Klasifikace genomických dat s bohatou apriorní znalostí

Jiří Kléma

Katedra počítačů,
FEL, ČVUT v Praze

Seminář strojového učení a modelování, MFF UK, 10.10.2013

Osnova přednášky

- Podstata strojového učení z příkladů
  - role apriorní znalosti,
- genové čipy (DNA microarrays)
  - širší kontext molekulární biologie,
  - genová exprese, centrální dogma molekulární biologie,
  - technologie, datový výstup,
- analýza a klasifikace dat z genových čipů
  - příklady úspěšných studií,
  - slabiny přímočarých řešení vycházejících pouze z měření,
    - dostupná apriorní znalost a možnosti jejího použití,
  - **využití množin genů pro klasifikaci**
    - jak apriorních tak indukovaných.
- možnosti integrace mRNA a miRNA dat pro predikci fenotypu
  - využití známých interakcí mezi mRNA a miRNA,
  - tři metody: subrakce, maticový rozklad, složený klasifikátor.

Učení z příkladů – klasifikace

- Předpokládáme existenci:
  - prostoru instancí X, obvykle vektory příznaků,
  - stavového prostoru Y, obvykle $Y \subset \mathbb{R}$, při klasifikaci kategoriální či $Y = \{0, 1\}$,
  - sdruženého rozdělení pravděpodobnosti P_{XY} na $X \times Y$.
- Pro učení je k dispozici:
  - (multi)množina trénovacích příkladů

$$T = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\}$$

  - příklady nezávisle vybírány z P_{XY}.
- Cílem klasifikace je:
  - z třídy funkcí $\mathcal{F} \subseteq \{f \mid f : X \to Y\}$
  - nalézt takové f, aby platilo

$$f(x) = \arg\max_{y \in Y} P_{Y|X}(y|\mathbf{x})$$

Učení z příkladů – klasifikace

- Obvyklý postup spočívá v minimalizaci:

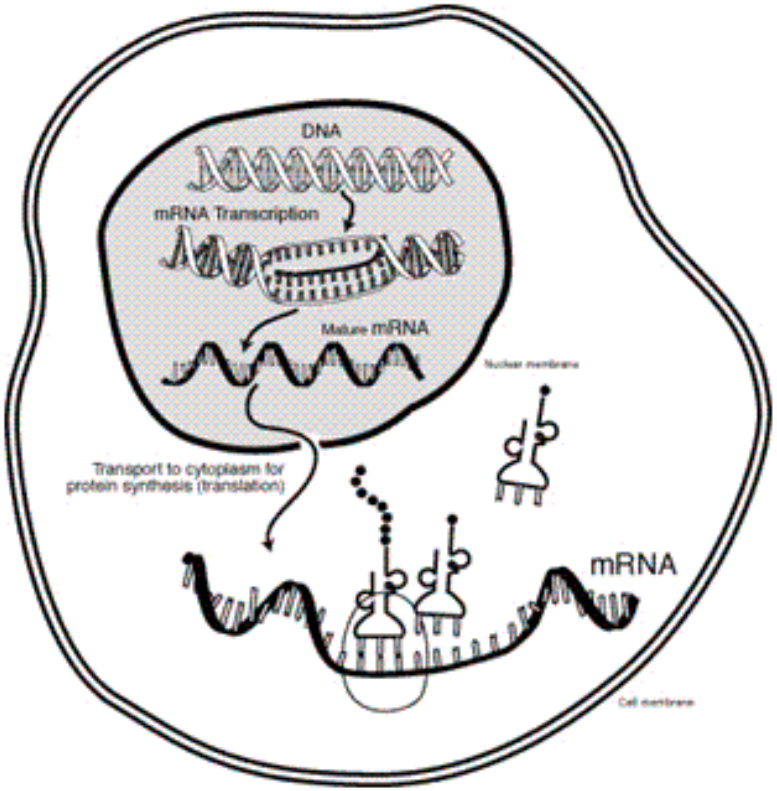
$$Cost(f) = \frac{1}{n} \sum_{\{\mathbf{x_i}, y_i\} \in T} I(y_i \neq f(\mathbf{x_i})) + \lambda \rho(f)$$

  - I je indikátorová funkce implementující 0-1 ztrátovou funkci,
  - $\rho(f)$ je regularizační term definující **volné zaujetí**,
  - $\lambda \in R$ je převodní poměr mezi ztrátou a regularizací,

- zaujetí a jeho role v učení
  - upřednostní některé funkce před jinými nezávisle na jejich empirickém hodnocení,
  - nutnou součástí každého algoritmu učení a podmínkou generalizační schopnosti,
  - brání přeučení,
  - $\mathcal{F}$ definuje **pevné zaujetí**.

- jak zaujetí definovat
  - na základě **apriorní znalosti**,
  - veškerá informace o daném problému nad rámec vlastních trénovacích dat,
  - konkrétní $\times$ obecná (Occamova břitva, hladkost, apod.).

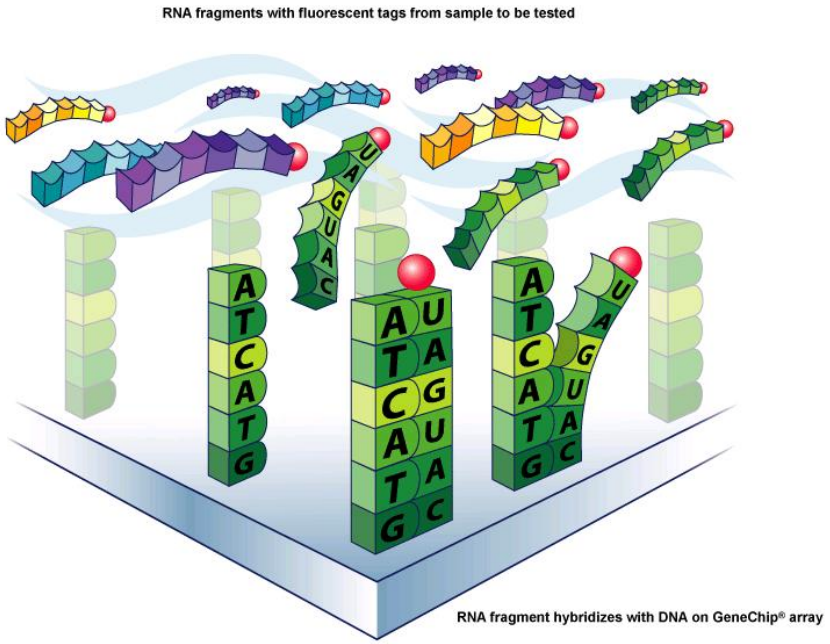
Centrální dogma molekulární biologie



National Human Genome Research Institute Genetic Illustrations

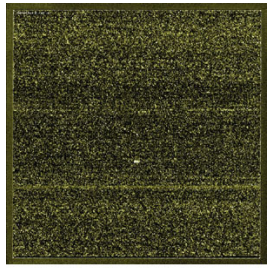
- genová exprese – proces, kterým je v genu uložená informace převedena v reálně existující buněčnou strukturu nebo funkci.

Genové čipy (DNA microarrays)



