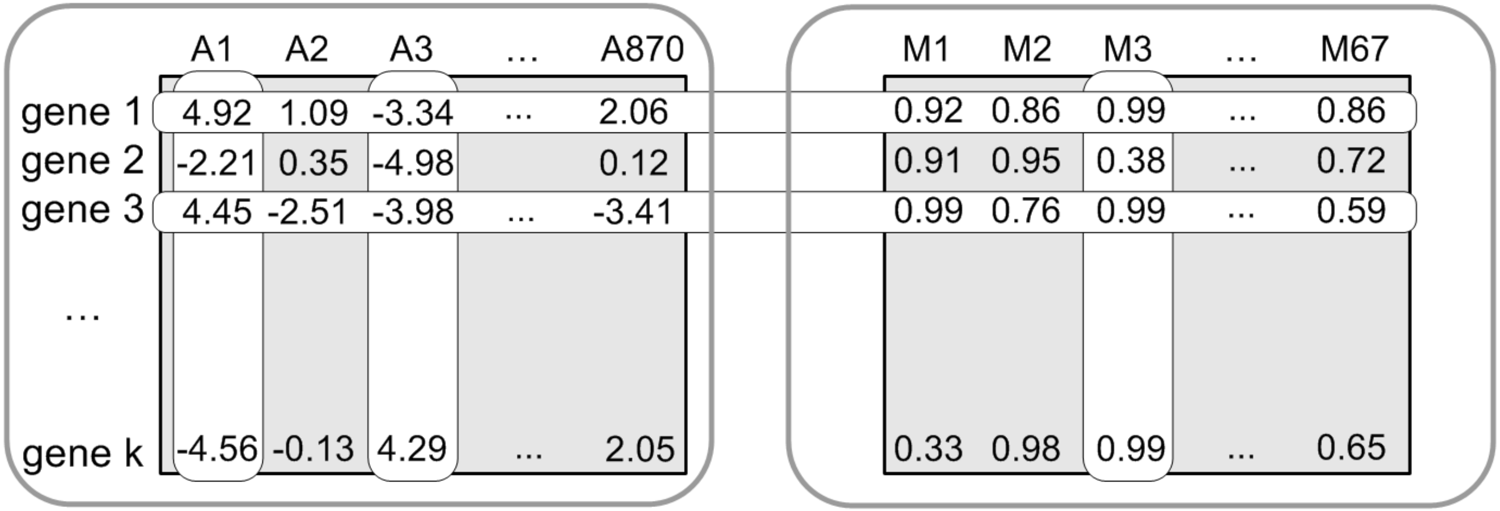
DISTILLER

Data Integration System
To Identify Links in Expression in Expression regulation

Lemmens et al. Genome Biol. 2009

Search co expression modules that meet minimal requirements:

- All genes are significantly co-expressed in a sufficiently large, *a priori* unspecified set of experimental conditions CC
- All genes contain motif instances for a sufficient number of common, a priori unspecified regulators CR



	A1	A2	A3	…	A870
gene 1	4.92	1.09	-3.34	…	2.06
gene 2	-2.21	0.35	-4.98		0.12
gene 3	4.45	-2.51	-3.98	…	-3.41
…					
gene k	-4.56	-0.13	4.29	…	2.05

M1	M2	M3	…	M67
0.92	0.86	0.99	…	0.86
0.91	0.95	0.38	…	0.72
0.99	0.76	0.99	…	0.59
0.33	0.98	0.99	…	0.65

Network inference

- Items = genes
- Transactions = conditions, motifs

Transactions	Items		
M1	G1	**G3**	**G5**
M2	**G3**	**G5**	G9
M3	**G3**	**G5**	G11
M4	G1	**G3**	**G5**

Itemset 1: G3, G5 supported by 4 motifs
Itemset 2: G1,G3,G5 supported by 2 motifs
Tidset t(G3,G5) = M1, M2, M3, M4
Tidset t(G1, G3,G5) = M1, M4
Minimal support = 3
Itemset G3, G5 is frequent

