Bioinformatické aplikace strojového učení

Jiří Kléma

Katedra kybernetiky,
FEL, ČVUT v Praze

IDA
Intelligent Data Analysis
RESEARCH GROUP

Informační technologie - aplikace a teorie 2012

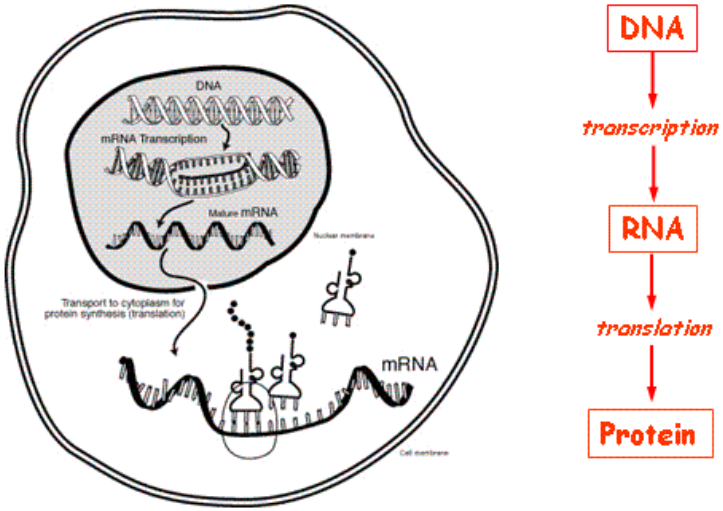
Přehled témat

- Bioinformatika a příbuzné obory
  - v čem se liší od biomedicínské informatiky?
  - možnosti pro aplikaci a vývoj metod strojového učení.
- Zajímavé a úspěšné bioinformatické
  - nástroje,
  - studie,
  - projekty.
- Výzkum ve skupině Inteligentní datové analýzy na FEL ČVUT
  - molekulární klasifikace využívající apriorní znalosti,
  - predikce schopnosti bílkovin vázat se na DNA,
  - současné projekty.

Co to je bioinformatika?

- Bioinformatika
  - reprezentace
  - shromažďování
  - vizualizace
  - vyhledávání
  - **analýza**

  } molekulárně biologických dat



National Human Genome Research Institute Genetic Illustrations

- Typické oblasti bioinformatického výzkumu
  - sekvenční analýza (sestavování, zarovnávání sekvencí),
  - anotace genomu (vyhledávání genů v DNA sekvenci),
  - evoluční biologie (stromy života, sdílení informací mezi druhy),
  - analýza genové a proteinové exprese,
  - predikce struktury bílkovin.

Příbuzné obory

- zjednodušené chápání **bioinformatika = počítače + biologie** není přesné,

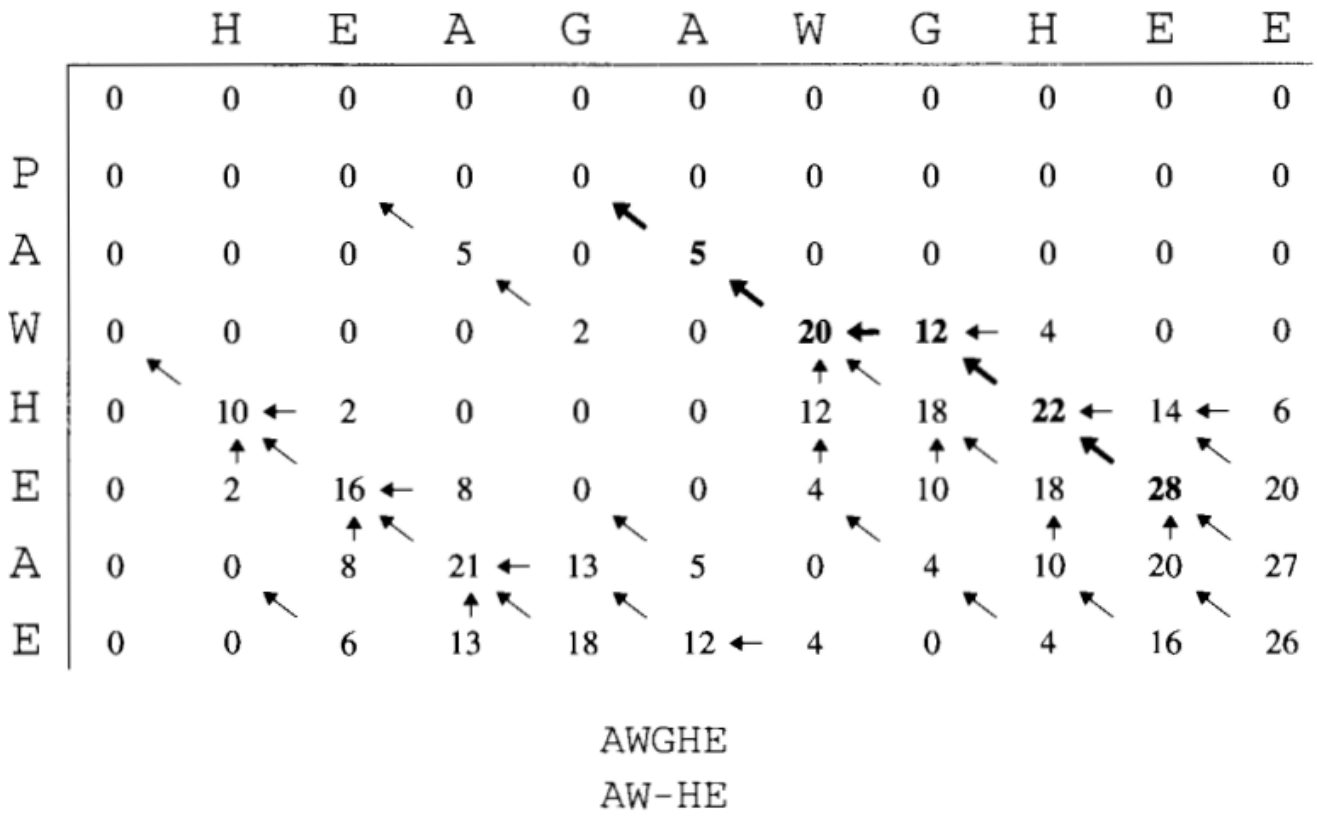
- biomedicínská informatika
  - obecnější, data mohou mít širší původ (EEG, obrázky apod.),
- přírodou inspirované výpočetní techniky
  - genetické algoritmy, mravenčí kolonie, DNA počítače, neuronové sítě,
- výpočetní biologie
  - zaměření na analýzu, modelování a simulace,
  - nemusí jít výhradně o molekulární data (neurovědy, sociální systémy apod.),
- systémová biologie
  - studium komplexních interakcí v biologických systémech, důraz na dynamiku dějů,
  - u genů mj. regulační sítě.

Aplikace strojového učení v bioinformatice – kategorizace

- kategorizace dle přehledového článku [Larranaga et al., 2005]
  - klasifikační problémy
    - anotace genomu (vyhledávání genů, DNA vazebních míst pro interakci s proteiny),
    - predikce funkce genů nebo sekundární struktury bílkovin,
    - klasifikace biologických vzorků (nemocní, zdraví apod.).
  - shlukovací problémy
    - tvorba fylogenetických stromů,
    - zjišťování funkční podobnosti genů z dat genové exprese,
  - pravděpodobnostní grafické modely
    - modelování DNA sekvencí (vyhledávání genů),
    - tvorba genových sítí v systémové biologii,
  - optimalizace
    - zarovnávání sekvencí.
    - zjednodušené modely zavinování proteinu.

