The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2415-12

Winter School on Quantitative Systems Biology

26 November - 7 December, 2011

Phenotypic constraints drive the architecture of biological networks

A. Samal
Institute for Systems Biology
Seattle
USA

Phenotypic constraints drive the architecture of biological networks

Areejit Samal

Price Lab

Institute for Systems Biology

Seattle USA

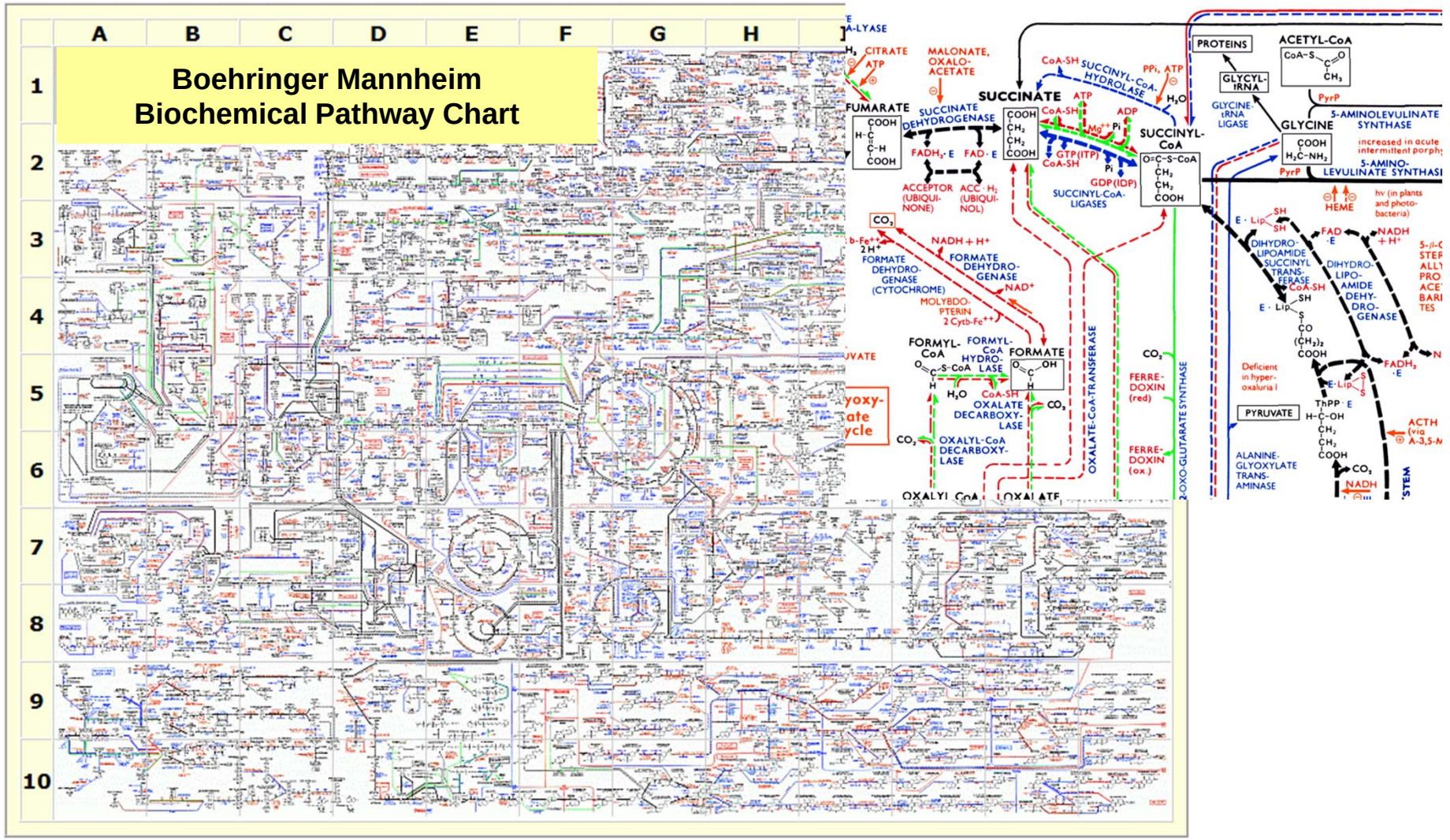
Work done @

CNRS, LPTMS, Univ Paris-Sud 11, Orsay, France

&

Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

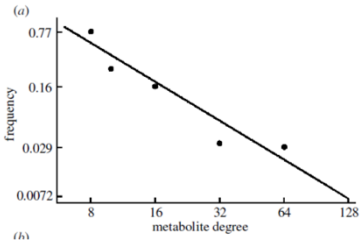
Metabolic network



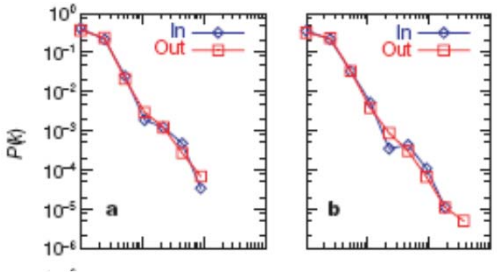
Metabolic network consists the set of biochemical reactions that convert nutrient molecules into key molecules required for growth and maintenance of the cell.

Large-scale structure of metabolic networks

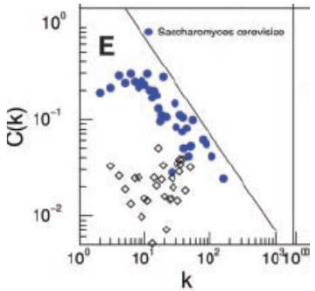
Small-world, Scale-free and Hierarchical Organization



Wagner & Fell (2001)

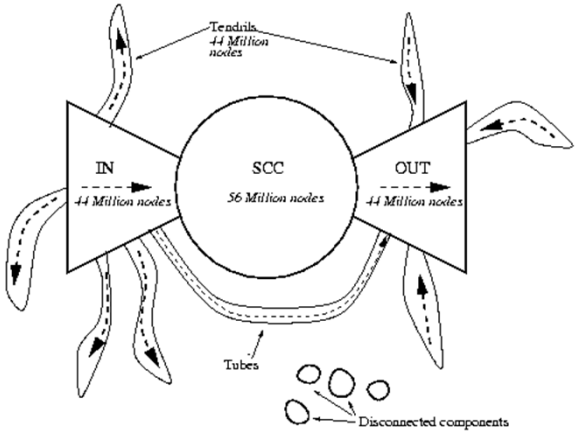


Jeong et al (2000)



Ravasz et al (2002)

Small Average Path Length, High Local Clustering, Power-law degree distribution



Ma and Zeng (2003);
Csete and Doyle (2004)

Bow-tie architecture

Directed graph with a giant component of strongly connected nodes along with associated IN and OUT component.

