MTA doktori tézisek

Evolúciógenetika a genomléptékű modellezés korában

Papp Balázs

Szegedi Biológiai Kutatóközpont, Eötvös Loránd Kutatási Hálózat

Biokémiai Intézet

Szintetikus és Rendszerbiológiai Egység

Szeged, 2021

I. Bevezetés

Az evolúcióbiológia számos megoldatlan kérdése a genotípus és a fenotípus közötti bonyolult összefüggéseket alakító mutációs hatásokkal kapcsolatos. Például miért van a legtöbb inaktiváló mutációnak csekély hatása a fenotípusra? A többszörös mutációk közötti kölcsönhatások hogyan hoznak létre új fenotípust? Az ilyen kérdések megoldásához szükség van annak megértésére, hogyan képezhető le a genotípus a fenotípusra. Míg a genotípus - fenotípus leképezésnek az egyes fehérjék szintjén történő vizsgálata már régóta folyik (Dean and Thornton 2007), a nagyobb génhálózatok vizsgálata le van maradva. Legalább két alapvető okból fontos lenne kvantitatív modellrendszerek kidolgozása a mutációs hatások molekuláris szintű hálózatokban történő vizsgálata céljából. Egyrészt egy ilyen modellrendszer mechanisztikus betekintést engedne bonyolult evolúciós jelenségekbe, a többszörös mutációktól függő evolúciós újdonságok megjelenésétől kezdve a minimalizált genomok evolúciójáig. Másrészt prediktívebbé, előrejelzésre képesebbé tenné az evolúcióbiológia tudományát azáltal, hogy lehetővé tenné az evolúció kimenetelének specifikus előrejelzését. A prediktív modellrendszertámpontot nyújtana arra vonatkozóan, hogy az evolúció során mely gének elvesztése, mutációja, vagy expresszióváltozása várható. Az alapkutatás serkentése mellett egy ilyen modellrendszergyakorlati hasznot is hozna. Egyebek mellett lehetővé tenné evolúciós változások előrejelzését kórokozó baktériumokban (Sommer, et al. 2017) és információt nyújtana új bioszintetikus reakcióutak létrehozásához (Johannes and Zhao 2006; Notebaart, et al. 2018).

A rendszerbiológia jelenkori fejlődése példátlan lehetőséget kínál olyan számítógépes modellek megalkotására, amelyek számíthatóvá teszik a fenotípust a mutációk és környezeti változások alapján. Ezek a számítógépes megközelítések realisztikus molekuláris rendszerek matematikai modelljein alapulnak, és sokfélék lehetnek, a kisebb metabolikus rendszerek (Teusink, et al. 1998) és szabályozási körök (Chen, et al. 2004) részletes kinetikus modelljeitől a genomléptékű metabolikus hálózatok kényszer alapú modelljeiig (Price, et al. 2004). A kényszer alapú metabolikus modellek különösen alkalmasak a genotípusok és fenotípusok közötti leképezés tanulmányozására nagy hálózatokban, és már eddig is értékes betekintést nyújtottak egyes mikroorganizmusok anyagcseregén-készletének és fenotípusának evolúciójába (Feist and Palsson 2008). Ezek a modellek jó minőségű metabolikus hálózati rekonstrukciókból indulnak ki (Price, et al. 2004). E rekonstrukciók tipikusan genomannotációs adatok, enzim-adatbázisok, például a KEGG (Kanehisa and Goto 2000) és a BRENDA (Schomburg, et al. 2002), valamint elsődleges irodalmi adatok összegzése útján épülnek fel, ami sok nem automatizálható emberi munkát igényel. A biokémiai reakciókból álló hálózatot ezután matematikai formába öntik és kényszer alapú módszerekkel elemzik (Price, et al. 2004). Különösen a ‘flux balance analysis’ (FBA) módszer használatos széles körben az anyagcseretermékek hálózaton keresztüli optimális áramlásának kiszámítására a környezetben rendelkezésre álló tápanyagok függvényében. Ezeket az előrejelzéseket már kiterjedten tesztelték, és az empirikus adatokkal jól megegyezőnek találták (Edwards, et al. 2001; Snitkin, et al. 2008; Oberhardt, et al. 2009).

Jelen disszertáció négy fő kutatási témára összpontosít, amelyek mindegyike genomléptékű metabolikus hálózatokat alkalmaz az evolúciógenetika egy-egy régóta vitatott kérdésének vizsgálatára. Először, hogyan befolyásolják a különböző mutációk egymás fenotípusos hatásait,

azaz hogyan keletkeznek a genetikai kölcsönhatások mechanisztikus szinten? Továbbá milyen pontossággal tudjuk számítógépes módszerrel előrejelezni a metabolikus hálózat részletes ismerete alapján azt, hogy mely génpárok között lép fel genetikai kölcsönhatás? Másodszor, előre tudjuk-e jelezni erősen redukált genommal rendelkező, endoszimbióta baktériumok génkészletét? Azaz előre tudjuk-e jelezni, hogy több millió éves reduktív genomevolúció során mely gének vesznek el és melyek maradnak meg? Harmadszor, számítógépes módszerrel előre tudjuk-e jelezni az új környezetekhez való alkalmazkodás kimenetelét és genetikai alapját? Közelebbről, hogyan jön létre egyes enzimek meglévő, alacsony szintű mellékaktivitásából az új tápanyagok felhasználásának képessége? Végül pedig hogyan jönnek létre az olyan evolúciós újdonságok, amelyekhez több mutáció egyidejű megszerzése szükséges?

II. Genetikai kölcsönhatások anyagcsere-hálózatokban

Egy mutáció fenotípusos hatása gyakran más mutációknak a genomban való jelenlététől függ; ezt a jelenséget genetikai kölcsönhatásnak vagy episztatikus kölcsönhatásnak nevezik. A genetikai kölcsönhatások ismerete több szempontból is kulcsfontosságú: segít megérteni a gének közötti funkcionális kapcsolatokat; azt, hogy mennyire képesek az élő szervezetek elviselni a káros mutációkat; valamint a komplex genetikai betegségek hátterét. Az elmúlt évtized során nagy áteresztőképességű vizsgálatokban átfogó térképek készültek a gének közötti genetikai kölcsönhatásokról számos élőlényben, így többek között sarjadzó élesztőben (*Saccharomyces cerevisiae*) (Costanzo, et al. 2010; Costanzo, et al. 2016), *E. coli* baktériumban (Babu, et al. 2014) és humán sejtvonalakban (Horlbeck, et al. 2018). Ezek a vizsgálatok a funkcióvesztéses mutációkra összpontosítottak, és a genetikai kölcsönhatások két fő formáját azonosították: (1) a *negatív genetikai kölcsönhatásokat* (synthetic sick vagy lethal / szinergisztikus), amikor két

mutáció fokozza egymás káros hatását, azt jelezve, hogy funkcionális kompenzáció lehetséges közöttük, valamint (2) *pozitív genetikai kölcsönhatásokat* (antagonisztikus), amikor egy mutáció hatása egy másik káros mutáció jelenlétében a vártnál kisebb. Bár a genetikai kölcsönhatásokra vonatkozóan már sok kísérleti adat gyűlt össze, az episztázis szerveződését és mechanisztikus hátterét illetően még számos nyitott kérdés maradt. Az elmúlt években munkámmal négy megoldatlan kérdés tisztázásához járultam hozzá: (1) Mi az oka annak, hogy a gének túlnyomó többsége kevés genetikai kölcsönhatásban vesz részt, míg néhány központi hálózati csomópontként ('hub'-ként) működő gén a genetikai kölcsönhatások hálózatának számos pontjával kapcsolatban áll? (2) A genetikai kölcsönhatások hátterében álló biokémiai hálózat mai ismerete alapján lehetséges-e számítógépes módszerrel előrejelezni, melyik konkrét génpár között létesül genetikai kölcsönhatás? (3) A kísérleti adatok és a modellek előrejelzései közötti eltérések felhasználhatók-e az anyagcseremodell automatikus finomítására? (4) Környezetfüggők-e a genetikai kölcsönhatások, és ha igen, miért?