BLAST – nejúspěšnější bioinformatický nástroj

- lokální zarovnávání sekvencí,
- problém optimálně řešitelný klasickým dynamickým programováním
  - Smith-Watermanův algoritmus pracující v $\mathcal{O}(nm)$.



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	**5**	0	0	0	0	0
W	0	0	0	0	2	0	**20**	**12**	4	0	0
H	0	10	2	0	0	0	12	18	**22**	14	6
E	0	2	16	8	0	0	4	10	18	**28**	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

AWGHE
AW-HE

Durbin et al.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids

BLAST – nejúspěšnější bioinformatický nástroj

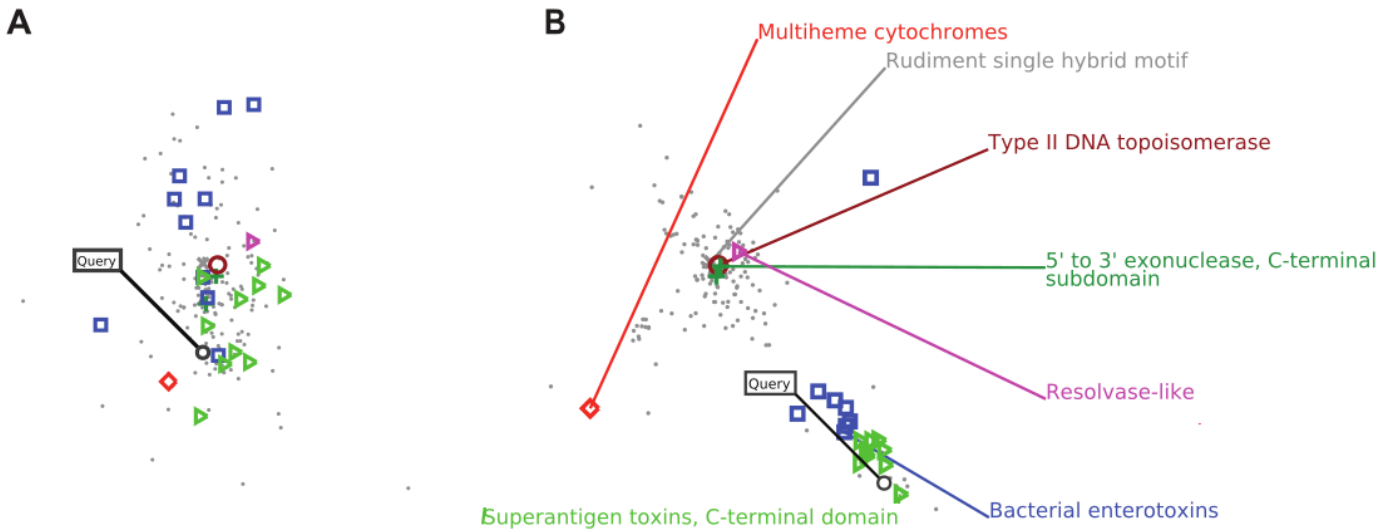
- Basic Local Alignment Search Tool (BLAST) [Altschul et al., 1997]
  - http://blast.ncbi.nlm.nih.gov,
  - obrovské využití: >500 tisíc dotazů denně,
  - velký vědecký dopad: článek o PSI-BLAST má >40 tisíc citací (Google Scholar),
- srovnání BLAST a Smith-Watermanova algoritmu
  - BLAST je heuristický, negarantuje optimální řešení,
  - je zhruba 10-50x rychlejší (zjištěno empiricky),
  - uvnitř pracovního toku BLAST omezeně využívá Smith-Watermanův algoritmus.
- ilustrativní běh pro nahodilou sekvenci: ACDEFGHIKLMNPQRSTVWY

Sequences producing significant alignments:

Accession	Description	Total score	Query coverage	E value	Max ident	Links
EFN87491.1	hypothetical protein EAI_07403 [Harpegnathos saltator]	32.0	60%	3.2	75%	
XP_002845759.1	predicted protein [Arthroderma otae CBS 113480] >gb|EEQ32809.1| predicted protein [Arthr	30.8	65%	9.0	69%	G
YP_004693596.1	transposase IS116/IS110/IS902 family protein [Nitrosomonas sp. Is79A3] >gb|AEJ00197.1| tr	30.8	55%	9.1	75%	G
XP_002008604.1	GI13587 [Drosophila mojavensis] >gb|EDW19080.1| GI13587 [Drosophila mojavensis]	30.8	40%	9.1	100%	G
ZP_10384399.1	YybS [Bacillus sp. 916] >gb|EJD68942.1| YybS [Bacillus sp. 916]	30.3	80%	12	65%	
EHM06687.1	YybS [Bacillus amyloliquefaciens IT-45]	30.3	80%	12	65%	
ZP_10045056.1	putative membrane protein (DUF2232) [Bacillus sp. 5B6] >gb|EIF15402.1| putative membra	30.3	80%	12	65%	
YP_005423273.1	hypothetical protein BANAU_3937 [Bacillus amyloliquefaciens subsp. plantarum YAU B9601	30.3	80%	12	65%	G
YP_005132496.1	hypothetical protein BACAU_3767 [Bacillus amyloliquefaciens subsp. plantarum CAU B946]	30.3	80%	12	65%	G

BLAST – nejúspěšnější bioinformatický nástroj

- rozšíření BLAST pro vyhledávání evolučně vzdálených bílkovin, echt strojové učení,
- videolecture: W.S.Noble: Machine Learning Methods for Protein Analysis,
  - BLAST a jeho následovníky chápe jako analogii Google,
  - využij lokální BLAST k výpočtu globální pravděpodobnostní evoluční sítě proteinů,
  - následně se implicitní vysokodimenzionální prostor transformuje do libovolné nižší dimenze,
  - vzniká "sémantická" mapa proteinů,
  - kvalitu detekce evoluční příbuznosti lze ověřit při použití strukturní informace o bílkovinách
    * nebo jako dodatečná informace v multitask learning.