princip

čip



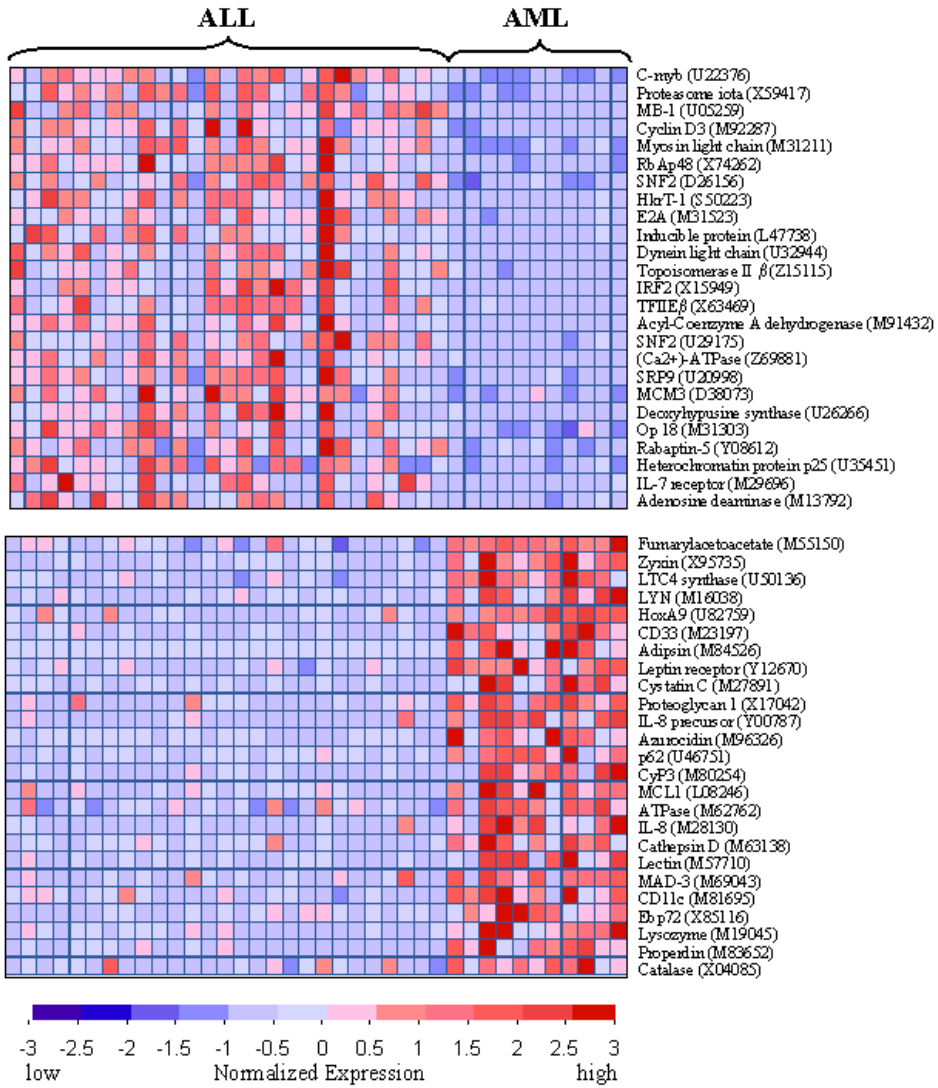
výsledná data

- dichotomie v použitelnosti a dostupnosti měření
  - měříme tam, kde to je nejsnazší, biologická relevance k fenotypu je slabší,
  - nové technologie – RNA-Seq.

Molekulární klasifikace leukémie – úspěšná studie

- tradiční klasifikace dle morfologie nádoru a dalších charakteristických znaků
  - přesto existují nerozlišitelné typy nádorů s odlišnou reakcí na léčbu,
- jednou z prvních studií molekulárního přístupu je [Golub et al., Science, 1999]
  - klasifikace na základě dat genové exprese (∼7000 genů, 38 vzorků, ALL a AML leukémie),
  - z pohledu strojového učení klasická úloha, zvolen intuitivní a dedikovaný postup
    - výběr 50 genů prokazatelně korelujících s fenotypem, tj. třídou vzorků,
    - následně každý z genů u testovacích vzorků rozhoduje o třídě,
    - konečná predikce váží volby genů,
    - k testování použito 34 nezávislých vzorků,
  - přínos hlavně jako aplikace na microarrays a biologické důsledky,
  - navíc navrženy nové podtypy leukémie na základě SOM shlukování
    - vedle class prediction i class discovery,
- paralelní studie, mj. [Alizadeh et al., Nature, 2000],
- postup není rutinně aplikovatelný na libovolná data genové exprese.

Molekulární klasifikace leukémie – úspěšná studie



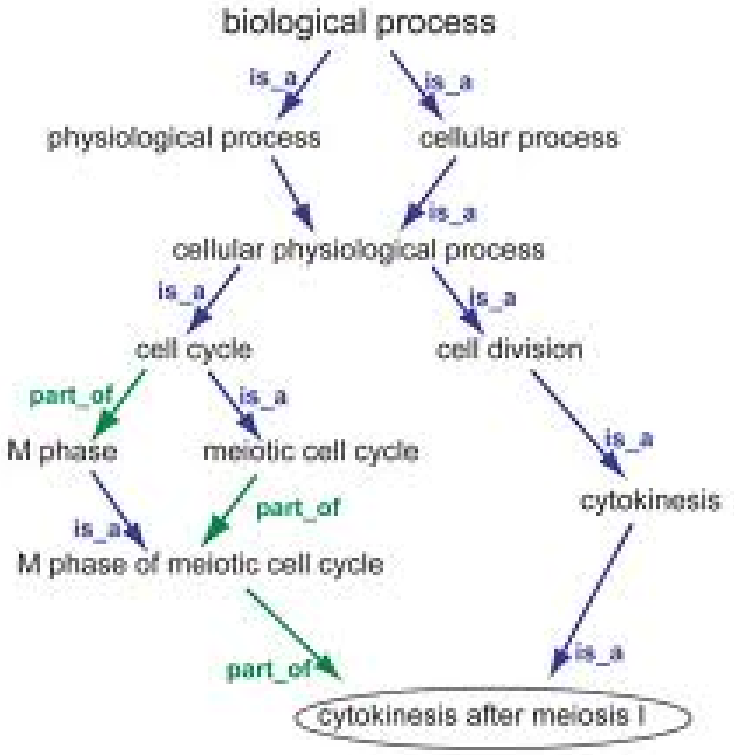
Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.

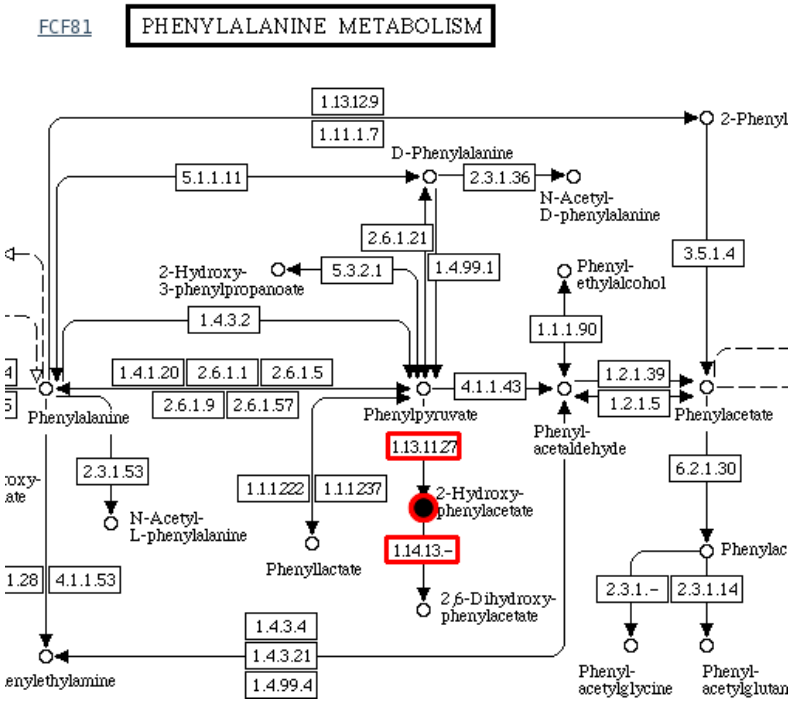
Motivace pro použití apriorní znalosti při klasifikaci dat genové exprese

- obvyklé vlastnosti dat genové exprese
  - šum, všeobecně nízká kvalita měření,
  - malá velikost vzorku, celogenomové měření s vysokým počtem příznaků/genů,
  - náchylnost k přeučení,
  - malá stabilita a interpretabilita výsledných modelů,
- k čemu může přispět znalost biologických procesů = množin genů
  - uvažujeme extrakci příznaků/agragaci odpovídajících procesům,
  - agregace může vést k filtorvání šumu,
  - před učením dojde k transformaci na prostor procesů $\rightarrow$ redukce dimenzionality,
  - vzroste zaujetí, pouze známé procesy mohou být použity,
  - stabilita a srozumitelnost vzroste.