A felsorolt kérdések megválaszolásához integrált rendszerbiológiai megközelítést alkalmaztunk: kísérletes adatok alapján részletes térképet készítettünk az élesztő anyagcsere genetikai kölcsönhatásairól, majd az adatokat egy genomléptékű anyagcserehálózat-modellel integráltuk. Elemzéseink négy fő következtetésre vezettek (Harrison, et al. 2007; Szappanos, et al. 2011):

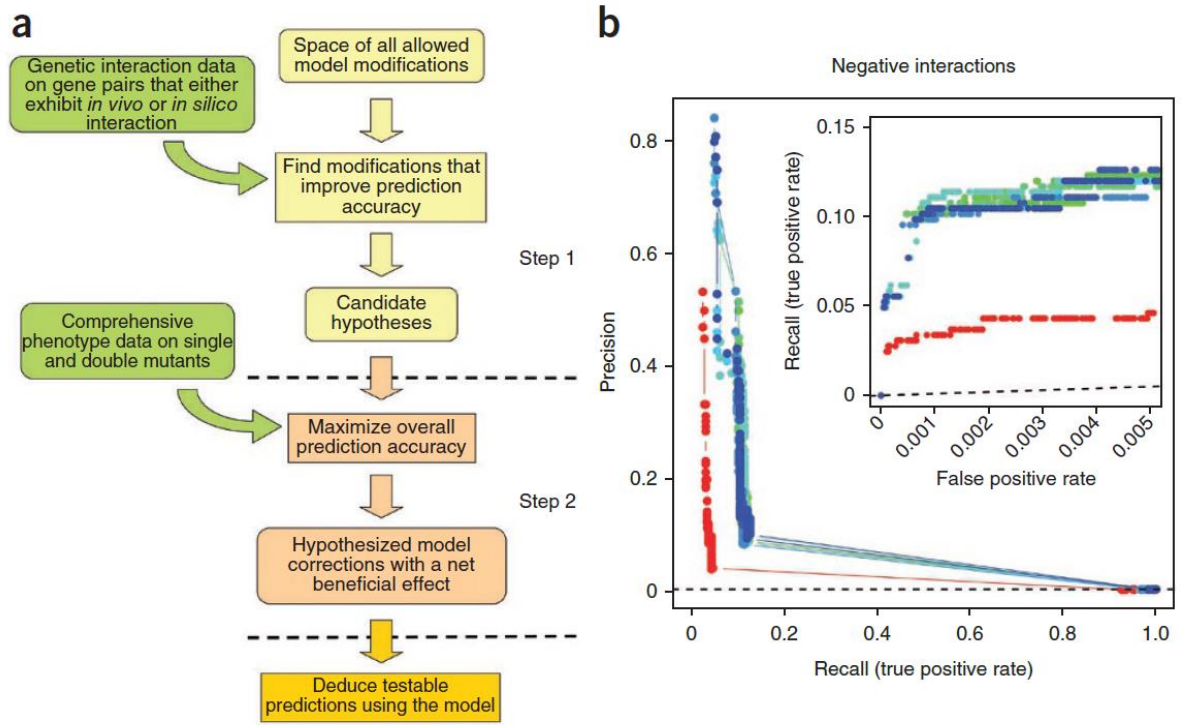
1) A számítógépes modell sikeresen előrejelezte, hogy a nagy fitnesz-hozzájárulású gének nagyszámú, mind pozitív, mind negatív genetikai kölcsönhatást mutatnak. Fontos megjegyezni, hogy a genomléptékű modell csak a biokémiai reakciók sztöchiometriájára és a sejtek növekedési követelményeire vonatkozó információt tartalmazza, anélkül, hogy figyelembe venné a génreguláció és az enzimkinetika hatásait. Ez az eredmény ezért azt sugallja, hogy a genetikai

kölcsönhatási hálózatok hub-jait az anyagcsere-hálózatok szerkezete hozza létre. A modellezés azt is megmutatta, hogy a hub-okat nagyszámú biológiai folyamatban résztvevő, pleiotróp enzimekhozzák létre azáltal, hogy hozzájárulnak számos kulcsfontosságú biomassza-összetevő szintéziséhez. Ennélfogva elvesztésük fenotípusos hatását a hálózat számos más enzime is módosíthatja, ami nagyszámú episztatikus kölcsönhatást eredményezhet.

2) A genetikai kölcsönhatások adatainak a genomléptékű anyagcseremodellel való összehasonlítása az első nagy léptékű kísérlet volt annak tesztelésére, hogy képesek vagyunk-e egyedi genetikai kölcsönhatásokat előrejelezni genomléptékű anyagcseremodellek alapján. Körülbelül 67 500 metabolikus génpár nagy megbízhatóságú adatainak elemzésével kimutattuk, hogy az *in vivo* kimutatott kölcsönhatások erősen feldúsulnak a számítógépes módszerrel előrejelzett kölcsönhatások között (a negatív kölcsönhatások dúsulása százszoros, a pozitív kölcsönhatásoké hatvanszoros). Azonban az anyagcseremodell nem képes előrejelezni az *in vivo* azonosított genetikai kölcsönhatások többségét (a negatív kölcsönhatások 97%-át és a pozitív kölcsönhatások 89%-át; 1. ábra).

3) A nagyszámú, kísérletesen kimutatott genetikai kölcsönhatás elvileg gazdag információforrást kínál a modell adatalapú feljavítása számára. E célból kifejlesztettünk egy gépi tanulási módszert, amely automatikusan olyan módosításokat javasol a modellnek, amelyek javítják a negatív genetikai kölcsönhatások előrejelzésére való képességét (1A. ábra). A módszer több módosítást javasolt, amelyek együttesen javították a modell illeszkedését az adatokhoz [a szenzitivitás (recall) 100–267%-kal, a pozitív találatok megbízhatósága (precision) 44–59%-kal emelkedett; 1B. ábra]. Az egyik javasolt módosítás az aszpartátból kiinduló de novo NAD

bioszintetikus reakcióút eltávolítása volt. Ez a reakcióút jelen van *E. coli* baktériumban (Flachmann, et al. 1988), de valószínűleg tévedésből került be az élesztő anyagcsere-hálózatába. Ellenőrző kísérleteink megerősítették, hogy e reakcióút célzott eltávolítása lehetővé teszi a kinurenin reakcióútban érintett mutánsok nikotinsav-auxotrófiájának helyes előrejelzését.



1. ábra. Az anyagcseremodell automatikus finomítása. *(A) A finomító algoritmus munkafolyamata. Mivel minden egyes modell kiértékelése nagy számítási igényű (azaz minden egyes modellnél nagy számú géndeléciót kell szimulálni), kétlépcsős eljárást alkalmaztunk annak érdekében, hogy minden rendelkezésre álló fenotípus-adatot felhasználjunk, de a számítógépes feldolgozás is kivitelezhető maradjon. Az első lépcsőben modelleket kerestünk: ehhez egy adott modellben csak azokat a génpárokat értékeltük, amelyek az eredeti modell szerint részt vesznek akár in vivo, akár in silico kölcsönhatásban. Mivel a genetikai kölcsönhatások mind in vivo, mind in silico igen ritkák, a jelen munkánk során*

megvizsgált génpárok többsége nem vesz részt kölcsönhatásokban, és kihagyásuk jelentősen felgyorsítja a hipotézistér átvizsgálását. A második lépcsőben új, nagyon korlátozott hipotézisteret definiáltunk az első lépcső legsikeresebb modelljei alapján, de olyan modelleket kerestünk, amelyekben javul az előrejelzés átfogó pontossága, amit a populáció mindegyik modelljének minden részletre kiterjedő értékelésével mértünk. (B) A modell finomításának hatása az előrejelzés pontosságára az algoritmus 8 független lefuttatása után. Az ábra a módosított (kékről zöldre) és az eredeti (piros) modelleknek a tapasztalati alapon azonosított genetikai kölcsönhatások adataival való egyezését mutatja mindprecision-recall, mind pedig részleges ROC görbék (betét) segítségével. A szaggatott vonalak az előrejelzés véletlenszerűen várt pontosságát mutatják. Megjegyzés: bár az ábra esetében a modellnek mind a finomításához, mint az értékeléséhez ugyanazt az adatkészletet használtuk, a keresztellenőrzés a modell jelentős javulását mutatta (Szappanos, et al. 2011). Az ábra forrása: (Szappanos, et al. 2011)

4) Végül, számítógépes módszerrel elemeztük hogyan változnak a genetikai kölcsönhatások különféle tápanyag-ellátottsági körülmények között és kimutattuk, hogy a genetikai kölcsönhatások gyakrankörnyezetfüggőek. A vizsgálat az úgynevezett synthetic lethal kölcsönhatásokra összpontosított, amely a negatív genetikai kölcsönhatás szélsőséges formája: itt a dupla génkiütés növekedésre képtelen fenotípust mutat, ami egyik szimpla delécios mutánsnál sem jelenik meg. A synthetic lethalkölcsönhatásokról úgy tartják, hogy génpárok közötti funkcionális redundanciát jeleznek, például alternatív reakcióutakat vagy génduplikációt (Hartman, et al. 2001). Lényeges, hogy munkánk kimutatta: számos, egy bizonyos környezetben synthetic lethal kölcsönhatásokban részt vevő gén egy másik környezetben létfontosságúvá válik, ami azt jelzi, hogy redundanciájuk inkább látszólagos, mintsem valódi. Általánosabban megfogalmazva, az anyagcsere-hálózatok genetikai perturbációkkal szembeni ellenállóképessége valószínűleg a változatos tápanyag-ellátottsági viszonyok közötti túléléshez szükséges alkalmazkodás egyik melléktermé ke lehet.

