



LÉKAŘSKÁ FAKULTA MASARYKOVY UNIVERSITY
Interní hematologická klinika LF MU a FN Brno
Centrum molekulární biologie a genové terapie



Zpracování dat

Boris Tichý
30.10.2012



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Převod obrazové informace na numerická data

Adresace

Segmentace

Extrakce intenzit

Příprava dat

Kontrola kvality

Normalizace

Filtry

Analýza dat

Class discovery vs. Class prediction

Supervised vs. Unsupervised

Integrace s dalšími zdroji informací

Meta-analýza



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Zpracování obrazu

Adresace (gridding)

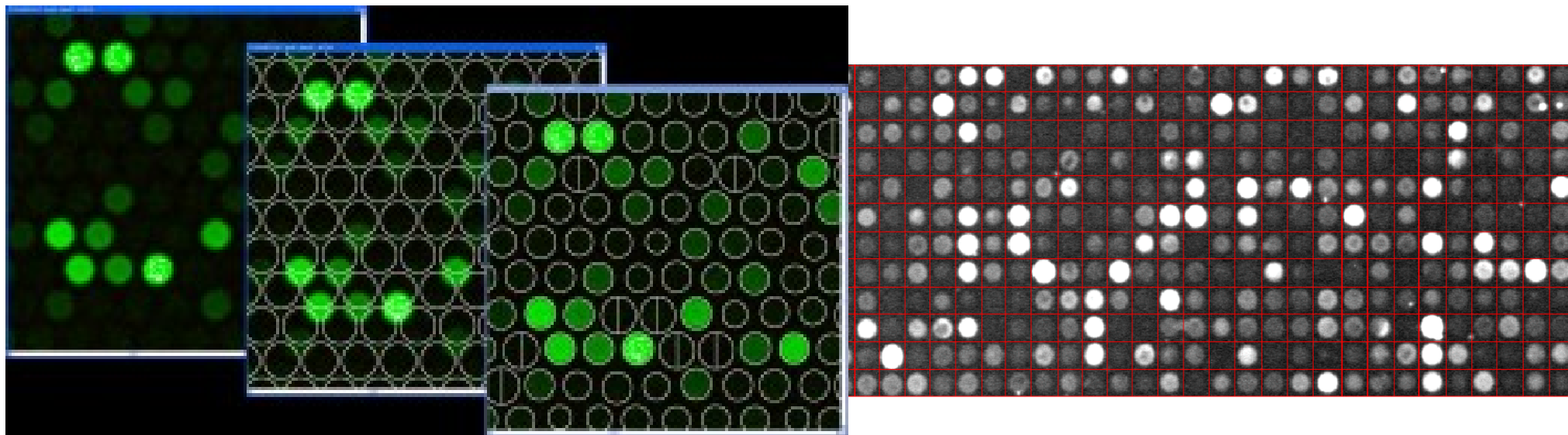
Určení pozice jednotlivých spotů

Většinou úprava předem daných očekávaných pozic