GE data a dostupná apriorní znalost

- genové ontologie, metabolické a signální dráhy, transkripční faktory,





Enrichment analysis – biomarkery v datech genové exprese

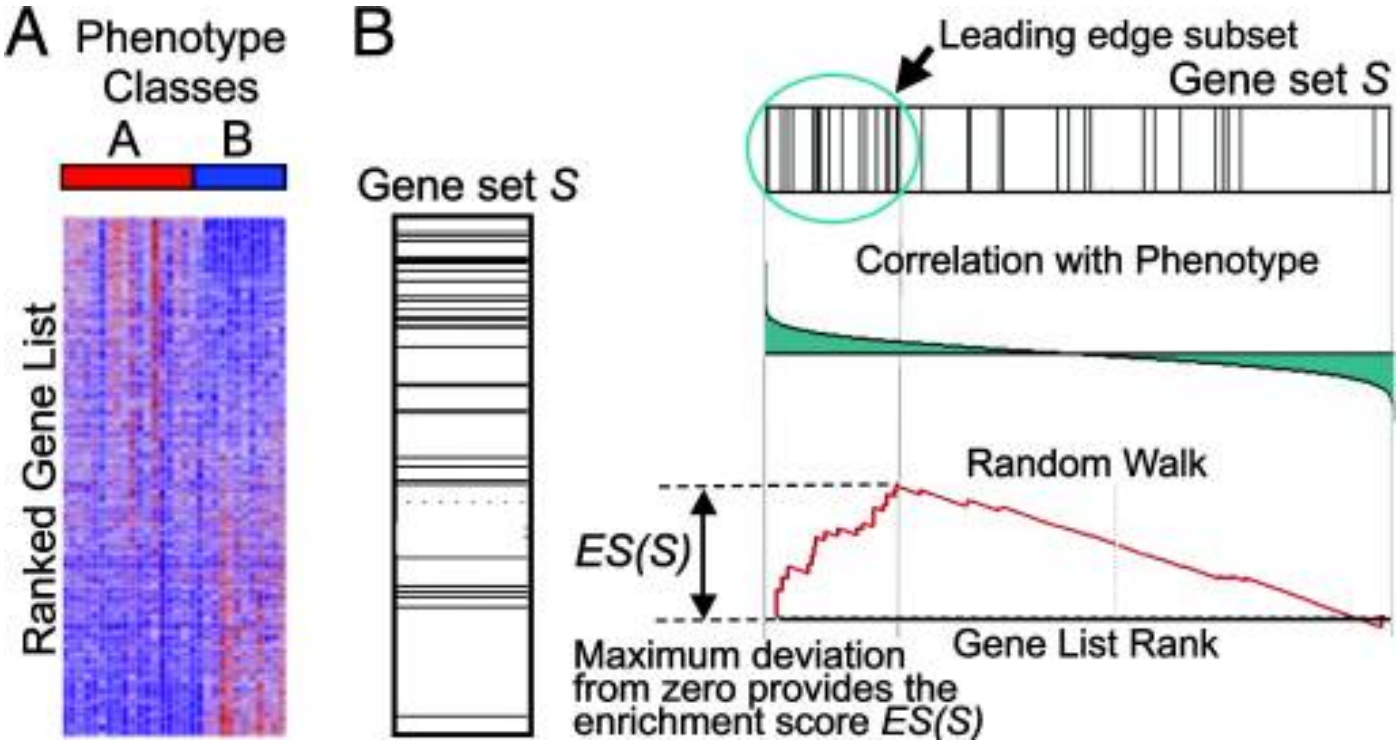
- výběr obohacených množin genů – existující statistický postup
  1. použij t-test (nebo jiný) k vyhledání diferenciálně exprimovaných genů,
  2. zvol prahovou hodnotu a vyber exprimované geny
     - volbou hladiny významnosti ($p \leq \alpha = 0.05$),
     - uspořádáním genů dle p-hodnot a výběrem k prvních,
  3. sestav čtyřpolní tabulku,
  4. proveď test nezávislosti mezi významností a příslušností k množině genů
     - χ^2 resp. Fisherův test,

	Differentially expressed gene	Non-differentially expressed gene	Total
In gene set	m_{GD}	m_{GD^c}	m_G
Not in gene set	m_{G^cD}	$m_{G^cD^c}$	m_{G^c}
Total	m_D	m_{D^c}	m

- nevýhody: ostrý práh, nedostatečná citlivost v případě slabé interakce velkého počtu genů.

Enrichment analysis – biomarkery v datech genové exprese

- výběr obohacených množin genů – dedikovaná metoda
  - Gene Set Enrichment Analysis (GSEA)

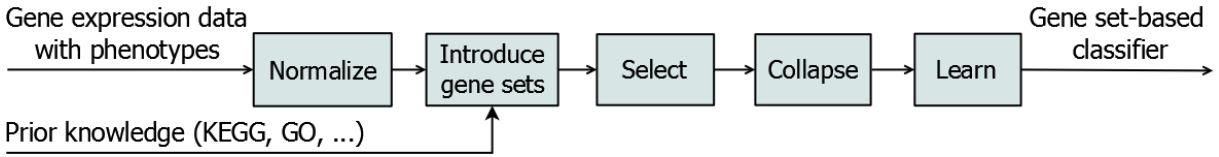


Subramanian et al: GSEA: a knowledge-based approach for interpreting genome-wide expression profiles.

  - Significance Analysis of Microarray for Gene Sets (SAM-GS), Global Test.

Klasifikace GE dat s dostupnou apriorní znalostí

- důležité otázky syntézy
  - jak **množiny** genů vytvářet
    - ∗ původ, optimální kardinalita,
  - jak počítat jejich aktivitu
    - ∗ funkce signatury množiny genů
      (nevážená, vážená, využívající topologii, založená na optimalizaci),
  - jak vybírat optimální množinu odvozených příznaků
    - ∗ výběr/řazení složitější než u původních příznaků,



- důležité biologické otázky následné analýzy
  - jaké problémy volit pro testování,
  - jak data normalizovat,
  - jak korektně převádět identifikátory mezi anotačními databázemi,
  - jak hromadně hodnotit srozumitelnost modelů.

BMC Bioinformatics

Comparative evaluation of set-level techniques in predictive classification of gene expression samples

Matěj Holec[1], Jiří Kléma[1*], Filip Železný[1] and Jakub Tolar[2]

Open Access

Analyzed factors	*Alternatives*	*#Alts*
1. Gene sets (Sec.)	Genuine, Random	2
2. Ranking algo (Sec.)	GSEA, SAM-GS, Global	3
*3. Set(s) forming features**	$1, 2, \ldots 10$, $n-9, n-8, \ldots n$, 1:10, $n-9:n$	22
4. Aggregation (Sec.)	SVD, AVG, SetSig, None	4
Product		528

Auxiliary factors	*Alternatives*	*#Alts*
5. Learning algo (Sec.)	svm, 1-nn, 3-nn, nb, dt	5
6. Dataset (Sec.)	$d_1 \ldots d_{30}$	30
7. Testing Fold	$f_1 \ldots f_{10}$	10
Product		1500

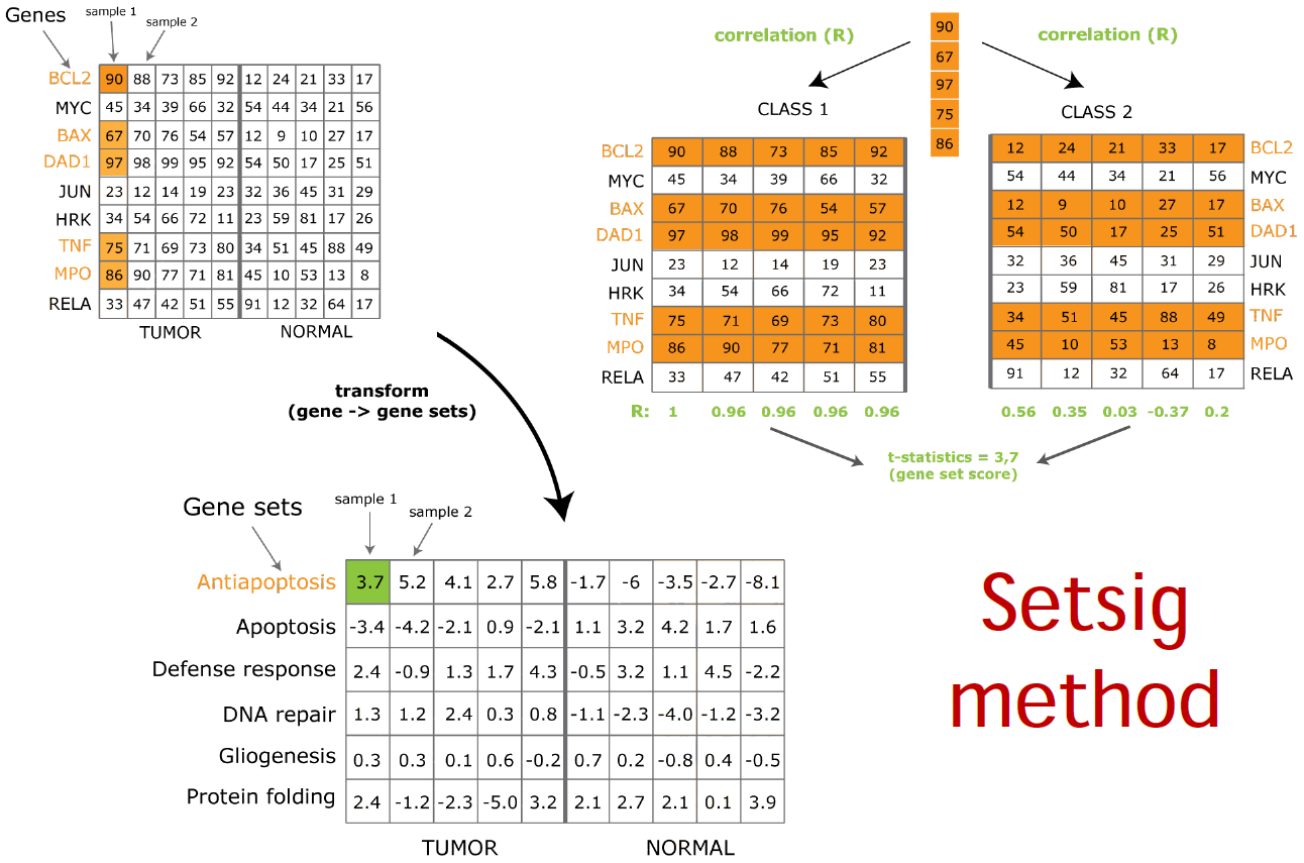
BMC Bioinformatics

Comparative evaluation of set-level techniques in predictive classification of gene expression samples

Matěj Holec[1], Jiří Kléma[1*], Filip Železný[1] and Jakub Tolar[2]

Open Access

- výpočet aktivity množin genů – metageny
  - průměrování, analýza hlavních komponent, bez agregace.



