TARTU ÜLIKOOL
UNIVERSITY OF TARTU

BIIT

Advances in Bioinformatics

2002-2007

Jaak Vilo
vilo@ut.ee

http://biit.cs.ut.ee

Timeline 2002-2007

- 2002 July – move to Estonia (EGeen Inc; EBC)
- 2003 Start teaching at UT
- 2004/09 Docent (0.5) at university
- 2006/01 Senior Researcher
- 2007/12 Professor
- BIIT has now (Jan 2008)
  - 1 postdoc
  - 1 guest lecturer
  - 10 PhD students
  - 10+ students

Year	BSc	MSc
2003	2	0
2004	6	1
2005	9	1
2006	4	3
2007	4	3
Total	**25**	**8**

Research Focus

- Algorithms (Data Mining & Bioinformatics)
- Tools (web based)
- Databases & information systems

- Gene regulation & Systems Biology
- Cancer; Stem Cells;
- Microarray & other high-throughput data

Grants and Projects Completed

- **BiGeR** Estonian Science Foundation ETF5724 (2003-2007)
- **DMMA** Estonian Science Foundation ETF5722 (2003-2006)
- "Base funding" (start-up grant for a new research group) from University of Tartu (2005-2006)
- ATD, **Alternative Transcript Diversity**, EU FP6 STREP (2004-2007) LSHG-CT-2003-503329
- FunGenES, **Functional Genomics of Embryonic Stem Cells**: EU FP6 Integrated Project, subcontractor (2006-2007). LSHG-CT-2003-503494
- Baltic GRID non-funded partner

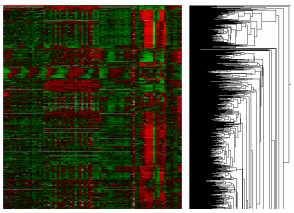
Current grants and projects

- COBRED -- **Colon and Breast cancer Diagnostics** (2007-2010) EU FP6 STREP, LSHB-CT-2007-037730
- ENFIN, **Enabling Systems Biology**. EU FP6 Network of Excellence (2005-2010) LSHG-CT-2005-518254 QureTEC
- ESNATS , Embryonic Stem Cell-Based Alternative Testing Strategies. Duration of Project: 5y (2008-2013). Currently under negotiation. QureTEC

- Target funding (2006-2011), The methods, environments, and applications for solving large and complex computational problems. (SF0182712s06)
- ETF7437 (2008-2011) Multi-experiment gene expression data matrix analysis (MEM) (Abstract)
- Estonian Language Technology Research Programme: Dictionary informatics (information retrieval) (2005-2008)

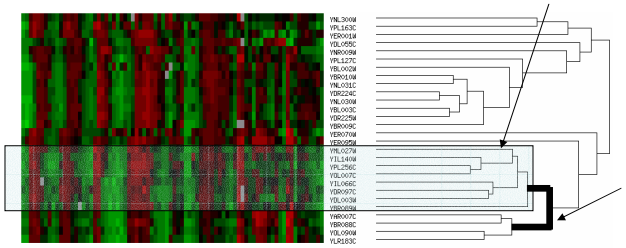
Fast Approximate Hierarchical Clustering using Similarity Heuristics

Hierarchical clustering is applied in gene expression data analysis, number of genes can be 20000+



Hierarchical clustering:

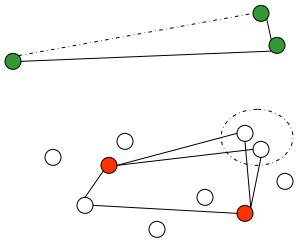
Each subtree is a cluster.



Hierarchy is built by iteratively joining two most similar clusters into a larger one.

Fast Hierarchical Clustering

Avoid calculating all $O(n^2)$ distances:



– Estimate distances

– Use pivots

– Find close objects

– Cluster with partial information

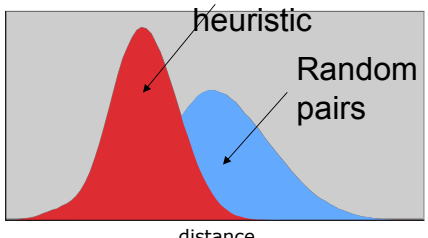
Meelis Kull

Approximate Hierarchical Clustering

• Full hierarchical clustering calculates similarities of each object with every other object (SLOW)

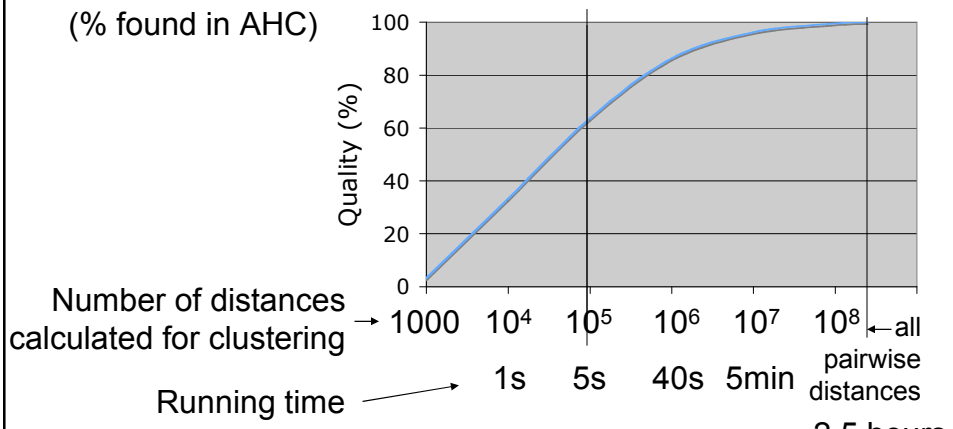
• Our approximate clustering calculates a small fraction of all similarities

• We have developed heuristics to find pairs of similar objects efficiently



Biological relevance

- Enrichment of (any) biological function in a cluster
- Find all enrichments in full clustering, compare quality



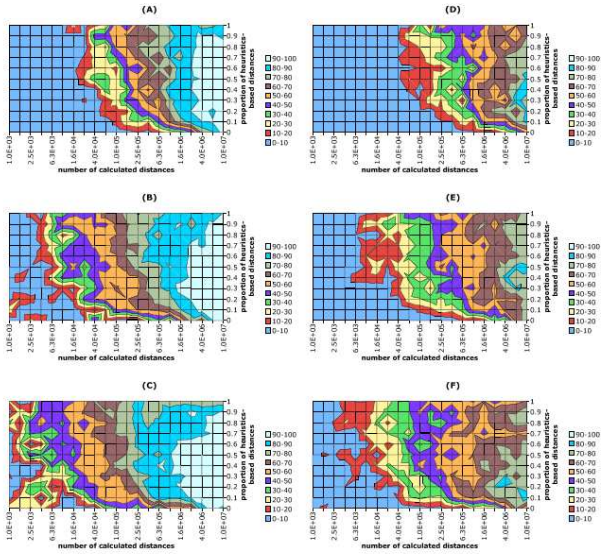
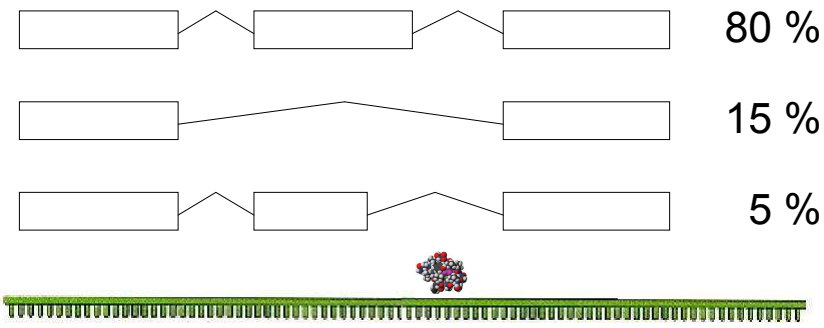
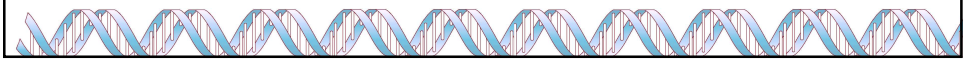


