Teadmiste kaevandamine bioloogilistest andmetest

Jaak Vilo

Teoreetilise arvutiteaduse talvekool,
Arula, 3. veebruar 2003

Eesmärgid

- Tutvustada bioloogiliste andmete analüüsi (kaevandamise) problemaatikat
- Kas bioinformaatikal on midagi pakkuda teoreetilisele arvutiteadusele?
- Näiteid bio-andmete kaevandamisest

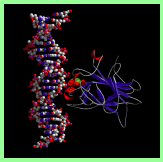
What is Bioinformatics?

- **Bioinformatics is the use of information technology to store and analyze genetic information**
- Bioinformatic researchers **develop** and apply **computing tools** to extract the **secrets of the life and death** of organisms from the **genetic** blueprints and **molecular** structure **stored in digital collections**

DNA määrab funktsiooni (?)

DNA → **VALK** → **STRUKTUUR**

GenBank / EMBL Bank — SwissProt/TrEMBL — PDB/Molecular Structure Database

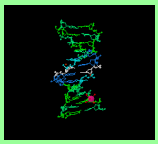


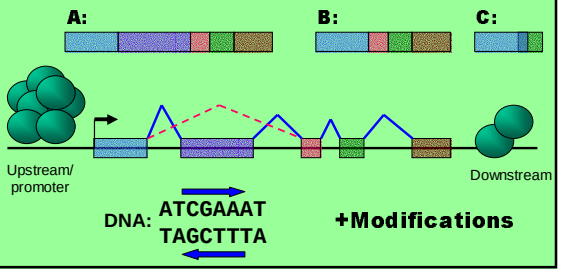
4 Nucleotides

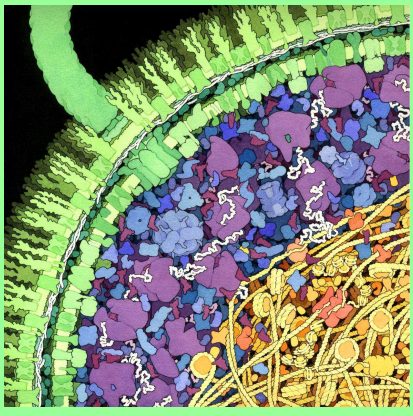
20+ Amino Acids
(3nt ➡ 1 AA)

Funktsioon?

A Simple Gene







David S. Goodsell
http://www.scripps.edu/pub/goodsell/

Uurimisküsimusi

- Milline on iga geeni ja geeniprodukti (valk, RNA) funktsioon?
- Kuidas täpselt toimuvad bioloogilised protsessid?
  - Transkriptsioon ja translatsioon ja nende reguleerimine
  - Millised valgud interakteeruvad ja miks
  - Metaboolsed rajad ja võrgud
  - Signaali ülekanne rakus ja organismis

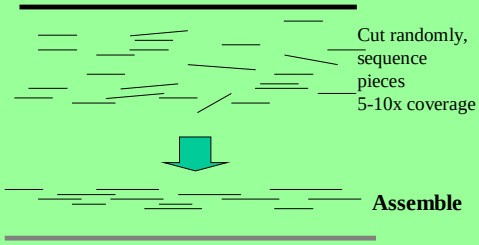
Kus vaja arvutiteadust?

- Andmete kogumise protsessi tugi
- Toorandmetest üldistatud andmete saamine
- Andmete haldamine
- Andmete analüüs
- Teadmiste esitamine
- Modelleerimine ja simulatsioon

Excerpts from curricula descriptions

- New experimental techniques generating mass data are being developed.
- Every new biological research idea requires a specifically tailored, algorithmic approach, raising an abundance of challenging questions in algorithm design, analysis and implementation.
- The demands of biosequence analysis require advances in many classical fields of algorithmics. Our recent work concentrates on suffix trees and related index structures, on efficient pattern matching in strings and trees, and on a new, algebraic style of dynamic programming.