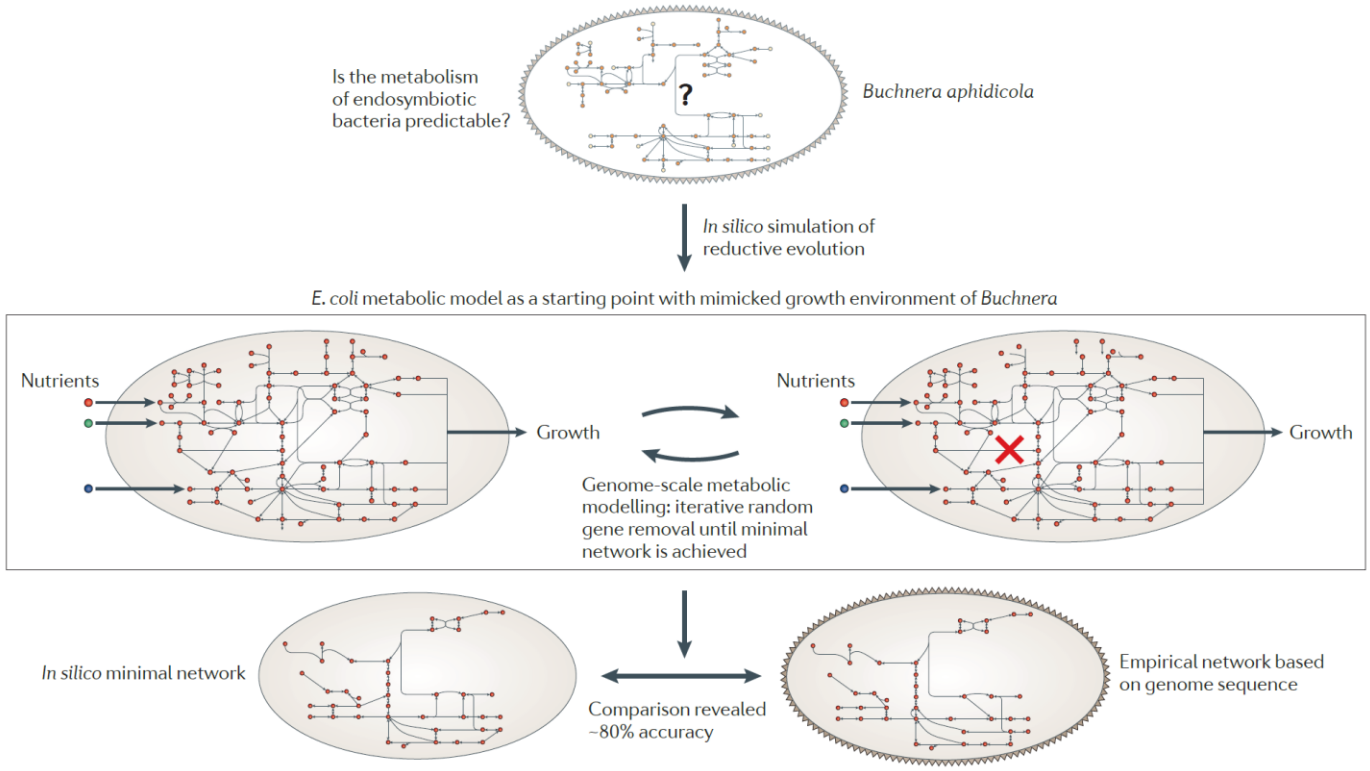
III. A genomredukció előrejelzése

A poszt-genomi korszak egyik központi kérdése annak megértése, milyen élőlények milyen génekkel rendelkeznek. Az ilyen megállapítások tipikusan „a posteriori” születnek, azaz miután felfedeztük, hogy egy adott élőlényben megvan az adott gén, hipotéziseket alkotunk annak ökológiájáról vagy biológiájáról. De lehetséges-e ennek a fordítottját megtenni, és ezáltal „a priori” prediktív, előrejelzés-szerű elméletet alkotni a genomról? Másképp fogalmazva, egy élőlény ökológiájának alapján megjósolhatjuk-e bármilyen pontossággal, hogy milyen génekkel kell rendelkeznie?

Arra, hogy megvizsgáljuk, megvalósítható-e a genom hosszú távú evolúciójának előrejelzése, azt a kérdést tettük fel, hogy az *E. coli* genomja alapján megjósolható-e a *Buchnera*, egy az *E. coli* baktériumból származó, erősen redukált genommal rendelkező endoszimbionta anyagcseréje (2. ábra). A *Buchnera* szabadon élő ősökből kiinduló evolúciója körülbelül 200 millió évvel ezelőtt kezdődött; ennek során génjeinek 75%-át elvesztette, és majdnem minimális, az élet fenntartásához szükséges génkészlettel (~600 génnel) rendelkezik. Az *E. coli* anyagcseréjének genomléptékű modelljéből indultunk ki, és a modellt úgyváltoztattuk meg, hogy a rendelkezésre álló élettani adatokalapján tükrözze az endoszimbióta életmódját. Flux balance analysisszimulációk sorozatával egymás után megvizsgáltunk véletlenszerűen kiválasztott géndelécióka t, és értékeltük, hogy a *Buchnera* ökológiáját figyelembe véve ezek hatásukat tekintve semlegesek vagy nem. Egymás után következő génkiütéseket szimuláltunk egészen addig, amíg már nem lehetett több gént kiütni az *in silico* növekedés (fitnesz) megzavarása nélkül, és így egy sor minimális hálózatot kaptunk (2. ábra). Figyelemre méltó, hogy ha

összehasonlítjuk ezeknek a minimális *in silico* hálózatoknak a génkészletét az akkor ismert három *B. aphidicola* genommal, látható, hogy az egyes gének jelenléte/hiánya nagy pontossággal előrejelezhető (ROC görbe alatti terület = 0,794 – 0,802, az endoszimbionta genomtól függően). Kétszáz millió évnyi reduktív genomevolúció tehát meglepően jól előrejelezhető az élőlény távoli ősének és jelenlegi életformájának ismerete alapján.

Még a közeli rokon *Buchnera* törzsek génkészlete is különböző. A reduktív genomevolúció ilyen változatosságát elvileg akár a szelekciós erők különbségei (ökológia), akár véletlen események is okozhatják, különbségeket hozva létre a génvesztések sorrendjében, ilyen módon választva a lehetséges alternatív reakcióutak között. Ismételt szimulációkból származó minimális hálózatokat összehasonlítva valóban találtunk is bizonyítékot véletlenszerű eseményekből eredő, részlegesen különböző evolúciós kimenetelekre.



2. ábra. Genomléptékű anyagcseremodellező megközelítés alkalmazása a genomredukció előrejelzésére endoszimbionta baktériumokban. *A számítógépes megközelítés a mai E. coli genomléptékű anyagcseremodelljével helyettesíti a B. aphidicola endoszimbionta szabadon élő őséét, és a mai endoszimbionta életformáját utánozza abból a célból, hogy előrejelezze az egyes génkiütési események hatását az evolúció előrehaladtával. Az evolúció szimulációi során egymást követő génkiütési események (piros kereszt) szimulációja útján minimális anyagcsere-hálózatokat hoztunk létre, addig folytatva a génkiütést, amíg már nem lehetett több gént eltávolítani az in silico növekedés megzavarása nélkül. A számítógépes módszerrel előrejelzett minimális hálózatok nagy átfedést mutattak a valódi* Buchnera *anyagcseregén-készletével (alsó ábrarész). Az ábra forrása: (Papp, et al. 2011).*

IV. Rejtett anyagcsere és az adaptív evolúció előrejelezhetősége

Az evolúció- és rendszerbiológia egyik központi témája annak megértése, hogy az adaptáció során hogyan alakulnak ki új molekuláris reakcióutak. A legjobban ismert hálózatokra, a

kismolekulájú anyagok metabolizmusára vonatkozóan az uralkodó paradigma szerint az evolúció a már létező enzimek gyenge mellékaktivitásait hasznosítja (Jensen 1976). Bár néhány sokat tanulmányozott enzim ilyen aktivitásainak biokémiai mechanizmusából sokat tanulhattunk, teljes homály fedi azt, hogy a sejt egész rendszerére vonatkoztatva a rejtett reakciók mennyire járulnak hozzá az új reakcióutak megjelenéséhez.