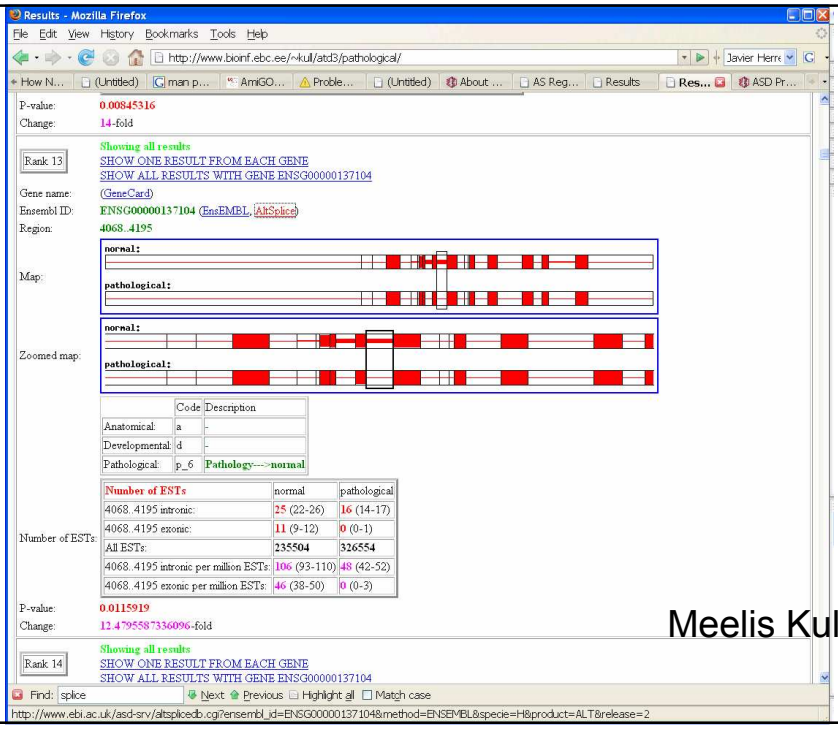
Fig. 4. GO50 and GO25 quality of HappieClust for different numbers of pivots. (A) $q = 5$, GO50; (B) $q = 10$, GO50; (C) $q = 20$, GO50; (D) $q = 5$, GO25; (E) $q = 10$, GO25; (F) $q = 20$, GO25.

Find disease-specific gene variants



Proteins can affect splicing by taking part in or interacting with the splicing complex

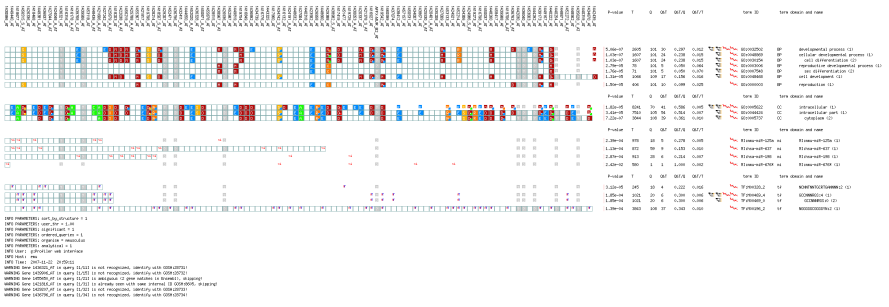




Meelis Kull

Tools: g:Profiler

- Characterise gene lists
- Manipulate gene ID-s
  - synonyms
  - orthologs
- Fast, rich GUI, public easy free access
- Machine-readable outputs (for integration)



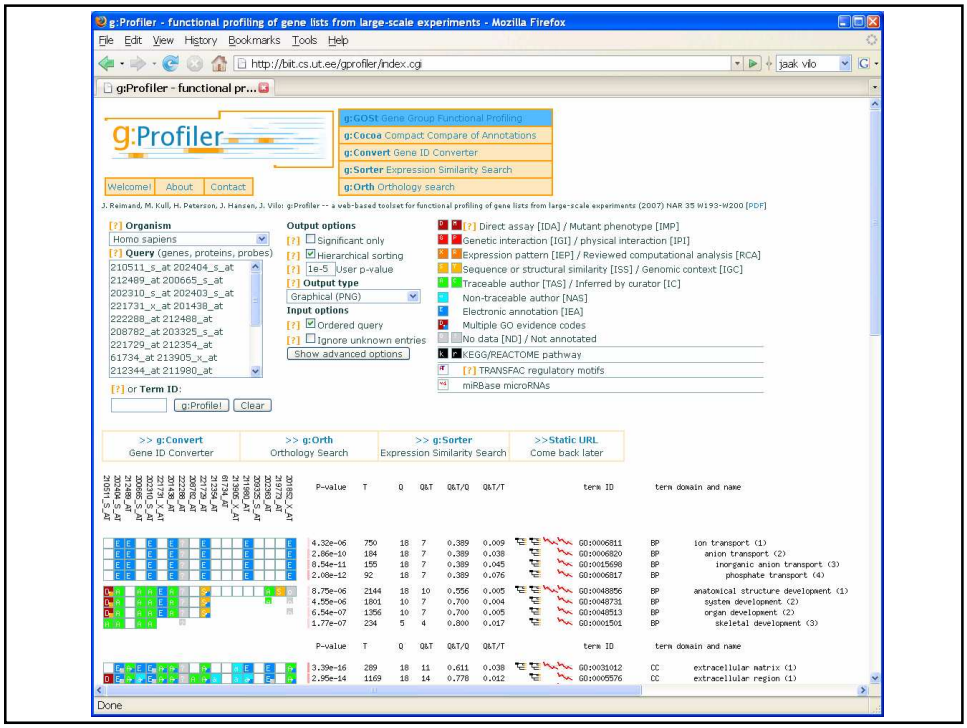
Nucleic Acids Research, 2007, **1–8**
doi:10.1093/nar/gkm226

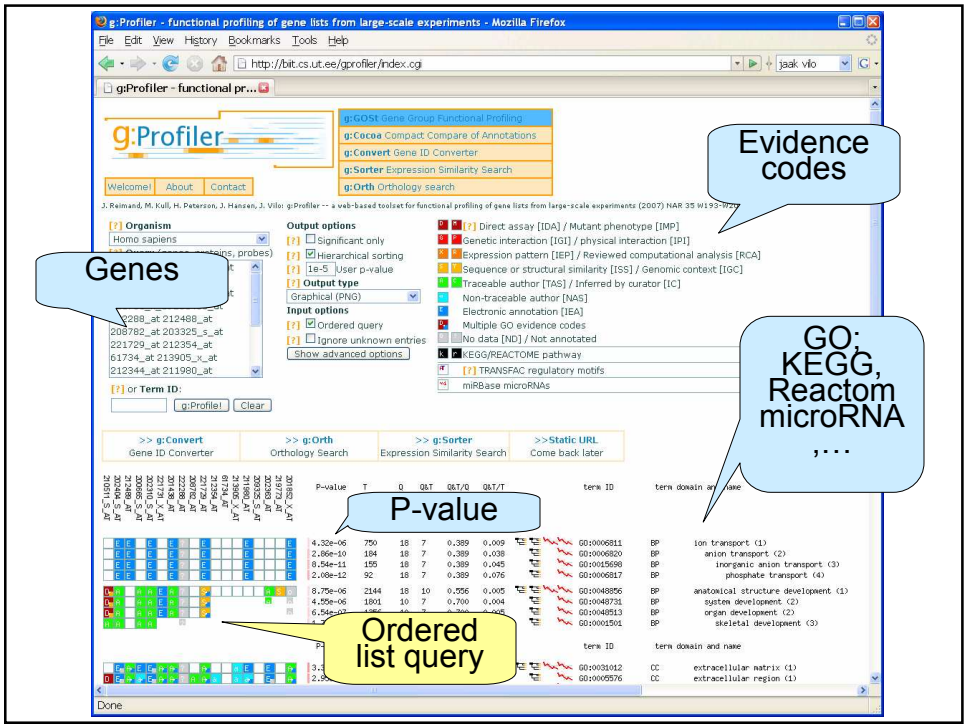
g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments

Jüri Reimand[1], Meelis Kull[1,2,3], Hedi Peterson[2,3], Jaanus Hansen[1] and Jaak Vilo[1,2,3,*]

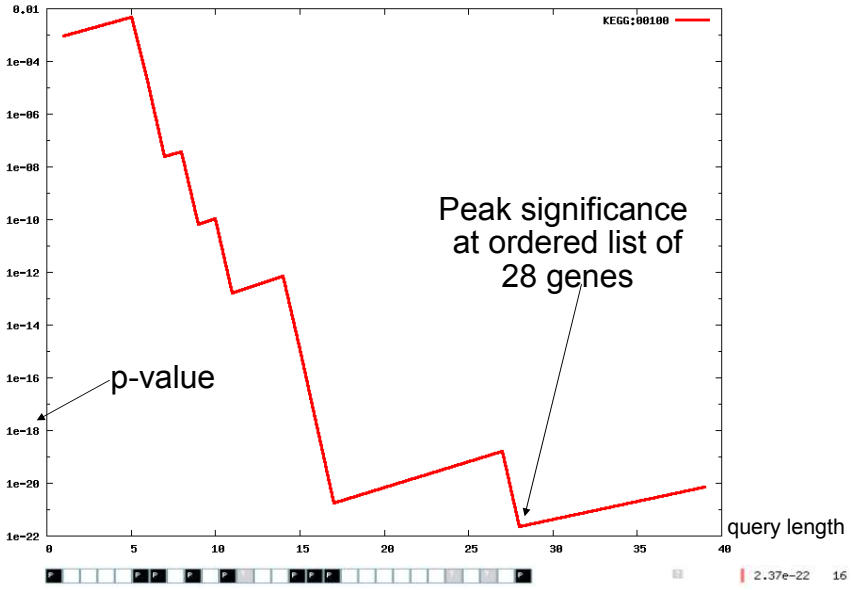
[1]Institute of Computer Science, University of Tartu, Liivi 2, 50409 Tartu, Estonia, [2]Estonian Biocentre, Riia 23b, 51010 Tartu, Estonia and [3]EGeen, Ülikooli 6a, 51003 Tartu, Estonia

