PB051: Výpočetní metody v bioinformatice a systémové biologii

David Šafránek

6.4.2012

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

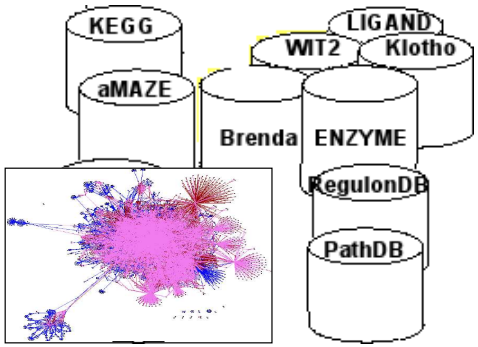
Obsah

Obsah

Průběh výzkumu v systémové biologii

rekonstrukce sítí

databáze biol. znalostí + literatura

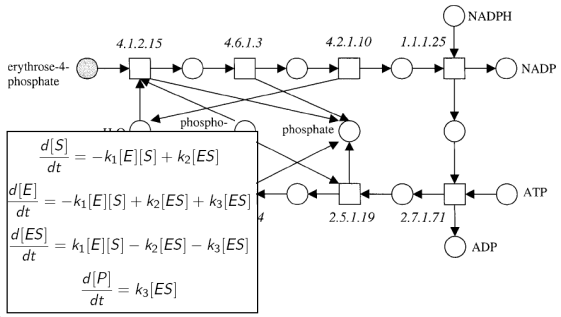


biologická síť

hypotézy

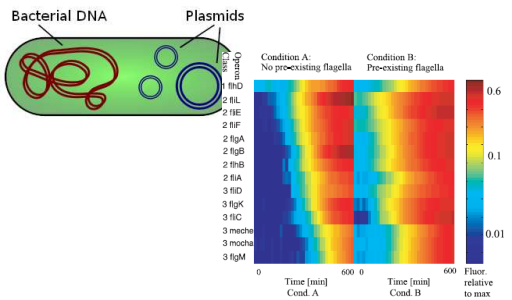
specifikace modelu

SBML, diferenciální rovnice,
boolovská síť, Petriho síť, ...



validace modelu

genové reportéry, DNA microarray,
hmotnostní spektrometrie, ...

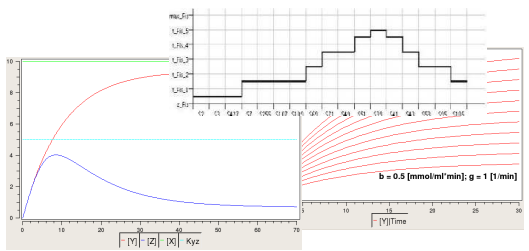


objevené vlastnosti

dotazy na model

analýza modelu

statická analýza, numerická simulace,
analytické metody, model checking



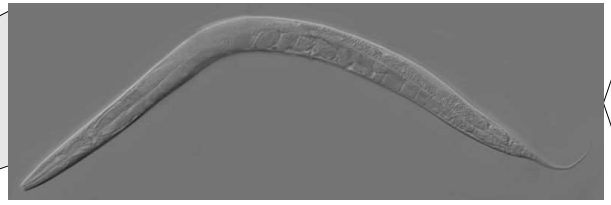
verifikace hypotéz, detekce vlastností
vyvození nových hypotéz