A fenti ismerethiány megszüntetésére olyan rendszerszintű megközelítést alkalmaztunk, amely bepillantást enged a rejtett anyagcsere felépítésébe, egyúttal pedig lehetővé teszi a rejtett aktivitások szerepének előrejelzését az új tápanyagokhoz való alkalmazkodásban (Notebaart, et al. 2014). Röviden, az *Escherichia coli* baktériumra koncentráltunk, amely az enzimaktivitások szempontjából a legrészletesebben jellemzett élőlény. Ezért adatbázisok és irodalmi áttekintés alapján elkészítettük az *Escherichia coli* rejtett anyagcsere-hálózatának számítógépes rekonstrukcióját, és azt integráltuk ugyanennek az élőlénynek a natív genomléptékű anyagcseremodelljével (Feist, et al. 2007). Hozzátettünk összesen 262 olyan rejtett reakciót és 277 olyan metabolitot, amelyek nincsenek jelen a natív hálózatban. Az így kapott rekonstrukció a valaha készült első átfogó számítógépes modell, amely leképezi egy élőlény rejtett anyagcseréjét. Ezután megvizsgáltuk a rejtett anyagcsere evolúciós potenciálját, és a következő eredményeket kaptuk:

1) A hálózat szerkezetének elemzése azt mutatta, hogy a rejtett reakciók jelentős része képes új, biológiai szempontból potenciálisan nagy jelentőségű reakcióutakat kialakítani.

2) *In silico* megbecsültük az *E. coli* evolúciós potenciálját több száz új tápanyaghoz való alkalmazkodásra. Számítógépes módszerrel előrejeleztük, hogy rejtett reakcióknak a natív

hálózathoz való hozzáadása milyen hatással van a maximális növekedésre különféle környezeti körülmények között. Mivel az új tápanyagok felhasználására irányuló evolúciós képesség érdekelt minket, ezért feltételeztük, hogy minden rejtett reakció hasznosulhat (azaz ennek nincsenek enzimkinetikai vagy szabályozási akadályai). Az elemzés több tucatnyi olyan esetet kimutatott, ahol rejtett reakciók, aktivitásuk fokozódása által, lehetővé teszik vagy elősegítik a növekedést korábban nem hasznosított tápanyagon. Az esetek többségében egyetlen rejtett reakció idézte elő a fokozott növekedést, és csak kevés esetben volt szükség egyidejűleg több reakció közreműködésére (lásd (Notebaart, et al. 2014)).

3) Egy nagy áteresztőképességű kísérleti szűrést is végeztünk a rejtett aktivitások képességének becslésére új tápanyagforrásokhoz való alkalmazkodásban. Teljes genomra kiterjedő géntúltermeléses kísérletet hajtottunk végre*E. coli* baktériumban, és 194 szénforrás jelenlétében mértük a növekedést (Patrick, et al. 2007; Kim, et al. 2010). Összesen 17 olyan gént azonosítottunk, amely túltermelése hatására 17 különböző szénforrás közül legalább egynek a jelenlétében fokozta a növekedést. A 17 gén közül 11 enzimeket kódolt, amelyek közül 9 ismert rejtett aktivitással rendelkezett. Ezek az elemzések erőteljesen alátámasztják azt az elképzelést, hogy az evolúció képes a rejtett reakciók felhasználására mind a növekedés elősegítésére változatlan környezeti feltételek között, mind pedig teljesen új tápanyagforrások kiaknázására.

4) Előre tudjuk-e jelezni az új környezeti feltételekhez való alkalmazkodás genetikai alapjait? A számítógépes előrejelzések figyelemreméltó egyezést mutattak a teljes genomra kiterjedő géntúltermeléses kísérlet eredményeivel. Nevezetesen, a számítógépes modell sikeresen

megjelölte azon szénforrások 44%-át, amelyek jelenlétében egy enzim kísérletes túltermelése lehetővé tette vagy fokozta a növekedést – ez az átfedés statisztikailag magas szinten szignifikáns ($P<10^{-13}$). Ezek az eredmények bizonyítják, hogy egy élőlény rejtett anyagcseréjének részletes ismerete alapján lehetséges előrejelezni az új tápkörülmények irányába tartóevolúció genetikai alapjait.

V. Egyszerű utak komplex adaptációkhoz

Az evolúciós újítások eredének magyarázata továbbra is az evolúcióbiológia egyik központi kérdése. Különösen komoly kihívást jelentenek az evolúcióbiológusok számára azok a tulajdonságok, amelyek létrejöttéhez több mutáció egyidejű megjelenésére van szükség, amelyeknek önmagukban látszólag egyike sem jelent előnyt az egyén számára. Az ilyen tulajdonságokat gyakran komplex adaptációnak nevezik, és evolúciójukat nem könnyű levezetni, mégpedig nem fizikai vagy kémiai akadályok, hanem azon törvényszerűségek miatt, amelyek a mutációk populációban való terjedésének dinamikáját határozzák meg.

A komplex adaptáció paradoxonjának feloldására egy egyszerű elvi modellt javasoltunk. Ez a forgatókönyv szorosan kapcsolódik a preadaptáció elképzeléséhez, és kizárólag az egymást követő, jótékony hatású mutációk felhalmozódására támaszkodik. Röviden összefoglalva: az időben változó környezeti feltételek olyan egylépéses adaptív mutációkra szelektálnak, amelyek – mintegy melléktermékként – ugródeszkául szolgálnak bonyolultabb fenotípusok kialakulásához. A komplex adaptációk tehát dinamikusan változó környezeti körülmények között felgyorsulhatnak. Megvizsgáltuk, hogyan jöhetnek létre új tápanyagforrásokat felhasználni képes fenotípusok a bakteriális anyagcsere-hálózatban új enzimreakciók hozzáadásával és háromféle bizonyítékot találtunk a felvázolt forgatókönyv működésére:

1) Először *in silico* tanulmányoztuk az *E. coli* anyagcsere-hálózat bővítését, új tápanyagok hasznosításának képességét keresve. Számítógépes elemzéseink azt mutatták, hogy új,

összetett reakcióutak képesek kialakulni olyan biokémiai reakciók egyenkénti, egymást követő megszerzése révén, amelyek mindegyike előnyt jelent specifikus környezeti feltételek mellett.

2) Másodszor, a változó környezetre épülő modellünk azt jósolja, hogy az új anyagcseregének megjelenésének meghatározott sorrendben kell történnie. Nevezetesen, ha egy új környezetben a növekedés fenntartásához két enzim együttes jelenléte szükséges, és az egyik enzim egy másik környezetben önmagában is előnyt biztosít, akkor ennek az utóbbi enzimnek kell először megjelennie. És csakugyan: 943 baktériumban a génszerzési események evolúciós történetének filogenetikai elemzése alátámasztotta ezt az előrejelzést.

3) Végül laboratóriumi evolúciós vizsgálatot végeztünk, amelynek célja az *E. coli* adaptációja volt két új szénforráshoz. Kísérleteink azt mutatták, hogy az egyik szénforráshoz történő evolúciós alkalmazkodás megkönnyítette a későbbi alkalmazkodást a másik szénforráshoz.

Összegezve, a fenti eredmények azt bizonyítják, hogy a komplex anyagcsere-adaptációk valóban kialakulhatnak az alkalmazkodást szolgáló, köztes mutációkon keresztül, a tápanyag-hasznosítási képesség lépésenkénti kibővítésével. Ez a következtetés fontos elvi előrelépést jelent, mivel nem szükséges feltételezni, hogy a semleges köztes mutációk felhalmozódásának lassú folyamata is szerepet játszik.

VI. A legfontosabb eredmények összefoglalása

Ez a disszertáció egy sor olyan közleményen alapul, amelyek anyagcsere-hálózatok számítógépes rendszerbiológiai modellezését alkalmazzák az evolúcióbiológia több lezáratlan kérdésének megoldására. A következő főbb eredményeket kaptuk, amelyeket az alább kiemelt cikkekben közöltünk:

1) Genomléptékű anyagcsere-hálózatok segítségével kísérletet tettünk a genetikai kölcsönhatások előrejelzése határainak tágítására, és kimutattuk, hogy a genetikai kölcsönhatási hálózat csomópontjai (‘hub’-jai) nagy mértékben előrejelezhetők, azonban ezek a számítógépes modellek gyakran elmulasztják az egyedi genetikai kölcsönhatások pontos előrejelzését. Ezeknek az eltéréseknek az alapján gépi tanulási módszert dolgoztunk ki, amely nagy léptékű genetikai kölcsönhatás-adatok alapján finomítja az anyagcserehálózati modellt.

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., Andrews, B.J., Boone, C., Oliver, S.G., Pál, C., **Papp, B.** (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genetics* **43:** 656

2) Kimutattuk, hogy a ’synthetic lethal’ genetikai kölcsönhatások gyakran a fennálló környezeti viszonyoktól függenek. Fontos megjegyezni, hogy ezt gyakran az okozza, hogy környezetváltozást követően a genetikai kölcsönhatásban részt vevő egyik vagy mindkét gén létfontosságúvá válik, ami arra utal, hogy a két gén csak részlegesen redundáns.

