**Univerzita Karlova v Praze**

**Přírodovědecká fakulta**

**Katedra filosofie a dějin přírodních věd**

Charles University in Prague, Faculty of Sciences, Department of Philosophy and History of Science

**Disertační práce**

PhD Thesis

# Architektura regulační sítě metabolismu

The architecture of regulatory network of metabolism

**Mgr. Jan Geryk**

Školitel/Supervisor: **Prof. RNDr. Jaroslav Flegr, CSc.**

**Praha, 2012**

Prague, 2012

**Prohlášení:**

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 10.12.2012 Podpis

**Poděkování**

Na tomto místě děkuji celému kolektivu Katedry filosofie a dějin přírodních věd za přátelské a inspirativní prostředí, které vytváří. Děkuji mému původnímu školiteli Františku Slaninovi za odborný dozor nad mými nápady a významný podíl na mém dovzdělávání se v oblasti matematiky. Dále děkuji mému současnému školiteli Jaroslavu Flegrovi, za poskytnutou tvůrčí volnost, která mi umožnila vypracování třetí práce o regulační kanalizaci metabolických sítí a za moudré vhledy do některých obecných otázek mé práce. Děkuji Aleši Kuběnovi za četné diskuze o matematických a statistických problémech, se kterými jsem byl konfrontován. Děkuji Tereze za inspiraci. Děkuji také rodičům za podporu, zejména v podobě nedělních obědů.

**Abstrakt**

Předkládaná disertační práce se zabývá modularitou metabolických sítí a především architekturou regulační sítě metabolismu, která reprezentuje přímé regulační interakce mezi metabolity a enzymy.

V první práci se zabývám problematikou tzv. "modularity measure", což je kvantitativní míra modularity sítě používaná pro účely identifikace modulů. Bylo zjištěno, že při maximalizaci této veličiny v síti může dojít k chybnému sloučení dvou jednoznačně vyjádřených modulů v jeden. Maximální velikost modulu u kterého existuje riziko, že je tvořen dvěma moduly je známa jako rozlišovací limit modularity measure. V mé první práci je tento rozlišovací limit zobecněn, což umožňuje nahlédnout jeho podstatu v použití nulového modelu. Zároveň je zde ukázáno, že riziko chybného sloučení existuje i v případě větších modulů, než bylo uváděno v původní práci.

Druhá práce je zaměřena na otázku, jak se změní modularita metabolické sítě *E.coli* po přidání regulačních vazeb. Bylo zde ukázáno, že modularita mírně nicméně signifikantně vzroste, zaměříme-li se na modulární jádro sítě. Identifikované moduly jsou funkčně interpretovatelné jako regulačně autonomí části metabolismu. Zvýšení modularity vzhledem k nulovému modelu lze považovat za nepřímý důsledek potřeby lokální regulace některých částí metabolické sítě. Vznik zpětnovazebné regulace na krátkou vzdálenost zvýší lokálně hustotu hran a může tedy přispět i k celkové asymetrii v distribuci hran sítě a v důsledku toho ke zvýšení modularity.

Poslední práce se zaměřuje na regulační kanalizaci metabolických cest. Zobecňuje tak pozorování specifické regulační struktury jednoho modulu identifikovaného v předchozí práci. Auto-kanalizace je definována z topologického hlediska jako potenciál metabolické cesty tlumit reakce odbočující z této cesty a zároveň netlumit vlastní aktivitu. Tento potenciál je realizován existencí inhibičních regulačních vazeb vedoucích od metabolitů produkovaných metabolickou cestou k reakcím, které odbočují z této metabolické cesty. V předkládané práci bylo ukázáno, že reálná regulační síť vykazuje signifikantně vyšší auto-kanalizaci metabolických cest než randomizované soubory regulačních sítí. Současně zde bylo prokázáno, že reakce odbočující z metabolické cesty jsou často inhibovány metabolitem, jež je produktem jiné cesty, začínající odbočující reakcí a tento metabolit neinhibuje žádnou z reakcí původní cesty. Absence inhibice původní cesty je selektována pravděpodobně proto, že by mohla zamezovat produkci jiných metabolitů závislých na aktivitě původní cesty.

Signifikance studovaných vlastností regulační sítě metabolismu E.coli vzhledem k nulovým modelům naznačuje, že jsou projevem evolučních adaptací regulačního systému. Potvrzuje se tak očekávání, že topologie regulační sítě by měla být do značné míry optimalizovaná a zároveň se otevírájí nové možnosti dalšího výzkumu.

**Abstract**

The thesis focus on the modularity of metabolic network and foremost on the architecture of regulatory network representing direct regulatory interactions between metabolites and enzymes.

I focus on the "modularity measure" in my first work. Modularity measure is quantitative measure of network modularity commonly used for module identification. It was showed that algorithms using this measure can produce modules that are composed of two clearly pronounced sub-modules. Maximum size of module for which there is a risk that is is composed of two sub-modules is called resolution limit of modularity measure. In my first work I generalize resolution limit of modularity measure. The generalized version provide insight to the origin of resolution limit in the null-model used by modularity measure. Moreover it is showed that the risk of omitting of sub-modular structures applies for bigger modules than mentioned in the original publication.

The second work is focused on the question how does the modular structure of *E. coli* metabolic network change if we add regulatory interactions. I find that the modularity of modular core of network slightly increase after regulatory edges addition. The modularity increase is significant with respect to randomized ensemble of regulatory networks. Identified modules contains many regulatory feedbacks and due to them are functionally interpretable. The higher modularity of combined network which respect to the null-model can be viewed as a consequence of the need of local feedback regulation in selected parts of metabolic network. Emergence of feedback regulation on short distance increase local edges density and can also increase asymmetry in edge density distribution and as a consequence modularity of the network.

Last work focus on the regulatory canalization of metabolic paths. The regulatory canalization is a generalization of the specific regulatory structure observed within one module identified in our previous work. Auto-canalization is defined from the topological point of view as a potential of metabolic path to inhibit reactions branching-out of this path and at the same time leave the path unaffected by this inhibition. This potential is realized by inhibitory interactions between metabolites produced by the path and reactions branching out of this path. I showed that real regulatory network of E. coli exhibit significantly increased auto-canalization of metabolic paths which respect to randomized counterpart. Moreover it was showed that reactions branching out of the metabolic path are typically inhibited by the metabolite which is produced by another simple pathway starting in the out-branching reaction and at the same time the inhibitory metabolite does not inhibits any reaction within the main path. The absence of inhibitory effect on the main path is probably selected, because saturated state of one out-branching pathway does not guarantee that other out-branching pathways are also saturated and potential inhibition of the main pathway can prevent production of other metabolites whose synthesis depend on the main pathway.

The significance of the studied properties of the regulatory network implies evolutionary adaptations imprinted on the topological scale. The results confirm expectation of optimized topology of regulatory network and open new ways for further research.

**Obsah**

## Věda o sítích

Názorně si síť můžeme představit jako množinu bodů, které jsou mezi sebou propojeny úsečkami. V matematice je síť definována jako množina libovolných entit spolu s informací, které dvojice z těchto entit jsou v relaci sousedství. Této nejjednodušší formalizaci sítě se v matematice říká graf, výše zmíněné entity nazýváme vrcholy nebo uzly a dvojice vrcholů, které jsou v relaci sousedství nazýváme hrany grafu. Je mnoho fenoménů, které nás obklopují a které jsou přirozeně reprezentovány jakožto síť respektive graf. Například lidské společenství je asi nejdéle známým příkladem systému, který lze reprezentovat grafem a bude možná první, u kterého pochopíme jeho dynamiku díky obrovskému množství dat pocházejících z mobilní komunikace, internetu nebo GPS (Barabasi, 2009). Jednotliví lidé mohou představovat vrcholy grafu a přítomnost hrany může být definována např. jako existence přátelství mezi dvěma lidmi. Jiným příkladem je protein-interakční síť v buňce. Vrcholy grafu zde reprezentují proteinové molekuly a hrana mezi dvojicí proteinů je položena tehdy, když je experimentálně potvrzeno, že tyto dva proteiny spolu vytváří vazbu. Věda o sítích zkoumá „strukturu propojení“ grafů, které reprezentují rozličné fyzikální, biologické, sociální nebo informační fenomény. Její masivní rozvoj byl podmíněn vznikem internetu, a pokroky v molekulární biologii a biochemii, které umožnují mapovat molekulární interakce probíhající v buňkách. Ve většině vysoce složitých systémů jako je lidský mozek, lidské společenství, internet, ale i jediná buňka nacházíme nějakou formu sítě. Soustředíme-li se na buňku, je zřejmé, že právě interakce mezi molekulami, které ji konstituují, jsou podstatou jejího adaptabilního chování. Grafová reprezentace nitrobuněčných interakcí tedy musí zachycovat podstatnou část složitosti buňky, esenciální pro její funkci. Můžeme tedy říci, že věda o sítích je zároveň vědou o komplexních systémech (Barabasi, 2005, Barabasi, 2007, Barabasi, 2012).

### *Distribuce konektivit*

Existuje celá řada kvantitativních vlastností, které graf charakterizují z různých hledisek. Důležitou charakteristikou je distribuce počtu sousedů jednotlivých vrcholů v grafu, dále jen distribuce konektivit. Právě distribuce konektivit byla jednou z prvních vlastností, která se na sítích z reálného světa studovala. Bylo opakovaně zjištěno, že většina těchto sítí má mocninné rozdělení konektivit vrcholů (Albert et al., 1999, Albert and Barabasi, 2002). Toto rozdělení

můžeme vyjádřit jako pravděpodobnost $p(k)$, že náhodně zvolený vrchol v grafu bude mít $k$ sousedů: $p(k) = ck^{-\gamma}$, kde $k$ je počet sousedů, $c$ je normalizační konstanta zajišťující, že součet $p(k)$ pro všechna $k$ bude roven jedné a $\gamma$ je parametr, který pro většinu doposud analyzovaných sítí leží v intervalu (2,3). Mocninná distribuce konektivit přiřazuje relativně vysokou pravděpodobnost výskytu vrcholů s vysokou konektivitou ve srovnání s distribučními funkcemi, kde hustota pravděpodobnosti klesá exponenciálně s konektivitou. Uzly s vysokou konektivitou mnohonásobně převyšující průměr jsou v sítích s mocninným rozložením konektivity relativně častým jevem a nazývají se huby.

Pomocí modelů můžeme ověřovat, zdali předpokládané růstové mechanismy generují sítě se stejnou nebo podobnou vlastností, jakou má síť, která je předmětem výzkumu. Minimálním předpokladem o růstovém mechanismu sítě je zcela náhodné přidávání hran v rámci množiny vrcholů. Tuto myšlenku rozvinuli dva maďarští matematici Erdős a Rényi do podoby tzv. ER modelu (Erdős and Rényi, 1959a, Erdős and Rényi, 1959b). ER model je množina všech grafů, které je možno vytvořit máme-li k dispozici *n* vrcholů, spolu s mírou, která přiřazuje každému grafu z této množiny hodnotu pravděpodobnosti. Náhodný graf z tohoto souboru vygenerujeme tak, že spojíme každou dvojici vrcholů se stejnou pravděpodobností *p*. Je zřejmé, že pro konstantní *p* budou grafy s různým počtem hran vznikat s různou pravděpodobností. Nejpravděpodobnější bude vznik grafů s počtem hran odpovídajícím střední hodnotě v tomto souboru. Naprostá většina grafů produkovaných ER modelem má binomiální distribuci konektivit, která je charakteristická tím, že většina uzlů bude mít počet sousedů blízký průměru. Vysoké hodnoty jsou zde extrémně nepravděpodobné. ER model může generovat i sítě s mocninným rozdělením konektivit ale jejich vznik náhodným pokládáním hran je velmi málo pravděpodobný, zvláště u větších sítí. Porovnání reálných sítí s ER modelem tedy vedlo k závěru, že za mocninnou distribucí konektivit nejspíš stojí složitější mechanizmus růstu než je implikován ER modelem.

Barabasi and Albert (1999) navrhli pro vysvětlení mocninné distribuce konektivit v reálných sítích jednoduchý růstový mechanismus tzv. „preferential attachment“. Tento mechanismus je možno simulovat tak, že začneme s malým počtem vrcholů (s nenulovými konektivitami), v každém kroce přidáme ke stávající síti 1 nový vrchol, který spojíme hranou s každým stávajícím vrcholem a to s pravděpodobností, která je přímo úměrná počtu sousedů stávajícího vrcholu. Přesněji řečeno, pravděpodobnost toho, že stávající vrchol $i$ bude spojen hranou s nově přidávaným vrcholem je, $p_i = k_i / \sum_j k_j$, kde $k_i$ je počet sousedů uzlu $i$ a suma ve jmenovateli jde přes všechny stávající vrcholy. Vrcholy s vyšší konektivitou budou tedy

získávat nové sousedy rychleji než vrcholy s nižší konektivitou. Sítě generované tímto modelem vykazují mocninné rozdělení konektivit.

Vraťme se opět k buňce a položme si otázku jestli se princip „preferential attachment" uplatňuje v nitrobuněčné protein-interakční síti. Eisenberg and Levanon (2003) analyzovali tuto otázku na protein-interakční síti kvasinky *Saccharomyces cerevisiae*. Autoři nejprve rozdělili proteiny do 4 skupin podle evolučního stáří odhadnutého na základě počtu fylogenetických skupin, které mají homologický protein. Zjistili, že průměrná konektivita proteinu roste s evolučním stářím, což je ve shodě s Barabasi-Albert modelem. Tento fakt by však vysvětloval i model, kde pravděpodobnost získání nového souseda závisí pouze na čase ve kterém byl stávající uzel přidán do sítě. Autoři se proto zaměřili na skupinu proteinů ze čtvrté, nejstarší skupiny, která by měla obsahovat zhruba stejně staré proteiny. Zjistili, že proteiny ze 4. skupiny mající vysokou konektivitu v rámci této skupiny mají i vysoký počet sousedů ve skupinách evolučně mladších. Vyneseme-li závislost konektivity proteinů ze 4. skupiny- počítanou v rámci 4. skupiny na počtu sousedů, které tyto proteiny mají v evolučně mladších skupinách dostaneme v logaritmických souřadnicích rostoucí lineární závislost. Tento fakt poukazuje k tomu, že v protein-interakčních sítích se uplatňuje lineární forma preferential attachment. Lineární forma preferential attachment byla detekována i v jiných sítích, např. v citačních sítích nebo v síti internetu (Newman, 2001, Jeong et al., 2003).

Preferential attachment v protein-interakčních sítích lze vysvětlit na základě genové duplikace. Za předpokladu, že geny jsou v evoluci duplikovány náhodně, proteiny s vysokou konektivitou budou mít větší pravděpodobnost, že protein produkovaný náhodně duplikovaným genem bude jedním z jejich interakčních sousedů. Tento nový protein pak může mutovat a funkčně se specializovat při zachování vazebné interakce s původním proteinem s vysokou konektivitou (Barabasi and Oltvai, 2004).

Nedávná práce ukázala, že preferential attachment může být důsledkem lokální optimalizace kompromisu mezi podobností a popularitou při výběru uzlu se kterým bude nově přidávaný vrchol spojen (Papadopoulos et al., 2012).

Distribuce konektivit ovlivňuje některé významné vlastnosti sítě. Např. bylo dokázáno, že v grafu s mocninnou distribucí konektivit se infekční proces šíří velmi efektivně. Tento proces je charakterizován hodnotou pravděpodobnosti přenosu infekce z infekčního vrcholu na sousední nenakažený vrchol. Např. v ER grafu existuje kritická hodnota této pravděpodobnosti, pod kterou nedojde k masovému rozšíření infekce, tj. epidemii. V grafu s mocninou distribucí konektivit je tato kritická hodnota rovná nule a tedy i choroba s velmi

malou infekčností může vyústit v epidemii (Pastor-Satorras and Vespignani, 2001). Albert et al. (2000) dále dokázali, že sítě s mocninnou distribucí konektivit jsou robustní vůči náhodným útokům. Robustnost byla kvantifikována jako velikost největší souvislé pod-sítě, kterou získáme po odstranění určitého počtu vrcholů a jako délka nejdelší z nejkratších cest mezi dvojicemi vrcholů. Při náhodném odstraňovaní vrcholů v síti s mocninnou distribucí konektivit velikost největší souvislé podsítě klesá lineárně s počtem odstraněných vrcholů. V ER grafu pozorujeme již po odstranění malého počtu vrcholů fázový přechod za kterým je velikost největší souvislé komponenty blízká jedné. Tento výsledek je snadno vysvětlitelný existencí hubů v sítích s mocninným rozdělením konektivit. Je relativně malá pravděpodobnost, že náhodně odstraněný vrchol bude hub, s největší pravděpodobností to bude periferní vrchol s nízkou konektivitou, jehož odstranění naruší celkovou souvislost grafu pouze minimálně. Huby tak v průběhu náhodných útoků dlouhou dobu udržují celkovou souvislost grafu. Je zřejmé, že cílený útok na huby naopak povede k rychlé ztrátě souvislosti.

Distribuce konektivit představuje základní grafovou vlastnost, která se zdá být universální pro reálné sítě. Chceme-li nalézt specifické vlastnosti grafu o kterých si můžeme být jisti, že nejsou důsledkem distribuce konektivit, je nutno tyto vlastnosti srovnávat s nulovým modelem, který generuje sítě se stejnou distribucí konektivit jako má zkoumaná síť.

### *Vzdálenost v grafu*

Jiná důležitá charakteristika grafu je průměrná nejkratší vzdálenost mezi dvojicemi vrcholů (dále jen průměrná vzdálenost). Tato veličina charakterizuje graf s hlediska efektivity přenosu informace, je zřejmé, že čím kratší bude průměrná vzdálenost mezi vrcholy tím rychleji a "levněji" je možno v síti transportovat různé "komodity". Nejspíše první experiment s cílem změřit tuto veličinu v reálné síti učinil americký psycholog Stanley Milgram.V Nebrasce rozdal náhodným lidem velké množství dopisů, které obsahovaly instrukce jakým způsobem mají být tyto dopisy doručeny jedinému adresátovi v Bostonu. Lidé, kteří obdrželi dopisy, měli poslat tento dopis někomu, koho znají osobně a kdo se podle jejich soudu nachází nejblíže adresátovi těchto dopisů. Po určitém čase byl experiment ukončen a Milgram spočítal, že průměrný počet osob přes které se dopis dostal až k adresátovi je 6. Z hlediska ER modelu není tento fakt překvapivý. Je známo, že průměrná vzdálenost mezi dvojicemi vrcholů v ER modelu je úměrná hodnotě $\ln(N)/\ln(k)$, kde $N$ je počet vrcholů v síti a $k$ je průměrná konektivita v síti. V Barabási-Albert modelu je průměrná vzdálenost mezi vrcholy úměrná $\ln(\ln(N))$ (Cohen and Havlin, 2003). Ani jeden z těchto modelů však nevysvětluje vysokou

hodnotu zhlukovacího koeficientu v reálných sociálních sítích. Tento koeficient vyjadřuje, jaká je v průměru pravděpodobnost, že dva sousední vrcholy nějakého dalšího vrcholu jsou spojeny hranou. Model Wattse a Strogatze (1998) velmi elegantním způsobem vysvětluje jak může koexistovat vysoký zhlukovací koeficient a logaritmická průměrná vzdálenost v síti s relativně malou hustotou hran. Autoři začali s pravidelným grafem, jehož vrcholy jsou uspořádány v kruhu a každý z nich je spojen malým počtem hran jen s nejbližšími $k$ vrcholy. V takovém grafu je vysoký zhlukovací koeficient a relativně vysoká hodnota průměrné vzdálenosti, rovná $N/2k$. Autoři náhodně přeuspořádali hrany tak, že pro každý vrchol $i$ a hranu, která z něj vychází s pravděpodobností $p$ odpojili tuto hranu od souseda vrcholu $i$ a spojili ji s náhodně vybraným vrcholem v grafu. Zjistili, že zhlukovací koeficient je pro nízké hodnoty $p$ konstantní zatímco průměrná vzdálenost vrcholů s rostoucím $p$ rychle klesá k hodnotě odpovídající logaritmu počtu vrcholů. Pro hodnotu $p$=1 dostaneme klasický ER graf. Existuje tedy interval hodnot $p$ ve kterém dostaneme sítě s vysokým zhlukovacím koeficientem a logaritmickou průměrnou vzdáleností mezi vrcholy, které dobře odpovídají reálným sociálním sítím.

Detailnější analýzy reálných sítí ukázaly, že většina těchto sítí vykazuje signifikantně větší průměrnou vzdálenost mezi vrcholy než randomizované verze těchto sítí, zachovávající hodnotu konektivity pro každý vrchol (Zhang and Zhang, 2009, Xu et al., 2011). Rozdíl vzhledem ke střední hodnotě randomizovaného souboru je ale relativně malý. Tedy platí, že reálné sítě vykazují logaritmickou závislost průměrné vzdálenosti na počtu vrcholů. Zvýšená hodnota průměrné vzdálenosti je však v rozporu s představou maximalizace efektivity přenosu informací. Na druhou stranu bylo by naivní se domnívat, že v tak složitých systémech, jaké reálné sítě bezpochyby jsou, je průměrná vzdálenost jediné optimalizační kritérium zajišťující jejich stabilitu. Bylo prokázáno, že zvyšování modularity grafu vede ke zvětšování průměrné vzdálenosti mezi vrcholy (Zhang and Zhang, 2009). Modularita je korelátem evolvability systému a podílí se i na jeho robustnosti (Kashtan and Alon, 2005). Vyšší průměrná délka cesty v reálných sítích tedy může být výsledkem kompromisu mezi efektivitou přenosu, evolvabilitou a robustností (Zhang and Zhang, 2009).

### *Modularita*

Modularita je další významnou vlastností, která je na reálných sítích intenzivně zkoumána. Graf je modulární, když je možné jej rozdělit na podgrafy tak, že v těchto podgrafech bude významně vyšší hustota hran než mezi těmito podgrafy. Podgrafy ve kterých je významně

vyšší hustota hran než v jejich okolí nazýváme moduly. V sociálních sítích jsou moduly tvořeny skupinami lidí, kteří sdílejí společné zájmy nebo například pracovní podmínky. V síti www jsou moduly tvořeny www stránkami s podobným tématem. V protein-interakčních sítích jsou nejsilněji vyjádřené moduly tvořeny skupinami proteinů, které spolu vytváří nějaký vazebný komplex. Tyto moduly je možno považovat za triviální. Kromě proteinových komplexů jsou v protein-interakčních sítích i slaběji vyjádřené moduly, které bývají považovány, za systémy vykonávající nějakou jasně definovanou buněčnou funkci. Tato interpretace je však problematická (Wang and Zhang, 2007, Lewis et al., 2010). Podobná situace je v metabolických sítích, kde se funkční interpretace modulů zakládá pouze na procentuálním zastoupení metabolických KEGG kategorií v jednotlivých modulech. Moduly detekované v metabolických sítích jsou typicky složeny s více KEGG kategorií, což dále komplikuje funkční interpretaci (Guimera and Amaral, 2005, Zhao et al., 2006).

Analýza modularity grafu je spojena s řadou technických problémů a je velmi intenzivně zkoumána. Asi nejznámější a nejvíce používaná kvantitativní míra modularity grafu je tzv. "modularity measure". Máme-li libovolné rozdělení grafu na podgrafy, modularity measure je definováno jako rozdíl frakce hran obsažené v těchto podgrafech (potenciálních modulech) a střední hodnoty stejné kvantity v nulovém modelu (Newman and Girvan, 2004, Newman, 2004, Clauset et al., 2004). Nulovým modelem zde myslíme soubor všech kombinatoricky možných grafů se stejnou sekvencí konektivit jako má zkoumaný graf. Obecně lze modularity measure vyjádřit jako: $Q = \sum_{i=1}^{m}(f_i - \langle f_i \rangle)$, kde $f_i$ je frakce hran grafu obsažená v podgrafu $i$, $\langle f_i \rangle$ značí střední frakci hran v podgrafu $i$ v nulovém modelu a suma probíhá všechny předem určené podgrafy. Budou-li hrany grafu koncentrovány v modulech v podobné hustotě jako v nulovém modelu dostaneme hodnoty $Q$ blízké nule, budou-li naopak hrany koncentrovány v modulech v mnohem větší míře než v nulovém modelu bude se hodnota $Q$ blížit jedné. Většina algoritmů pro identifikaci modulů v síti nějakým způsobem maximalizuje $Q$ v rámci prostoru všech možných rozdělení grafu na podgrafy. Za relevantní modulární strukturu považujeme rozdělení sítě na podgrafy, kterému přísluší nejvyšší hodnota $Q$ dosažená v průběhu algoritmu, označme ji $Q_{max}$. Nejlepších výsledků bylo dosaženo pomocí stochastických optimalizačních metod jako je simulované žíhání (Guimera and Amaral, 2005).

Je známo, že hodnota modularity measure pouze omezeně vypovídá o "skutečné modularitě sítě. Například v případě metabolických a protein-interakčních sítí dostáváme relativně vysoké hodnoty $Q_{max} \approx 0.8$ (Zhao et al., 2006, Zhang and Zhang, 2009). Randomizované

soubory, které zachovávají sekvenci konektivit těchto metabolických a protein-interakčních sítí však v průměru vykazují hodnotu $Q_{max} \approx 0.7$. Vysoké hodnoty $Q_{max}$ jsou typické pro náhodné sítě s nízkou průměrnou konektivitou (Hu et al., 2010b). Řídce propojená síť umožňuje zvolit rozdělení sítě na moduly tak, že bude většina hran koncentrována v modulech, tím maximalizujeme člen $f_i$. Zároveň je v souboru grafů příslušného nulového modelu velmi nízká střední frakce hran v typickém podgrafu $\langle f_i \rangle$, což je způsobeno nízkou průměrnou konektivitou zkoumané sítě. Dokladem budiž práce Bagrowa (2012), která ukazuje, že graf typu strom, který má nejmenší možnou hustotu hran při zachování spojitosti, může dosahovat velmi vysokých hodnot $Q$. Při analýzách je tedy nutné srovnat hodnotu $Q_{max}$ změřenou v reálné síti s randomizovaným souborem sítí. Touto cestou je možno učinit odhad pravděpodobnosti s jakou nalezneme síť se stejnou nebo vyšší hodnotou než je $Q_{max}$ v souboru všech grafů s danou sekvencí konektivit, který představuje nulový model. Pro tyto účely je možno použít z-skóre, $z = (Q_{max} - \langle Q_{max} \rangle_{NM})/\sigma_{NM}$, kde $Q_{max}$ je modularita zkoumané sítě, $\langle Q_{max} \rangle_{NM}$ je průměrná modularita v randomizovaném souboru a $\sigma_{NM}$je směrodatná odchylka $Q$ v tomto souboru (Fortunato, 2010). Nedávné práce navrhují nové indexy lépe odrážející míru modularity sítě. Tyto indexy jsou odvozeny ze stability modulární struktury vzhledem k náhodným poruchám (Hu et al., 2010b, Hu et al., 2010a)

Existuje další problém spojený s modularity measure, známý jako rozlišovací limit. Fortunato a Barthelemy (2007) dokázali, že za určitých podmínek je hodnota $Q$ vyšší pro rozdělení sítě, které zcela odporuje přirozené představě modulu, oproti rozdělení, které je v souladu s touto představou. Autoři uvažují dva podgrafy, každý se stejným počtem hran ($l$), Tyto dva podgrafy jsou vzájemně spojeny jednou hranou a každý z nich je dále spojen jednou hranou se zbytkem grafu. Dále uvažujme dvě možná rozdělení grafu na moduly. V rozdělení $A$ bude každý ze dvou podgrafů představovat samostatný modul, v rozdělení $B$ budou tyto dva podgrafy sjednoceny do jednoho modulu. Těmto rozdělením přísluší hodnoty $Q(A)$ a $Q(B)$. Fortunato a Barthelemy analyzovali podmínky, za kterých platí, že $Q(A) < Q(B)$. Ukázali, že tato nerovnost je splněna, když $l < l_r = \sqrt{L/2} - 1$, kde $L$ je počet hran v grafu. Hodnota $l_r$ nezávisí striktně na počtu vrcholů v uvažovaných podgrafech, můžeme tedy zvolit počet uzlů v podgrafech $(n)$ tak aby měly maximální počet hran tj. $l = 0.5n(n-1)$ a zároveň platilo $l < l_r$. V takovém případě je autonomie podgrafů jednoznačná, přesto bude mít rozdělení, kde jsou sloučeny v jeden modul vyšší hodnotu $Q$. Maximalizujeme-li $Q$ za účelem identifikace modulů v síti, je zde riziko, že nalezené moduly velikosti $\leq 2l_r$ budou složeny ze dvou jednoznačně vyjádřených modulů. Je samozřejmě možné maximalizovat $Q$ v modulech s

kritickou velikostí a tímto způsobem nalézt moduly v rámci modulů. Není ale jasné kdy jsou submoduly relevantnější než modul ve kterém byly nalezeny a navíc jsou při aplikaci algoritmu na různé pod-sítě používány různé nulové modely.