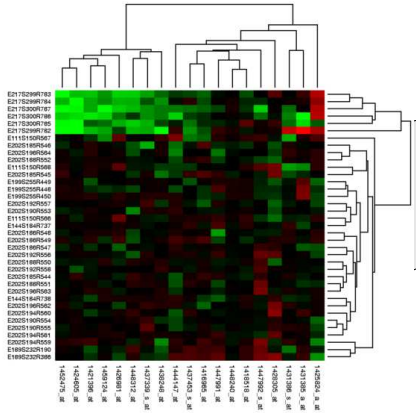
Metody systémového měření

komponentová biologie

high−througput tenchologie

genomika
proteomika



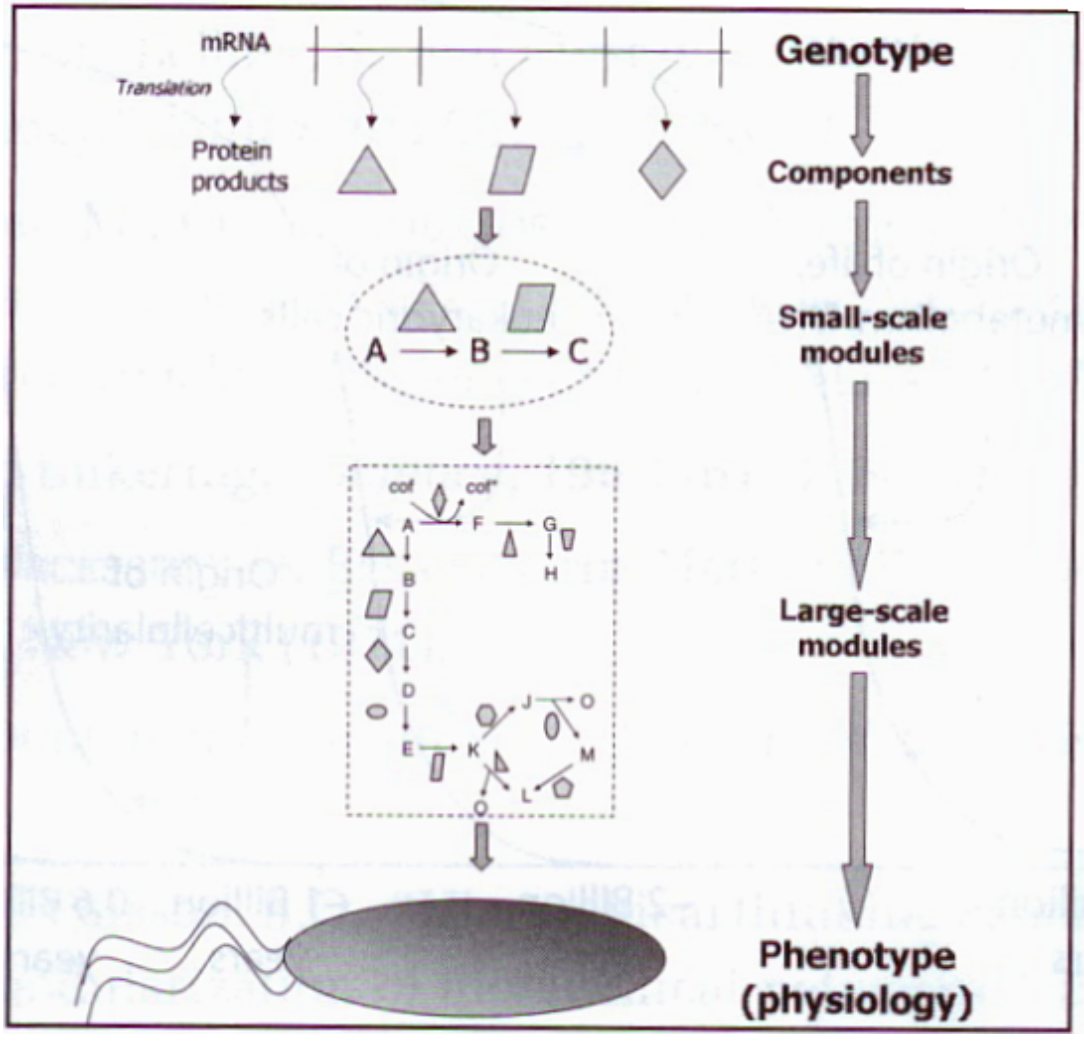


systémová biologie

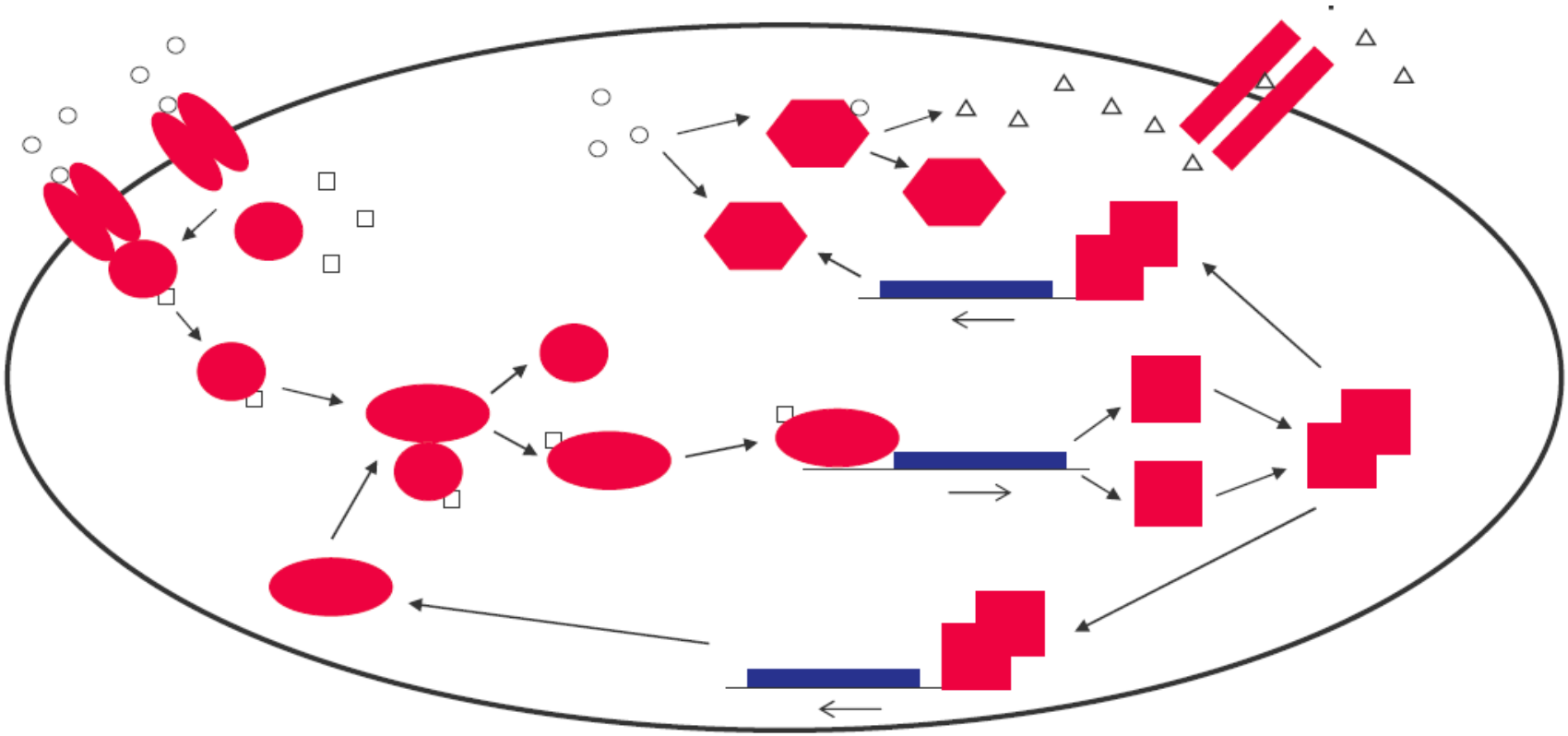
integrativní analýza

bionformatika
modely (in silico)
simulace

Koncept hierarchie

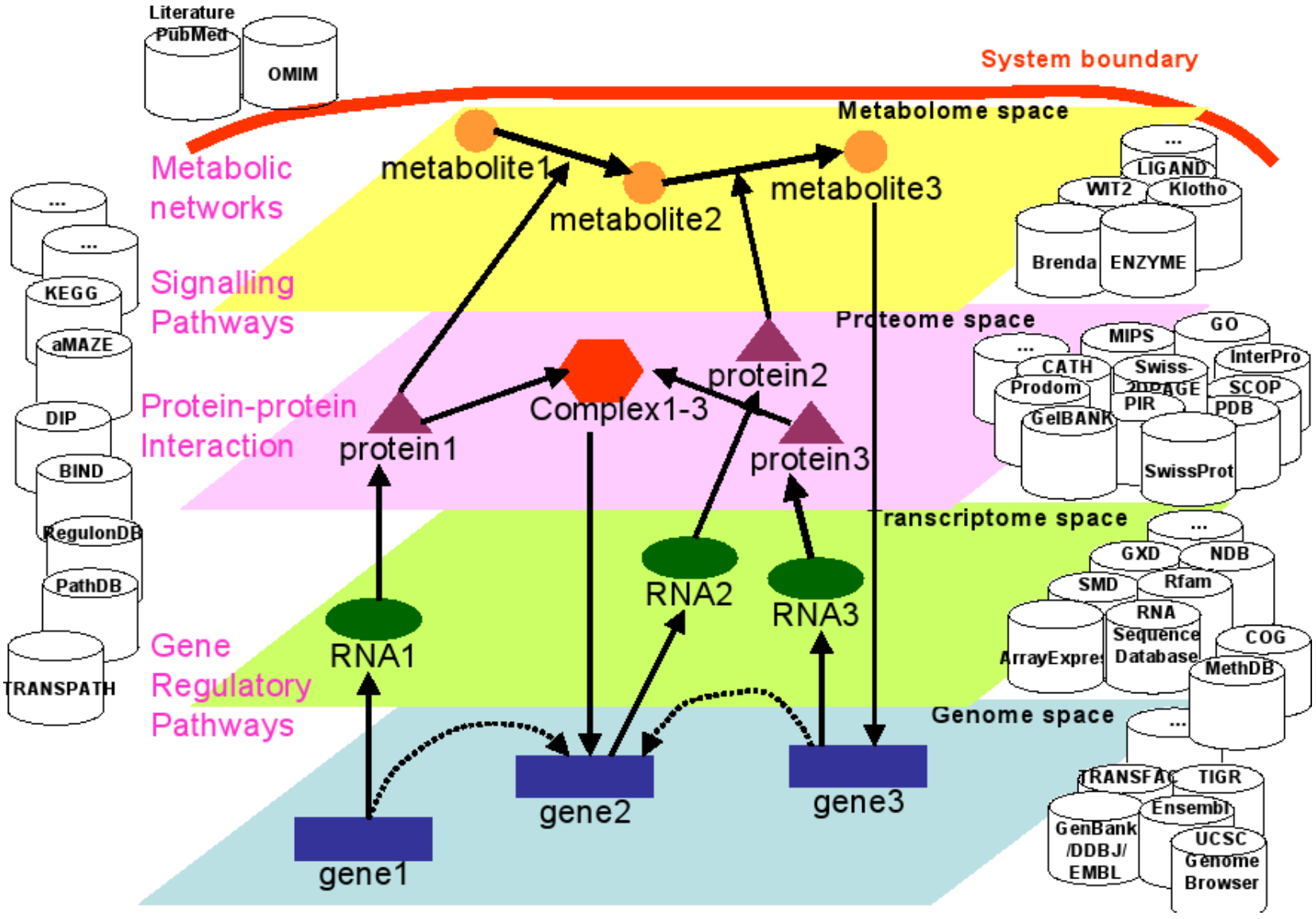