Received January 31, 2007; Revised March 22, 2007; Accepted March 28, 2007

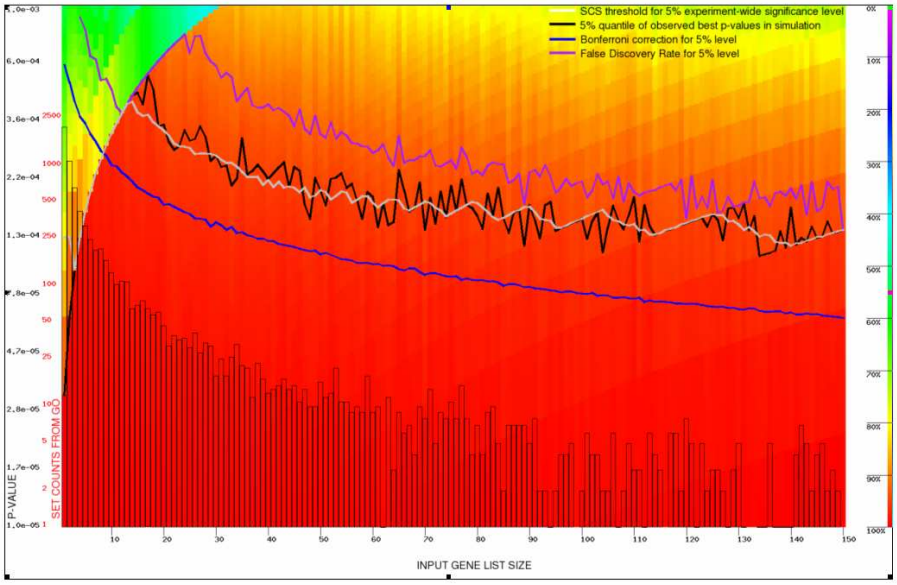

g:Profiler - functional profiling of gene lists from large-scale experiments - Mozilla Firefox
http://biit.cs.ut.ee/gprofiler/index.cgi
g:Profiler
g:GOSt Gene Group Functional Profiling
g:Cocoa Compact Compare of Annotations
g:Convert Gene ID Converter
g:Sorter Expression Similarity Search
g:Orth Orthology search
Welcome! About Contact
Organism
Homo sapiens
Query (genes, proteins, probes)
Output options
Significant only
Hierarchical sorting
1e-5 User p-value
Output type
Graphical (PNG)
Input options
Ordered query
Ignore unknown entries
Show advanced options
or Term ID:
g:Profile! Clear
Direct assay [IDA] / Mutant phenotype [IMP]
Genetic interaction [IGI] / physical interaction [IPI]
Expression pattern [IEP] / Reviewed computational analysis [RCA]
Sequence or structural similarity [ISS] / Genomic context [IGC]
Traceable author [TAS] / Inferred by curator [IC]
Non-traceable author [NAS]
Electronic annotation [IEA]
Multiple GO evidence codes
No data [ND] / Not annotated
KEGG/REACTOME pathway
TRANSFAC regulatory motifs
miRBase microRNAs
>> g:Convert Gene ID Converter
>> g:Orth Orthology Search
>> g:Sorter Expression Similarity Search
>>Static URL Come back later
P-value
term ID
term domain and name
Done


g:Profiler - functional profiling of gene lists from large-scale experiments - Mozilla Firefox
Evidence codes
Genes
GO: KEGG, Reactom microRNA ,...
P-value
Ordered list query
Done