Existence rozlišovacího limitu je důsledkem použití nulového modelu. Střední hustota hran v určitém podgrafu totiž závisí, kromě lokálních vlastností jako je konektivita vrcholů v tomto podgrafu, i na globálních vlastnostech grafu, v případě modularity measure na počtu hran grafu ($L$).

Přes všechny nedostatky je modularity measure v praxi použitelné, jak dokládají testy na syntetických sítích s definovanou modulární strukturou.

Analýzy modularity ($Q$) reálných sítích ukazují, že většina těchto sítí vykazuje signifikantně vysokou hodnotu $Q_{max}$ vzhledem k randomizovanému souboru. Metabolické a protein-interakční sítě dosahují velmi vysokých hodnot modularity $Q_{max} \approx 0.8$, ale rozdíl vzhledem ke střední hodnotě v randomizovaném souboru je relativně malý ($\approx 15\%$) (Guimera et al., 2007, Zhang and Zhang, 2009). Naproti tomu neuronová síť *Caenorhabditis elegans* s modularitou $Q_{max} \approx 0.4$ vykazuje 80% nárust vzhledem ke střední hodnotě. Sociální sítě jsou charakteristické spíše nižší hodnotou $Q \approx 0.5$ a velkým rozdílem vzhledem ke střední hodnotě v randomizovaném souboru ($\approx 100\%$) (Zhang and Zhang, 2009). Tyto výsledky jsou zhruba konzistentní s nově zavedenými indexy "robustnosti modulární struktury", které ukazují, že sociální sítě mají nejrobustněji vyjádřenou modularitu, metabolické a protein-interakční sítě vykazují jen slabě vyjádřenou modulární strukturu (Hu et al., 2010b, Hu et al., 2010a).

### *Univerzalita sítí*

Většina poznatků o komplexních sítích, nashromážděných za poslední dvacetiletí poukazuje k univerzální struktuře charakteristické mocninným rozdělením konektivit, "logaritmickou průměrnou vzdáleností mezi vrcholy" a modularitou.

Princip prefferential attachment je rozpoznatelný ve většině studovaných sítí. V případě sítě www je všeobecně známo, že čím větší počet www stránek na danou stránku odkazuje, tím výše bude tato stránka hodnocena vyhledávacími stroji. Čím lepší pozice ve výsledcích vyhledávání, tím vyšší počet návštěvníků a tedy i pravděpodobnost toho, že nově zakládané www stránky budou na tuto stránku odkazovat. V případě sociálních sítí má člověk s vysokým počtem přátel vyšší pravděpodobnost, že získá nového přítele protože se díky množství svých přátel dostává do kontaktu s novými lidmi častěji než kdyby měl přátel málo.

V případě protein-interakčních sítí má protein s vysokou konektivitou větší šanci získat nového interakčního partnera než protein s nízkou konektivitou protože má více sousedů, jejichž geny jsou k dispozici náhodné duplikační události a následné diversifikaci při zachování interakce s původním proteinem. Zdá se tedy, že prefferential attachment je obecnou zákonitostí charakteristickou pro komplexní systémy.

Logaritmická závislost průměrné vzdálenosti v síti na počtu vrcholů sítě je vlastností většiny nulových modelů, je ji tedy možno považovat za vedlejší produkt růstových mechanizmů sítě.

Modularita je v případě sociálních sítí a internetu důsledkem obecné lidské tendence sdružovat se do skupin na základě sdílení společného zájmu respektive problému. V případě biologických sítí se často na modularitu nahlíží jako na vlastnost, která zvyšuje evolvabilitu organismu. Moduly jsou v této hypotéze považovány za systémy, vykonávající nějakou elementární funkci, která je potřebná ve všech kontextech, evoluční adaptace vzniká změnou propojení těchto elementárních modulů. Evoluce tedy podobně jako programátor nemusí pokaždé znovu vymýšlet elementární operace, spíše zadaný problém vyřeší tak, že zkombinuje již hotové základní moduly. Největší experimentální podporu má tato představa v oblasti výzkumu ontogeneze. Grafová reprezentace regulačních sítí uplatňujících se v ontogenezi je však studována minimálně nebo vůbec. V případě metabolických a protein-interakčních sítí je tato hypotéza přinejmenším problematická. Její problematičnost spočívá v pochybnosti, jestli moduly nalezené v těchto sítích odpovídají evolučním modulům, konzistentním s touto hypotézou (Zhao et al., 2007, Wagner, 2009).

Nabízí se otázka, zda-li existují mezi sítěmi z různých oblastí kvalitativní rozdíly, nebo je jejich struktura zcela univerzální. V rámci omezení jako je mocninná distribuce konektivit a modularita zbývá ještě dosti prostoru pro variabilitu. Jednou z prvních charakteristik, která diversifikovala reálné sítě je korelace konektivit sousedních vrcholů. Bylo celkem jednoznačně prokázáno, že síť internetu vykazuje negativní korelace v konektivitě sousedních vrcholů jak na úrovni www stránek tak na úrovni autonomních agentů - tento typ sítí je znám jako disassortativní (Hao and Li, 2011, Nelly Litvak and Hofstad, 2012). Naopak pozitivní korelace mezi konektivitou sousedních vrcholů je typická pro sítě spoluautorství ve vědeckých publikacích (Hao and Li, 2011). Detailnější vhled do lokální struktury poskytuje klasifikace uzlů do tzv. rolí z hlediska konektivity v rámci modulární struktury. Tato klasifikace ukázala, že v rámci množiny uzlů se stejnou konektivitou existuje korelace mezi mírou v jaké propojují různé moduly, jejich evoluční konzervací a průměrnou konektivitou

jejich sousedů (Guimera and Amaral, 2005, Guimera et al., 2007). Tato zjištění dále umožnila rozdělit známé sítě na transportní a signalizační na základě korelace v profilu abundancí spojení mezi jednotlivými rolemi (Guimera et al., 2007).

Výsledků, které poukazují ke specificitě jednotlivých sítí je stále méně než těch, které naznačují universalitu. Obecnost a specifičnost však nemusejí být v rozporu. Komplexita reálných sítí umožňˇuje aby některé jejich charakteristiky byly formovány jedinou zákonitostí a zároveň jiné charakteristiky odrážely funkční a specifika jednotlivých sítí. Síťová věda je velmi mladá a sítě velmi složité, je tak možno očekávat spoustu dalších poznatků.

## Metabolické sítě

První studie topologie metabolických sítí se objevily na počátku 21. století. Prokázaly, že metabolické sítě mají mocninné rozdělení konektivit metabolitů (Jeong et al., 2000, Wagner and Fell, 2001). Konektivitu metabolitu můžeme chápat jako počet reakcí, do kterých tento metabolit vstupuje (je jejich substrátem) nebo jako počet reakcí, které jej produkují. V obou těchto případech rozdělení četností konektivit odpovídá klesající mocninné funkci s exponentem $\gamma \approx 2.2$ (Jeong et al., 2000). Pozorované mocninné rozdělení je dáno velkým počtem metabolitů s malou konektivitou a relativně malým počtem metabolitů s vysokou konektivitou, tzv. "currency metabolites" (Tanaka, 2005). Currency metabolites jsou donory energie, redukčních ekvivalentů a jiných funkčních skupin pro velký počet reakcí.

Analogickým způsobem jako ve studii (Eisenberg and Levanon, 2003) byl detekován preferential attachment v metabolické síti (Light et al., 2005). Autoři reprezentovali metabolickou síť baktérie *E. coli* jako orientovaný graf ve kterém vystupují pouze enzymy. Enzym 1 je spojen hranou s enzymem 2, když enzym 1 produkuje metabolit, který je substrátem enzymu 2. Metabolické enzymy byly rozděleny do skupin podle odhadovaného evolučního stáří. Průměrná konektivita enzymů v těchto skupinách roste s jejich evolučním stářím což je v souladu s principem preferential attachment. Dále byla extrahována metabolická pod-síť *E.coli* složená pouze s enzymů, které jsou sdíleny všemi fylogenetickými doménami. Tato pod-síť je odhadem toho jak mohl vypadat metabolismus společného předka. Konektivita každého enzymu této ancestrální podsítě byla rozdělena na počet sousedů, keří jsou obsaženi v ancestrální pod-síti a na počet sousedů, kteří jsou mimo tuto podsíť a jsou tedy evolučně mladší. Bylo zjištěno, že tyto dvě konektivity spolu lineárně korelují. Evolučně staré enzymy, které pravděpodobně měly vysokou konektivitu již ve společném předkovi nakumulovaly více nových sousedů v průběhu evoluce než stejně staré enzymy, které měly

nízkou konektivitu ve společném předkovi (Light et al., 2005). Tyto výsledky naznačují, že i v případě metabolických sítí se uplatňuje lineární forma preferential attachment.

V prvních studiích metabolických sítí je uváděna průměrná délka cesty blízká hodnotě 3,2, navíc je tato hodnota konstantní pro 43 organizmů reprezentující různé zástupce bakterií, archeí a eukaryot (Jeong et al., 2000). Průměrná délka cesty na základě této studie nezávisí na velikosti sítě. Autoři prokázali, že se zvětšující se velikostí sítě roste i průměrná konektivita, větší metabolické sítě jsou tedy hustěji propojeny, což umožňuje udržet průměrnou vzdálenost konstantní.

Definice cesty mezi dvěma metabolity ve studii (Jeong et al., 2000) neodpovídá v řadě případů nejmenšímu počtu chemických reakcí nutných k transformaci jednoho metabolitu na druhý. Mějme reakci, ve které metabolit B vzniká z metabolitu A a donorem energie pro tuto reakci je ATP. V reprezentaci Jeonga et al.(2000) je metabolit A i ATP spojen hranou s produktem B. Minimální vzdálenost od ATP k metabolitu B je tedy 1. Nemůžeme však říci, že stačí jedna reakce k transformaci ATP na metabolit B. Právě interpretace vzdálenosti v metabolické síti jako nejmenší počet reakcí nutných k vzájemné transformaci dvou metabolitů je biologicky relevantním ukazatelem flexibility metabolismu.

V práci (Arita, 2004) byla metabolická cesta vedoucí od A do B definována jako sekvence reakcí skrze které se alespoň jeden uhlíkový atom z metabolitu A dostane do metabolitu B. Bylo zjištěno, že průměrná nejkratší délka takto definované cesty je 8,4 pro metabolickou síť *E. coli*. Jinou strategii použili autoři Ma a Zeng (2003). Pro každou reakci určili, které její substráty figurují jako "currency metabolites", tedy donory energie nebo funkčních skupin pro tuto reakci. Dále odstranili hrany, které spojují tyto currency metabolites s produkty reakce, kterým pouze dodávají funkční skupinu nebo energii. Autoři takto upravili metabolické sítě 80 zástupců pocházejících ze všech tří domén života. Zjistili, že průměrná délka cesty pro bakterie, archea a eukaryota je 7,22, 8,5 a 9,57. Průměrná délka cesty v této studii není konstantní, závisí na velikosti sítě, což je v přímém rozporu se studií (Jeong et al., 2000). Vzhledem k tomu, že je ve studii (Ma and Zeng, 2003) použita biologicky relevantnější definice metabolické cesty, je možno i její výsledky považovat za směrodatnější.

Průměrná vzdálenost v metabolické síti je vyšší než střední hodnota v randomizovaném souboru, který zachovává sekvenci konektivit výchozího grafu (Zhang and Zhang, 2009, Basler et al., 2012). Existují však dobré důvody, které zpochybňují relevanci standardní randomizační procedury, která zachovává konektivitu vrcholů. Náhodné grafy, které tato procedura produkuje mohou obsahovat reakce, které nesplňují zákon zachování hmoty.

Nedávno byla publikována randomizační metoda, která generuje pouze reakce, pro které platí, že suma atomů každého prvku v množině substrátů je rovna sumě atomů stejného prvku v množině produktů reakce (Basler et al., 2011). Průměrná vzdálenost v metabolické síti je mírně nižší než její střední hodnota v randomizovaném souboru, který respektuje zákon zachování hmoty (Basler et al., 2012). Autoři zřejmě měřili vzdálenosti v kompletních sítích (obsahujících currency metabolites). Není tedy jasné jaký výsledek bychom dostali po odstranění currency metabolites. Tato metoda zachovává konektivitu reakcí, nikoliv konektivitu metabolitů a souvislost grafu, její ambice je generovat náhodné sítě, které neodporují fyzikálním zákonům a poskytnout tak minimální plausibilní nulový model. Je zřejmé, že spousta náhodných grafů produkovaných tímto modelem bude reprezentovat zcela nesmyslný metabolismus neslučující se s přežitím buňky. Je tedy možno zajít ještě dál a konstruovat náhodné metabolické sítě, které umožňují produkci všech esenciálních složek buněčné biomasy. Autoři studie (Samal and Martin, 2011) vygenerovali několik randomizovaných souborů, které se lišily v počtu různých složení růstového média ve kterém byly sítě z těchto souborů schopny produkovat prekurzory biomasy. Sítě byly generovány kombinováním existujících reakcí obsažených v KEGG databázi, čímž je garantováno splnění zákona zachování hmoty. Ukázalo se, že průměrná vzdálenost v síti je v rámci těchto souborů takřka konstantní a jen velmi zanedbatelně menší než ve skutečné metabolické síti *E. coli*. Současný stav poznání tedy naznačuje, že vzdálenost v metabolických sítích není významně redukována z hlediska nulových modelů.

První studie modulární struktury metabolických sítí používaly hierarchický klastrovací algoritmus k detekci modulů, který je založen na definici lokální denzity hran (Ravasz et al., 2002, Ma et al., 2004). V těchto studiích je uváděno, že identifikované moduly odpovídají metabolickým funkčním kategoriím, známým z učebnic biochemie a definovaných např. v databázi KEGG. Tento závěr má však spíše intuitivní charakter, korespondence mezi moduly a biochemickými kategoriemi nebyla v těchto studiích kvantifikována a statisticky vyhodnocena. V práci Ravasz et. al. (2002) bylo zjištěno, že hodnota klastrovacího koeficientu, který měří hustotu hran mezi sousedy daného uzlu, inverzně závisí na konektivitě toho uzlu, přesněji $C(k) \sim k^{-1}$. Autoři ve stejné práci postulují deterministický model růstu sítě, který kombinuje malé moduly s maximální denzitou hran do větších méně denzních modulů. Takto generované sítě mají jasnou hierarchickou strukturu a platí pro ně $C(k) \sim k^{-1}$. Autoři Ravasz et. al. (2002) navrhují inverzní závislost klastrovacího koeficientu na konektivitě jako indikátor hierarchické modulární struktury sítě. Nedávno však bylo zjištěno,

že velmi podobná závislost je produkována i randomizovanými sítěmi zachovávajícími sekvenci konektivit, ve kterých by hierarchická struktura neměla být přítomna. Jediný rozdíl oproti náhodnému modelu je v tom, že reálné metabolické sítě mají mírně vyšší hodnotu klastrovacího koeficientu pro nízké hodnoty konektivity. V případě metabolických sítí je inverzní závislost klastrovacího koeficientu na konektivitě dána přítomností currency metabolites, které jsou vzájemně signifikantně málo propojeny, (vykazují disassortativitu). Po odstranění těchto metabolitů získáme assortativní síť, kde má klastrovací koeficient konstantní hodnotu v závislosti na konektivitě (Hao et al., 2012). Závislost klastrovacího koeficientu na konektivitě tedy není zárukou hierarchické struktury sítě jak bylo původně předpokládáno.

Jak již bylo řečeno výše, v pracích používajících modularity measure ($Q$) bylo prokázáno, že hodnota $Q$ je signifikantně vyšší v reálných metabolických sítích než střední hodnota $Q$ v randomizovaných souborech zachovávajících konektivitu. Rozdíl oproti střední hodnotě se pohybuje okolo 15% (Zhao et al., 2006, Guimera et al., 2007, Zhang and Zhang, 2009). Metabolické sítě vykazují spíše slaběji vyjádřenou modulární strukturu ve srovnání s se sociálními nebo neurálními sítěmi (Hu et al., 2010b, Hu et al., 2010a). Moduly identifikované pomocí maximalizace $Q$ jsou typicky složeny z více biochemických KEGG kategorií a jednotlivé KEGG kategorie jsou typicky distribuovány mezi více modulů. V některých modulech sice jednoznačně převažuje jediná KEGG kategorie, z globálního pohledu je vzájemná korespondence mezi moduly a KEGG kategoriemi slabá (Guimera and Amaral, 2005, Zhao et al., 2006).

Kashtan a Alon (2005) simulovali evoluci booleovských sítí. Sítě byly selektovány tak aby vykonávaly předem definovanou logickou funkci vstupů. Fitness sítě byla definována jako frakce všech kombinatoricky možných vstupů, pro kterou má tato logická funkce správnou hodnotu na výstupu. Bylo zjištěno, že když se periodicky mění logické funkce pro které jsou booleovské sítě selektovány a tyto logické funkce jsou vždy sestaveny s určitého počtu fixních subfunkcí s fixními proměnnými, výsledkem selekčního procesu jsou sítě s vysokou modularitou ($Q$). Tato studie jasně ukazuje, že modularita může být důsledkem proměnlivosti selekčních tlaků na komplexní funkce, při konstantnosti elementárních funkcí, ze kterých je vždy možno tyto komplexní funkce složit.

Parter et. al. (2007) empiricky ověřoval predikce předchozí studie na 117 bakteriálních druzích. Rozdělil tyto druhy do 6 kategorií podle variability prostředí ve kterém jsou schopny žít. Průměrná modularita metabolických sítí v jednotlivých souborech roste s variabilitou

životního prostředí baktérií v těchto souborech což je v souladu se studií (Kashtan and Alon, 2005). Podobné výsledky byly získány ve studii (Kreimer et al., 2008).

Samal et. al. (2011) definoval moduly jako skupiny reakcí, jejichž reakční rychlosti jsou ve fixním vzájemném poměru za všech okolností. Pomocí stejné metody jako ve studii (Samal and Martin, 2011) generoval náhodné metabolické sítě s fixním počtem reakcí, které jsou schopny produkovat složky biomasy s využitím různých zdrojů uhlíku. V této studii bylo zjištěno, že čím více zdrojů uhlíku je schopna metabolická síť využít k syntéze biomasy tím větší část sítě je tvořena moduly s fixním poměrem reakčních rychlostí. S flexibilitou metabolické sítě rostl i počet takto definovaných modulů. Autoři potvrdili, že flexibilita z hlediska využívaní zdrojů vede ke vzniku nových metabolických cest, které zapojují různé vstupní metabolity do sítě a tedy ke zvýšení modularity. Tyto cesty díky linearitě splňují definici modulu použitou v této studii. Implicitní poselství této studie říká že konstantní selekce na schopnost využívat více zdrojů uhlíku vede k modularitě metabolické sítě.

Kashtan a Alon (2005) uvádějí, že moduly vyselektované v jedné periodě zůstávají z převážné většiny zachovány při vstupu do další periody s novou selektovanou funkcí. Zároveň potvrzují, že pokud budeme selektovat síť na jedinou konstantní funkci, modularita zanikne. Je tedy zřejmé, že existuje nějaká kritická délka periody ve které nestačí modulární struktura zaniknout a systém zůstává evolvabilní. Připustíme-li, že modularita definovaná v (Samal et al., 2011) koreluje s veličinou $Q$, pak se může tvrzení, že modularita zaniká při konstantním selekčním tlaku (Kashtan and Alon, 2005) jevit v rozporu se studií (Samal et al., 2011). Evoluce metabolických sítí je však natolik specifická, že zřejmě nelze očekávat, že všechny závěry získané s booleovskými sítěmi budou platné i pro metabolické sítě. Vyjasnění situace by poskytla studie, která by potvrdila 1) že modularita definovaná v (Samal et al., 2011) koreluje s veličinou $Q$ (a případně, že selekce na schopnost růstu v mnoha podmínkách vede ke zvýšení $Q$) 2) provedla simulaci evolučního procesu s periodicky se měnícími selekčními tlaky na schopnost růstu v různých podmínkách v rámci stejného genotypového prostoru jako ve studii (Samal et al., 2011).

V reálných podmínkách lze očekávat jak konstantní selekční tlak na schopnost růstu v různých podmínkách tak i periodicky se měnící selekční tlaky. Ve studii (Parter et al., 2007) bylo potvrzeno, že s růstem variability životního prostředí roste i velikost metabolické sítě. Adaptace na více růstových podmínek tedy nejspíše probíhá vznikem nebo získáním nových metabolických cest napojujících specifické nutrienty z externího prostředí na metabolickou síť, jak se ukazuje v (Samal et al., 2011) .

Nedávná studie potvrdila, že pozorovaná modularita metabolických sítí je produkována velmi jednoduchým modelem růstu sítě. V tomto modelu se v každém kroku s pravděpodobností $p$ přidá jeden nový metabolit a spojí se hranou s náhodně vybraným uzlem. S pravděpodobností $1-p$ je náhodně vybraný uzel $u$ spojen hranou s jiným náhodně vybraným uzlem $v$. Pravděpodobnost výběru uzlu $v$ exponenciálně klesá s jeho vzdáleností vzdálenosti od $u$. Sítě generovaté tímto algoritmem vykazují stejnou hodnotu $Q$ jako reálné metabolické sítě srovnatelné velikosti. Tento algoritmus ve velice zjednodušené podobě odpovídá představě růstu metabolické sítě a ukazuje, že pozorovanou modularita je vysvětlitelná velmi jednoduchým principem (Samal et al., 2011).

## Regulační sítě metabolismu

Existují dvě základní úrovně regulací metabolismu. První úrovní je transkripční regulace. Transkripční regulace je realizována aktivací transkripčního faktoru buď interním signálem jako např. zvýšenou koncentrací intracelulárního metabolitu nebo externím signálem, který je přenesen z povrchového receptoru buňky do jejího nitra. Aktivovaný transkripční faktor dále aktivuje expresi genu nebo celého operonu, který kóduje enzym nebo skupinu enzymů. Zvýšená exprese enzymu má za následek zvýšení koncentrace tohoto enzymu v cytoplazmě nebo jiném kompartmentu a to v důsledku vede ke zrychlení reakce katalyzované tímto enzymem. Transkripční regulace metabolismu je považována za hlavní mechanismus který zajišťuje fyziologickou adaptaci buněčného metabolismu. Reálné metabolické sítě obsahují stovky regulačních vazeb tohoto typu. Můžeme tedy mluvit o transkripčně regulační síti metabolismu. Tato síť je v nejjednodušší podobě reprezentována bipartitním grafem, který má dva typy vrcholů: transkripční faktory a enzymy, jejichž expresi ovlivňují. Hranu pokládáme mezi transkripční faktor *tf* a enzym *e*, když je experimentálně potvrzeno, že *tf* reguluje expresi enzymu *e*.

Druhou úrovní jsou přímé regulace enzymové aktivity metabolitem. Tento typ regulací je nejčastěji realizován vazbou metabolitu na enzym mimo jeho aktivní místo, tato vazba způsobí změnu konformace enzymu a v důsledku toho změnu reakční rychlosti. Tento typ regulací je znám jako tzv. alosterická regulace. Regulační síť tohoto typu reprezentujeme bipartitním grafem, který má dva typy vrcholů: metabolity a enzymy. Hrana je položena mezi metabolit *m* a enzym *e*, když je experimentálně potvrzeno, že metabolit *m* reguluje aktivitu enzymu *e* přímou interakcí *m* s *e*.

Je zřejmé, že lze volit alternativní reprezentace, kde např. místo enzymů vystupují reakce. Transkripčně regulační síť je možno dále rozšiřovat o interakce mezi transkripčními faktory a metabolity, které tyto transkripční faktory regulují.

Regulační sítě samy o sobě, tak jak byly definovány v předchozích odstavcích poskytují pouze omezenou informaci o regulaci metabolismu. Chceme-li pochopit komplikovanou mašinérii řízení metabolismu musíme regulační sítě studovat v kontextu metabolické sítě ve kterém evolvovaly do dnešní podoby.

### *Transkripčně regulační síť metabolismu*

Bylo opakovaně prokázáno, že transkripčně regulační sítě mají hierarchickou strukturu (Ihmels et al., 2004, Seshasayee et al., 2009, Samal and Jain, 2008). Ve studii (Seshasayee et al., 2009) bylo ukázáno, že v transkripčně regulační síti metabolismu *E.coli* existuje malý počet (~10) globálních transkripčních faktorů (dále jen TFs), které regulují velký počet metabolických genů v různých funkčních kategoriích nejčastěji geny energetického metabolismu. Pro každý globální TF je však možno určit funkční kategorií, kterou reguluje dominantně. Uveďme příklad CRP, který je alostericky aktivován cAMP v nepřítomnosti glukózy v růstovém médiu a dále aktivuje expresi celé řady metabolických genů (~200), které odpovídají metabolickým cestám utilizace alternativních zdrojů uhlíku. Dominantně reguluje metabolismus cukrů, ve smyslu přechodu z glykolýzy na glukoneogenesi a respirační metabolismus.

Zbylá část metabolických TFs má výrazně nižší konektivitu než globální transkripční TFs a regulují metabolické geny spadající do jediné funkční kategorie nebo do jediné metabolické cesty, v práci (Seshasayee et al., 2009) jsou nazývány specifické TFs. Aktivita specifických TFs je často regulována metabolitem, který je produktem enzymů kontrolovaných těmito TFs. Pro specifické TFs je tedy charakteristická lokální regulace zpětnou vazbou.