Pozice spotů bývají součástí anotace dodané výrobcem

Různé automatické a poloautomatické algoritmy

Většinou možnost manuální úpravy



Zpracování obrazu

Segmentace

Klasifikace jednotlivých pixelů

Popředí

Pozadí

Různé algoritmy

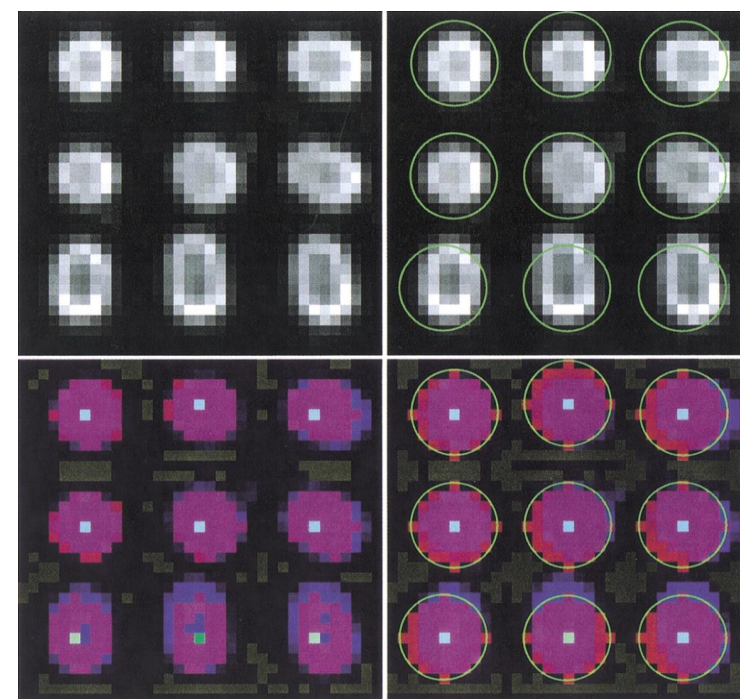
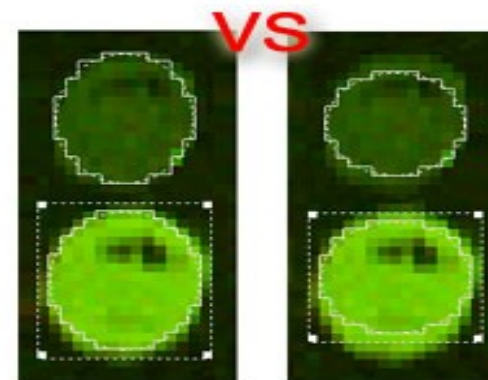
Fixed circle – velmi jednoduchý, neadaptivní

Adaptive circle

Histogram

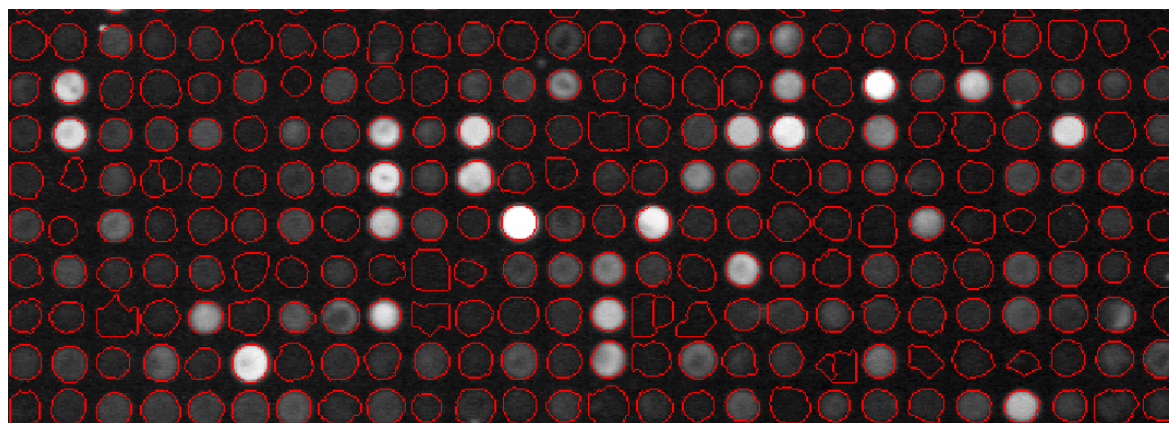
Adaptive shape – např. Watershed algoritmus

Shluková analýza



Spot default segmentation

Segmentation using circles



Zpracování obrazu

Extrakce intenzit

Různé vyjádření

Medián

Průměr

Poměr (červená/zelená)

Průměr poměrů jednotlivých pixelů

Výpočet parametrů kvality

Poměr signál/šum

Rozptyl intenzit pixelů (SD)

