Υπολογιστικές προσεγγίσεις για την ανακάλυψη και παραγωγή γνώσης από ετερογενείς πηγές: Μεθοδολογία και Εφαρμογή σε βάσεις Βιολογικών και Μοριακών Δεδομένων

Λευτέρης Κουμάκης

Υποψήφιος Διδάκτωρ ΜΠΔ

Επιβλέπων Καθηγητής: Βασίλειος Μουστάκης

Μέλος Τριμελούς: Μιχαήλ Ζερβάκης (Καθηγητής)

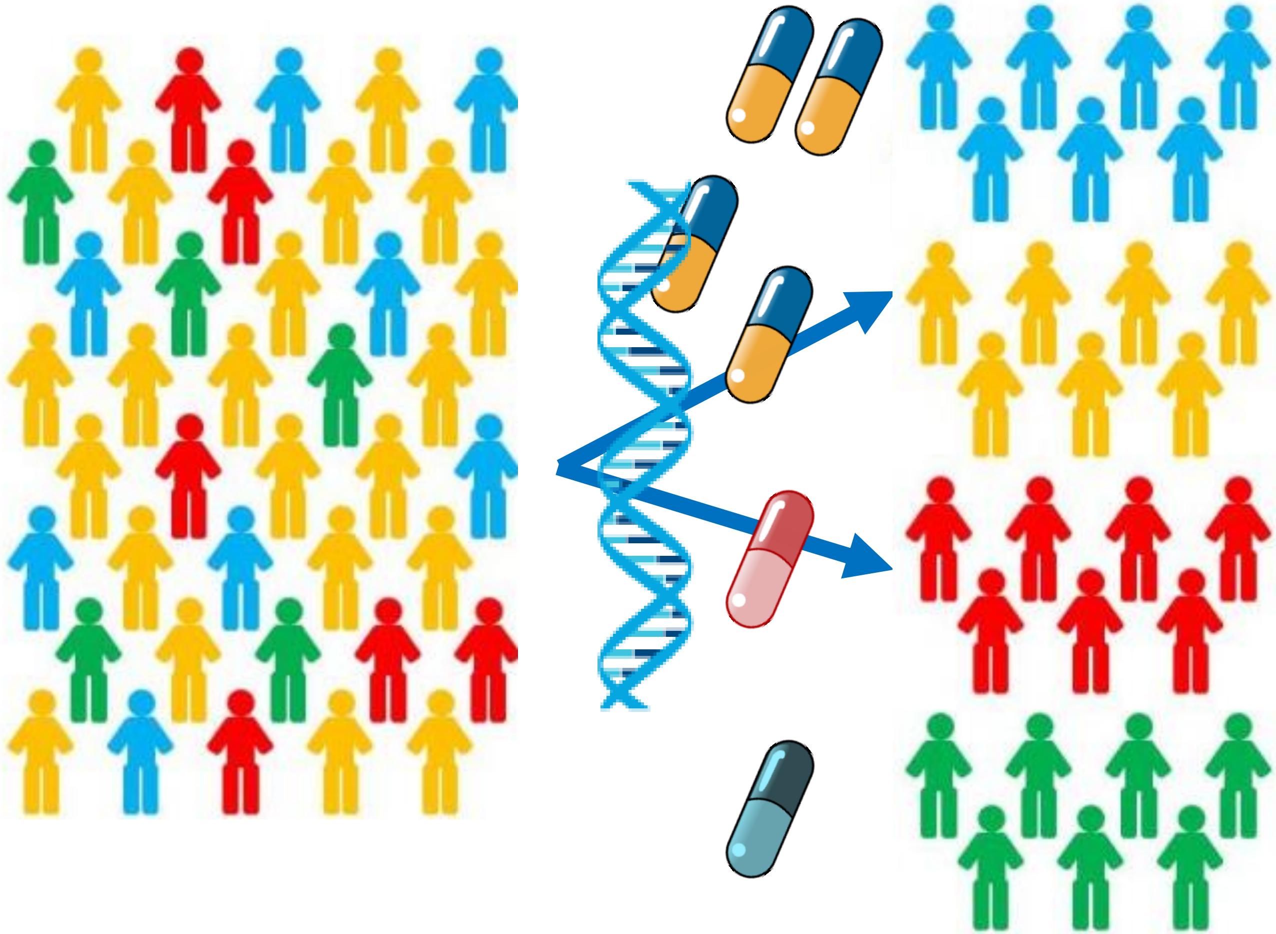
Μέλος Τριμελούς: Γεώργιος Ποταμιάς (Ερευνητής Β)

Outline

- Introduction
  - Microarrays and Gene Regulatory Networks
  - Problem definition
- Methodology
  - MinePath algorithm
  - Web based implementation (www.minepath.org)
- Experiments
  - Comparison study
  - Biological Validation
  - Discovery of new knowledge
  - miRNAs
- Conclusions

Introduction

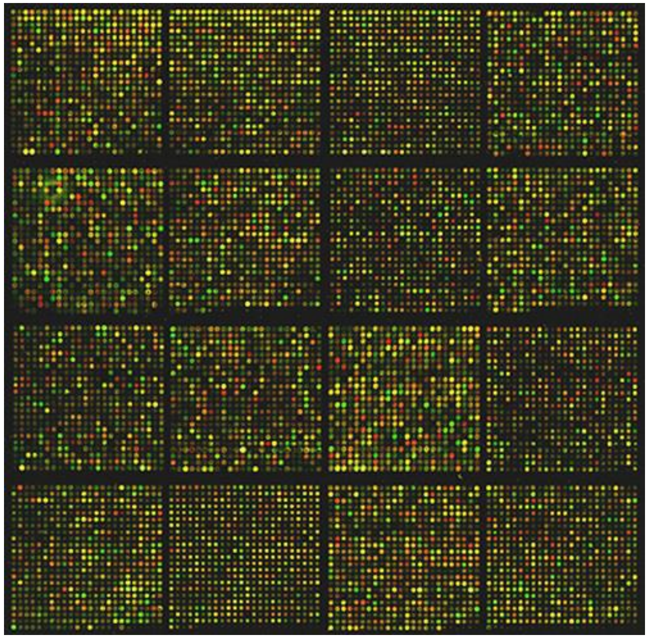
Personalized Medicine



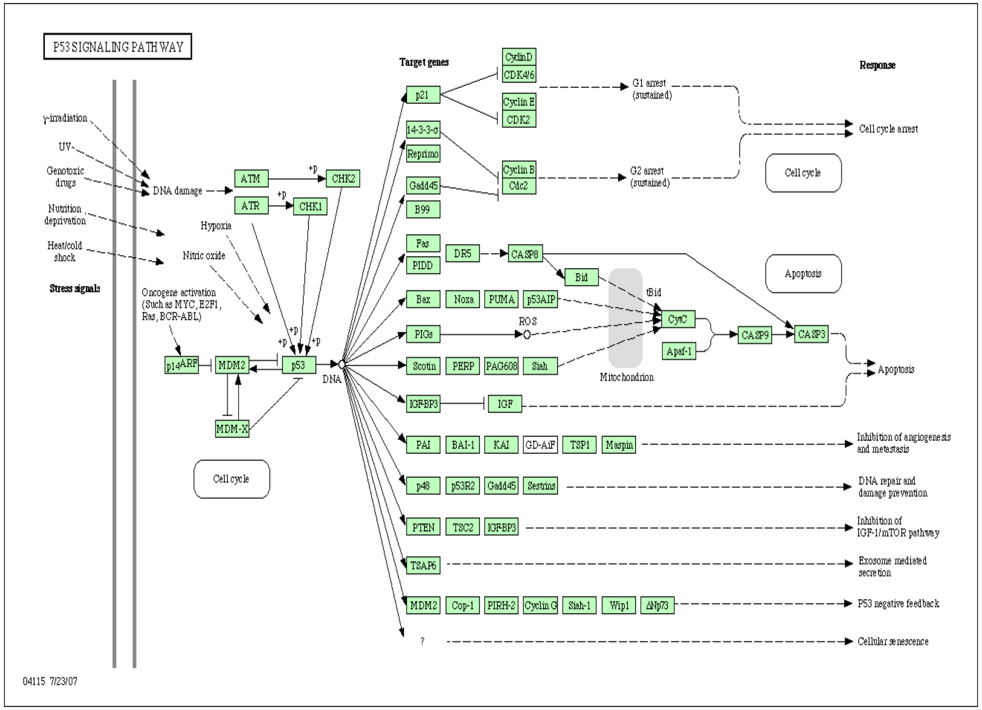
Genomic data sources

The two of the most important:

Microarray gene-expression experiments

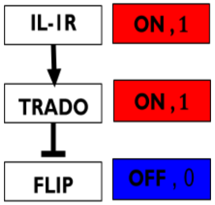


Molecular pathways and gene regulatory networks (GRNs)



Problem definition

Genes



cases

	POS			NEG	
	case1	case2	case3	case4	case5
IL-IR	ON	ON	ON	ON	ON
TRADO	ON	ON	ON	OFF	ON
FLIP	OFF	OFF	OFF	OFF	ON
MyD88	ON	ON	ON	ON	ON
NIK	ON	OFF	OFF	ON	OFF

Sub-Paths

cases

	POS			NEG	
	case1	case2	case3	case4	case5
IL-IR→TRADO	ON	ON	ON	OFF	ON
✓ IL-IR→TRADO--|FLIP	ON	ON	ON	OFF	OFF
IL-IR→MyD88	ON	ON	ON	ON	ON
IL-IR→MyD88→NIK	ON	OFF	OFF	ON	OFF

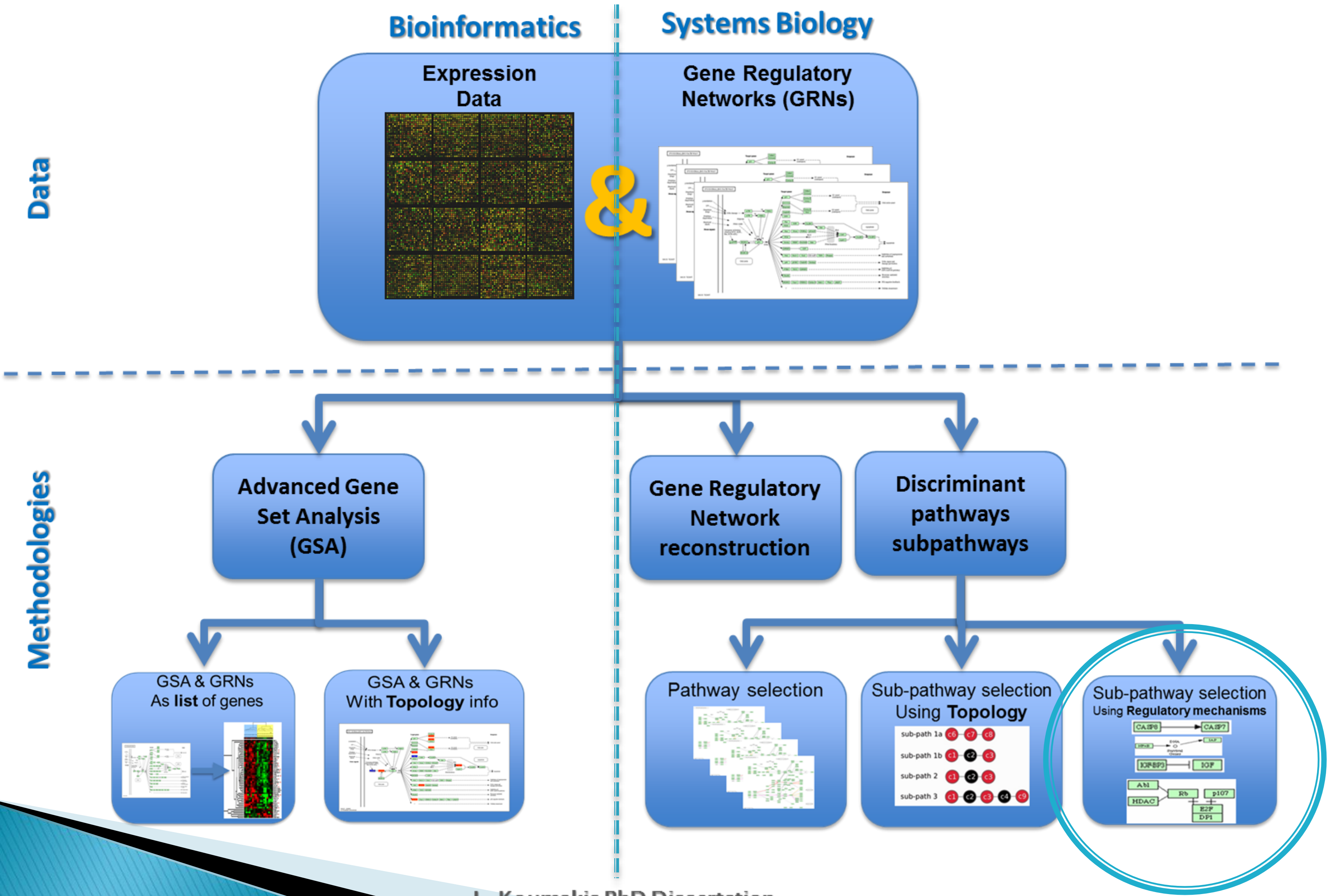
Initial expectation: microarrays would reveal specific gene signatures for various phenotypes

But it seems to be bounded to a number of limitations mainly because of the complexity and the individual variations and heterogeneities associated with the induced gene-signatures

Barabási et al*: “*Given the functional interdependencies between the molecular components in a human cell, **a disease is rarely a consequence of an abnormality in a single gene**, but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems.*”

* Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease." *Nature Reviews Genetics* 12, no. 1 (2011): 56-68.