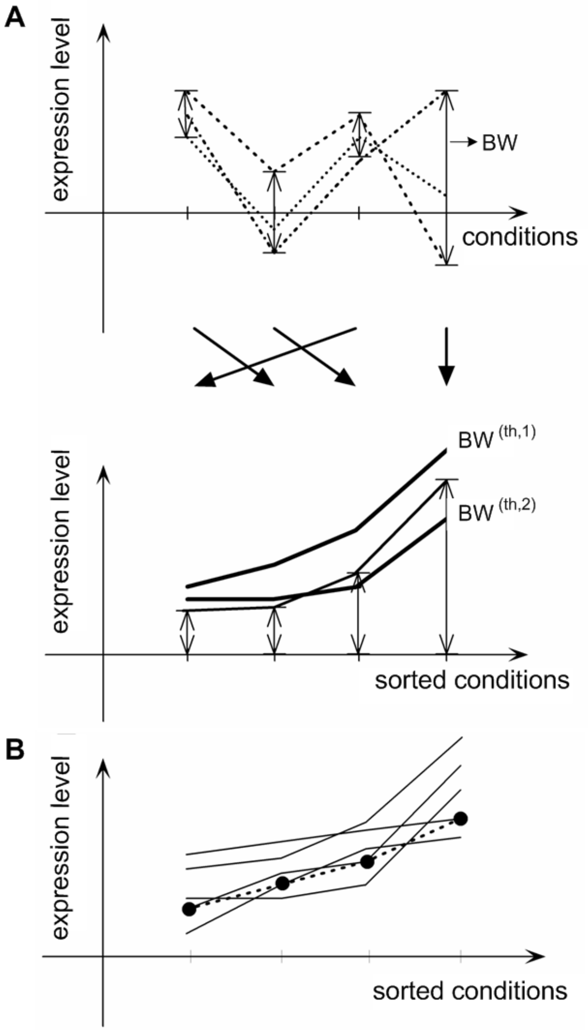
- Supports:
  - All genes in the module (itemsets) are significantly co-expressed in a sufficiently large **(minimum support)**, a priori unspecified set of experimental conditions
  - All genes in the module (itemsets) contain motif instances for a sufficient number *R* **(minimum support)** of common, a priori unspecified regulators

Network inference

Expression support

co-expression in a significant number of experimental conditions

- BW for each condition
- Rank BW in increasing order
- Check if BW sequence is within threshold BW sequence
- BW threshold sequence is determined by randomization



Network inference

Rank modules by assigning interest score

- p(module motif content):

  The chance that a module with the same number of genes and the same number of motifs is found at random

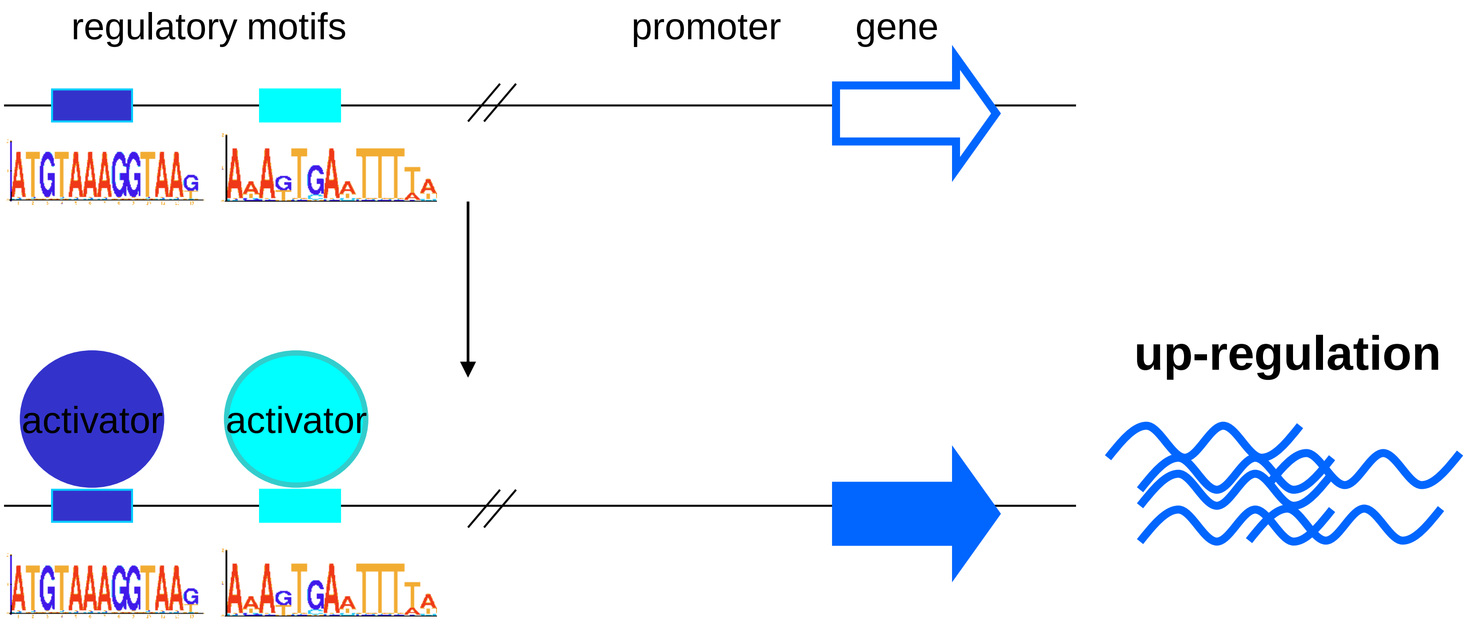
- p(module expression pattern):