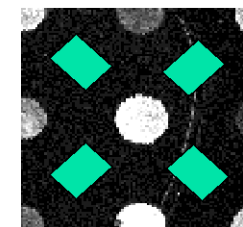
Cirkularita

Posun od očekávané pozice

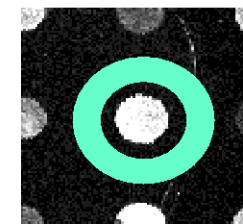
Affymetrix

Probe sets = více sond pro jeden gen

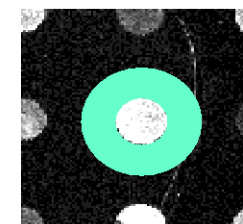
Speciální metody pro výpočet výsledné hodnoty



Spot, GenePix



ImageGene



ScanAlyze

Příprava dat

Normalizace

Dovoluje porovnávat naměřené hodnoty mezi sebou

V rámci jedné array

Mezi arrays

Jedna array

Mezi kanály (červená/zelená)

Mezi regiony

Mezi arrays

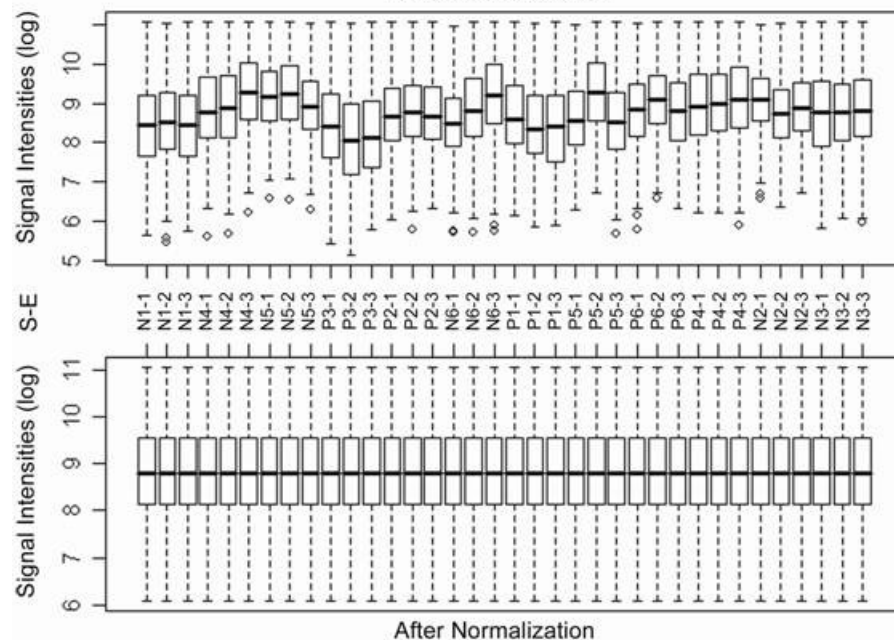
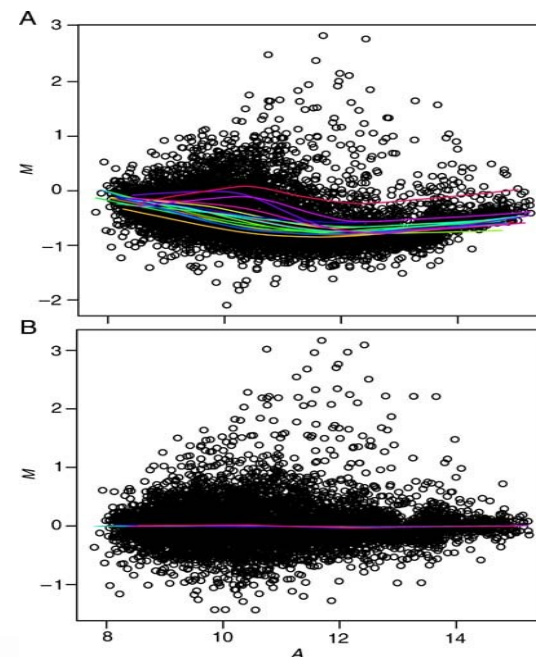
Dvoukanálové => poměry R/G