Microarrays and GRNs



Bioinformatics

Systems Biology

Data

Expression Data

&

Gene Regulatory Networks (GRNs)

Methodologies

Advanced Gene Set Analysis (GSA)

Gene Regulatory Network reconstruction

Discriminant pathways subpathways

GSA & GRNs As **list** of genes

GSA & GRNs With **Topology** info

Pathway selection

Sub-pathway selection Using **Topology**

Sub-pathway selection Using **Regulatory mechanisms**

MA & GRNs methodologies

	Advanced Gene Set Analysis												
	Siu et al [27]	Wang et al [28]	Braun et al [29]	Tai et al [30]	Sfakianakis et al [31]	Beltrame et al [32]	KEGG color mapper	Genoscape [34]	PiNGO [35]	Cline et al [36]	DDN [37]	Ibrahim et al [38]	TopoGSA [39]
Use of microarray data	√	√	√	√	√	√	X	√	√	√	√	√	√
Use GRNs	√	√	√	√	√	√	√	√	√	√	√	√	√
Use pathway genes	√	√	√	√	√	√	X	√	√	√	√	√	√
Use sub-paths	X	X	X	X	X	X	X	X	X	X	X	X	X
Use topology	X	X	X	X	X	X	√	√	√	√	√	√	√
Use regulatory mechanisms	X	X	X	X	X	X	X	X	X	X	X	X	X
Identify discriminant genes	√	√	√	√	√	√	X	√	√	√	X	√	√
Identify discriminant pathways	X	X	√	X	X	X	X	X	X	√	√	√	√
Identify discriminant sub-paths	X	X	X	X	X	X	X	X	X	X	X	X	X
Web based	X	X	X	X	X	X	√	X	X	X	X	X	X
Visualization support	X	X	X	X	X	X	√	√	√	√	√	X	X

	Discriminant pathways & sub-paths																				
	Draghici et al [40]	Oncomine [41]	Eu.Gene [42]	Adewale et al [43]	Ma et al [44]	PathBLAST [45]	GeneMANIA [46]	Nacu et al [48]	Chen et al [49]	DEGAS [51]	KeyPathwayMiner [52]	Ideker et al. [54]	Wu and Stein [56]	CLiPPER algorithm [57]	Kazmi et al [58]	SubpathwayMiner [59]	Graphite Web [61]	GGEA [16]	SPIA [60]	TEAK [15]	PATHOME [13]
Use of microarray data	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Use GRNs	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Use pathway genes	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Use sub-paths	X	X	X	X	X	X	X	X	√	√	√	√	√	√	√	√	√	√	√	√	√
Use topology	X	X	X	X	X	X	X	X	√	√	√	√	√	√	√	√	√	√	√	√	√
Use regulatory mechanisms	X	X	X	X	X	X	X	X	X	X	X	X	X	X	√*	X	√**	√	√	√	√
Identify discriminant genes	X	X	X	X	X	X	√	√	X	X	X	X	X	X	X	X	√	X	X	√	√
Identify discriminant pathways	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Identify discriminant sub-paths	X	X	X	X	X	X	X	X	√	√	√	√	√	√	√	X	X	X	X	√	√
Web based	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	√	√	X	X	X	X
Visualization support	X	X	X	X	X	√	√	X	X	X	√	X	X	√	√	√	√	X	X	√	X

** takes advantage only of the activations between genes*
*** A web server which uses SPIA*

Advanced Gene Set Analysis

- All neglect the regulatory mechanisms of GRNs
- None can identify discriminant sub-paths
- Limited support of visualization features
- only one supports web based interface

Discriminant pathways &sub-paths

- Five methods can handle effectively the regulatory mechanisms
- Two out of them can identify discriminant sub-paths in GRNs

Most of the methodologies lack of visualization features and support for web based platform

MinePath Methodology

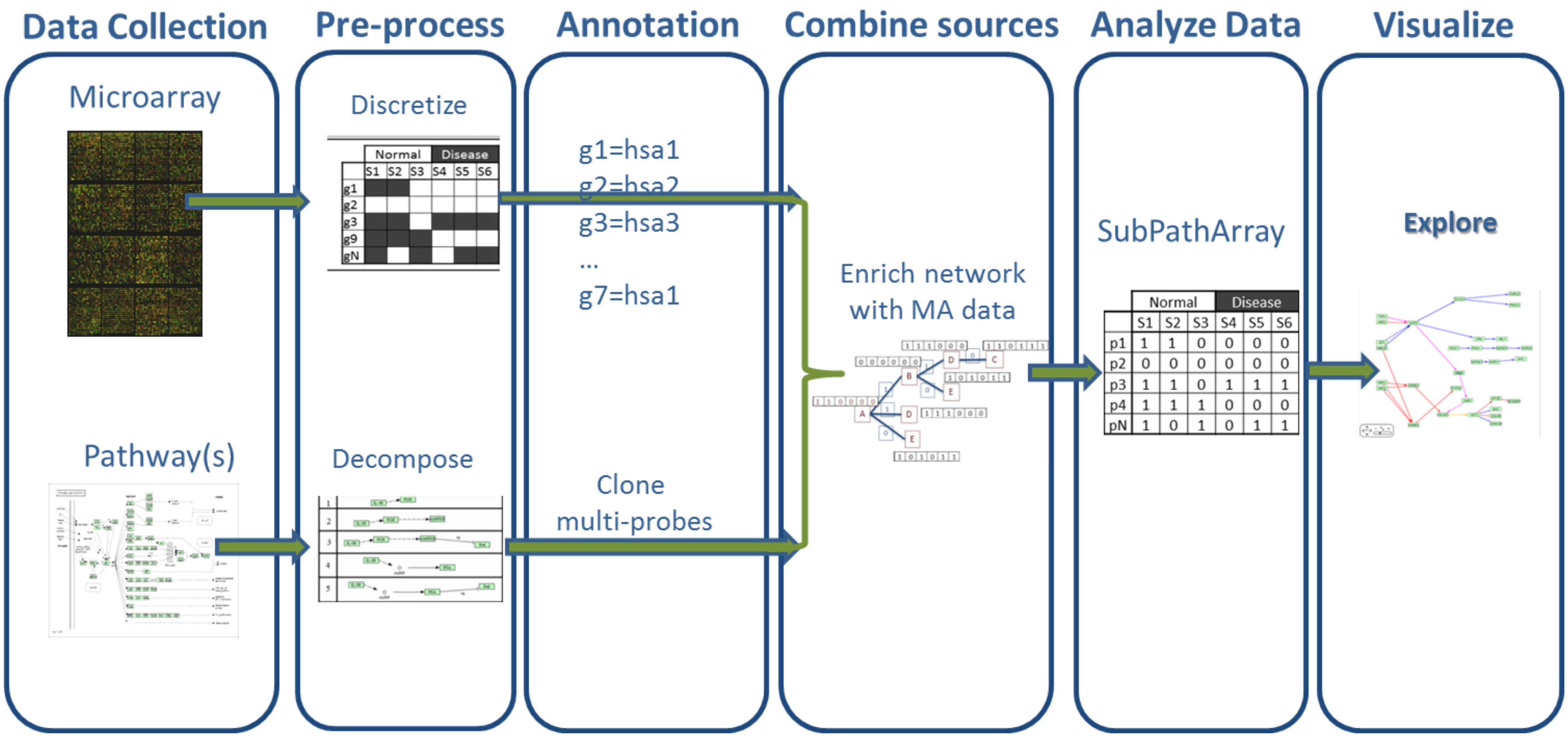
MinePath approach

MinePath introduces a new methodology for the identification of differentially expressed functional paths or sub-paths within a gene regulatory network (GRN) using microarray data analysis.

Innovative features & benefits:

- MinePath takes advantage of the regulatory mechanisms in a GRN such as the direction and the type of interaction (activation/inhibition) between genes for each sub-pathway.
- Contrary to similar efforts which visualize the state of genes on a pathway, MinePath identifies and visualizes differentially expressed regulatory mechanisms and sub-pathways of GRNs.
- MinePath is a web based application (no setup is needed) which can compute, identify and visualize differentially expressed paths from your expression data within seconds

MinePath flow of operations



Data Collection

Microarray

Pathway(s)

Pre-process

Discretize

	Normal			Disease		
	S1	S2	S3	S4	S5	S6
g1	■	■	□	□	□	□
g2	□	□	□	□	□	□
g3	■	■	□	■	■	■
g9	■	■	■	□	□	□
gN	■	□	■	□	■	■

Decompose

Annotation

g1=hsa1
g2=hsa2
g3=hsa3
…
g7=hsa1

Clone
multi-probes

Combine sources

Enrich network
with MA data

Analyze Data

SubPathArray

	Normal			Disease		
	S1	S2	S3	S4	S5	S6
p1	1	1	0	0	0	0
p2	0	0	0	0	0	0
p3	1	1	0	1	1	1
p4	1	1	1	0	0	0
pN	1	0	1	0	1	1

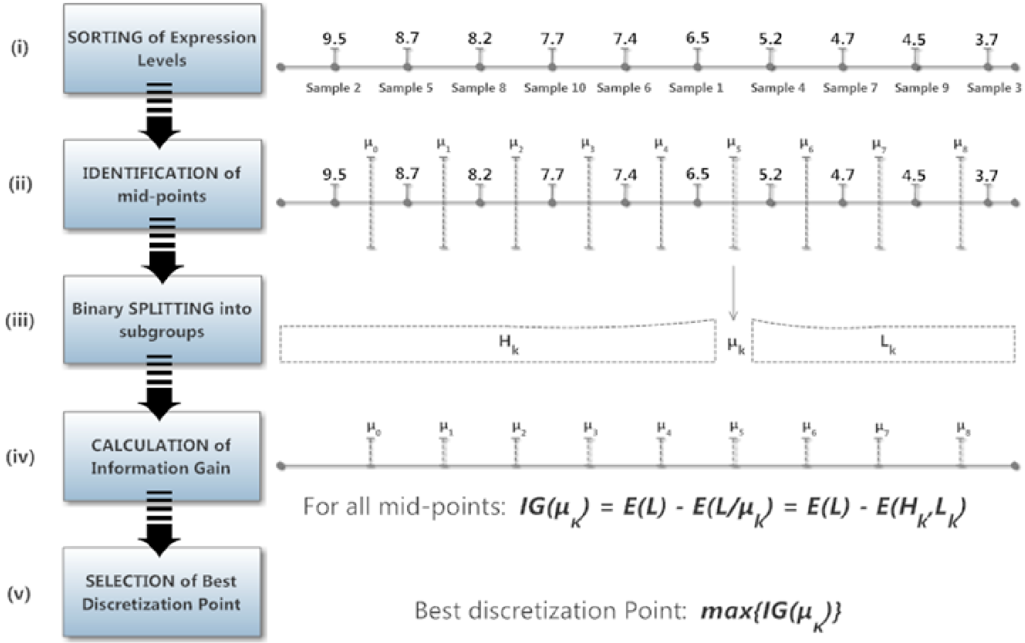
Visualize

Explore

Pre–processing (Microarrays)

Based on *Information Gain & Entropy**

- 0 indicates a non-expressed or under-expressed gene
- 1 indicates over-expressed gene



SORTING of Expression Levels

IDENTIFICATION of mid-points

Binary SPLITTING into subgroups

CALCULATION of Information Gain

SELECTION of Best Discretization Point

For all mid-points: $IG(\mu_k) = E(L) - E(L/\mu_k) = E(L) - E(H_k,L_k)$

Best discretization Point: $max\{IG(\mu_k)\}$

* *Potamias G., Koumakis L., & Moustakis V.* "Gene selection via discretized gene-expression profiles and greedy feature-elimination." In *Methods and Applications of Artificial Intelligence*, pp. 256-266. Springer Berlin Heidelberg, 2004.

genes

	Normal			Disease		
	S1	S2	S3	S4	S5	S6
A	98	78	23	43	1	9
B	34	23	3	22	11	12
C	79	66	12	80	82	67
D	89	91	77	12	43	33
E	80	20	78	12	89	99

Binary representation →

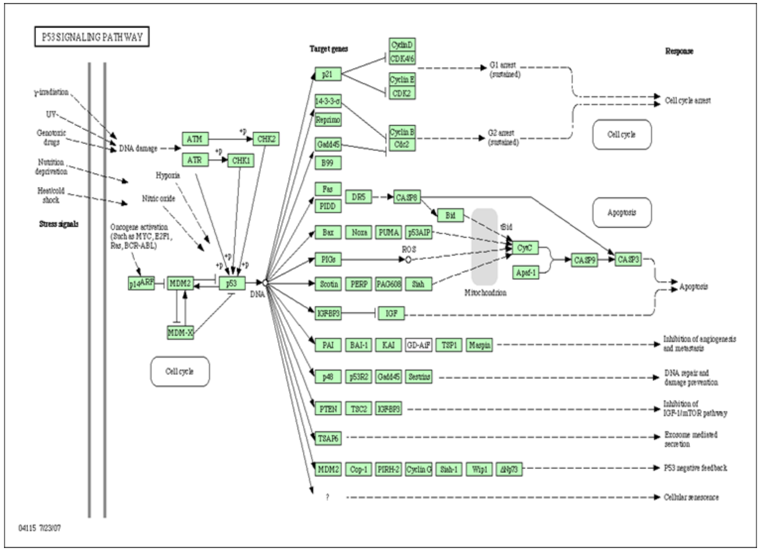
	Normal			Disease		
	S1	S2	S3	S4	S5	S6
A	1	1	0	0	0	0
B	1	1	0	1	0	0
C	0	0	0	1	1	0
D	1	1	1	0	0	0
E	0	0	0	0	1	1

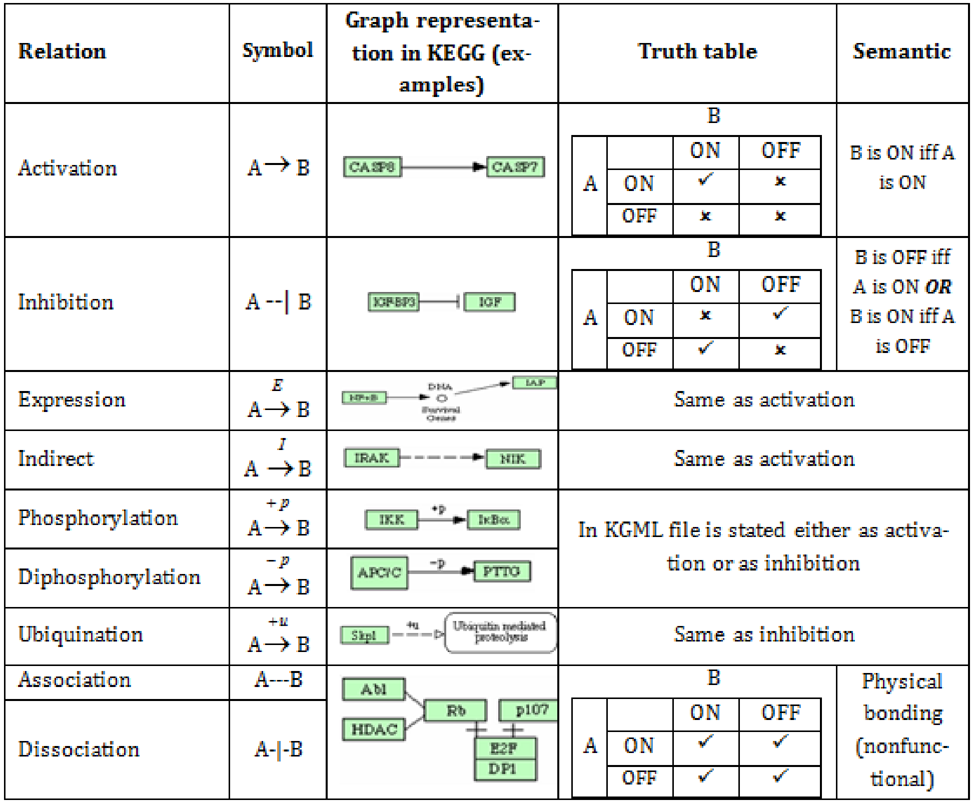
Thresholds
60.5
17
79.5
60
84.5

Pre-processing (GRNs)

GRNs are described through standard graph annotations.