*Harrison, R., **Papp, B.**, Pál, C., Oliver, S.G., Delneri, D. (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A.* **104**: 2307-12.

3) Az *E. coli* genomléptékű anyacseremodelljében a nem esszenciális gének fokozatos eltávolításának szimulációjával előrejeleztük olyan, szoros rokonságban álló endoszimbionta baktériumok erősen lecsökkentett génkészletét, amelyek kb. 200 millió évvel ezelőtt különültek el az *E. coli*-tól.

*Pál, C., **Papp, B.**, Lercher, M.J., Csermely, P., Oliver, S.G. and Hurst, L.D. (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**: 667-70.

4) Elkészítettük az *E. coli* baktériumban ismert enzim-mellékaktivitások (azaz rejtett reakciók) átfogó hálózatának rekonstrukcióját, amely a valaha készült első ilyen rekonstrukció. Számítógépes szimulációk és egy nagy áteresztőképességű kísérletek kombinálásával, a tápkörülmények több száz különféle variációja mellett előrejeleztünk és igazoltunk olyan környezeti körülményeket, ahol a rejtett reakciók fokozott aktivitása fitneszelőnyt biztosít. Eredményeink bizonyítják, hogy az evolúció során rejtett reakciók közreműködésével létrejövő adaptációk genetikai alapjai számítógépes módszerrel előrejelezhetők.

Notebaart, R.A.*, Szappanos, B., Kintses, B., Pál, F., Györkei, A., Bogos, B., Lázár, V., Spohn, R., Csörgő, B., Wagner, A., Ruppin, E., Pál, C.*, **Papp, B.*** (2014) Network-level architecture and the evolutionary potential of underground metabolism. *Proc Natl Acad Sci U S A.* **111**: 11762-11767.

5) Új modellt állítottunk fel a komplex adaptációk paradoxonjának feloldására – ezek olyan bélyegek, amelyek előfeltétele több mutáció egyidejű megjelenése, amelyeknek látszólag egyike sem jelent egymagában előnyt. Az új tápanyagok hasznosítására való képesség evolúcióját anyagcsere-hálózatokban vizsgálva kimutattuk, hogy az egyetlen reakció hozzáadásával elérhető fenotípusok ugródeszkaként szolgálnak komplex anyagcsere-fenotípusok későbbi megjelenéséhez egy másik környezetben. Tehát az időben változó környezeti feltételek lehetővé teszik a tápanyagfelhasználási képességek lépésenkénti kibővítését anélkül, hogy feltételezni kellene nem adaptív köztes lépéseket.

Szappanos, B., Fritzemeier, J.C., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R.A., Lercher, M.J., Pál, C.*, **Papp, B.*** (2016) Adaptive evolution of complex innovations through stepwise metabolic niche expansion. *Nat Commun.* **7**:11607

Hivatkozások

Babu M, Arnold R, Bundalovic-Torma C, Gagarinova A, Wong KS, Kumar A, Stewart G, Samanfar B, Aoki H, Wagih O, et al. 2014. Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of protein complexes in Escherichia coli. PLoS Genet 10:e1004120.

Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ. 2004. Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15:3841-3862.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al. 2010. The genetic landscape of a cell. Science 327:425-431.

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. Science 353.

Dean AM, Thornton JW. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. Nat Rev Genet 8:675-688.

Edwards JS, Ibarra RU, Palsson BO. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125-130.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121.

Feist AM, Palsson BO. 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26:659-667.

Flachmann R, Kunz N, Seifert J, Gutlich M, Wientjes FJ, Laufer A, Gassen HG. 1988. Molecular biology of pyridine nucleotide biosynthesis in Escherichia coli. Cloning and characterization of quinolinate synthesis genes nadA and nadB. Eur J Biochem 175:221-228.

Harrison R, Papp B, Pal C, Oliver SG, Delneri D. 2007. Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci U S A 104:2307-2312.

Hartman JLt, Garvik B, Hartwell L. 2001. Principles for the buffering of genetic variation. Science 291:1001-1004.

Horlbeck MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, Adamson B, Chow ED, Kampmann M, Peterson TR, et al. 2018. Mapping the Genetic Landscape of Human Cells. Cell 174:953-967 e922.

Jensen RA. 1976. Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409-425.

Johannes TW, Zhao H. 2006. Directed evolution of enzymes and biosynthetic pathways. Curr Opin Microbiol 9:261-267.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27-30.

Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD. 2010. Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. Mol Syst Biol 6:436.

Notebaart RA, Kintses B, Feist AM, Papp B. 2018. Underground metabolism: network-level perspective and biotechnological potential. Curr Opin Biotechnol 49:108-114.

Notebaart RA, Szappanos B, Kintses B, Pal F, Gyorkei A, Bogos B, Lazar V, Spohn R, Csorgo B, Wagner A, et al. 2014. Network-level architecture and the evolutionary potential of underground metabolism. Proc Natl Acad Sci U S A 111:11762-11767.

Oberhardt MA, Palsson BO, Papin JA. 2009. Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320.

Papp B, Notebaart RA, Pal C. 2011. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet 12:591-602.

Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. 2007. Multicopy suppression underpins metabolic evolvability. Mol Biol Evol 24:2716-2722.

Price ND, Reed JL, Palsson BO. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886-897.

Schomburg I, Chang A, Schomburg D. 2002. BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30:47-49.

Snitkin ES, Dudley AM, Janse DM, Wong K, Church GM, Segre D. 2008. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. Genome Biol 9:R140.

Sommer MOA, Munck C, Toft-Kehler RV, Andersson DI. 2017. Prediction of antibiotic resistance: time for a new preclinical paradigm? Nat Rev Microbiol 15:689-696.

Szappanos B, Kovacs K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, Gelius-Dietrich G, Lercher MJ, Jelasity M, Myers CL, et al. 2011. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet 43:656-662.

Teusink B, Walsh MC, van Dam K, Westerhoff HV. 1998. The danger of metabolic pathways with turbo design. Trends Biochem Sci 23:162-169.

MTA Doctoral Thesis

Evolutionary genetics in the era of genome-scale modelling

Balázs Papp

Synthetic and Systems Biology Unit

Institute of Biochemistry

Biological Research Center, Eötvös Loránd Kutatási Hálózat

Szeged, 2021

I. Introduction

Many outstanding questions in evolutionary biology depend on the mutational effects that govern the complex relationship between genotype and phenotype. For example, why are most inactivating mutations have little phenotypic effects? And how do multiple mutations interact with each other to produce a novel phenotype? Resolving such issues requires an understanding of how genotypes map onto phenotypes. While genotype-phenotype maps have long been investigated at the level of individual proteins (Dean and Thornton 2007), analyses of larger gene networks lag behind. Developing quantitative frameworks to interrogate mutational effects in large cellular networks would be important for at least two fundamental reasons. First, such a framework would provide mechanistic insights into complex evolutionary phenomena, from the emergence of evolutionary novelties that hinge on multiple mutations to the evolution of minimized genomes. Second, it would transform evolutionary biology into a more predictive discipline by allowing specific predictions on the outcome of evolution. A predictive framework would give us a clue which genes are likely to be lost, to mutate, or change expression during evolution. Beyond catalyzing basic research, such a framework would also have practical relevance. Among others, it would allow forecasting evolutionary changes in pathogenic microbes (Sommer, et al. 2017) and inform the engineering of novel biosynthetic pathways (Johannes and Zhao 2006; Notebaart, et al. 2018).

Recent advances in systems biology provide an unprecedented opportunity to build computational models that map from mutations and environmental changes to phenotypes. These computational approaches rely on mathematical models of specific molecular systems and come in different flavors. These models range from detailed kinetic models of smaller metabolic systems (Teusink,

et al. 1998) and regulatory circuits (Chen, et al. 2004) to constraint-based models of genome-scale metabolic networks (Price, et al. 2004). Constraint-based metabolic models are especially appealing for studying the relationship between genotypes and phenotypes in large networks and have already provided valuable insights into the evolution of metabolic gene contents and phenotypes of microbial species (Feist and Palsson 2008). These models start from high-quality metabolic network reconstructions (Price, et al. 2004). These reconstructions are typically built through integrating genome annotation data, information from enzyme databases, such as KEGG (Kanehisa and Goto 2000) and BRENDA (Schomburg, et al. 2002)) and the primary literature and involve extensive manual curation. The network of biochemical reactions is then converted into a mathematical representation and analyzed using constrained-based methods (Price, et al. 2004). In particular, a widely used method termed flux balance analysis (FBA) calculates the optimal flow of metabolites through the network as a function of available nutrients in the environment. These predictions have been extensively tested and showed high agreement with empirical data (Edwards, et al. 2001; Snitkin, et al. 2008; Oberhardt, et al. 2009).