Jednokanálové => intenzity

Transformace

Log2 (poměry, intenzity)

Log10 (intenzity)



Příprava dat

Normalizace

Metody

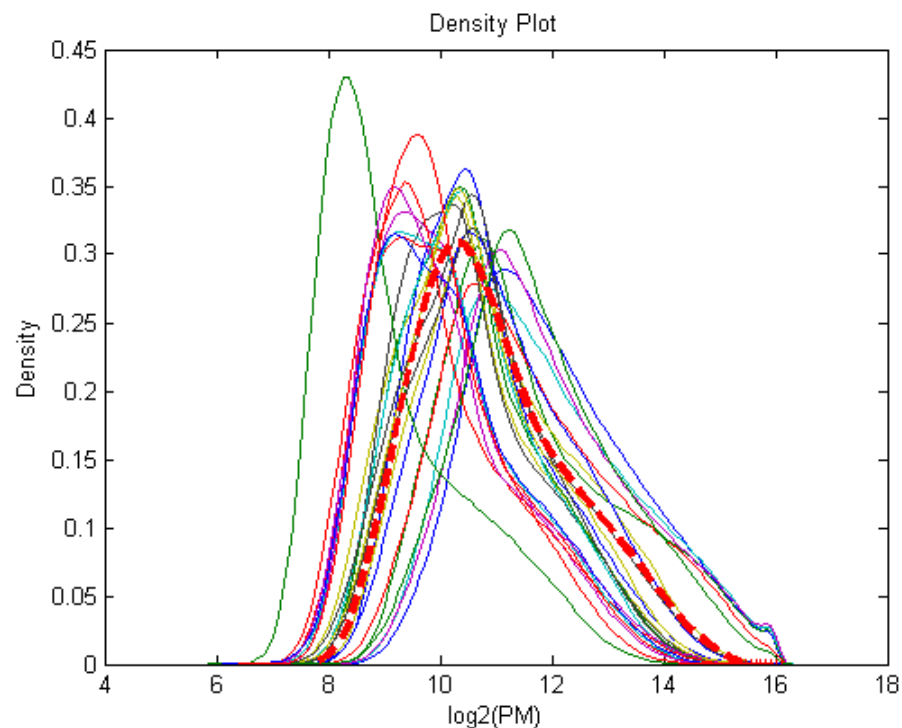
Lineární vs. nelineární

Vydělení průměrem

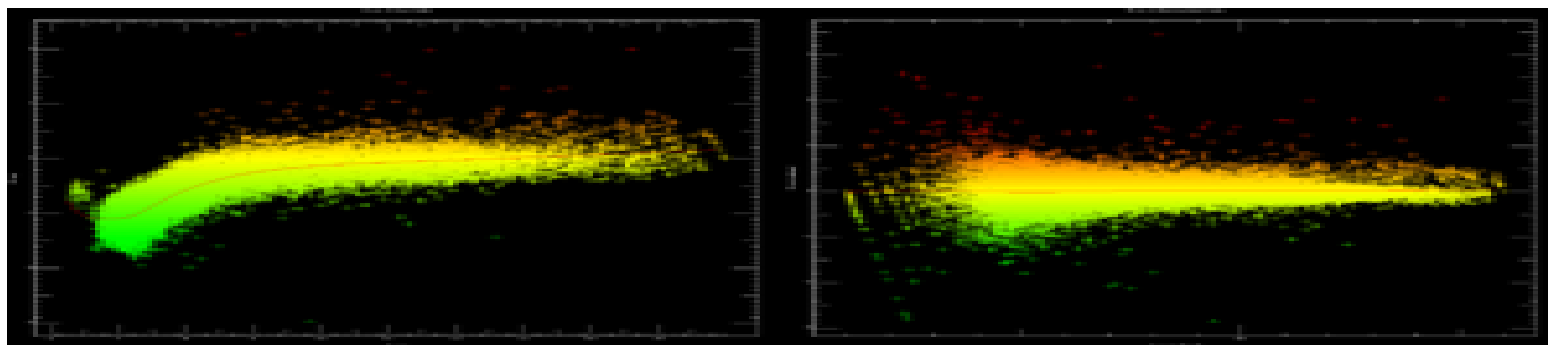
Scaling

Quantile

LOESS



Scaling s využitím housekeeping genů/spike-in kontrol



Příprava dat

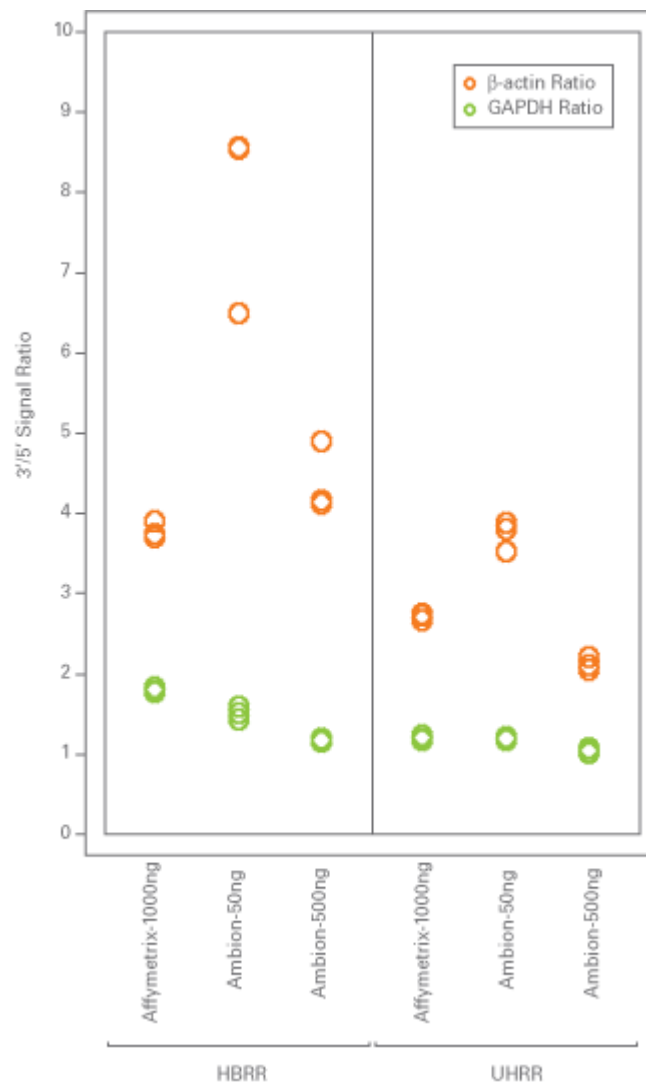
Kontrola kvality

Obecné
Pozadí