Seshasayee et. al. analyzoval koregulaci enzymových párů *E.coli* a zjistil, že největší míru koregulace vykazují dvojice enzymů v lineárních metabolických cestách, bez odboček jak z hlediska sdílení stejných TFs tak z hlediska změřené korelace v expresním profilu. Když se tato cesta připojuje k jiné cestě, expresní korelace mezi dvojicemi enzymů takovými, že jeden enzym z této dvojice je před připojením a druhý po připojení druhé cesty, se zeslabuje. V situaci, kdy se jedna cesta rozděluje na dvě, je expresní korelace mezi dvojicemi enzymů (jeden z dvojice před rozdělením a druhý po rozdělení cesty) nejslabší. Autoři tento fakt vysvětlují tím, že divergence metabolické cesty představuje kontrolní bod, ve kterém je

možno podle aktuální situace nasměrovat metabolický tok jedním nebo druhým směrem. Proto je možno očekávat rozdílnou regulaci divergujících cest.

V práci (Ihmels et al., 2004) autoři analyzovali metabolickou transkripčně regulační síť *S. cerevisiae*. Zjistili, že expresní profil enzymu před rozdělením metabolické cesty na dvě, často signifikantně koreluje pouze s jedním s enzymů následujících dvou reakcí za rozdělením. Dále zjistili, že v situacích kdy existuje více izoenzymů katalyzujících reakci před rozdělením metabolické cesty, existuje pro každou ze dvou reakcí za rozdělením specifický izoenzym (katalyzující reakci před rozdělením), který je s ní signifikantně koregulován. Tyto výsledky poukazují k tomu, že transkripční regulace usměrňuje metabolický tok do lineárních cest za účelem zefektivnění metabolismu ve specifických podmínkách.

V práci (Seshasayee et al., 2009) je uváděn pouze medián expresní korelace enzymových párů. V případě divergentních rozboček jsou mezi sebou korelovány všechny dvojice, které splňují podmínku, že jeden enzym z této dvojice je před rozdělením a jeden za rozdělením. Některé z těchto dvojic mohou být korelované - ty odpovídají dvojicím identifikovaným v práci (Ihmels et al., 2004), jiné korelované nejsou. Zaměříme-li se na medián všech dvojic, tato informace zůstane zakryta. V práci (Ihmels et al., 2004) autoři detekovali počet divergentních rozboček ve kterých je signifikantně koexprimovaný pouze jeden enzymový pár, tvořící lineární cestu a porovnali tento počet se stejnou kvantitou v případě randomizace identit vrcholů metabolické sítě. Výsledky těchto dvou studií nejsou v rozporu, Seshasayee et al.(2009) uvádí, že 52% rozboček obsahuje alespoň jeden enzymový pár, který sdílí stejné TFs, což umožňuje aby se fenomén detekovaný u *S. cerevisiae* (Ihmels et al., 2004) mohl uplatňovat i u *E.coli*.

Je možno studovat i dynamické chování transkripčně regulačních sítí pomocí booleovské aproximace. V této aproximaci každý gen nabývá pouze dvou stavů 1/0, simulace probíhá v diskrétních krocích, přechod ze stavu v čase *t* do stavu v čase *t* +1 určuje přechodová funkce, která všem genům přiřadí novou hodnotu na základě hodnot jejich sousedů v čase *t*. V práci (Barrett et al., 2005) byla simulována transkripčně regulační síť metabolismu *E.coli*. Pro simulace byla vybrána jen taková složení růstového média, která umožňují dvojení buňky za $\leq$12 h, jejich celkový počet je 15580. Vývoj stavů regulační sítě byl reprezentován vektorem logických hodnot všech interakcí pro všechny časové okamžiky simulace. Bylo ukázáno, že takto definované vektory - aktivitní profily se koncentrují ve třech jasně separovaných klastrech. Každému klastru odpovídá specifický terminální elektronový akceptor nebo jejich skupina ($O_2$ -klastr1, Fumarát/DMSO/TMAO -klastr2, $NO_3$ /$NO_2$-klast3). Tyto klastry lze

dále dělit na sub-klastry podle toho je-li zdrojem uhlíku glukóza nebo glukonát. Tato studie ukazuje, že velmi komplexní regulační systém metabolismu baktérie *E.coli* se chová nečekaně jednoduše z hlediska stavů které nabývá a že elektronové akceptory jsou hlavní determinující faktor těchto stavů.

V práci (Samal and Jain, 2008) byla studována transkripčně regulační síť metabolismu *E.coli* z hlediska robustnosti atraktorů booleovského modelu a jeho flexibility vzhledem ke změně vnějších podmínek. Bylo ukázáno, že nezávisle na počáteční konfiguraci genů tato síť konverguje $v \leq 4$ krocích k fixnímu atraktoru. Tento atraktor je determinován přítomností/nepřítomností metabolitů regulujících aktivitu příslušných TFs. Tyto metabolity poskytují informaci o složení růstového média. Bylo prokázáno, že regulační nastavení exprese genů odpovídající atraktoru, ke kterému síť konvergovala z nějakého výchozího stavu zvyšuje růstovou rychlost (produkci biomasy) typicky až na 80-90% maxima.

Tyto vlastnosti jsou triviálně determinovány tím, že zkoumaná síť je acyklická, (tedy strom) s maximální hloubkou 4, kde nejvýše v hierarchii jsou regulační metabolity a nejníže enzymy. Atraktory, jsou fixní body - stav genů je v atraktorech neměnný. Z hlediska Kaufmanových náhodných sítí se tato nachází v zamrzlém regionu, přesto velice flexibilně reaguje na změnu růstového prostředí a je robustní vzhledem k počáteční konfiguraci genů.

### *Regulační síť metabolismu reprezentující přímé interakce metabolitů s enzymy*

Tento typ regulační sítě je hlavním předmětem mého výzkumu. Kromě publikací, které jsou výstupem této disertační práce existuje pouze jediná publikace, která se zabývá topologií sítí reprezentujících přímé regulační interakce metabolitů s enzymy (Gutteridge et al., 2007). Většina prací týkajících se tohoto typu regulací se zaměřuje na detailní dynamické modelování regulačních motivů, jako je např. lineární metabolická cesta se zpětnou vazbou nebo na méně detailní dynamické modelování větších metabolických celků.

Autoři práce (Gutteridge et al., 2007) studovali topologii sítě reprezentující přímé regulační interakce mezi enzymy a metabolity u čtyř organismů (*E. coli*, *S. cerevisiae*, *P. falciparum*, *H. sapiens*). Bylo zjištěno, že distribuce počtu regulovaných enzymů jednotlivými metabolity odpovídá mocninné distribuční funkci pro všechny studované organismy ($1.64<\gamma<1.82$). Byla také pozorována inverzní závislost klastrovacího koeficientu na regulační konektivitě metabolitů. Toto zjištění však není příliš informativní ve světle studie (Hao et al., 2012). Autoři dále nalezli slabou korelaci mezi regulační a reakční konektivitou metabolitů ($\rho_{xy}$= 0.42, 0.53). Tato korelace je z převážné části způsobena currency metabolites, které vykazují

vysokou hodnotu obou typů konektivit. Byla nalezena slabá korelace mezi chemickou a regulační podobností v rámci všech dvojic regulačních metabolitů. Tento poznatek má spíše charakter omezení, které říká, že metabolity, které regulují převážně stejné enzymy jsou zároveň chemicky příbuzné. V rámci dvojic metabolitů, které regulují převážně nestejné enzymy je jejich chemická příbuznost rovnoměrně distribuována v celé škále.

## Cíle práce a shrnutí výsledků

*Hlavním cílem práce je charakterizovat strukturu sítě reprezentující přímé regulační interakce mezi enzymy a metabolity a odpovědět na otázku, zdali je tato struktura nahlížená z hlediska teorie grafů náhodná nebo vykazuje vlastnosti, které jsou odrazem evoluční optimalizace.*

V případě metabolických sítí je zřejmé, že evoluční trajektorie nemůže probíhat celým prostorem všech kombinatoricky možných grafů s danou sekvencí konektivit. Prvním omezením je zákon zachování hmoty. Máme-li fixní množinu metabolitů, reakce může vzniknout jen mezi skupinami substrátů a produktů, které mají stejný počet atomů od každého prvku. Metabolity, které mohou existovat, jsou dále omezeny kritériem stability. Minimální nulový model zkoumané sítě by měl produkovat náhodné sítě, které vyhovují elementárním fyzikálním omezením. Jestliže je zkoumaná vlastnost signifikantní z hlediska takto definovaného nulového modelu, je pravděpodobné, že je produktem adaptivní evoluce organismu.

Hlavním argumentem této práce je hypotéza, že regulační síť metabolit-enzymových interakcí vykazuje oproti metabolické síti mnohem vyšší evoluční flexibilitu, tj. že její topologie není omezena fyzikálními zákony v takové míře jako metabolismus. Proto je možné očekávat, že její struktura bude evolučně optimalizovaná. Tento argument je založen na komplexitě a z ní vyplývající flexibilitě proteinových molekul. Je známo, že vhodnou kombinací aminokyselin, lze na povrchu proteinu vytvořit vazebné místo pro libovolný molekulární tvar. Typickým příkladem jsou protilátky. Dalším krokem je spojení vazby metabolitu na enzym se změnou konformace enzymu a tím i rychlosti katalýzy.

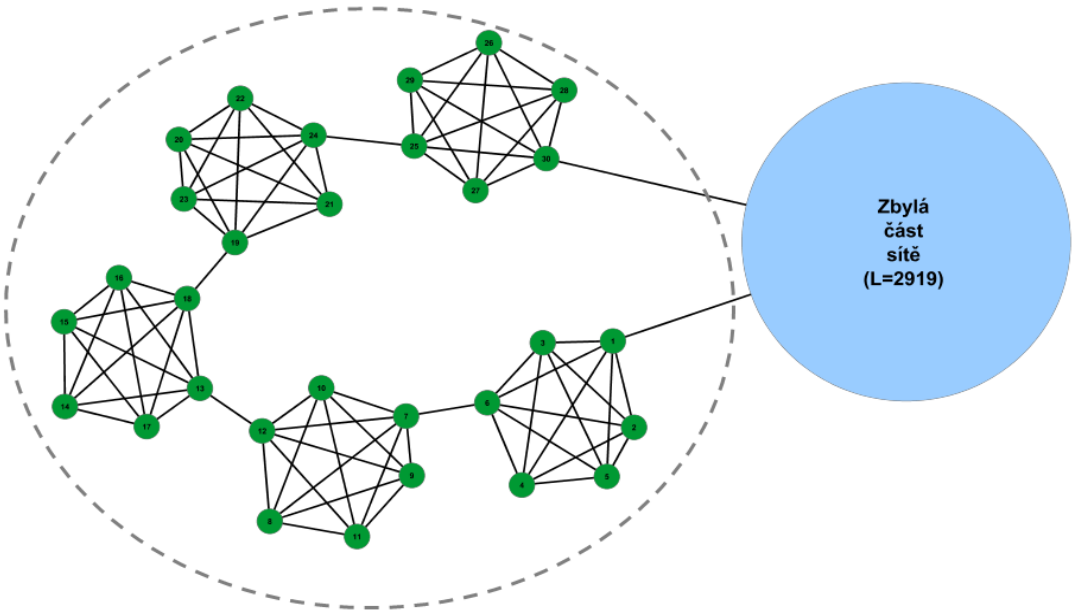
Možnost takto propojit libovolné dva procesy nezávisle na jejich chemické podstatě je známa jako gratuita a je základem evolvability všech organizmů.

***První práce*** je technického charakteru a zabývá se problémem rozlišovacího limitu modularity measure. Jak již bylo uvedeno, autoři Fortunato a Bathelemy (2007) ukázali, že

maximalizace $Q$ může vést k indentifikaci modulů, které jsou složeny z jednoznačně vyjádřených sub-modulů, které je přirozené považovat za samostatné moduly. Autoři ve své práci uvažovali pouze speciální případ dvou modulů, na základě kterého odvodili kritickou velikost modulu ($2l_r = \sqrt{2L} - 2$), pod kterou existuje riziko, že je tento modul složen ze dvou jednoznačně vyjádřených sub-modulů - v extrémním případě mohou být tyto sub-moduly kliky.

V první práci s názvem "Hledání (správné) míry modularity" jsem na základě obecné formulace modularity measure odvodil obecný rozlišovací limit. Tento obecný rozlišovací limit říká za jakých podmínek je hodnota $Q$ vyšší pro rozdělení libovolného podgrafu na jediný modul než pro rozdělení tohoto podgrafu na libovolný počet menších modulů. Obecný rozlišovací limit požaduje aby střední hodnota počtu hran mezi moduly byla menší než počet hran v těchto modulech. Je-li tato podmínka splněna, bude mít rozdělení podgrafu na jediný modul vyšší hodnotu $Q$. V náhodných modelech závisí střední počet hran v podgrafu na celkovém počtu hran sítě. Při konstantní lokální struktuře skupiny modulů tak můžeme přidáváním hran do zbytku sítě měnit možnost identifikace těchto modulů.

Uvažujme pět úplných podgrafů, propojených tak aby se zbytkem grafu tvořily kružnici, kde celkový počet hran sítě je $L$ = 3000 Obr.1. Na základě obecného rozlišovacího limitu je možno pro tuto situaci odvodit, že $l_r = \sqrt{L/m} - 1$, kde $L$ je počet hran v síti a $m$ je počet modulů v uvažovaném podgrafu.




Obr. 1. Přerušovaná šedá čára vymezuje sloučení pěti modulů na obrázku do jednoho.

Sloučíme-li všech pět podgrafů na obr.1. v jediný modul bude jeho modularita $Q = 0{,}0263$, zatímco součet modularit těchto pěti modulů bude $Q = 0{,}0249$. Rozdělení, ve kterém budou tyto moduly sloučeny v jeden bude při maximalizaci $Q$ preferováno, přesto že jsou tyto moduly jednoznačně separovány od zbytku sítě.

***Druhá práce*** (Geryk and Slanina) je zaměřená na otázku, jak se změní modulární struktura metabolické sítě *E.coli*, obohatíme-li tuto síť o regulační interakce mezi metabolity a enzymy. V dalším textu budu síť obohacenou o regulační interakce nazývat kombinovaná síť. Motivací pro tuto otázku byla hypotéza modulu jako funkčně autonomní jednotky. Typická vlastnost autonomního systému je schopnost samořízení a/nebo snížená možnost kontroly tohoto systému zvenčí. Přidání regulačních vazeb může významně změnit poměry hustot hran v síti a odkrýt tak novou modulární strukturu. Moduly detekované v kombinované síti mohou obsahovat zpětné regulační vazby, které zajišťují jejich autonomii z hlediska řízení.

Pro detekci modulů v metabolické síti byly použity dva algoritmy. První z nich je hierarchický klastrovací algoritmus. Je založen na lokální podobnosti dvojic vrcholů definované jakožto hustota hran mezi sousedy této dvojice. Na začátku procedury je množina samotných vrcholů bez hran. Každý vrchol sítě v této počáteční fázi představuje samostatnou komponentu. Algoritmus v každém kroku nalezne dvojici vrcholů, kde každý z nich je v jiné komponentě, která má nejvyšší hodnotu podobnosti a spojí je hranou. Tato operace se opakuje až do okamžiku, kdy jsou všechny vrcholy součástí jedné komponenty. Komponenty v daném kroku algoritmu určují rozdělení sítě na moduly. V každém kroku algoritmu je spočítána hodnota modularity measure ($Q$) metabolické sítě rozdělené na moduly, které odpovídají komponentám příslušejícím tomuto kroku. Pro další analýzy je vybráno takové rozdělení na komponenty/moduly, kterému přísluší nejvyšší hodnota $Q$.

Popsaný algoritmus jsem v práci (Geryk and Slanina) modifikoval ve dvou bodech.

1) Byla zavedena nová definice podobnosti vrcholů a cíleně omezena jen na dvojice vrcholů spojených hranou. Tato podobnost je definována pro hranu (*u*,*v*) jako frakce maximálního počtu hran přítomná mezi sousedy vrcholů *u* a *v*, s tím, že z těchto sousedů vylučujeme *u* a *v*.

Tato definice vyjadřuje jak je hustota hran v lokálním podgrafu vymezeném vrcholy *u* ,*v* a jejich sousedy, vzdálena od hustoty hran stromu, což je minimální hustota hran jakou tento podgraf může mít.

2) Druhý bod spočíval v modifikaci modularity measure ($Q$) pro bipartitní graf.

Druhý algoritmus pro detekci modulů použitý v naší práci je standartní optimalizace $Q$ pomocí simulovaného žíhání.

Pomocí metody simulovaného žíhání bylo dosaženo vyšší modularity ($Q = 0{,}66$ v případě samotné metabolické sítě a $Q = 0{,}6$ v případě kombinované sítě) než pomocí klastrovacího algoritmu ($Q = 0{,}31$ v případě samotné metabolické sítě a $Q = 0{,}38$ v případě kombinované sítě). V naší práci bylo ukázáno, že klastrovací metoda i metoda simulovaného žíhání produkují velmi podobný soubor modulů s vysokou denzitou hran, tento soubor tvoří modulární jádro sítě, (jeho modularita odpovídá modularitě dosažené klastrovacím algoritmem). Metoda simulovaného žíhání navíc oproti klastrovací metodě produkuje soubor modulů s velmi nízkou denzitou hran. Tyto moduly s nízkou denzitou mají typicky stromovou strukturu, přesto přispívají k celkové modularitě poměrně vysokou hodnotou tak, že je dosaženo celkové modularity $Q{\sim}0{,}6$ v případě simulovaného žíhání.

V případě klastrovacího algoritmu jsme zaznamenali signifikantní nárust modularity po přidání regulačních interakcí ($p < 0{,}01$), v případě simulovaného žíhání naopak pokles. Extrahujeme-li z rozdělění produkovaného metodou simulovaného žíhání modulární jádro a spočítáme pro něj hodnotu modularity, dostaneme $Q = 0{,}293$ v případě modulárního jádra samotné metabolické sítě a $Q = 0{,}38$ v případě modulárního jádra kombinované sítě. Tyto výsledky ukazují, že vzrůst modularity je lokalizován do modulárního jádra.

Detekované moduly lze rozdělit na dva základní typy. Moduly prvního typu mají vysokou denzitu hran reprezentujících reakce a jsou díky tomu detekovány jak v samotné metabolické síti tak v kombinované síti. Druhý typ modulů má nízkou denzitu reakčních hran a teprve po přidání regulačních hran se jejich denzita zvýší. Tyto moduly jsou detekovatelné pouze v kombinované síti. Právě díky regulačním zpětným vazbám je možno nalezené moduly interpretovat jako autonomní systémy. Jedním z modulů druhého typu je např. metabolická cesta odštěpení glukózy z maltotetraosy a následná vazba nukleotidfosfátů na glukózu. Aktivované formy glukózy zpětně inhibují odštěpování glukózy a vytvářejí tak autonomní regulaci tohoto modulu. Jiným příkladem je glykolytická cesta přeměny fruktózy 6-fosfátu na glyceraldehyd 3-fosfát. Některé metabolity z této cesty inhibují reakce konzumující intermediáty této cesty a některé reakce dodávající intermediáty této cesty. Tento modul je možno interpretovat jako auto-kanalizační systém. Je-li jednou tato cesta aktivní, má tendenci udržovat svou aktivitu.

***Třetí práce*** je inspirována předchozí prací (Geryk and Slanina). Konkrétně regulační strukturou posledního popisovaného modulu, která byla v předchozí práci interpretována jako

auto-kanalizační. Třetí práce se zaměřuje na otázku, zdali jsou v metabolické regulační síti *E.coli* podobné kanalizační regulační motivy abundantní.

V této práci je použita bipartitní reprezentace regulační sítě, ve které vystupují reakce a enzymy. Za účelem zodpovězení výše položené otázky jsem definoval několik veličin, které formalizují pojem kanalizace metabolické cesty. V tomto shrnutí budou popsány jen tři nejdůležitější. Tyto veličiny jsou definovány pro metabolickou cestu ($p$) a analogicky pro minimální hypercestu ($P$).

První z veličin je kanalizace metabolické cesty $p$, značená $C(p)$. $C(p)$ je definováno jako frakce ireversibilních reakcí nepatřících do $p$, které konzumují alespoň jeden metabolit obsažený v $p$ a každá z těchto reakcí, $r_i,\ i = 1,..,n$ splňuje následující podmínku:

1) Existuje alespoň jeden metabolit, který inhibuje $r_i$ a zároveň neinhibuje žádnou z reakcí obsažených v $p$.

Veličina $C(p)$ tedy vyjadřuje do jaké míry má regulační síť potenciální možnost utlumit aktivitu reakcí odbočujících z cesty $p$, a zároveň ponechat reakce v této cestě aktivní.

Další veličina je auto-kanalizace, značená $C_a(p)$. Jediný rozdíl oproti předchozí definici je v podmínce 1, kterou nyní označme jako $1_a$:

$1_a$) Existuje alespoň jeden metabolit produkovaný alespoň jednou reakcí obsažených v $p$, který inhibuje $r_i$ a zároveň neinhibuje žádnou z reakcí obsažených v $p$.

$C_a(p)$ tedy vyjadřuje do jaké míry má samotná cesta $p$ možnost utlumit aktivitu reakcí odbočujících z cesty $p$ a zároveň ponechat reakce v této cestě aktivní.

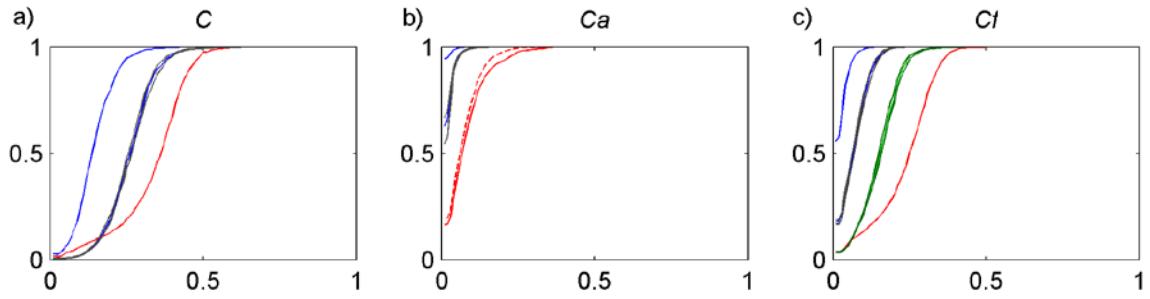
Poslední veličina je označena $C_f(p)$ a opět se liší od veličiny $C(p)$ pouze v podmínce 1, kterou nyní označme $1_f$.

$1_f$) Existuje alespoň jeden metabolit který inhibuje $r_i$, zároveň neinhibuje žádnou z reakcí obsažených v $p$ a dále existuje cesta od $r_i$ do tohoto metabolitu, která nemá žádné odbočující reakce.

Veličina $C_f(p)$ udává do jaké míry jsou odbočující reakce cesty $p$ inhibovány zpětnou vazbou metabolitem, ke kterému vede z $r_i$ cesta bez odboček.

Výše definované veličiny byly spočítány pro všechny metabolické cesty délky $l = 6$. Výsledek je možno znázornit jako kumulativní distribuční funkci těchto veličin, která udává pravděpodobnost: $P(C(p) \leq x)$. Aby bylo možno rozhodnout zdali je zkoumaná vlastnost reálné sítě signifikantní, je nutno jí porovnat s randomizovaným souborem regulačních sítí. Pro tento účel byly zkonstruovány 2 randomizované soubory regulačních sítí, které představují nulové modely. První soubor je náhodným vzorkem z množiny všech regulačních

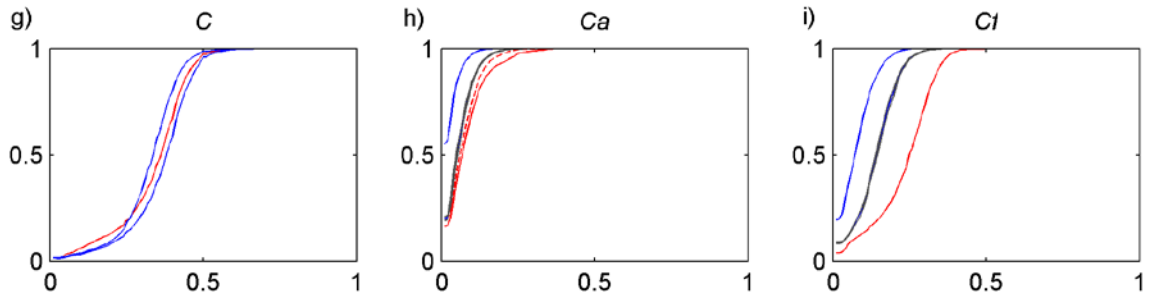
sítí se stejným počtem regulovaných reakcí, regulátorových metabolitů a regulačních interakcí jako má reálná síť (regulátory a regulované reakce jsou voleny náhodně z množiny všech metabolitů a reakcí). Maximální počet regulátorů na jednu reakci musí být v tomto souboru $\leq 15$, což je maximální konektivita pozorovaná v reálné regulační síti. Druhý soubor je představován náhodnými regulačními sítěmi, ve kterých jsou oproti předchozímu souboru fixovány regulované reakce, regulátorové metabolity jejich konektivity. Kumulativní funkce příslušející reálné síti jsou vzhledem k rozsahu kumulativních funkcí stejných veličin v případě obou randomizovaných souborů posunuty doprava, kromě veličiny $C(p)$ a druhého rand. souboru Obr.2,3. Tento fakt indikuje, že metabolické cesty v kontextu reálné regulační sítě vykazují signifikantně vyšší hodnoty zkoumaných veličin než v kontextu náhodné regulační sítě. Aby bylo možno vyloučit, že signifikance veličiny $C_f(p)$ je determinována signifikancí jednodušší veličiny, byl zkonstruován třetí rand. Soubor, který je vymezen stejně jako první s dodatečným omezením. Ve třetím rand. souboru je navíc oproti prvnímu fixován počet regulačních interakcí, pro které existuje cesta bez odbočujících reakcí od regulované reakce do regulátorového metabolitu. Na obr. 2 c) je znázorněn jen 5 percentil kumulativní funkce veličiny $C_f(p)$ ve třetím rand. souboru. Vidíme, že je $C_f(p)$ signifikantní i vzhledem k třetímu rand. souboru.




Obr.2. Porovnání zkoumaných veličin s prvním randomizovaným souborem. Do oblasti mezi modrými křivkami spadne 95% kumulativních funkcí příslušejících náhodným regulačním sítím z prvního souboru. Šedou barvou jsou označeny konfidenční intervaly pouze pro 5 percentil kumulativní funkce náhodné regulační sítě. Zeleně je označen 5 percentil kumulativní fukce náhodných sítí ze 3. rand. souboru.

Ve druhé části studie jsem se zaměřil na (auto) kanalizaci hypercest. Metabolickou hypercestu je možno chápat jako soubor reakcí, které umožňují přeměnu určité substrátové množiny na množinu produktů. Hypercesta je minimální, když po odstranění libovolné reakce z této hypercesty, zanikne možnost přeměny substrátové množiny na produktovou. Z důvodu

značné výpočetní náročnosti enumerace hypercest jsem se zaměřil pouze na minimální hypercesty vedoucí do glukózy.