- Nodes can be either genes, groups of genes, compounds or other networks.
- Edges can be one of the gene relations known from the biology theory



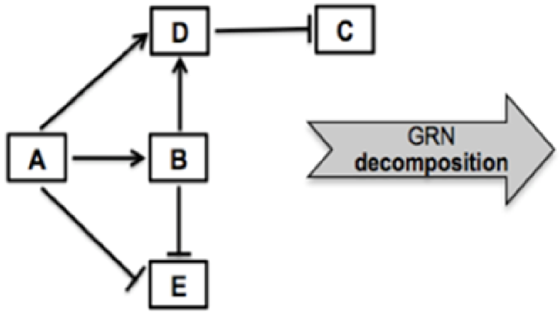


Relation	Symbol	Graph representation in KEGG (examples)	Truth table	Semantic
Activation	A→ B	CASP8 → CASP7	B ON OFF A ON ✓ ✗ OFF ✗ ✗	B is ON iff A is ON
Inhibition	A --| B	IGFBP3 —| IGF	B ON OFF A ON ✗ ✓ OFF ✓ ✗	B is OFF iff A is ON ***OR*** B is ON iff A is OFF
Expression	$A \xrightarrow{E} B$	NF-κB → DNA ○ Survival Genes → IAP	Same as activation	
Indirect	$A \xrightarrow{I} B$	IRAK - - - → NIK	Same as activation	
Phosphorylation	$A \xrightarrow{+p} B$	IKK —+p→ IκBα	In KGML file is stated either as activation or as inhibition	
Diphosphorylation	$A \xrightarrow{-p} B$	APC/C —-p→ PTTG		
Ubiquination	$A \xrightarrow{+u} B$	Skp1 - -+u- → Ubiquitin mediated proteolysis	Same as inhibition	
Association	A---B	Abl HDAC Rb p107 E2F DP1	B ON OFF A ON ✓ ✓ OFF ✓ ✓	Physical bonding (nonfunctional)
Dissociation	A-|-B			

Pre-processing (GRNs)

Sub-paths decomposition:

- KGML (KEGG XML) processing
- All possible GRN sub-paths are extracted



1 A → B → D ⊣ C

2 A → B → D

3 A → B

4 A → D ⊣ C

5 A → D

6 A → B ⊣ E

7 A ⊣ E

8 B → D ⊣ C

9 B → D

10 B ⊣ E

11 D ⊣ C

Extension (optional):

- take into account the starting and ending points of each sub-path as a new sub-path
- In our example 2 more sub-paths:
  - **A--|C**
  - **B--|C**

Data Annotation (mapping)

MinePath provides two options to cope with the one to many (probe to gene) issue:

- ***Max Probe***: selection of the value of the probe with the highest intensity out of all the probes that map to the same gene (default option).
- ***Probes clones***: produce all the possible combinations of sub-paths based on probes and not on gene ids.

Platform	Affy-U133A
Probes	22283
Annotated to KEGG	20967

Pathway	Description	Genes in U133A plat.	Sub-paths	Sub-Paths after clones
hsa04010	MAPK signaling	481	1291	21109
hsa04012	ErbB signaling	164	486	4277
hsa04020	Calcium	335	157	189
hsa04110	Cell cycle	231	161	437
hsa04115	p53 signaling	123	277	1939
hsa04150	mTOR signaling	91	65	365
hsa04210	Apoptosis	157	145	1505
hsa04310	Wnt signaling	256	277	371
hsa04350	TGF-beta signaling	140	57	79
hsa04370	VEGF signaling	129	61	187
hsa04510	Focal adhesion	404	420	1275
hsa04520	Adherens junction	179	442	10873
hsa04912	GnRH signaling	205	145	1488
hsa05200	Pathways in cancer	634	988	16014

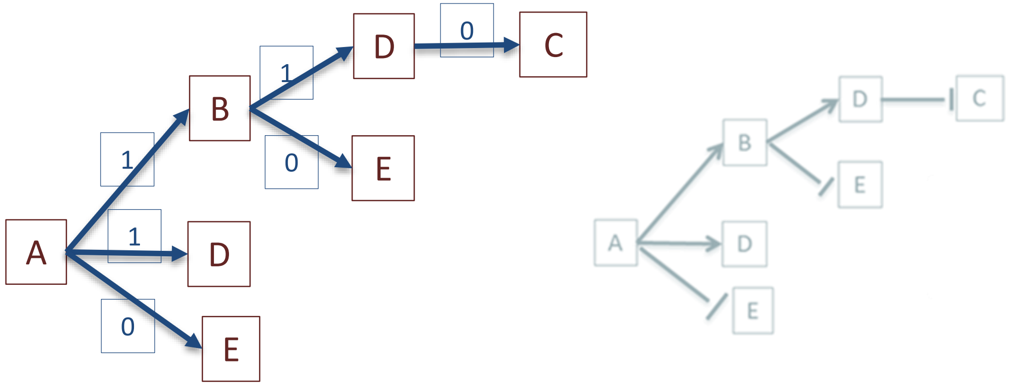
3*3*1*3*2 = 54 sub-paths in this example

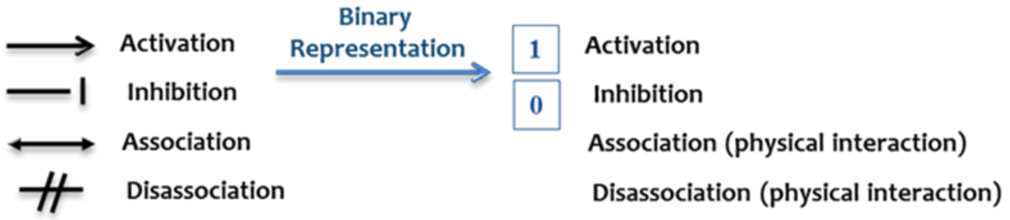
Binary representation of Data

Microarrays

	Normal			Disease		
	S1	S2	S3	S4	S5	S6
A	1	0	1	0	1	1
B	1	1	1	1	1	1
C	0	0	0	1	1	1
D	1	1	1	1	1	1
E	1	0	1	0	1	1

Gene Regulatory Networks





Activation

Inhibition

Association

Disassociation

Binary Representation

1 Activation

0 Inhibition

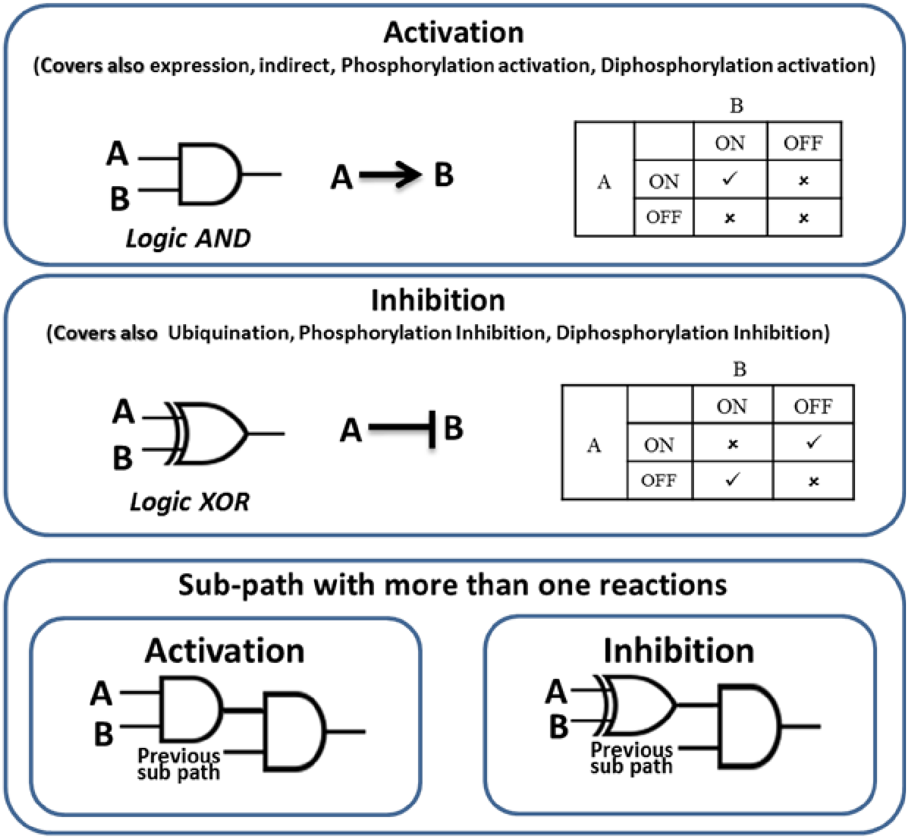
Association (physical interaction)

Disassociation (physical interaction)

Mapping gene interactions using logic gates

Sub-paths are extracted from the graph using basic **Boolean operations** for optimization

- Activation is mapped as a **logic AND**
- Inhibition as a **logic XOR**
- sub-paths with more than one reaction require the combination of previous sub-path and the last relation using a logic AND



Activation

(Covers also expression, indirect, Phosphorylation activation, Diphosphorylation activation)

		B	
		ON	OFF
A	ON	✓	×
	OFF	×	×

Inhibition

(Covers also Ubiquination, Phosphorylation Inhibition, Diphosphorylation Inhibition)

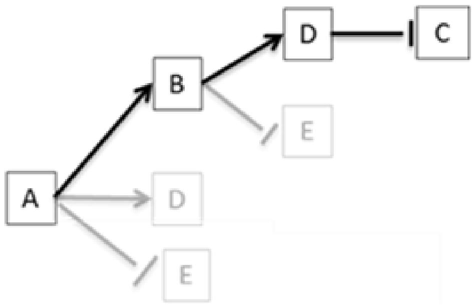
		B	
		ON	OFF
A	ON	×	✓
	OFF	✓	×

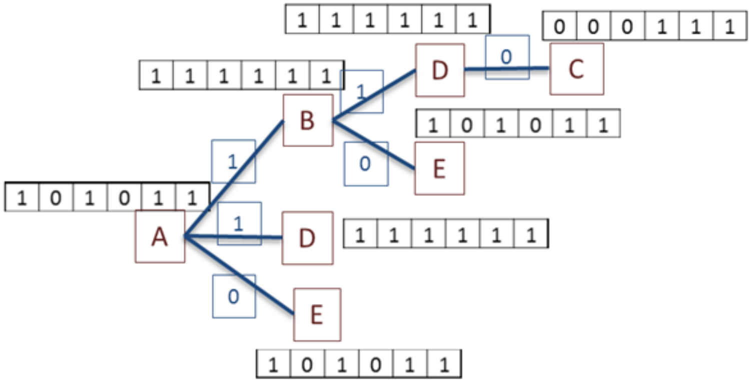
Sub-path with more than one reactions

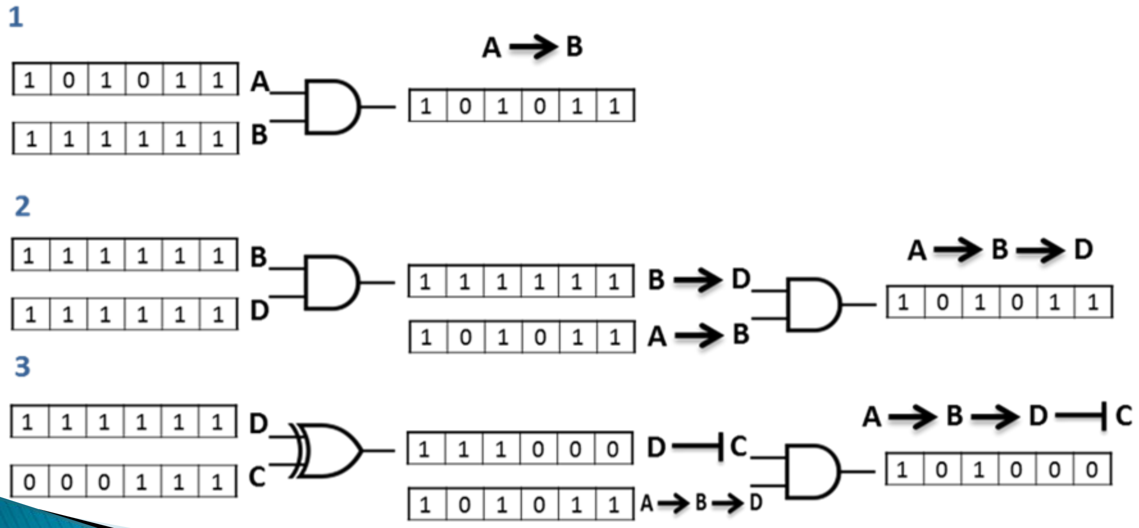
Where:

A: source Gene(s)
B: target Gene(s)

Calculating functional status of a sub-path







	Pheno-1			Pheno-2		
	S1	S2	S3	S4	S5	S6
A → B	1	0	1	0	1	1
A → B → D	1	0	1	0	1	1
A → B → D ⊣ C	1	0	1	0	0	0

The result is an array of sub-paths with binary values for every sample in the form of a discretized microarray

Analysis

MinePath produces a binary matrix containing information about the sub-paths (active or not) for the specific samples

- Transformation does not aim to reduce the dimensionality issue of microarrays
    - e.g. U133A (22.283 probes) & all hsa KEGG pathways produce more than 30.000 sub-paths
- MinePath analysis identifies:
    - The “best” or in our case the most discriminant features (sub-paths) using two different filtering/ranking methodologies:
        - the discriminant ranking
        - the polarity ranking
    - The “best” common sub-paths (sub-paths that appear to be functional for both phenotypes)

Sub-paths ranking

- Assume the two phenotypic classes ***P*** (positive), ***N*** (negative). The following quantities are computed:
  - $\boldsymbol{H_P}$ = number of ***P*** samples that the sub-path holds.
  - $\boldsymbol{L_P}$ = number of ***P*** samples that the sub-path does not hold.
  - $\boldsymbol{H_N}$ = number of ***N*** samples that the sub-path holds.
  - $\boldsymbol{L_N}$ = number of ***N*** samples that the sub-path does not hold.