Bow-tie architecture of the metabolism is similar to that found in the WWW by Broder et al.

Large-scale structure of metabolic networks is very different from random networks and is similar in many respects to real world networks.

Are the observed structural features of a metabolic network ‘unusual’ or ‘atypical’?

- *Extremely popular* to study network measures such as degree distribution, clustering coefficient, assortativity, motifs, etc. in biological networks.

- *But mere observation* of scale-free, small-world or other striking structural features in biological networks do not mean these properties are *non-obvious* or *atypical* features of cellular networks.

- A *distinction* needs to be made between *observation* of these structural properties in real networks and our understanding of the *generative principles* that may have led to these properties in real networks.

Questions of Interest

- Are the observed structural properties of biological networks *frozen accidents*?
- What are the *adaptive* features of a network?
- Which features of a biological network are *under selection* and which are mere *byproducts* of selection on other traits?

To answer these questions adequately, one needs *proper controls* or *null models* of biological networks.

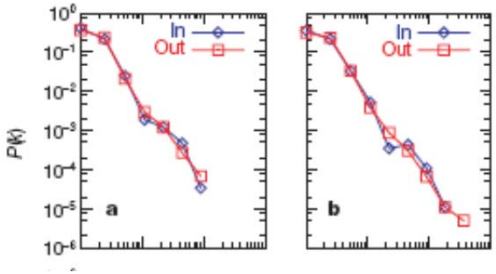
- *Current controls* are *not well posed* for biological networks, especially, *metabolic networks*.
- Furthermore, *proper controls* should account for *biological function.*

What are the *non-obvious* or *atypical* properties of metabolic networks once *biological function* is accounted in the control or null model?

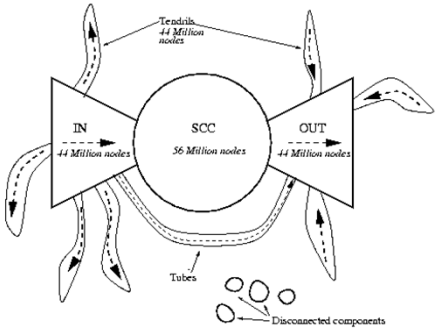
Is the large-scale structure of *E. coli* metabolic network atypical?

We decided to compare the following structural properties of *E. coli* metabolic network with those in randomized metabolic networks:

- Metabolite Degree distribution
- Clustering Coefficient
- Average Path Length
- P_c: Probability that a path exists between two nodes in the directed graph
- Largest strongly component (LSC) and the union of LSC, IN and OUT components



Scale Free and Small World
Ref: Jeong et al (2000); Wagner & Fell (2001)

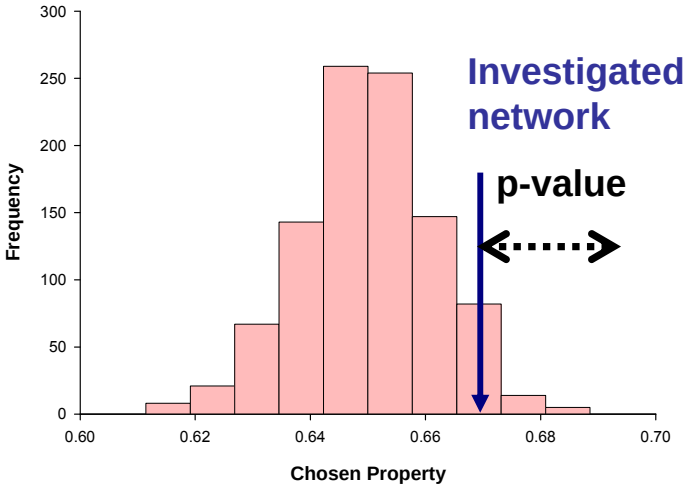


Bow-tie architecture of the Internet and metabolism
Ref: Broder et al (1999); Ma and Zeng (2003)

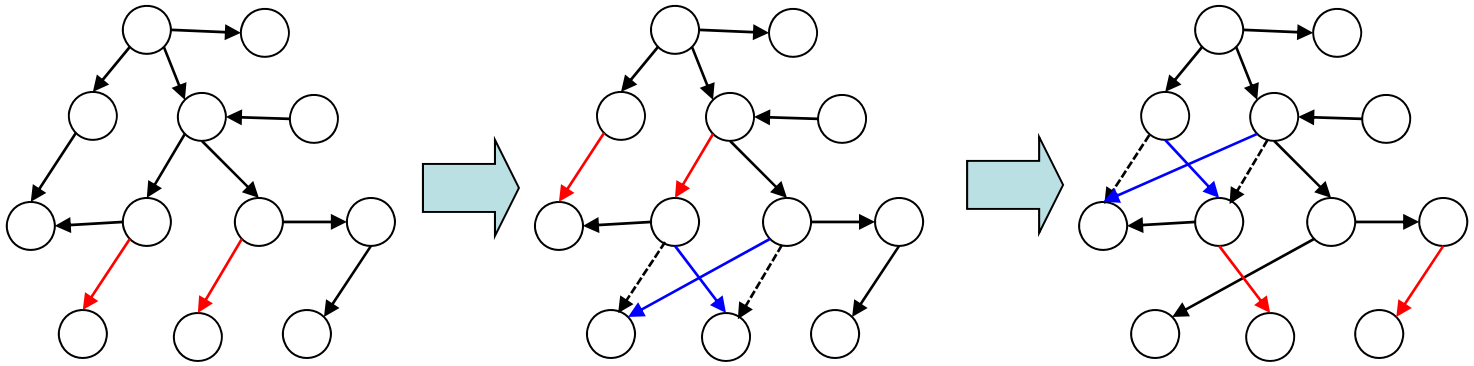
However, the widely used null model to generate randomized metabolic networks is not well posed to answer this question!

Edge-randomization algorithm: widely-used null model

1) Measure the 'chosen property in the investigated real network.
2) Generate randomized networks with structure similar to the investigated real network using edge-randomization.
3) Use the distribution of the 'chosen property' for the randomized networks to estimate a p-value.