DNA (shotgun) sequencing



shortest supersequence

Data Mining
Andmete kaevandamine

- Data Mining ja Knowledge Discovery from Databases (KDD) on uued arvutiteaduse alad, palju ühist statistikaga
- Eesmärgiks on väga suurte andmekogude analüüs
- Otsi (lihtsaid) reegleid ja seoseid mis kehtivad (vähemalt teatud osas) andmetes
- Statistika, masinõppimine, andmebaasiteooria, + domain knowledge

Bio-andmete kaevandamine

1. Millised bio-andmed on olemas ja kuidas neid võiks kasutada
2. Andmete kogumine, puhastamine, ettevalmistamine ja ühendamine
3. Analüüs arvutil (peamine algoritm)
4. Tulemuste esitamine, reeglite tuvastamine, visualiseerimine
5. Tulemuste tõlgendamine

TGTTCTTTCTTCTTTCATACATCCTTTTCCTTTTTTTCC
TTCTCCTTTCATTTCCTGACTTTTAATATAGGCTTACCA
TCCTTCTTCTCTTCAATAACCTTCTTACATTGCTTCTTC
TTCGATTGCTTCAAAGTAGTTCGTGAATCATCCTTCAAT
GCCTCAGCACCTTCAGCACTTGCACTTCATTCTCTGGAA
GTGCTGCACCTGCGCTGTCTTGCTAATGGATTTGGAGTT
GGCGTGGCACTGATTTCTTCGACATGGGCGGCGTCTTCT
TCGAATTCCATCAGTCCTCATAGTTCTGTTGGTTCTTTT
CTCTGATGATCGTCATCTTTCACTGATCTGATGTTCCTG
TGCCCTATCTATATCATCTCAAAGTTCACCTTTGCCACT
TTCCAAGATCTCTCATTCATAATGGGCTTAAAGCCGTAC
TTTTTTCACTCGATGAGCTATAAGAGTTTTCCACTTTTA
GATCGTGGCTGGGCTTATATTACGGTGTGATGAGGGCGC
TTGAAAAGATTTTTTCATCTCACAAGCGACGAGGGCCCG
AGTGTTTGAAGCTAGATGCAGTAGGTGCAAGCGTAGAGT
CTTAGAAGATAAAGTAGTGAATTACAATAGATTCGATAC

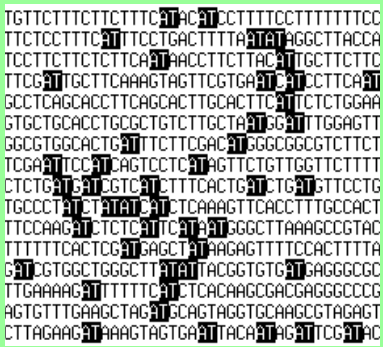
A Challenge Problem (P. Pevzner, 2000)

- Insert into every sequence a 15-mer where 4 positions have been randomly changed
- Challenge: discover what was the original sequence inserted
- Why? 4^15 sequences in total, variants can differ as much as in 8 positions out of 15

Tekstist mustriga otsimine

- Kas **ATGCAGA** esineb tekstis?
- Kas **ATGCAGA** esineb tekstis ligilähedselt (ATCCAGA, ATGCGA)?
- Kas **A[TA].C[CG].{3,7} A[TA].C[CG]** esineb tekstis?
- Kui kaua võtab aega ülaltoodud päringutele vastamine?
- Kuidas eeltöödelda ja indekseerida?
  - Suffiksipuud ja massiivid

Patterns: AT



Algoritmid

- Tehniline: loenda kõik etteantud stringides esinevad mustrid ja nende sagedused
- Bioloogiline: kogu kokku sarnaselt ekspresseerunud geenide promootorid ja otsi võimalikke transkriptsiooni-faktorite seondumiseks soodsaid mustreid

Pattern Discovery

1. Choose the language (formalism) to represent the patterns
2. Choose the rating for patterns, to tell that one pattern is “better” than other
3. Design an algorithm that **finds the best patterns** from the pattern class, **fast.**

Patterns: AT

Patterns: **WHAT** ([AT][ACT]AT)

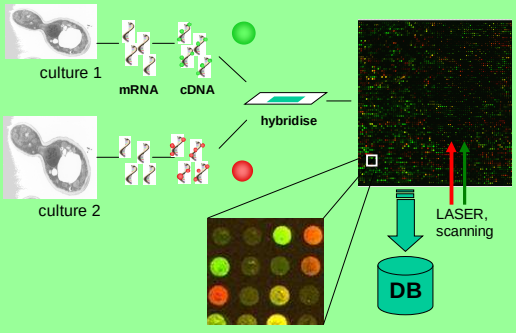
Bioinformaatika algoritmid

- Mitte ainult arvutiressursside küsimus
- Viisid kuidas kombineerida andmeid
- Kas saadav tulemus on bioloogiliselt relevantne
- Kas see edendab meie arusaamist bioloogilistest fenomenidest?