Discriminant rank for each sub-path (r_{sb}): $$\boldsymbol{r_{sb} = (H_P \times L_N) - (H_N \times L_P)}$$

Polarity rank for each sub-path (r_{sb}): $$\boldsymbol{r_{sb} = \frac{(H_P - H_N)}{(H_P + H_N)}}$$

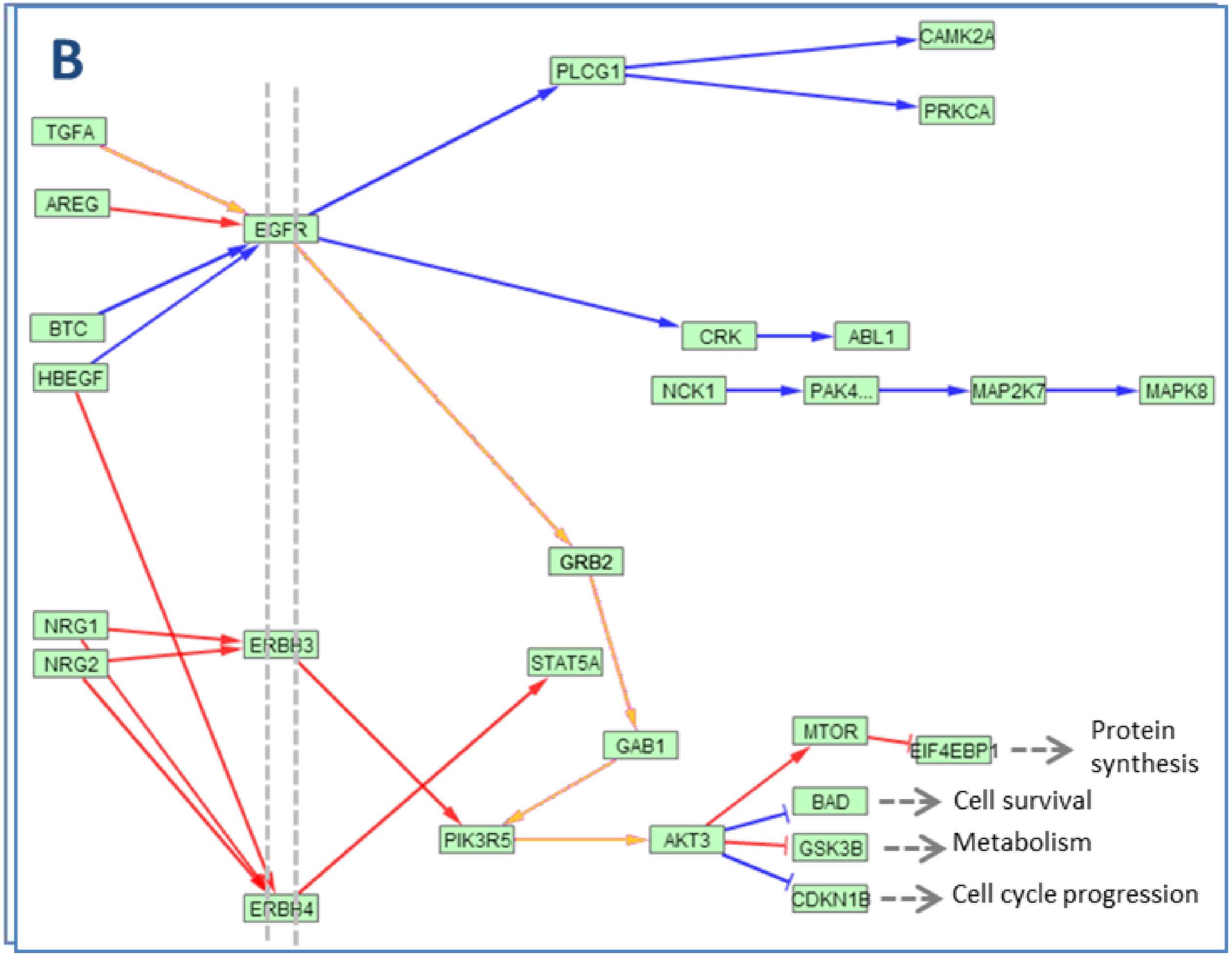
- expresses a *differentiation* characteristic
- represents the descriptive power of the sub-path per phenotypic class
- Ordering the positive ranks in descending order and the negative ranks in ascending order we may identify the most discriminant sub-path with respect to phenotypic classes P and N.

Common sub-paths

Sub-paths which are **always activated** may fill-in the gap (functional interaction) between two sub-paths and reveal a complete functional and biologically valuable route.

Colour coding:

- **Red**: sub-paths active at class 1
- **Blue**: sub-paths active at class 2
- **Orange:** sub-paths that are always active.



Validation

MinePath provides mechanisms that validate the best sub-paths against the different phenotypes using well-known algorithms and validation procedures from the area of machine learning:

- Decision tree learning (C4.5)
- Naïve Bays
- Support Vector Machines (Linear kernel)

By default MinePath computes, stores and reports 10-fold cross-validation results, but additional modelling experiments could be conducted and evaluated

- *e.g. following a train vs. independent test experimentation mode*

Implementation details

MinePath is Java based

  - More than 5500 lines of code

- Uses open source libraries:
  - Cytoscape for the handling of the graphs
  - Weka for the validation of the best sub-paths
- Provides as output:
  - the matrix (sub-paths vs samples) of the dataset
  - the best (according to the ranking) sub-paths
  - the best sub-paths that are always functional

MinePath web-server (Web 2.0 application):

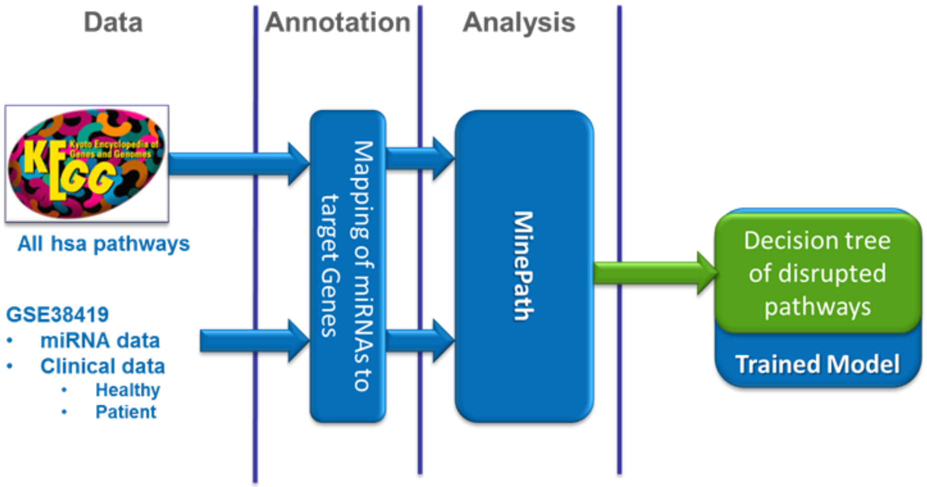
  - frontend-backend software design using AJAX calls
  - Use of Ext-JS library and pure JavaScript
  - Use of Cytoscape Web library for the visualization

miRNA extension

MinePath has been implemented to be modular and to be easily extended to support more algorithms and different clinical scenarios

- *e.g. Find disrupted pathways in nephroblastoma using miRNA expression data*.

1. Initially we collect the data
2. we identify the target genes from the miRNAs,
3. we analyse using MinePath
4. finally we train the model using the disrupted sub-paths



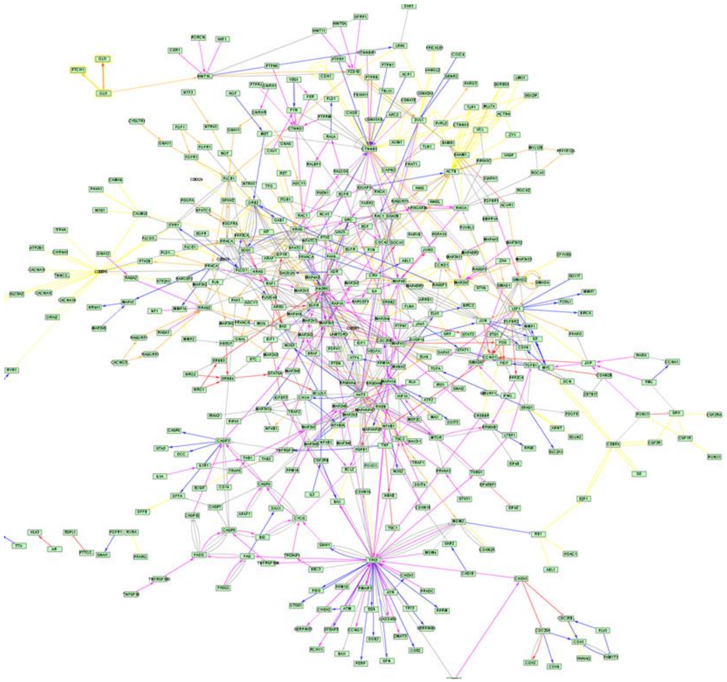
For the miRNA scenario we assume that all the KEGG pathways are fully functional.

Merging GRNs extension

This extra functionality provides the possibility to merge GRNs into one graph for further analysis.

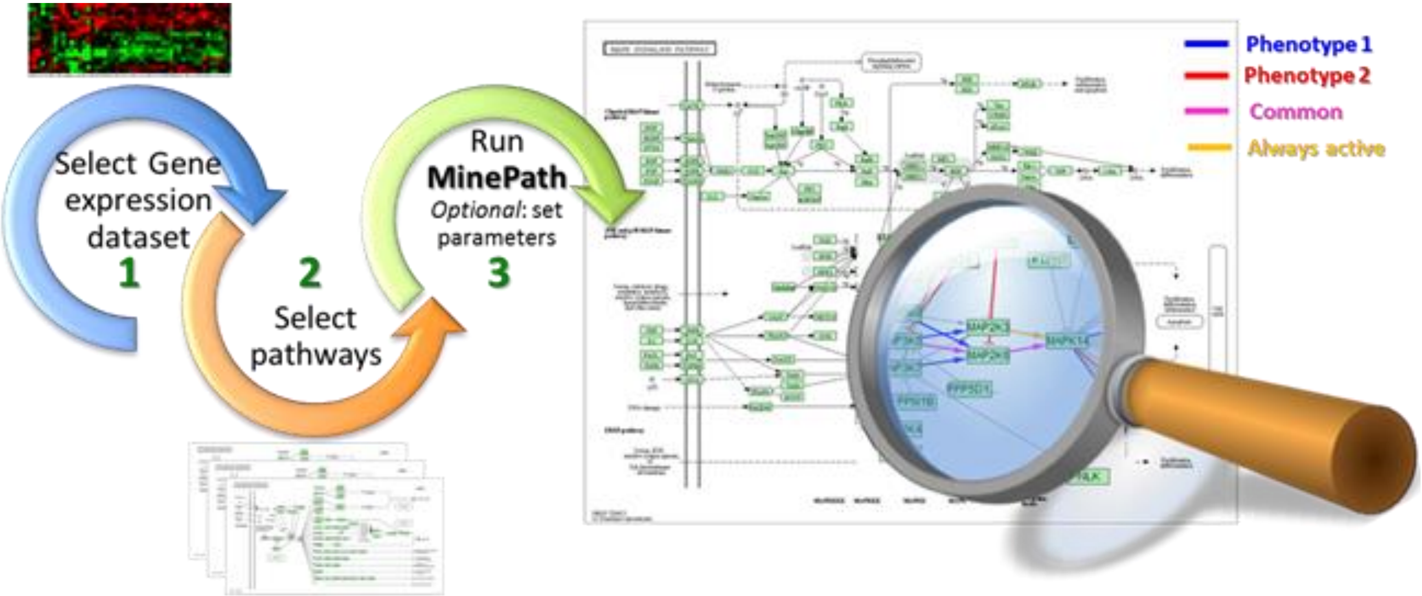
Is an of-line functionality that can be used only from the standalone tool of MinePath.

- Using this extra functionality we created an artificial pathway, which is the merged pathway of the 14 cancer related pathways



	KEGG Id	Pathway description
1	has04310	Wnt signalling
2	hsa04010	MAPK signalling
3	hsa04012	ErbB signalling
4	hsa04060	Cytocin-cytocin receptor interaction
5	hsa04110	Cell cycle
6	hsa04115	p53 signalling
7	hsa04150	mTOR signalling
8	hsa04210	Apoptosis
9	hsa04350	TGF-β signalling
10	hsa04370	VEGF signalling
11	hsa04510	Focal adhesion
12	hsa04512	ECM-receptor interaction
13	hsa04520	Adherens junction
14	hsa04630	Jak-STAT signalling

Web Based MinePath



Select Gene expression dataset 1

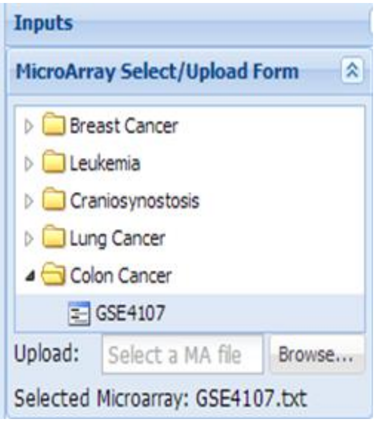
2 Select pathways

Run **MinePath** *Optional*: set parameters 3

http://minepath.org

Web Based MinePath

Select or upload gene expression dataset



Inputs

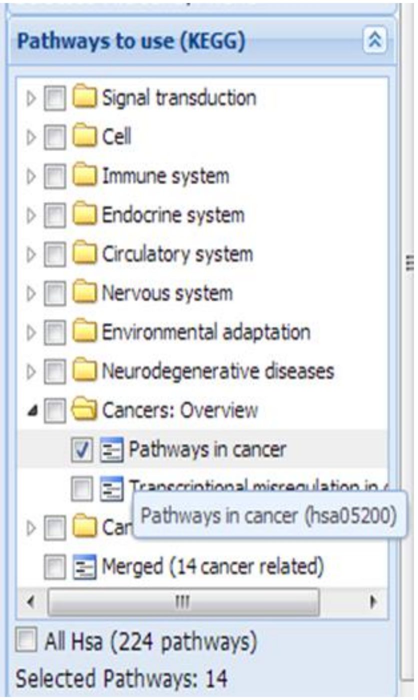
MicroArray Select/Upload Form

- ▷ Breast Cancer
- ▷ Leukemia
- ▷ Craniosynostosis
- ▷ Lung Cancer
- ◢ Colon Cancer
  - GSE4107

Upload: Select a MA file Browse...