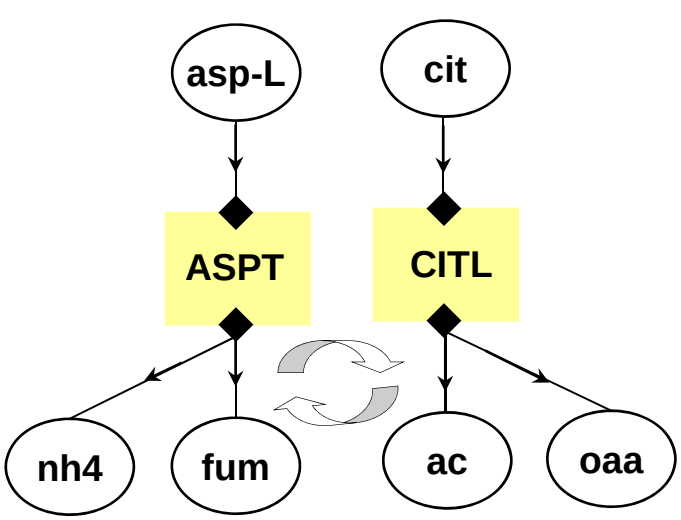
BUT this null model does not account for phenotypic or functional constraints central to cellular networks!

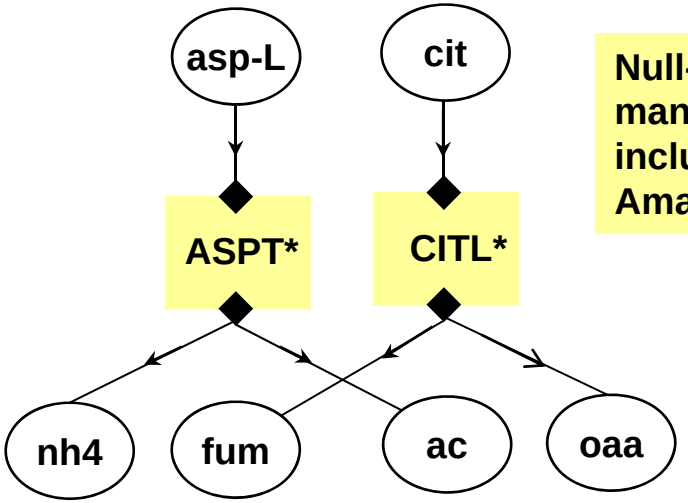


Investigated Network | After 1 exchange | After 2 exchanges

Reference:
Milo *et al* (2002,2004); Maslov & Sneppen (2002)

Edge-randomization: unsuitable for metabolic networks





Null-model used in many studies including: Guimera & Amaral Nature (2005)

ASPT: asp-L → fum + nh4
CITL: cit → oaa + ac

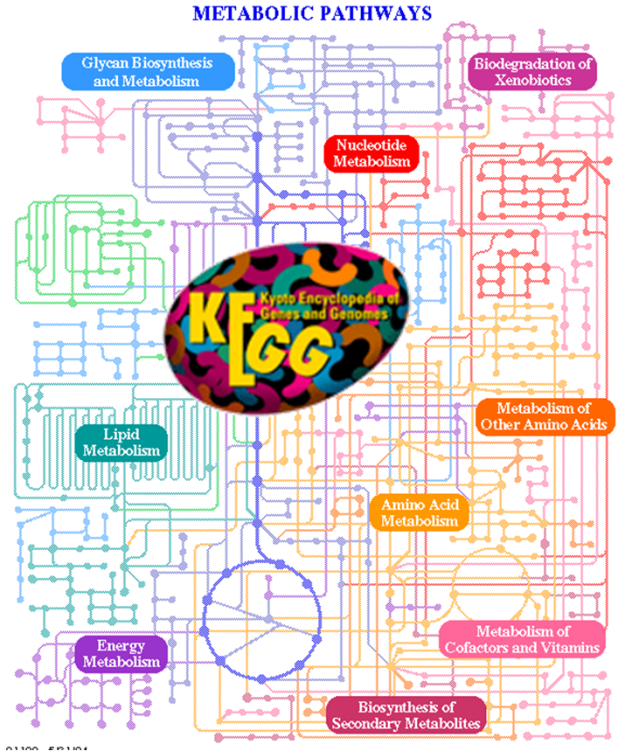
ASPT*: asp-L → ac + nh4
CITL*: cit → oaa + fum

Preserves degree of each node in the network but generates fictitious reactions that violate mass, charge and atomic balance satisfied by real chemical reactions!!
Note that *fum* has 4 carbon atoms while *ac* has 2 carbon atoms in the example shown.

Biochemically meaningless randomization inappropriate for metabolic networks

We decided to develop a proper null model for metabolic networks that accounts for basic biochemical and functional constraints.

Framework: Global reaction set



KEGG Database + *E. coli* iJR904

Problem 1: Edge-randomization generates fictitious reactions which violate mass and atomic balance.

Solution: We decided to overcome this problem by limiting the set of reactions in random metabolic networks to those within KEGG database.

➢We have use a curated database of **5870 mass balanced reactions** derived from KEGG by Rodrigues and Wagner (2009).

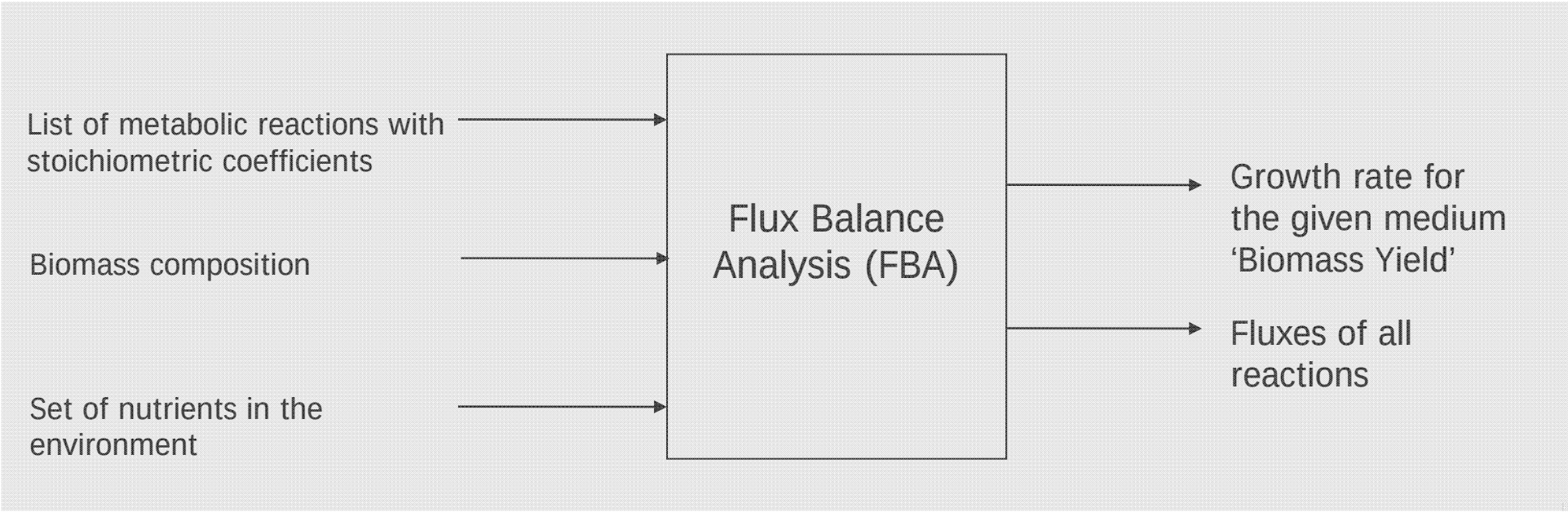
➢*E. coli* metabolic network iJR904 is a subset of reactions in KEGG with *n*=931 reactions. Number of possible networks with *n*=931 reactions like *E. coli* within KEGG is:

$$\binom{5870}{931} \sim 10^{1113}$$

which is > 10^{78} (estimated number of atoms in universe!)

➢ One can implement Flux Balance Analysis (FBA) within this database unlike KEGG.

Constraint-based Flux Balance Analysis (FBA)



Advantages

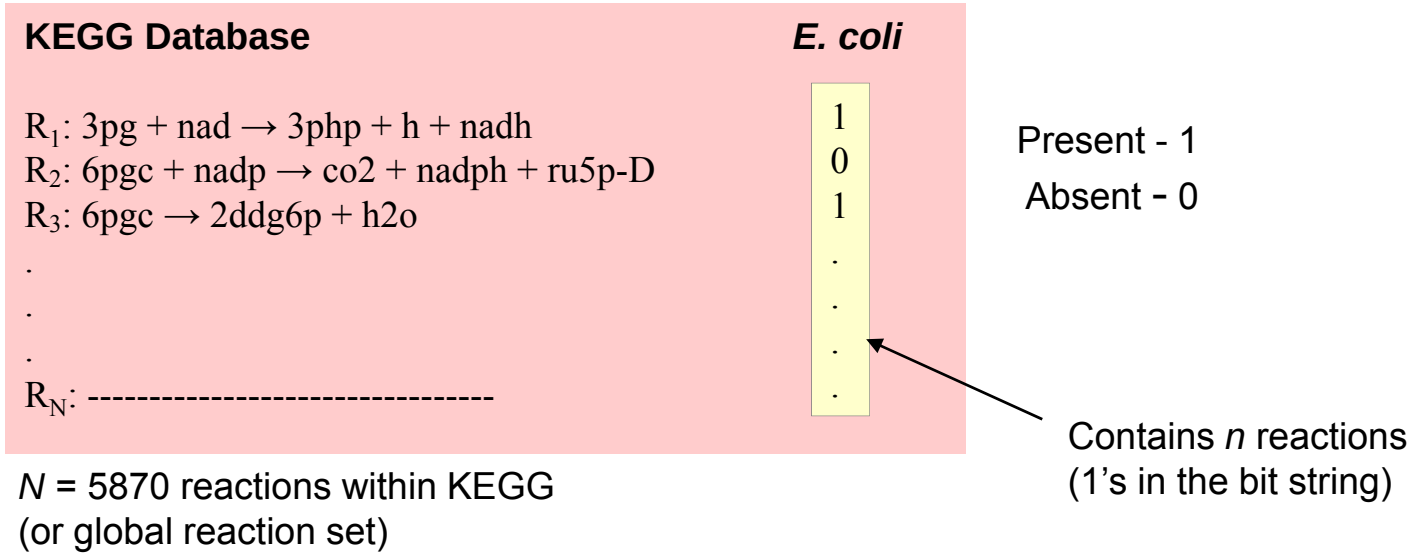
FBA does not require enzyme kinetic information which is not known for most reactions.

Disadvantages

FBA cannot predict internal metabolite concentrations and is restricted to steady states.
Basic models do not account for metabolic regulation.

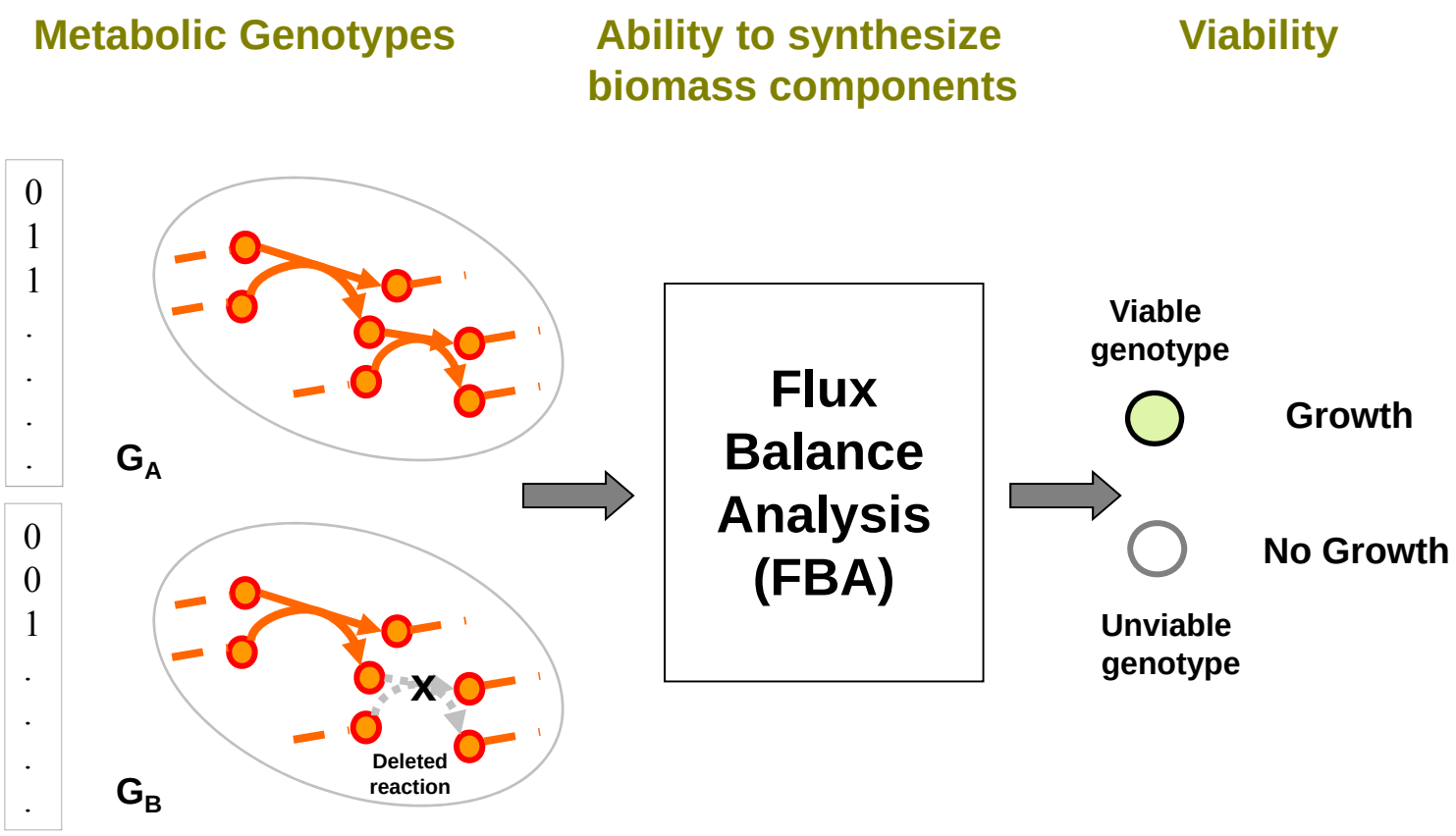
Reference: Varma and Palsson, Biotechnology (1994); Price et al (2004)