KEGG: Biosynthesis of



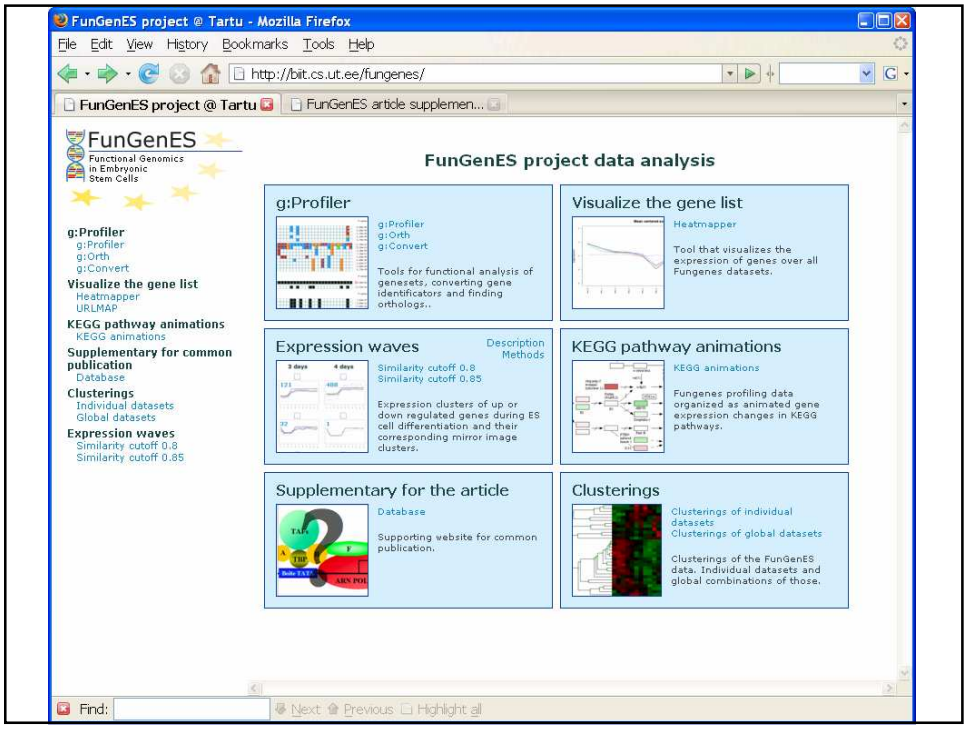
SCS - Set Counts and Sizes

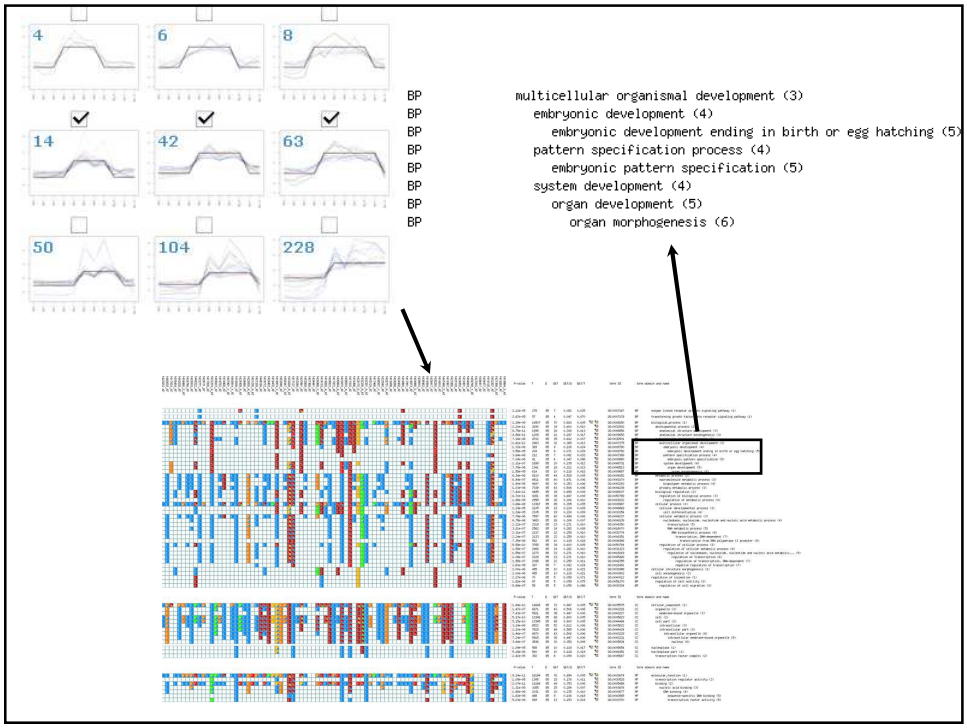


			27			7	1.76e-11	581	GO:0007155	BP	cell adhesion (1)
						3	4.46e-05	23	GO:0006775	BP	fat-soluble vitamin metabolic process (1)
						3	2.08e-05	18	GO:0006776	BP	vitamin A metabolic process (2)
		7				2	1.44e-05	49	GO:0045597	BP	positive regulation of cell differentiation (1)
		13					5.92e-06	173	GO:0045892	BP	negative regulation of transcription, DNA-dependent (1)
		7					4.21e-06	41	GO:0043406	BP	positive regulation of MAPK activity (1)
	42	109	59			21	4.60e-08	4146	GO:0065007	BP	biological regulation (1)
	42	100	56			20	3.56e-07	3834	GO:0050789	BP	regulation of biological process (2)
	14	34					4.87e-06	895	GO:0048519	BP	negative regulation of biological process (3)
13	43	95	58			19	7.19e-15	2605	GO:0032502	BP	developmental process (1)
12	30	73	43			14	2.48e-14	1733	GO:0048856	BP	anatomical structure development (2)
9	22	51	26			11	3.59e-13	997	GO:0009653	BP	anatomical structure morphogenesis (2)
2	3	7	4				3.95e-05	57	GO:0001763	BP	morphogenesis of a branching structure (3)
2	3	7	4	1			2.44e-05	53	GO:0048754	BP	branching morphogenesis of a tube (4)
		20	11		8		3.94e-05	437	GO:0032989	BP	cellular structure morphogenesis (3)
		20	11		8		3.94e-05	437	GO:0000902	BP	cell morphogenesis (4)
	5	14	12			3	2.95e-07	185	GO:0048646	BP	anatomical structure formation (2)
		19					3.93e-08	252	GO:0050793	BP	regulation of developmental process (2)
13	35	77	51	4		14	8.39e-14	1937	GO:0007275	BP	multicellular organismal development (1)
5	11	16		2			4.54e-05	398	GO:0009790	BP	embryonic development (2)
3	9			2			3.24e-06	189	GO:0007389	BP	pattern specification process (2)
	7						2.93e-05	139	GO:0003002	BP	regionalization (3)
	6						2.15e-05	89	GO:0009952	BP	anterior/posterior pattern formation (4)
	8	11	5				3.74e-06	144	GO:0035295	BP	tube development (2)
11	27	61	40			12	1.95e-11	1491	GO:0048731	BP	system development (2)
8	24	51	36			11	3.42e-10	1202	GO:0048513	BP	organ development (3)
	6	16	14			3	1.72e-08	207	GO:0001944	BP	vasculature development (4)
	6	16	14			3	1.43e-08	204	GO:0001568	BP	blood vessel development (5)
7	14	25	16			6	8.25e-08	441	GO:0009887	BP	organ morphogenesis (4)
	5	13	13			3	2.14e-08	177	GO:0048514	BP	blood vessel morphogenesis (5)
		11	10			3	1.15e-06	139	GO:0001525	BP	angiogenesis (6)
10	23	56	27		22	16	5.05e-08	1607	GO:0048869	BP	cellular developmental process (1)
10	23	56	27		22	16	5.05e-08	1607	GO:0030154	BP	cell differentiation (2)
	41	90	48				3.12e-06	3490	GO:0050794	BP	regulation of cellular process (1)
6	13	32					8.11e-06	836	GO:0048523	BP	negative regulation of cellular process (2)
2	7						6.23e-07	78	GO:0003006	BP	reproductive developmental process (1)
	6						5.82e-06	71	GO:0007548	BP	sex differentiation (2)
	14	38			15	9	5.72e-06	1066	GO:0048468	BP	cell development (1)
	7	18					3.91e-08	226	GO:0007167	BP	enzyme linked receptor protein signaling pathway (1)
		12					7.64e-06	151	GO:0007169	BP	transmembrane receptor protein tyrosine kinase signaling pathway (2)
					24	6	7.26e-11	647	GO:0002376	BP	immune system process (1)
					13	6	4.08e-06	370	GO:0006952	BP	defense response (1)
				2	14		2.22e-07	336	GO:0042221	BP	response to chemical stimulus (1)
					12		1.29e-07	226	GO:0001775	BP	cell activation (1)
					10		3.99e-06	210	GO:0045321	BP	leukocyte activation (2)
			9		20	6	5.24e-11	425	GO:0009605	BP	response to external stimulus (1)
					12		1.27e-06	280	GO:0007610	BP	behavior (1)
					12		8.57e-09	177	GO:0007626	BP	locomotory behavior (2)
			4		11		2.37e-10	101	GO:0042330	BP	taxis (3)
			4		11		2.37e-10	101	GO:0006935	BP	chemotaxis (4)
					19	5	2.34e-10	412	GO:0006955	BP	immune response (1)
					13	5	1.58e-07	277	GO:0009611	BP	response to wounding (1)

Data Analysis Environments

- FunGenES consortium – data analysis environment
- ~10 labs produced 150+ conditions (3-5x)
  - Affymetrix Gene Chips (22,000 probesets; ~30MB file each)
  - Normalise, extract data for analysis
  - study, analyse, visualise, ...
  - compare
- Clustering, visualisation, searching, etc.


FunGenES project @ Tartu - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://bit.cs.ut.ee/fungenes/
FunGenES project @ Tartu
FunGenES article supplemen...
FunGenES
Functional Genomics in Embryonic Stem Cells
FunGenES project data analysis
g:Profiler
g:Profiler
g:Orth
g:Convert
Tools for functional analysis of genesets, converting gene identificators and finding orthologs..
Visualize the gene list
Heatmapper
Tool that visualizes the expression of genes over all Fungenes datasets.
Expression waves
Description
Methods
Similarity cutoff 0.8
Similarity cutoff 0.85
Expression clusters of up or down regulated genes during ES cell differentiation and their corresponding mirror image clusters.
KEGG pathway animations
KEGG animations
Fungenes profiling data organized as animated gene expression changes in KEGG pathways.
Supplementary for the article
Database
Supporting website for common publication.
Clusterings
Clusterings of individual datasets
Clusterings of global datasets
Clusterings of the FunGenES data. Individual datasets and global combinations of those.
g:Profiler
g:Profiler
g:Orth
g:Convert
Visualize the gene list
Heatmapper
URLMAP
KEGG pathway animations
KEGG animations
Supplementary for the publication
Database
Clusterings
Individual datasets
Global datasets
Expression waves
Similarity cutoff 0.8
Similarity cutoff 0.85
Find:
Next
Previous
Highlight all


4
6
8
14
42
63
50
104
228
BP multicellular organismal development (3)
BP embryonic development (4)
BP embryonic development ending in birth or egg hatching (5)
BP pattern specification process (4)
BP embryonic pattern specification (5)
BP system development (4)
BP organ development (5)
BP organ morphogenesis (6)

Regulators for ES gene clusters

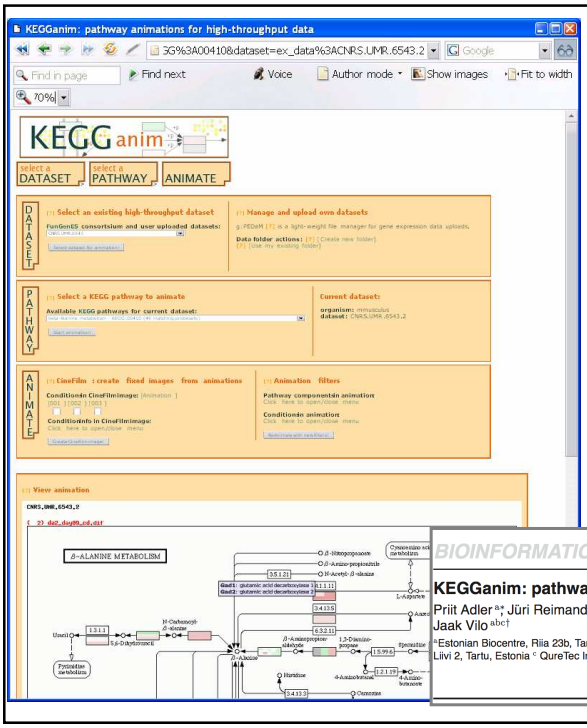
DELETED

FunGenES

- Analysed data for 4-5 groups; others used web
- 4 people visited Estonia (for a week) to analyse their data
- Consortium data – public analysis environment
- Couple of leads in functional validation