Selected Microarray: GSE4107.txt

Select pathways



Pathways to use (KEGG)

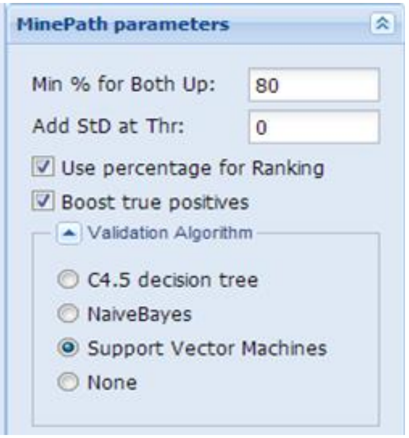
- ▷ □ Signal transduction
- ▷ □ Cell
- ▷ □ Immune system
- ▷ □ Endocrine system
- ▷ □ Circulatory system
- ▷ □ Nervous system
- ▷ □ Environmental adaptation
- ▷ □ Neurodegenerative diseases
- ◢ □ Cancers: Overview
  - ☑ Pathways in cancer
  - □ Transcriptional misregulation in c
- ▷ □ Car
- □ Merged (14 cancer related)

Pathways in cancer (hsa05200)

□ All Hsa (224 pathways)

Selected Pathways: 14

Run MinePath



MinePath parameters

Min % for Both Up: 80

Add StD at Thr: 0

☑ Use percentage for Ranking

☑ Boost true positives

Validation Algorithm

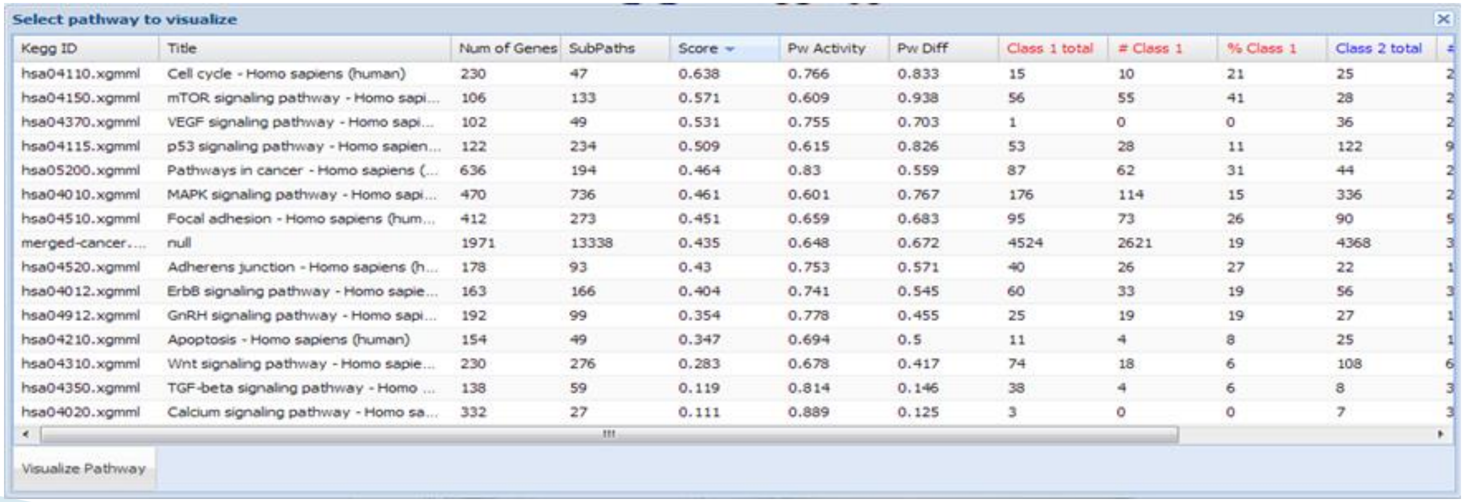
- ○ C4.5 decision tree
- ○ NaiveBayes
- ● Support Vector Machines
- ○ None

Web Based MinePath

Statistics for each pathway participated in the experiment such as the number of genes, the number of sub-paths, and number of sub-paths for each class and for the common sub-paths, percentages and three scores:

- Pathway power (*pwA*): is the sum of the significant sub-paths in the pathway (including the common sub-paths) divided by the number of the total sub-paths of the pathway.
- Pathway discriminant power (*pwDS*): is the number of the significant sub-paths for the two classes divided by the number of the total sub-paths of the pathway.
- The pathway score (*Score*):Score = pwA * pwDS

The user can also short the results based on any of these features.



Select pathway to visualize

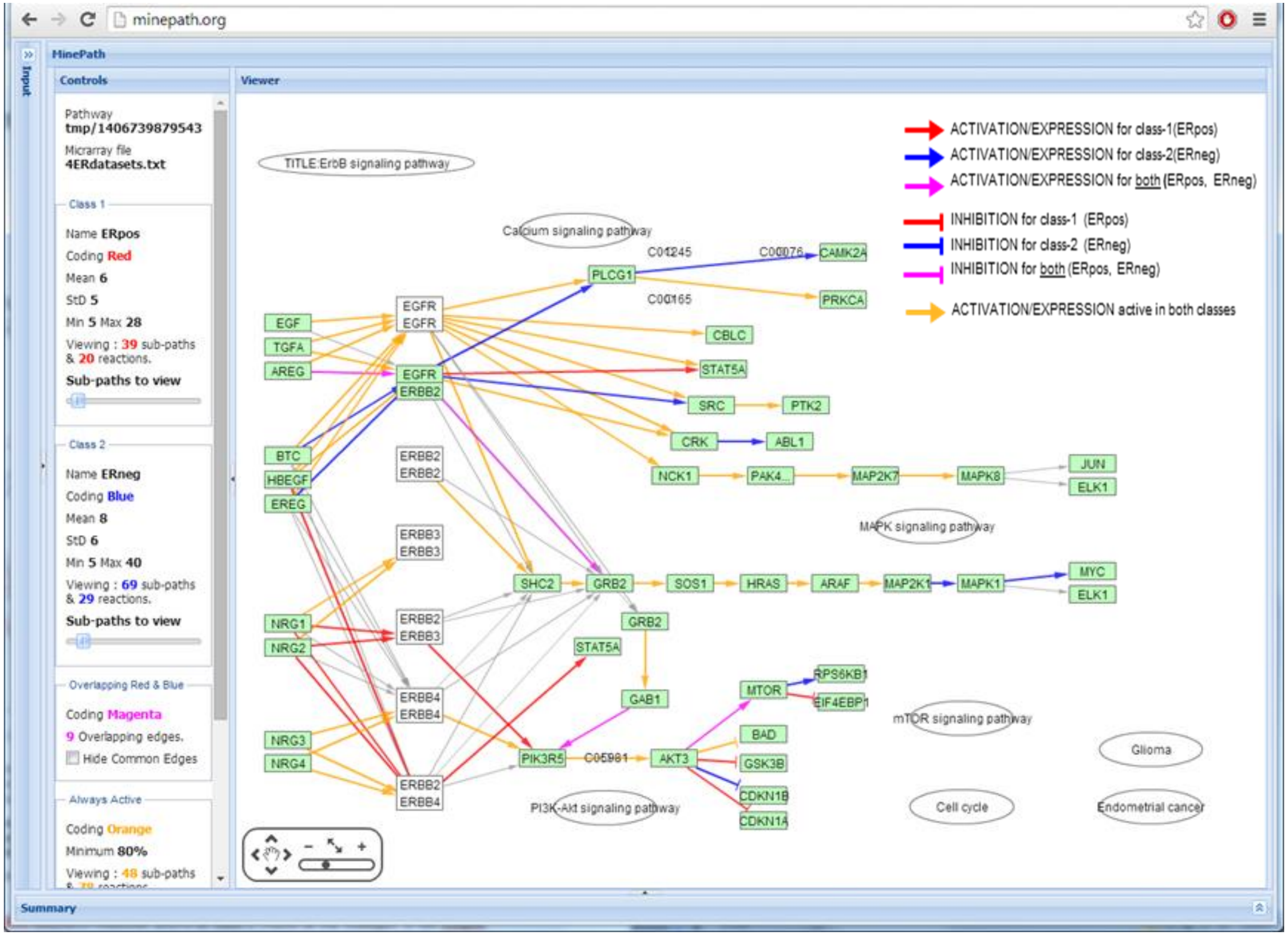
Kegg ID	Title	Num of Genes	SubPaths	Score	Pw Activity	Pw Diff	Class 1 total	# Class 1	% Class 1	Class 2 total	#
hsa04110.xgmml	Cell cycle - Homo sapiens (human)	230	47	0.638	0.766	0.833	15	10	21	25	2
hsa04150.xgmml	mTOR signaling pathway - Homo sapi...	106	133	0.571	0.609	0.938	56	55	41	28	2
hsa04370.xgmml	VEGF signaling pathway - Homo sapi...	102	49	0.531	0.755	0.703	1	0	0	36	2
hsa04115.xgmml	p53 signaling pathway - Homo sapien...	122	234	0.509	0.615	0.826	53	28	11	122	9
hsa05200.xgmml	Pathways in cancer - Homo sapiens (...	636	194	0.464	0.83	0.559	87	62	31	44	2
hsa04010.xgmml	MAPK signaling pathway - Homo sapi...	470	736	0.461	0.601	0.767	176	114	15	336	2
hsa04510.xgmml	Focal adhesion - Homo sapiens (hum...	412	273	0.451	0.659	0.683	95	73	26	90	5
merged-cancer....	null	1971	13338	0.435	0.648	0.672	4524	2621	19	4368	3
hsa04520.xgmml	Adherens junction - Homo sapiens (h...	178	93	0.43	0.753	0.571	40	26	27	22	1
hsa04012.xgmml	ErbB signaling pathway - Homo sapie...	163	166	0.404	0.741	0.545	60	33	19	56	3
hsa04912.xgmml	GnRH signaling pathway - Homo sapi...	192	99	0.354	0.778	0.455	25	19	19	27	1
hsa04210.xgmml	Apoptosis - Homo sapiens (human)	154	49	0.347	0.694	0.5	11	4	8	25	1
hsa04310.xgmml	Wnt signaling pathway - Homo sapie...	230	276	0.283	0.678	0.417	74	18	6	108	6
hsa04350.xgmml	TGF-beta signaling pathway - Homo ...	138	59	0.119	0.814	0.146	38	4	6	8	3
hsa04020.xgmml	Calcium signaling pathway - Homo sa...	332	27	0.111	0.889	0.125	3	0	0	7	3

Web Based MinePath

Colour coding:

- **Red**: sub-paths active at class 1
- **Blue**: sub-paths active at class 2
- **Magenta:** overlapping sub-paths in the two classes
- **Orange:** sub-paths that are always active.

MinePath supports active interaction and immediate visualization when the end user sets new thresholds for the two phenotypes or for the always active sub-paths, as well as to hide/show the overlapping relations and hide/show the association-dissociations of the pathway from the **control panel**



Pathway **tmp/1406739879543**
Micrarray file **4ERdatasets.txt**

Class 1
Name **ERpos**
Coding **Red**
Mean **6**
StD **5**
Min **5** Max **28**
Viewing : **39** sub-paths & **20** reactions.
Sub-paths to view

Class 2
Name **ERneg**
Coding **Blue**
Mean **8**
StD **6**
Min **5** Max **40**
Viewing : **69** sub-paths & **29** reactions.
Sub-paths to view

Overlapping Red & Blue
Coding **Magenta**
9 Overlapping edges.
□ Hide Common Edges

Always Active
Coding **Orange**
Minimum **80%**
Viewing : **48** sub-paths & **78** reactions.

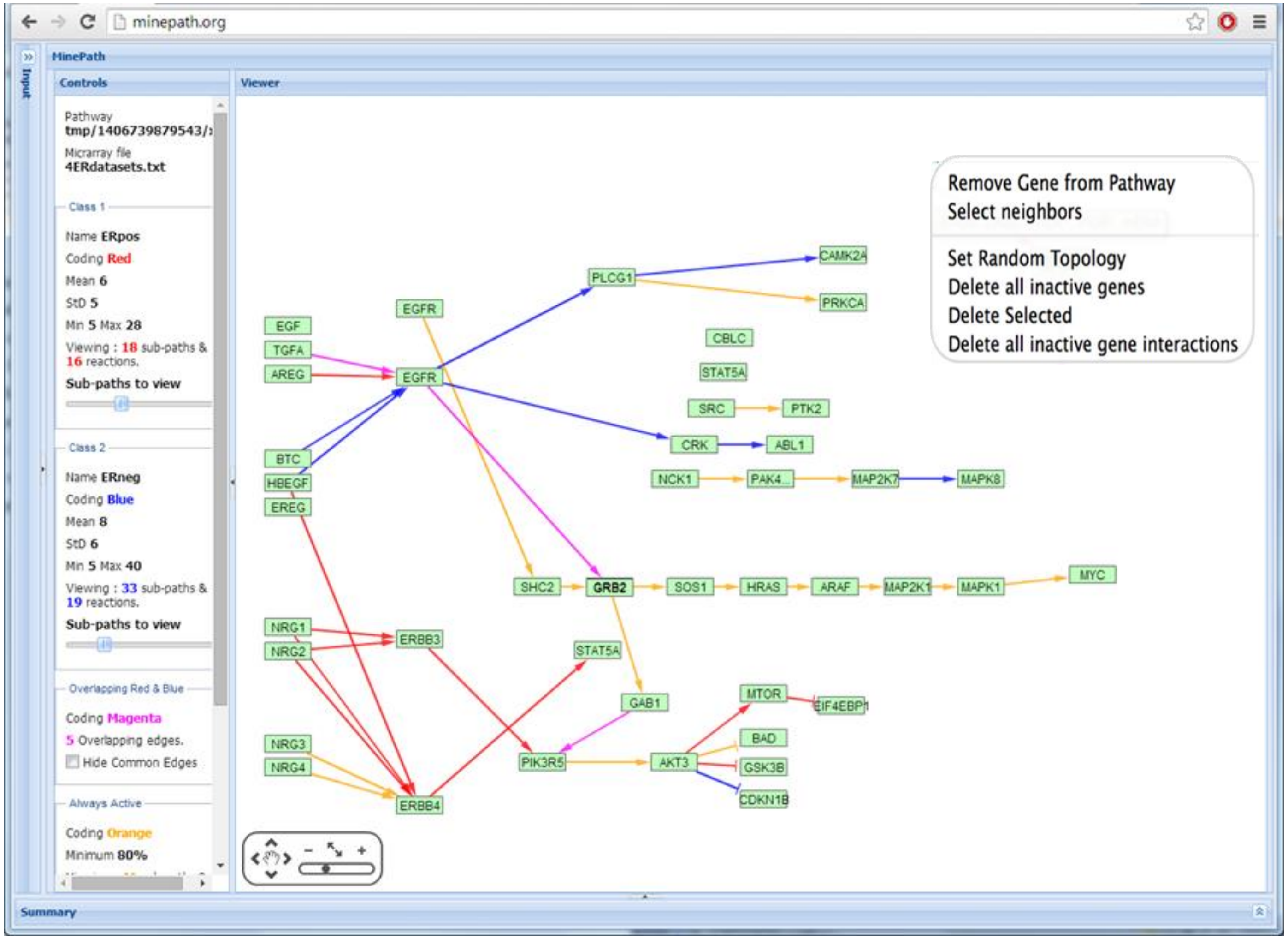
An example of the ErbB pathway for the '4ERdatasets' dataset

Web Based MinePath

MinePath is equipped with special functionality that enables the reduction of network's complexity:

- deletion of genes
- deletion of relations
- deletion of parts of the network
- re-orientation of its topology.

The functionality is available with *a right click (in the viewer)*.



minepath.org

MinePath

Controls

Pathway
tmp/1406739879543/

Micrarray file
4ERdatasets.txt

Class 1

Name ERpos

Coding Red

Mean 6

StD 5

Min 5 Max 28

Viewing : 18 sub-paths & 16 reactions.

Sub-paths to view

Class 2

Name ERneg

Coding Blue

Mean 8

StD 6

Min 5 Max 40

Viewing : 33 sub-paths & 19 reactions.

Sub-paths to view

Overlapping Red & Blue

Coding Magenta

5 Overlapping edges.

□ Hide Common Edges

Always Active

Coding Orange

Minimum 80%

Thresholds:13 for class 1 (ERpos), 13 for class 2 (ERneg), 95% for always active sub-paths

Experiments

MinePath comparison study

GGEA* :

Glioma cases from the GSE4271 (100 samples) versus the control cases from the GSE1133 (158 samples)

- most of the selected pathways from GGEA have been identified as highly discriminant using MinePath
- 17 pathways listed in the FiDePa also occur in the top 25 of the GGEA ranking
- MinePath ranked **Glioma** pathway **as highly discriminant** (score 1) while using **FiDePa** is ranked in **20th** position and using GGEA in **12th** position.