This thesis focuses on four major research topics, each of which employs genome-scale metabolic networks to address a long-standing issue in evolutionary genetics. First, how do mutations modulate each other's phenotypic effects, that is, how do genetic interactions arise at the mechanistic level? And how accurately can we computationally predict which gene pairs show a genetic interaction based on a detailed knowledge of the metabolic network? Second, can we predict the gene content of endosymbiotic bacteria that have highly reduced genomes? That is, can we predict which genes are lost and which are kept during millions of years of reductive genome evolution? Third, can we computationally predict the outcome and genetic basis of adaptation to

new environments? More specifically, how does the ability to utilize new nutrients arise from existing low-level enzymatic side activities? Finally, how do evolutionary novelties arise that demand the simultaneous acquisition of multiple mutations?

II. Genetic interactions in metabolic networks

The phenotypic effect of a mutation often depends on the presence of other mutations in the genome, a phenomenon termed genetic interaction or epistatic interaction. Genetic interactions are the key to understand the functional relationships between genes, the extent to which organisms tolerate deleterious mutations, as well as the underpinnings of complex genetic diseases. In the past decade, high-throughput studies have generated comprehensive maps of genetic interactions between genes in several organisms, including budding yeast (*Saccharomyces cerevisiae*) (Costanzo, et al. 2010; Costanzo, et al. 2016), *E. coli* (Babu, et al. 2014) and human cell lines (Horlbeck, et al. 2018). These works focused on loss-of-function mutations and revealed two main forms of genetic interactions: (i) *negative genetic interactions* (synthetic sick or lethal / aggravating) when two mutations enhance each other's harmful effects, potentially indicating functional compensation between them, and (ii) *positive genetic interactions* (antagonistic / diminishing) when a mutation has a smaller than expected deleterious effect in the presence of another deleterious mutation. However, despite the rapid accumulation of experimental data on genetic interactions, several questions remain open about the organization and mechanistic underpinnings of epistasis. In the past years, I contributed to four outstanding issues: (i) Why do the vast majority of genes show few genetic interactions, while a small number of 'hub' genes are highly connected in the genetic interaction network? (ii) Is it possible to computationally predict which specific gene pair would show a genetic interaction based on our knowledge of the

underlying biochemical network?, (iii) Can we make use of the discrepancies between empirical data and model predictions to automatically refine the metabolic model?, and (iv) Are genetic interactions environment dependent and if so, why?

To tackle the above questions, we applied an integrated systems biology approach by constructing a large-scale empirical genetic interaction map of yeast metabolism and integrating the data with a genome-scale metabolic network model. Our analyses yielded four major insights (Harrison, et al. 2007; Szappanos, et al. 2011):

1) The computational model successfully captured the high genetic interaction connectivity, for both positive and negative interactions, of genes that have a large contribution to fitness. Importantly, the genome-scale model incorporates information only on the stoichiometry of biochemical reactions and growth requirements of the cell without explicitly accounting for gene regulation and enzyme kinetic details. Therefore, this result suggests that genetic interaction hubs emerge from the structure of metabolic networks. Modelling also showed that hubs are driven by pleiotropic enzymes that participate in multiple biological processes by contributing to the biosynthesis of multiple key biomass precursors. As a result, the phenotypic impact of their loss can potentially be shaped by several other enzymes in the network, yielding numerous epistatic interactions.

2) By comparing genetic interaction data with a genome-scale model of metabolism, we provided the first large-scale assessment of our ability to predict individual genetic interactions using genome-scale metabolic models. Analysis of high-confidence experimental data across

~67,500 metabolic gene pairs uncovered a strong enrichment of *in vivo* interactions among computationally predicted ones (100-fold and 60-fold enrichment for negative and positive genetic interactions, respectively). However, the metabolic model fails to capture the majority of *in vivo* detected genetic interactions (97% and 89% of the negative and positive interactions, respectively; see Figure 1).

3) In principle, the large number of experimentally observed genetic interactions offers a rich source of information to modify the model in a data-driven way. To this end, we developed a machine learning method that automatically suggests modifications to the model that improve its ability to predict negative genetic interactions (Figure 1A). Overall, the method proposed several modifications that together improved the fit of the model to the data (i.e. 100–267% increase in recall and 44–59% increase in precision; see Figure 1B). Among the suggested modifications, we found the removal of the *de novo* NAD biosynthesis pathway starting from aspartate. This pathway is present in *E. coli* (Flachmann, et al. 1988), but was probably erroneously included in the yeast network. Indeed, a follow-up experiment confirmed that removing this pathway specifically allows the correct prediction of nicotinic acid auxotrophy of mutants affecting the kynurenine pathway.

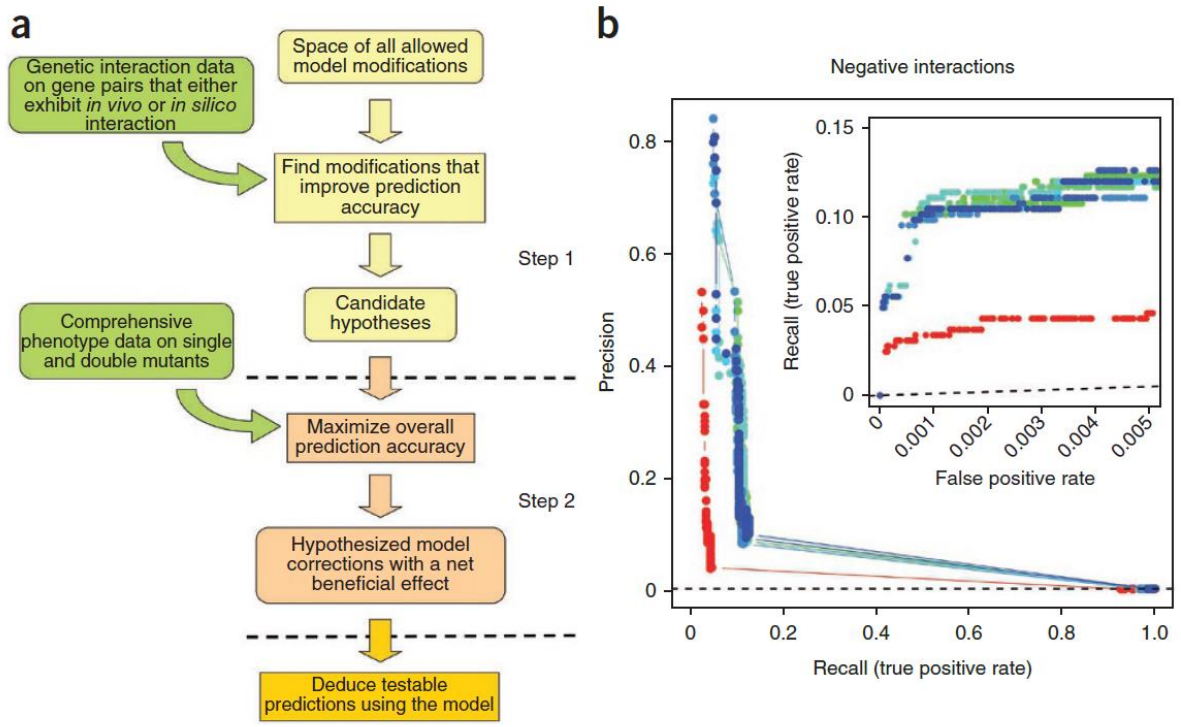


Figure 1. Automated refinement of the metabolic model. *(A) Workflow of the refinement algorithm. Because evaluating each model is computationally intensive (i.e. a large number of gene deletions should be simulated for each individual model), we employed a two-step procedure to make use of all available phenotypic data while maintaining computational feasibility. In the first step, we searched for models by evaluating a model on only those gene pairs that display either in vivo interaction or in silico interaction according to the original model. Because genetic interactions are very rare both in vivo and in silico, most gene pairs examined in this study show no interaction and omitting them significantly speeds up the exploration of the hypothesis space. In the second step, we defined a new, very restricted hypothesis space based on the most successful models from the first step, but searched for models that improve overall prediction accuracy as assessed by a comprehensive evaluation of each model in the population. (B) Impact of model refinement on prediction accuracy using 8 independent runs of the algorithm. Figure shows the congruency of the modified (blue to green) and original (red) models to the empirical genetic interaction data by both precision recall and partial ROC curves (inset). Dashed lines represent prediction accuracy expected by chance. Note that while the same dataset was used for both model refinement and evaluation in this plot, a cross-validation procedure confirmed significant model improvement (Szappanos, et al. 2011). Figure reproduced from (Szappanos, et al. 2011).*

4) Finally, by computationally analyzing how genetic interactions change across dozens of nutrient conditions, we found that genetic interactions often depend on the prevailing environments. The study focused on synthetic lethal interactions, which is an extreme form of negative genetic interactions where the double gene deletant shows a no-growth phenotype that is not displayed be either single deletion mutant. Synthetic lethal interactions are often thought to indicate functional redundancy between gene pairs, e.g. through alternative pathways or gene duplicates (Hartman, et al. 2001). Crucially, our work shows that many genes involved in synthetic interactions in one environment become essential in another environment, indicating that their redundancy is more apparent than real. More generally, the robustness of metabolic networks against genetic perturbations is likely to be a by-product of adaptation to survive in a large variety of nutrient conditions.