Tools: KEGGanim

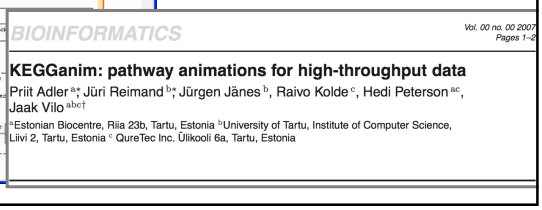
- Animate large-scale mRNA, proteomics, metabolomics data on KEGG pathways
- Create visualisations
  - animations
  - freeze panels
- Manipulate private data



biit.cs.ut.ee/KEGGanim
Adler et al
Bioinformatics, 2007

Overlay HT data over KEGG pathways

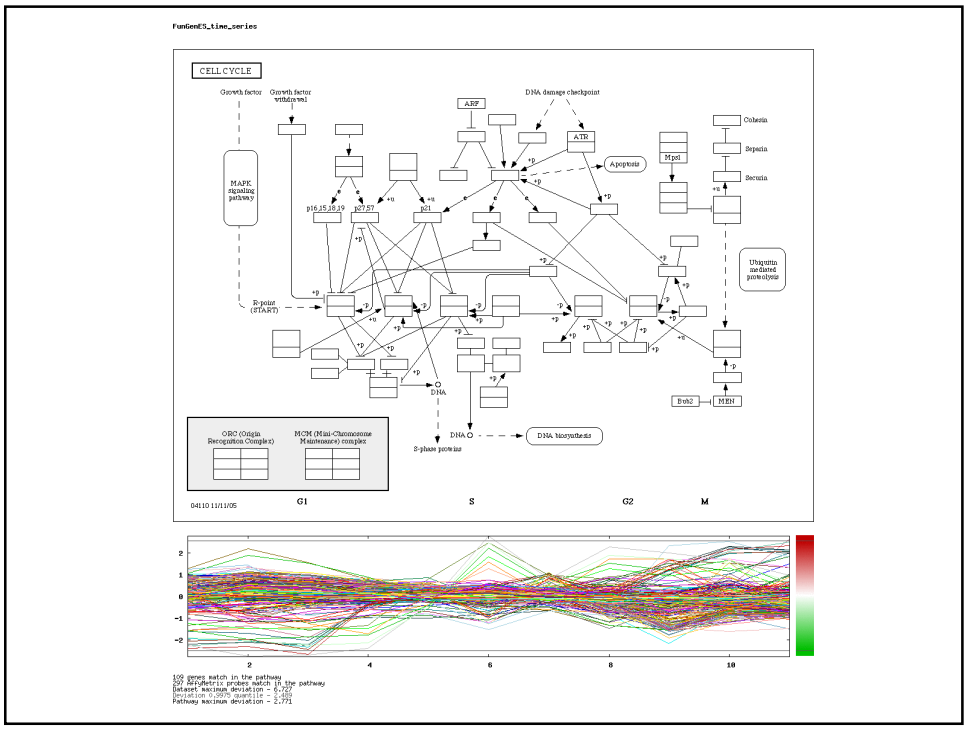
- Animate
- "Freeze panes" (cinefilm)

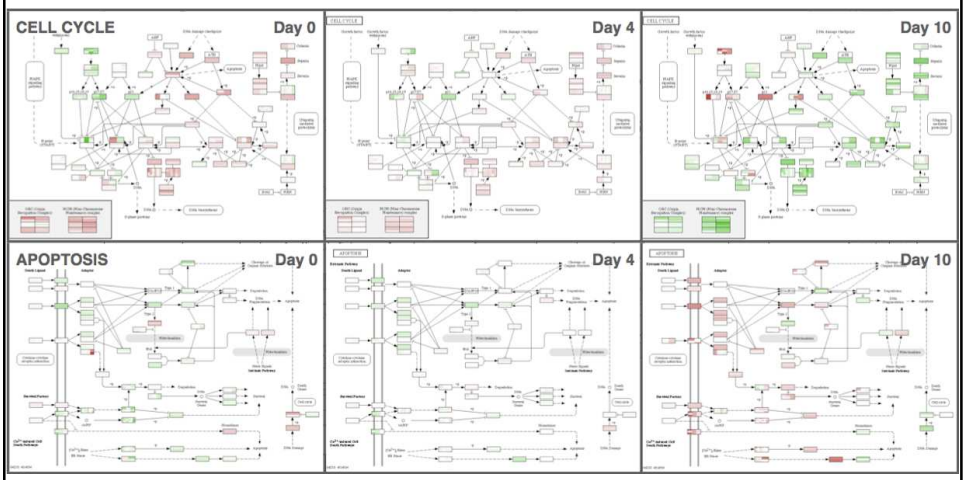

BIOINFORMATICS

KEGGanim: pathway animations for high-throughput data

Priit Adler [a*], Jüri Reimand [b*], Jürgen Jänes [b], Raivo Kolde [c], Hedi Peterson [ac], Jaak Vilo [abc†]

[a]Estonian Biocentre, Riia 23b, Tartu, Estonia [b]University of Tartu, Institute of Computer Science, Liivi 2, Tartu, Estonia [c] QureTec Inc. Ülikooli 6a, Tartu, Estonia

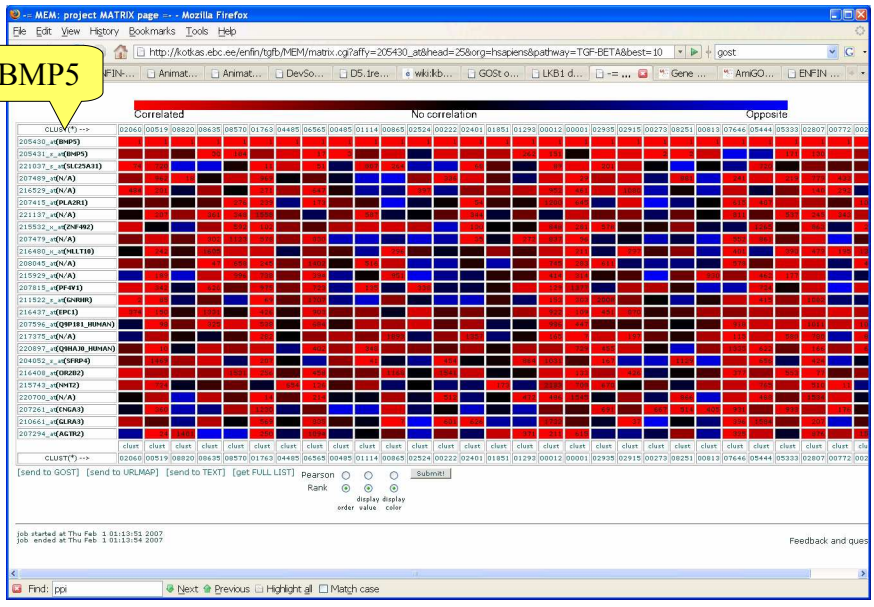

FunGenES_time_series
CELL CYCLE
Growth factor
Growth factor withdrawal
DNA damage checkpoint
ARF
ATR
Apoptosis
MAPK signaling pathway
p16,15,18,19
p27,57
p21
Mpd
Cohesin
Separin
Securin
Ubiquitin mediated proteolysis
R-point (START)
DNA
DNA O
DNA biosynthesis
S-phase proteins
Evi2
MEH
ORC (Origin Recognition Complex)
MCM (Mini-Chromosome Maintenance) complex
04110 11/11/05
G1
S
G2
M
2
1
0
-1
-2
2
4
6
8
10
106 genes match in the pathway
297 Affymetrix probes match in the pathway
Dataset maximum deviation = 6.727
Deviation 0.9975 quantile = 2.403
Pathway maximum deviation = 2.771