Bit string representation of metabolic network



The *E. coli* metabolic network or any *random* network of reactions within KEGG can be represented as a bit string of length *N* with exactly *r* entries equal to 1 where *n* is the number of reactions in *E. coli.*

Definition of Growth Phenotype



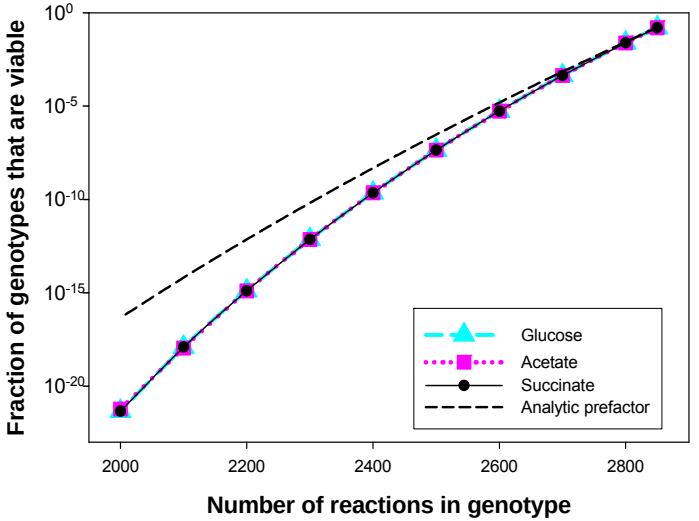
Bit string and equivalent network representation of genotypes

Fraction of networks satisfying functional constraints

Question: What fraction of possible metabolic networks within KEGG with exactly *n* reactions can grow under glucose minimal media like *E. coli*?

Answer:
To estimate this fraction:

- Generate random networks within KEGG with exactly *n* reactions.
- Determine the fraction of networks that can grow under glucose minimal media using FBA.

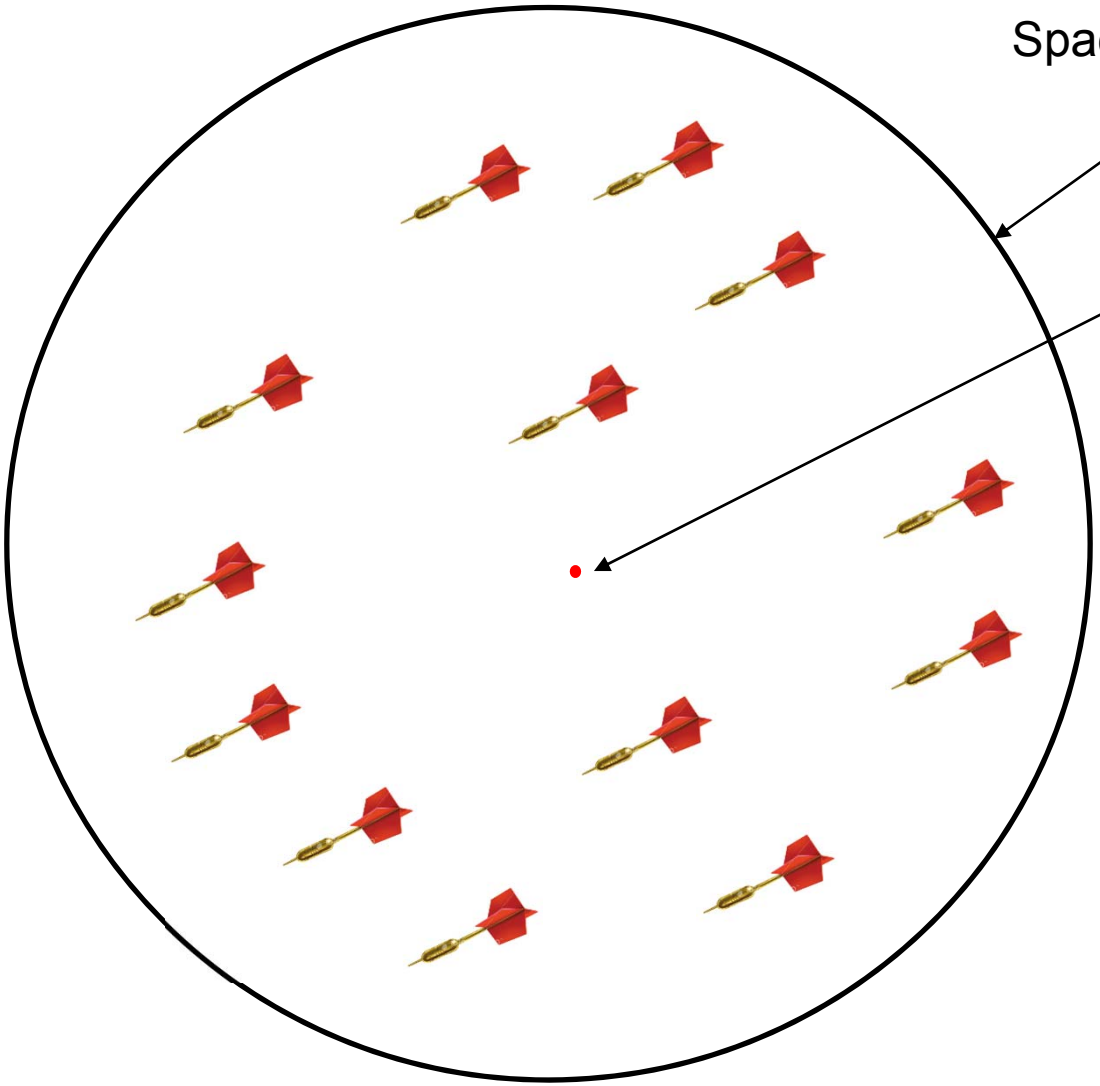


We estimate the fraction of possible networks with *n*=931 reactions like *E. coli* that have this phenotype is: ~ 10^{-50}

Only a tiny fraction of possible networks satisfy functional constraints!