  The chance that a module with at least the same number of genes and containing the same number of conditions is found by chance

Interest score:
p(M|m,g)Xp(M|e,g) → Rank the modules

- Modules are selected iteratively such that they add as much as possible new information to the already selected modules

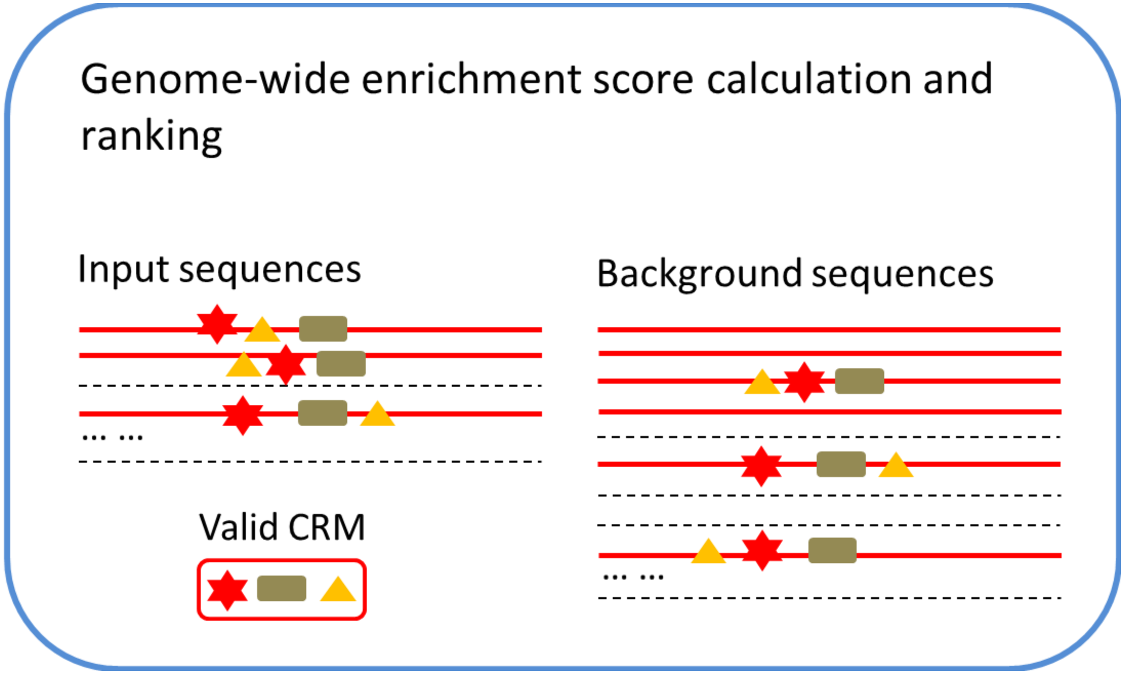
CRM detection



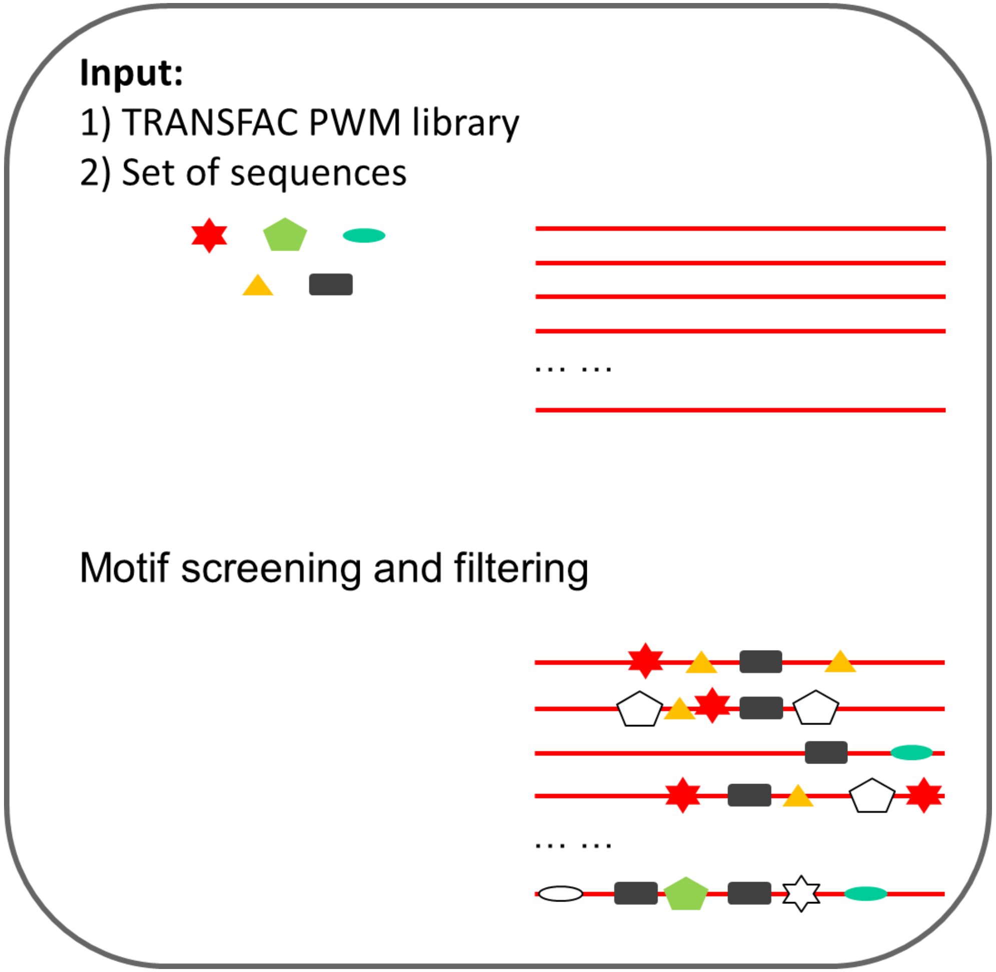
CRM: cis acting regulatory module
Combination of TF binding sites

CRM detection

Genes that are needed together in the cell usually are activated together = coregulation