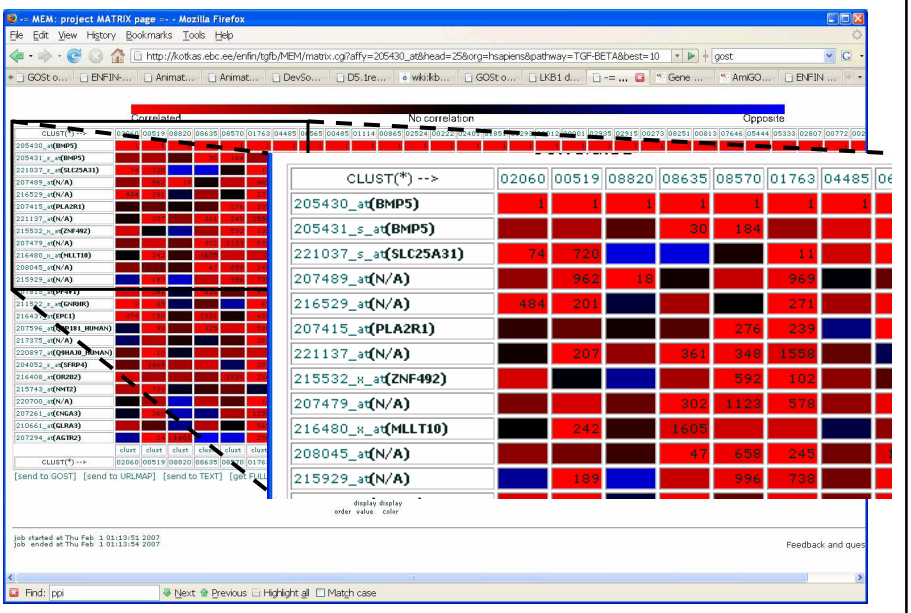

CELL CYCLE
Day 0
Day 4
Day 10
APOPTOSIS
Day 0
Day 4
Day 10

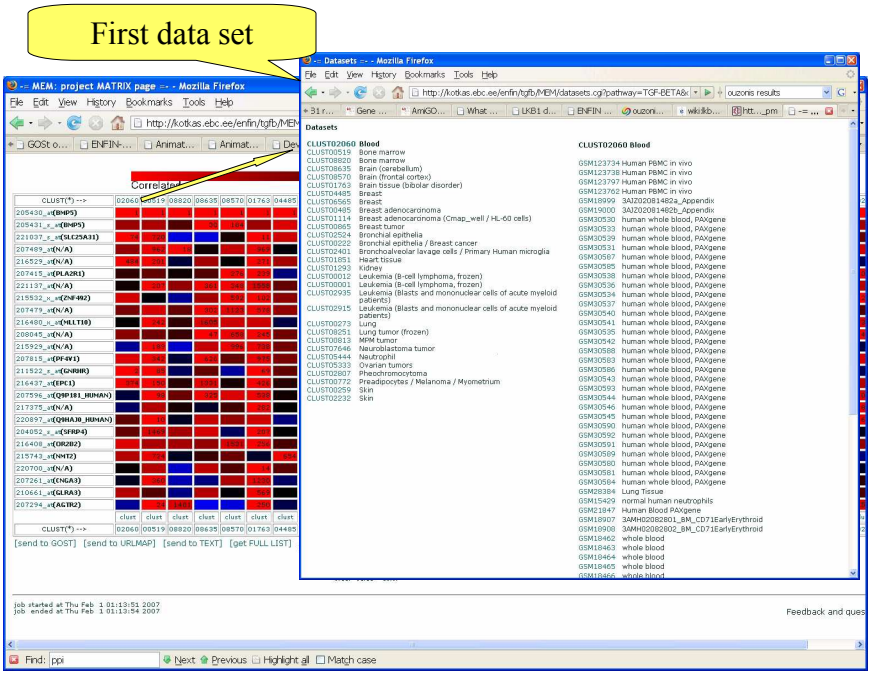
More large-scale experimental data

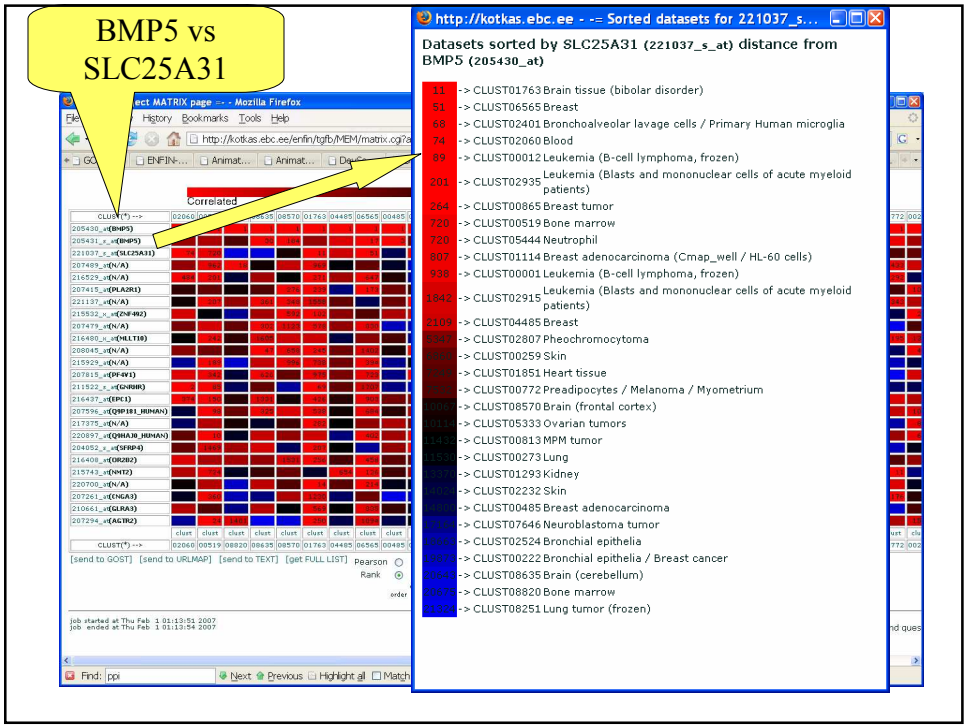
- Gene expression – thousands of experiments characterising 20K+ genes
- Protein-Protein interactions (complexes, signal transduction, ...)
- Protein-DNA interactions
- proteomics, metabolomics, genetics, ...

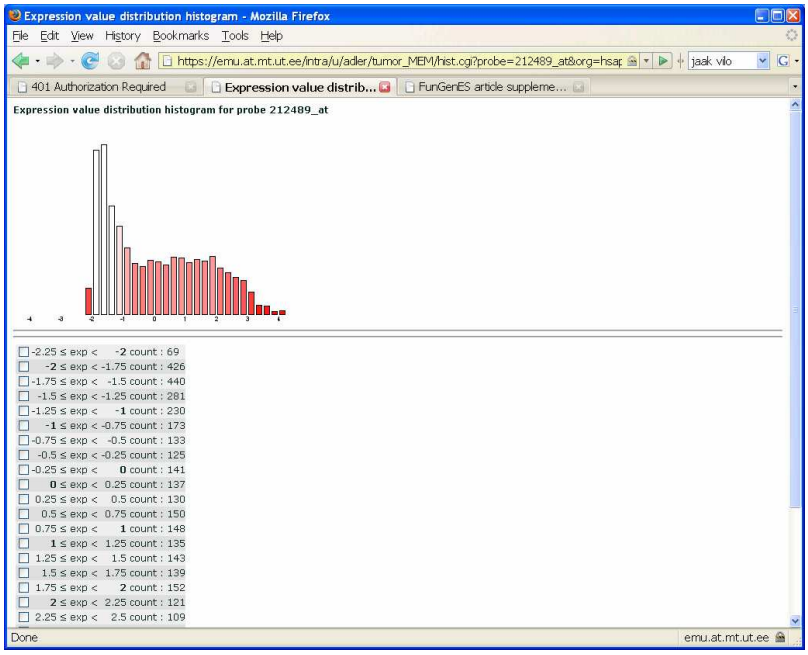
MEM – Multiple Experiment Matrix
Co-expression of TGFβ gene(s) across multiple datasets




-- MEM: project MATRIX page -- - Mozilla Firefox
Correlated
No correlation
Opposite
CLUST(*) -->
02060 00519 08820 08635 08570 01763 04485
205430_at(BMP5) 1 1 1 1 1 1 1
205431_s_at(BMP5) 30 184
221037_s_at(SLC25A31) 74 720 11
207489_at(N/A) 962 18 969
216529_at(N/A) 484 201 271
207415_at(PLA2R1) 276 239
221137_at(N/A) 207 361 348 1558
215532_x_at(ZNF492) 592 102
207479_at(N/A) 302 1123 578
216480_x_at(MLLT10) 242 1605
208045_at(N/A) 47 658 245
215929_at(N/A) 189 996 738
[send to GOST] [send to URLMAP] [send to TEXT] [get FULL LIST]
job started at Thu Feb 1 01:13:51 2007
job ended at Thu Feb 1 01:13:54 2007
Feedback and ques
Find: ppi
Next Previous Highlight all Match case


First data set
-- MEM: project MATRIX page -- - Mozilla Firefox
-- Datasets -- - Mozilla Firefox
Datasets
CLUST02060 Blood
CLUST02060 Blood
Correlated
[send to GOST] [send to URLMAP] [send to TEXT] [get FULL LIST]
job started at Thu Feb 1 01:13:51 2007
job ended at Thu Feb 1 01:13:54 2007
Feedback and ques
Find: ppi
Next Previous Highlight all Match case