Reference: Samal et al., BMC Systems Biology (2010)

Necessity for Markov Chain Monte Carlo (MCMC) Sampling



Space of possible networks ~ 10^{1113}

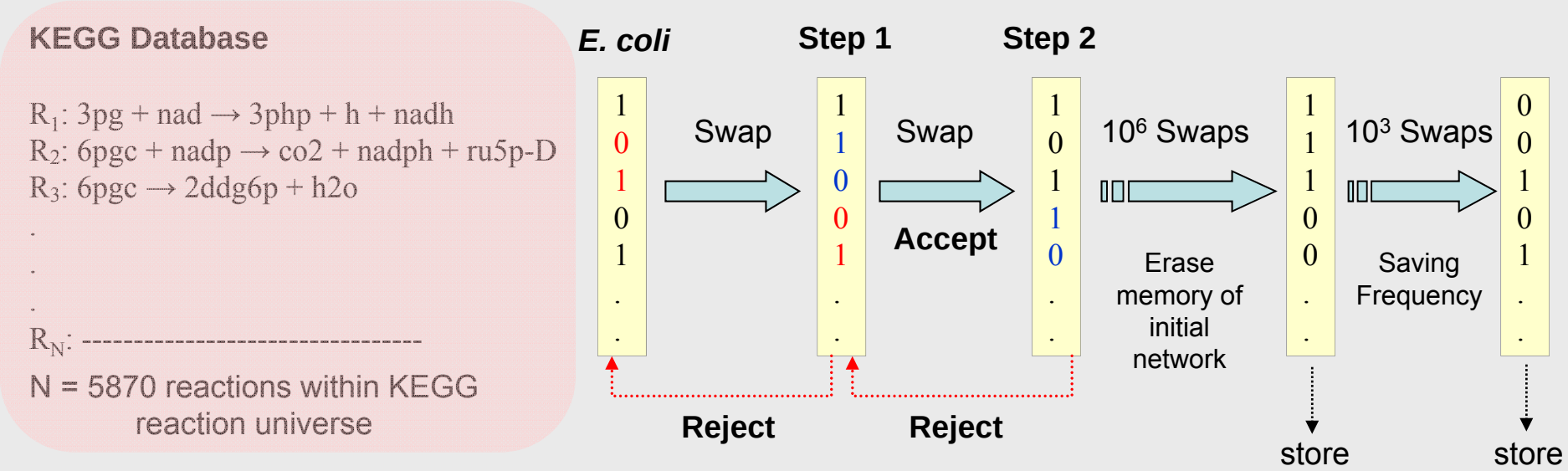
Space of networks with desired phenotype ~ 10^{1050}

We wish to uniformly sample the space of viable metabolic networks within KEGG having the same number of reactions as *E. coli.*

But one cannot sample the viable space by generating random bit strings followed by a test for phenotype.

We resort to Markov Chain Monte Carlo (MCMC) method to sample the space of networks satisfying functional constraints.

MCMC sampling of metabolic networks with growth phenotype



Accept/Reject Criterion of reaction swap:

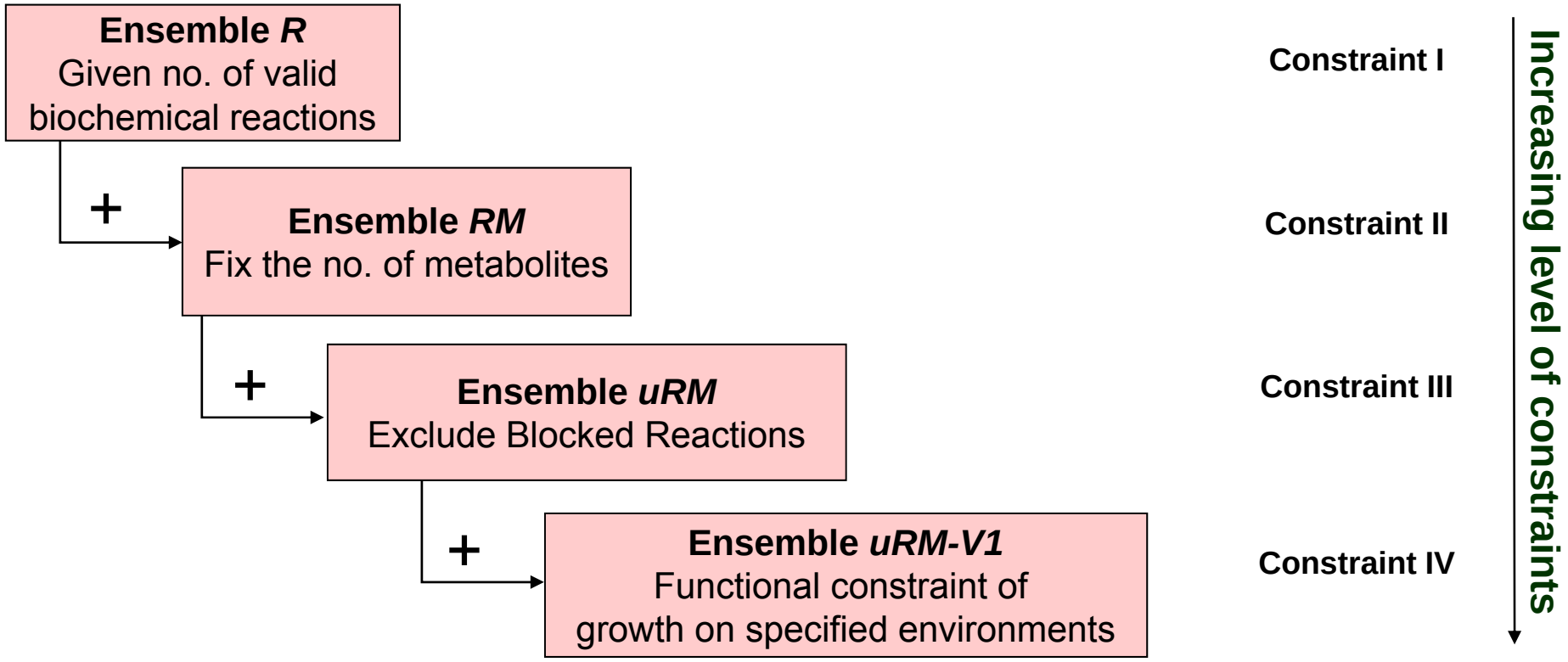
Accept if

(a) No. of metabolites in the new network is less than or equal to *E. coli*

(b) New network is able to grow on the specified environment(s)