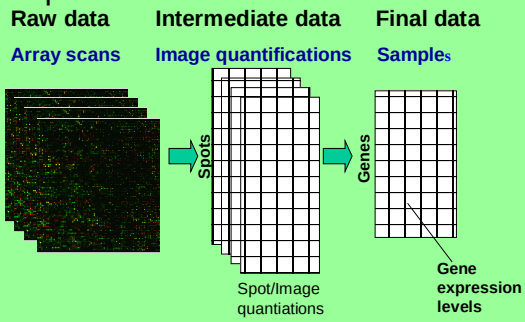
Näidis-meetodid

- Geenieksprессiooni analüüs
- Transkriptsioonifaktorite seondumiskohtade ennustamine
- Valk-valk interaktsioonide analüüs
- G-valk retseptorite ja G-valkude seoste analüüs
- Teadustekstide analüüs

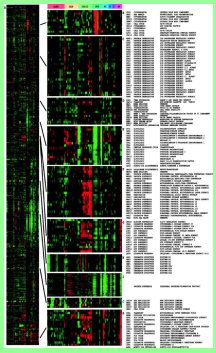
Analysis of biological samples with microarrays



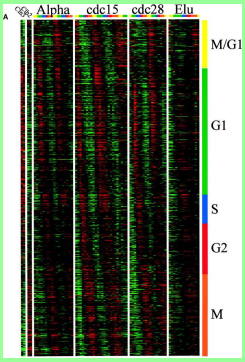
From microarray images to gene expression data

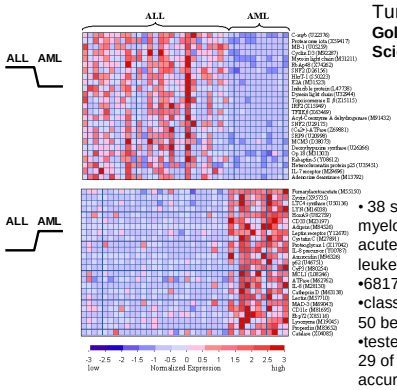


Eisen etal, PNAS 98



Spellman etal Mol Biol Cell 98



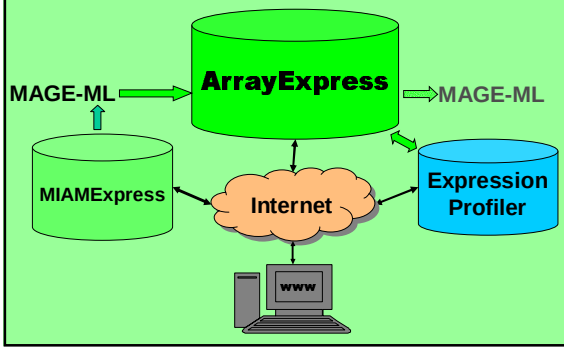


Tumor classification
Golub et al,
Science Oct 15th 1999

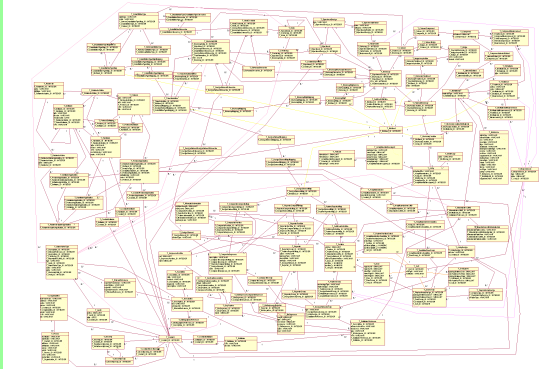
- 38 samples of acute myeloic leukemia (AML) and acute lymphoblastic leukemia (ALL)
- 6817 genes
- classificator built based on 50 best correlated genes
- tested on 34 new samples, 29 of them predicted accurately

ArrayExpress

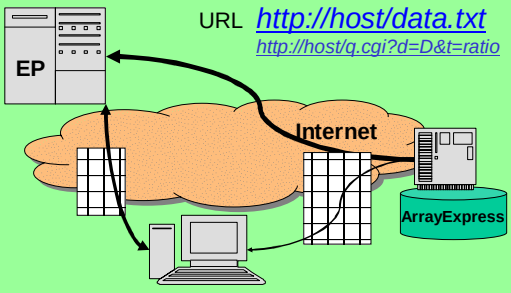
a public repository for microarray gene expression data