Biochemické procesy v buňce

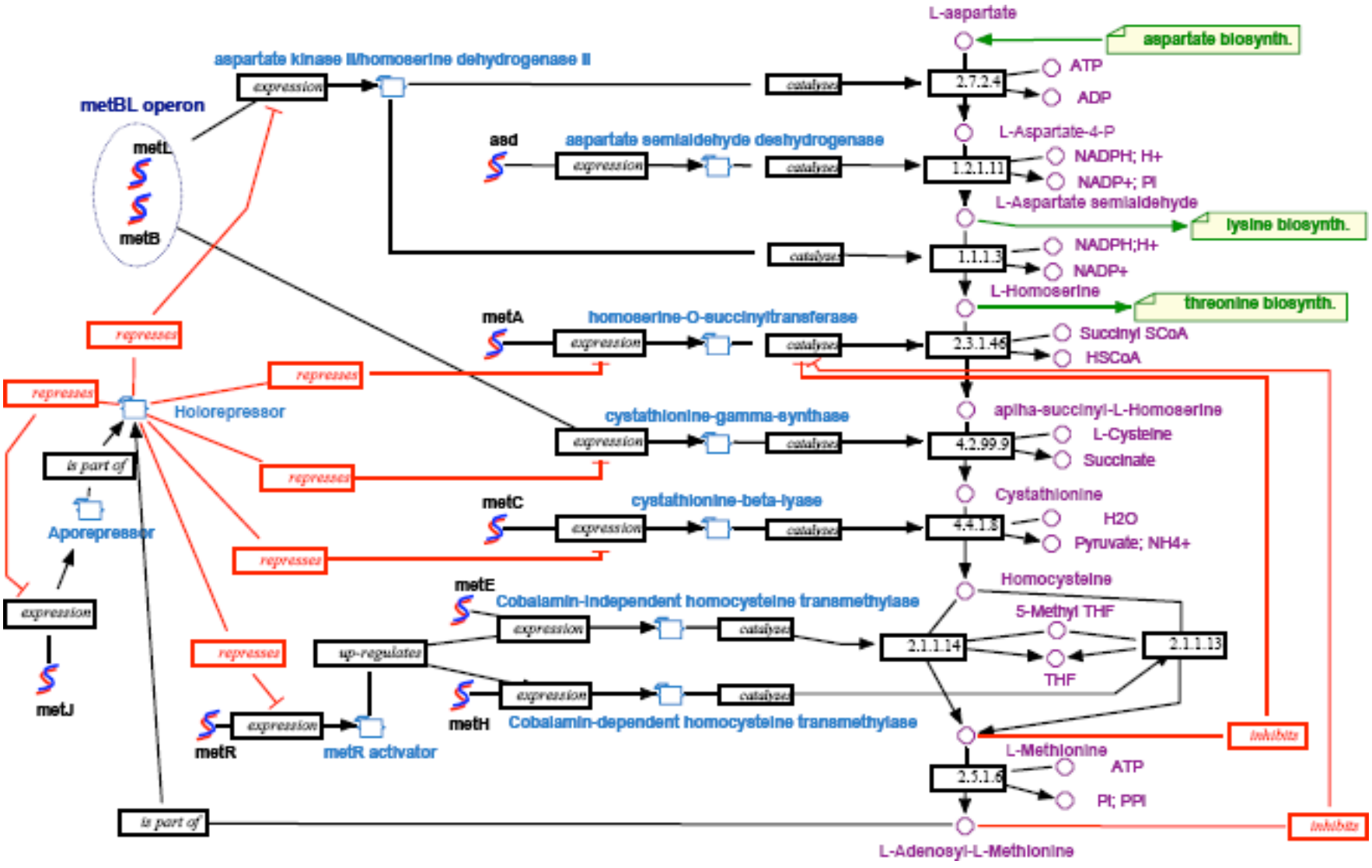


- molekulární komponenty – proteiny, DNA, RNA,. . .
- interakce na různých úrovních (transkripce, metabolismus,. . .)
- příjem signálů a živin (nutrientů) na membráně

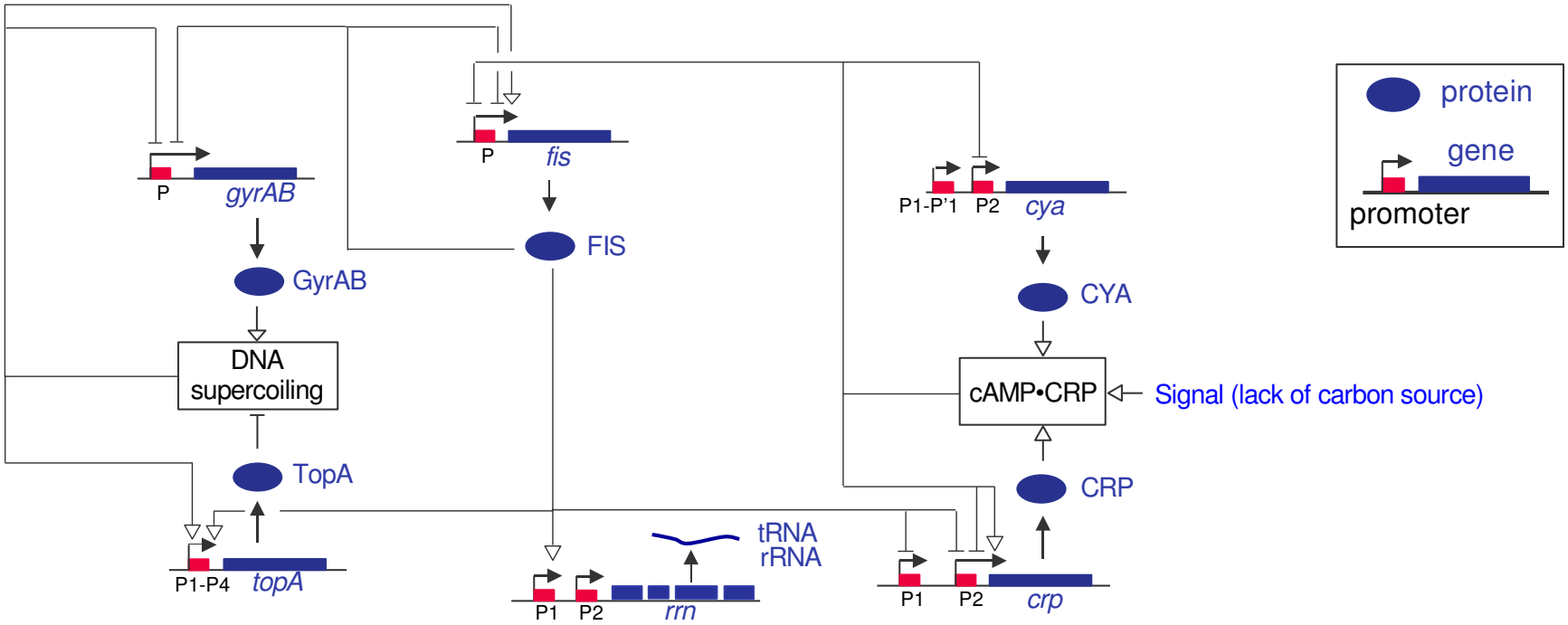
Funkční vsrtvy buňky



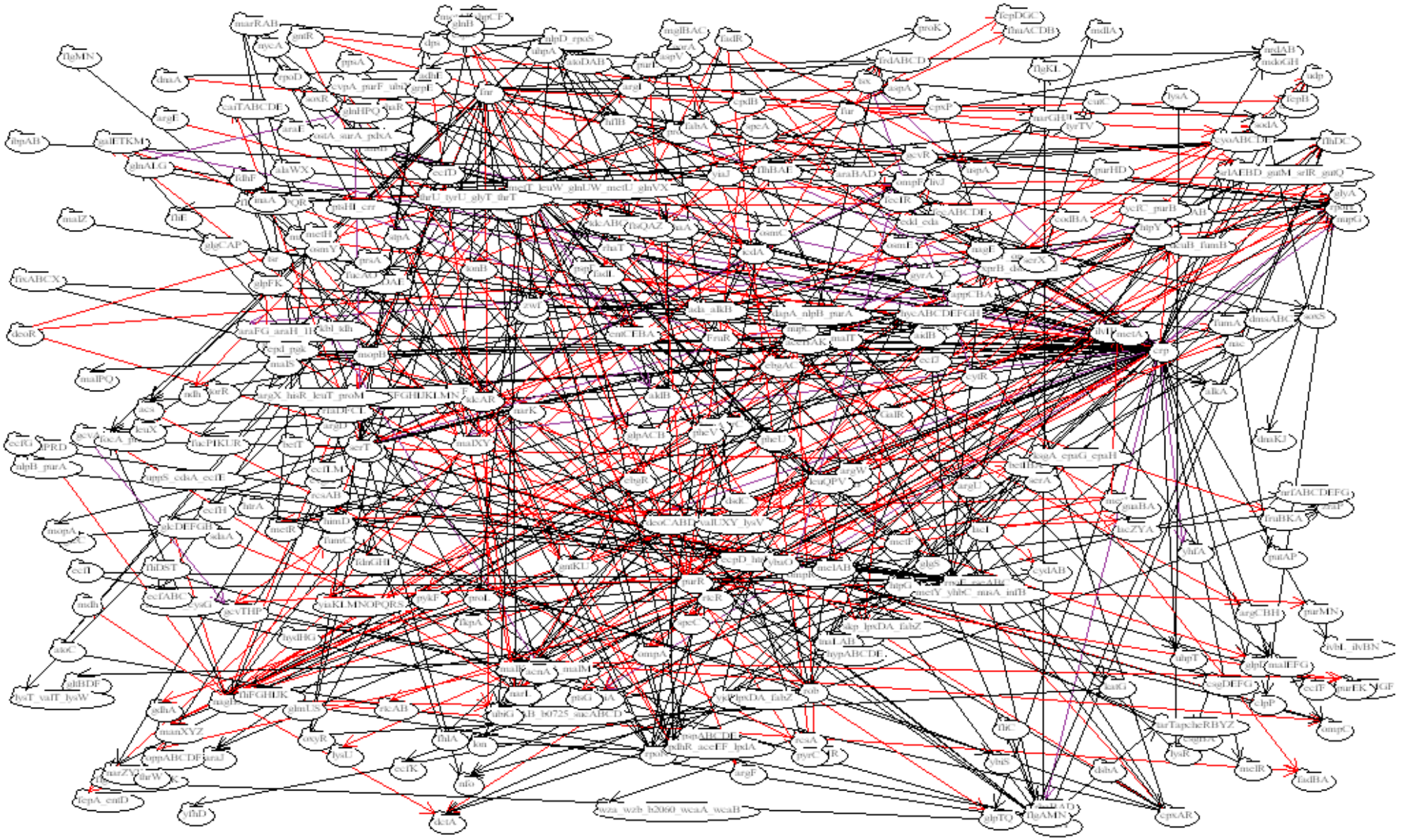
Příklad geneticky řízené metabolické dráhy