Mramor: On utility of gene set signatures in gene expression-based class prediction.

- závěry studie (výsledky statistických testů)
  - empiricky ověřené předpoklady
    - apriorní množiny genů překonávají náhodné,
      - zejména menší množiny a ty reprezentující chemické a genetické poruchy,
    - metody selekce fungují rozumně, množiny s nižším indexem překonávají ty s vyšším,
      - výchozí = použij všechny geny, tendence k přeučení,
    - použití 10 množin je výhodnější než použití jediné nejlepší,
  - biologicky zajímavé závěry
    - Global test překonává GSEA a SAM-GS,
    - SVD a SetSig překonává průměrování,
    - optimální množinový pracovní tok jednoznačně překonává výchozí genový přístup
      - výchozí = použij všechny geny, tendence k přeučení,
    - po zařazení selekce příznaků jsou co do přesnosti srovnatelné
      - informační zisk a SVM-RFE,
      - počty genů 22 a 228 odpovídají průměrnému počtu unikátních genů v 1 a 10 množinách.

IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
May-June 2012 (vol. 9 no. 3) pp. 788-798

Empirical Evidence of the Applicability of Functional Clustering through Gene Expression Classification

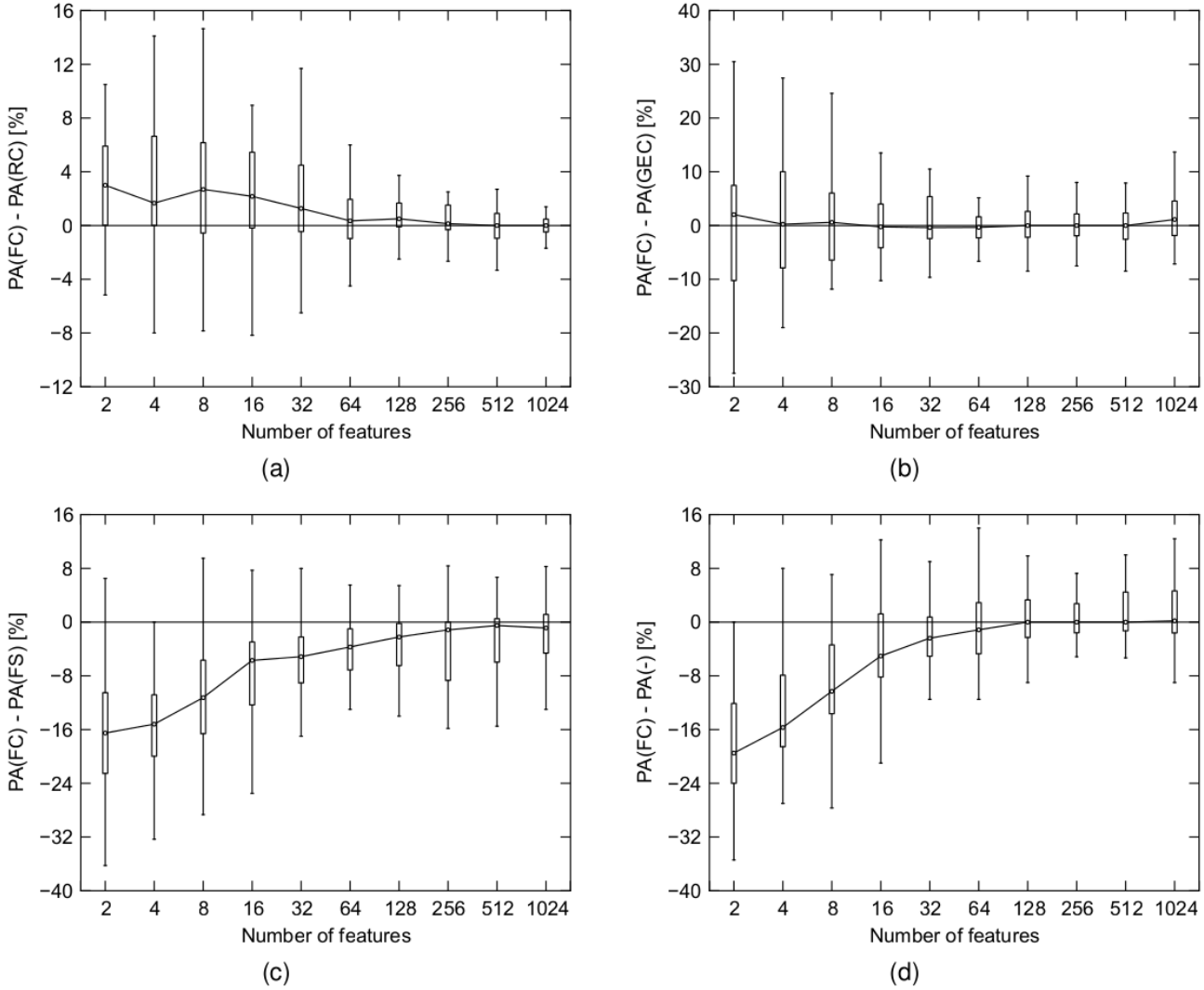
Miloš Krejník, **Jiří Kléma**, Czech Technical University, Prague

- opět klasifikace GE dat založená na množinách genů
  - množiny genů jsou vytvářeny na základě **shlukování**,
    - vedle tradičních metod jako k-means nebo k-medoids i fuzzy shlukování,
      - odráží multifunkční povahu genů,
    - definice funkční podobnosti genů vychází z nástroje pro funkční anotaci DAVID
      - vychází z binárních anotačních vektorů genů, více shodných pojmů = větší podobnost,
  - srovnává a kombinuje shlukování
    - náhodně vytvářené rozklady genomu (RC),
    - založené čistě na datech genové exprese (GEC),
    - založené čistě na genových anotacích (FC),
    - kombinující oba vstupy, tedy GE i anotace (FCi).
- menší důraz na absolutní klasifikační přesnost, důležitá vzájemná porovnání.
  - aktivita shluků určena průměrováním, resp. jako medoid shluku,

IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Empirical Evidence of the Applicability of Functional Clustering through Gene Expression Classification

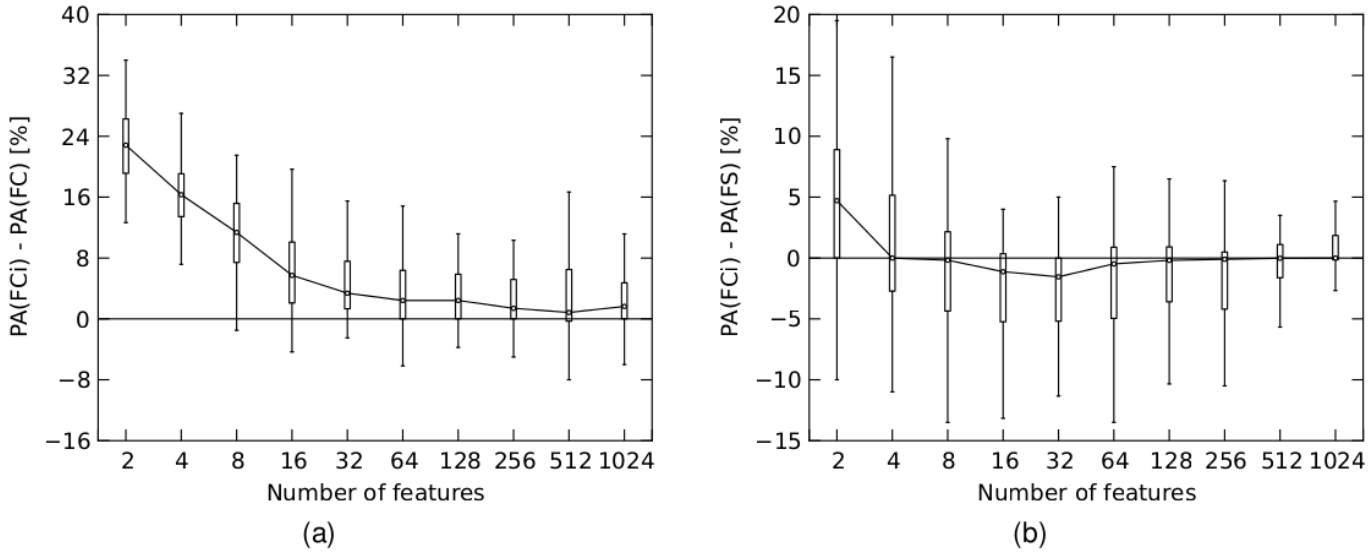
May-June 2012 (vol. 9 no. 3) pp. 788-798 **Miloš Krejník**, **Jiří Kléma**, Czech Technical University, Prague



IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
May-June 2012 (vol. 9 no. 3) pp. 788-798

Empirical Evidence of the Applicability of Functional Clustering through Gene Expression Classification

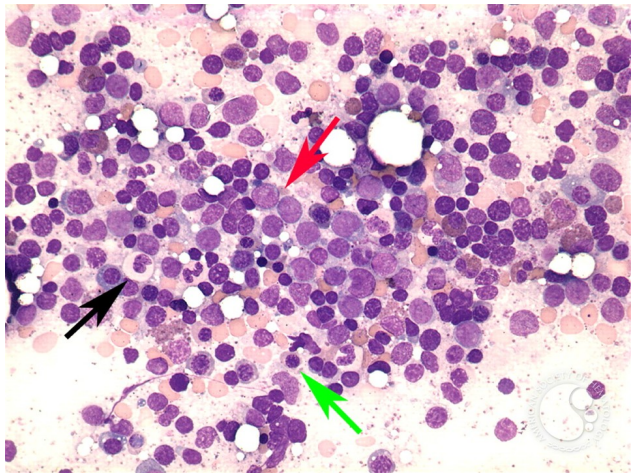
Miloš Krejník, **Jiří Kléma**, Czech Technical University, Prague



- FC je lepší než RC, ale rozdíl je zejména u menšího počtu shluků,
- FC je prediktivním výkonem srovnatelné s GEC
  - je třeba vzít v úvahu, že je nezávislé na konkrétních datech a tedy univerzální,
- po sloučení do FCi je shlukování kompetitivní s FS.

Integrace mRNA, miRNA a metylačních dat

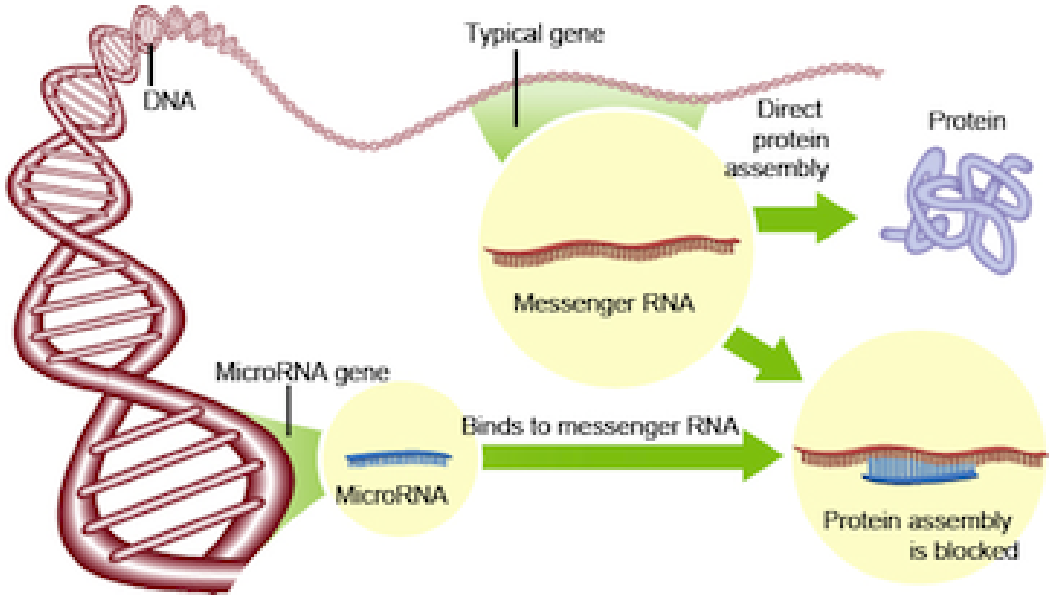
- projekty IGA MZ 2013-2015
  - XGENE.ORG – nástroj integrované analýzy transkripčních, miRNA a metylačních dat,
  - Predikce odpovědi na demetylační léčbu u pacientů s myelodysplastickým syndromem s využitím integrativní genomiky,
  - řešeny ve spolupráci s Ústavem hematologie a krevní transfúze.
- myelodysplastický syndrom (MDS)
  - krevní onemocnění (jejich rodina),
  - poruchy vývoje a vyzrávání krevních buněk,
  - s potenciálním přechodem v leukémii,
  - projevy: abnormality v kostní dřeni, chromozomální aberace, únava, dušnost, nevolnost, krvácivé projevy, náchylnost k infekcím,
  - výskyt častější ve vyšším věku.



The American Society of Hematology Image Bank

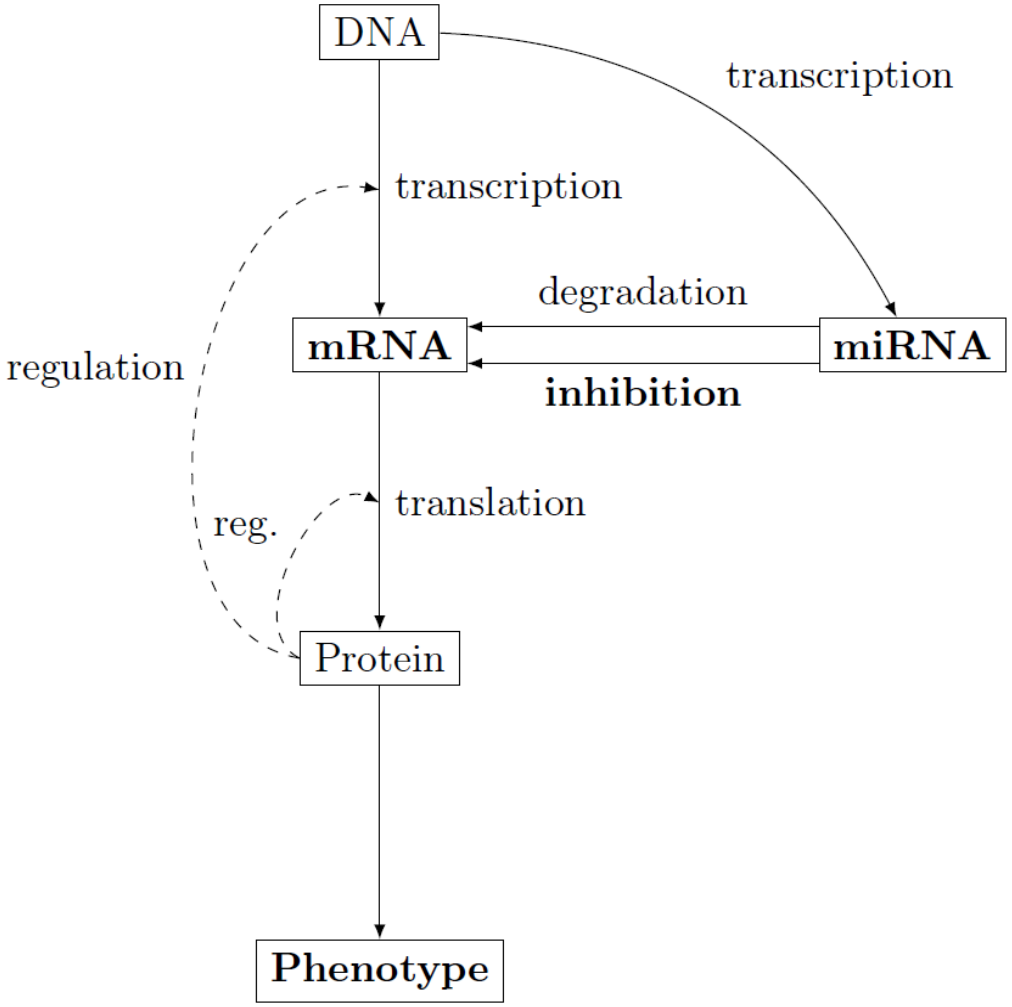
miRNA resp. microRNA data

- jednovláknové řetězce nekódující RNA o délce 21-23 nukleotidů,
- podílejí na regulaci genové exprese,
- regulují výrobu proteinů kódovaných k nim komplementárním mRNA.
- komplementarita predikována nebo validována, my ji pokládáme za danou m-n relaci.