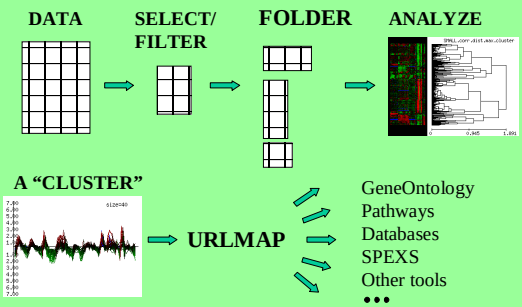
Simplified ArrayExpress model



Getting data into EP



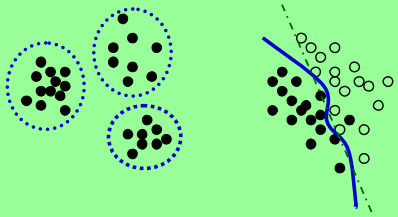
Expression Profiler: **EPCLUST**



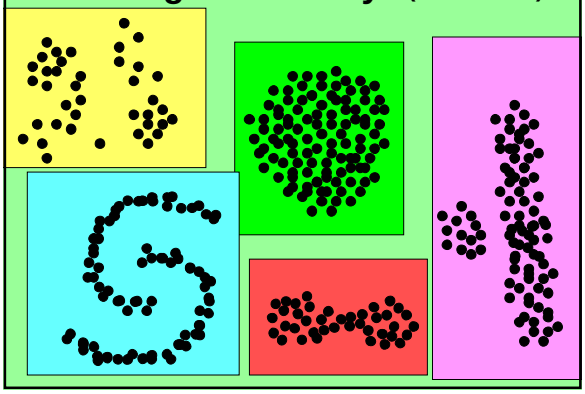
Unsupervised vs. Supervised

Find groups inherent to data (clustering)

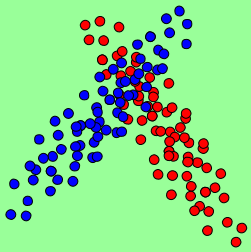
Find a "classifier" for known classes



Clustering – it's "easy" (for humans)



Clustering cont…

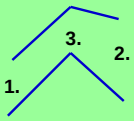


Distance measures:

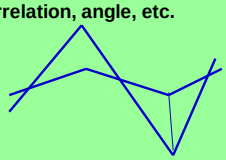
which two profiles are similar to each other?

Euclidean , Manhattan etc.

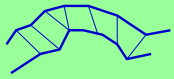
Rank correlation



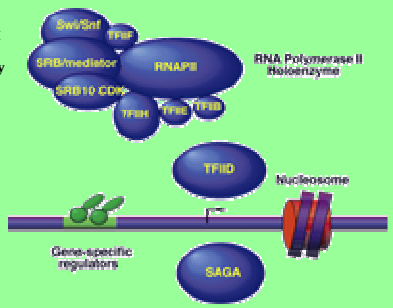
Correlation, angle, etc.



Time warping

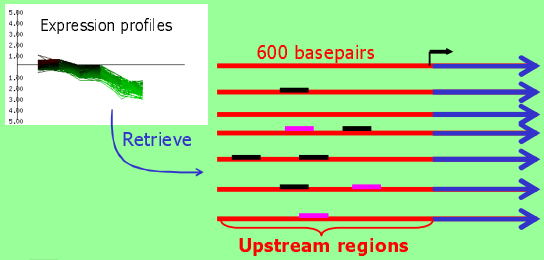


Model of RNA Polymerase II Transcription Initiation Machinery. **The machinery depicted here encompasses over 85 polypeptides in ten (sub) complexes**: core RNA polymerase II (RNAPII) consists of 12 subunits; TFIIH, 9 subunits; TFIIE, 2 subunits; TFIIF, 3 subunits; TFIIB, 1 subunit, TFIID, 14 subunits; core SRB/mediator, more than 16 subunits; Swi/Snf complex, 11 subunits; Srb10 kinase complex, 4 subunits; and SAGA, 13 subunits.