Příklad modulu genetické regulace v E. Coli



Kompletní transkripční síť E. Coli



Biologická síť jako obecný graf

Definition

Nechť V je konečná množina uzlů a $E \subseteq V \times V$ relace. *Biologickou sítí* nazveme graf G reprezentovaný uspořádanou dvojicí $G \equiv (V, E)$.

- Pokud $\forall \langle a, b \rangle \in E. \langle a, b \rangle \in E \rightarrow \langle b, a \rangle \in E$, G nazýváme *neorientovaný*.
- V ostatních případech hovoříme o *orientovaném* grafu.

typ sítě	V	E	G
genové	geny (resp. proteiny)	regulace exprese	or.
proteinové	proteiny	proteinové interakce	neor.
metabolické	metabolity, enzymy	enzymové reakce	or.
signální	molekuly	aktivace/deaktivace	or.

Biologické sítě – výpočetní problémy

- (re)konstrukce sítí
  - identifikace interakcí
    $\Rightarrow$ z experimentálních dat
    $\Rightarrow$ integrací dat z databází znalostí
- zpracování sítí
  - vizualizace
  - integrace atributů uzlů, hran
- analýza sítí
  - analýza statistických vlastností
    $\Rightarrow$ rozložení konektivity, detekce motivů, ...
  - porovnání sítí

Ontologie genů a genových produktů

- nutnost systematicky uchopit genom, genové produkty a funkce (interaktom)
- akumulace biologických dat
  - decentralizovaný proces
  - paralelní proces
- problém: nejednoznačné popisy téhož objektu, procesu
  - např. proteolýza vs. (řízená degradace proteinů)
- od roku 1998 vývoj Gene Ontology
  $\Rightarrow$ `http://geneontology.org`

Gene Ontology

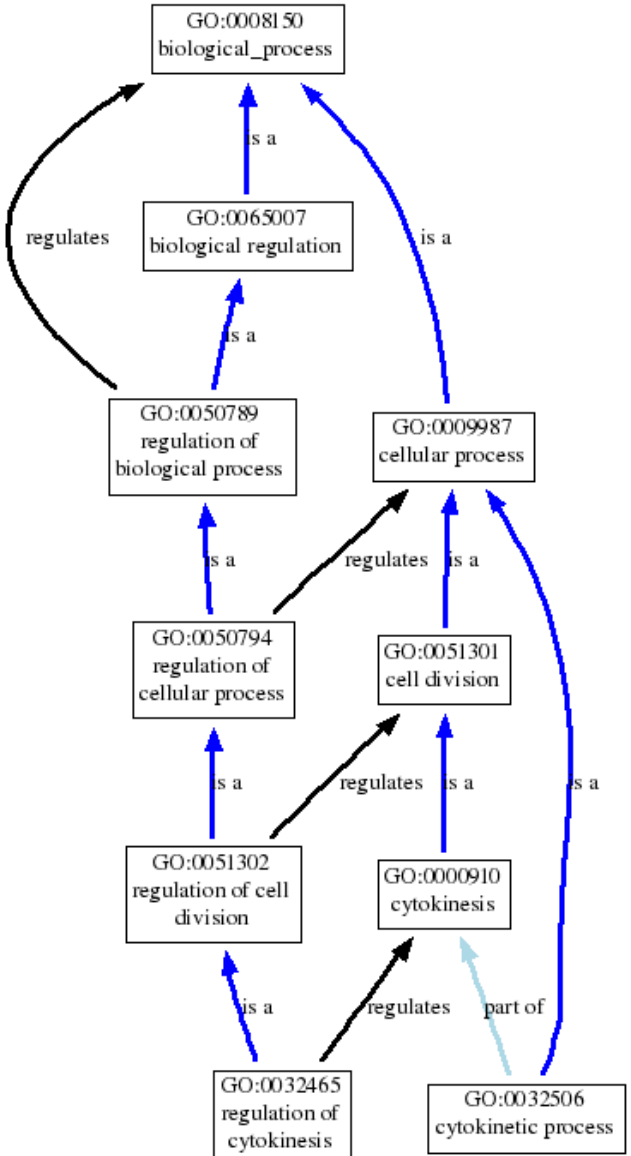
- cílem je ucelený systematický popis genových produktů a jejich funkcí
- ontologie představuje systematický slovník pojmů
  - hierarchický pohled na data včetně vazeb mezi nimi (DAG)
  - zachycení synonym
- GO obsahuje 3 kategorie (DAGy):
  - biological process
  - cellular component
  - molecular function

Gene Ontology

- respektuje standard OBO a OWL (W3C) pro strukturu a reprezentaci ontologií
  - `http://obofoundry.org/`
  - `http://www.w3.org/TR/owl-features/`
- každý uzel představuje jednoznačný pojem (množinu synonym) $\Rightarrow$ jednoznačně reprezentován ID (tzv. GO termem)
- různé typy relací mezi uzly:
  - part_of, is_a, located_in, derived_from, ...
  - viz `http://obofoundry.org/ro/`

Gene Ontology

Příklad podgrafu GO



Nástroje pro Gene Ontology

- on-line i off-line vyhledávače v GO
- statistické testy na overreprezentaci dané množiny genů v pojmech GO
  - GOstat – `http://gostat.wehi.edu.au/cgi-bin/goStat.pl`
  - DAVID – `http://david.abcc.ncifcrf.gov/`
  - BiNGO – `http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html`
- mapování microarray dat na ontologický strom
  - ~~eGOn – `http://www.genetools.no/`~~
  - High-Throughput GoMiner – `http://discover.nci.nih.gov/gominer/htgm.jsp`
  - ~~Meta Gene Profiler (MetaGP) – `http://metagp.ism.ac.jp/`~~
- ...

KEGG – Kyoto Encyclopedia of Genes and Genomes

- integrativní databáze genových produktů
- genový prostor
  $\Rightarrow$ GENES, GENOME, ORTHOLOGY, Organisms
- chemický prostor
  $\Rightarrow$ COMPOUND, GLYCAN, REACTION, ENZYME, LIGAND
- systémový prostor
  $\Rightarrow$ PATHWAY, BRITE, DISEASE, DRUG
- každý pojem má jednoznačné ID
- zachycena ortologie (podobnost) genů
- genový prostor čerpá aktuálně z několika zdrojů