Distribuce intenzit
Distribuce poměrů

Specifické
Negativní, pozitivní kontroly
Spike-in
3'/5' poměr

Filtry

Odstranění nekvalitních/nepotřebných dat



Analýza dat

Class discovery

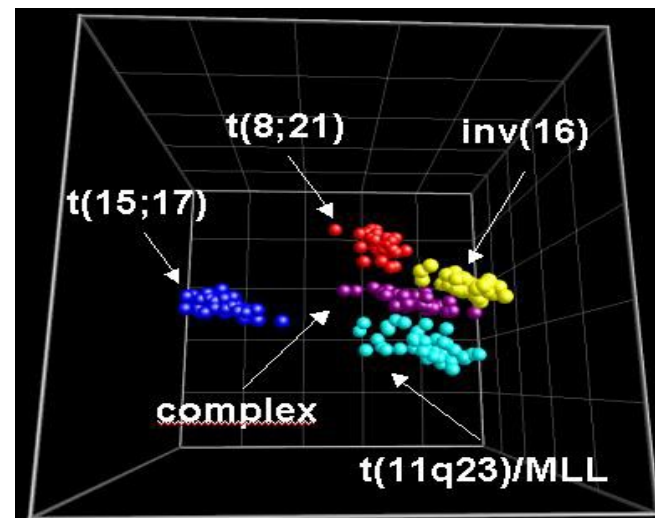
Nalezení vztahů v datech – skupiny vzorků, genů

Metody

Shluková analýza

Hierarchické shlukování, k-means, self organising maps,..

Analýza hlavních komponent (PCA)



Class prediction