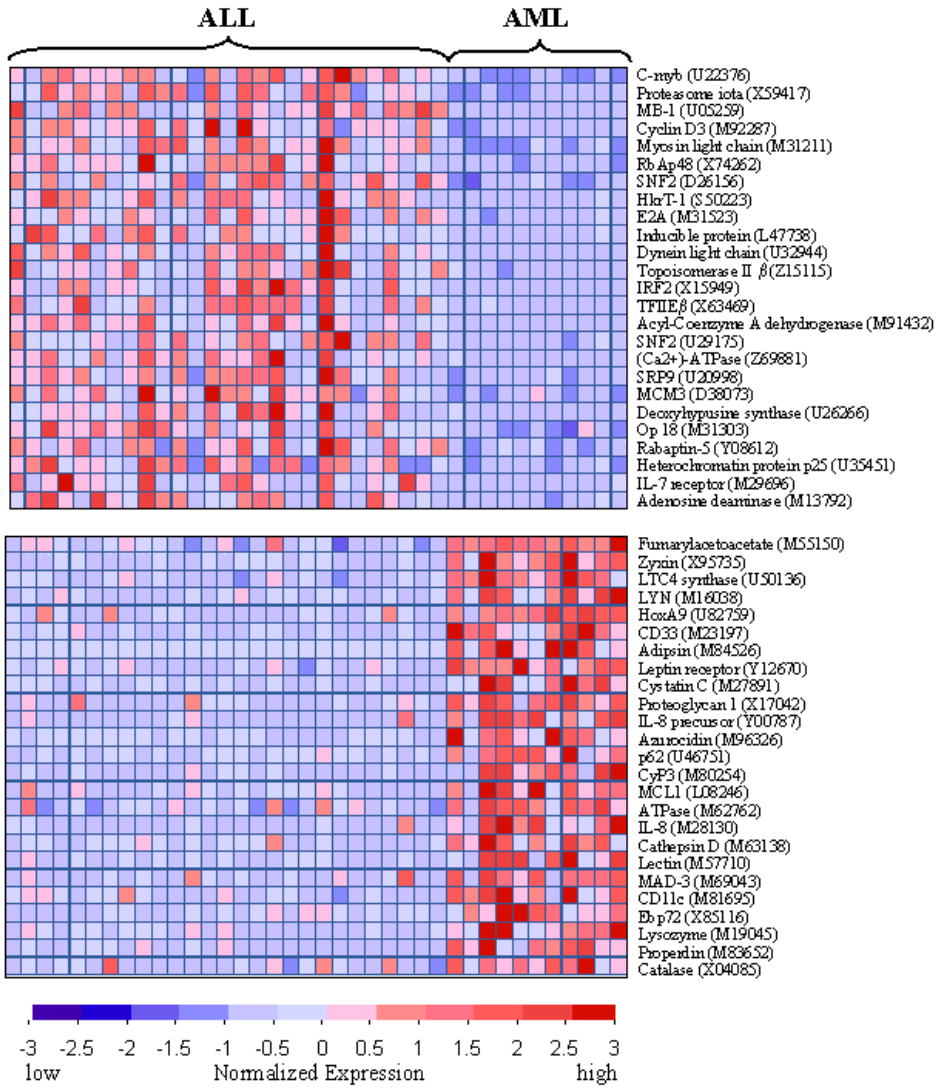


Melvin et al.: Detecting Remote Evolutionary Relationships among Proteins by Large-Scale Semantic Embedding

Molekulární klasifikace rakoviny

- 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, 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,
- postup není rutinně aplikovatelný na všechny typy nádorů resp. libovolná data genové exprese
  - predikce rekurence nádorů močového měchýře s 2.LF UK a Nemocnicí na Karlově náměstí.

Molekulární klasifikace rakoviny

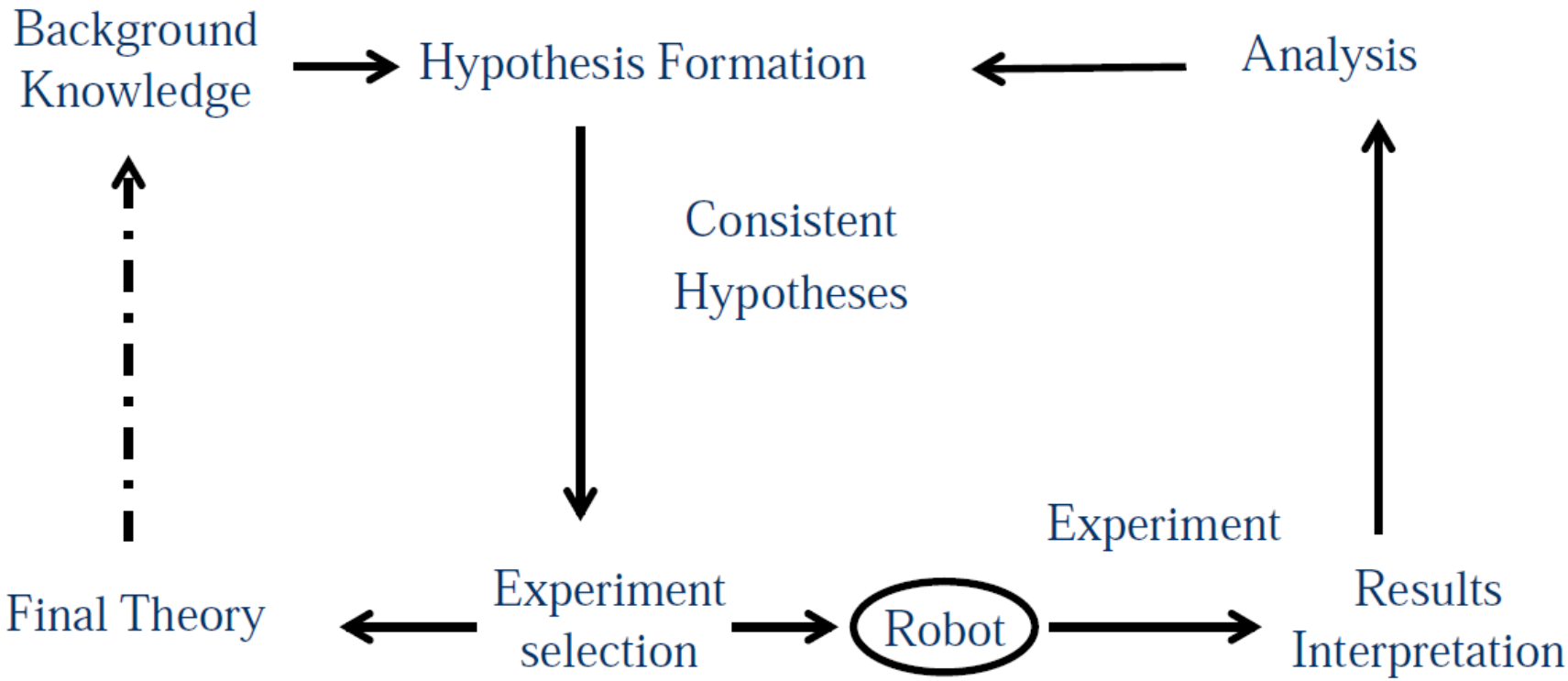


Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.

Robot vědec

- bioinfomatika jako oblast pro vývoj principiálně nových obecných metod
  - učení,
  - získávání znalostí,
- 2008-10 – Robot Scientist Adam [King, 2009]
  - nejenom fyzicky provádí biologické experimenty, sám je i navrhuje,
  - tvoří hypotézy, je schopen samostatně dojít k vědeckým závěrům,
  - vstupem obecné zadání a strukturovaný popis domény,
  - nikoli náhodou je aplikační oblastí funkční genomika
    - zprůmyslnění měření volá po zprůmyslnění analýzy těchto dat,
    - konkrétně pivní kvasinky, stále stovky genů s neznámou funkcí,
  - http://www.aber.ac.uk/en/cs/research/cb/projects/robotscientist/.

Robot vědec



King: Automating Science.