Cvičení

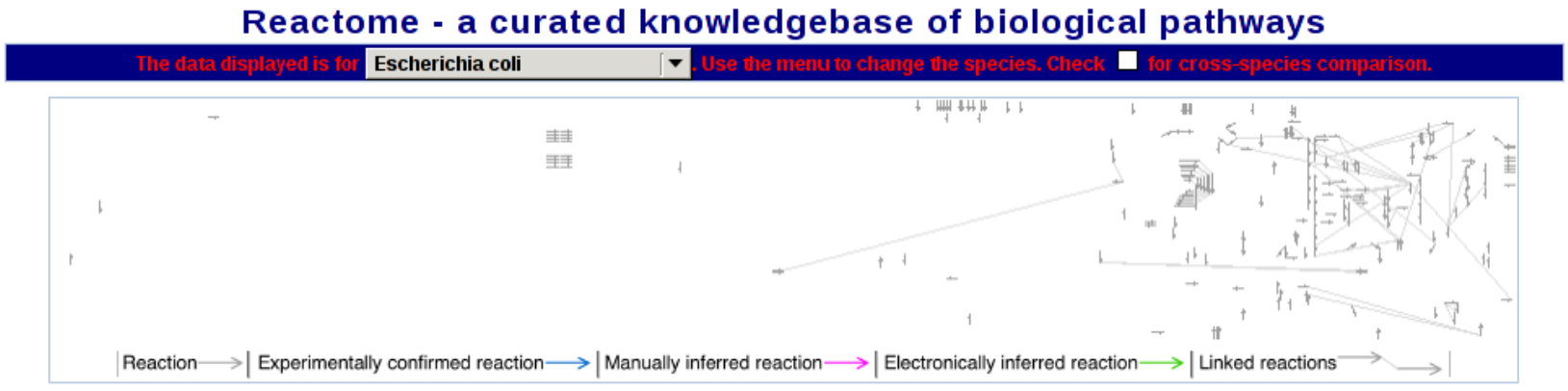
Uvažujme sekvenci aminokyselin (FASTA formát):

```
SVVEEHGQLSISNGELVNERGEQVQLKGMSSHGLQWYGQ
FVNYESMKWLRDDWGINVFRAAMYTSSGGYIDDPS
VKEKVKEAVEAAIDLDIYVIIDWHILSDNDPNIYK
EEAKDFFDEMSELYGDYPNVIYEIANEPNGSDVTW
GNQIKPYAEEVIPIIRNNDPNNIIIVGTGTWSQDV
HHAADNQLADPNVMYAFHFYAGTHGQNLRDQVDYA
LDQGAAIFVSEWGTSAATGDGGVFLDEAQVWIDFM
DERNLSWANWSLTHKDESSAALMPGANPTGGWTEA
ELSPSGTFVREKIRES
```

pomocí vhodné služby odhadněte, o jaký protein/enzym se jedná a k němu dohledejte informaci na IntEnz

- najděte na `http://www.ebi.ac.uk/intenz/`
- prozkumejte relevantní dráhy na `http://www.genome.jp/kegg/`

Reactome



Apoptosis	Axon guidance	Biological oxidations	Botulinum neurotoxicity
Cell junction organization	Cell Cycle Checkpoints	Cell Cycle, Mitotic	Chromosome Maintenance
DNA Repair	DNA Replication	Diabetes pathways	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
Gene Expression	Hemostasis	HIV Infection	Interactions of the immunoglobulin
Influenza Infection	Integration of energy metabolism	Integrin cell surface interactions	
Metabolism of amino acids and derivatives	Metabolism of carbohydrates	Metabolism of lipids and lipo	
Metabolism of nucleotides	Metabolism of porphyrins	Metabolism of prot	
Metabolism of vitamins and cofactors	Muscle contraction	mRNA Process	
Pyruvate metabolism and Citric Acid (TCA) cycle	Regulation of beta-cell development	Regulatory RNA path	
Signaling by EGFR	Signaling by FGFR	Signaling by GPCR	
Signaling in Immune system	Signaling by Insulin receptor	Signalling by NGF	

starý look reactome;
v současné verzi používá
SBGN standard,
doporučuji proklikat

http://www.reactome.org/

Specificky zaměřené zdroje dat

- EcoCyc — `http://ecocyc.org` - E. coli K12
- SGD — `http://www.yeastgenome.org/`
- WORMBASE ~~— `http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/`~~ **https://wormbase.org/**
- FlyBase — `http://flybase.org/`
- ...

Nástroje pro analýzu biologických sítí

- Cytoscape
  `http://www.cytoscape.org`
  - vizualizační layouty
  - mapper vizuálních prvků na data
  - filtry
  - pluginy pro práci s biologickými sítěmi
- VisANT
  `http://visant.bu.edu/`
  - prakticky tatáž funkčnost jako Cytoscape
  - podpora hierarchického zanořování (MetaNodes)
  - méně flexibilní prostředí
  - více biologicky-relevantních funkcí

Cytoscape

- analyzátor komplexních interakčních sítí
- původně zaměřený na bioinformatická data
- přístup k webovým službám (databáze)
- funkce pro integraci dat
- platforma pro pluginy
- open source politika

Cytoscape – formát dat

- SIF – Simple Interaction Format (interní formát)
- GML – Graph Modelling Language (obecný formát pro popis grafu, zahrnuje i vizuální informace)
- XGMML – eXtensible Graph Markup and Modeling Language (možnost anotací uzlů a hran daty)
- BioPAX – formát využívající OWL (standard pro popis ontologií)
- SBML – standard pro popis dynamických modelů
- dále OBO, text format CSV, MS Excel XLS format, ...

Cytoscape – základní principy

- sítě organizovány v tzv. sessions (formát cys)
- v rámci session lze spravovat několik sítí
- data o uzlech/hranách jsou sdílena v rámci session
- session zahrnuje též vizuální data včetně nastavení stylů, vnitřní stavy některých pluginů

Cytoscape – základní principy

- možnost vyhledávání a filtrace dat
- datový panel zobrazuje filtrovaná/vyhledaná data
- jednoduše lze vytvořit podsíť z aktuálně vybraných uzlů
- možnost topologických filtrů