10-fold cross validation using the best sub-paths is 100%.

Pathway	MinePath score (pw diff)	ORA *P* (GGEA)	Rank (FiDePa)
Neurotrophin signalling	1	5.5E-15	–
Pancreatic cancer	1	3.8E-14	12
Renal cell carcinoma	1	1.3E-13	–
Chronic myeloid leukaemia	1	6.3E-13	8
Glioma	**1**	**5.1E-12**	**20**
Insulin signalling	1	3.2E-11	18
Adherens junction	1	4.9E-11	6
MAPK signalling	0.977	0.0000044	1
Cell cycle	0.966	---	19
Adipocytokine signaling pathway	0.964	---	14
Toll-like receptor signalling	0.962	1.2E-09	10
Acute myeloid leukaemia	0.957	0.00000039	–
Apoptosis	0.955	0.04	3
Leucocyte transendothelial migration	0.952	3.9E-11	24
Nature killer cell mediated cytotoxicity	0.938	6.5E-11	2
Pathways in cancer	0.93	1.8E-24	–
T cell receptor signalling	0.926	1.2E-17	7
ErbB signalling	0.926	8.9E-13	–
mTOR signalling	0.92	0.0000012	15
B cell receptor signalling	0.917	4.2E-12	17
Colorectal cancer	0.875	1.1E-14	11
Focal adhesion	0.855	1.4E-18	5
Wnt signalling	0.851	1.2E-10	–
GnRH signalling	0.829	6.5E-11	16
VEGF signalling	0.8	1.5E-13	22
Non-small cell lung cancer	0.8	0.00000034	–
Fc epsilon RI signalling	0.44	4.1E-13	9
Endometrial Cancer	---	0.00000016	–

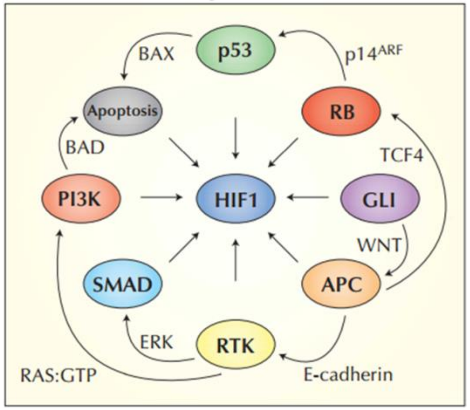
* Geistlinger, Ludwig, et al. "From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems." *Bioinformatics* 27.13 (2011): i366-i373.

MinePath comparison study

PATHOME*

- For validation, authors compared the performance with DAVID and GSEA based on a reference set of known cancer related pathways**.
- Gastric cancer study GSE13861
  - 65 Tumor samples
  - 19 non-tumor samples

Core reference set



Reference Standard**	KEGG Pathway	Title	PATHOME*	DAVID	GSEA	MinePath
HIF1	hsa04150	mTOR signaling	X	X	X	0
	hsa05200	Pathways in cancer	0	X	X	0
	hsa05211	Renal cell carcinoma	X	X	X	X
P53	hsa04115	P53 signaling	X	X	X	X
RB(cell cycle)	hsa04110	Cell cycle	X	X	0	X
Apoptosis	hsa04210	Apoptosis	X	X	X	X
GLI	hsa04340	Hedgehog signaling	X	X	X	X
APC	hsa04310	Wnt signaling	0	X	X	0
RTK	hsa04012	ERBB signaling	X	X	X	X
	hsa05200	Pathways in cancer	0	X	X	0
SMAD	hsa04350	TGF-βsignaling	X	X	X	0
PI3K	hsa04012	ERBB signaling	X	X	X	X
	hsa05200	Pathways in cancer	0	X	X	0
	hsa04150	mTOR signaling	X	X	X	0
	hsa04010	MAPK signaling	0	X	X	0
	hsa04910	Insulin signaling	0	X	X	0
	hsa04510	Focal adhesion	0	0	X	0
	hsa04062	Chemokine signaling	0	X	X	0
	hsa04370	VEGF signaling	X	X	X	X
	19	Hits	8	1	1	**11**
		Selected	27	15	17	19

where X not detected, 0 Detected

* Nam et al. "PATHOME: an algorithm for accurately detecting differentially expressed subpathways." Oncogene (2014).

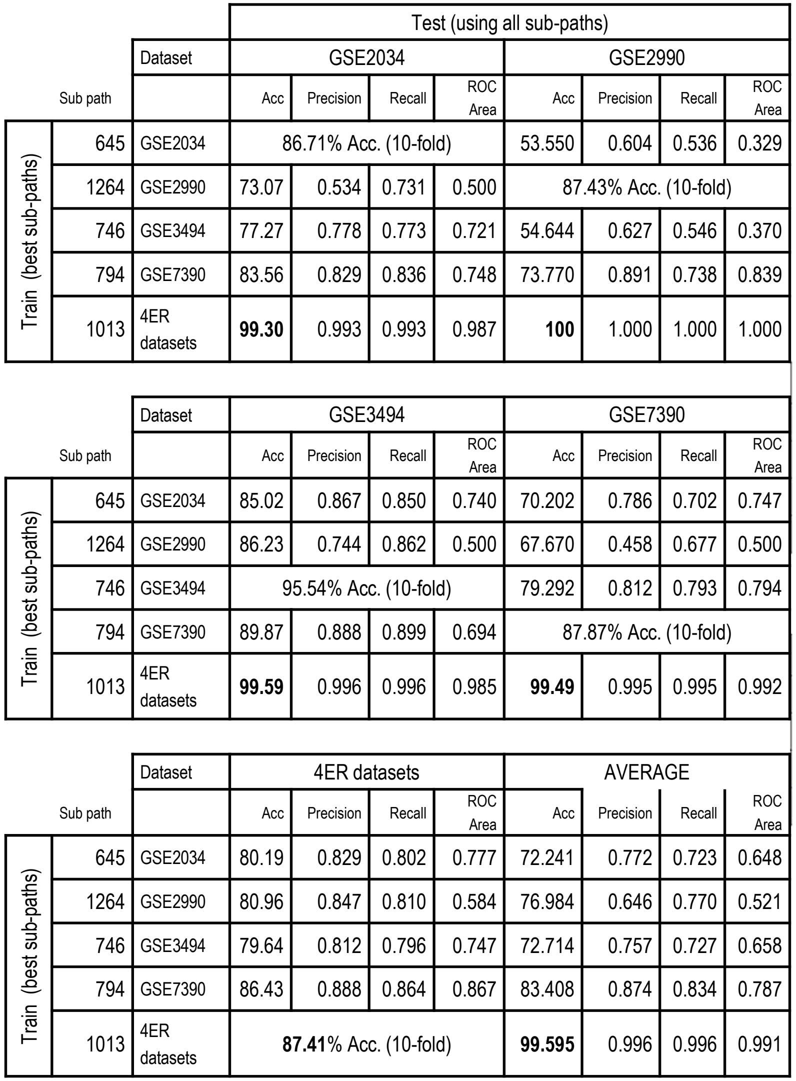
** Vogelstein, Bert, & Kenneth W. Kinzler. "Cancer genes and the pathways they control." *Nature medicine* 10.8 (2004): 789-799.

Validation on independent datasets

Dataset	GSE2034	GSE2990	GSE3494	GSE7390	4ER datasets
Platform	Affy-U133A				
Class	ER+ vs ER-				
ER+ samples	209	149	213	134	705
ER- samples	77	34	34	64	209
Probes	22283				
KEGG Ids	20967				

The merged dataset (4ER) performed the best accuracies overall. Even though the merged dataset actually contains the test subset each time, its trained model provided very high accuracies (over 99%) overall the datasets.

- *"Xu et al* "Integrating data from multiple studies to obtain more samples appears to be a promising way to overcome the prevalence of study-specific signatures and difficulties in validating the prognostic tests constructed from these signatures on independent data."*



		Test (using all sub-paths)							
	Dataset	GSE2034				GSE2990			
Sub path		Acc	Precision	Recall	ROC Area	Acc	Precision	Recall	ROC Area
Train (best sub-paths) 645	GSE2034	86.71% Acc. (10-fold)				53.550	0.604	0.536	0.329
1264	GSE2990	73.07	0.534	0.731	0.500	87.43% Acc. (10-fold)			
746	GSE3494	77.27	0.778	0.773	0.721	54.644	0.627	0.546	0.370
794	GSE7390	83.56	0.829	0.836	0.748	73.770	0.891	0.738	0.839
1013	4ER datasets	**99.30**	0.993	0.993	0.987	**100**	1.000	1.000	1.000

	Dataset	GSE3494				GSE7390			
Sub path		Acc	Precision	Recall	ROC Area	Acc	Precision	Recall	ROC Area
Train (best sub-paths) 645	GSE2034	85.02	0.867	0.850	0.740	70.202	0.786	0.702	0.747
1264	GSE2990	86.23	0.744	0.862	0.500	67.670	0.458	0.677	0.500
746	GSE3494	95.54% Acc. (10-fold)				79.292	0.812	0.793	0.794
794	GSE7390	89.87	0.888	0.899	0.694	87.87% Acc. (10-fold)			
1013	4ER datasets	**99.59**	0.996	0.996	0.985	**99.49**	0.995	0.995	0.992

	Dataset	4ER datasets				AVERAGE			
Sub path		Acc	Precision	Recall	ROC Area	Acc	Precision	Recall	ROC Area
Train (best sub-paths) 645	GSE2034	80.19	0.829	0.802	0.777	72.241	0.772	0.723	0.648
1264	GSE2990	80.96	0.847	0.810	0.584	76.984	0.646	0.770	0.521
746	GSE3494	79.64	0.812	0.796	0.747	72.714	0.757	0.727	0.658
794	GSE7390	86.43	0.888	0.864	0.867	83.408	0.874	0.834	0.787
1013	4ER datasets	**87.41**% Acc. (10-fold)				**99.595**	0.996	0.996	0.991

* Xu, Lei, et al. "Merging microarray data from separate breast cancer studies provides a robust prognostic test." *BMC Bioinformatics* 9.1 (2008): 125.

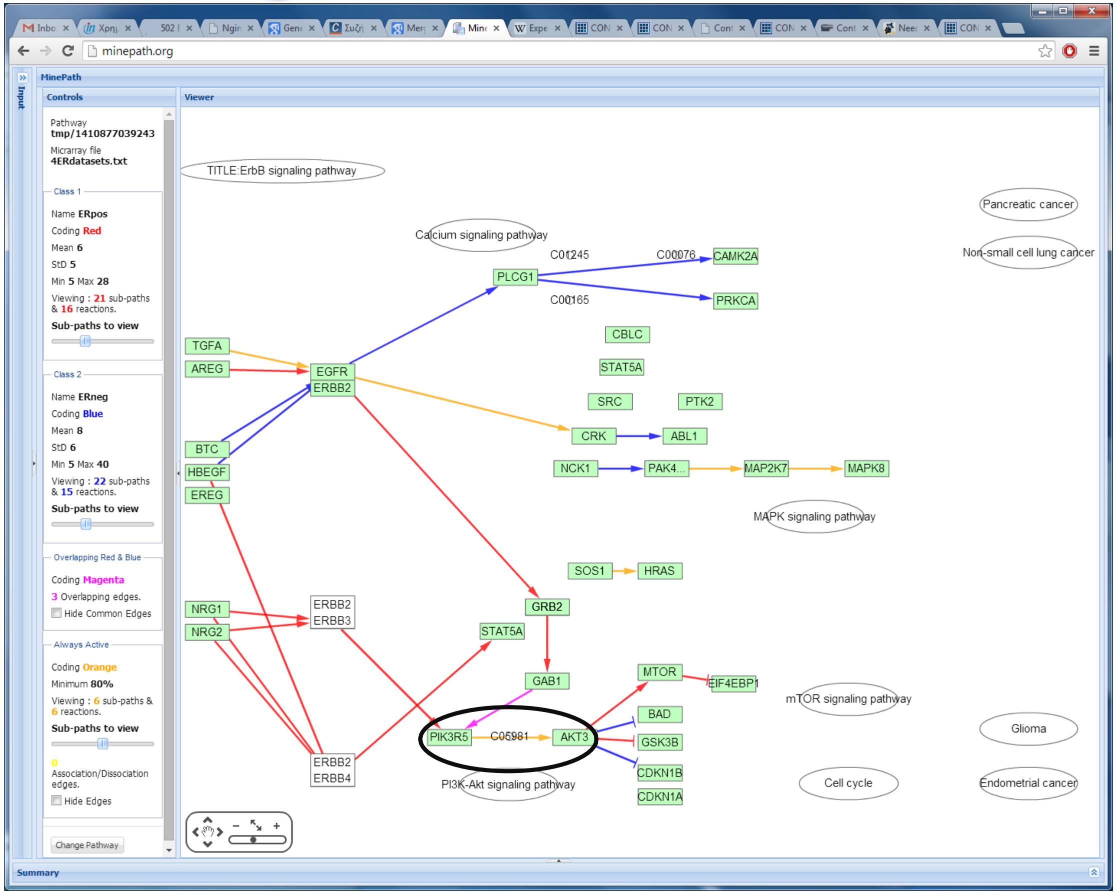
MinePath – discovery of New Biological Knowledge

Merged ER datasets & 14 cancer related pathways

1. Load/visualize ErbB signaling pathway
2. Double the thresholds (ER+ from 6 to 12, ER- from 8 to 16 & common to 90%)
3. Delete inactive genes and relations

According to the literature, the results are quite relevant to the estrogen-receptor status.