CRM detection

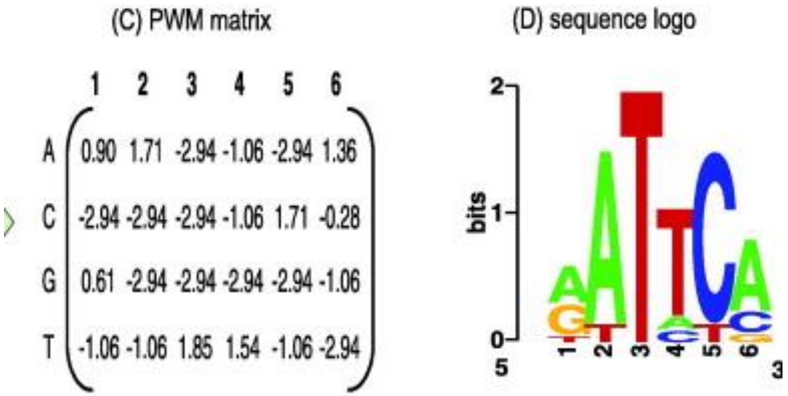


Input:
1) TRANSFAC PWM library
2) Set of sequences

… …

Motif screening and filtering

… …



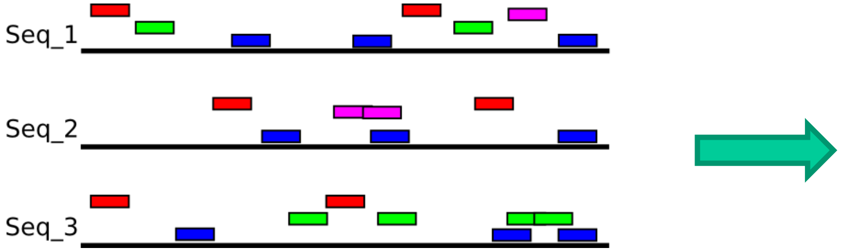
(C) PWM matrix

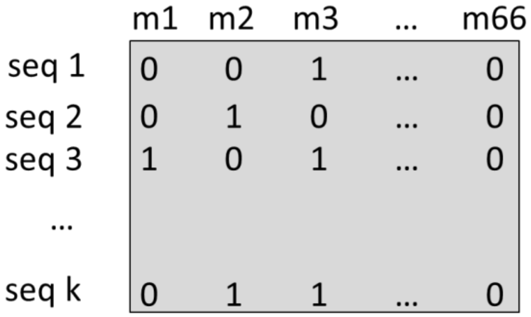
$$\begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline A & 0.90 & 1.71 & -2.94 & -1.06 & -2.94 & 1.36 \\ C & -2.94 & -2.94 & -2.94 & -1.06 & 1.71 & -0.28 \\ G & 0.61 & -2.94 & -2.94 & -2.94 & -2.94 & -1.06 \\ T & -1.06 & -1.06 & 1.85 & 1.54 & -1.06 & -2.94 \end{array}$$

(D) sequence logo

- CRM is valid if motifs occur in each others neighbourhood
- Order needs to be conserved?

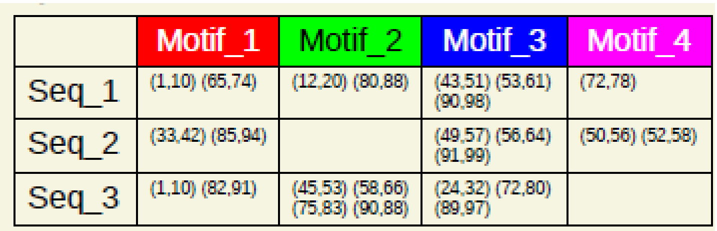
CRM detection





	m1	m2	m3	...	m66
seq 1	0	0	1	...	0
seq 2	0	1	0	...	0
seq 3	1	0	1	...	0
...					
seq k	0	1	1	...	0

DISTILLER



	Motif_1	Motif_2	Motif_3	Motif_4
Seq_1	(1,10) (65,74)	(12,20) (80,88)	(43,51) (53,61) (90,98)	(72,78)
Seq_2	(33,42) (85,94)		(49,57) (56,64) (91,99)	(50,56) (52,58)
Seq_3	(1,10) (82,91)	(45,53) (58,66) (75,83) (90,88)	(24,32) (72,80) (89,97)	

CPMODULE

Convert screening results in table with (start, stop) positions, for every sequence/motif pair

CRM detection

Constraint Programming (CP)

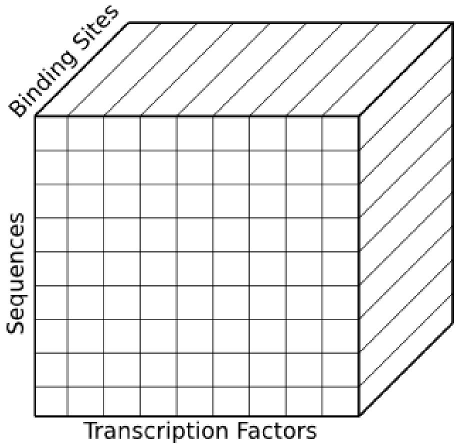
Model (by user):

Problem specification in terms of constraints

Search (by solver):

Propagation: in which a constraint is used to remove values from the domain of variables that would violate it