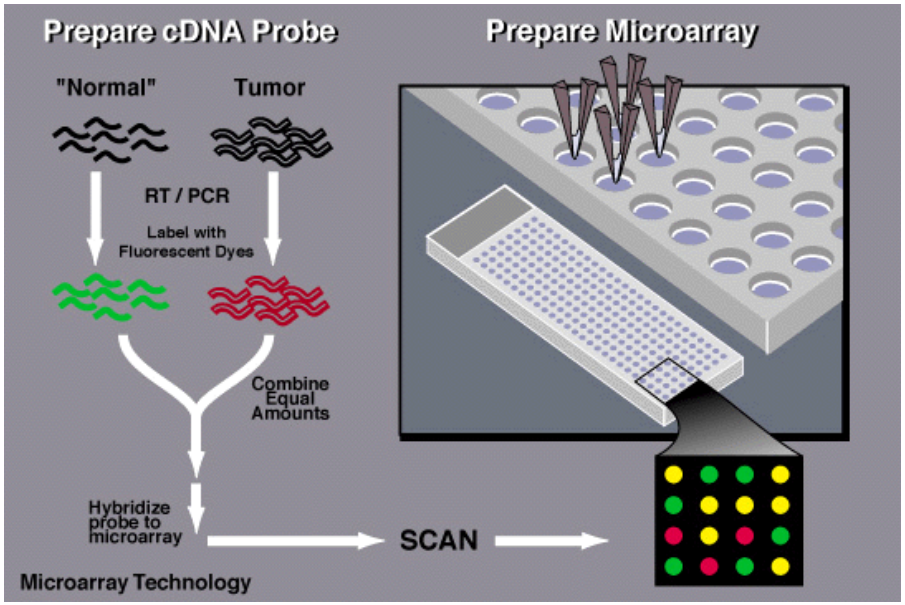
Obsah

Detekce regulačních interakcí

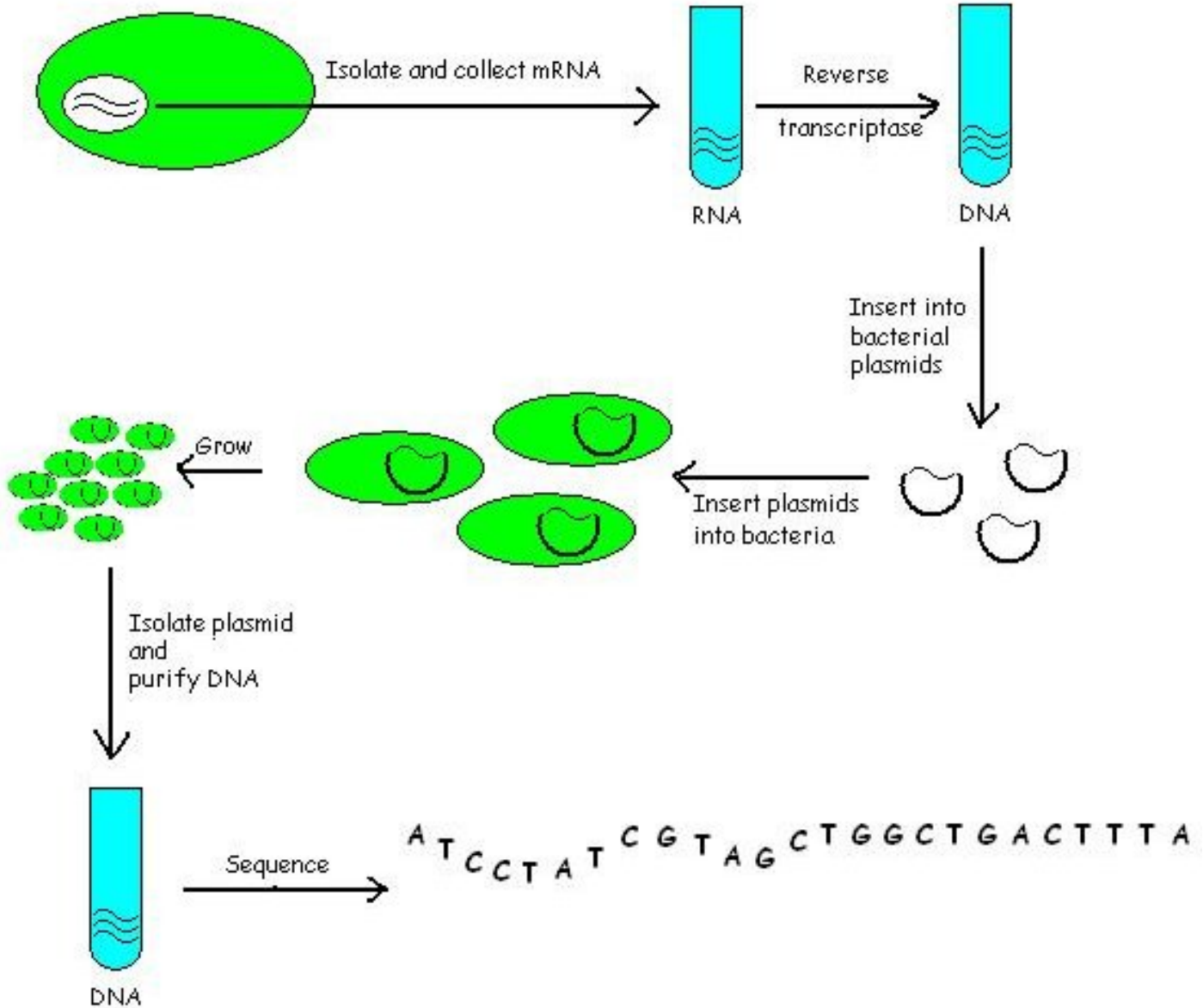
- využití databází promotorových sekvencí
  - prohledávání promotorových sekvencí na přítomnost známých TFBSs
  - TRANSFAC, MATCH, PromoterScan, RegulonDB promoter analysis, ...
- využití DNA mikročipů
  - identifikace genů s podobnými profily exprese a jejich agregace do skupin (tzv. klastrů)
- analýza promotorů ortologických genů (napříč různými druhy)
  - tzv. phylogenetic footprinting
  - viz např. `http://bayesweb.wadsworth.org/binding_sites/index.html` (E. coli)

Měření genové exprese

- nejpoužívanějším nástrojem je technologie DNA microarray
- umožňuje tzv. high-throughput analýzu
  - v daném okamžiku je paralelně nasamplována exprese všech genů v genomu příslušného organismu
  - postaveno na relativním srovnání minimálně dvou různých vzorků
    - exprese v přítomnosti vs. nepřítomnosti O_2
    - exprese při knock-outu určitého genu vs. normální stav
    - ...



Reverzní transkriptáza a cDNA



Reverzní transkriptáza a cDNA

- enzym EC 2.7.7.49 (druh DNA polymerázy)
- objevena v retrovirech [Temin, Baltimore, 1970]
- přepisuje mRNA na jednořetězcovou (komplementární) DNA (tzv. cDNA)
- umožňuje vytvořit knihovnu DNA

Polymerase Chain Reaction (PCR)

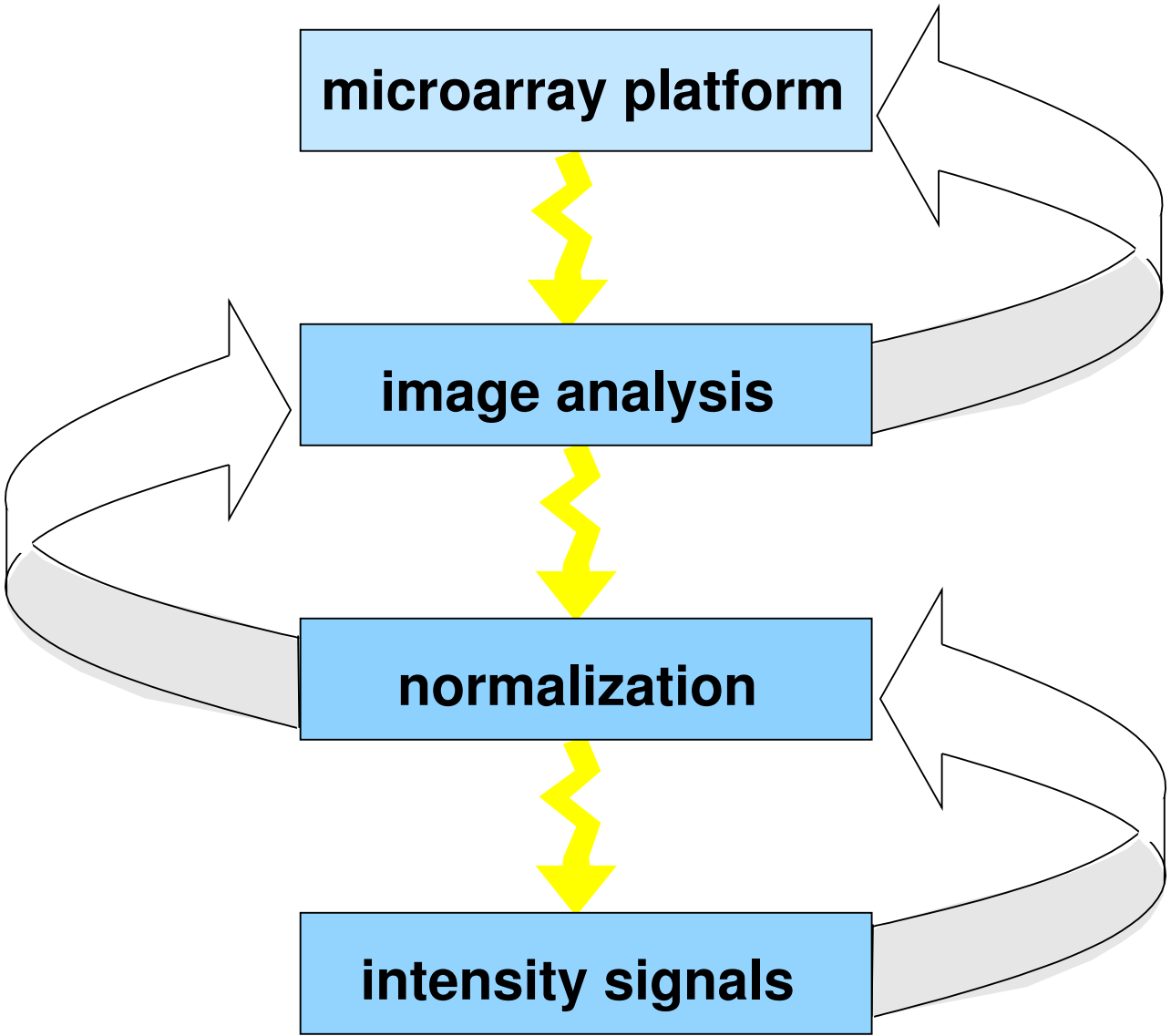
- umožňuje replikaci určité části DNA (tzv. amplifikace)
- DNA je zahřátím rozdělena
- úsek DNA je označen párem oligonukleotidů (15-25 bazí)
  - při snížení teploty hybridizace oligonukleotidů s řetězcem DNA
  - doplnění zbývající sekvence DNA prostřednictvím RNA polymerázy
- `http://www.dnalc.org/resources/animations/pcr.html`
- lze využít i pro mRNA: RT-PCR (reverse transcription PCR)
  - reverzní transkripce mRNA do cDNA
  - amplifikace cDNA (PCR)