Klasifikace vzorků na základě expresních dat

Metody

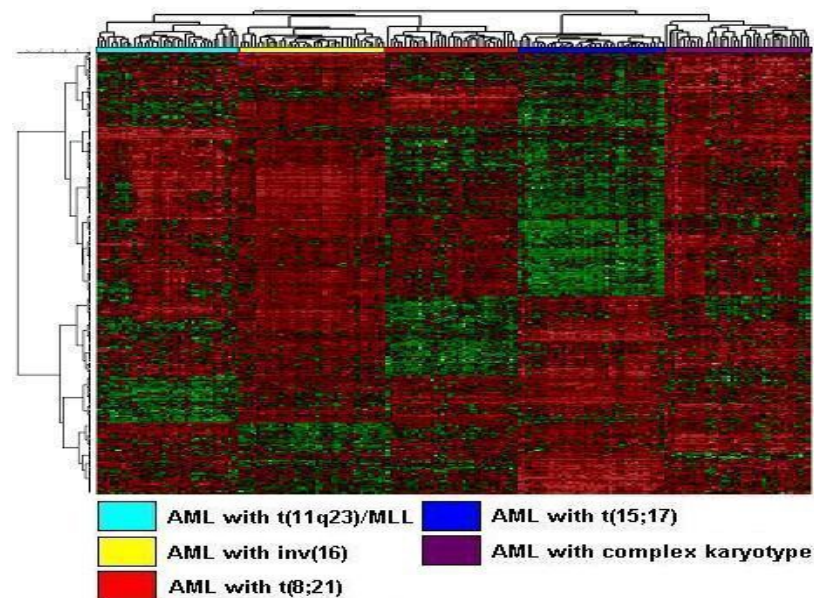
Support vector machines

K-nearest neighbor

Neural networks

Decision trees

Nearest shrunken centroids



Analýza dat

Unsupervised

Bez informace o vzorcích (genech)
(= Class discovery)

Supervised

Vzorky (geny) předem rozděleny do skupin

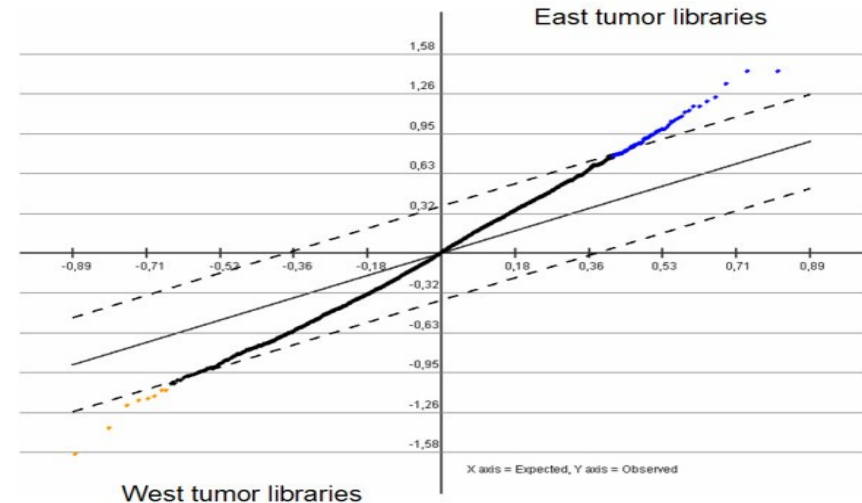
Class prediction – algoritmy se učí na známých vzorcích