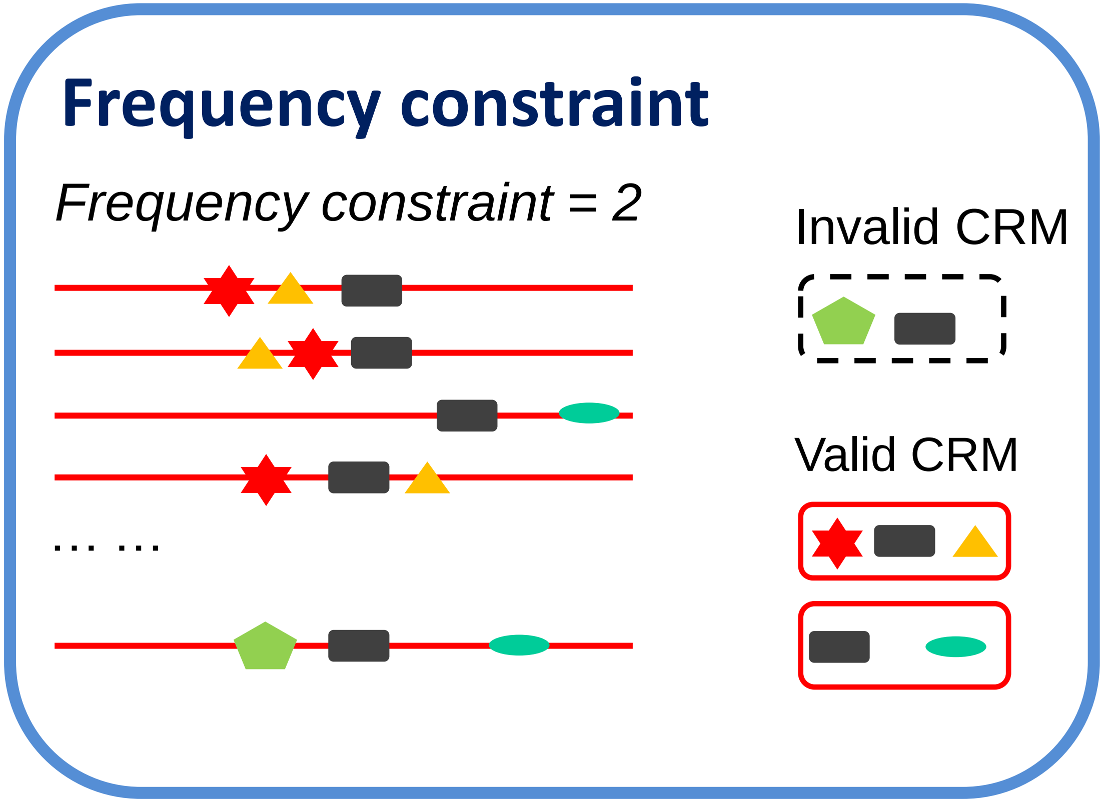
Branching: in which a variable is assigned a value from its domain D(v)



De Raedt et al. 2008 *KDD*

CRM detection

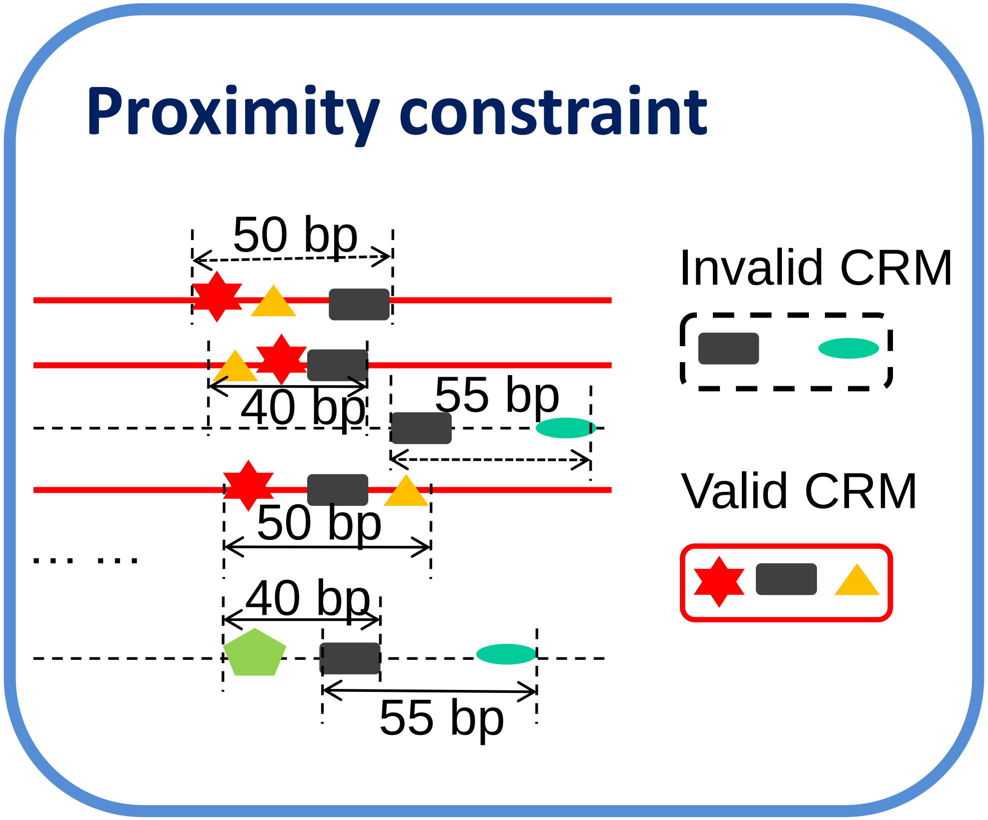
Frequency constraint



The motif set should occur in a sufficient number of sequences (but not all) to be considered valid