Cox: The Art of the PFUG

Možnosti integrace mRNA a miRNA dat pro predikci fenotypu



Zjednodušené schéma genové exprese

Subtraktivní agregace mRNA a miRNA dat

- vychází z představy inhibice a degradace příslušných mRNA regulující miRNA,
- vzhledem k (obvykle) malému počtu vzorků minimalizuje počet parametrů (pouze c),
- SubAgg:

$$x_g^{sub} = x_g^G - \frac{c}{|\mathcal{R}_g|} \sum_{r \in \mathcal{R}_g} \frac{x_g^G}{\sum_{t \in \mathcal{G}_r} x_t^G} x_r^\mu$$

  x_g^G naměřené množství mRNA pro gen g, $\mathcal{R}_g$ množina miRNA cílících na g,
  x_r^μ naměřená exprese miRNA sekvence r, $\mathcal{G}_r$ množina cílových genů pro miRNA r,

- měřítkem kvality klasifikační přesnost dosažená s danou množinou příznaků
  - SubAgg srovnávána s mRNA a miRNA expresemi a jejich prostým sloučením (merge),
  - klasifikační pracovní tok: agregace, sloučení, selekce příznaků (konstantní velikost), SVM/NB.

:: Kléma, Zahálka, Anděl, Krejčík: *Knowledge-Based Subtractive Integration of mRNA and miRNA Expression Profiles to Differentiate Myelodysplastic Syndrome.* Submitted to Bioinformatics 2014.

Subtraktivní agregace mRNA a miRNA dat vycházející z SVD

- lze i z malého vzorku "učit" rozdílnou aktivitu jednotlivých miRNA vzhledem k dané mRNA?
  - nad rámec p-hodnot u predikovaných mRNA-miRNA interakcí,
- prostor exprese všech miRNA cílících na danou mRNA před subtrakcí redukován
  - projekcí na svůj první singulární vektor,
- SVDAgg:

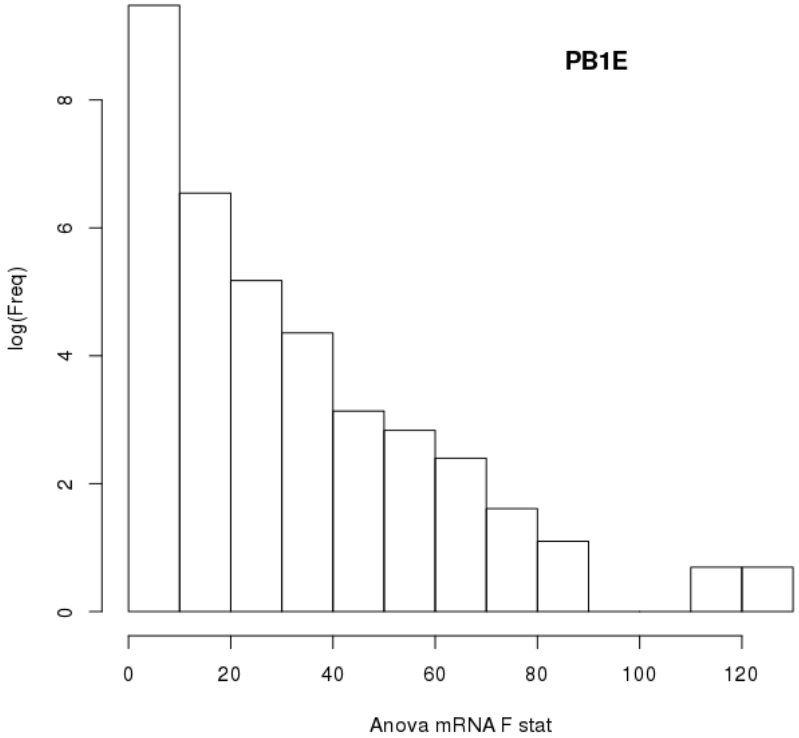
$$\mathbf{x}^{\mu,\mathbf{svd}} = \mathbf{X}^{\mu}_{1\ldots s,\mathcal{R}_g}\mathbf{V}_{1\ldots|R_g|,1} \quad \mathbf{x}^{G,svd}_{g} = \left[\mathbf{x}^{G}_{g}, \mathbf{x}^{\mu,svd}\right]\mathbf{U}_{1\ldots 2,1}$$

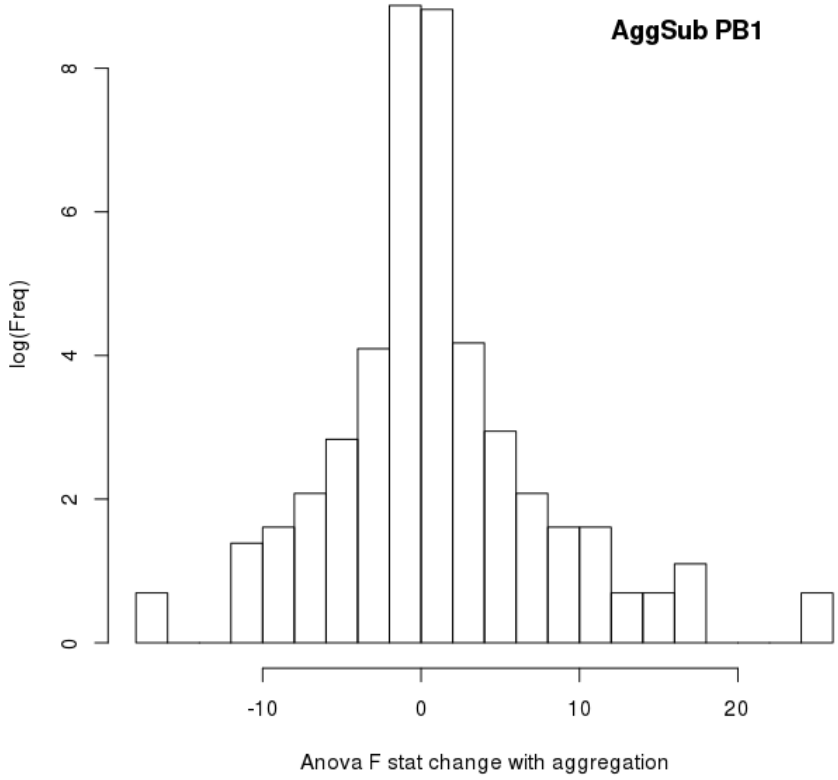
  $\mathbf{V}$ the singular vector matrix of targeting miRNAs,
  $\mathbf{U}$ the first singular vector of the mRNa and miRNA concatenated vectors,
- měřítkem kvality klasifikační přesnost dosažená s danou množinou příznaků
  - SVDAgg srovnávána s mRNA a miRNA expresemi a jejich prostým sloučením (merge),
  - klasifikační pracovní tok: agregace, sloučení, selekce příznaků (konstantní velikost), SVM/NB.

:: Kléma, Zahálka, Anděl, Krejčík: *Knowledge-Based Subtractive Integration of mRNA and miRNA Expression Profiles to Differentiate Myelodysplastic Syndrome.* Submitted to Bioinformatics 2014.

Subtraktivní agregace mRNA a miRNA dat – slučování a selekce

- slučování příznaků – agregované příznaky doplňují a nenahrazují ty původní,
- selekce příznaků – o jejich použití rozhodne až SVM-RFE,
- robustní vzhledem k odlišné prediktivní síle mRNA a miRNA profilů napříč úlohami.