- Hutcheson et al*: "…fulvestrant treatment is sensitive to the actions of the ErbB3/4 ligand HRGb1 (NRG1) with enhanced ErbB3/4-driven signaling activity, and significant increases in cell proliferation …"



minepath.org

MinePath

Input

Controls

Pathway
tmp/1410877039243
Micrarray file
4ERdatasets.txt

Class 1

Name ERpos
Coding Red
Mean 6
StD 5
Min 5 Max 28
Viewing : 21 sub-paths & 16 reactions.
Sub-paths to view

Class 2

Name ERneg
Coding Blue
Mean 8
StD 6
Min 5 Max 40
Viewing : 22 sub-paths & 15 reactions.
Sub-paths to view

Overlapping Red & Blue

Coding Magenta
3 Overlapping edges.
□ Hide Common Edges

Always Active

Coding Orange
Minimum 80%
Viewing : 6 sub-paths & 6 reactions.
Sub-paths to view
0
Association/Dissociation edges.
□ Hide Edges

Change Pathway

Viewer

Summary

Exploring ErBb for the 4ERdatasets using MinePath

* Hutcheson, I.R., et al.: Heregulin beta1 drives gefitinib-resistant growth and invasion in tamoxifen-resistant MCF-7 breast cancer cells. *Breast Cancer Research* 9(4):R50, (2007)

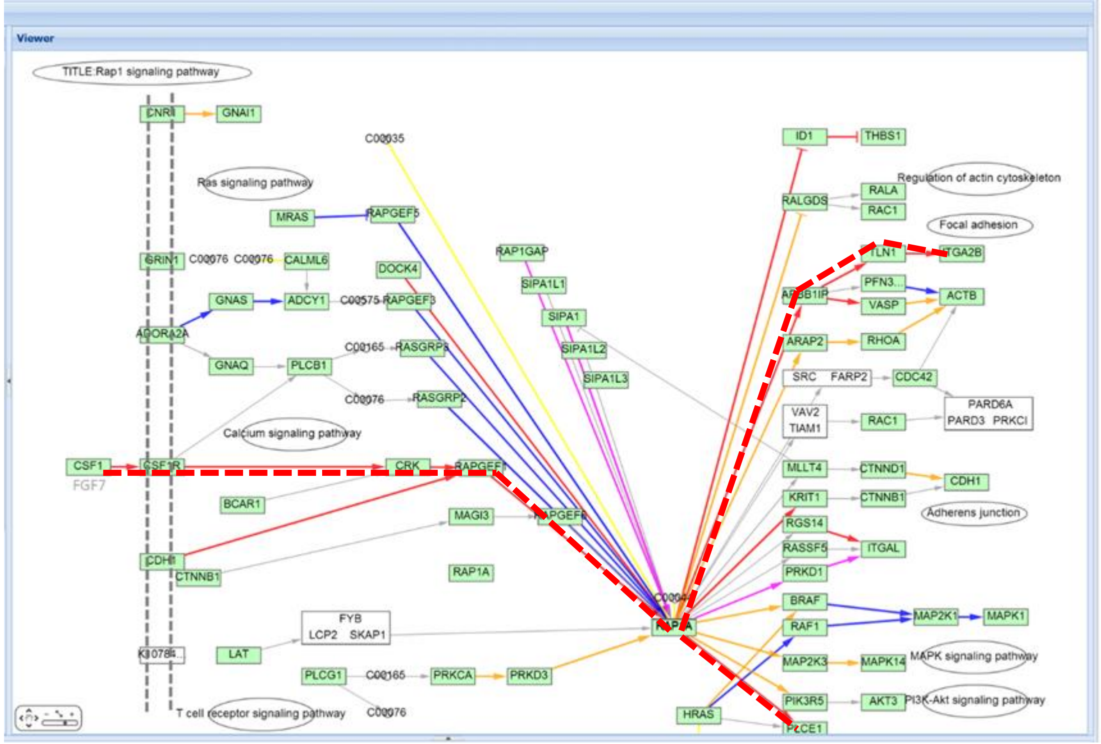
MinePath Biological Validation

Craniosynostosis (GSE27976*)

- 199 patients compared against a control population (n = 50)

MinePath identified **Rap1** signaling pathway as one of the most discriminant pathways and the most informative for Synostosis:

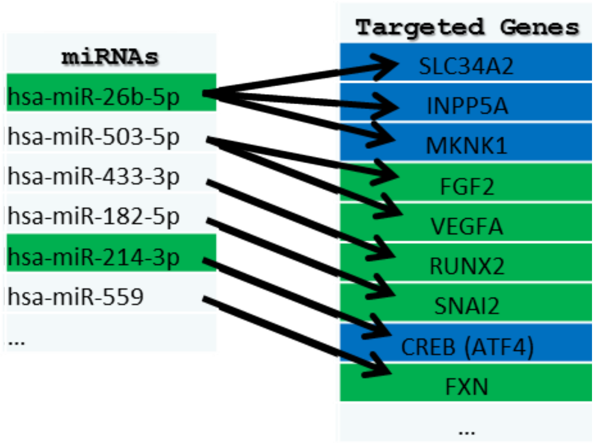
- **CSF1→CSF1R→CRK→RAPGEF1→RAP1A→APBB1P→TLN1→ITGA2B** leading to Focal Adhesion
  - Stamper et al* :
    - **FGF7/CSF1** (the most discriminant gene)
    - Focal adhesion pathway (the most discriminant pathway)
- **CSF1→CSF1R→CRK→RAPGEF1→RAP1A→PLCE1** leading to the PI3K-Akt signaling pathway.
  - Dufour et al** identified that PI3K/Akt attenuation plays important role in the control of osteoblast survival by FGFR2 signaling (member of the fibroblast growth factor FGFR family).



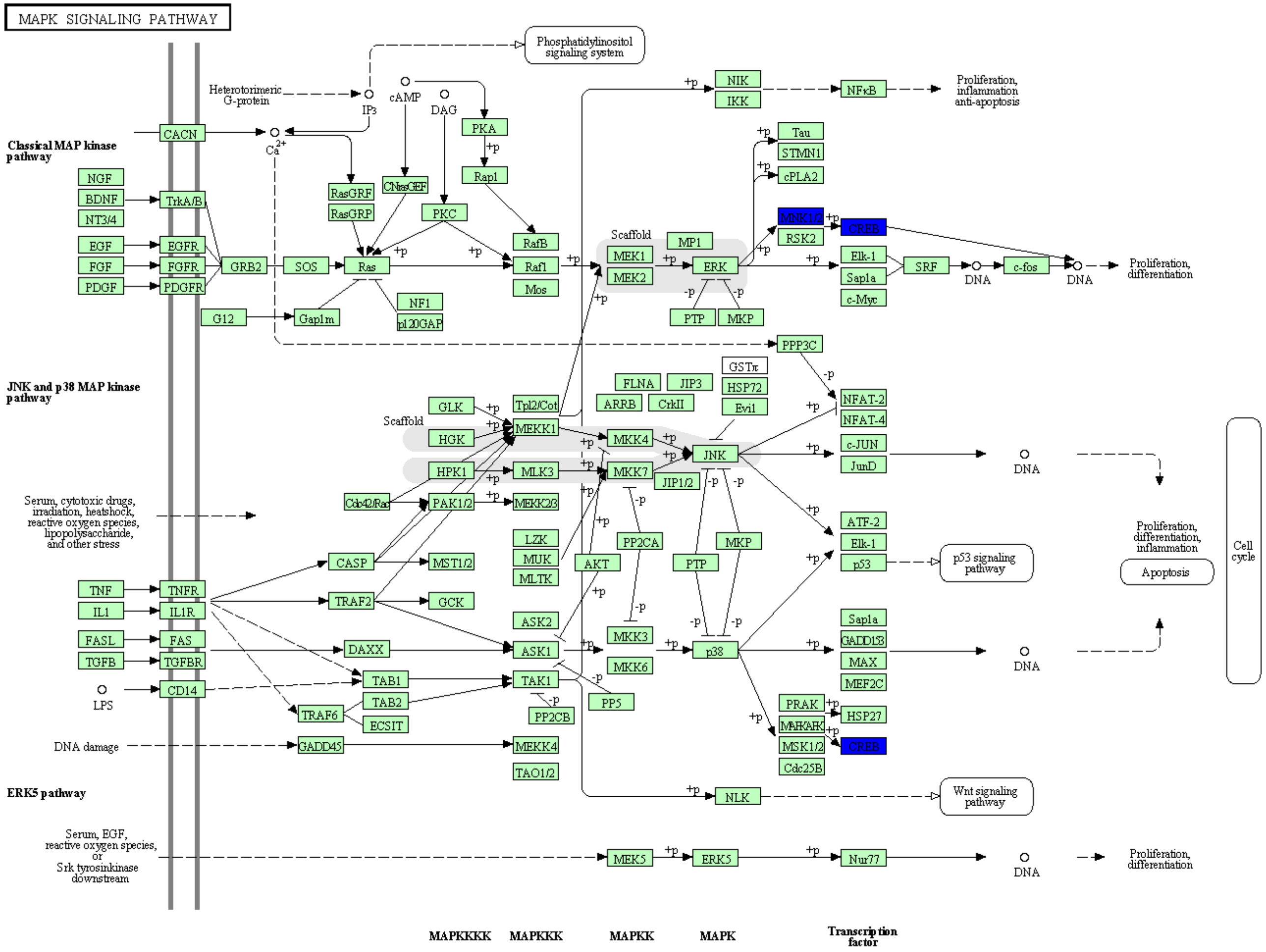
* Stamper, Brendan David, et al. "Transcriptome correlation analysis identifies two unique craniosynostosis subtypes associated with IRS1 activation."*Physiological genomics* 44.23 (2012): 1154-1163

** Dufour, Cécilie, et al. "FGFR2-Cbl interaction in lipid rafts triggers attenuation of PI3K/Akt signaling and osteoblast survival." *Bone* 42.6 (2008): 1032-1039.

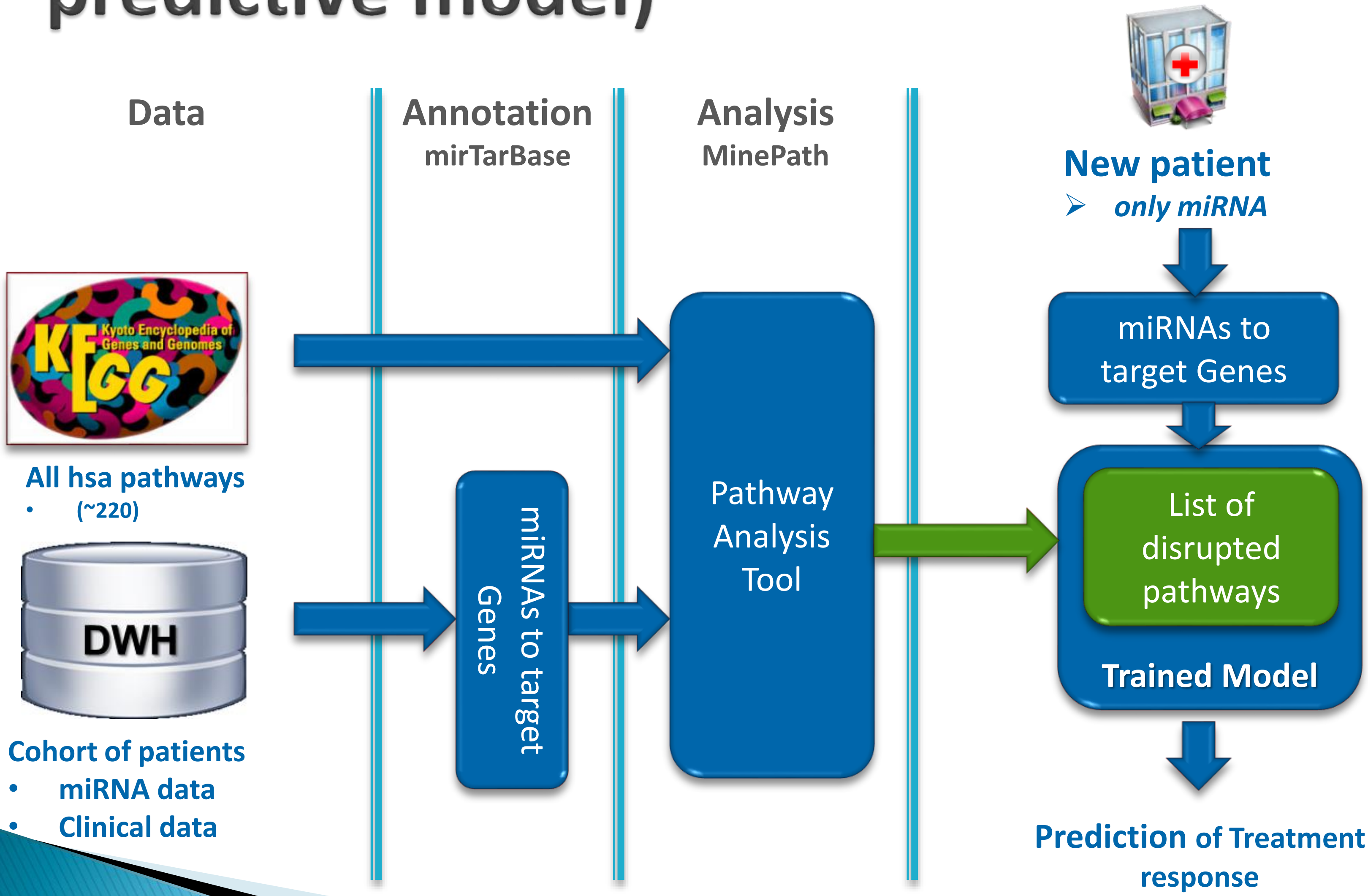
MinePath using miRNAs (a clinical predictive model)



miRNAs	Targeted Genes
hsa-miR-26b-5p	SLC34A2
hsa-miR-503-5p	INPP5A
hsa-miR-433-3p	MKNK1
hsa-miR-182-5p	FGF2
hsa-miR-214-3p	VEGFA
hsa-miR-559	RUNX2
…	SNAI2
	CREB (ATF4)
	FXN
	…



MinePath using miRNAs (a clinical predictive model)



Data

Annotation
mirTarBase

Analysis
MinePath

All hsa pathways
- (~220)

Cohort of patients
- miRNA data
- Clinical data

DWH

miRNAs to target Genes

Pathway Analysis Tool

New patient
➢ *only miRNA*

miRNAs to target Genes

List of disrupted pathways

Trained Model

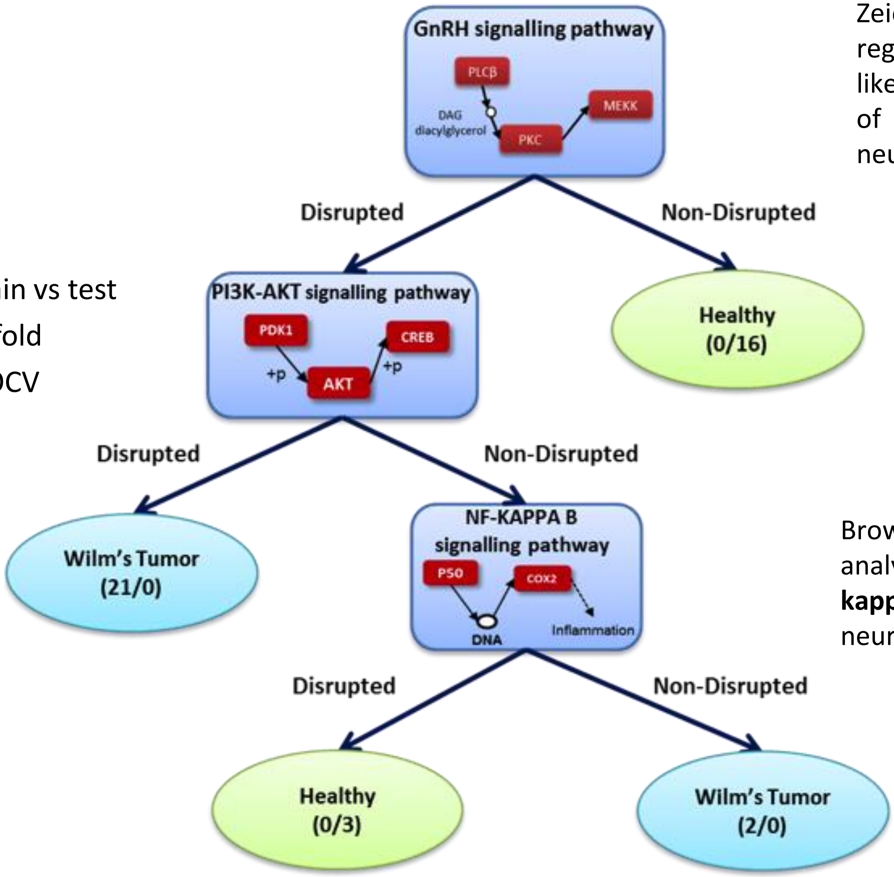
Prediction of Treatment response

Predictive models

Decision tree

(3 sub-paths)

- Accuracy
  - 100% in train vs test
  - 80% in 10-fold
  - 78% in LOOCV



Zeidman et al* proved that **PKCε** through its regulatory domain can induce immature neurite-like processes via a mechanism that appears to be of importance for neurite outgrowth during neuronal differentiation in neuroblastoma cells

Santo et al** identified the forkhead transcription factor FOXO3a as a key target of the PI3K/AKT pathway in neuroblastoma and concluded that the inactivation of FOXO3a by **AKT** was essential for neuroblastoma cell survival.