Randomizing metabolic networks

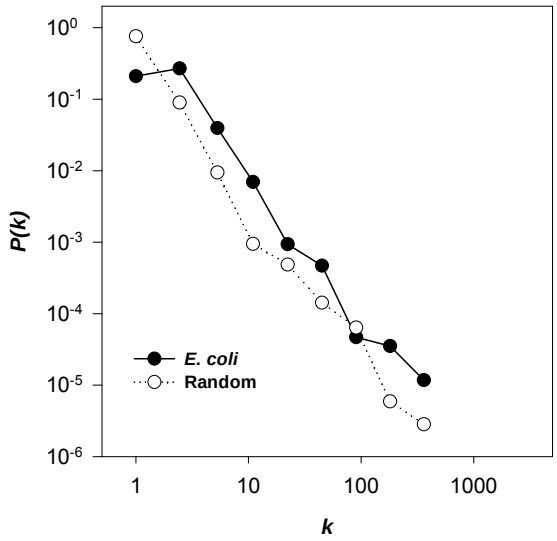
We have developed a new method using Markov Chain Monte Carlo (MCMC) sampling and Flux Balance Analysis (FBA) to generate meaningful randomized ensembles for metabolic networks by successively imposing constraints.



Reference: Samal et al., BMC Systems Biology (2010); Samal and Martin, PLoS ONE (2011)

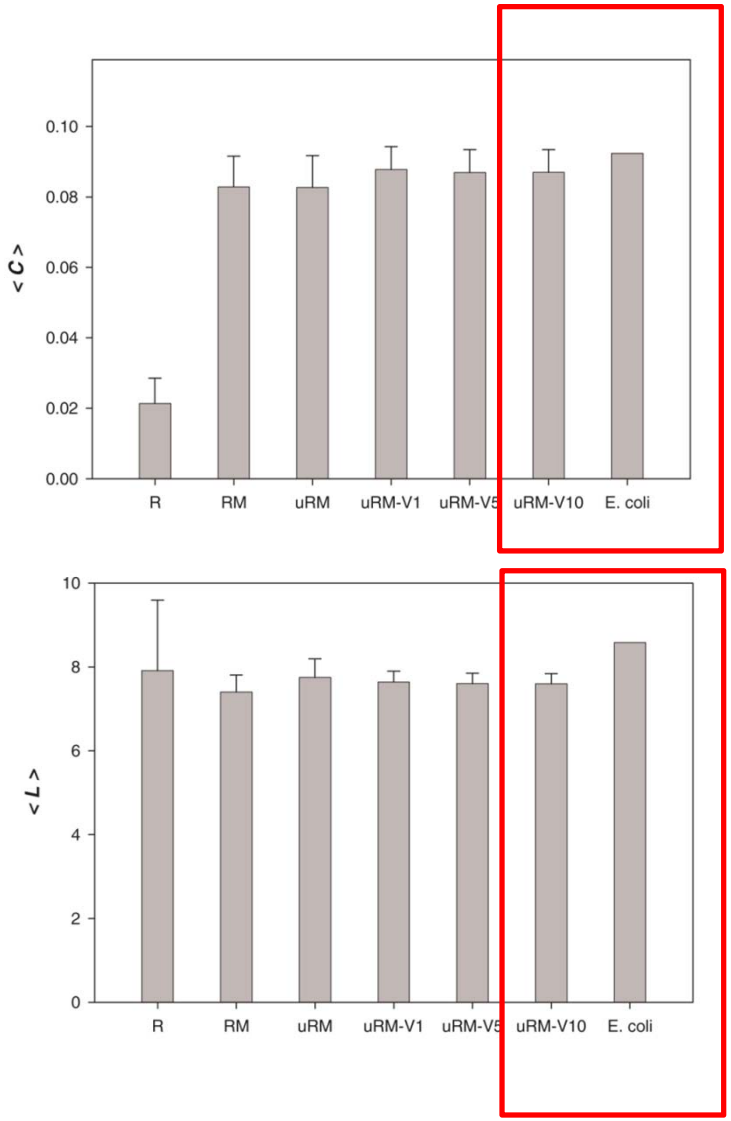
Metabolite degree distribution

Degree of a metabolite is the number of reactions in which the metabolite participates in the network.



Sampled networks are scale-free like *E. coli* !

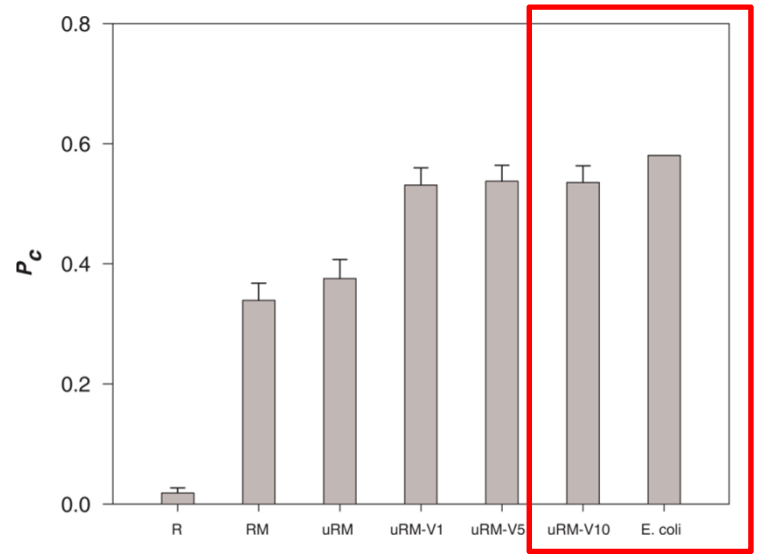
Clustering coefficient and Average Path Length