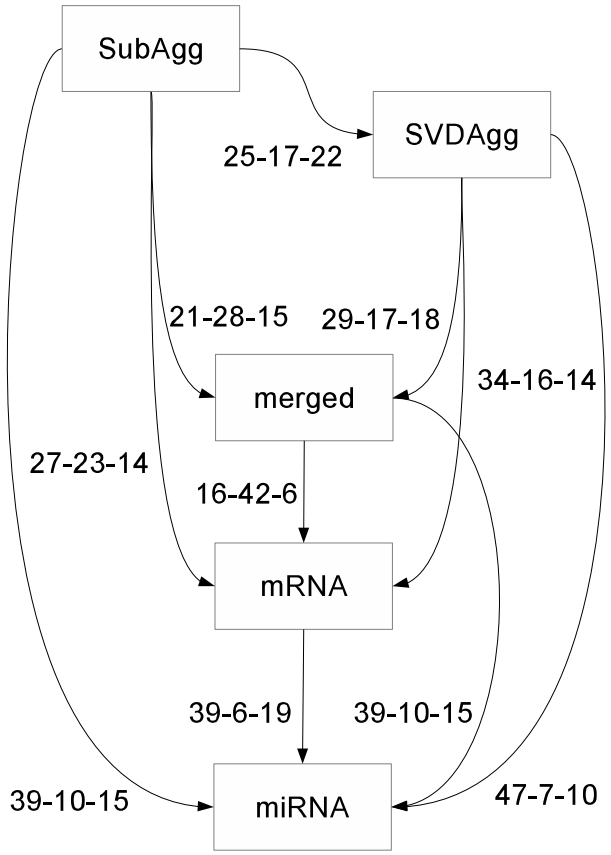




Subtraktivní agregace mRNA a miRNA dat – výsledky

T	#	R	Feature set				
			mR	miR	mer	Sub	SVD
PB	1	E	0.990	0.873	0.990	0.995	0.990
PB	2	E	0.965	0.930	0.965	0.965	0.991
PB	3	E	1.000	0.947	0.992	1.000	1.000
PB	4	E	1.000	0.879	1.000	1.000	1.000
PB	5	E	0.935	0.968	0.968	0.974	0.968
PB	6	E	1.000	0.974	1.000	1.000	1.000
PB	7	E	0.914	0.871	0.914	1.000	1.000
PB	8	E	0.864	0.845	0.818	0.800	0.818
BM	1	E	0.965	0.971	0.971	0.982	0.988
BM	2	E	0.963	0.993	0.963	0.951	0.963
BM	3	E	1.000	1.000	1.000	1.000	1.000
BM	4	E	1.000	0.894	1.000	1.000	1.000
BM	5	E	0.908	0.950	0.908	0.958	0.908
BM	6	E	0.933	0.933	0.933	0.933	0.933
BM	7	E	0.975	1.000	0.988	1.000	0.950
BM	8	E	0.763	0.850	0.763	0.825	0.825
PB	1	V	0.976	0.907	0.976	0.976	0.980
PB	2	V	0.939	0.817	0.939	0.939	0.939
PB	3	V	1.000	0.853	1.000	0.979	1.000
PB	4	V	1.000	0.893	1.000	1.000	1.000
PB	5	V	0.935	0.968	0.961	0.968	0.968
PB	6	V	1.000	0.957	1.000	0.991	1.000
PB	7	V	0.914	0.657	0.857	0.929	0.986
PB	8	V	0.864	0.827	0.855	0.836	0.818
BM	1	V	0.988	0.935	0.988	0.982	0.976
BM	2	V	0.978	0.993	0.993	0.985	0.993
BM	3	V	1.000	0.952	1.000	1.000	1.000
BM	4	V	1.000	0.894	1.000	1.000	1.000
BM	5	V	0.908	0.958	0.908	0.917	0.958
BM	6	V	0.933	0.933	0.933	0.947	0.933
BM	7	V	0.938	0.938	0.938	1.000	0.938
BM	8	V	0.775	0.813	0.825	0.825	0.788
Avg. ranking			3.14	3.73	2.91	2.58	2.64

Absolutní SVM přesnosti přes různá nastavení



Párová porovnání (+|0|-)

Agregace mRNA a miRNA dat založená na maticovém rozkladu

- NMF: non-negative matrix factorization

$$\mathbf{X} \in \mathbb{R}^{N \times M} \quad \rightarrow \quad \mathbf{X} \approx \mathbf{WH}, \quad \mathbf{W} \in \mathbb{R}^{N \times K}, \quad \mathbf{H} \in \mathbb{R}^{K \times M}, \quad K \ll N, M$$

- optimalizační kritérium

$$min_{\mathbf{WH}} \left(||\mathbf{X} - \mathbf{WH}||_F^2 + \gamma(||\mathbf{W}||_F^2 + ||\mathbf{H}||_F^2) \right)$$

- motivace: doporučovací systémy, obrovské řídké databáze typu Netflix
  - N uživatelů, M filmů, $\mathbf{X}$ uživatelské preference,
  - $\mathbf{W}$ váhová matice přiřazující uživatele k jisté (skryté) podskupině filmů (žánru),
  - $\mathbf{H}$ váhová matice přiřazující filmy k jisté (skryté) podskupině filmů,
  - γ regularizační parametr,
  - doporučení: příslušná položka $\mathbf{WH}$ nahrazující NA hodnotu v $\mathbf{X}$.

Agregace mRNA a miRNA dat založená na maticovém rozkladu

- analogie v doméně genové exprese
  - vzorky $\sim$ uživatelé, geny $\sim$ filmy,
  - množiny genů resp. biologické procesy odpovídají žánrům,
  - chtěli bychom pracovat s mRNA i miRNA, použít existující znalost (proteinové interakce, mRNA-miRNA regulaci),
- sparse network-regularized multiple non-negative matrix factorization (SNMNMF),
  - komplikovanější varianta NMF: dva typy měření, relační data, interakční data,
  - $\mathbf{X}^g \in \mathbb{R}^{N \times M^g}$: genová transkripční aktivita (mRNA), N vzorků a M^g transkriptů,
  - $\mathbf{X}^\mu \in \mathbb{R}^{N \times M^\mu}$: miRNA transkripční aktivita, M^μ různých miRNA,
  - $\mathbf{X}^g \approx \mathbf{W}\mathbf{H}^\mathbf{g}$ & $\mathbf{X}^\mu \approx \mathbf{W}\mathbf{H}^\mu$,
  - $\mathbf{W} \in \mathbb{R}^{N \times K}$: váha k-tého **komodulu** v n-tém vzorku,
  - $\mathbf{H}^g \in \mathbb{R}^{K \times M^g}$: soft přiřazení genů ke K regulačním komodulům,
  - $\mathbf{H}^\mu \in \mathbb{R}^{K \times M^\mu}$: soft přiřazení miRNA ke K regulačním komodulům.

:: Zhang, Li, Liu, Zhou: *A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules.* Bioinformatics, 27(13):401–409, 2011.

Agregace mRNA a miRNA dat založená na maticovém rozkladu

- apriorní znalost: transkripční faktory a proteinové interakce (komplexy), miRNA cíle

$$\mathbf{A} \in \mathbb{B}^{M^g \times M^g}, \mathbf{B} \in \mathbb{B}^{M^\mu \times M^g}$$

- SNMNMF optimalizační kritérium

$$\begin{aligned} &\left\|\mathbf{X}^g - \mathbf{W}\mathbf{H}^g\right\|_F^2 + \left\|\mathbf{X}^\mu - \mathbf{W}\mathbf{H}^\mu\right\|_F^2 \\ &-\lambda_g Tr\left(\mathbf{H}^g\mathbf{A}\mathbf{H}^{gT}\right) - \lambda_\mu Tr\left(\mathbf{H}^\mu\mathbf{B}\mathbf{H}^{gT}\right) \\ &+\gamma_1 \left\|\mathbf{W}\right\|_F^2 + \gamma_2 \left(\sum \left\|h_j^\mu\right\|_1^2 + \sum \left\|h_j^g\right\|_1^2\right), \end{aligned}$$

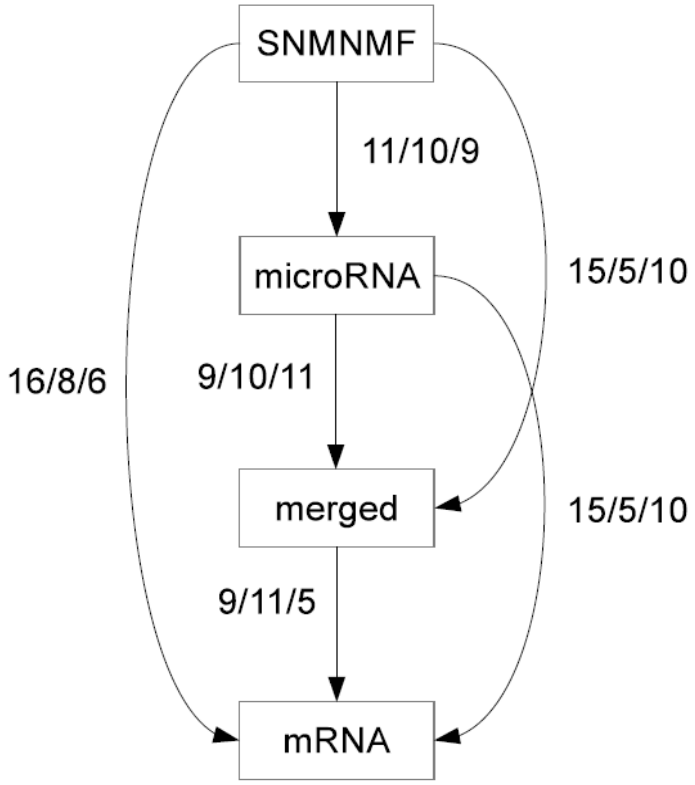
- použití pro klasifikaci
  - klasifikace v prostoru komodulů,
  - regularizační parametry $\lambda_g, \lambda_\mu, \gamma_1, \gamma_2$ optimalizovány křížovou validací,
  - testovací příklady lze projektovat do prostoru komodulů fixací $\mathbf{H}$.