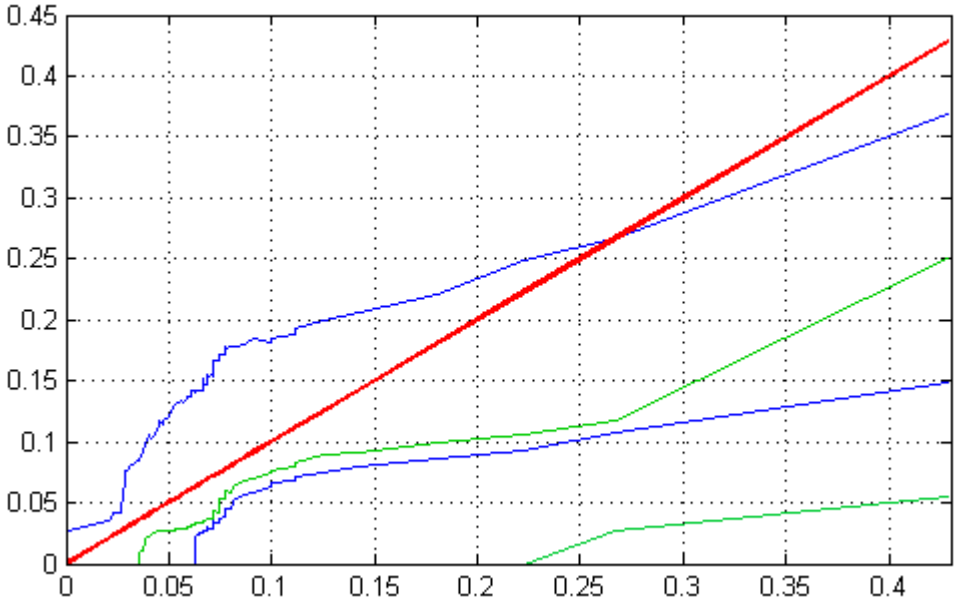



Obr.3. Porovnání zkoumaných veličin s druhým randomizovaným souborem. Do oblasti mezi modrými křivkami spadne 95% kumulativních funkcí příslušejících náhodným regulačním sítím z druhého rand. souboru. Šedou barvou jsou označeny konfidenční intervaly pouze pro 5 percentil kumulativní funkce náhodné regulační sítě (spodní modrá/šedá křivka).

Po provedení korekce, která je detailně popsána v článku (Geryk, submitted), je signifikance analyzovaných veličin je slabší než v předchozím případě. Zaměříme-li se na auto-kanalizaci hypercest je možno rozpoznat hraniční hodnotu $C_a(p)$ za kterou jsou hodnoty této veličiny signifikantně abundantní z hlediska prvního i druhého rand. souboru, tato hodnota je $\sim 0.25$ Obr.4.

Nejvyšší hodnota $C_a(p)$ přísluší hypercestám, které odpovídají anaerobní asimilaci glycerolu. Glycerol je oxidován glycerol dehydrogenázou na dihydroxyaceton a poté fosforylován dihydroxyaceton kinázou na dihydroxyaceton fosfát, který následně vstupuje do glukoneogeneze. Nedávno bylo experimentálně potvrzeno, že *E.coli* dokáže anaerobně asimilovat glycerol a že glycerol dehydrogenázoa a dihydroxyaceton kináza jsou esenciální pro tento proces (Gonzalez et al., 2008).

Jiná hypercesta odpovídá klasické glukoneogenezi. V této hypercestě je zajímavá inhibice fosfoglycerát dehydrogenázy glycinem a serinem. Tato inhibice nejspíš hraje důležitou roli v situaci, kdy jsou pro baktérii zdrojem uhlíku a energie aminokyseliny. Taková situace může nastat v případě uropatogenní *E.coli*, pro kterou jsou peptidy a aminokyseliny hlavním zdrojem v močovém systému. Aminokyseliny serin, glycin a arginin jsou konvertovány na pyruvát a oxaloacetát. Oxaloacetát je dále konvertován na fosfoenol-pyruvát, který vstupuje do glukoneogenetické hypercesty (Alteri et al., 2009).




Obr.4. Na obou osách jsou vyneseny percentily veličiny $C_a(p)$. Červená přímka odpovídá porovnání reálné regulační sítě sama se sebou. Ostatní křivky odpovídají porovnání percentilů reálné regulační sítě s odpovídajícími percentily v rand. souborech. Pro x-tý percentil v reálné síti, jsou spočítány konfidenční meze 90% rozsahu x-tého percentilu v randomizovaném souboru, x=1,...,100 a tyto meze vyneseny do grafu. Modré křivky odpovídají konfidenčním mezím v druhém rand. souboru. Zelené křivky odpovídají konfidenčním mezím v prvním rand. souboru.

Výše zmíněná inhibice fosfoglycerát dehydrogenázy brání tomu aby byly glycin a serin zpětně syntetizovány z glukoneogenetického intermediátu, čímž by se vytvořil tzv. futilní cyklus, který by pouze disipoval energii. Tato inhibice kanalizuje metabolický tok v glukoneogenezi a zároveň je ji možno považovat za zpětnovazebnou inhibici syntézy glycinu a serinu.

## Závěr

V druhé práci (Geryk and Slanina) bylo ukázáno, že část regulačních interakcí metabolické sítě *E.coli* se v některých oblastech koncentruje a vytváří tak denzně propojené moduly. Tato koncentrace, kvantifikovaná veličinou $Q$, je signifikantní z hlediska randomizovaného souboru regulačních sítí. Bližší inspekce detekovaných modulů ukázala, že se jedná o funkčně interpretovatelné systémy lokální zpětnovazebné inhibice nebo kanalizace metabolického toku. Koncentraci regulačních vazeb do modulů lze považovat za nepřímý důsledek evoluční optimalizace lokální regulace v určitých místech metabolické sítě. Vznik zpětnovazebné regulace na krátkou vzdálenost zvýší lokálně hustotu hran a může tedy přispět i k celkové

asymetrii v distribuci hran sítě a v důsledku toho ke zvýšení modularity. Je zřejmé, že moduly definované jako oblasti sítě s relativně vyšší hustotou hran mohou poskytnout jen velmi hrubý vhled do topologické organizace metabolické sítě. Třetí práce je zaměřena na jemnější topologický aspekt - regulační kanalizaci metabolických cest. V práci (Geryk, submitted) bylo prokázáno, že i v tomto aspektu se reálná regulační síť *E.coli* signifikantně odlišuje od randomizovaných souborů regulační sítě. Bylo ukázáno, že metabolické cesty jsou kanalizovány inhibicí odbočujících reakcí a) metabolity, které jsou produkovány reakcemi v těchto cestách b) metabolity, které jsou produktem jiných metabolických cest začínajících v odbočujících reakcích a vytvářejících tak nejjednodušší zpětnou vazbu. V práci (Nishikawa et al., 2008) je ukázáno, že konfigurace metabolických toků, která maximalizuje růstovou rychlost baktérie je charakteristická velkým množstvím reakcí, které mají nulový tok. Aby baktérie maximalizovala svůj růst musí umět vypnout celou řadu reakcí. V kontextu těchto výsledků se nabízí hypotéza, že alosterické regulace se podílejí na tlumení reakcí, které jsou v daném růstovém prostředí nežádoucí. Efekt transkripční regulace se projevuje řádově v minutách, alosterická regulace je téměř okamžitá, může tedy hrát roli zejména v počáteční fázi adaptace metabolismu na nové prostředí.

Signifikance studovaných vlastností regulační sítě vzhledem k nulovým modelům naznačuje, že jsou tyto vlastnosti projevem evolučních adaptací. Prezentované výsledky také ukazují, že má smysl analyzovat regulační síť reprezentující přímé regulační. interakce na vyšší úrovni abstrakce než jsou dynamické modely. Můžou ták být odkryty důležité vlastnosti, které by v komplexitě dynamického modelu zůstaly nepovšimnuty.

## Použitá literatura


ALBERT, R. & BARABASI, A. L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics,* 74**,** 47-97.

ALBERT, R., JEONG, H. & BARABASI, A. L. 1999. Internet - Diameter of the World-Wide Web. *Nature,* 401**,** 130-131.

ALBERT, R., JEONG, H. & BARABASI, A. L. 2000. Error and attack tolerance of complex networks. *Nature,* 406**,** 378-82.

ARITA, M. 2004. The metabolic world of Escherichia coli is not small. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 1543-1547.

BAGROW, J. P. 2012. Communities and bottlenecks: Trees and treelike networks have high modularity. *Physical Review E,* 85.

BARABASI, A. L. 2005. Taming complexity. *Nature Physics,* 1**,** 68-70.

BARABASI, A. L. 2007. The Architecture of complexity. *Ieee Control Systems Magazine,* 27**,** 33-42.

BARABASI, A. L. 2009. Scale-free networks: a decade and beyond. *Science,* 325**,** 412-3.

BARABASI, A. L. 2012. The network takeover. *Nature Physics,* 8**,** 14-16.

BARABASI, A. L. & ALBERT, R. 1999. Emergence of scaling in random networks. *Science,* 286**,** 509-512.

BARABASI, A. L. & OLTVAI, Z. N. 2004. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics,* 5**,** 101-U15.

BARRETT, C. L., HERRING, C. D., REED, J. L. & PALSSON, B. O. 2005. The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 19103-19108.

BASLER, G., EBENHOH, O., SELBIG, J. & NIKOLOSKI, Z. 2011. Mass-balanced randomization of metabolic networks. *Bioinformatics,* 27**,** 1397-1403.

BASLER, G., GRIMBS, S., EBENHOH, O., SELBIG, J. & NIKOLOSKI, Z. 2012. Evolutionary significance of metabolic network properties. *J R Soc Interface,* 9**,** 1168-76.

CLAUSET, A., NEWMAN, M. E. J. & MOORE, C. 2004. Finding community structure in very large networks. *Physical Review E,* 70.

COHEN, R. & HAVLIN, S. 2003. Scale-free networks are ultrasmall. *Physical Review Letters,* 90**,** 058701.

EISENBERG, E. & LEVANON, E. Y. 2003. Preferential attachment in the protein network evolution. *Physical Review Letters,* 91.

ERDŐS, P. & RÉNYI, A. 1959a. On Random Graphs. *Publicationes Mathematiae,* 6**,** 290-297.

ERDŐS, P. & RÉNYI, A. 1959b. On the Evolution of Random Graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl,* 5.

FORTUNATO, S. 2010. Community detection in graphs. *Physics Reports-Review Section of Physics Letters,* 486**,** 75-174.

FORTUNATO, S. & BARTHELEMY, M. 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 36-41.

GERYK, J. & SLANINA, F. Modules in the metabolic network of *E.coli* with regulatory interaction. *International Journal of Data Mining and Bioinformatics,* "in press".

GONZALEZ, R., MURARKA, A., DHARMADI, Y. & YAZDANI, S. S. 2008. A new model for the anaerobic fermentation of glycerol in enteric bacteria: trunk and auxiliary pathways in Escherichia coli. *Metab Eng,* 10**,** 234-45.

GUIMERA, R. & AMARAL, L. A. N. 2005. Functional cartography of complex metabolic networks. *Nature,* 433**,** 895-900.
GUIMERA, R., SALES-PARDO, M. & AMARAL, L. A. 2007. Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics,* 3**,** 63-69.
GUTTERIDGE, A., KANEHISA, M. & GOTO, S. 2007. Regulation of metabolic networks by small molecule metabolites. *BMC Bioinformatics,* 8**,** 88.
HAO, D. & LI, C. 2011. The dichotomy in degree correlation of biological networks. *PLoS One,* 6**,** e28322.
HAO, D., REN, C. & LI, C. 2012. Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. *Bmc Systems Biology,* 6**,** 34.
HU, Y., DING, Y., FAN, Y. & DI, Z. 2010a. How to Measure Significance of Community Structure in Complex Networks. *arXiv:1002.2007 [physics.soc-ph]*.
HU, Y., NIE, Y., YANG, H., CHENG, J., FAN, Y. & DI, Z. 2010b. Measuring the significance of community structure in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys,* 82**,** 066106.
IHMELS, J., LEVY, R. & BARKAI, N. 2004. Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nature Biotechnology,* 22**,** 86-92.
JEONG, H., NEDA, Z. & BARABASI, A. L. 2003. Measuring preferential attachment in evolving networks. *Europhysics Letters,* 61**,** 567-572.
JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. & BARABASI, A. L. 2000. The large-scale organization of metabolic networks. *Nature,* 407**,** 651-654.
KASHTAN, N. & ALON, U. 2005. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 13773-13778.
KREIMER, A., BORENSTEIN, E., GOPHNA, U. & RUPPIN, E. 2008. The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 6976-6981.
LEWIS, A. C., JONES, N. S., PORTER, M. A. & DEANE, C. M. 2010. The function of communities in protein interaction networks at multiple scales. *Bmc Systems Biology,* 4**,** 100.
LIGHT, S., KRAULIS, P. & ELOFSSON, A. 2005. Preferential attachment in the evolution of metabolic networks. *Bmc Genomics,* 6.
MA, H. W. & ZENG, A. P. 2003. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics,* 19**,** 270-277.
MA, H. W., ZHAO, X. M., YUAN, Y. J. & ZENG, A. P. 2004. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics,* 20**,** 1870-1876.
NELLY LITVAK & HOFSTAD, R. V. D. 2012. Uncovering disassortativity in large scale-free networks. *arXiv:1204.0266 [physics.soc-ph]*.
NEWMAN, M. E. J. 2001. Clustering and preferential attachment in growing networks. *Physical Review E,* 64.
NEWMAN, M. E. J. 2004. Fast algorithm for detecting community structure in networks. *Physical Review E,* 69.
NEWMAN, M. E. J. & GIRVAN, M. 2004. Finding and evaluating community structure in networks. *Physical Review E,* 69.
NISHIKAWA, T., GULBAHCE, N. & MOTTER, A. E. 2008. Spontaneous Reaction Silencing in Metabolic Optimization. *Plos Computational Biology,* 4.
PAPADOPOULOS, F., KITSAK, M., SERRANO, M. A., BOGUNA, M. & KRIOUKOV, D. 2012. Popularity versus similarity in growing networks. *Nature,* 489**,** 537-40.

PARTER, M., KASHTAN, N. & ALON, U. 2007. Environmental variability and modularity of bacterial metabolic networks. *Bmc Evolutionary Biology,* 7.
PASTOR-SATORRAS, R. & VESPIGNANI, A. 2001. Epidemic spreading in scale-free networks. *Physical Review Letters,* 86**,** 3200-3.
RAVASZ, E., SOMERA, A. L., MONGRU, D. A., OLTVAI, Z. N. & BARABASI, A. L. 2002. Hierarchical organization of modularity in metabolic networks. *Science,* 297**,** 1551-1555.
SAMAL, A. & JAIN, S. 2008. The regulatory network of E. coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *Bmc Systems Biology,* 2.
SAMAL, A. & MARTIN, O. C. 2011. Randomizing genome-scale metabolic networks. *PLoS One,* 6**,** e22295.
SAMAL, A., WAGNER, A. & MARTIN, O. C. 2011. Environmental versatility promotes modularity in genome-scale metabolic networks. *Bmc Systems Biology,* 5.
SESHASAYEE, A. S. N., FRASER, G. M., BABU, M. M. & LUSCOMBE, N. M. 2009. Principles of transcriptional regulation and evolution of the metabolic system in E. coli. *Genome Research,* 19**,** 79-91.
TANAKA, R. 2005. Scale-rich metabolic networks. *Physical Review Letters,* 94.
WAGNER, A. 2009. Evolutionary constraints permeate large metabolic networks. *Bmc Evolutionary Biology,* 9.
WAGNER, A. & FELL, D. A. 2001. The small world inside large metabolic networks. *Proceedings of the Royal Society of London Series B-Biological Sciences,* 268**,** 1803-1810.
WANG, Z. & ZHANG, J. Z. 2007. In search of the biological significance of modular structures in protein networks (vol 3, pg 1011, 2007). *Plos Computational Biology,* 3**,** 1404-1404.
WATTS, D. J. & STROGATZ, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature,* 393**,** 440-2.
XU, K., BEZAKOVA, I., BUNIMOVICH, L. & YI, S. V. 2011. Path lengths in protein-protein interaction networks and biological complexity. *Proteomics,* 11**,** 1857-67.
ZHANG, Z. & ZHANG, J. 2009. A big world inside small-world networks. *PLoS One,* 4**,** e5686.
ZHAO, J., DING, G. H., TAO, L., YU, H., YU, Z. H., LUO, J. H., CAO, Z. W. & LI, Y. X. 2007. Modular co-evolution of metabolic networks. *Bmc Bioinformatics,* 8**,** 311.
ZHAO, J., YU, H., LUO, J. H., CAO, Z. W. & LI, Y. X. 2006. Hierarchical modularity of nested bow-ties in metabolic networks. *Bmc Bioinformatics,* 7.

# Hledání (správné) míry modularity

**Jan Geryk**

Katedra filosofie a dějin přírodních věd
Viničná 7, Praha 2
geryk.jan@seznam.cz

**Abstrakt**

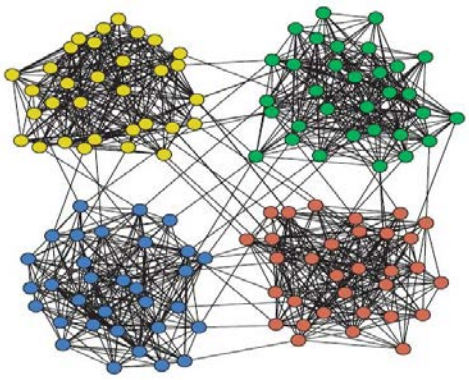
Složité sítě různých vztahů nacházíme ve všech oblastech přístupných lidskému poznání. Při maximální abstrakci je možno na takový systém nahlížet jako na soustavu bodů propojených čárami - hranami. Jednou z typických charakteristik reálných sítí je modularita. O modulární síti hovoříme v případě, když ji můžeme rozdělit na podsítě (moduly), ve kterých je významně vyšší hustota hran než mezi těmito podsítěmi. Tento příspěvek se zaměřuje na nedostatky, resp. vlastnosti standardně používané míry modularity (*Q*). Je zde ukázáno, že i relativně velké moduly identifikované na základě funkce *Q* mohou být složeny z většího počtu menších, silněji vyjádřených submodulů. Z hlediska zachycení lokální, modulární struktury sítě tedy míra modularity *Q* není optimálním řešením.

## 1 Úvod

Žijeme ve světě interakcí. Ať už se jedná o interakce mezi lidmi, www stránkami nebo mezi molekulami v buňce, ve všech případech tyto interakce vytváří efektivně propojené sítě s velmi složitou architekturou. Přítomnost komplexních sítí je typickým atributem živých organismů a jejich produktů. Jediná buňka mnohobuněčného organismu obsahuje řádově tisíce interagujících molekul, což ji umožňuje vykonávat celou řadu složitých úloh. Lidský mozek je tvořen desítkami miliard neuronů, které jsou opět propojeny do sítě pomocí synapsí a dendritů. Konečně internet, produkt lidského mozku, spojuje přibližně 4,2 miliard počítačů a zásadním způsobem odstraňuje geografická omezení prapůvodní sítě lidských kontaktů (z očí do očí).

Nejjednodušší formální reprezentace sítě je neorientovaný graf. Ten je definován jako množina objektů, kde pro každou dvojici těchto objektů - vrcholů je určeno, jestli spolu interagují (či nikoliv). Když spolu interagují, říkáme, že jsou spojeny hranou. Graf tedy zachycuje jakousi elementární architekturu sítě. Grafová reprezentace reálného systému smělým způsobem odhlíží od celé řady jeho relevantních vlastností. Právě tato extrémní abstrakce přináší naději na objevení nových souvislostí, které by jinak zůstaly nepovšimnuty v informační záplavě detailního modelu. Existují již některé doklady toho, že topologie sítě zásadním způsobem ovlivňuje její dynamiku [1,2,3]. Jestliže je předpoklad o významu topologie pro funkci sítě správný, měly by především biologické sítě, řídící aktivitu buněk, orgánů a jedinců, vykazovat specifickou architekturu, optimalizovanou pro zajištění životních funkcí. Cílem současného výzkumu je nalézt topologické charakteristiky reálných sítí, ve kterých se tyto odlišují od náhodného modelu. V případě nitrobuněčných regulačních sítí je na místě předpokládat, že jejich architektura má adaptační význam a tento předpoklad ověřovat pro všechny nalezené topologické odchylky od náhodného modelu. Nalezení biologicky relevantní funkce, pro kterou je zkoumaná topologická odchylka optimálním řešením, je možno považovat za indikátor selekčně-evolučního původu této odchylky.

Typickou charakteristikou většiny reálných sítí, včetně sítí nitobuňečných, je modularita. Intuitivně ji lze chápat jako přítomnost hustě propojených oblastí sítě, kde hustota hran mezi těmito oblastmi je nižší než uvnitř těchto oblastí, tzv. modulů obr.1. Přestože pojem modularity je intuitivně dobře přístupný, jeho fomalizace není jednoznačná. Existuje celá řada různých definic modulu. Ty lze rozdělit na definice lokální a globální.



Obr.1. Modulární graf. (převzato z [9])

Lokální definice uvažují pouze vyšetřovaný podgraf a jeho blízké okolí, nezávisle na zbytku grafu. Modul může

být definován v silném pojetí jako podgraf, kde pro každý vrchol toho podgrafu platí, že počet jeho sousedů ležících ve vyšetřovaném podgrafu je větší než počet jeho sousedů ležících mimo tento podgraf. V slabém pojetí je modul definován jako podgraf, pro který platí, že dvojnásobek počtu hran ležících uvnitř tohoto podgrafu je větší než počet hran, které jej spojují se zbytkem grafu [4].

Globální definice nahlíží na daný podgraf z hlediska struktury celého grafu. V jejich základu je porovnání struktury vyšetřovaného podgrafu s náhodným modelem zkoumaného grafu. Náhodný model, který zpravidla zachovává některé elementární vlastnosti původního grafu, zanáší do definice modulu globální aspekt. Typicky se v náhodném modelu fixuje sekvence konektivit - tj. pro každý vrchol fixujeme počet jeho sousedů. V rámci těchto omezení pak uvažujeme náhodnou distribuci hran. Modul můžeme pak definovat jako podgraf, který vykazuje signifikantně vyšší hustotu hran, než bychom očekávali v randomizované verzi grafu na stejné množině vrcholů [5].

V současnosti nejpoužívanější definice modulu/modularity je implikována funkcí $Q$ (je známa jako „modularity measure“ [6,7,8]. Pro libovolné rozdělení grafu na $M$ podgrafů, je $Q$ mírou modularity tohoto rozdělení.:

$$Q = \sum_{i=1}^{M} \left[ \frac{l_i}{L} - \left( \frac{d_i}{2L} \right)^2 \right], \qquad (1)$$

kde sčítáme přes všechny podgrafy/moduly. $L$ je počet hran v grafu, $l_i$ je počet hran v podgrafu $i$, $d_i$ je součet stupňů vrcholů v podgrafu $i$, tj. $d_i = \sum_{v \in i} k_v$, kde $k_v$ značí počet sousedů uzlu $v$.

$Q$ vyjadřuje rozdíl frakce hran grafu obsažených v podgrafech (1,..,$M$) a stejné kvantity v případě náhodné verze studovaného grafu. Náhodný model zde představuje soubor všech grafů ($\mathbf{G}_d$), které lze vytvořit, při zachování sekvence konektivit původního grafu. Člen $(d_i/2L)^2$ vyjadřuje odhad průměrné frakce hran v podgrafu $i$ v souboru $\mathbf{G}_d$.

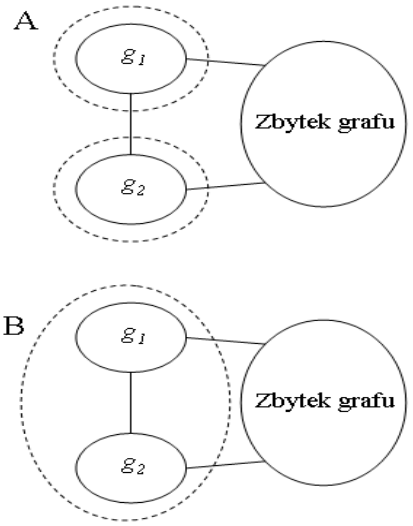
Výhoda srovnání intramodulární frakce hran s náhodnou verzí grafu se zachovanou sekvencí konektivit spočívá v tom, že odfiltrujeme vliv této sekvence konektivit na modularitu sítě. Nalezneme-li modulární strukturu s použitím $Q$, můžeme se domnívat, že má vlastní příčinu a není pouze nutným důsledkem procesů, které vedou k ustavení příslušné distribuce konektivit ve zkoumaném grafu.

Nejčastěji používaná metoda identifikace modulů v grafu spočívá v optimalizaci $Q$ v prostoru všech možných rozdělení grafu na podgrafy. Počet možných rozdělení grafu na podgrafy roste exponenciálně s počtem vrcholů grafu a tak je nutno použít stochastické optimalizační techniky (např. metodu simulovaného žíhání) [9,10].

## 2 Rozlišovací limit modularity measure

Definice $Q$ se zdá být zcela přirozená, některé její důsledky však odporují základní intuitivní představě modulu. Tento paradox je možno ukázat na speciální situaci, která je detailně popsána v práci [11].

Autoři této práce uvažují dva podgrafy ($g_1$, $g_2$), oba se stejným počtem hran $l_1 = l_2 = l$. Tyto podgrafy jsou vzájemně spojeny jednou hranou. Každý z nich je dále spojen jednou hranou se zbytkem grafu. Základem je porovnání hodnot $Q$ pro dvě různá rozdělení grafu na podgrafy: A) podgrafy $g_1$ a $g_2$ budou každý tvořit samostatný podgraf - tomuto rozdělení bude příslušet hodnota $Q_A$, B) podgrafy $g_1$ a $g_2$ budou brány jako jeden podgraf - tomuto rozdělení bude příslušet hodnota $Q_B$ obr.2. Volba zbytku grafu a jeho rozdělení je arbitrární s tím, že je pro oba uvažované případy (A, B) stejná.




Obr.2. Modelová situace pro odvození rozlišovacího limitu. V rozdělení A jsou $g_1$ a $g_2$ samostatné podgrafy, v rozdělení B tvoří jeden podgraf.

Ptejme se nyní za jakých podmínek je $Q_A < Q_B$. Autoři Fortunato a Barthélemy [11] ukázali, že tato nerovnost je splněna, když:

$$l < l_R = \sqrt{\frac{L}{2}} - 1 \ ^{1}. \qquad (2)$$

Vidíme, že tato kritická hodnota $l_R$ nezávisí na počtu vrcholů v podgrafech $g_1$ a $g_2$. Tyto podgrafy proto můžeme volit tak aby byly klikami[2] a zároveň splňovaly podmínku (2). V takovém případě představují podgrafy $g_1$ i $g_2$ samostatné moduly z hlediska silné i slabé lokální definice. Autonomie podgrafů $g_1$ a $g_2$ je v tomto extrémním případě na první pohled zřejmá. Míra modularity $Q$ však rozdělení A (které bychom intuitivně preferovali) přiřazuje nižší hodnotu než rozdělení B.