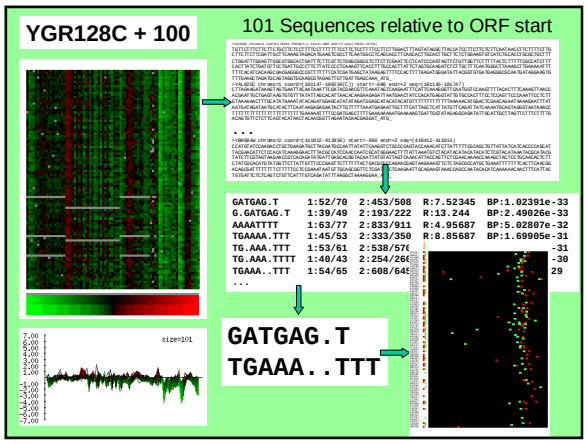
F.C.P. Holstege, E.G. Jenning, J.J. Wyrick, Tong Ihn Lee, C.J. Hengartner, M.R. Green, T.R. Golub, E.S. Lander, and R.A. Young
Dissecting the Regulatory Circuitry of a Eukaryotic Genome
Cell 95: 717-728 (1998)

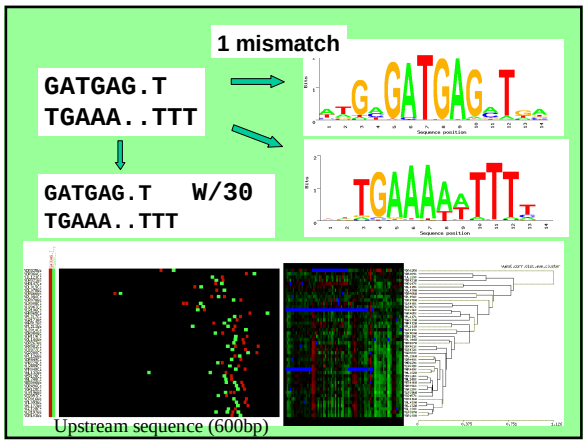
Cluster of co-expressed genes, pattern discovery in regulatory regions

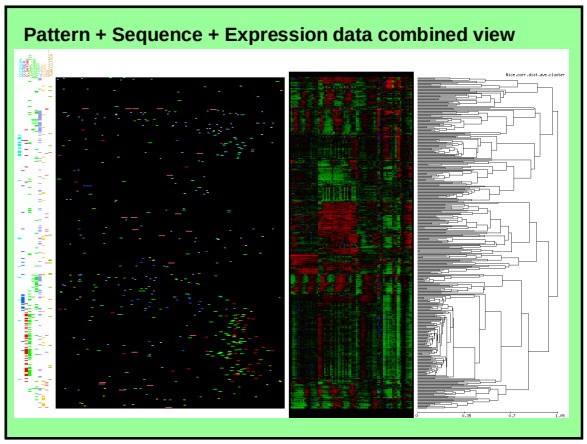


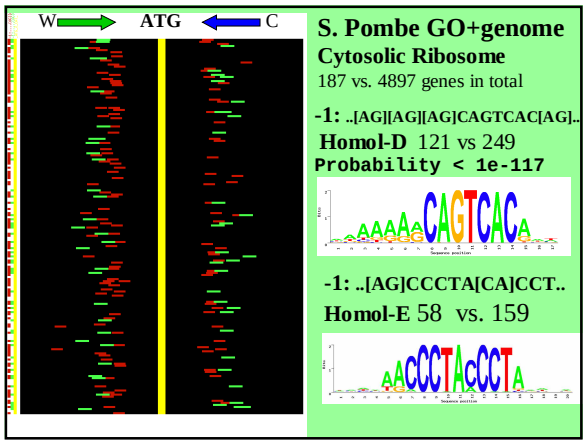
Find patterns over-represented within cluster