:: Zhang, Li, Liu, Zhou: *A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules.* Bioinformatics, 27(13):401–409, 2011.

SNMNMF – výsledky klasifikace pro MDS

task	mRNA	miRNA	merged	SNMNMF
PB1	**0.76**	**0.76**	**0.76**	**0.76**
BM1	0.74	**0.77**	0.74	**0.77**
PB2	**0.90**	**0.77**	**0.90**	**0.90**
BM2	0.93	1.00	0.96	**1.00**
PB3	**1.00**	0.90	**1.00**	0.88
BM3	0.95	**1.00**	0.95	**1.00**
PB4	0.76	0.76	**0.80**	0.72
BM4	**0.95**	**0.95**	**0.95**	0.93
PB5	**0.71**	**0.71**	**0.71**	**0.71**
BM5	0.79	**0.87**	0.79	**0.87**

Absolutní SVM přesnosti přes různá nastavení



Párová porovnání (+|0|-)

:: Anděl, Kléma, Krejčík: *Integrating mRNA and miRNA expressions with interaction knowledge to predict myelodysplatic syndrome.* In ITAT 2013: Information Technologies – Applications and Theory, Workshop on Bioinformatics in Genomics and Proteomics, pp. 48-55, 2013.

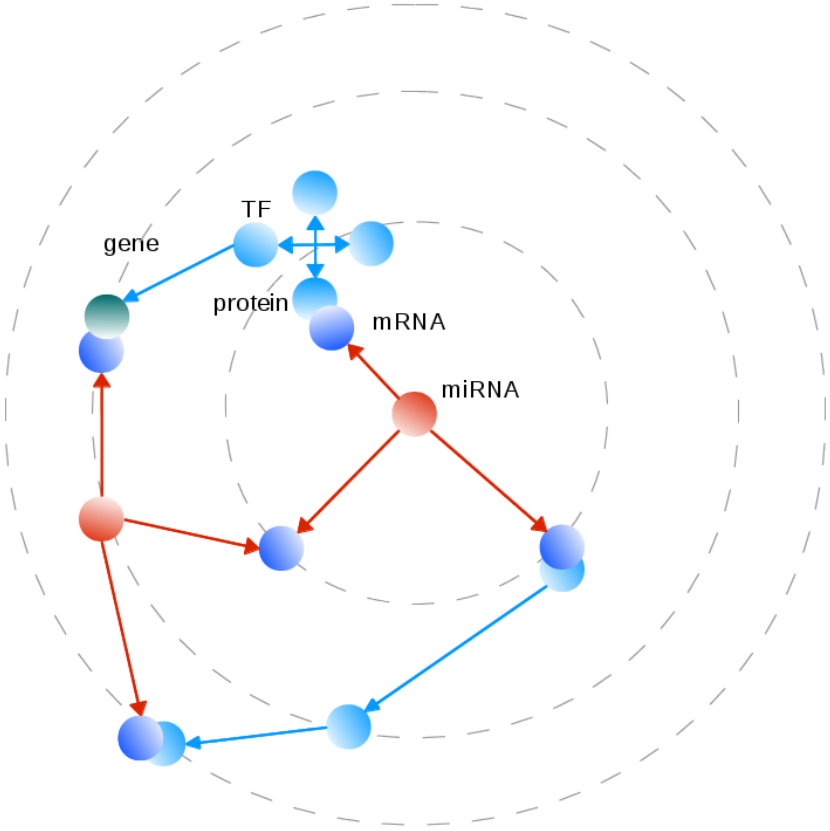
Klasifikace mRNA a miRNA profilů složeným klasifikátorem

- Složené klasifikátory
  - panel expertů, resp. chytrý dav,
  - nutná podmínka funkce: dekorelovanost, přesnost dílčích slabých klasifikátorů,
  - AdaBoost, random forest,
  - populární a úspěšná oblast strojového učení,
- Hypotéza o využití apriorní znalosti při snaze o dekorelaci slabých klasifikátorů
  - vztahy mezi mRNA, miRNA a proteiny buď implikují zvýšenou pst korelace,
  - nebo funkční souvislost, tj. latentní potenciálně silné prediktory na neměřené úrovni,
  - předchozí experimenty $\rightarrow$ nespolehlivé vztahy, ohraničené množiny příznaků umělé,
  - šlo by využít stochastických množin odpovídajících slabým klasifikátorům?
    - dekorelované a potenciálně prediktivní fragmenty procesů,

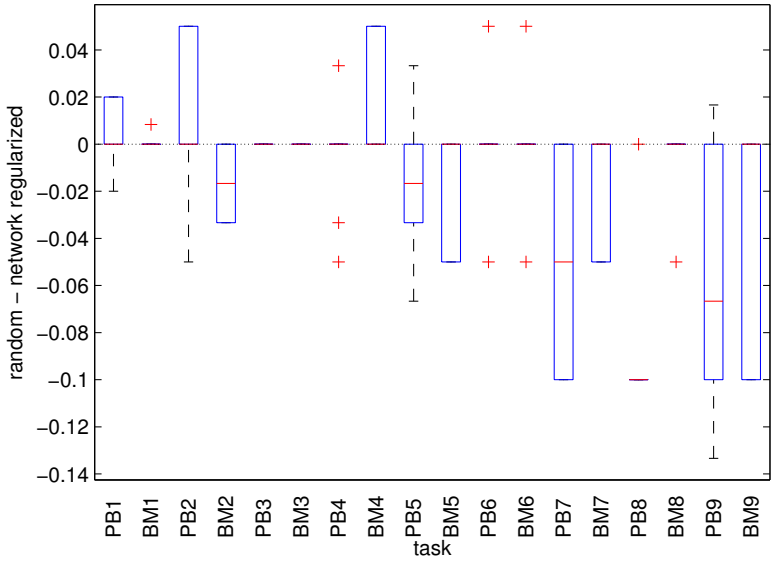
:: Dutkowski, Ideker: Protein networks as logic functions in development and cancer. PLoS Comput Biol 7, e1002180, 2011.

Klasifikace mRNA a miRNA profilů složeným klasifikátorem

- Možné schéma: náhodná procházka, pst zastavení 0.5, upřednostní blízké vrstvy před vzdálenými číslo vrstvy, pst výběru uzlu z dané vrstvy (medián), počet uzlů v dané vrstvě (medián)
  1, 0.65, 185; 2, 0.27, 1659; 3, 0.08, 12478; 4, 0,001, 466



Interakční síť pro tvorbu slabých klasifikátorů

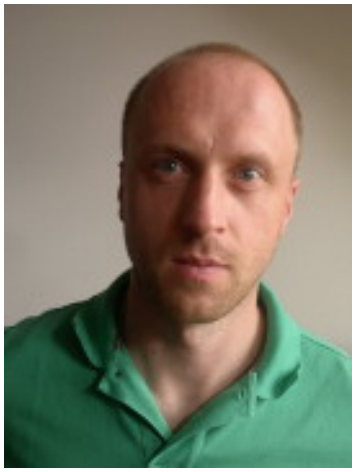


RF vs NWRF, předběžné výsledky

Shrnutí

- molekulární biologie je atraktivní pro strojové učení
  - velké objemy veřejných dat,
  - měření jsou často zašuměná, data neanotovaná a rozmanitá,
  - apriorní znalost je netriviální, strukturovaná a rozmanitá,
  - důležitý, zajímavý a živý obor,
  - prostor pro vznik nových aplikačně specifických ML algoritmů,
  - prostor pro novou aplikaci těch stávajících,
- aplikace strojového učení napomohla
  - obecně posoudit věrohodnost, významnost a možnosti použití apriorní znalosti,
  - lépe porozumět konkrétním nemocem (ALL/AML, MDS)
    - prostřednictvím výkladu komplexních příznaků získaných učením.

Spolupracovníci, spoluautoři



Matěj Holec

Miloš Krejník

Michael Anděl

Jan Zahálka