Fortunato a Barthélemy na základě tohoto faktu upozorňují na riziko chybné identifikace funkčních modulů. Získáme-li rozdělení grafu na moduly pomocí optimalizace $Q$, je zde jisté riziko, že nalezené moduly s počtem hran $\leq 2l_R = \sqrt{2L} - 2$ budou složeny ze dvou, či více jasně vyjádřených modulů. A to vyjádřených zřetelněji než moduly identifikované pomocí optimalizace $Q$.

Existence rozlišovacího limitu je zřejmě důsledkem použití náhodného modelu studovaného grafu jakožto nulové hypotézy.

## 2.1 Obecný rozlišovací limit

Přejděme k zobecnění rozlišovacího limitu $Q$ pro arbitrární náhodný model. Tento náhodný model je nyní libovolná množina grafů (**G**) na množině vrcholů výchozího grafu. Zaveďme obecnou formulaci modularity measure:

$$Q = \frac{1}{L}\sum_{i=1}^{M}\left(l_i - \langle l_i \rangle_{\mathbf{G}}\right), \qquad (3)$$

kde $\langle l_i \rangle_{\mathbf{G}}$ je střední hodnota počtu hran v **G** na množině vrcholů podgrafu $i$. Platí, že:

$$\langle l_i \rangle_{\mathbf{G}} = \sum_{\{u<v\}\in i} p(u,v)\,, \qquad (4)$$

kde $p(u,v) = |\mathbf{G}(u,v)| / |\mathbf{G}|$. $|\mathbf{G}|$ je velikost množiny **G** a $|\mathbf{G}(u,v)|$ je počet grafů z **G**, ve kterých jsou vrcholy $u$ a $v$ spojeny hranou. $p(u,v)$ tedy vyjadřuje pravděpodobnost existence hrany $\{u,v\}$ v náhodně vybraném grafu z **G**.

Uvažujme opět dvě možnosti rozdělení grafu na podgrafy. Rozdělení A´ je nyní tvořeno $m$ samostatnými podgrafy ($g_1$ ,..., $g_m$) a zbylou částí grafu ($g_0$), jejíž rozdělení je arbitrární. V rozdělení B´ jsou podgrafy $g_1$ ,..., $g_m$ sloučeny v jeden, rozdělení zbylé části grafu $g_0$ je zde stejné jako v rozdělení A´. Z definice (3) dostáváme:

$$Q_{A'} = Q_0 + \frac{1}{L}\left[\sum_{i=1}^{m} l_i - \sum_{i=1}^{m}\langle l_i \rangle_{\mathbf{G}}\right], \qquad (5)$$

$$Q_{B'} = Q_0 + \frac{1}{L}\left[\sum_{i=1}^{m} l_i - \sum_{i=1}^{m}\langle l_i \rangle_{\mathbf{G}} + \sum_{0<i<j} l_{ij} - \sum_{0<i<j}\langle l_{ij} \rangle_{\mathbf{G}}\right], \qquad (6)$$

kde $l_i$ je počet hran v podgrafu $i$, $l_{ij}$ je počet hran spojujících podgrafy $i$ a $j$, $\langle l_{ij} \rangle_{\mathbf{G}}$ je střední hodnota počtu hran v **G** mezi vrcholy podgrafů $i$ a $j$ (hodnoty $l_{ij}$ a $\langle l_{ij} \rangle_{\mathbf{G}}$ se sčítají přes všechny neuspořádané dvojice podgrafů $g_1$ ,..., $g_m$), $Q_0$ je modularita zbylé části grafu $g_0$. Pro rozdíl $\Delta Q = Q_{A'} - Q_{B'}$ zřejmě platí:

$$\Delta Q = \sum_{i<j}\langle l_{ij} \rangle_{\mathbf{G}} - \sum_{0<i<j} l_{ij}\ . \qquad (7)$$

Podmínka $\Delta Q < 0$ je splněna pro:

$$\sum_{0<i<j}\langle l_{ij} \rangle_{\mathbf{G}} < \sum_{0<i<j} l_{ij}\ . \qquad (8)$$

Nerovnost (8) představuje obecný rozlišovací limit modularity measure.

V typickém náhodném modelu hodnota $\sum_{0<i<j}\langle l_{ij} \rangle_{\mathbf{G}}$ závisí na počtu hran ve zkoumaném grafu. Při konstantní lokální struktuře skupiny podgrafů $g_1$ ,..., $g_m$ můžeme vhodnou volbou zbytku grafu měnit hodnotu $\Delta Q = Q_A - Q_B$ a tedy i samotnou možnost identifikace $g_1$ ,..., $g_m$ jakožto samostatných modulů.

## 2.2 2 modely náhodného grafu

Uvažujme nejprve použití Erdős-Rényi (E-R) náhodného modelu v definici (3). V tomto modelu je pravděpodobnost existence hrany stejná pro všechny dvojice vrcholů: $p(u,v) = p = L\binom{N}{2}^{-1}$. Předpokládejme, že podgrafy $g_1$ ,..., $g_m$ mají každý stejný počet vrcholů

[1] V práci [11] je jednička v nerovnosti (2) zanedbána.

[2] Klika je podgraf s maximálním počtem hran takový, že každé jeho zvětšení povede k podgrafu, který nebude mít maximální možný počet hran.

($n$). Mezi každou dvojici těchto podgrafů položme 1 hranu. Střední hodnota počtu hran mezi každou dvojicí podgrafů z množiny $g_1$ ,...,$g_m$ je nyní: $\langle l_{ij} \rangle_{\mathbf{G}\,ER} = pn^2$. Dosazením do (8) dostáváme:

$$\frac{m(m-1)}{2} pn^2 < \frac{m(m-1)}{2} \qquad (9)$$

a tedy:

$$n < \sqrt{\frac{1}{p}} \quad . \qquad (10)$$

Tato nerovnost představuje rozlišovací limit při použití E-R modelu jakožto nulové hypotézy. Bude-li splněna podmínka (8), bude množina podgrafů $g_1$ ,..., $g_m$ považována za jeden modul (na základě hodnoty $Q$). Zvolíme-li vhodně zbytek grafu, mohou být podgrafy $g_1$ ,..., $g_m$ klikami a přesto bude rozdělení, kde jsou sloučeny v jeden podgraf, preferováno.

V případě náhodného modelu, který zachovává sekvenci konektivit, předpokládejme stejný počet hran ($l$) v jednotlivých podgrafech $g_1$ ,..., $g_m$ a každý z nich spojme jednou hranou se zbytkem grafu. Mezi každou dvojici podgrafů $g_1$ ,..., $g_m$ položme 1 hranu. Odhad pravděpodobnosti existence hrany v $\mathbf{G}_d$ (tj. souboru všech grafů se stejnou sekvencí konektivit jako má vyšetřovaný graf) je: $p(u,v) \approx k_v k_u / 2L$, kde $k_v$ je počet sousedů uzlu $v$. Na základě (4) platí, že $\langle l_{ij} \rangle_{\mathbf{G}\,d} = \frac{d_i d_j}{2L}$, kde $d_i = \sum_{v \in i} k_v$. Z našich předpokladů plyne, že $d_i = 2l + m$ pro $i$=1,...,$m$. Dosazením do nerovnosti (8) dostáváme:

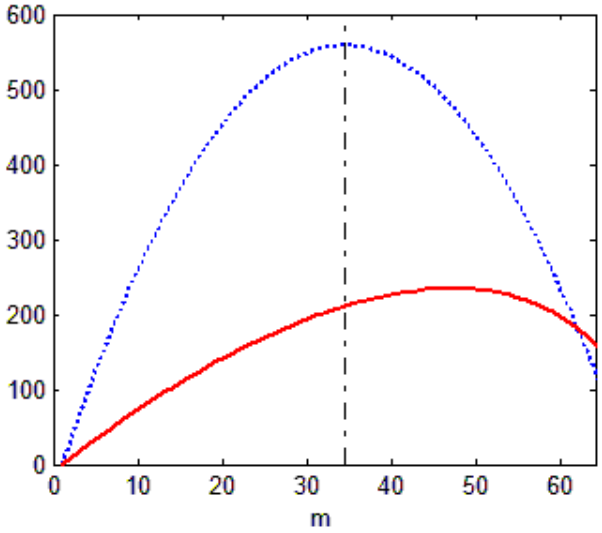
$$\frac{m(m-1)}{2} \frac{(2l+m)^2}{2L} < \frac{m(m-1)}{2}, \qquad (11)$$

a po úpravě:

$$l < l_{R'} = \sqrt{\frac{L}{2}} - \frac{m}{2}. \qquad (12)$$

Pro $m$ =2 a nerovnost (12) přechází v nerovnost (2) odvozenou v [11]. V limitním případě může být nerovnost (12) splněna pro $m$ stejných klik, kde každá je spojena jednou hranou se zbytkem sítě a mezi každou dvojicí těchto klik je také jedna hrana. Z toho plyne, že k "chybnému" sloučení jednoznačně vyjádřených funkčních submodulů může dojít i v modulech větších než je hodnota $2l_R = \sqrt{2L} - 2$.

Známe-li hodnotu $l_{R'}$ pro dané $m$, můžeme odhadnout počet vrcholů podgrafu ($n_{R'}$ ) tak aby byl klikou. Stačí řešit rovnici $\frac{1}{2} n_{R'}(n_{R'} - 1) = l_{R'}$. Na obr. 3 je vynesena hodnota $mn_{R'}$ v závislosti na $m$ (spojitá čára). Celkový počet hran grafu $L$ =2234 odpovídá počtu hran proteininterakční sítě kvasinky *S. cerevisae*. Hodnota $mn_{R'}$ udává maximální velikost podgrafu složeného s $m$ klik, tak aby tyto kliky nebyly identifikovány jakožto samostatné moduly při použití modularity measure. Nalezené moduly o velikosti $\leq mn_{R'}$ tak mohou sestávat s $\geq m$ relativně řídce propojených klik, které splňují lokální definici modulu. Z hlediska slabé lokální definice (pro $L$ =2234) představují kliky $g_1$ ,..., $g_m$ moduly až do hodnoty $m$ = 34 - této hodnotě odpovídá hodnota $mn_{R'} \approx 209$.



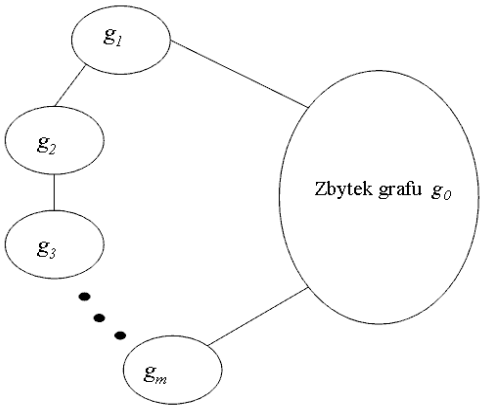

Obr.3. Spojitá čára znázorňuje závislost $mn_{R'}$ na $m$, přerušovaná čára odpovídá závislosti $ml_{R'}$ na $m$. Z hlediska slabé definice jsou kliky $g_1$ ,..., $g_m$ moduly až do hodnoty $m$ = 34 - tato hranice je vyznačena čerchovaně. (Funkce $mn_{R'}$ a $ml_{R'}$ jsou vykresleny jen pro $l_{R'} > 0$.)

Uvažujme ještě jeden typ propojení podgrafů $g_1$ ,..., $g_m$ (každý se stejným počtem hran ($l$)). Spojme tyto podgrafy tak aby tvořily spolu se zbytkem grafu $g_0$ kružnici obr.4. V takovém případě bude $d_i$ =2$l$ + 2, pro $i$=1,...,$m$. Dosazením do nerovnosti (8) získáváme:

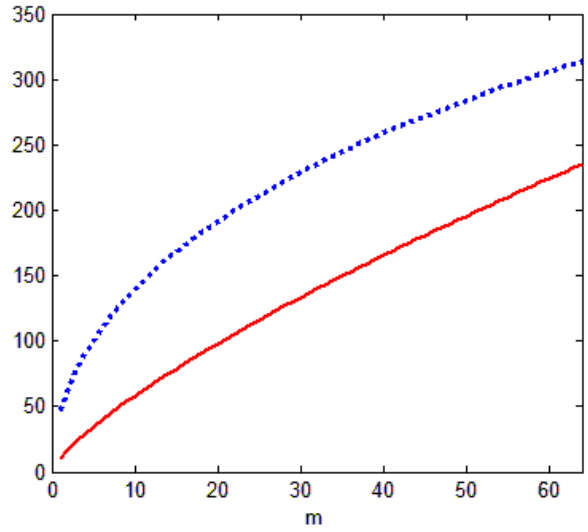
$$l < l_{R''} = \sqrt{\frac{L}{m}} - 1. \qquad (13)$$

Pro $m$ = 2 dostáváme opět nerovnost (2). Nyní je systém podgrafů $g_1$ ,..., $g_m$ propojen minimálním počtem hran tak aby ještě zůstal souvislý. Analogicky předchozímu

případu uvažujme hodnoty $mn_{R''}$ a $ml_{R''}$ v závislosti na $m$ obr.5. Z tohoto obrázku je vidět, že riziko „přehlédnutí“ skupiny relevantních modulů přetrvává i v případě, kdy jsou podgrafy $g_1$ ,..., $g_m$ propojeny minimálním počtem hran při zachování souvislosti.




Obr.4. Situace pro odvození rozlišovacího limitu (13).




Obr. 5. Spojitá čára znázorňuje závislost $mn_{R''}$ na $m$, přerušovaná čára odpovídá závislosti $ml_{R''}$ na $m$.

## 3 Závěr

Rozlišovací limit je důsledkem použití náhodného modelu jakožto nulové hypotézy, jak již bylo naznačeno v [8]. Z nerovnosti (8) a z faktu, že v typickém náhodném modelu je lokální hustota hran závislá na hustotě hran grafu, který je předmětem randomizace, je tato skutečnost zřejmá. Konkrétně bylo ukázáno, že lze zkonstruovat systém podgrafů sestávající s relevantních modulů (z hlediska lokálních definic) tak aby tyto moduly nebyly odhaleny při použití modularity measure. Libovolný modul nalezený pomocí $Q$ tak může být tvořen více jednoznačně vyjádřenými submoduly. Autoři Fortunato a Barthélemy v práci [11] zmiňují, že u větších modulů než $2l_R = \sqrt{2L} - 2$ je menší riziko, že sestávají s více submodulů protože tyto submoduly by musely být slaběji vyjádřeny. Nerovnost (13) však ukazuje, že i moduly $> 2l_R$ mohou být silně (jednoznačně) vyjádřeny při splnění podmínky $Q_{A'} < Q_{B'}$.

V případě proteininterakční sítě *S. cerevisae* se kritické hodnoty $n_{R'}$ a $n_{R''}$ pohybují na spodní hranici velikosti funkčních modulů. (Jsou však stále relevantní až do hodnoty $m \approx 40$.). Pro slaběji vyjádřené moduly, které však stále mohou splňovat lokální definice modulu, bude hodnota rozlišovacího limitu vyšší.

Problém rozlišovacího limitu úzce souvisí s hierarchickým uspořádáním sítě. Menší moduly, zajišťující elementární funkce, se mohou kombinovat do větších modulů s komplexnější funkcí [12,13]. Jestliže moduly na různých hierarchických úrovních vykazují funkční autonomii, úloha detekce modulů pak spočívá v nalezení optimálního rozdělení sítě pro každou z těchto úrovní. Stále však zůstává problém porovnání míry modularity mezi jednotlivými úrovněmi.

Identifikace modulů na základě míry modularity $Q$ vykazuje vážné nedostatky. Pro detekci funkčních modulů je vhodnější použít lokální metody - ty mohou přinést lepší výsledky. V současnosti se problémy spojené s modularitou reálných sítí intenzivně zkoumají a tak je možno očekávat stále nové poznatky.

## Literatura


[1] R. J. Prill, P. A. Iglesias, A. Levchenko: Dynamic properties of network motifs contribute to biological network organization. *Plos biology* 3 (2005) 1881-1892.

[2] P. Fernández, R. V. Solé: The role of computation in complex regulatory networks. *Scale-free networks and genome biology*, Landes Bioscience, 2003.

[3] R. Albert, H. G. Othmer: The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster. Journal of theoretical biology* 223 (2003) 1-18.

[4] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi: Defining and identifying communities in networks. *PNAS* 101 (2004) 2658-2663.

[5] S. Fortunato, C. Castellano: Community Structure in Graphs. *arXiv:0712.2716v1 - physics* (2007) 1-42.

[6] M. E. J. Newman, M. Girvan: Finding and evaluating community structure in networks. *Physical Review E* 69 (2004) 1-16.

[7] A. Clauset, M. E. J. Newman, Ch. Moore: Finding community structure in very large networks. *Physical Review E* 70 (2004) 1-6.

[8] S. Fortunato: Quality functions in community detection. *arXiv:07054445v1-physics* (2007) 1-10.

[9] J. Duch, A. Arenas: Community detection in complex networks using Extremal Optimization. (2005) 1-4.

[10] R. Guiméra, L. A. N. Amaral: Functional cartography of complex metabolic networks. *Nature* 433 (2005) 895-900.

[11] S. Fortunato, M. Barthélemy: Resolution limit in community detection. *PNAS* 104 (2007) 36-41.

[12] E. Ravasz, A.-L. Barabási: Hierarchical organization of complex network. *Physical Review* E 67 (2003) 1-7.

[13] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L. Barabasi: Hierarchical organization of modularity in metabolic networks. *Science* 297 (2002) 1551-1555.

# Modules in the metabolic network of *E.coli* with regulatory interactions

Jan Geryk*

Faculty of Science, Department of Philosophy and
History of Sciences,
Charles University,
Viničná 7, 128 44 Prague, Czech Republic
E-mail: geryk.cz@gmail.com
*Corresponding author

František Slanina

Institute of Physics,
Academy of Sciences of the Czech Republic,
Na Slovance 2, CZ-18221 Praha 8, Czech Republic
E-mail: slanina@fzu.cz

**Abstract:** We examine the modular structure of the metabolic network when combined with the regulatory network representing direct regulation of enzymes by small metabolites in *E.coli*. We introduce novel clustering algorithm and compare it with mainstream module detection method based on simulated annealing. Both methods identify the similar modular core. Slight but significant increase in modularity is observed after regulatory interactions addition. We also identify new functional modules in the combined network, which cannot be detected in the metabolic network only. Regulatory loops in the modules are the source of their autonomy, and allow us to hypothesize about module function.

**Keywords:** regulation; metabolic network; bipartite graph; modularity; vertex similarity measure; community structure; clustering; regulatory network; allosteric; metabolic pathway.

**Reference** to this paper should be made as follows: Geryk, J. and Slanina, F. (xxxx) 'Modules in the metabolic network of *E.coli* with regulatory interactions', *Int. J. Data Mining and Bioinformatics*, Vol. x, No. x, pp.xxx–xxx.

**Biographical notes:** J. Geryk studied medicine for 3 years, then Theoretical and evolutionary biology finished with M.S. degree. At present he is PhD student.

F. Slanina obtained PhD in theoretical physics in 1995. Since then, he has been working on various problems in biological physics.

## 1 Introduction

One of the most studied properties of real networks is their modularity. The idea of modularity is widely accepted in diverse fields (neurophysiology, computer science, evolutionary biology, etc.). In this context, a module represents a relatively autonomous system with an elementary function. It remains a challenging problem to find cellular modules solely on the basis of the network topology representing molecular interactions within the cell. We can expect auto-regulation and robustness in the functional modules. In the graph model, these properties are represented by high density of edges inside modules. The relative autonomy of modules implies a low density of edges between modules in the graph representation. In the graph theory, module identification is transformed into the question of how to find a partition of a graph with maximum density of edges inside subgraphs and minimum density of edges between subgraphs. There are a number of methods that solve the question and provide efficient algorithms for detection of modules in the networks, but open questions still remain (Barber, 2007; Ding et al., 2006; Fortunato, 2010; Guimera et al., 2008; Lancichinetti and Fortunato, 2009; Newman, 2004; Newman and Girvan, 2004; Palla et al., 2005; Rosvall and Bergstrom, 2007; Zhang et al., 2009).

One of the studied problems in the field of metabolic network research is the distribution of classical metabolic pathways among the modules. These metabolic pathways are defined on the basis of biochemical knowledge and are accessible in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database. It was demonstrated that one metabolic pathway is typically distributed among more than just one module at the same time. Within the module, there is typically more than one metabolic pathway (Guimera and Amaral, 2005; Zhao et al., 2006). These results show that it is impossible to assign the known metabolic pathways unambiguously to modules identified on the basis of the network topology. One may hypothesise that topological modules have specific functions that cannot be satisfactorily captured by classical metabolic pathway categorisation. However, this hypothesis has never been tested. Several analyses performed during the last few years are in accord with the hypothesis of evolutionary autonomy of modules. It was confirmed that the modularity measure depends on the variability of the bacteria's live environment. Bacterial strains living in variable (unpredictable) environments have more modular metabolic networks in comparison with strains that live under constant environmental conditions (Kreimer et al., 2008; Parter et al., 2007). Moreover, Alon and Kashtan (2005) predicted these findings by modelling the evolution of boolean networks. It was also shown that enzymes within a topological module have a tendency to co-occur in the set of metabolic networks of 54 taxa, implying evolutionary conservation of modules (Zhao et al., 2007). But from another point of view, modules or network partitions obtained solely on the basis of reaction co-occurrence within a phylogenetic system (Wagner, 2009) were not systematically compared with topological modules. In yet another study, Guimerá and Amaral (2005) show that non-hub nodes (metabolites) which provide interface between modules are evolutionary more conserved than the rest of the network nodes. Current knowledge can offer several indications supporting the relevancy of metabolic networks' modular structure, but functional interpretation of detected modules is still insufficient or missing.

This paper is focused on the modular structure of a bipartite representation of the *E. coli* metabolic network. A standard metabolic network representation is unipartite, i.e., the network has a single type of nodes only – typically the metabolites (Ma and Zeng, 2003a, 2003b). In the bipartite representation used in this work, two types of nodes are

present – metabolites and enzymes. Edges can be placed only between metabolites and enzymes. Bipartite representation allows integration of the regulatory interactions together with the metabolic network in a straightforward way. The effect of the addition of the regulatory interactions on the modular structure is especially analysed. Assumption of the functional autonomy of modules implies their auto-regulation. We hypothesise that regulatory interactions are concentrated in the functional modules. We compare the modular structures identified within the metabolic network with and without regulatory interactions from the quantitative and qualitative point of view.

## 2 Methods

### *2.1 Data extraction*

The dataset from a previously published paper, where the metabolic network of *E.coli* was reconstructed using the EcoCyc 9.0 database, is used (Seshasayee et al., 2009). The complete list of removed currency metabolites is available in the paper mentioned above. We constructed a bipartite graph where the nodes in one set are metabolites and those in the second set are enzyme genes, using this dataset. An edge is placed between the metabolite and the enzymatic gene if the metabolite is a substrate or product of the enzyme coded by this gene. As a second step, we reduced the complexity of the network by replacing every set of enzymatic genes with the same neighbours by one node. Enzymatic genes with the same neighbours correspond in most cases to an enzymatic complex which catalyses one reaction or to isoenzymes. In the bipartite graph representation, they form complete subgraphs which are expected to be in the centres of modules. We avoid the impact of these complete subgraphs on the detected modular structure by representing genes with the same neighbours by a single vertex. For the subsequent analysis, the largest connected component from the reconstructed bipartite graph is used.

Regulatory interactions were extracted from the EcoCyc 9.0 database, particularly from the file ‘regulation’. The metabolites in this file are assigned to reactions which they regulate. With another file (‘reaction’) from the same database, it is possible to assign enzymes (or EC numbers respectively) to reactions. The enzymes are represented as the Blattner ID of their corresponding genes in the dataset of Seshasayee et al. (2009). To assign EC nubres to Blattner IDs the ‘eco_enzyme.list’ file from the KEGG database was used.

The extracted regulatory interactions are then placed into the bipartite graph in the form of additional edges. If a non-metabolite node corresponds to the set of enzymatic genes with the same neighbours, a regulatory edge is placed between the metabolite and the non-metabolic node if the metabolite regulates at least one enzyme or enzyme subunit coded by some of these genes. By this procedure, the metabolic network combined with the regulatory network is constructed. In the following, we talk about enzymes or enzyme nodes to mean non-metabolite nodes in our bipartite representation.

### *2.2 The module identification algorithms*

The procedure is centred on the quantity measuring local density of edges (vertex similarity measure). The portions of the graph where this quantity is larger are more likely to belong to the inside of modules. Let us denote $E(u, v)$ as the number of edges within the induced

subgraph that is determined by the neighbours of nodes $u$ and $v$, $u \in U$ ($U$ is the set of metabolites), $v \in V$ ($V$ is the set of enzymes); $k_v$ is the number of node $v$ neighbours and $k_u$ is the number of node $u$ neighbours. In the following, we concentrate only on the local density of edges in the neighbourhood of two vertices connected by the edge, so the simplest definition of the local density of edges would be:

$$\tau'(u,v) = \frac{E(u,v)}{k_u k_v}. \tag{1}$$

However, this definition has a serious drawback. The local density defined by equation (1) attains relatively high values if the induced subgraph of the connected nodes $\{u, v\}$ is a small tree. The maximum value of $\tau'$ is attained in the case of a star with arbitrary size. The tree structures do not correspond to the intuitive idea of modules. Therefore, throughout this work, we use the following definition of the local density of edges in the neighbourhood of two connected vertices.

$$\tau(u,v) = \frac{E(u,v) - k_v - k_u + 1}{(k_u - 1)(k_v - 1)}. \tag{2}$$

For $k_v = 1$ or $k_u = 1$, $\tau(u, v) = 0$ is defined. From the above definition, it becomes clear that for any tree subgraph will be $\tau(u, v) = 0$. To measure the density of edges in the identified modules, we use formally the same equation as equation (2). Let us denote the number of edges inside module $s$ $E_g$, number of metabolites in $s$ $nu_s$ and number of enzymes in $s$ $nv_s$. The normalised density of module $s$ is defined as follows:

$$D_s = \frac{E_s - nu_s - nv_s + 1}{(nu_s - 1)(nv_s - 1)}. \tag{3}$$

For $nu_s = 1$ or $nv_s = 1$, $D_s = 0$ is defined as in the previous case. The mean of $D_s$ over all modules is denoted by $D$.

As in several other procedures to find modules, in the course of our algorithm we shall need a measure to quantify how good the partitioning of vertices among modules is. To this end, we use the standard modularity measure (Newman and Girvan, 2004), with a slight modification, to take into account the bipartite character of the network. The modification is explained in detail in Appendix A. Therefore, we define

$$Q^B = \frac{1}{L}\sum_{s=1}^{m}\left(l_s - \frac{d_s h_s}{L}\right) \tag{4}$$

where $L$ is the number of edges in the bipartite graph, $l_s$ is the number of edges inside the module $s$, $m$ is the number of modules in the bipartite graph, $d_s$ is the sum of metabolite degrees in the module $s$ and $h_s$ is the sum of enzyme degrees in the module $s$. The modularity measure is the difference between the number of edges inside the modules and the expected value of this quantity inside a random graph ensemble with the same degree sequence as in the original graph.