Statistické testy

2 a více skupin (vzorků) – hledání rozdílů (dif. exprimovaných genů)

t-test, ANOVA, neparametrické testy, SAM (significance analysis of microarrays)

! korekce mnohonásobného testování (Bonferoni,...), odhad chyby (FDR)



Analýza dat

Integrace s dalšími daty

Funkce genů

Gene Ontology, KEGG, BioCarta

Genové interakce, možné funkční vztahy

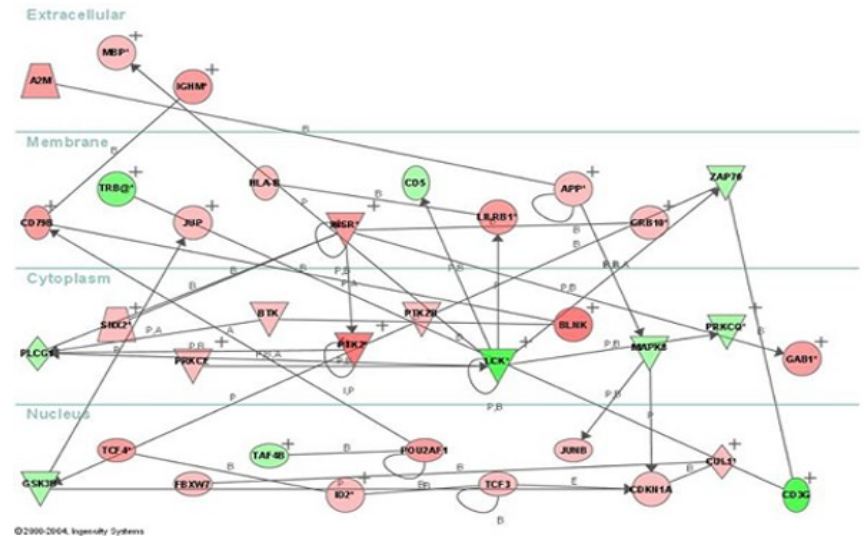
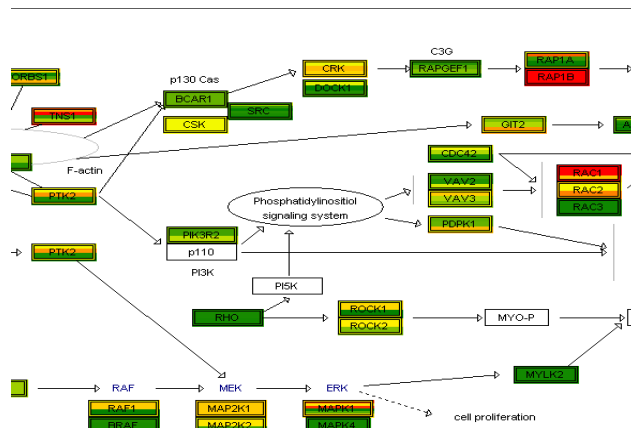
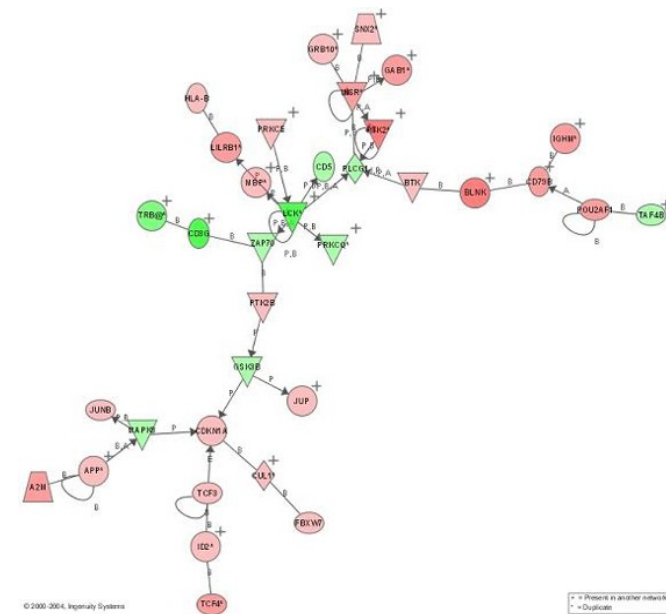
Transkripční faktory, microRNA motivy, publikace