Brown et al*** using morphoproteomic analysis revealed the activation of the **NF-kappaB** pathway in high risk neuroblastoma cases

* Zeidman, et al. "PKCε, via its regulatory domain and independently of its catalytic domain, induces neurite-like processes in neuroblastoma cells." The Journal of cell biology 145, no. 4 (1999): 713-726

** Santo, et al. "FOXO3a is a major target of inactivation by PI3K/AKT signaling in aggressive neuroblastoma." Cancer research 73, no. 7 (2013): 2189-2198.

*** Brown et al. "Morphoproteomic confirmation of constitutively activated mTOR, ERK, and NF-kappaB pathways in high risk neuro-blastoma, with cell cycle and protein analyte correlates." Annals of Clinical & Laboratory Science 37, no. 2 (2007): 141-147.

Conclusions

Pathway selection methodologies

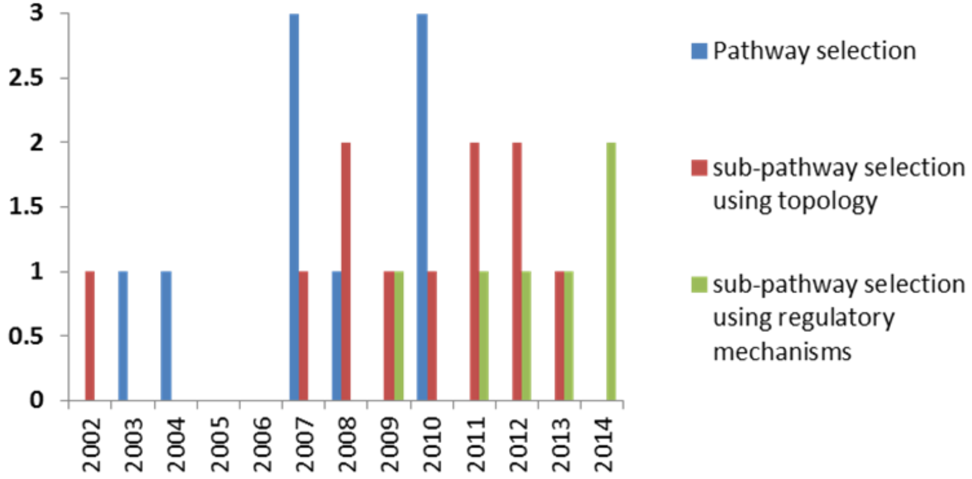
Pathway selection methodologies show similarities with gene signatures in terms of level of information used over the years

Three categories of methodologies focuses on the identification and selection of discriminant pathways and sub-paths:

1. identification of differentially expressed pathways using microarrays
2. Pathways or sub-paths selection using topology
3. Pathways or sub-paths selection using regulatory mechanisms

The most advanced and newer category is the third one which seems to be at its first steps and could possibly gain a momentum.

This assumption amplifies with the similarities we can find between the discriminant gene regulatory (sub)-networks and microarray gene selection methodologies.



Similar efforts

Only four tools take advantage of the underlying GRN gene regulation mechanisms, naming GGEA, SPIA, TEAK and PATHOME.

The main differences are:

- Methodologies count the activations and inhibitions (most of them with +1 and -1 respectively) and each sub-path gets a final score per phenotype which is also used as a ranking. Contrary, our approach strictly checks and takes into account only **sub-paths that are functional for each phenotype**
- Even though these methodologies take into account sub-paths none of them **report sub-paths**. They sum up and provide a ranking for each pathway as a whole.
- MinePath is the only methodology which takes into account and **visualizes sub-paths fully functional** in both phenotypes. These sub-paths have no discriminant power but can link the gap (functional interaction) between two sub-paths and reveal a complete functional root, which is biologically valuable
- MinePath offers a **complete solution** based on a **productive environment** with efficient, interactive and user-friendly visualization that offers rich exploratory capabilities
- **Web based** implementation

Conclusion

MinePath serves the users' exploratory needs to reveal the regulatory mechanisms that underlie and putatively govern the expression of target phenotypes

- The phenotype information is extracted from microarrays and all the selected GRNs are evaluated for **the identification of the most informative sub-paths at the specific phenotype**.
- These sub-paths present evidential molecular mechanisms that govern the disease itself, its type, its state or other targeted disease phenotypes

MinePath introduces a new and **efficient representation** of the differentially expressed sub-paths over a **Web-based** human-computer interface.

- supports live interaction, immediate visualization of regulatory relations and it is equipped with special topological and network-adjustment functionalities

The methodology was applied on gene-expression studies and results were quite indicative and strongly supported by the relevant biomedical literature

Future work

The modular implementation gives us the ability to "build on demand" new tools based on end user scenarios

- miRNA scenario/extension
- Validate candidate sub-paths (GRN reconstruction validation)

Additional functionality:

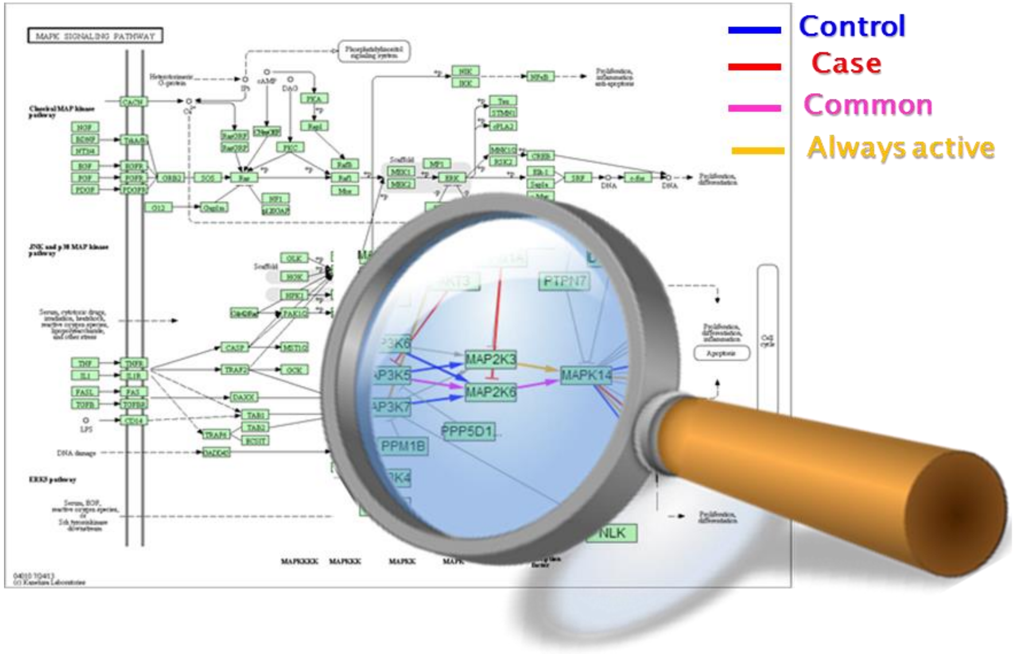
- For the methodology
  - Introduce new ranking algorithms
  - Introduce other pre-processing methodologies (apart discretization)
  - Support multi-class datasets
  - Support other quantified gene-expression data (e.g., RNA-seq)
- For the platform
  - automated uploading of microarray data from public sources (e.g., GEO)
  - merging of gene-expression datasets (to serve meta-analysis needs)
  - visualization of two or more pathways in order to enrich exploratory quests

Publications

- Koumakis L., Potamias G., Tsiknakis M., Zervakis M. and Moustakis V. Integrating Microarray Data and GRNs. Methods in Molecular Biology (under review)

- Koumakis L., Potamias G., Sfakianakis S., Moustakis V., Zervakis M., Graf N. and Tsiknakis M. "miRNA based pathway analysis tool in nephroblas-toma as a proof of principle for other cancer domains." In Bioinformatics and Bioengineering (BIBE), 2014 14th IEEE International Conference on BioInformatics and BioEngineering.
- Koumakis, L., Moustakis, V., Zervakis, M., Kafetzopoulos, D., & Potamias, G. Coupling Regulatory Networks and Microarays: evealing Molecular Regulations of Breast Cancer Treatment Responses. Artificial Intelligence: Theories and Applications. Lecture Notes in Computer Science, 7297, 239-246 (2012).
- Koumakis, L., Potamias, G., Zervakis, M., & Moustakis, V. (2011). Integrating microarray data and gene regulatory networks: Survey and critical considerations. 10th International Workshop on Biomedical Engineering. Kos, Greece 5-7 October 2011.
- K. Kalantzaki, L. Koumakis, E. Bei, M. Zervakis, G. Potamias and D. Kafetzopoulos. Experimental Model Construction and Validation of the ErbB Signaling Pathway. 13th IEEE International Conference on Bioinformatics and Bioengineering. Chania, Greece, November 10-13, 2013

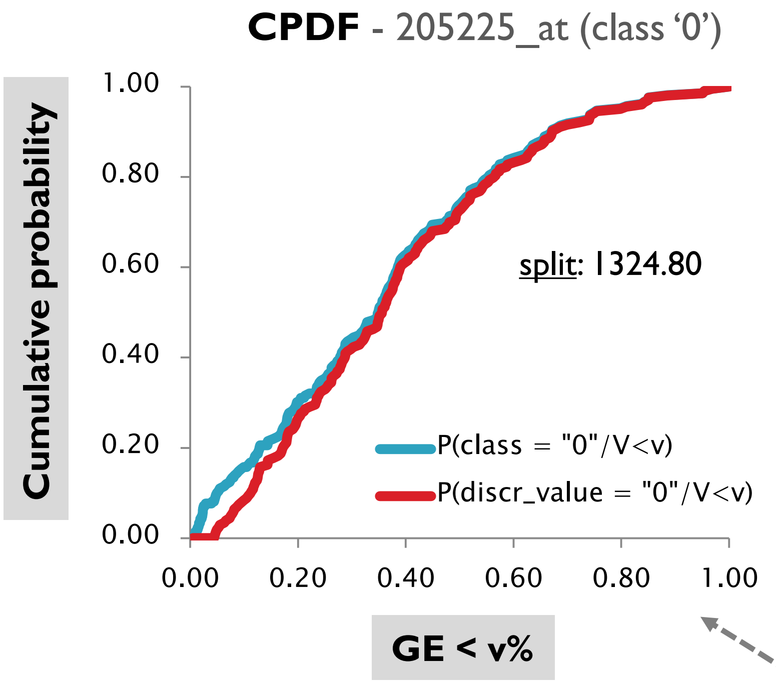
Ευχαριστώ

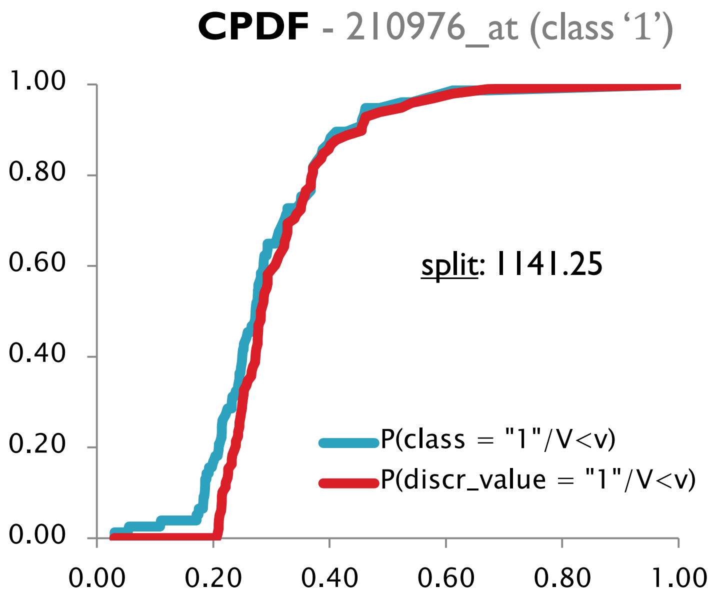


http://minepath.org

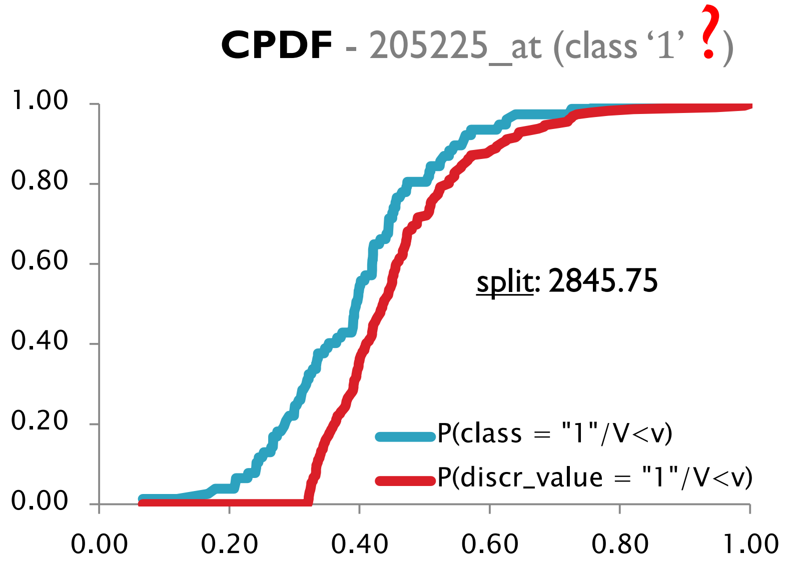
Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) – Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ . Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.

Discretization – a probabilistic evaluation


CPDF - 205225_at (class ‘0’)


CPDF - 210976_at (class ‘1’)

- MinePath Entropy-based discretization fits well the highly-discriminant genes
- … does not fit well the (very-)low discriminant genes


CPDF - 205225_at (class ‘1’ ?)

GSE2034 / 209 ER+ (“**0**”), 77 ER- (“**1**”) BRCA cases

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005 Feb 19-25;365(9460):671-9.