Our algorithm for finding modules in the bipartite graph is based on the idea that edges with higher $\tau$ are more likely to be placed within the modules. In some sense, it is an inverse procedure to the algorithm used in (Newman and Girvan, 2004) and its variants. The algorithm starts with a bare set of vertices and no edges. We add edges one by one, starting with the edge with the largest $\tau$ and continuing in the order of decreasing $\tau$. If more than one edge has the same value of $\tau$, all of them are placed at once. At each step, we obtain a graph composed of one or more components representing potential modules. For the partitioning we obtained, we calculate the modified modularity measure (3). In the course of the algorithm, the value of $Q^B$ evolves. For the subsequent analysis, we use the modules which emerged at such steps, in which $Q^B$ attained maximum value.

To compare our method with the mainstream module detection method, we also applied the simulated annealing module identification method to the studied metabolic network. The simulated annealing for module identification is a stochastic optimisation method where the optimised quantity is modularity measure $Q$ (Guimera and Amaral, 2005). The procedure starts with arbitrary partition ($A$) of the network. In the next step, the neighbouring partition ($B$) of the starting partition is generated, typically by moving one node from one module to another module, and the modularity measure $Q$ ($B$) for the newly generated partition is computed. If $Q$ ($A$) $\leq Q$ ($B$), the partition $B$ is accepted as a new starting partition. If $Q$ ($A$) $> Q$ ($B$), the partition $B$ is accepted with probability $p = exp$ ($-Q$ ($A$) $- Q(B)/T$). $T$ is a parameter that controls the probability of accepting partitions with decreasing modularity. During the procedure, $T$ is continuously decreasing. This allows a broader search of partition space at the beginning, continues to be more stringent and results to $p \sim 0$ in the last steps of the procedure. Modules from the last partition with the highest $Q$ are considered as relevant modules of the network.

### *2.3 Significance of maximum modularity value of the network*

The randomisation method described in Maslov et al. (2004) is used to assess the significance of the maximum modularity value. The principle of the method is to apply local randomisation repeatedly in the graph. In each local randomisation step, two edges $\{a, b\}$ and $\{c, d\}$ are randomly selected, removed from the graph and new edges: $(a, d)$ and $\{c, b\}$ are created, provided that edges $\{a, d\}$ or $\{c, b\}$ are not already present. If edges $\{a, d\}$ or $\{c, b\}$ are already present, the random selection of the two edges is repeated until it is possible to swap them.

During randomisation of the metabolic network, the graph connectivity is controlled, and only randomisations that conserve the connectivity of the graph are accepted. To obtain one randomised version of the metabolic network, 30,000 local randomisation steps were applied as described above. Sixty randomised metabolic networks were generated, and a maximum modularity value $\max\left(Q^B_{\text{rand}}\right)$ was computed for each of them by applying the clustering algorithm. The null hypothesis that the maximum modularity value max ($Q^B$) obtained with the original metabolic network is smaller than the random sample from the normal distribution with the expected value and standard deviation computed from the ensemble of 60 randomised networks is tested.

In the case of the regulatory network, the connectivity constraint during randomisation is relaxed because this network is disconnected in itself. As in the previous case, 30,000 local randomisations were applied to obtain one randomised regulatory network and 60 randomised regulatory networks were generated in total. Every randomised regulatory

network was assembled with the original metabolic network, and $\max\left(Q^{B}_{\text{rand}}\right)$, was computed by applying the clustering algorithm. With this ensemble, it is possible to test the statistical significance of the modularity increase after assembling a metabolic network with a regulatory network. As in the previous case, a *z*-test is used to test whether the maximum modularity value obtained with the original metabolic network combined with the original regulatory network is smaller than the random sample from normal distribution with mean and standard deviation computed from the randomised ensemble of the regulatory networks combined with the non-randomised metabolic network.

### *2.4 Significance of the KEGG category content in the identified modules*

A commonly used model to test the statistical significance of the functional category content in the module is hypergeometric distribution. This model does not reflect the way the modules or a network partition were obtained, assuming the nodes in the module are sampled quite randomly from the set of network nodes. A typical module detection algorithm implicitly prefers connected subgraphs as modules. In the clustering algorithm that is used in this work, the modules are defined as connected subgraphs in the network partitions obtained by successive reconstruction of the network. The effect of connectedness should be filtered out to test the significance of the metabolic category content in the modules.

For each module size obtained by the clustering algorithm, 100,000 connected subgraphs (of that size) were randomly sampled from the metabolic network with or without regulatory interactions, and the KEGG category distribution in the randomly sampled subgraphs was recorded. For each module identified by the clustering algorithm (Section 2.3) and the KEGG category dominant in the module, the empirical *p*-value was computed by counting the fraction of randomly sampled connected subgraphs of corresponding size with content larger or equal to the content of the KEGG category (that is dominant in the identified module). The KEGG categories correspond to 11 general metabolic classes or maps defined in the KEGG webpage.

## 3 Results

### *3.1 Quantitative comparison*

We applied our clustering algorithm and simulated annealing method both on the metabolic network without regulatory interactions and on the metabolic network combined with regulatory network. In the second case, there are two alternatives for controlling the algorithm flow. In the first alternative, the regulatory and reactionary edges are not distinguished, and $\tau$ is computed for every edge in the graph. In the second alternative, $\tau$ is computed only for edges that represent reactionary (and not regulatory) relationships between the metabolite and enzyme node, ensuring the reactionary connectedness of identified modules. We investigate both possibilities.

All quantities we used for comparing the mentioned methods and modular structures identified in the network with and without regulations, are summarised in Table 1. The main difference between both module identification methods is the value of modularity maximum max ($Q$). The reason is that not all network edges are partitioned in to the modules after applying our clustering method. This is due to $\tau$ constraints that determine the way modules are constructed. There is no constraint in the local density of edges in case of the simulated

annealing method. As a result, all edges are partitioned into the modules, increasing the number of positive summands in equation (3).

**Table 1** Comparison of module detection methods

| | *D* | *Max* ($Q^B$) | *Mean/std of* $max(Q^B_{rand})$ | *Mean of max. fraction of nodes in one KEGG category* | *% of modules with significant KEGG category content* |
|---|---|---|---|---|---|
| | | | *Metabolic network without regulatory interactions* | | |
| Clustering | 0.422 | 0.310 | 0.084/0.007 | 0.681 | 38 |
| Simulated annealing | 0.210 | 0.658 | – | 0.635 | 36 |
| | | | *Metabolic network with regulatory interactions* | | |
| Clustering | 0.271 | 0.381 | 0.231/0.010 | 0.673 | 37 |
| Clustering with reactionary connectedness | 0.309 | 0.341 | 0.233/0.006 | 0.685 | 36 |
| Simulated annealing | 0.181 | 0.604 | – | 0.657 | 38 |

Consider a network with a dense core subgraph and sparse rest of the network, the periphery. Even if the periphery of the network will be absolutely non-modular (for example, created by one linear chain of nodes) it may have a relatively high value of modularity for many possible partitions. We show this more precisely in supplementary materials. This idealised situation is similar to our results with the metabolic network. The clustering method identifies core modules, leaving the rest of the network non-partitioned. The simulated annealing identifies similar core modules, but also many other modules with very low edge density. These low edge modules are source of higher modularity approached by simulated annealing.

To prove this proposition quantitatively we use two partitions of the metabolic network combined with the regulatory network. The first partition is generated by simulated annealing and the second partition by the clustering method without constraint of reactionary connectedness. The results obtained by using partitions produced by the clustering algorithm with imposed reactionary connectedness and the results obtained by using the metabolic network without regulations are similar. First, we subtract all nodes not contained in the modules identified by the clustering method from the partition produced by simulated annealing. Thus, we obtain reduced partition $P_r$, which divides into the modules the same subset of network nodes as the clustering method. The normalised mutual information (Guimerà et al., 2006) between the partition $P_r$ and the partition produced by the clustering algorithm is $I_{norm}=0.814$, a value confirming relatively high similarity. The modularity of $P_r$ is 0.378, which is very close to the one obtained by the clustering algorithm (0.381), see Table 2. The modularity of the remaining modules not contained in $P_r$ is 0.142. In total, we got a modularity of 0.520, approaching the value of 0.604 produced by simulated annealing. (The difference 0.84 is due to the fact that some of the modules are broken by the division of the partition generated by simulated annealing, according to the partition generated by the clustering algorithm). We also compute a mean of normalised density ($D_{core}$) of modules defined by the partition $P_r$ and the same quantity, denoted ($D_{periphery}$) for the rest of the modules not contained in $P_r$. We obtain $D_{core}$ = 0.388 and $D_{periphery}$ = 0.0029.

For the partition generated by the clustering algorithm, we obtain $D = 0.271$ (Table 2). These results confirm a high-density modular core identified by both methods, and a sparse non-modular periphery partitioned only by the simulated annealing method. The same fact is also reflected by the mean density of modules ($D$) produced by simulated annealing, which is significantly smaller than the same quantity produced by the clustering algorithm in all considered variants (Table 2).

**Table 2** Comparison of the partition $P_r$ with the partition obtained by the clustering algoritm (Explanation in the text)

| | *Metabolic network with regulatory interactions* | | |
|---|---|---|---|
| | $D_{core}$ | $D_{periphery}$ | $Q^B$ |
| $P_r$ | 0.388 | 0.0029 | 0.378 |
| Clustering | 0.271 | – | 0.381 |

The main difference between both methods in terms of partitioning of the core is the resolution level modules are detected on. Simulated annealing tends to generate smaller modules than the clustering method. Some of the core modules detected by the clustering algorithm are divided into smaller modules by simulated annealing.

The observed increase of modularity after addition of regulatory edges is significant on the basis of a $z$-test ($p < 0.01$.). Combination of the randomised regulatory network with the non-randomised metabolic network leads to modularity decrease on average (Table 1). The effect of modularity increase is not observed after applying the simulated annealing method. If we reduce the partition produced by simulated annealing (applied on the metabolic network without regulations) to the nodes contained in the partition obtained by the clustering method and compute modularity, we obtain a value of 0.293. The same procedure using the metabolic network combined with regulatory interactions leads to the value 0.378, implying that the effect of modularity increase is localised in the modular core.

The analysis of KEGG category content shows a weaker consistency of identified modules with traditional partitioning of metabolism into the functional units. Both quantities used are similar for both applied methods as well as for the metabolic network with and without regulatory interactions (Table 1).

### *3.2 The biochemical structure of modular core*

During comparison of partitions produced by the module identification algorithms used in this work, we recognised four types of modules.

The modules of the first type exhibit a high density of reactionary edges, and are identified by both considered methods and in both cases, with and without regulatory interactions. A typical example of this type is the module corresponding to the metabolism of vitamin B6. This module corresponds to module 1.4 (Figure 1). And to module 2.3 (Figures 2 and 3). A second example is the module corresponding to the synthesis of s-adenosyl-L-homocysteine from the L-methionine. This module corresponds to module 1.2 (Figure 1). And to module 2.2.2 (Figures 2 and 3).

Modules of the second type are significantly similar in both methods and in comparison of the networks with and without regulatory interactions. These modules are typically divided into small number of dense submodules by simulated annealing. The example is the biggest identified module corresponding to the metabolism of nucleotides. This module corresponds to module 1.5 (Figure 1) and module 2.5 (Figure 2).

Modules of the first and second type cause significant similarity between the partitions of the metabolic network without regulatory interactions and partitions of the network with regulatory interactions. Modules of the third type are identified by both methods only in case of a metabolic network combined with regulatory interactions. A typical example is module 2.1 (Figures 2 and 3), corresponding to the synthesis of activated forms of glucose from the maltotetraose. A second example is module 2.4 (Figures 2 and 3) corresponding to the linear synthesis pathway of D-glucuronate from D-galacturonate.

**Figure 1** The modular core in the metabolic network without regulatory interactions identified by the clustering algorithm. Enzymes are represented as black nodes. White nodes with black borders represent metabolites

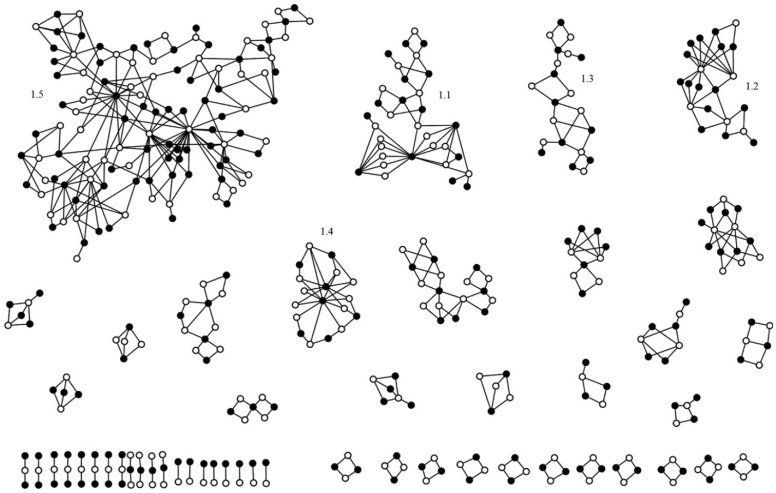



**Figure 2** The modular core in the metabolic network with regulatory interactions identified by the clustering algorithm. Three modules (2.2.1, 2.2.2 and 2.2.3) are detected as one module by the clustering algorithm without constraint of reactionary connectedness. In all other cases, these modules are detected separately. Enzymes are represented as black nodes. White nodes with black borders represent metabolites. Regulatory interactions are represented as dotted lines

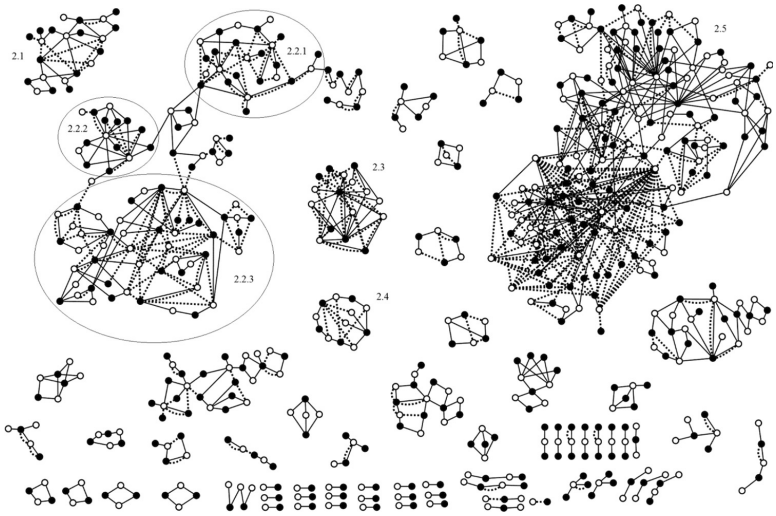
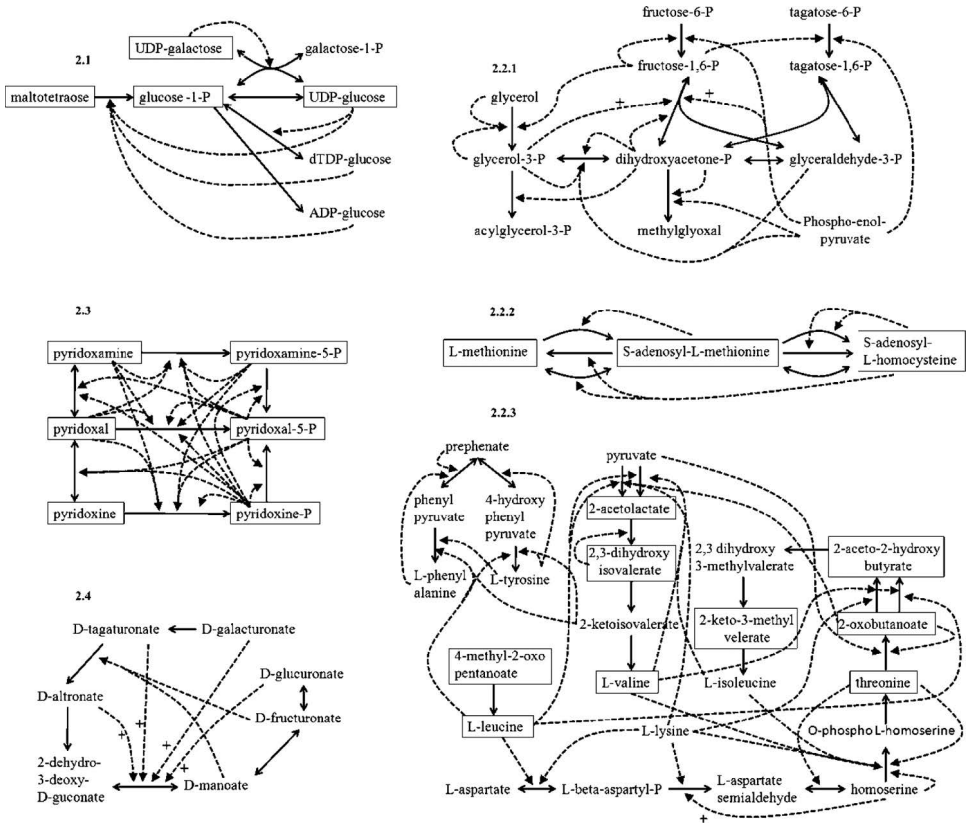


**Figure 3** Essential reactions of four modules (2.1, 2.2, 2.3 and 2.4) described in the text. Dotted lines without sign represent inhibitory regulations, dotted lines with + sign represent activations




Modules of the fourth type are recognisable only in case of metabolic networks combined with regulatory interactions. These modules are divided into a small number of dense modules by simulated annealing. Typical examples are module 2.2.1 (Figures 2 and 3) corresponding to the part of glycolysis where fructose-1, 6-bisphosphate is cleaved to dihydroxyacetone phosphate and glyceraldehyde-3-phosphate, and module 2.2.3 (Figures 2 and 3) corresponding to the metabolism of aminoacids.

Modules of the third and fourth type correspond to the sparse tree structures in the metabolic network. After addition of regulatory edges, they became denser and so, detectable by the module identification algorithms. In the metabolic network without regulatory interactions, these modules look like arbitrary or random parts of the network, and there is no reason why they would be relatively autonomous functional units. We argue that the regulatory loops contained in these modules are sources of autonomy and functional interpretability.

Let us concentrate on module 2.2.1. The module contains important glycolytic reactions, especially fructose-1,6-bisphosphate cleavage to dihydroxyacetone phosphate and glyceraldehyde-3-phosphate. There is also cleavage reaction of tagatose-1,6-bisphosphate to the same products and reaction converting dihydroxyacetone phosphate to glycerol phosphate. If we consider regulatory interactions, this system becomes closed due to its many

regulatory loops, and thus, physiologicaly interpretable. Concentration of dihydroxyaceton phosphate and glyceraldehyd 3-phosphate is elevated during glycolysis activation, and the regulatory effect they provide within the module is pronounced. Both molecules inhibit their alternative utilisation in pathways other than glycolysis. In the same time, fructose 1,6 phosphate inhibits generation of dihydroxyaceton phosphate and glyceraldehyd 3-phosphate from the alternative sources. Phosphoenol pyruvate is also elevated when flux through glycolysis increases. Phospho-enol pyruvate inhibits alternative utilisation of dihydroxyaceton phosphate and glyceraldehyd 3-phosphate too. We propose that the function of this module is to fix and facilitate metabolic flux through glycolysis when glycolysis is elevated, for example, in an environment with increased concentration of glucose. In principle, the investigated module represents sophisticated positive feedback regulation of glycolysis.

## 4 Summary and discussion

Our results suggest that the metabolic network is composed of modular core and non-modular sparse periphery. A similar result was reported by Zhao et al. (2006). In contrast to the simulated annealing method our clustering algorithm is capable of selectively identifying the modular core, leaving the rest of the network non-partitioned. We observe a statistically significant increase of modularity of the core after addition of regulatory edges. However, the modularity difference is very small (0.07). It seems important to perform similar analyses on different graph representations to make a decision about the effect of regulatory interactions addition on network modularity. It is a well known fact that modularity of the metabolic network depends strongly on graph representation. Our bipartite representation leads to a sparser graph than the conventionally used unipartite representation. It results in smaller modularity value approached by simulated annealing.

The metabolism of nucleotides is dominant and a significantly abundant category in the biggest identified modules. This is true for both methods and for the metabolic network with and without regulatory interactions. Nucleotides have two crucial functions in the living cell. They are donors of energy in the majority of cellular processes, and also precursors of DNA and RNA synthesis. Our analysis shows that metabolism of nucleotides is the most integrated part of the metabolic network of both reactionary and regulatory perspectives. This result is in accordance with their crucial importance for the cell.

In the metabolic network without regulations, it is difficult to interpret the identified modules as autonomous functional units. The situation will change after regulatory edges addition. In Section 3.2, we demonstrated that due to regulatory loops within the module, it is possible to generate a hypothesis about module function. The hypothesis about a positive feedback system within the glycolysis pathway formulated in Section 3.2 is testable by dynamic modelling, but this ambition is out of the scope of our paper. It was recently shown that it is possible to explain specific experimental behaviour of *E.coli* on the basis of a relatively simple metabolic subsystem with regulatory feedbacks. In addition, the clearly defined function of the studied subsystem emerged as a consequence of considering regulatory interactions (Kotte et al., 2010).

The relative autonomy of the modules identified in the metabolic network with regulatory interactions is not only by virtue of the sparse connection to the rest of the network implied

by the definition of the module, but also by virtue of auto-regulations included in the modules. This conclusion follows from the obvious idea that the system can be autonomous only if it manifests some kind of self-control. The graph representation captures this notion very coarsely, but it is important to investigate what we can say about it from the graph perspective.

## Acknowledgements

J. Geryk thanks all the members of the Department of Philosophy and History of Sciences for inspirational discussions and suggestions. This work was carried out within the project AV0Z10100520 of the Academy of Sciences of Czech Republic and was supported by the MSMT of the Czech Republic, Grant No. OC09078 and by the Czech Ministry of Education MSM 0021620845.

## References

Alon, U. and Kashtan, N. (2005) 'Spontaneous evolution of modularity and network motifs', *PNAS*, Vol. 102, pp.13773–13778.

Barber, M.J. (2007) 'Modularity and community detection in bipartite networks', *Physical Review*, Vol. E 76: 066102, p.9.

Ding, Ch., He, X., Xiong, H., Peng, H. and Holbrook, S.R. (2006) 'Transitive closure and metric inequality of weighted graphs: detecting protein interaction modules using cliques', *Int. J. Data Mining and Bioinformatics*, Vol. 1, pp.162–177.

Fortunato, S. (2010) 'Community detection in graphs', *Physics Reports*, Vol. 486, pp.75–174.

Guimera, R. and Amaral, L.A.N. (2005) 'Functional cartography of complex metabolic network', *Nature*, Vol. 433, pp.895–900.

Guimera, R., Sales-Pardo, M. and Amaral, L.A.N. (2008) 'Module identification in bipartite and directed networks', *Physical Review E*, Vol. 76: 036102, p.8

Kotte, O., Zaugg, J.B. and Heinemann, M. (2010) 'Bacterial adaptation through distributed sensing of metabolic fluxes', *Molecular Systems Biology*, Vol. 6, pp.1–9.

Kreimer, A., Borenstein, E., Gophna, U. and Ruppin, E. (2008) 'The evolution of modularity in bacterial metabolic networks', *PNAS*, Vol. 105, pp.6976–6981.

Lancichinetti, A. and Fortunato, S. (2009) 'Community detection algorithms: a comparative analysis', *Physical Review E*, Vol. 80: 056117, p.11.

Ma, H. and Zeng, A.P. (2003a) 'Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms', *Bioinformatics*, Vol. 19, pp.270–277.

Ma, H.W. and Zeng, A.P. (2003b) 'The connectivity structure, giant strong component and centrality of metabolic networks', *Bioinformatics*, Vol. 19, pp.1423–1430.

Maslov, S., Sneppen, K. and Zaliznyak, A. (2004) 'Detection of topological patterns in complex networks', *Physica A*, Vol. 333, pp.529–540.

Newman, M.E.J. (2004) 'Fast algorithm for detecting community structure in networks', *Physical Review E*, Vol. 69: 066133, p.5.

Newman, M.E.J. and Girvan, M. (2004) 'Finding and evaluating community structure in networks', *Physical Review E*, Vol. 69: 026113, p.19.

Palla, G., Derényi, I., Farkas, I. and Vicsek, T. (2005) 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature*, Vol. 435, pp.814–818.

Parter, M., Kashtan, N. and Alon, U. (2007) 'Environmental variability and modularity of bacterial networks', *BMC Evolutionary Biology*, Vol. 7, p.169.

Rosvall, M. and Bergstrom, C.T. (2007) 'An information-theoretic framework for resolving community structure in complex networks', *PNAS*, Vol. 104, pp.7327–7331.

Seshasayee, A.S.N., Gillian, M., Fraser, M. and Babu, M.M. (2009) 'Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*', *Genome Research*, Vol. 19, pp.79–91.

Wagner, A. (2009) 'Evolutionary constraints permeate large metabolic networks', *BMC Evolutionary Biology*, Vol. 9, p.231.

Zhang, S., Liu, H.-W., Ning, X.-M. and Zhang, X.-S. (2009) 'A hybrid graph-theoretic method for mining overlapping functional modules in large sparse protein interaction networks', *Int. J. Data Mining and Bioinformatics*, Vol. 3, pp.68–84.

Zhao, J., Yu, H., Luo, J.H., Cao, Z.W. and Li, X.Y. (2006) 'Hierarchical modularity of nested bow-ties in metabolic networks', *BMC Bioinformatics*, Vol. 7, p.386.

Zhao, J., Ding, G.H., Tao, L., Yu, H., Yu, H.Y., Luo, J.H. and Cao, Z.W. (2007) 'Modular co-evolution of metabolic networks', *BMC Bioinformatics*, Vol. 8, p.311.