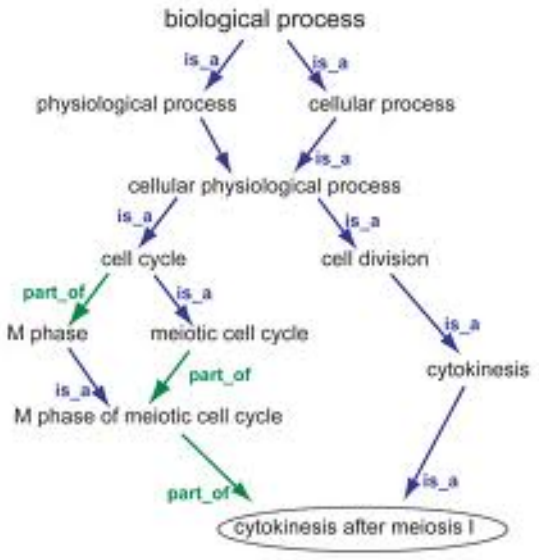
Robot vědec

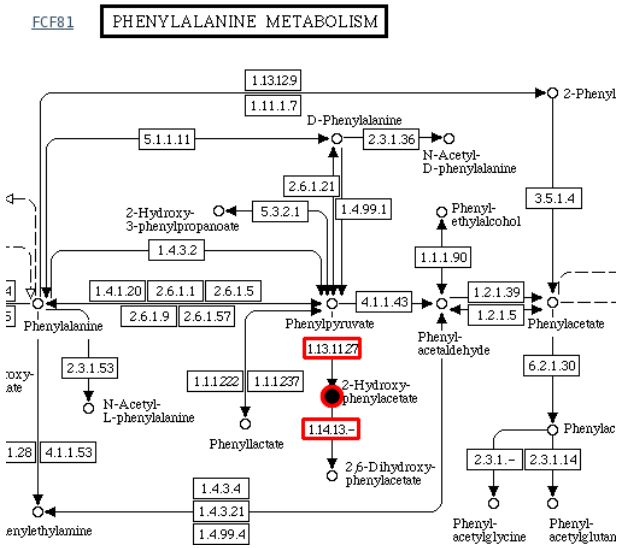
King: Automating Science.

Skupina IDA na FEL ČVUT

Klasifikace GE dat s využitím apriorní znalosti

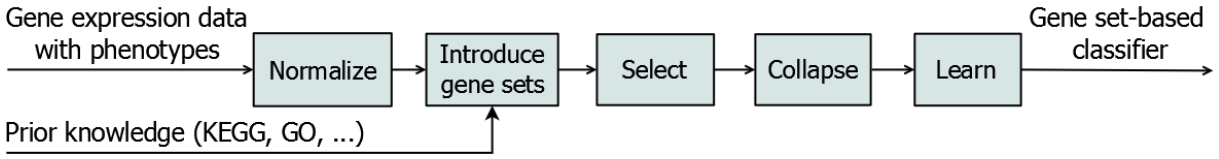
- molekulární klasifikátory založené přímo na expresi genů mohou být
  - nepřesné – šum v datech, nevyhovující poměr mezi počtem příznaků a příkladů,
  - nesrozumitelné – nahodilé vazby bez biologického významu, komplikované modely,
- nabízí se využít apriorní znalosti
  - genové ontologie, metabolické a signální dráhy, transkripční faktory,
- v termínech strojového učení odpovídá extrakci příznaků,
- analogií je rozšíření metod pro výběr signifikantně deregulovaných množin genů.





Klasifikace GE dat s využitím apriorní znalosti

- 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 otázky následné analýzy
  - jaké problémy volit pro testování,
  - je biologicky podložená agregace lepší než nahodilé agregáty stejných parametrů,
  - s čím srovnávat výslednou přesnost,
  - jak automaticky hodnotit srozumitelnost modelů.

BMC Bioinformatics

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

Open Access

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

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

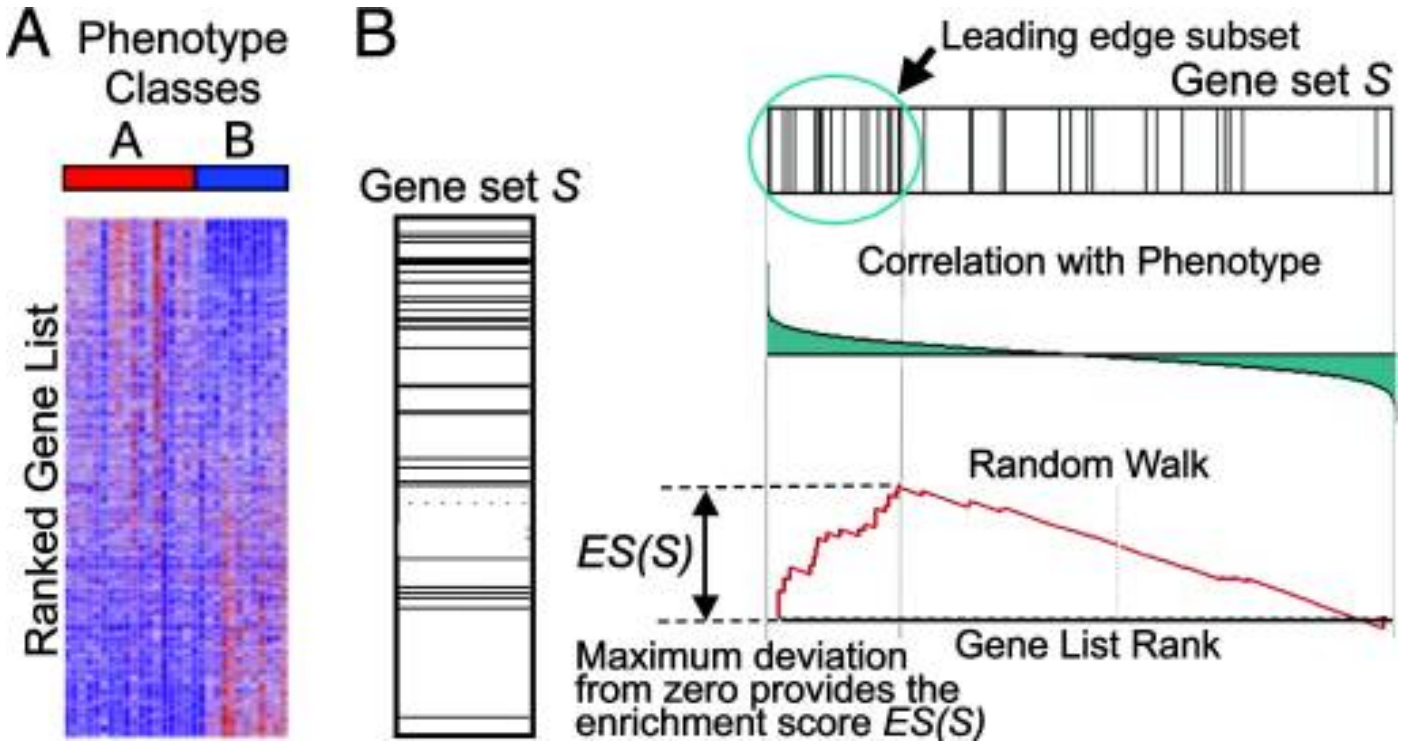
BMC Bioinformatics

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

- výběr optimálních množin genů
  - 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.