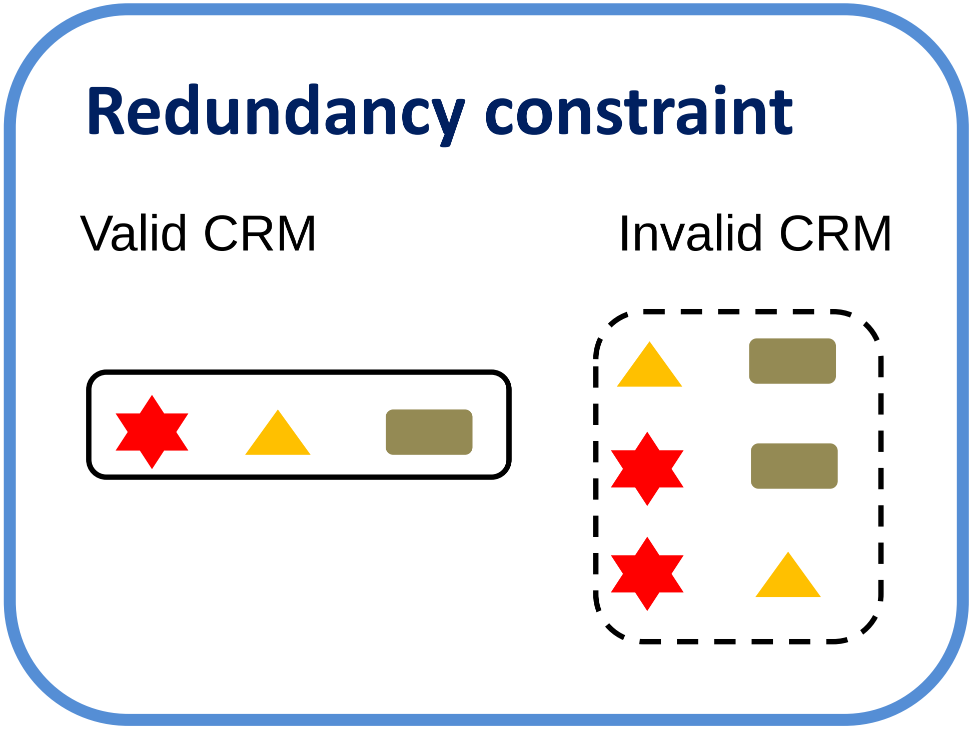
(support in itemset mining)

CRM detection



Only motif instances that occur in each others proximity can contribute to a valid motif set (CRM)

CRM detection



Redundancy constraint

Valid CRM

Invalid CRM

When enumerating all motif sets that meet the frequency and proximity constraint many subsets will occur in the same sequences and be composed of the same instances

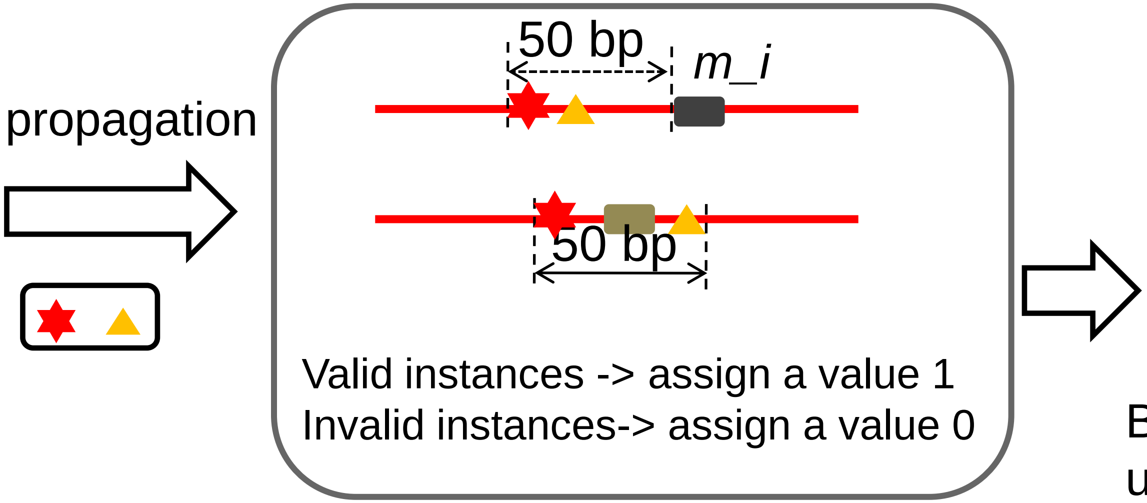
Only the motif set with more motifs will be retained