Postup při DNA microarray experimentu

1. konstrukce čipu z cDNA knihovny (amplifikace a rozmístění)
2. odběr celkové mRNA z experimentálních vzorků (typicky 2)
3. reverzní transkripce do cDNA asociované s fluorescenčním barvivem
4. hybridizace odebrané cDNA s cDNA na čipu
5. omytí čipu a oskenování výsledku
6. analýza dat
7. komerční čipy používají místo cDNA knihovny skupinu oligonukleotidů pro každý gen
   $\Rightarrow$ pouze jeden vzorek mRNA je analyzován na jednom čipu (porovnání více identicky připravených čipů)

`http://www.bio.davidson.edu/courses/genomics/chip/chip.html`

Workflow produkce výsledků DNA microarray experimentu



Validace a zpracování výsledků

- validace dat separátním měřením koncentrací mRNA nepřímo (pomocí RT-PCR)
  - RT-PCR spuštěna pro oba vzorky (shodný počet kroků PCR)
  - porovnání koncentrací příslušných cDNA
- klastrování dat
  - zjišťování podobnosti mezi datovými vektory
  - agregace do specifických skupin (klastrů)

Databáze microarray dat

- Stanford Microarray Database – různé pohledy na data, filtrace

  `http://smd.stanford.edu/cgi-bin/cluster/drpGetData.pl`
- ArrayExpress – statisticky zpracovaná data

  `http://www.ebi.ac.uk/gxa/`
- Gene Expression Omnibus (GEO)

  `http://www.ncbi.nih.gov/geo/`
- MUSC DNA Microarray Database

  `http://proteogenomics.musc.edu/ma/`
- GenExpDB (E. Coli specifická data)

  `http://genexpdb.ou.edu/`

Klastrování microarray dat

- předpokládejme matici se sondami pro n genů
- uvažujme sadu p experimentů
- pro každý gen i dostáváme vektor $\mathbf{x_i} = (x_{i1}, ..., x_{ip})$ zachycující posloupnost výsledků (tzv. *expresní profil*)
- definujeme míru vzdálenosti $d : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$:

$$d(\mathbf{x_n}, \mathbf{x_m}) = (\sum_{i=1}^{p} |x_{ni} - x_{mi}|^q)^{\frac{1}{q}}$$

- pro $q = 2$ dostáváme Euklidovskou vzdálenost

Klastrování microarray dat

- existují dva hlavní přístupy ke klastrování
  - partitioning – cílem je najít jedno **nejvhodnější** rozdělení do klastrů (parametrem je počet požadovaných klastrů)
    $\Rightarrow$ Self-organizing maps, K-means
  - hierarchické metody – vytvořen celý strom hierarchie
    $\Rightarrow$ kořen – klastr obsahující všechny experimenty, v listech
    $\Rightarrow$ listy – jednoprvkový klastr pro každý experiment
- klastry mohou být identifikovány i pro vektory tvaru $\mathbf{x_j}' = (x_{1j}, ..., x_{nj})$

Algoritmus pro hierarchické klastrování

- nejpoužívanější metoda je tzv. aglomerativní (zdola-nahoru)
- parametrem je míra podobnosti hodnot $d(x_i, x_j)$
- postup (t značí aktuální úroveň):
  1. $t = n \Rightarrow$ inicializuj pro každý gen $i \leq n$: $C_i^n = \{x_i\}$
  2. spoj dva klastery C_k^t a C_l^t s minimální vzdáleností $D(C_k^t, C_l^t)$
  3. update D dle nového rozdělení
  4. $t := t - 1$
  5. iteruj (2-4) dokud $t > 1$

Algoritmus pro hierarchické klastrování

Při aglomeraci se používá míra podobnosti dvou klastrů na téže úrovni t:

- $D(C_k^t, C_l^t) = min_{x_i \in C_k^t, x_j \in C_l^t} d(x_i, x_j)$ (single linkage)
- $D(C_k^t, C_l^t) = max_{x_i \in C_k^t, x_j \in C_l^t} d(x_i, x_j)$ (complete linkage)
- $D(C_k^t, C_l^t) = \frac{1}{|C_k^t||C_l^t|} \sum_{x_i \in C_k^t, x_j \in C_l^t} d(x_i, x_j)$ (average linkage)

Algoritmus pro hierarchické klastrování

- update míry vzdálenosti (krok (3)):

$$D(C_m^{t-1}, C_k^t \cup C_l^t) = \alpha_k D(C_m^t, C_k^t) + \alpha_l D(C_m^t, C_l^t) \\ +\gamma |D(C_m^t, C_l^t) - D(C_m^t, C_k^t)|$$

- single linkage: $\alpha_k = \alpha_l = 0.5$, $\gamma = -0.5$
- complete linkage: $\alpha_k = \alpha_l = 0.5$, $\gamma = 0.5$
- average linkage: $\alpha_i = \frac{|C_i^t|}{|C_k^t| + C_l^t|}$, $i \in \{k, l\}$, $\gamma = 0$

Metoda K-means

- založeno na optimalizaci odchylky mezi expresními profily vzhledem ke středu (průměrnému profilu) klastru
- nejčastěji je tato optimalizace reprezentována minimalizací
- pevně dán počet požadovaných klastrů
- náhodně se inicializují střední profily
  - metoda je přesnější při větším počtu pokusů
- klastry jsou průběžně modifikovány při minimalizaci odchylek (Euklidovské vzdálenosti) od středových profilů

Metoda K-means

- algoritmus K-means má dvě základní fáze
  - výpočet vzdáleností jednotlivých vektorů od vektoru středových hodnot
  - update vzhledem k optimalizační funkci
- nejpoužívanější metrikou je Euklidovská vzdálenost
- vektor středových hodnot je vypočítán jako aritmetický průměr vektorů aktuálně přiřazených danému klasteru

Algoritmus K-means

- vstup: počet iterací (inicializací), počet klastrů K, práh přesnosti ϵ
- náhodně inicializuj rozdělení do klastrů $C_1^1, \ldots C_K^1$ se středy $c_1^1, \ldots, c_K^1$ a vypočítej hodnotu optimalizační funkce W^1
- v i-tém kroku proveď:
  - výpočet $C_1^{i+1}, \ldots, C_K^{i+1}$ – přiřaď každý datový vektor x ke klastru s nejmenší vzdáleností středového vektoru od x
  - přepočítej středové vektory $c_1^{i+1}, \ldots, c_K^{i+1}$ a minimalizuj W^{i+1}
- dokud $\exists k$, $|c_k^i - c_k^{i+1}| \geq \epsilon$, iteruj

Nástroje pro cluster-based analýzu

- klastrování lze využít pro detekci skupin shodně regulovaných genů
- kombinace klastrování dle genů a experimentů
  - odhady regulátorů jednotlivých klastrů
  - odhady programů regulace
- nástroje
  - Genomica (dříve GeneXPress)

    `http://genomica.weizmann.ac.il/`
  - FunCluster (balík pro R)