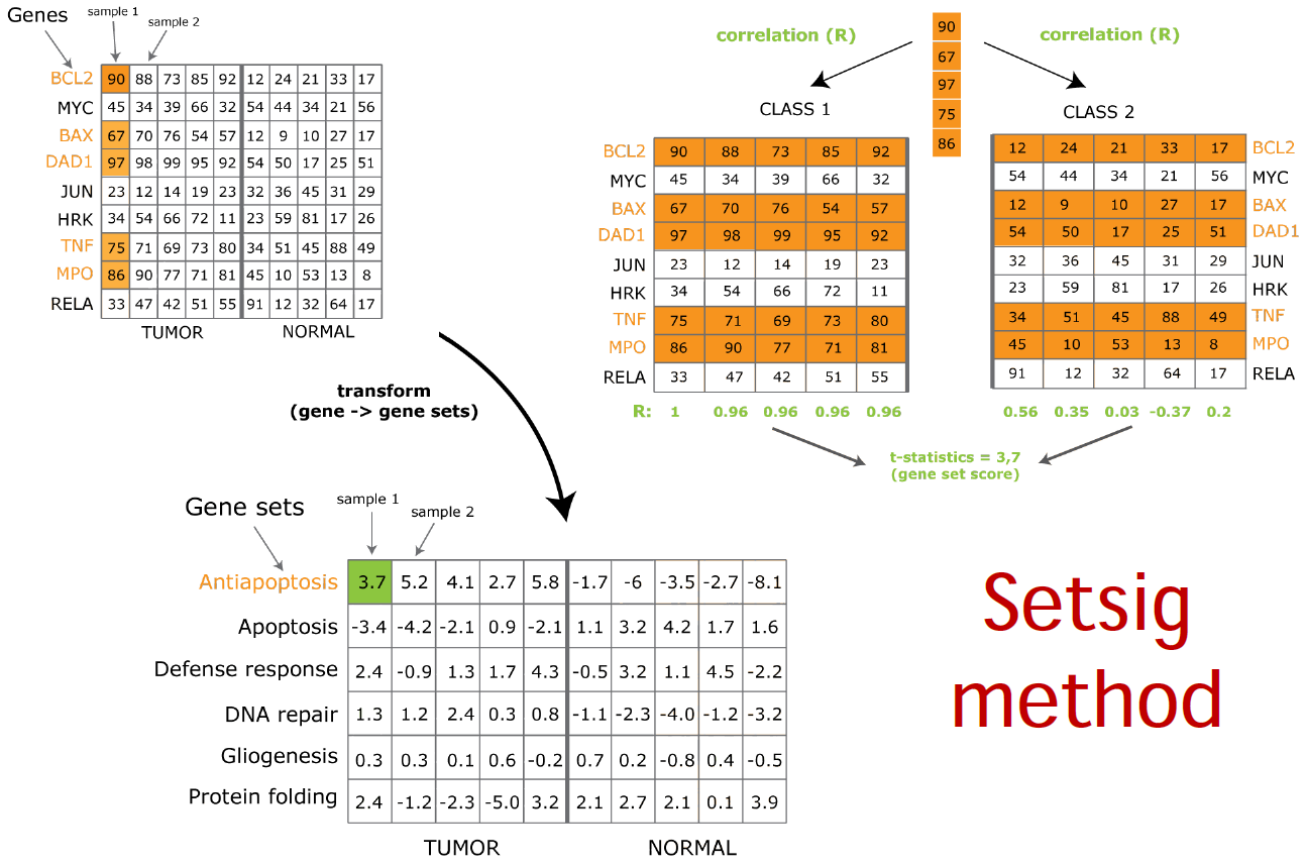
BMC Bioinformatics

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

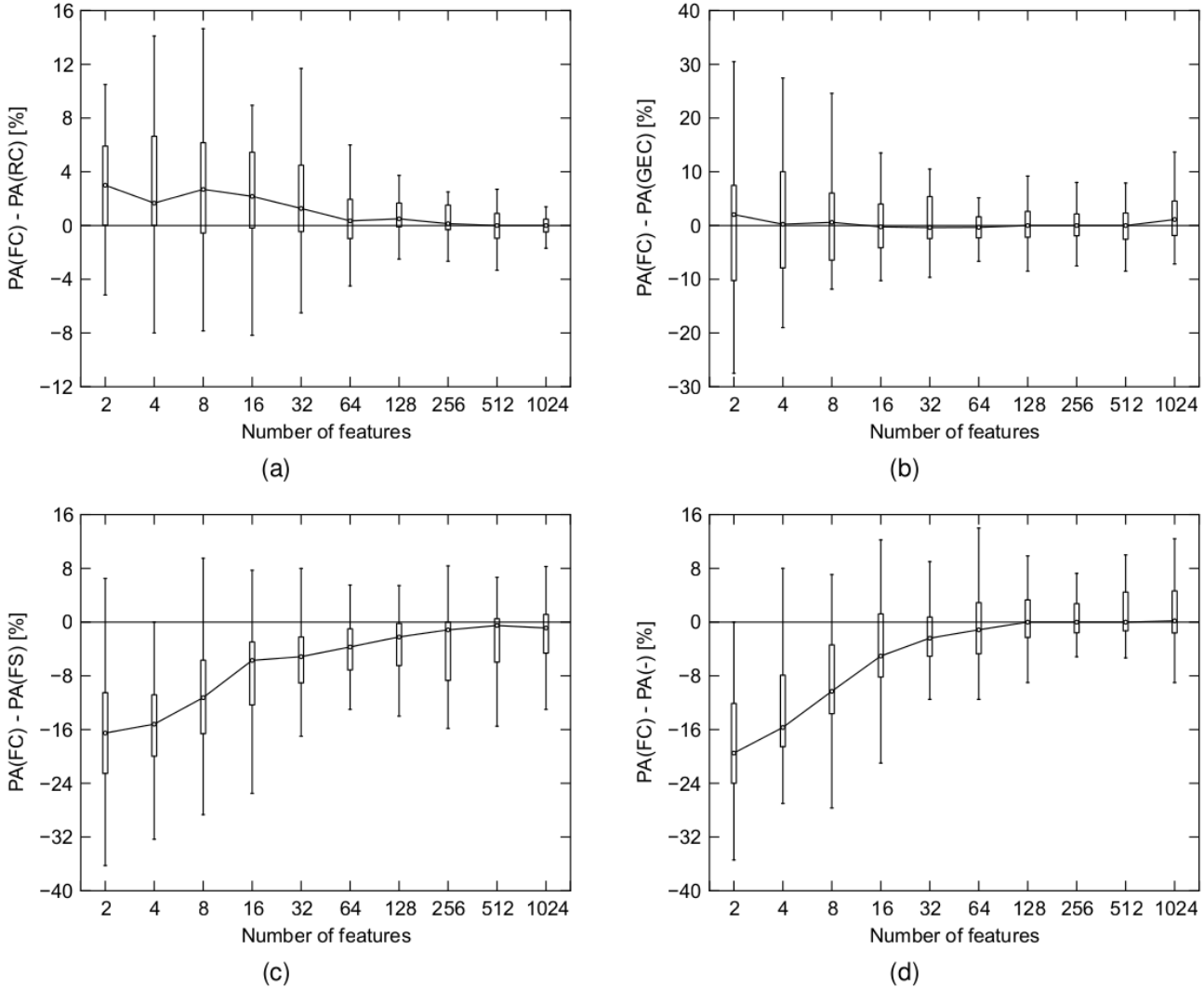
Miloš Krejník, **Jiří Kléma**, Czech Technical University, Prague