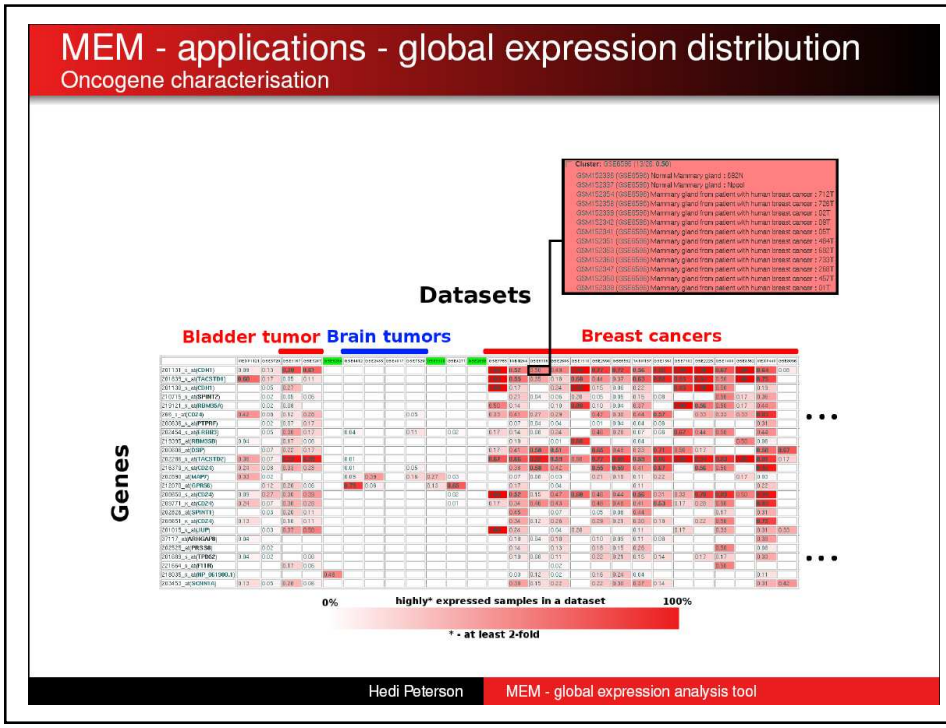

BMP5 vs SLC25A31
http://kotkas.ebc.ee - -= Sorted datasets for 221037_s...
Datasets sorted by SLC25A31 (221037_s_at) distance from BMP5 (205430_at)
11 -> CLUST01763 Brain tissue (bibolar disorder)
51 -> CLUST06565 Breast
68 -> CLUST02401 Bronchoalveolar lavage cells / Primary Human microglia
74 -> CLUST02060 Blood
89 -> CLUST00012 Leukemia (B-cell lymphoma, frozen)
201 -> CLUST02935 Leukemia (Blasts and mononuclear cells of acute myeloid patients)
264 -> CLUST00865 Breast tumor
720 -> CLUST00519 Bone marrow
720 -> CLUST05444 Neutrophil
807 -> CLUST01114 Breast adenocarcinoma (Cmap_well / HL-60 cells)
938 -> CLUST00001 Leukemia (B-cell lymphoma, frozen)
1842 -> CLUST02915 Leukemia (Blasts and mononuclear cells of acute myeloid patients)
2109 -> CLUST04485 Breast
5347 -> CLUST02807 Pheochromocytoma
-> CLUST00259 Skin
-> CLUST01851 Heart tissue
-> CLUST00772 Preadipocytes / Melanoma / Myometrium
-> CLUST08570 Brain (frontal cortex)
-> CLUST05333 Ovarian tumors
-> CLUST00813 MPM tumor
-> CLUST00273 Lung
-> CLUST01293 Kidney
-> CLUST02232 Skin
-> CLUST00485 Breast adenocarcinoma
-> CLUST07646 Neuroblastoma tumor
-> CLUST02524 Bronchial epithelia
-> CLUST00222 Bronchial epithelia / Breast cancer
-> CLUST08635 Brain (cerebellum)
-> CLUST08820 Bone marrow
-> CLUST08251 Lung tumor (frozen)


Expression value distribution histogram - Mozilla Firefox
Expression value distribution histogram for probe 212489_at
-2.25 ≤ exp < -2 count : 69
-2 ≤ exp < -1.75 count : 426
-1.75 ≤ exp < -1.5 count : 440
-1.5 ≤ exp < -1.25 count : 281
-1.25 ≤ exp < -1 count : 230
-1 ≤ exp < -0.75 count : 173
-0.75 ≤ exp < -0.5 count : 133
-0.5 ≤ exp < -0.25 count : 125
-0.25 ≤ exp < 0 count : 141
0 ≤ exp < 0.25 count : 137
0.25 ≤ exp < 0.5 count : 130
0.5 ≤ exp < 0.75 count : 150
0.75 ≤ exp < 1 count : 148
1 ≤ exp < 1.25 count : 135
1.25 ≤ exp < 1.5 count : 143
1.5 ≤ exp < 1.75 count : 139
1.75 ≤ exp < 2 count : 152
2 ≤ exp < 2.25 count : 121
2.25 ≤ exp < 2.5 count : 109
Done
emu.at.mt.ut.ee

MEM goals

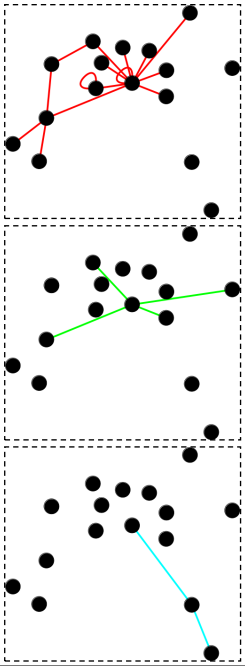
- Query across hundreds of datasets simultaneously
- Characterise gene expression across all datasets
- Study in detail all genes within one pathway
- Find new genes possibly linked to the given pathway
- Identify relevant “conditions” for gene activity

- Priit Adler; Raivo Kolde; Hedi Peterson



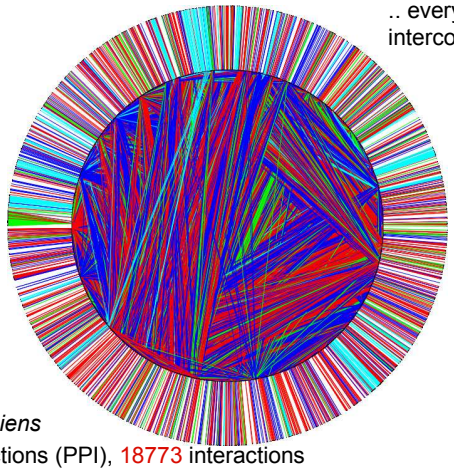
GraphWeb: mining biological networks for submodules with functional significance

- Genes as nodes
- *-omics* define edges
  - expression correlation
  - protein-protein interactions
  - literature co-occurrence
  - regulation
  - binding site discovery



Data as graphs

.. everything is interconnected

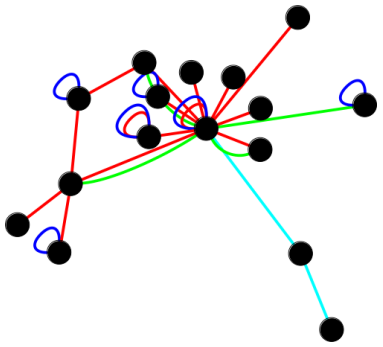


Public datasets for *H.sapiens*

- IntAct: Protein interactions (PPI), 18773 interactions
- IntAct: PPI via orthologs from IntAct, 6705 interactions
- MEM: gene expression similarity over 89 tumor datasets, 46286 interactions
- Transfac: gene regulation data, 5183 interactions

Gene modules

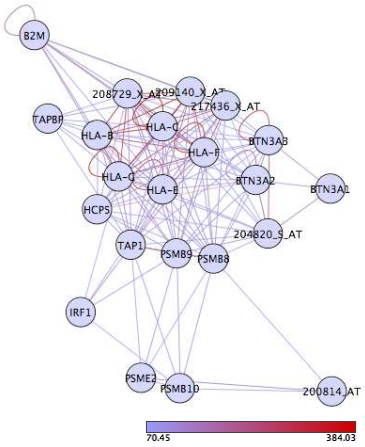
- Integrate data sources as graph *layers*
- Find well-connected subgraphs
- Combine evidence to infer knowledge about regulation and function



GO: cell cycle, regulation, growth.
KEGG: Alzheimer's disease

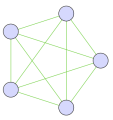
Weighing the evidence I

- Edges are not born equal
  - e.g. stronger vs weaker correlation
- Assign *local* weights to rank edges within a layer
- Look for *heavy* subgraphs

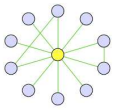


Expression similarity search across 89 human tumor-related datasets *(*MEM, *Adler et al, in prep.)*
GO: immune system, proteasome. Reactome: cell cycle, DNA replication, HIV infection

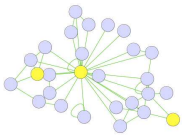
Finding the modules



- Cliques
  - Fully connected graphs ~ protein complexes

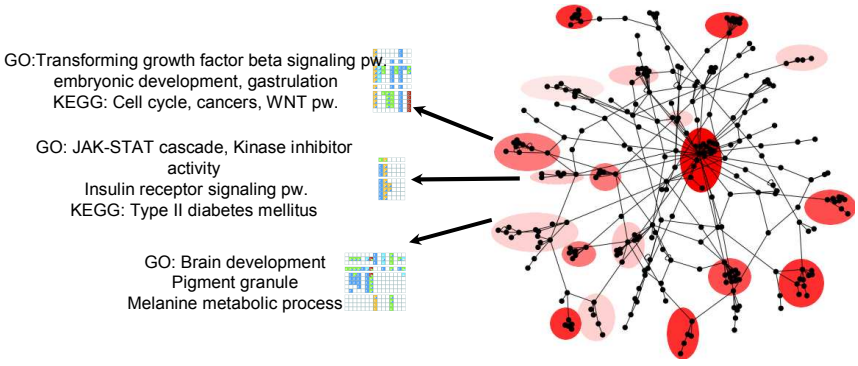


- Hubs
  - Highly connected nodes ~ transcriptional regulators

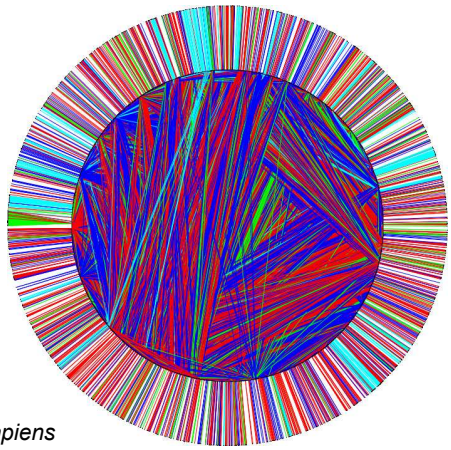


- Sets of neighbors
  - Specific genes of interest + near neighbors
- Graph clustering
  - MCL: Markov clustering *(van Dongen, 2000)*, betweenness centrality clustering