## Supplementary file

### *Modularity measure for bipartite graph*

The general formula for the modularity measure is:

$$Q = \frac{1}{L}\sum_{s=1}^{m}[l_s - E(l_s)] \tag{s.1}$$

where $L$ is the number of edges in the graph, $m$ is the number of modules, $l_s$ is the number of edges inside the module $s$ and $E(l_s)$ is the expected number of edges between nodes of the module $s$ in the random graph ensemble. We obtain $E(l_s)$ as the sum of probabilities that an edge exists between nodes in the module $s$. In the case of a bipartite graph:

$$E(l_s) = \sum_{u \in U_s, v \in V_s} p(u,v) \tag{s.2}$$

where $U_s$ is the set of all metabolites within the module $s$ and $V_s$ is the set of all enzymes within the module $s$. The probability $p(u, v)$ can be interpreted as a number of graphs in the random graph ensemble that contain an edge $\{u, v\}$, divided by the number of all graphs in this ensemble. $p(u, v)$ in the random bipartite graph ensemble with prescribed degree sequence is estimated in the following text. Virtually, we can construct the bipartite graphs from this ensemble by connecting the 'stubs' (or half of edges) arising from the metabolites and enzymatic genes. There are $2L$ stubs in the graph, $L$ arising from metabolites and $L$ from enzymes. There are $L!$ possibilities for how to construct a bipartite graph. If vertices $u$ and $v$ are connected by an edge, the number of possibilities for how to construct the graph is reduced. There are $k_u k_v$ possible realisations of an edge $\{u, v\}$ and after one of these realisations is chosen there is $(L - 1)$ number of remaining edges to be placed. The number of possibilities to construct a bipartite graph with an imposed constraint that between vertices

$u$ and $v$ must be an edge is estimated as $k_u k_v (L-1)!$. The probability $p(u, v)$ is estimated as follows:

$$p(u,v) \cong \frac{k_u k_v (L-1)!}{L!} = \frac{k_u k_v}{L}. \tag{s.3}$$

The same result was obtained in Barber (2007). The expected number of edges inside subgraph $s$ is then:

$$E(l_s) \cong \sum_{u \in U_s, v \in V_s} \frac{k_u k_v}{L} = \frac{d_s h_s}{L} \tag{s.4}$$

where $d_s = \sum_{u \in U_s} k_u$ and $h_s = \sum_{v \in V_s} k_v$. With this estimate, it is possible to define modularity measure for the bipartite graph:

$$Q^B = \frac{1}{L} \sum_{s=1}^{m} \left( l_s - \frac{d_s h_s}{L} \right). \tag{s.5}$$

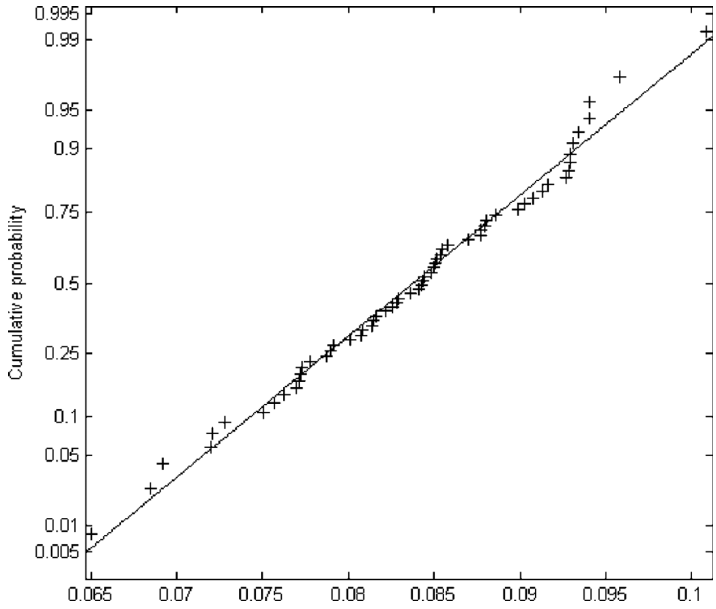
*Modularity of sparse periphery*

Consider a partition of the bipartite network on two parts. The first part is arbitrary. The second part is one linear chain of nodes connected by two ends with the first part of the network. Let us denote the number of nodes in the second part $N_p$ and the number of network edges $L$. For simplicity, we divide the second part to the $N_p/n$ modules with the same sizes $n$. From equation (s.5), it directly follows that the modularity of the second part is:

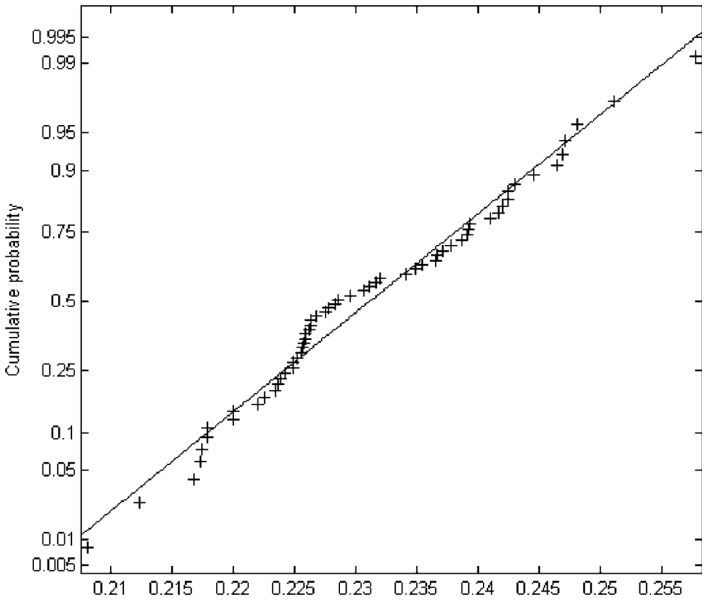
$$Q_p = \frac{N_p}{L} \left( 1 - \frac{1}{n} - \frac{n}{L} \right). \tag{s.6}$$

If we have fixed values of $N_p$ and $L$ we can still manipulate the value $Q_p$ by choosing the size of modules $n$. We used values $L = 2213$ and $N_p = 516$ corresponding to the application of our clustering algorithm on the metabolic network with regulatory interactions ($N_p$ is the number of nodes not partitioned into the modules by the clustering algorithm). In this case, if we choose $n = 20$, than $Q_p = 0.22$. This clearly demonstrates that non-modular structures may contribute significantly to the total modularity of the network.

**Supplementary Figure 1** The ordered values of $\max\left(Q^{B}_{\text{rand}}\right)$ (*x*-axis) from 60 randomised metabolic networks are plotted against their observed cumulative frequency. The *y*-axis is scaled for normal distribution




**Supplementary Figure 2** The ordered values of $\max\left(Q^{B}_{\text{rand}}\right)$ (*x*-axis) computed from 60 randomised regulatory networks combined with the original metabolic network are plotted against their observed cumulative frequency. The *y*-axis is scaled for normal distribution



Prohlášení spoluautora práce „Modules in the metabolic network of E.coli with regulatory interactions“ o podílu doktoranda na této práci:

J. Geryk je autorem základní ideje této práce včetně formálního zpracování. Dále implementoval finální verze všech analýz a z převážné části napsal zmíněnou publikaci.

F. Slanina implementoval první verzi klastrovacího algoritmu a podílel se na formulaci technických částí publikace.

Souhlasím, aby se práce „Modules in the metabolic network of E.coli with regulatory interactions“, ve které jsem spoluautorem, stala součástí disertační práce J. Geryka.

František Slanina

# Regulatory Canalized Paths in Metabolic Network of *E. coli*

Jan Geryk
Department of Philosophy and History of Sciences,
Faculty of Science, Charles University.
Viničná 7, 12844, Prague, Czech Republic.
E-mail: geryk.cz@gmail.com

**Abstract**

The metabolic regulatory network representing direct modulations of enzymes activities by metabolites has been studied very sparsely up to present days. I studied the tendency of regulatory network to canalize and auto-canalize metabolic paths in the present work. Canalization measure to what extend can regulatory network inhibits reactions branching out of the metabolic path and at the same time leave reactions within the metabolic path unaffected. Auto-canalization express to what extend is above mentioned canalization mediated by the metabolites produced within the metabolic path. I show that the regulatory network of *Escherichia coli* exhibits significantly higher auto-canalization of metabolic paths in comparison with two random ensembles with different degree of constraints. Moreover, second property which is responsible for canalization of metabolic paths was found as significantly overrepresented within real regulatory network. The reactions branching out of the metabolic paths are often inhibited by the metabolites produced by the secondary paths starting in these out-branching reactions and these regulatory metabolites does not inhibit any reactions within metabolic path. The regulatory auto-canalization of minimal hyperpaths leading to glucose is investigated in detail. It was found that minimal hyperpaths with significantly higher auto-canalization corresponds to the experimentally confirmed states of metabolic network of *E. coli* cultivated in the specific growth conditions.

**Author summary**

Regulation of metabolic network can be partitioned into transcriptional regulation which probably plays the most important role and regulation mediated by the direct modulation of enzymes activity by the metabolites. Metabolic regulation of the latter type is expected to be responsible for short time regulation of metabolic network and has not yet been analyzed at a genome-scale level except one publication. I study the structural properties of second type regulatory network in the present work. The studied properties quantitatively express to what extent can regulatory network inhibits reactions branching out of the metabolic paths and at the same time leave reactions within the metabolic paths unaffected. I find that the metabolic paths within *E. coli* exhibit significantly increased tendency to positively regulate themselves, indirectly by the inhibition of out-branching reactions. Secondly, the reactions branching out of the metabolic paths are often regulated by the metabolites produced by the secondary paths starting in this out-branching reaction and these regulatory metabolites does not inhibit any reactions within metabolic path. The discovered structural properties are probably manifestations of evolutionary adaptations in the architecture of metabolic-regulatory network.

**Introduction**

The connectivity structure of metabolic networks has been intensively studied for a decade. It was shown that metabolic networks exhibit power-law degree distribution and small-world property like many others real networks [1,2]. Important differences between currency metabolites typically sitting on the tail of the power-law distribution and others metabolites were also recognized [3-5]. Metabolic networks exhibit rather weak degree of modularity i.e. tendency to form sub-networks with increased density of edges [6-9]. The functional interpretation of the modules is still unresolved. Nevertheless some analyses showed that modularity of the bacterial metabolic networks increases with variability of living environment [10,11]. These findings are in accord with evolutionary simulation experiments [12,13].

There is a relatively large amount of work about the structure of transcriptional regulation of metabolic network [14-19]. It was repeatedly shown that metabolic transcriptional regulatory network of known model microorganisms is organized in the hierarchical manner, co-regulated sets of reactions typically form linear chains and metabolites often regulate activity of metabolic genes forming local regulatory loops. [16-19]. Metabolic junctions are often regulated differentially reflecting the fact that junctions are expected to be decision points [18].

In contrast to transcriptional regulation, the architecture of regulatory networks representing direct interactions of small metabolites with the enzymes is investigated insufficiently. Up to date and my knowledge, there exists only one study that systematically analyzed topological aspects of the regulatory network of metabolism mediated by the direct modulation of enzyme activity by small metabolites [20]. In this work, four regulatory networks of *E. coli*, *S. cerevisiae*, *P. falciparum* and *H. sapiens* were investigated. Authors applied some conventional measures to the dataset and found that the number of regulated enzymes by single metabolite follow power-law for all studied organisms. They also observed inverse dependency of clustering coefficient on the metabolite degree in the regulatory network, the property characteristic for modular hierarchy. Now, this finding is not very informative in the light of study of Hao et. al. [21]. Authors also found weak correlation between regulation and reaction connectivity of metabolites and concluded that the correlation is mainly determined by the small number of highly connected metabolites. The last finding of the study is weak correlation between chemical similarity and regulation similarity within pairs of regulation metabolites [20].

In the present work I focus on the topological properties of the combined metabolic and regulatory network of the bacterium *E. coli*. The regulatory network used in this study represents direct modulations of enzyme activity by small metabolites. I argue that the structure of the regulatory network is much more evolutionary flexible than metabolic network. Based on this assumption we can expect that evolutionary adaptations are in some way imprinted in the structure of the regulatory network. I want to generalize one specific observation from our previous publication, dealing with modularity of combined metabolic and regulatory network of *E. coli* [22], in the present work. In [22] we recognize one module containing linear chain of glycolytic reactions with interesting kind of positive self-regulation. Reactions branching out of this chain are inhibited by one metabolite contained in this chain, one reaction from the chain is positively regulated by its product forming positive feedback

loop. There are also two regulations mediated by the metabolite from this chain that inhibits reactions supplying the chain from other sources. I want to test the hypothesis that similar, positively auto-regulated structures are abundant in metabolic-regulatory network. I will use the shorter term auto-canalization instead of positive self-regulation in the following text. In the present work I formalize the concept of canalization of metabolic path and show that the real regulatory network exhibit significantly higher degree of regulatory canalization and auto-canalization of the metabolic paths and hyperpaths than its randomized counterpart.

**Results**

For subsequent analysis I used the oriented hypergraph representation of metabolic model of *E. coli* metabolism iAF1260 [23]. The hypergraph representation of metabolic network can be viewed as a list of reactions, together with information which metabolites are substrates and which are products of individual reactions. In the case of an oriented bipartite graph representation, this information is missing for reversible reactions. This can lead to a misidentified path, containing two substrates of a single reversible reaction where one substrate acts as a substrate and the other as a product within the path. Regulatory interactions between metabolites and reactions were extracted from Ecocyc database [24] and represented as a bipartite graph where one set of nodes are metabolites and a second set are reactions. An edge is placed between a reaction and a metabolite if this metabolite inhibits the reaction.

***Regulatory canalization: definition***

I define regulatory canalization for metabolic path and hyperpath in the present work. The term metabolic path used in this work refers to simple sequence of alternating metabolites and reactions, where every reaction within the path consumes the preceding metabolite and produces a following metabolite in the sequence (Fig. 1a). A metabolic hyperpath leading from the source subset to the target subset of metabolites is a sub-hypergraph which is sufficient to produce a target set from the source set (Fig.1b). I consider only minimal hyperpaths in this work. A minimal hyperpath cannot be further reduced by removing any reaction so that the definition of a hyperpath is preserved. Note that a metabolic path leading from a source metabolite to a target metabolite is sufficient to produce a target from the source only if every reaction in the path has only one substrate. The precise definition of the metabolic path and hyperpath used in this work can be found in the methods section.

In the following, I only define the quantities for the metabolic paths because definitions for minimal hyperpaths are analogical. If we replace the term metabolic path with minimal hyperpath, and the associated symbol of metabolic path ($p$) with the symbol for minimal hyperpath ($P$), we obtain definitions for minimal hyperpaths.

I define regulatory canalization ($C$) as a fraction of irreversible reactions branching out of the metabolic path with the property that at least one metabolite exists that inhibits an out-branching reaction and does not inhibit any reaction within the metabolic path. The out-branching reactions of the metabolic path ($p$) are defined as all irreversible reactions consuming at least one metabolite produced and/or consumed within $p$, and these reactions are not members of $p$ (fig. 1).

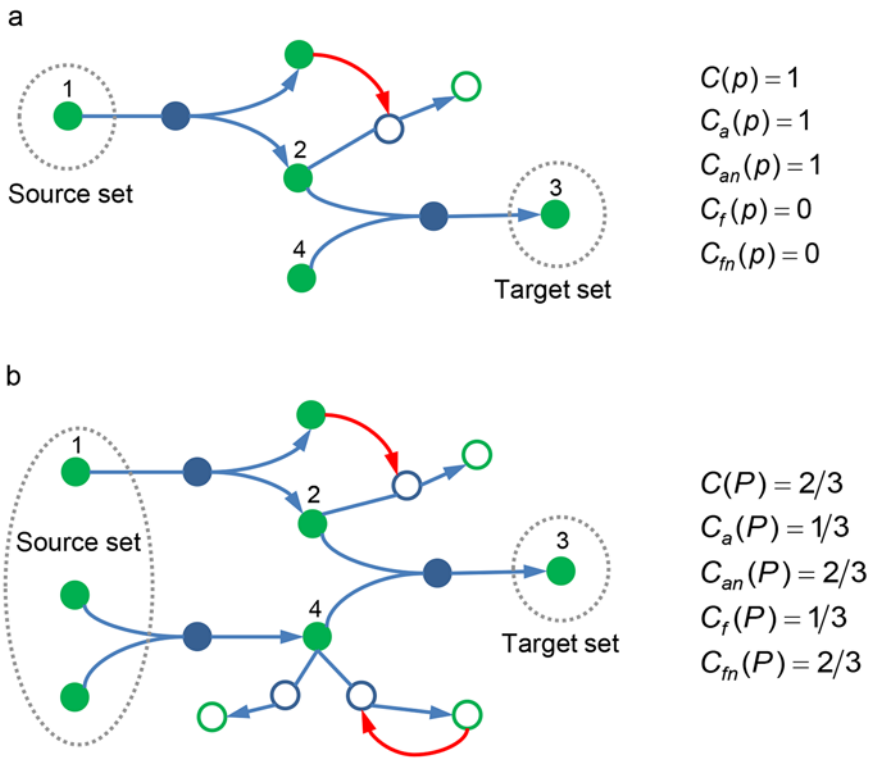



**Figure 1. Metabolic path and minimal hyperpath.**
a) Metabolic path formed by two hyperedges marked by blue filled circles and three metabolites (1,2,3). There is a single reaction branching-out of the metabolic path (empty circle with blue border) which is inhibited by metabolite produced by first hyperedge of the metabolic path. Note that after extending the source set of metabolite 4, we get a hyperpath because now is possible to produce target from source set. Five quantities defined within this work were computed for this path and reproduced on the right.
b) Hyperpath containing metabolic path from the figure 1a. By removing any of three hyperedges contained within this hyperpath the possibility to produce target set from source set is lost ie. hyperpath is minimal. The hyperpath has three out-branching reactions, one of them is inhibited by the metabolite produced within the hyperpath (upper) and one is inhibited by product of out-branching reaction (lower).

I denote $E_{out}(p)$ as the set of reactions branching out of the *p*. A rigorous definition of the out-branching reactions is contained in the methods section. I denote $E_{out}^{C}(p)$ subset of reactions belonging to $E_{out}(p)$ such that for each reaction $r \in E_{out}^{C}(p)$ at least one regulatory metabolite exists which inhibits $r$ and does not inhibit any reaction within the *p*. The most general definition of regulatory canalization of metabolic path is:

$$C(p) = \frac{\left|E_{out}^{C}(p)\right|}{\left|E_{out}(p)\right|} \qquad (1)$$

Where $|x|$ denotes number of elements in *x*. *C* expresses the potential of the metabolic regulatory system to inhibit reactions consuming metabolites from the metabolic path and in the same time leaves the metabolic path unaffected. I consider only irreversible reactions as

out-branching reactions because reversible reactions can potentially supply substrates to the metabolic path, depending on conditions.

In order to determine to what extent the regulatory canalization is mediated by the metabolites contained within the metabolic path, I denote $E_{out}^{Ca}(p)$ as subset of reactions belonging to $E_{out}^{C}(p)$ such that for each reaction $r \in E_{out}^{Ca}(p)$ at least one regulatory metabolite exists which inhibits $r$ and does not inhibit any reaction within the *p* and the regulatory metabolite is produced by the reactions contained in the *p*. I define regulatory auto-canalization of the metabolic path as:

$$C_a(p) = \frac{\left|E_{out}^{Ca}(p)\right|}{\left|E_{out}(p)\right|} \tag{2}$$

I also define normalized version of auto-canalization:

$$C_{an}(p) = \frac{\left|E_{out}^{Ca}(p)\right|}{\left|E_{out}^{C}(p)\right|} \tag{3}$$

The measure of regulatory auto-canalization is constructed to test the hypothesis that the metabolic path positively regulates itself, indirectly by inhibition of out-branching reactions. The second hypothesis is that the out-branching reactions are the beginnings of specific pathways and products of these pathways inhibit out-branching reactions forming negative feedback loops. I denote $E_{out}^{Cf}(p)$ subset of reactions belonging to $E_{out}^{C}(p)$ such that for each reaction $r \in E_{out}^{Cf}(p)$ at least one regulatory metabolite exists which inhibits $r$ and does not inhibit any reaction within the *p* and the metabolic path leading from the out-branching reaction to that regulatory metabolite exists and have no out-branching reactions. In order to test the second hypothesis I define measure:

$$C_f(p) = \frac{\left|E_{out}^{Cf}(p)\right|}{\left|E_{out}(p)\right|} \tag{4}$$

I also define the normalized version of this measure:

$$C_{fn}(p) = \frac{\left|E_{out}^{Cf}(p)\right|}{\left|E_{out}^{C}(p)\right|} \tag{5}$$

***Randomization of the regulatory network***

In order to assess the significance of the introduced quantities, the regulatory network must be properly randomized. The structure of the regulatory network is not constrained by the physico-chemical laws as are metabolic networks. For example, the substrates and products of a reaction must obey mass balance conditions, i.e., the sum of atoms of a particular type in the substrate set must be the same as in the product set. Another important constraint is the thermodynamic feasibility that can exclude many combinatorial possible reactions. A reasonable approach is to compare properties of real networks with randomized ensembles that express what is physically possible. In such a way it is possible to distinguish what properties of networks are evolutionary tuned and what are forced by the physical laws [25,26].

There are practically no limits in which a molecule can control the activity of an enzyme. From the view-point of evolution, it is possible to connect an arbitrary reaction with an arbitrary metabolite by a regulatory interaction via the flexibility of a catalyzing enzyme. This property of metabolic networks is called gratuity and was first recognized by Jacques Monod [27]. Obviously there must be a constraint in the number of metabolites that can regulate a single enzyme. Despite this, I observe no correlation of molecular weight of the enzyme with its number of regulators. This indicates that molecular weight is no limiting factor for the range of regulatory degree observed in the real regulatory network.

I introduce here three randomized ensembles of the regulatory network. 1000 networks were generated from each ensemble in order to assess significance of introduced quantities. The first one is the ensemble of all regulatory networks with the same number of regulatory metabolites, regulated reactions and regulatory interactions as the real regulatory network. In this ensemble, regulatory metabolites and regulated reactions are selected randomly for each random network. Edges are placed randomly between selected reactions and metabolites with the only constraint being that the number of metabolites which regulate single enzyme cannot exceed the maximum number observed in the real regulatory network (which is $k_{\max}(r) = 15$).

The second ensemble is the same as the first, except for one additional constraint: networks within the second ensemble are allowed to have a value greater than or equal to the specific quantity ($q$) exhibited by the real network. $q$ is the number of regulatory interactions for which a metabolic path, excluding out-branching reactions, exists leading from the regulated reaction to the metabolite regulating this reaction.

The third ensemble is much more conservative. The sets of regulatory metabolites and regulated reactions are the same as in the real regulatory network. In addition, regulatory degree distribution is fixed for metabolites and for reactions. Regulatory edges are randomly distributed within these constraints.

***Regulatory canalization of E. coli metabolic network***

As a first analysis, I constructed all metabolic paths of length 6 and computed five quantities defined in the first section for every constructed path. The cumulative distribution functions of all measured quantities are plotted on Figure 2. The upper six plots in Figure 2 (a-f) correspond to the first and second random ensemble and lower six plots (g-l) correspond to the third random ensemble.

The cumulative distribution function of the regulatory canalization ($C$) is significantly shifted to the right from the 95% range computed from the first randomized ensemble (Fig. 2a). This indicates that the real regulatory network exhibits significantly higher canalization of metabolic paths than randomized networks from the first ensemble. The same is not true if we compare cumulative distribution of $C$ with the third randomized example, indicating that the significance of $C$ diminishes after maintaining the regulatory nodes (Fig. 2g).

The regulatory auto-canalization $C_a$ and normalized auto-canalization $C_{an}$ exhibit a significant shift to higher values in comparison with both the first (Fig. 2b,e) and third randomized ensembles (Fig. 2h,k).

The quantity $C_f$ and its normalized version exhibit significant shifts to higher values in comparison with the first and third ensembles, representing the strongest signal measured

(Fig. 2c,f,i,l). The second randomized ensemble was constructed in order to determine whether the observed significance of the quantity $C_f$ is simply a consequence of a significantly larger value of *q*, observed in the real regulatory network (Fig. 2d,j). On the figure 2c we can see that after fixing the *q*, $C_f$ still exhibits significantly higher values than random networks from the second ensemble. This indicates that the significance of higher $C_f$ values is not simply a consequence of a higher *q* value. In contrast to $C_f$ the significance of $C_{fn}$ diminishes in comparison with the second randomized ensemble (Fig. 2f).

***Regulatory canalized hyperpaths to glucose***

A second more detailed analysis focused on the hyperpaths leading to glucose. Analogically to the previous case I constructed all minimal hyperpaths leading to glucose up to length 6 and computed five quantities defined in the first section for every constructed hyperpath. The cumulative distribution functions of all measured quantities are plotted on figure 3, in comparison with 95% range computed from the first randomized ensemble (Fig. 3a-e), second randomized ensemble (Fig. 3c,e) and from the third randomized ensemble (Fig. 3f-j).

The regulatory canalization (*C*) behaves similarly as in the case of the metabolic paths (Fig 3a,f). The quantity $C_f$ exhibits significantly larger values than the networks from the first ensemble. In comparison with the second and third ensembles, only the end of the distribution of $C_f$ exhibits significant deviation (Fig. 3c,h). Deviation of cumulative functions of $C_a$ and $C_{an}$ from the first and third random ensembles is still more pronounced than in the case of metabolic paths (Fig. 3b,d,g,i).

The course of cumulative function $C_a$ and $C_{an}$ exhibit a remarkable plateau phase. This plateau phase indicates that the distribution of $C_a$ and $C_{an}$ has approximately bimodal character. I further investigated the source of this bimodality and found that the second modus in distribution $C_a$ and $C_{an}$ is mostly determined by the hyperpaths leading from the glycerol to glucose, containing glycerol 3-phosphate. These hyperpaths represents aerobic assimilation of glycerol starting with the phosphorylation of the glycerol by glycerol kinase, and then oxidizing of glycerol-3-phosphate (G3P) to dihydroxyacetone phosphate by glycerol-3-phosphate dehydrogenase. G3P enters to the reaction with CDP-diacylglycerol which resultes in the formation of phosphatidylglycerol-phosphate. G3P can react with CDP-diacylglycerols with different numbers of carbon atoms in the acyl group creating many out-branching reactions which are all inhibited by the G3P. Relatively high $C_a$ values of hyperpaths representing aerobic assimilation of glycerol are caused due to the high fraction of neighboring reactions of G3P inhibited by G3P. If we plot the metabolite out-connectivity against the fraction of reactions consuming the metabolite inhibited by single metabolite, we can see that the points obtained fall under the decreasing power-law function. The only exception is G3P, which exhibits an anomaly large fraction of its neighboring reactions inhibited by itself in comparison with metabolites with same out-connectivity. This fact can be viewed as an artifact of the regulatory network representation. For this reason, I removed the interactions between G3P and the phosphatidyl-glycerol synthase reactions from the network, and repeated the auto-canalization analysis.