Genome Research 1998;
ISMB (Intelligent Systems in Mol. Biol.) 2000



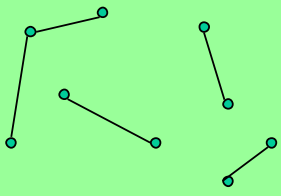


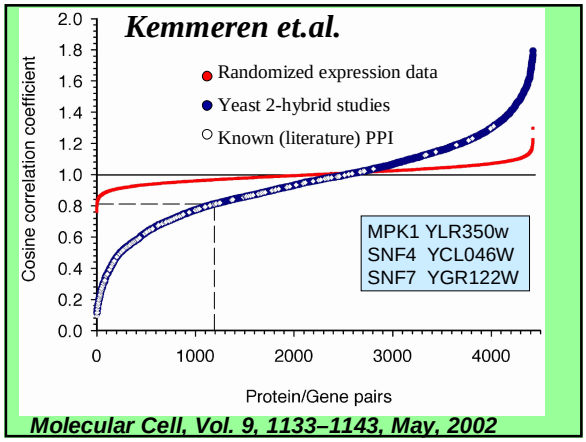




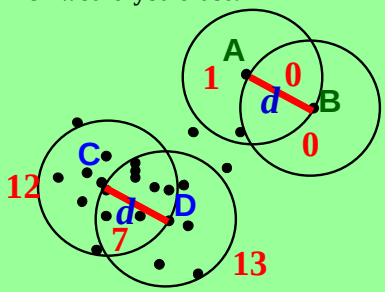
Valk-valk interaktsioonid (PPI pairs)

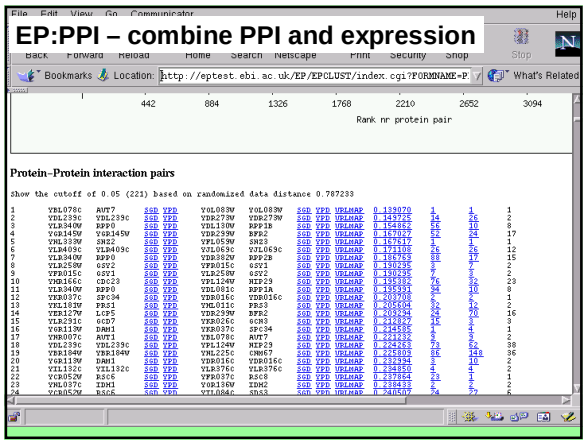
Probleem: uued tehnoloogiad võimaldavad leida tuhandeid potentsiaalseid interaktsioone. Millised neist on rohkem usaldusväärsed?





Interacting pairs of proteins
A and B; C and D
Which would you trust?

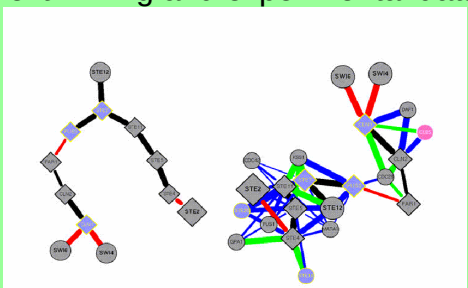




Text mining

- Teadusartiklid sisaldavad tohutul hulgal teavet
- Kuidas seda süstemaatiliselt kokku koguda ja esitada?
- Tuvasta Medline abstraktides esinevad geenide nimed ja seosed nende vahel
- Moodusta selle info põhjal graafilised esitlused seoste kohta

Signal transduction pathway from Text mining and experimental data

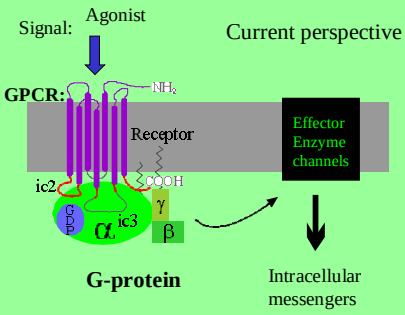


Lappe, Schlitt, Dietmann, Holm, submitted (2003)

GPCR

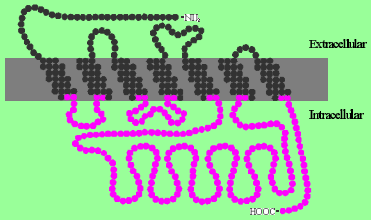
- G-valk retseptorid on ühed tähtsamad ravimite sihtmärgid
- Küsimus – kas saab ennustada retseptori põhjal millise signaali ülekande raja see käivitab?
- Millised G-valgud seonduvad selle GPCR valguga?

GPCR coupling

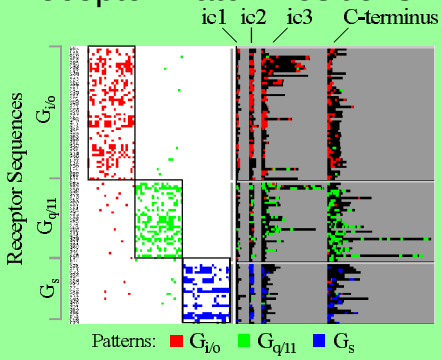


Our Computational Approach

- Using a new membrane topology prediction algorithm (designed specifically for GPCRs), we constrained our pattern search to the intracellular domains of ≈ 100 receptor sequences with well-characterised, and non-promiscuous coupling (split into G_s, $G_{i/o}$ and $G_{q/11}$)