III. Predicting genome reduction

One of the central questions in the post genomic era is understanding which organisms have which genes. Typically, such inferences are drawn *a posteriori*, that is having discovered that an organism has a given gene we then construct hypothesizes about its ecology or biology. But is it possible to do the inverse and hence have an *a priori* and predictive theory for a genome? That is, can we take an organism's ecology and predict which genes it should have with any accuracy?

As first attempt to probe the feasibility of predicting long-term genomic evolution, we asked whether, given the genome of *E. coli*, we can predict the metabolism of *Buchnera*, an intracellular

symbiont with a heavily reduced genome, that was derived from *E. coli* (Figure 2). *Buchnera* have evolved from its free-living ancestors approximately 200 million years ago and lost 75% of their genes, reaching nearly minimal gene sets (~600 genes) needed to sustain life. We used a genome-scale model of *E. coli* metabolism, and setup the model to mimic the lifestyle of the endosymbiont based on available physiological evidence. Using a series of flux balance analysis simulations, we considered sequentially the fate of randomly selected gene deletions and asked, given the ecology of *Buchnera*, whether these would be effectively neutral or not. Repeatedly simulating successive gene loss events until no further genes could be deleted without impairing *in silico* growth, we obtained a set of minimal networks (Figure 2). Remarkably, comparison of the gene complements of these in silico minimal networks with three then available *B. aphidicola* genomes revealed that gene presence / absence can be predicted with high accuracy (area under the ROC curve =0.794 – 0.802, depending on the endosymbiont genome). Thus, 200 million years of reductive genome evolution is surprisingly well predictable based on knowledge of the organism's distant ancestor and its current lifestyle.

Even closely related *Buchnera* strains vary in their gene complements. In principle, such variation in reductive genome evolution may reflect both differences in selective forces (ecology) and chance events, yielding differences in the order of gene deletions and hence a choice between alternative cellular pathways. Comparing the minimal networks from repeated simulations, we indeed found support for partially different evolutionary outcomes that arise from chance events.

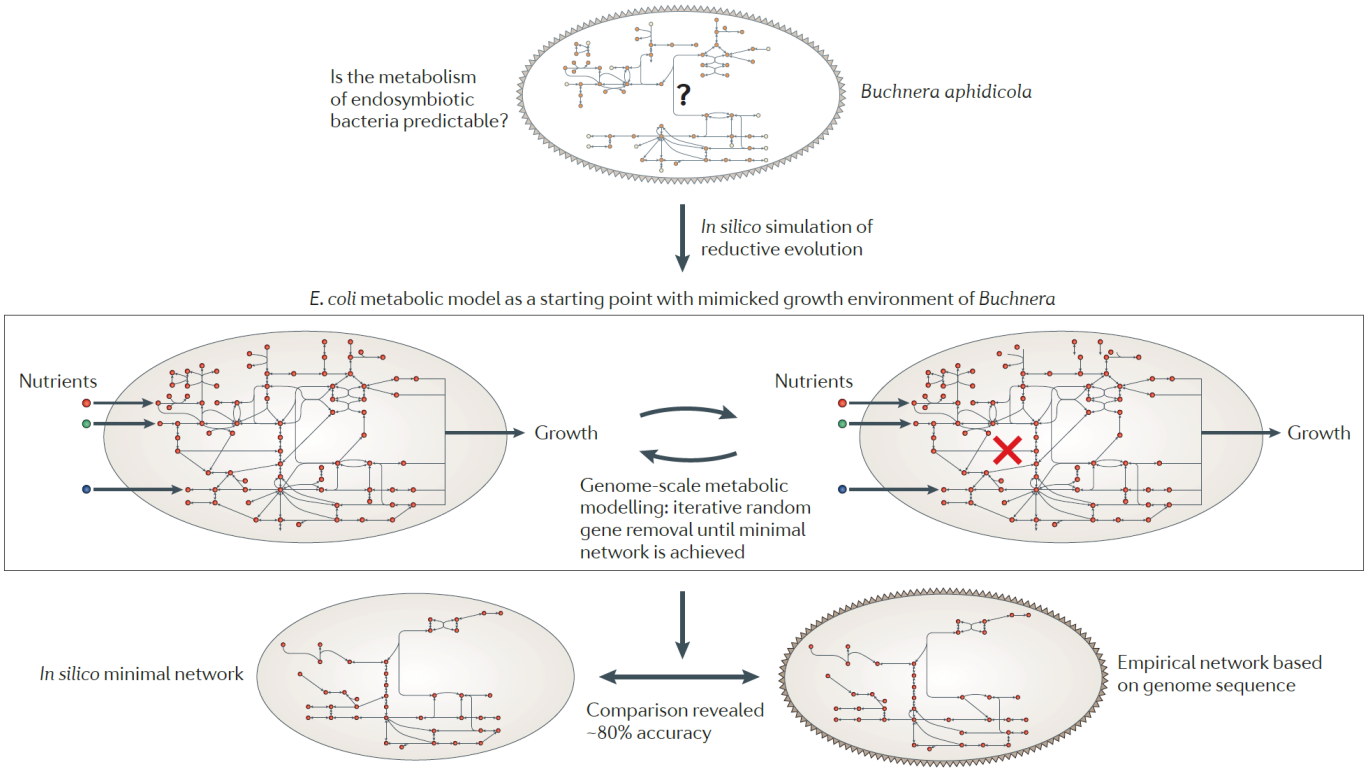


Figure 2. A genome-scale metabolic modelling approach to predict genome reduction in endosymbiotic bacteria. *The computational approach uses the genome-scale metabolic model of present day E. coli as a proxy for the free-living ancestor of the endosymbiont B. aphidicola and mimicks the lifestyle of the present day endosymbiont to predict the impact of individual gene loss events along an evolutionary trajectory. During the evolutionary simulations, minimal metabolic networks were generated by repeatedly simulating gene loss events (red cross) until no further genes could be removed without impairing in silico growth. The computationally predicted minimal networks showed high overlap with the metabolic gene contents of real Buchnera (lower panel). Figure reproduced from (Papp, et al. 2011).*

IV. Underground metabolism and the predictability of adaptive evolution

Understanding how new molecular pathways emerge during adaptation is one of the central issues in evolutionary and systems biology. In the most well-understood networks, small-molecule metabolism, the prevailing paradigm is that evolution capitalizes on the weak side activities of pre-

existing enzymes (Jensen 1976). While much has been learned about the biochemical mechanisms of these activities in a few well-studied enzymes, the extent to which underground reactions provide novelties in the context of the entire cellular system remains completely unexplored.

To address the above gap, we applied a systems-level approach that provides insights into the architecture of underground metabolism and also enables the prediction of the role of underground activities in adaptation to nutrient conditions (Notebaart, et al. 2014). In brief, we focused on *Escherichia coli*, which is the most comprehensively characterized organism in terms of enzymatic activities. Based on databases and literature survey, we therefore built an *in silico* underground metabolic network reconstruction of *Escherichia coli* and integrated it with the native genome-scale metabolic model of this organism (Feist, et al. 2007). Overall, we included 262 underground reactions and 277 metabolites that are not present in the native network. The resulting reconstruction is the first comprehensive computational model of the underground metabolism of any organism. We next explored the evolutionary potential of underground metabolism as follows:

1) Analysis of the structure of the network showed that a substantial proportion of underground reactions can form new pathways with high potential biological relevance.

2) We conducted an *in silico* survey to characterize the evolutionary potential of *E. coli* to adapt to hundreds of novel nutrient conditions. We computationally predicted the impact of adding underground reactions to the native network on maximum growth across a variety of environments. Because we were interested in the potential to evolve towards new nutrients, we assumed that all underground reactions can be utilized (i.e. there are no enzyme kinetic or regulatory constraints). The analysis revealed dozens of cases where underground reactions

allow or improve growth in previously uncharacterized growth environments when their activity is increased. Most of the growth improvements were conferred by single underground reactions, with only a minority requiring multiple reactions simultaneously (see (Notebaart, et al. 2014)).