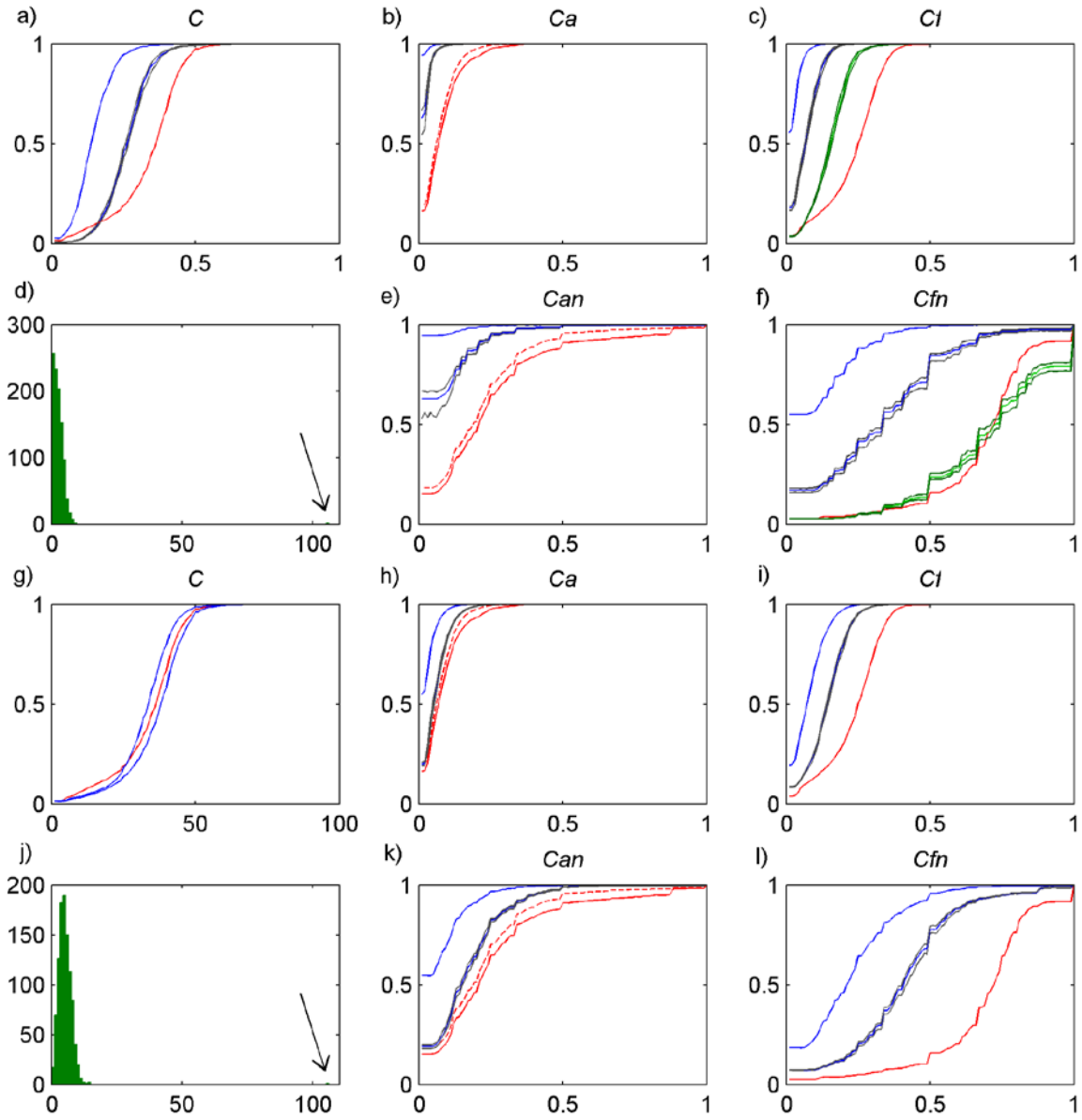



**Figure 2. Canalization of metabolic paths within real network and randomized ensembles.**
Red curves represent cumulative distribution functions of quantities listed above plots computed from real network. Red dashed curves represents cumulative distribution functions of $C_a$ and $C_{an}$ computed from real network without regulatory interactions between G3P and phosphatidyl-glycerol synthase reactions. Blue curves represent 5 and 95 percentile of values of cumulative distribution functions computed from first ensemble (plots a-f) and third ensemble (plots g-l). Green curves (in plots c and f) correspond to 5 percentile of cumulative function of $C_f$ and $C_{fn}$ computed from second ensemble. Confidence bounds are computed only for 5 percentile of cumulative functions and are represented as gray or dark green curves. On the plots d and j are visualized distributions of quantity $q$ within the first and third random ensembles and the arrow denote value of $q$ computed from real network.

Removing these interactions has no impact on the quantities $C, C_f$ and $C_{fn}$, so I will discuss only $C_a$ and $C_{an}$. The cumulative functions shift slightly to the left but still preserve the significance of $C_a$ and $C_{an}$ in the case of metabolic paths (Fig.2b,e,h,k). Removing these interactions has a strong impact on the $C_a$ and $C_{an}$ values of the minimal hyperpaths leading to glucose. The bimodality of the distribution of $C_a$ and $C_{an}$ is completely lost but preserving

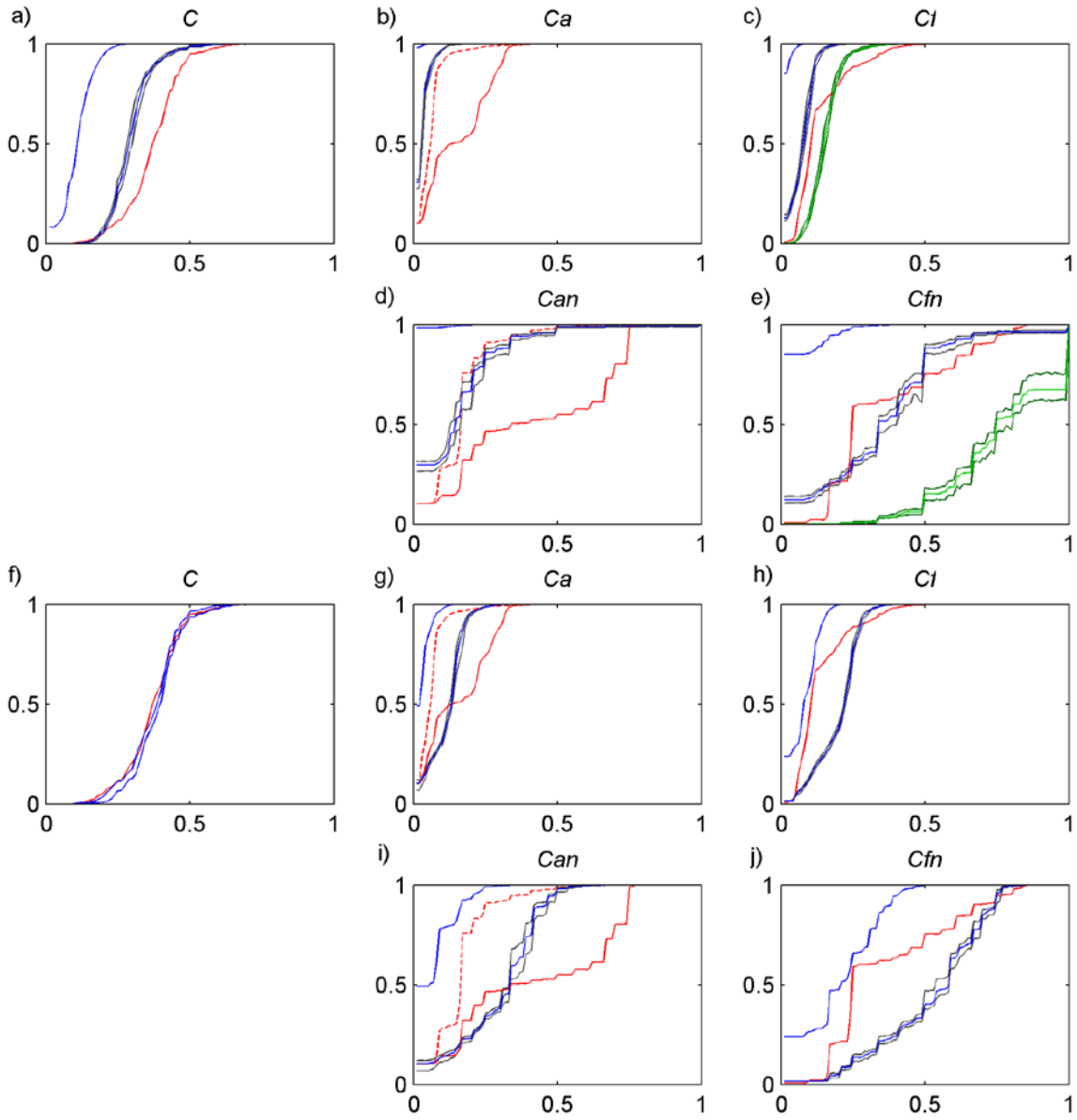



**Figure 3. Canalization of minimal hyperpaths leading to glucose within real network and randomized ensembles.**

Upper five plots (a-e) correspond to comparison of real network with first and second randomized ensemble, lower five plots (f-j) correspond to the comparison of real network with third randomized ensemble. Curves have same meanings as in the figure 2.

significance of $C_a$ and $C_{an}$ values in comparison with first randomized ensemble (Fig.b,d). After visual inspection it can be seen that the significance of $C_a$ and $C_{an}$ values diminish in comparison with third randomized ensemble (Fig. 3g,i). But we can see that high values ($C_a$ > 0.25) are still significantly abundant with respect to the third ensemble if we plot confidence bounds within the percentile-percentile plot. See details in methods.

Figure 4 visualizes the metabolic sub-network containing minimal hyperpaths leading to glucose with significantly high values of $C_a$ > 0.25 together with all regulatory interactions between reactions and metabolites within the sub-network.

The largest values of the regulatory auto-canalization ($C_a$) correspond to the hyperpaths representing anaerobic (fermentative) assimilation of glycerol. Glycerol is oxidized by glycerol dehydrogenase to dihydroxyacetone and then phosphorylated by dihydroxyaceton kinase to dihydroxyacetone phosphate in these hyperpaths. Dihydroxyacetone phosphate can be then used as a gluconeogenetic substrate (Fig. 4). It was recently confirmed by experimental methods that *E.coli* can anaerobically fermentate glycerol and the glycerol dehydrogenase and dihydroxyacetone kinase are essential enzymes for the glycerol assimilation. In this process methylglyoxal is reduced to 1,2-propanediol compensating production of reductive elements by glycerol dehydrogenase [28].

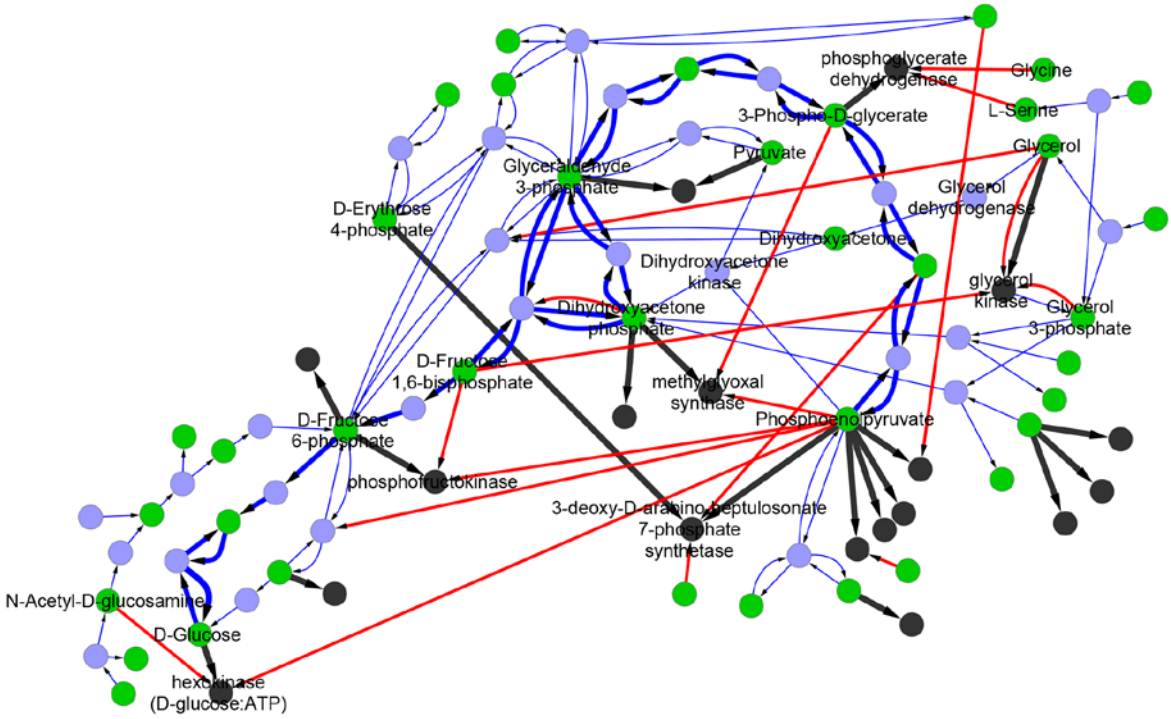



**Figure 4. Minimal hyperpaths leading to glucose with high degree of auto-canalization ($C_a$).** Sub-network containing minimal hyperpaths leading to glucose exhibiting significantly high value of $C_a$. Out-branching reactions are represented as black nodes, metabolites are represented as green nodes and other reactions are represented as blue nodes. All regulatory interactions contained within the sub-network are represented as red edges. The classical gluconeogenetic pathway is highlighted by bold edges.

Another hyperpath which is highlighted by the bold line on the figure 6 represents classical gluconeonenesis. Alteri *et al.* [29] studied metabolism of uropatogenic *E.coli* and found that during a urinary tract infection, peptides and amino acids are the primary carbon source for *E.coli*. Authors proved that gluconeogenesis is active in this condition and contributed significantly to the fitness of *E.coli* during a urinary tract infection. Amino acids serine, glycine and arginine are converted to pyruvate and oxaloacetate. Oxaloacetate is converted to phosphoenolepyruvate (PEP) which enters to gluconeogenesis (Fig.4). Figure 4 shows the inhibition of phosphoglycerate dehydrogenase by glycine and serine which is consistent with concept of autocanalization and with experimental results in [29], despite glycine and serine are not produced by the highlighted minimal hyperpath. Inhibition of phosphoglycerate

dehydrogenase by serine and glycine prevent occurrence of futile cycle because 3-phospho-D-glycerate is generated by the gluconeogenesis from the glycine and serine, and at the same time 3-phospho-D-glycerate is a substrate for serine and glycine synthesis by another path. The hyperpaths generated are too short to capture the entire path from serine to glucose.

3-phospho-D-glycerate inhibits methylglyoxal synthase within the highlighted hyperpath in figure 4 together with PEP. This double inhibition is consistent with the presence of oxygen in the cultivation conditions and in the urinary tract. There is no need to compensate redox balance by the production of 1,2-propanediol from methylglyoxal because of external electron acceptors (oxygen). In the case of hyperpaths representing fermentative assimilation of glycerol, the inhibition of methylglyoxal synthase by 3-phospho-D-glycerate is missing, relaxing production of 1,2-propanediol. Note that the inhibition of methylglyoxal synthase by PEP is present within these hyperpaths. It was proved experimentally that the replacement of *E.coli* PEP-dependent dihydroxyacetone kinase with *C.freundi* ATP-dependent dihydroxyacetone kinase, increased succinate yield two-fold during fermentative growth on glycerol [28].

**Discussion**

The present work brings first evidence that the regulatory network representing direct regulatory interactions between enzymes and metabolites exhibits structural properties that significantly deviates from the relevant random models. I show that the metabolic regulatory system has an increased potential to inhibit reactions consuming metabolites from metabolic path (or hyperpath) and in the same time leaves the metabolic path (or hyperpath) unaffected – a property formalized as quantity *C*. This property C can be further decomposed on two more specific properties ($C_a$ and $C_f$) that tightens the definition of *C*. It was shown that both are significantly increased within real regulatory network implying evolutionary adaptations. I observed that metabolic paths and hyperpaths exhibits an increased tendency to positively regulate itself, indirectly by inhibition of out-branching reactions – expressed by the quantity $C_a$. At the same time reactions branching out of metabolic paths are significantly more often regulated by the metabolite for which a path leading from the out-branching reaction to this metabolite exists and this path has no branching – expressed by the quantity $C_f$ . A simple explanation of the significantly higher values of $C_f$ within the real network is that the out-branching reactions are the beginnings of specific paths, and products of these paths inhibit out-branching reactions forming the simplest type of negative feedback loops. These regulatory products of the paths starting in out-branching reactions do not inhibit reactions of the central path because the central path can supply many other paths and saturated state of one out-branching path does not guarantee that other paths consuming intermediates from the central path are also saturated.

While focusing on minimal hyperpaths leading to glucose, I discovered some hyperpaths exhibiting a significantly higher degree of auto-canalization which corresponds to experimentally documented metabolic states of *E. coli.* Experiments with uropatogenic and glycerol fermentative *E. coli* show that the canalization structures discovered in this work can play some role in real conditions. Auto-canalization of metabolic paths and hyperpaths is a non-trivial property and should be further investigated. It is widely accepted that metabolism

is mainly controlled by the transcriptional regulation and this work does not oppose this concept. The regulatory auto-canalization mediated by direct regulatory interactions of metabolites with enzymes can work in harmony with transcriptional regulation. One can hypothesize that in a regulatory-metabolic network there are larger and more complex auto-canalization structures than metabolic paths and hyperpaths that cannot be discovered by the enumerative approach used in this work. I plan to use linear optimization techniques and flux balance analysis in order to find such general structures in further research.

The flexibility of proteins allow organisms to explore the entire realm of possible combinations of regulatory configurations. This fact known as gratuity, makes it possible to use relatively unconstrained random models to assess the significance of network properties. The first randomized ensemble represents a minimal model which only constraints the maximum number of regulators that can act on single protein molecule. The third random ensemble is a very conservative random model fixing all elementary properties of real regulatory network. The two most important quantities measured in this work ($C_a$ and $C_f$) exhibit significant deviations from all considered ensembles in the case of metabolic paths. This indicate that measured properties are not simple consequences of elementary network variables and have probably more complex causes connected with evolutionary adaptations of the bacteria.

## Methods

### *Datasets used*

Metabolic network structure was extracted from the metabolic model of *E. coli* metabolism iAF1260 [23]. I used only cytoplasmic reactions and the largest connected component of resulting metabolic hypergraph to subsequent analysis. Regulatory interactions between metabolites and reactions was extracted from Ecocyc database [24]. Following list of currency metabolites was removed from the network: ATP, ADP, AMP, Pi, NAD, NADH, NADP, NADPH, FAD, FADH2, NH3, NH4, CO2, H2O2, O2, H2, CoA, H2O, PPi, H.

### *Basic hypergraph definitions*

Metabolic network was represented as oriented hypergraph.

Oriented hypergraph is pair $H = (V, E)$, where $V$ is set of vertices (metabolites), $E$ is set of hyperedges (reactions). Oriented hyperedge is pair $e = (e^+, e^-)$, where $e^+, e^- \subset V$ and $e^+ \cap e^- = \varnothing$.

$e^+$ and $e^-$ represents substrate set and product set of the reaction $e$. Reversible reaction is represented as two hyperedges where substrate set of one of them is equal to product set of second a vice versa. In addition to oriented bipartite graph, directed hypergraph contains information which metabolites are substrates and which are products of reversible reactions. Absence of this information in oriented bipartite representation can lead to a misidentified path, containing two substrates of a single reversible reaction where one substrate acts as a substrate and the other as a product within the path.

Elementary path *p* is sequence of metabolites and reactions $p = (v_0, e_1, v_1, e_2, v_2, ..., e_k, v_k)$ such that all metabolites and reactions are unique in *p* and $v_{i-1} \in e_i^+ \wedge v_i \in e_i^-$, for $i = 1, ..., k$.

I extend the definition of the two conditions:

1) $e_i^- \neq e_j^+ \vee e_i^+ \neq e_j^-, \qquad \forall i < j : e_i, e_j \in p$

2) $v_j \notin e_i^+, \qquad \forall i < j : e_i^+, v_j \in p$

First condition ensures that one reversible reaction can be used only in one direction in the path. This prevent occurrence of cycle. Second condition ensures that no metabolite in the path is in the same time substrate of any reaction in path with is required to produce this metabolite. I call the elementary path fulfilling definition of the elementary path and conditions 1) and 2) defined above the metabolic path.

Hyperpath *P* going from a source subset $S_P \subset V$ to target subset $T_P \subset V$ is hypergraph $H_P = (V_P, E_P)$ with $V_P \subseteq V$, $E_P \subseteq E$, such that it is possible to sort hyperedges of *P* in to the ordering $(e_1, ..., e_k)$ with the following properties:

1) $e_j^+ \subseteq S_P \cup \left( \bigcup_{i<j} e_i^- \right)$, for $j = 1, ..., k$

2) $T_P \subseteq S_P \cup \left( \bigcup_{i=1}^{k} e_i^- \right)$

First condition ensures that every substrate of reactions participating in the hyperpath is produced by other reaction in the hyperpath or/and is member of source set $S_P$.

Hyperpath $P(V_P, E_P)$ is minimal if it has no proper subset, $P'(V_P', E_P')$, where $V_P' \subseteq V_P$ and $E_P' \subseteq E_P$ with the same source and target subsets.

***Reactions branching out of the metabolic path and hyperpath***

I define set of reactions branching out of the metabolic path, $E_{out}(p)$, as all irreversible reactions that consume at least one metabolite from the path and are not members of the path *p*. Formally, $E_{out}(p) = \{e : e^+ \cap p(v) \neq \varnothing \wedge e \notin p \wedge e \in E_{ir}\}$, where $E_{ir}$ is set of hyperedges representing irreversible reactions and $p(v)$ is set of metabolites which are members of *p*.

Similarly, I define set of reactions branching out of the minimal hyperpath $E_{out}(P)$ as all irreversible reactions that consume at least one metabolite which is produced and consumed within the minimal hyperpath or belong to $S_P$ or $T_P$ and in the same time these reactions are not members of the *P*.

Formally, $E_{out}(P) = \{e : e^+ \cap V_P^{es} \neq \varnothing \wedge e \notin E_P \wedge e \in E_{ir}\}$, where

$V_P^{es} = S_P \cup T_P \cup \left( \bigcup_{e_i, e_j \in E_P} (e_i^- \cap e_j^+) \right)$.

Quantities measured in this work are defined as a size of fraction of $E_{out}(p)$ or $E_{out}(P)$ that fulfills some special conditions.

***Confidence bounds for cumulative distribution functions and percentiles***

Each random network within the randomized ensemble corresponds to single cumulative function of studied quantity. For every possible value of the studied quantity I computed 5 and 95 percentile of the value of its cumulative function within randomized ensemble. I apply bootstrap test (bias corrected and accelerated percentile method [30]) in order to estimate confidence bounds for 5th percentile of cumulative function values.

In the case of minimal hyperpaths a graphical method is used in order to highlight the abundance of highest values of $C_a$. For every xth percentile observed in the real network the corresponding set of xth percentiles within the randomized ensemble are computed. The confidence bounds for 95th percentile of this set are computed using the same bootstrap method as above. If the curve corresponding to upper bound of this confidence interval fall under the linear function with slope 1 on the percentile-percentile plot, the corresponding xth percentile within real network is significantly bigger than xth percentiles from the random networks Fig.5.

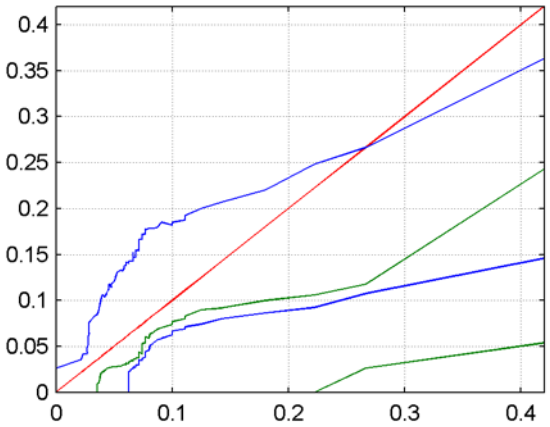



**Figure 5. Confidence bounds for percentiles of $C_a$ within randomized ensembles in comparison with real network.**
By red color is denoted linear function with slope 1 corresponding to the comparison of percentiles of real network with its self. Confidence bounds for 5 and 95 percentiles within first randomized ensemble compared with real network are green. Confidence bounds for 5 and 95 percentiles within third randomized ensemble compared with real network are blue.


## Acknowledgements

J. Geryk thanks all the members of the Department of Philosophy and History of Sciences for inspirational discussions and suggestions. Special thanks for the help in statistical evaluation of results goes to Aleš Kuběna.

## References


1. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. Nature 407: 651-654.

2. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proceedings of the Royal Society of London Series B-Biological Sciences 268: 1803-1810.

3. Ma HW, Zeng AP (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19: 270-277.

4. Huss M, Holme P (2007) Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. Iet Systems Biology 1: 280-285.

5. Gerlee P, Lizana L, Sneppen K (2009) Pathway identification by network pruning in the metabolic network of Escherichia coli. Bioinformatics 25: 3282-3288.

6. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297: 1551-1555.

7. Zhao J, Yu H, Luo JH, Cao ZW, Li YX (2006) Hierarchical modularity of nested bow-ties in metabolic networks. Bmc Bioinformatics 7.

8. Ma HW, Zhao XM, Yuan YJ, Zeng AP (2004) Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. Bioinformatics 20: 1870-1876.

9. Guimera R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433: 895-900.

10. Parter M, Kashtan N, Alon U (2007) Environmental variability and modularity of bacterial metabolic networks. Bmc Evolutionary Biology 7.

11. Kreimer A, Borenstein E, Gophna U, Ruppin E (2008) The evolution of modularity in bacterial metabolic networks. Proceedings of the National Academy of Sciences of the United States of America 105: 6976-6981.

12. Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. Proceedings of the National Academy of Sciences of the United States of America 102: 13773-13778.

13. Samal A, Wagner A, Martin OC (2011) Environmental versatility promotes modularity in genome-scale metabolic networks. Bmc Systems Biology 5.

14. Barrett CL, Herring CD, Reed JL, Palsson BO (2005) The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. Proceedings of the National Academy of Sciences of the United States of America 102: 19103-19108.

15. Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proceedings of the National Academy of Sciences of the United States of America 102: 2685-2689.

16. Ihmels J, Levy R, Barkai N (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nature Biotechnology 22: 86-92.

17. Samal A, Jain S (2008) The regulatory network of E. coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. Bmc Systems Biology 2.

18. Seshasayee ASN, Fraser GM, Babu MM, Luscombe NM (2009) Principles of transcriptional regulation and evolution of the metabolic system in E. coli. Genome Research 19: 79-91.

19. Goelzer A, Brikci FB, Martin-Verstraete I, Noirot P, Bessieres P, et al. (2008) Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis. Bmc Systems Biology 2.

20. Gutteridge A, Kanehisa M, Goto S (2007) Regulation of metabolic networks by small molecule metabolites. BMC Bioinformatics 8: 88.

21. Hao D, Ren C, Li C (2012) Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. Bmc Systems Biology 6: 34.

22. Geryk J, Slanina F Modules in the metabolic network of *E.coli* with regulatory interaction. International Journal of Data Mining and Bioinformatics "in press".

23. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular Systems Biology 3.

24. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, et al. (2011) EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Research 39: D583-D590.

25. Basler G, Grimbs S, Ebenhoh O, Selbig J, Nikoloski Z (2012) Evolutionary significance of metabolic network properties. J R Soc Interface 9: 1168-1176.

26. Samal A, Martin OC (2011) Randomizing genome-scale metabolic networks. PLoS One 6: e22295.

27. Monod J (1971) Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology: New York: Vintage.

28. Gonzalez R, Murarka A, Dharmadi Y, Yazdani SS (2008) A new model for the anaerobic fermentation of glycerol in enteric bacteria: trunk and auxiliary pathways in Escherichia coli. Metab Eng 10: 234-245.

29. Alteri CJ, Smith SN, Mobley HLT (2009) Fitness of Escherichia coli during Urinary Tract Infection Requires Gluconeogenesis and the TCA Cycle. PLoS Pathog 5.

30. Efron B (1987) Better Bootstrap Confidence-Intervals. Journal of the American Statistical Association 82: 171-185.