- 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

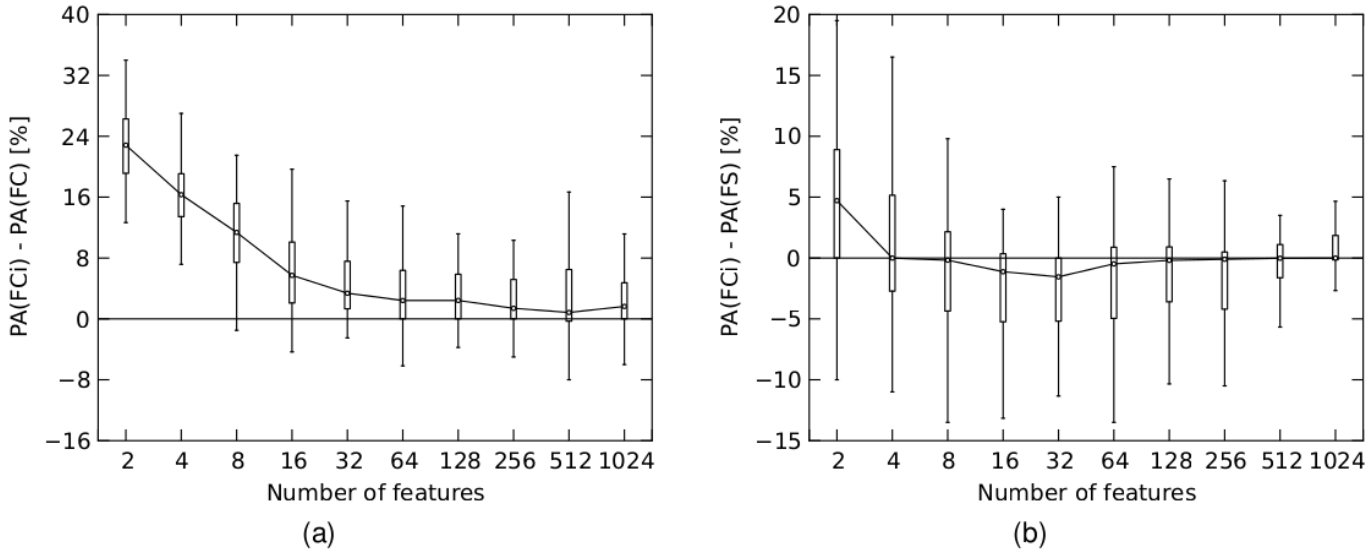
May-June 2012 (vol. 9 no. 3) pp. 788-798
Miloš Krejník, **Jiří Kléma**, Czech Technical University, Prague



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

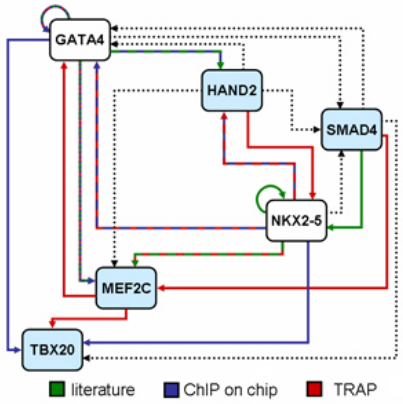
Miloš Krejník, **Jiří Kléma**, Czech Technical University, Prague



- 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.

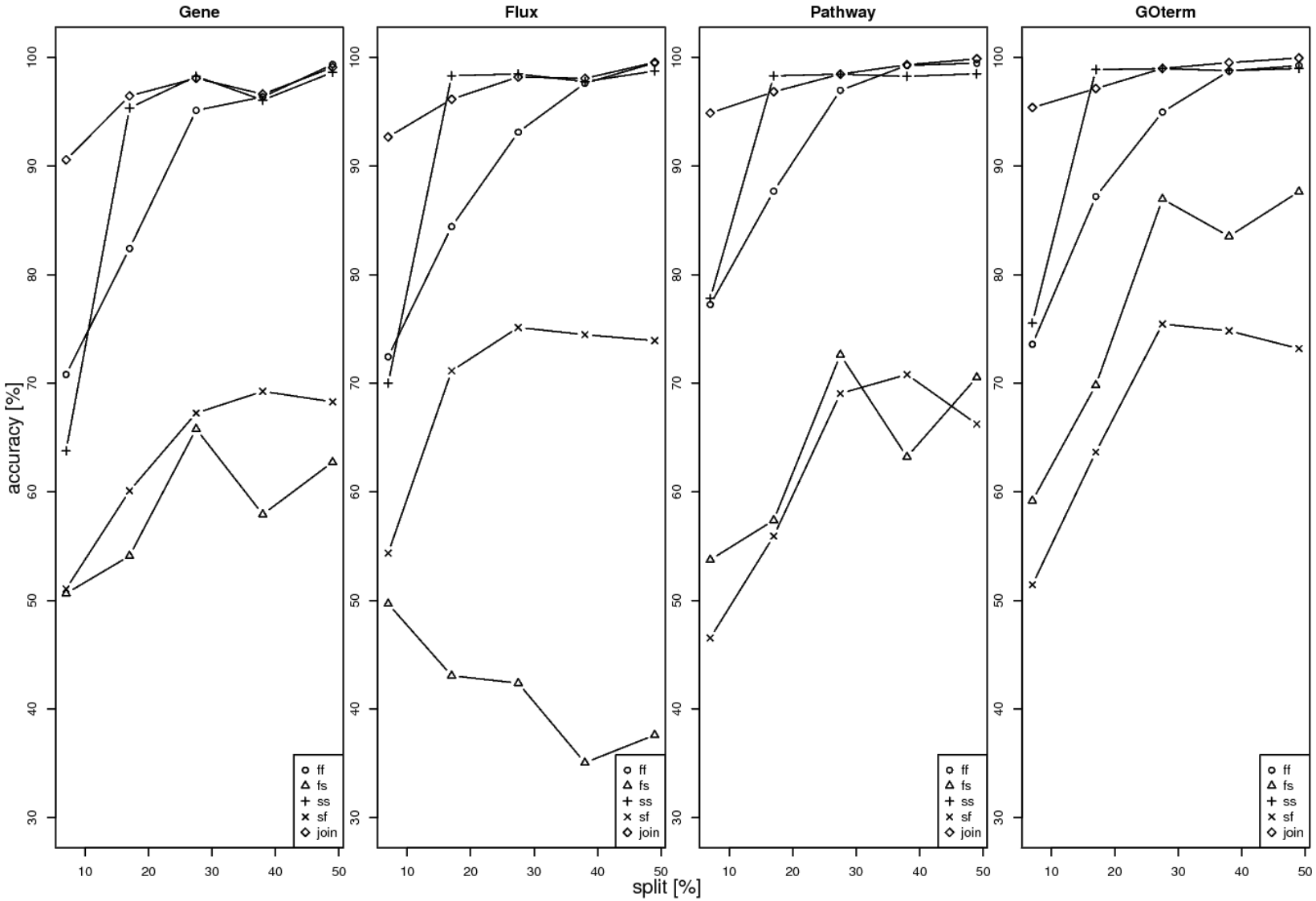
Současný a budoucí výzkum

- širší studium využití pro integraci dat z více platforem a více biologických druhů
  - jsou obecnější naddruhové klasifikátory v případě malého počtu vzorků výhodné?
    - jakou škálu druhů lze pokrýt?
    - kolik zhruba vzorků
  - mají klasifikátory biologický význam?
- učení pracovních toků
  - spolu s předzpracováním GE dat je tvorba množinového klasifikátoru netriviální tok,
  - v širším kontextu řeší evropský projekt e-Lico,
  - ne ontologie, existuje šablona, je třeba ji instanciovat,
  - využití klasifikační přesnosti.
- iterativní konstrukce příznaků ze známé regulační sítě
  - využívá i GE dat k inicializaci zrn,
  - postupně zrna rozšiřuje, dokud se příznaky "zlepšují",
  - nevede na mohutné příznaky, ale zvyšuje přesnost.