Receptor Match Positions

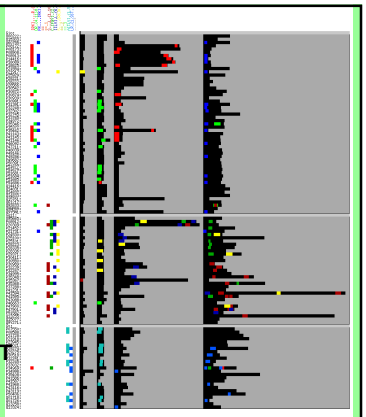


Croning, Vilo, Möller, *ISMB 2001*

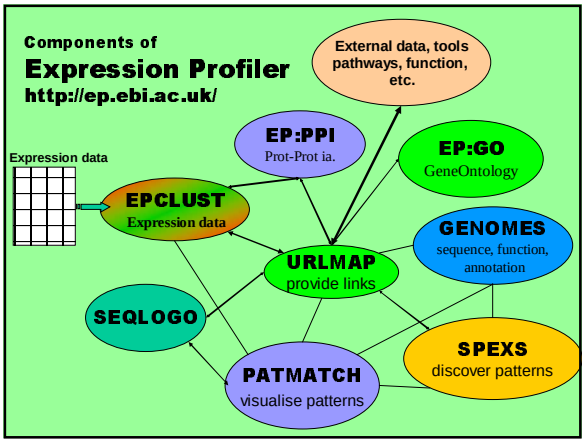
[RK]....R.{0,9}EK
DR.{4,11}H...[AGS]
FR....[RK].{0,3}L

S...L.{1,10}T[ILV]
C.[FWY].{2,11}K
[ILV].L.{6,10}A.T
S....[RK]A.{3,10}S

A[ILV].{1,5}Y..[ILV].T
LR.{1,9}T...[ILV]



Components of
Expression Profiler
http://ep.ebi.ac.uk/

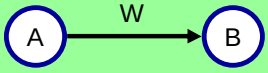


Networks

- Graphical models
- Directed labelled graph

- Nodes ⇔ genes
- Arcs/Edges ⇔ relationships
- Labels ⇔ types of relationships

Graph drawing



Start node (gene) | Connection weight, w | End node (gene)

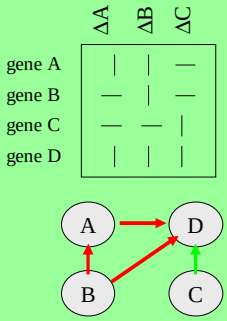
Different interpretation of arcs

- Edges can have different meanings, hence different networks

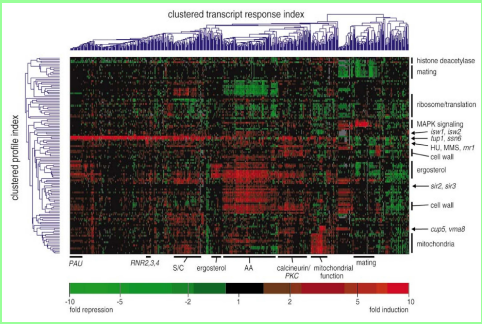
- Binding site for A is in front of B
- Proteins A and B interact
- Deletion of gene A affects expression of B (is somewhere in regulation cascade)
- “Literature” mentions genes together

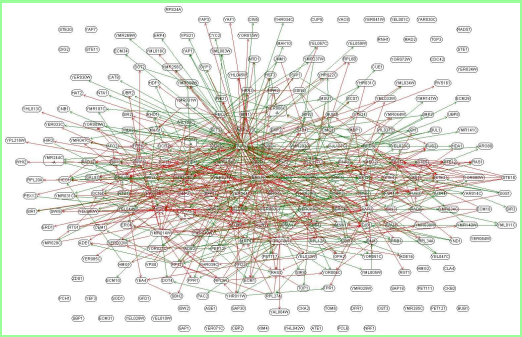
Features/distributions that do not depend on discretisation thresholds

- Visual inspection, biological interpretation
- General statistics and features of the graphs
- Indegree/Outdegree
- Complexity of the networks
- What is the modularity?
  - How many components?
  - Deletion of hot-spots, does it break the net?