Removing redundant motif sets (CRMs) drastically increases the computation time (closeness in itemset mining)

CRM detection

Propagation: using a constraint to remove values from the domain of variables that would violate the constraint

Illustrate with the proximity constraint



Valid instances -> assign a value 1
Invalid instances-> assign a value 0

	m1	m2	m3	...	m66
seq 1	0	0	1	...	0
seq 2	0	1	0	...	0
seq 3	1	0	1	...	0
...					
seq k	0	1	1	...	0

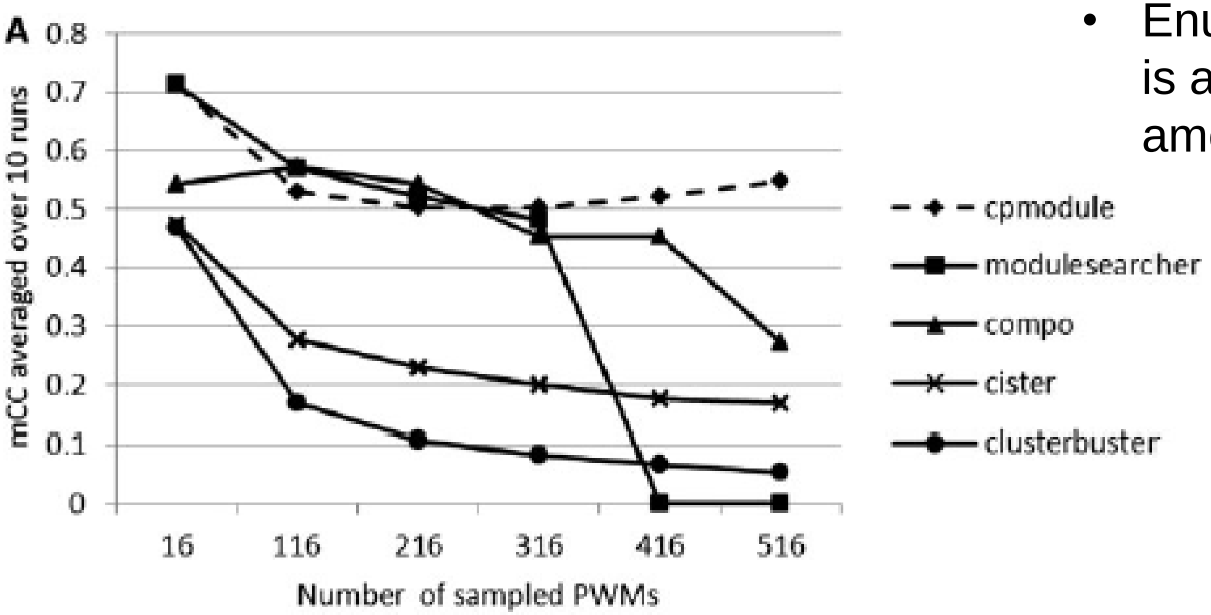
Binary matrix dynamically updated

Whether motif m_i is in the proximity of the motifs in motifset on sequence j?

CRM detection

Benchmarked on a synthetic dataset

- Xie et al. 2008 (22 sequences)
- 516 TRANSFAC PWMs
- Motifs inserted from 3 known PWMs



CPModule: performances similar to state-of-the-art algorithms on a synthetic dataset

BUT

- Able to deal with much larger sequence sets
- Enumerating all solutions allows it is able to rank the true solution amongst all solution

Guns et al. 2010 BIBM
Sun et al. NAR, 2011

Bioinformatics and datamining

Optimal bioinformatics tool

- Right heuristics
- Proper biological assumptions
- Room for experimenting with different assumptions

- Modular code
- Sustainable code
- High performance
- Latest algorithmic developments

Usefulness of declarative framework (Problog, ASP, Constrained based programming)

Acknowledgements

KUL/CMPG

- Ivan Ischukov*
- Hong Sun*
- Valerie Storms*
- Pieter Meysman*
- Kristof Engelen*
- Lore Cloots*
- Peyman Zarrineh*
- Riet De Smet*
- Karen Lemmens*
- Abeer Fadda*

UGENT/KUL/DBN

- Carolina Fierro
- Yan Wu
- Aminael Sanchez
- Marleen Claeys
- Dries De Maeyer
- Sergio Pullido
- Qiang Fu

University of Bristol

- Tijl Debie

KUL/Computer science

- Luc De Raedt
- Siegfried Nijssens
- Joris Renkens
- Tias Guns
- Tan Levan

http://bioinformatics.psb.ugent.be/DBN/