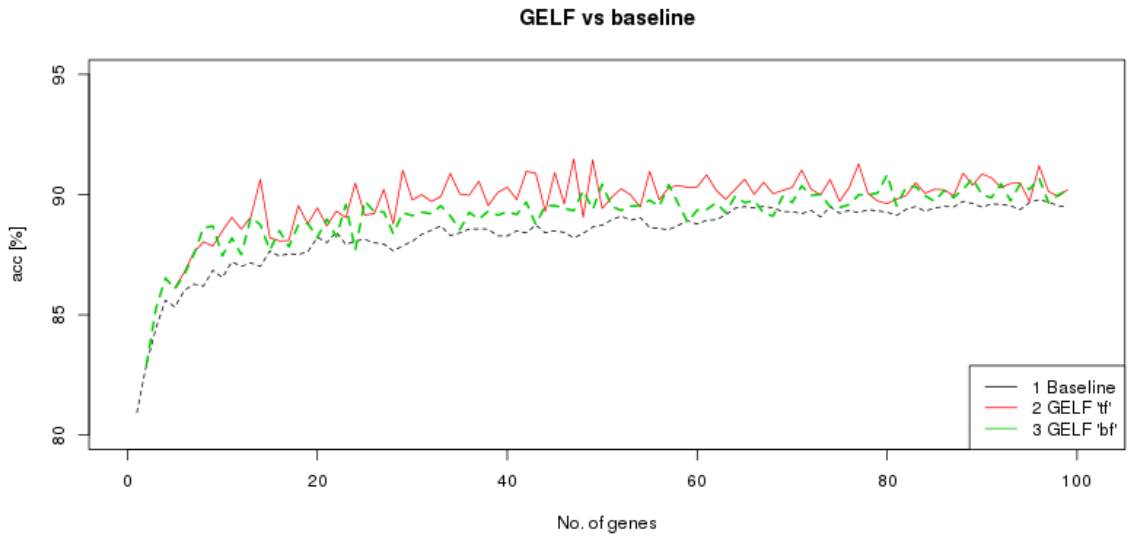
Toenjes: Prediction of cardiac TNs ...

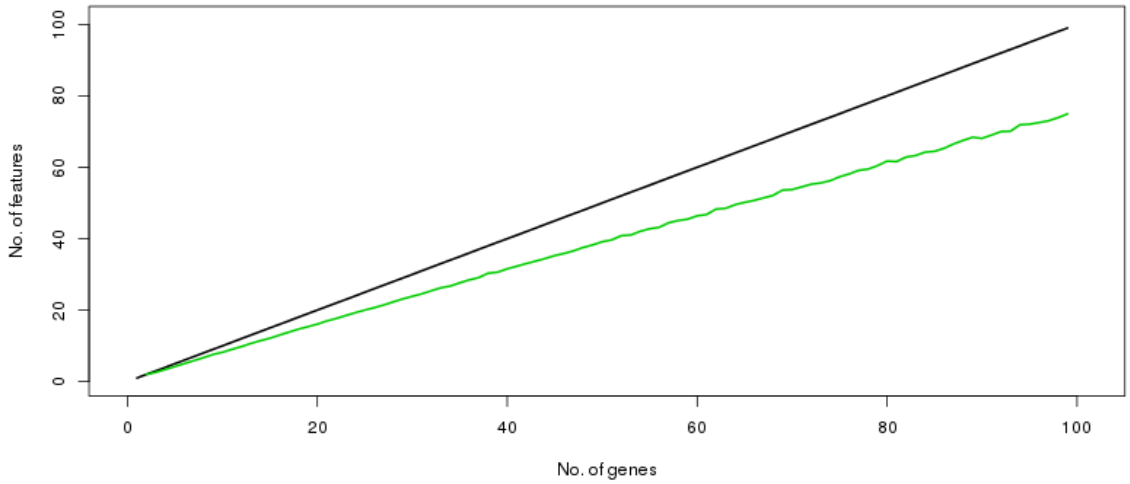
Učení z dat z více platforem a více biologických druhů



Holec, Klema et al.: Cross-Species and Cross-Platform Classification of Expression Data through Gene-Set Features, dosud nepublikováno.

Iterativní konstrukce příznaků podle transkripčních faktorů





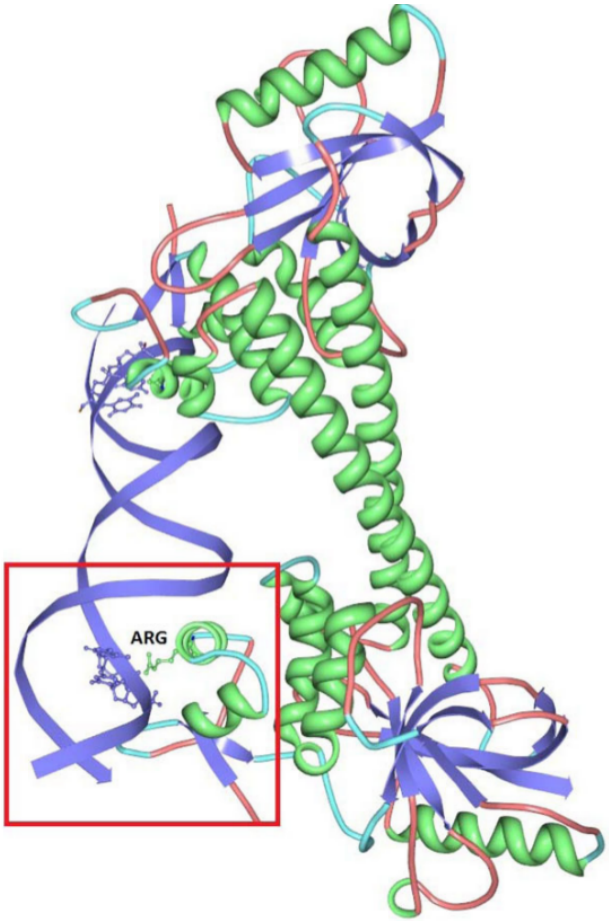
Holec, Kuzelka: GELF, dosud nepublikováno.

BMC Bioinformatics

Prediction of DNA-binding propensity of proteins by the ball-histogram method using automatic template search

Andrea Szabóová[1*], Ondřej Kuželka[1], Filip Železný[1] and Jakub Tolar[2]

- predikce schopnosti bílkovin vázat se na DNA
  - je dána 3-D struktura bílkoviny (PDB soubory)
    - u každé aminokyseliny dána pozice jejího alfa-uhlíku,
  - učení s učitelem, k dispozici binární anotace,
- povaha vazby
  - elektrostatická vazba
    - díky asymetrii náboje na povrchu proteinu,
- stávající metody predikce
  - dle manuálně konstruovaných vlastností proteinu
    - sekvenční, strukturní i evoluční,
    - stále otevřený problém.