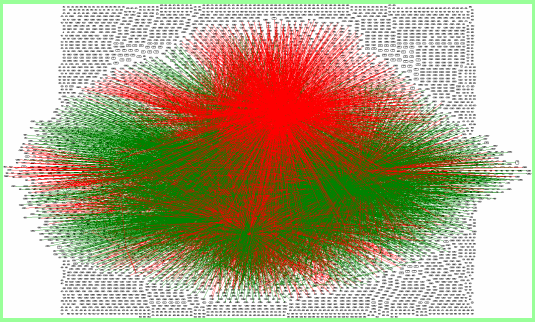
Hughes, T. R. et al: “Functional Discovery via a Compendium of Expression Profiles”, Cell 102 (2000), 109-126.





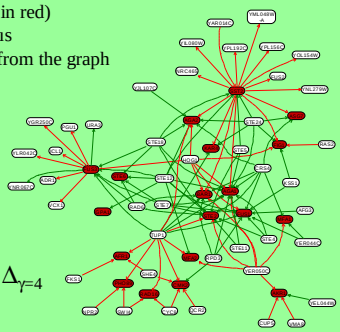
Green arrows - upregulation
Red arrows - downregulation
Thickness of arrow represents certainty of direction (up/down)

A complete graph

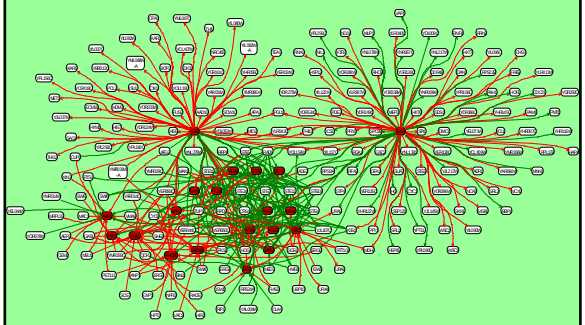


Filter

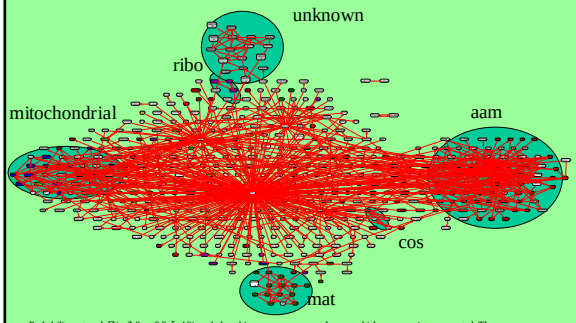
- choose a list of genes
  (MATING, marked in red)
- filter for these genes plus
  neighbouring genes from the graph



Mutation network $\Delta_{\gamma=4}$



Mutation network $\Delta_{\gamma=2}$



Probability network Π(γ=2.0, τ=0.8, ξ=10), underlayed in green are groups of genes which are more interconnected. The genes are coloured according to annotation in YPD ("cellular role"). The genes which are more interconnected are involved in the same cellular processes, like mating behaviour (mat, green), aminoacid metabolism (aam, red), cos gene family (cos, light blue), mitochondrial function (mitochondrial, dark blue), ribosome (ribo, purple) and a group of genes of unknown function (unknown, grey)

Kokkuvõtteks

- (Molekulaar-) Bioloogilisi andmeid on palju erinevaid tüüpe
- Bioloogiliste andmete mahud kasvavad praegu eksponentsiaalse kiirusega
- Arvutianalüüs on ainuke reaalne viis neid andmeid ära kasutada
- See eeldab loomingulist lähenemist nii küsimusepüstituses kui ka tehniliselt heade algoritmide kasutust või väljamõtlemist

Bio-andmete kaevandamine

1. Millised bio-andmed on olemas ja kuidas neid võiks kasutada
2. Andmete kogumine, puhastamine, ettevalmistamine ja ühendamine
3. Analüüs arvutil (peamine algoritm)
4. Tulemuste esitamine, reeglite tuvastamine, visualiseerimine
5. Tulemuste tõlgendamine

Acknowledgements

Alvis Brazma + the EBI microarray team
Misha Kapushesky, EBI (EP development)
Patrick Kemmeren, **Frank Holstege**, Utrecht U. (PPI)
Esko Ukkonen, Kimmo Palin, Helsinki Univ.

Meelis Kull
Tanel Kaart
Hedi Peterson
Kristo Käärmann
jne.

Maido Remm
Reidar Andreson
…