    `http://cran.r-project.org/web/packages/FunCluster/index.html`
  - STEM

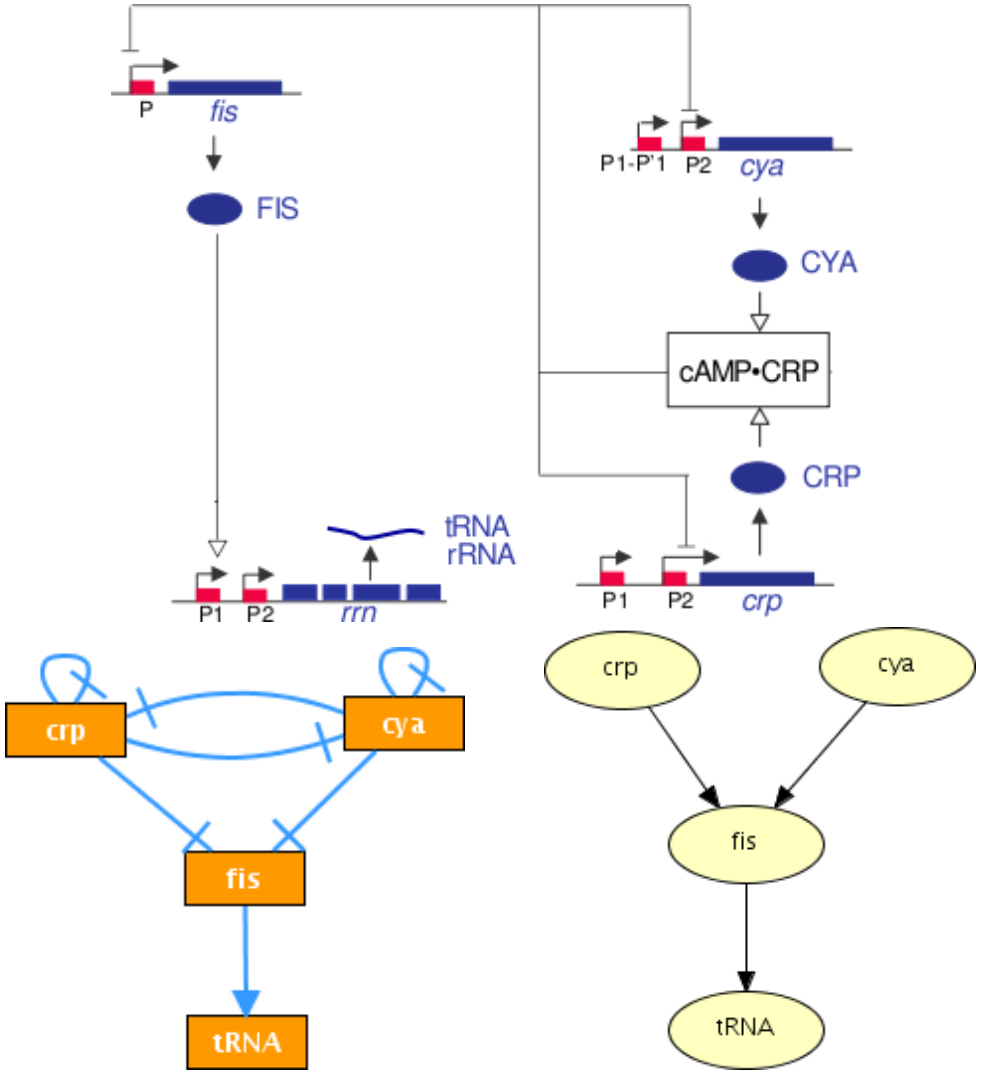
    `http://www.cs.cmu.edu/~jernst/stem/`
  - EisenLab tools

    `http://rana.lbl.gov/EisenSoftware.htm`

Předpověď (reinženýring) regulačních sítí

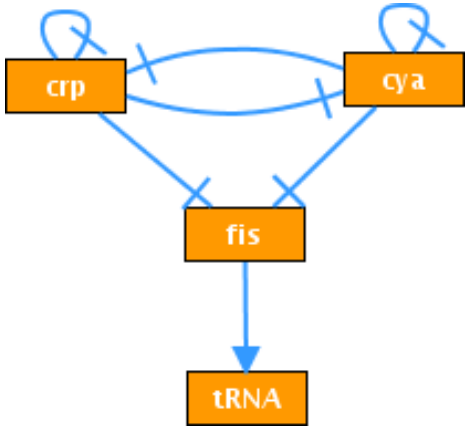
- regulační sítě lze předpovídat z microarray dat
- předpověď struktury sítě
  - detekce podmíněných závislostí proměnných
  - charakter korelace proměnných
- předpověď dynamiky proměnných
  - fitování naměřených dat na (spojitý) model
  - pravděpodobnostní rozložení diskrétních hodnot

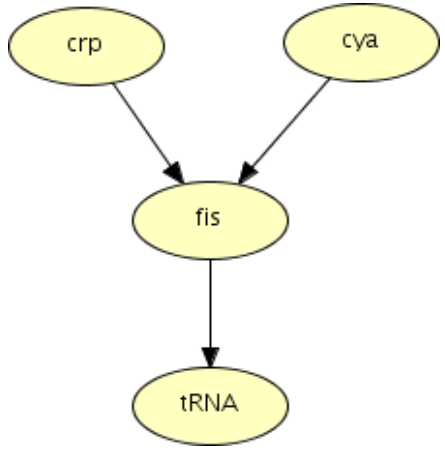
Předpověď (reinženýring) regulačních sítí



Předpověď (reinženýring) regulačních sítí

Boolovské vs. Bayesovské sítě





$$crp(t+1) = \neg crp(t) \wedge \neg cya(t)$$
$$cya(t+1) = \neg cya(t) \wedge \neg crp(t)$$
$$fis(t+1) = \neg crp(t) \wedge \neg cya(t)$$
$$tRNA(t+1) = fis(t)$$

$$P(X_{crp})$$
$$P(X_{cya})$$
$$P(X_{fis}|X_{crp}, X_{cya})$$
$$P(X_{tRNA}|X_{fis})$$

Předpověď (reinženýring) regulačních sítí

Bayesovské sítě

$$P(V|W) = \frac{P(V,W)}{P(W)}$$

$$P(W|V) = \frac{P(W,V)}{P(V)}$$

$$P(V,W) = P(W,V) = P(V|W) \cdot P(W) = P(W|V) \cdot P(V)$$

Bayesův vzorec:

$$P(V|W) = \frac{P(W|V) \cdot P(V)}{P(W)}$$

Obecně pro pravděpodobnost současných jevů platí řetězové pravidlo:

$$\begin{aligned} P(V,W,Y) &= P(V|W,Y) \cdot P(W,Y) \\ &= P(V|W,Y) \cdot P(W|Y) \cdot P(Y) \end{aligned}$$

Předpověď (reinženýring) regulačních sítí

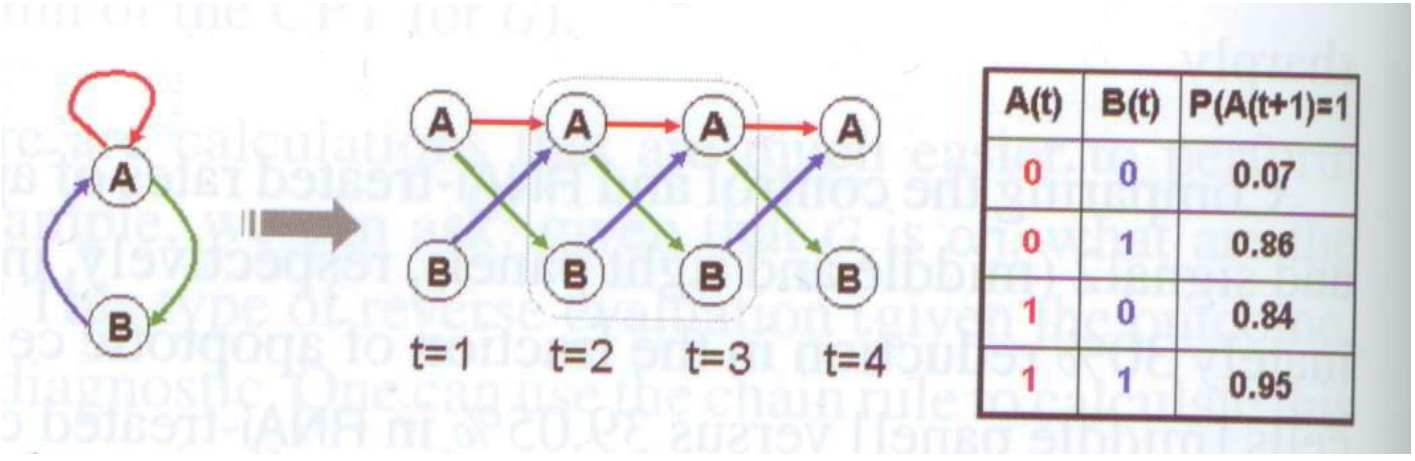
Algoritmy pro bayesovské sítě

- strojové učení z experimentálních dat:
  - algoritmy učení struktury
  - algoritmy učení pravděpodobnostního rozložení
    např. Expectation Maximization (EM) – iterativní metoda maximalizující $P(data|model)$
  - kombinované algoritmy
- pro úspěšný výsledek vyžadována rozsáhlá sada dat
- nástroje:
  - Hugin (`http://www.hugin.com/`)
  - Genomica (`http://genomica.weizmann.ac.il/`)

Předpověď (reinženýring) regulačních sítí

Algoritmy pro bayesovské sítě

- problémem jsou zpětné vazby (cykly v síti)
- řešením je unfolding v diskrétním čase:



A(t)	B(t)	P(A(t+1)=1
0	0	0.07
0	1	0.86
1	0	0.84
1	1	0.95

- původní síť s n uzly je nahrazena sítí s $2n$ uzly
- tabulka podmíněných pravděpodobností charakterizuje pravděpodobnost přechodů mezi jednotlivými konfiguracemi