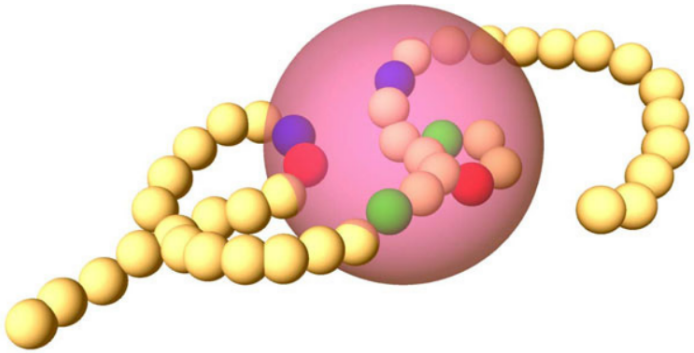


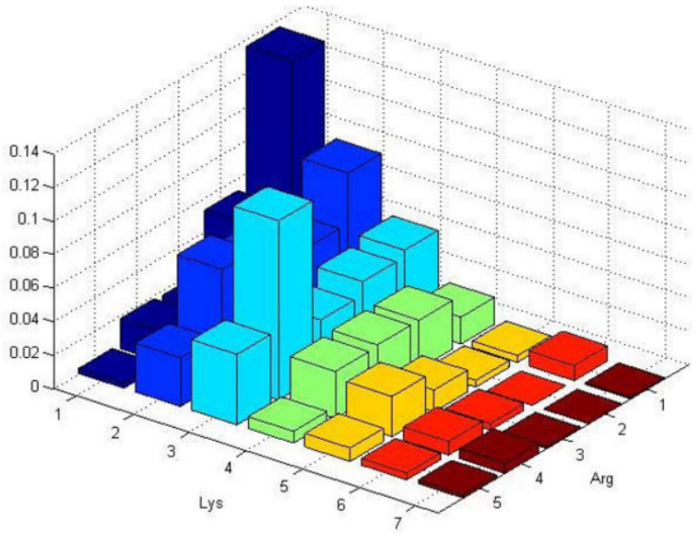
MBT protein viewer.

Prediction of DNA-binding propensity of proteins by the ball-histogram method using automatic template search

Andrea Szabóová[1*], Ondřej Kuželka[1], Filip Železný[1] and Jakub Tolar[2]

- metoda kulových histogramů
  - modelování rozdělení aminokyselin v prostoru
    - Monte-Carlo metoda, invariantní k rotaci proteinu,
  - 4 základní kroky
    - vyhledání šablon (k-tic sdružených vlastností), jejichž rozdělení bude sledováno,
    - sestavení kulových histogramů pro všechny trénovací proteiny,
    - propozicionalizace histogramů, převod do AVL formy,
    - učení náhodných lesů.





BMC Bioinformatics

Prediction of DNA-binding propensity of proteins by the ball-histogram method using automatic template search

Andrea Szabóová[1*], Ondřej Kuželka[1], Filip Železný[1] and Jakub Tolar[2]

- závěry
  - nalezeny zajímavé šablony,
  - překonána přesnost SOA metody,
  - autoři pracují na metodě založené na klasickém relačním učení.

Table 3 DNA-binding proteins:

	Lys				Gly				Gly		
Arg	**0.5**	**0**	0.5	Arg	**0.4**	**0.1**	0.5	Lys	**0.4**	**0.1**	0.5
	0	**0.5**	0.5		**0.2**	**0.3**	0.5		**0.2**	**0.3**	0.5
	0.5	0.5			0.6	0.4			0.6	0.4	

Table 4 Non-DNA-binding proteins:

	Lys				Gly				Gly		
Arg	**0**	**0.5**	0.5	Arg	**0.1**	**0.4**	0.5	Lys	**0.1**	**0.4**	0.5
	0.5	**0**	0.5		**0.3**	**0.2**	0.5		**0.3**	**0.2**	0.5
	0.5	0.5			0.4	0.6			0.4	0.6	

Table 2 Accuracies estimated by 10-fold cross-validation.

	Classifier	PD138/ NB110	PD138/ NB843
Ball Histograms	Random Forest	**0.87 ± 0.08**	**0.88 ± 0.01**
	SVM	0.84 ± 0.07	0.87 ± 0.01
	Logistic Regression	0.81 ± 0.05	0.87 ± 0.01
Szilágyi and Skolnick [7]	Random Forest	0.82 ± 0.07	0.87 ± 0.02
	SVM	0.81 ± 0.05	0.87 ± 0.01

Další bionformatický výzkum ve skupině IDA

- před dokončením

Learning protein-protein interactions with Markov logic networks

Přemysl Vítovec, Filip Železný and Jiří Kléma*

- spolupráce s biologickými pracovišti

Open Access

IMPACT FACTOR 7.25

Global gene expression changes in human embryonic lung fibroblasts induced by organic extracts from respirable air particles

Helena Líbalová[1,2], Kateřina Uhlířová[1], Jiří Kléma[3], Miroslav Machala[4], Radim J Šrám[1], Miroslav Ciganek[4] and Jan Topinka[1*]

1 Department of Genetic Ecotoxicology, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, 142 20 Prague 4, Czech Republic

2 Department of Biochemistry, Faculty of Science, Charles University, Albertov 2030, 128 40 Prague 2, Czech Republic

3 Czech Technical University in Prague, Prague 2, Czech Republic

4 Veterinary Research Institute, Brno, Czech Republic

Differential Regulation of the Nuclear Factor-κB Pathway by Rabbit Antithymocyte Globulins in Kidney Transplantation

Urbanova, Mariana[1,2]; Brabcova, Irena[2]; Girmanova, Eva[2]; Zelezny, Filip[3]; Viklicky, Ondrej[1,2,4]

Transplantation®

THE OFFICIAL JOURNAL OF THE TRANSPLANTATION SOCIETY

Shrnutí

- bioinformatika, resp. molekulární biologie je atraktivní pro strojové učení
  - velké objemy dat,
  - měření jsou často zašuměná,
  - data často neanotovaná, strukturovaná a rozmanitá,
  - důležitý, zajímavý a živý obor,
  - prostor pro vznik nových aplikačně specifických ML algoritmů,
  - prostor pro novou aplikaci těch stávajících,
- strojové učení je atraktivní pro bioinformatiku
  - z charakteristik dat uvedených výše,
  - automatizace generování dat volá po automatizaci jejich výkladu.

Zdroje přednášky

- Larranaga et al. (2005) **Machine Learning in Bioinformatics**. Briefings in Bioinformatics, vol. 7, no. 1, pp.86–112.
- Altschul et al. (1997) **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** Nucleic Acids Res. 25:3389-3402.
- Golub et al. (1999) **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** Science, Oct 15;286(5439):531-7.
- King et al. (2009) **The Automation of Science.** Science 324 (5923), pp.85–89.
- Holec, Klema et al. (2012) **Comparative Evaluation of Set-Level Techniques in Predictive Classification of GE Samples.** BMC Bioinformatics, 13, Suppl. 10, S15.
- Krejnik, Klema (2012) **Empirical Evidence of the Applicability of Functional Clustering through Gene Expression Classification.** IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9:3, pp. 788-798.
- Szaboova et al. (2012) **Prediction of DNA-binding Propensity of Proteins by the Ball-Histogram Method using Automatic Template Search.** BMC Bioinformatics 13, Suppl 10, S3.