Metody

Testy vyššího výskytu termínů mezi zvolenými geny (GSEA)

Sítě (Networks)

Dráhy (Pathways)



Analýza dat

Metaanalýza

Analýza více experimentů

Úložiště dat

GEO – gene expression omnibus (NCBI)

ArrayExpress (EBI)

Standardy

MAGE-ML – formát dat (jazyk)

MIAME – Minimum Information About Microarray Experiment

Informace umožňuje zopakování experimentu

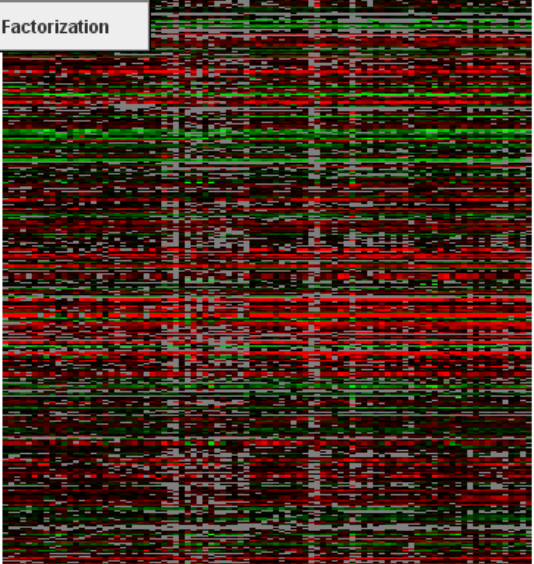
Nedostatek informací o vzorcích

Porovnatelnost platforem, laboratoří

Multiple Array Viewer

File Adjust Data Metrics Analysis Display Utilities

Clustering	Statistics	Classification	Data Reduction	Meta Analysis	Visualization	Miscellaneous
HCL Hierarchical Clustering	PTM Pavlidis Template Matching	SVM Support Vector Machines	RN Relevance Networks	GSEA Gene set enrichment analysis	LEM Linear Expression Map	GSH Gene Shaving
TEASE Tree EASE	TTEST t Tests	USC Uncorrelated Shrunk Cent	PCA Principal Component Analysis	EASE EASE Cluster Analysis	GDM Gene Distance Matrix	BN Bayesian Network
ST HCL Support Trees	BRIDGE BRIDGE	KNN K-Nearest Neighbors Classifi	COA Correspondence Analysis			LM Literature Mining
SOTA Self Organizing Tree Algorithm	SAM Significance Analysis for Micro	DAM Discriminant Analysis Classi	TRN Expression Terrain Map			
KMC k-Means/Medians Clustering	ANOVA One-way ANOVA					
KMS KMC Support	2-Fact. ANOVA Two-factor ANOVA					
CAST Cluster Affinity Search Techn	Nonpar Nonparametric Tests					
FOM Figure of Merit	BETR Bayesian Estimation of Temporal Regulation					
QTC QT Cluster	LIMMA Linear Models for Microarray Data					
SOM Self Organizing Map	RP Rank Products					
NMF Non-Negative Matrix Factorization						



MultiExperiment Viewer