Module evaluation



Finding the modules

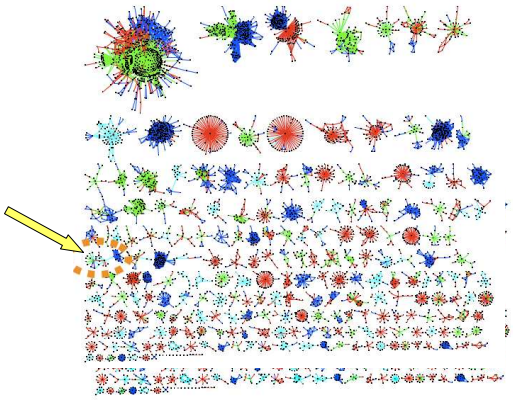


Public datasets for *H.sapiens*

- IntAct: Protein interactions (PPI), 18773 interactions
- IntAct: PPI via orthologs from IntAct, 6705 interactions
- MEM: gene expression similarity over 89 tumor datasets, 46286 interactions
- Transfac: gene regulation data, 5183 interactions

Jüri Reimand: GraphWeb. Genome Informatics, CSHL. Nov 1 2007

Finding the modules



Public datasets for *H.sapiens*

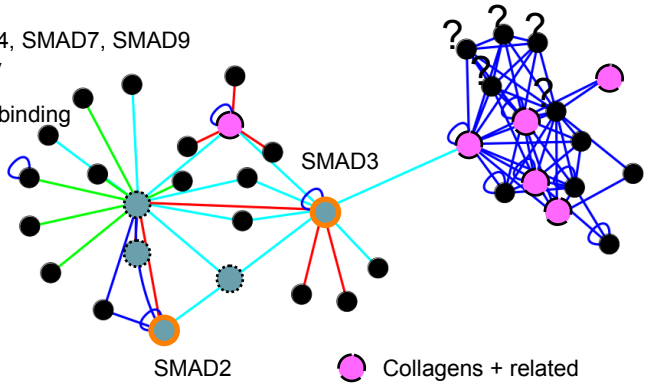
- IntAct: Protein interactions (PPI), 18773 interactions
- IntAct: PPI via orthologs from IntAct, 6705 interactions
- MEM: gene expression similarity over 89 tumor datasets, 46286 interactions
- Transfac: gene regulation data, 5183 interactions

Example

SMAD2, SMAD3, SMAD4, SMAD7, SMAD9 transcription factor family

SMAD2, SMAD3?: DNA binding

Transforming Growth Factor Beta signalling, Development, Cancers

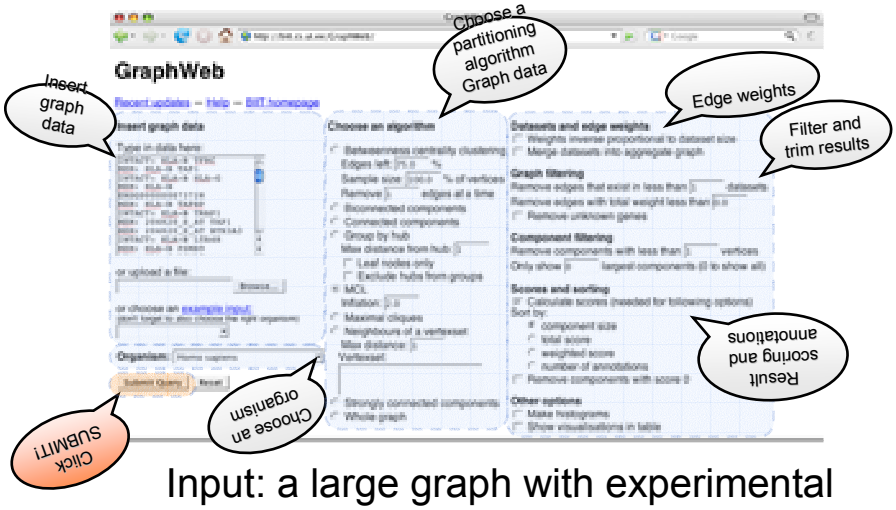


Public datasets for *H.sapiens*

- IntAct: Protein interactions (PPI), 18773 interactions
- IntAct: PPI via orthologs from IntAct, 6705 interactions
- MEM: gene expression similarity over 89 tumor datasets, 46286 interactions
- Transfac: gene regulation data, 5183 interactions

GraphWeb

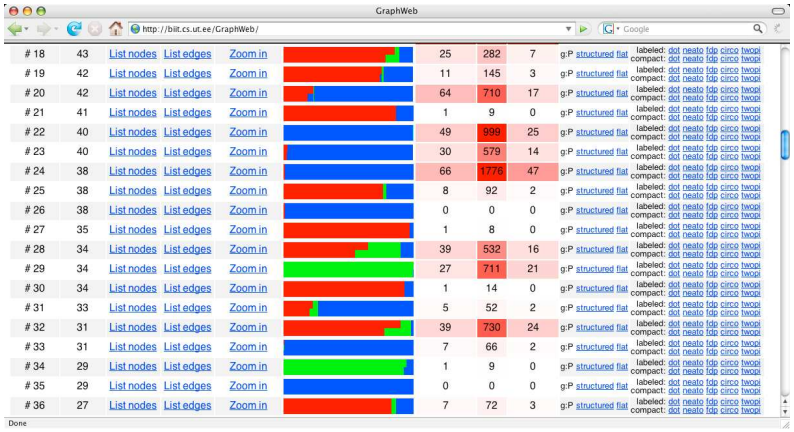
http://biit.cs.ut.ee/graphweb



Input: a large graph with experimental data

GraphWeb

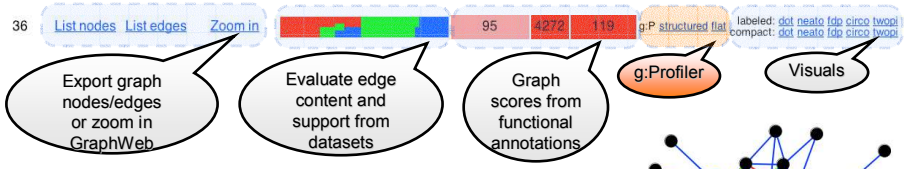
http://biit.cs.ut.ee/graphweb



Output: a list of tightly connected gene modules

GraphWeb

http://biit.cs.ut.ee/graphweb



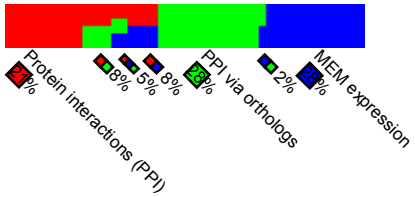
Export graph nodes/edges or zoom in GraphWeb

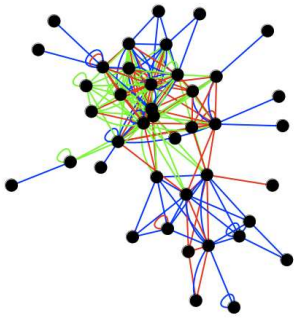
Evaluate edge content and support from datasets

Graph scores from functional annotations

g:Profiler

Visuals





GO: catabolic process, proteasome
KEGG: DNA replication, HIV infection, Cell cycle checkpoints

microRNA - discovery

- Alexander Stark, Pouya Kheradpour, **Leopold Parts**, Julius Brennecke, Emily Hodges,Gregory J. Hannon, and Manolis Kellis. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Research*, 2007
- Alexander Stark, Michael F. Lin, Pouya Kheradpour, Jakob S. Pedersen, **Leopold Parts**, et. al. Discovery of functional elements in 12 fly genomes using evolutionary signatures. *Nature* **450**, 219-232 (8 November 2007)
- Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* , 2007

Next 5 years

- Databases and analysis of human health
  - Colon & Breast cancer; disease relapse
    - mRNA, proteomics, metabolomics data
  - Estonian Genome Project
    - DB of 10's of thousands of patients with ~1000 attributes worth of health and lifestyle profiles
  - ES cell based toxicology profiling for new drug candidates
  - Disease/gene associations

Next 5 years

- Data query and visualisation methods
  - MEM, PATMATCH, ...
- Data mining
  - Motif discovery (SPEXS, Trie*Tools, ...)
  - Gene regulation modeling
  - Phenotype, Genotype, Expression, PPI, ChIP-chip, ... data joint analysis
- Data analysis environments
- Analysis of data

Anno 2007 (BIIT and Quretec)

Institute of Computer Science

- ☆ Software Engineering (Prof. Marlon Dumas)
- High-Performance (distributed) computing (Prof. Eero Vainikko)
- Language Technology (Prof. Mare Koit)
- ☆ Programming Languages (Prof. Varmo Vene)
- Cryptography (Prof. Ahto Buldas)
- ☆ Bioinformatics and DM (Prof. Jaak Vilo)

- 2 Target funding projects (HPC and LT)

- ☆ new appointments in Nov & Dec 2007