3) We also performed a high-throughput experimental screen to estimate the potential of underground activities in adaptation to new nutrient sources. We carried out a genome-wide gene overexpression screen in *E. coli* and measured growth under 194 carbon sources (Patrick, et al. 2007; Kim, et al. 2010). Overall, we identified 17 genes that improved growth upon overexpression in at least one of 17 specific carbon sources. Out of the 17 genes, 11 encoded enzymes and 9 of these had known underground activities. Together, these analyses strongly support the notion that evolution can capitalize on underground reactions both to enhance growth in existing environments and to exploit completely new nutrient sources.

4) Can we predict the genetic basis of evolutionary adaptation to new environments? Comparison of the computational predictions with the genome-wide overexpression experiment showed a remarkable agreement. Specifically, the computational model successfully predicted 44% of the carbon sources on which amplification of an enzyme conferred or improved growth, an overlap that is statistically highly significant ($P<10^{-13}$). These results demonstrate that it is possible to predict the genetic basis of evolution towards new nutrient environments based on a detailed knowledge of an organism's underground metabolism.

V. Simple paths to complex adaptations

Explaining the origin of evolutionary novelties remains a central challenge in evolutionary biology. Traits that require the simultaneous emergence of multiple mutations, none of which seemingly confer a benefit individually, pose an especially daunting challenge for evolutionists. Such traits are often referred to as complex adaptations and might be difficult to evolve, not because of physical or chemical constraints, but because of the dynamics of how mutations spread in the population.

We proposed a conceptually simple model to resolve the paradox of complex adaptation. This scenario is closely related to the notion of pre-adaptation and purely relies on the successive accumulation of beneficial mutations. In brief, temporally varying environmental conditions select for single adaptive mutations that, as a by-product, serve as stepping stones towards the establishment of more complex phenotypes. Thus, complex adaptations can be accelerated in dynamically changing environments. By studying how novel nutrient utilization phenotypes can be established in a bacterial metabolic network by adding new enzymatic reactions to it, we provided three lines of evidence in support of this scenario:

1) We first studied *in silico* the expansion of the *E. coli* metabolic network to utilize novel nutrients. Our computational analyses revealed that new complex pathways can evolve via the successive acquisition of single biochemical reactions that each confer a benefit under specific

environmental conditions.

2) Second, the varying environment model predicts that gain of new metabolic genes should occur in a defined order. In particular, if two enzymes are jointly required to support growth in a novel environment, and one enzyme confers a benefit on its own in another environment, then the latter enzyme should be gained earlier. Indeed, phylogenetic analysis of the evolutionary history of gene gain events in 943 bacteria supported this prediction.

3) Last, we carried out a laboratory evolution study to adapt *E. coli* to two novel carbon sources and showed that evolving the ability to grow on one of them facilitated subsequent adaptation to the other.

Taken together, the above results demonstrate that complex metabolic adaptations can evolve through adaptive intermediate mutations by stepwise expansion of nutrient utilization capabilities. This conclusion represents an important conceptual advance as there is no need to invoke the slow process of accumulating neutral intermediate mutations.

VI. Summary of key results

This thesis is based on a series of publications that utilize computational systems biology modelling of metabolic networks to address several outstanding issues in evolutionary genetics. We reached the following major results, as published in the highlighted papers:

1) We probed the limits of predicting genetic interactions using genome-scale metabolic networks and showed that genetic interaction hubs are highly predictable, but individual genetic interactions are often missed by these computational models. Building on these discrepancies, we developed a machine learning method that refines the metabolic network model based on large-scale genetic interaction data.

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., Andrews, B.J., Boone, C., Oliver, S.G., Pál, C., **Papp, B.** (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genetics* **43:** 656

2) We showed that synthetic lethal genetic interactions often depend on the prevailing environments. Importantly, this is often caused by one or both genes of the synthetically interacting pairs becoming essential upon environmental change, indicating that the two genes are only partly redundant.

*Harrison, R., ***Papp, B.**, Pál, C., Oliver, S.G., Delneri, D. (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A.* **104**: 2307-12.

3) By simulating the repeated loss of non-essential genes in a genome-scale metabolic model of *E. coli*, we showed that it is possible to predict the highly reduced gene content of closely related endosymbiotic bacteria that diverged ~200 million years ago.

*Pál, C., ***Papp, B.**, Lercher, M.J., Csermely, P., Oliver, S.G. and Hurst, L.D. (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**: 667-70.

4) We reconstructed a comprehensive network of known enzyme side activities (i.e. underground reactions) in *E. coli*, which is the first such reconstruction in any organism. By combining computational simulations and a high-throughput experimental survey across hundreds of nutrient environments, we predicted and confirmed new environments where enhanced activity of underground reactions confer growth. Our results demonstrate that the genetic basis of evolutionary adaptations via underground metabolism is computationally predictable.

Notebaart, R.A.*, Szappanos, B., Kintses, B., Pál, F., Györkei, A., Bogos, B., Lázár, V., Spohn, R., Csörgő, B., Wagner, A., Ruppin, E., Pál, C.*, **Papp, B.*** (2014) Network-level architecture and the evolutionary potential of underground metabolism. *Proc Natl Acad Sci U S A.* **111**: 11762-11767.

5) We proposed a new model to resolve the paradox of complex adaptations, i.e. new traits that require the simultaneous emergence of multiple mutations, none of which seemingly confer a benefit individually. By studying the evolution of new nutrient utilization capabilities in metabolic networks, we showed that phenotypes accessible through the addition of a single reaction serve as stepping stones towards the later establishment of complex metabolic

phenotypes in another environment. Thus, temporally varying environmental conditions enable the step-by-step expansion of nutrient utilization capacities without the need to invoke non-adaptive processes.

Szappanos, B., Fritzemeier, J.C., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R.A., Lercher, M.J., Pál, C.*, **Papp, B.*** (2016) Adaptive evolution of complex innovations through stepwise metabolic niche expansion. *Nat Commun.* **7**:11607

References

Babu M, Arnold R, Bundalovic-Torma C, Gagarinova A, Wong KS, Kumar A, Stewart G, Samanfar B, Aoki H, Wagih O, et al. 2014. Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of protein complexes in Escherichia coli. PLoS Genet 10:e1004120.

Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ. 2004. Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15:3841-3862.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al. 2010. The genetic landscape of a cell. Science 327:425-431.

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. Science 353.

Dean AM, Thornton JW. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. Nat Rev Genet 8:675-688.

Edwards JS, Ibarra RU, Palsson BO. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125-130.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121.

Feist AM, Palsson BO. 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26:659-667.

Flachmann R, Kunz N, Seifert J, Gutlich M, Wientjes FJ, Laufer A, Gassen HG. 1988. Molecular biology of pyridine nucleotide biosynthesis in Escherichia coli. Cloning and characterization of quinolinate synthesis genes nadA and nadB. Eur J Biochem 175:221-228.

Harrison R, Papp B, Pal C, Oliver SG, Delneri D. 2007. Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci U S A 104:2307-2312.

Hartman JLt, Garvik B, Hartwell L. 2001. Principles for the buffering of genetic variation. Science 291:1001-1004.

Horlbeck MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, Adamson B, Chow ED, Kampmann M, Peterson TR, et al. 2018. Mapping the Genetic Landscape of Human Cells. Cell 174:953-967 e922.

Jensen RA. 1976. Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409-425.

Johannes TW, Zhao H. 2006. Directed evolution of enzymes and biosynthetic pathways. Curr Opin Microbiol 9:261-267.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27-30.

Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD. 2010. Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. Mol Syst Biol 6:436.

Notebaart RA, Kintses B, Feist AM, Papp B. 2018. Underground metabolism: network-level perspective and biotechnological potential. Curr Opin Biotechnol 49:108-114.

Notebaart RA, Szappanos B, Kintses B, Pal F, Gyorkei A, Bogos B, Lazar V, Spohn R, Csorgo B, Wagner A, et al. 2014. Network-level architecture and the evolutionary potential of underground metabolism. Proc Natl Acad Sci U S A 111:11762-11767.

Oberhardt MA, Palsson BO, Papin JA. 2009. Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320.

Papp B, Notebaart RA, Pal C. 2011. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet 12:591-602.

Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. 2007. Multicopy suppression underpins metabolic evolvability. Mol Biol Evol 24:2716-2722.

Price ND, Reed JL, Palsson BO. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886-897.

Schomburg I, Chang A, Schomburg D. 2002. BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30:47-49.

Snitkin ES, Dudley AM, Janse DM, Wong K, Church GM, Segre D. 2008. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. Genome Biol 9:R140.

Sommer MOA, Munck C, Toft-Kehler RV, Andersson DI. 2017. Prediction of antibiotic resistance: time for a new preclinical paradigm? Nat Rev Microbiol 15:689-696.

Szappanos B, Kovacs K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, Gelius-Dietrich G, Lercher MJ, Jelasity M, Myers CL, et al. 2011. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet 43:656-662.

Teusink B, Walsh MC, van Dam K, Westerhoff HV. 1998. The danger of metabolic pathways with turbo design. Trends Biochem Sci 23:162-169.