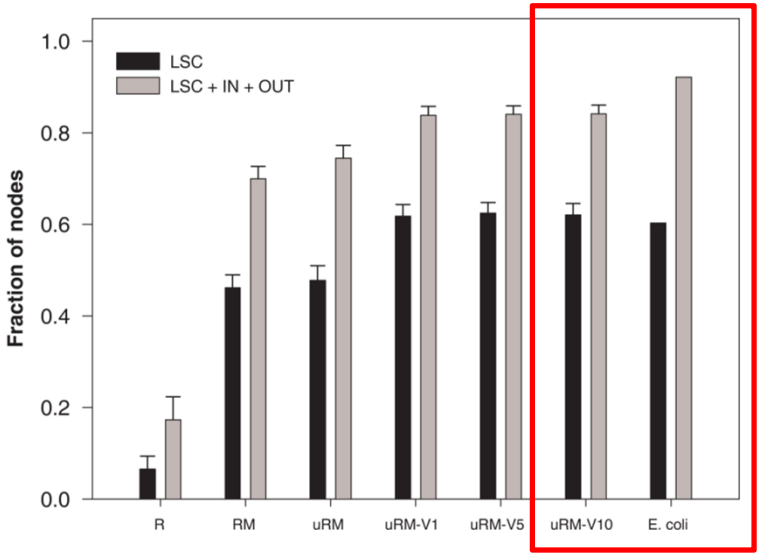


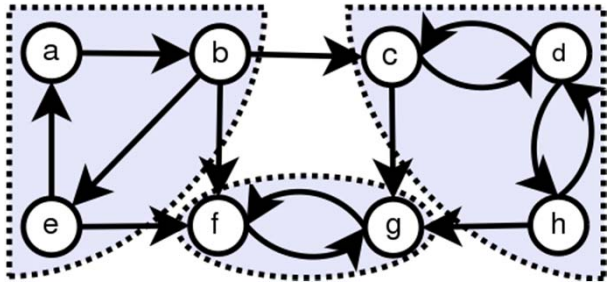
R: Fixed no. of reactions
RM: Fixed no. of reactions and metabolites
uRM: Fixed no. of unblocked reactions and metabolites
uRM-V1: Fixed no. of unblocked reactions and metabolites and viable in one environment
uRM-V5: Fixed no. of unblocked reactions and metabolites and viable in five environments
uRM-V10: Fixed no. of unblocked reactions and metabolites and viable in ten environments

Sampled networks have small-world and hierarchical architecture!

Probability that a path exists between two nodes and Size of largest strong component





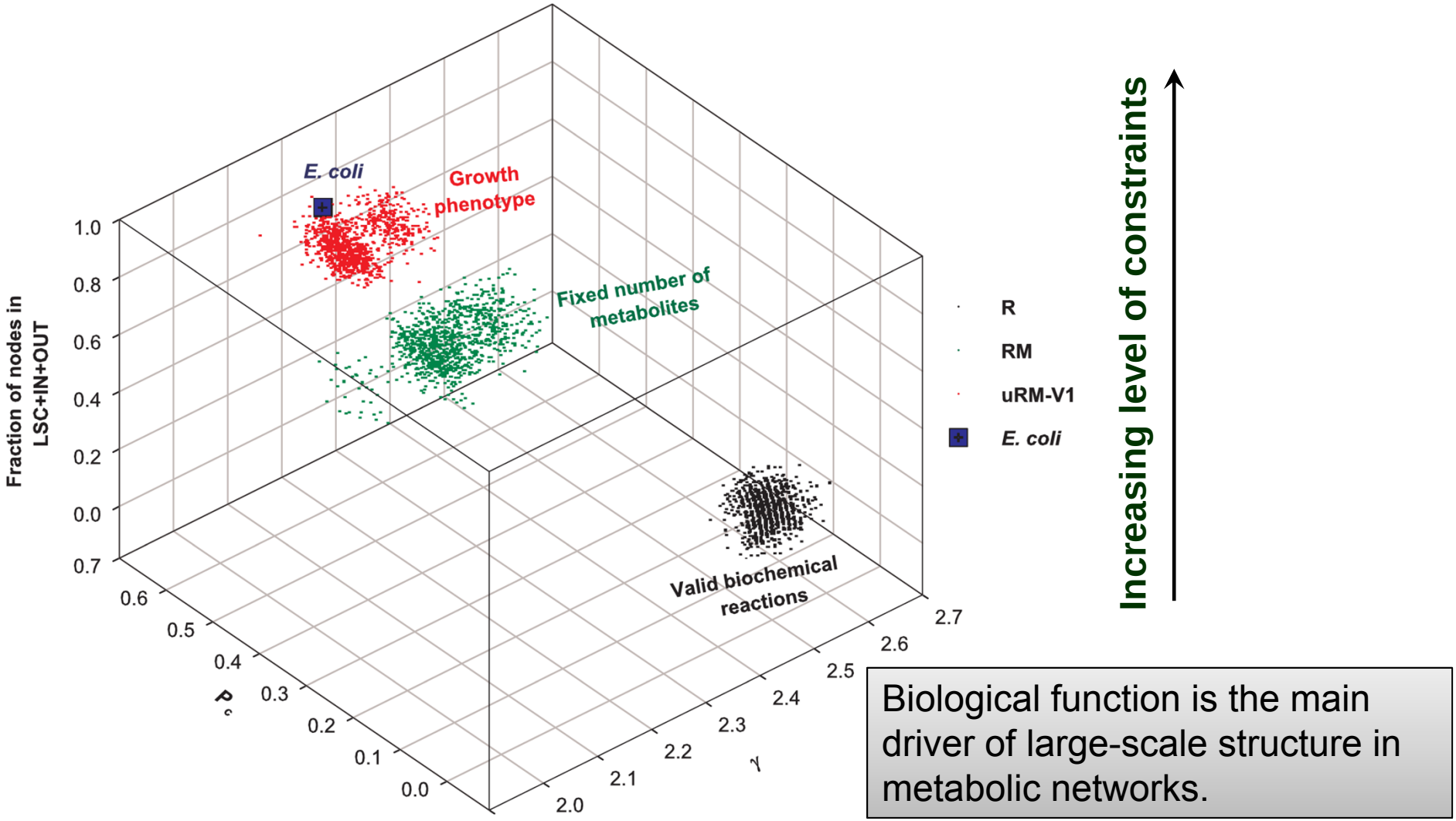


In a directed network, the may exist path from node *a* to *f* but lack a path back from node *f* to *a*. Probability that a path exists between two nodes is an important quantity characterizing directed networks.

A strongly connected component is a maximal set of nodes such that for any pair of nodes *a* and *b* in the set there is a directed path from *a* to *b* and from *b* to *a.*
The size of largest strong component is an important characteristic of directed networks.

Sampled networks have a bow-tie architecture similar to real network!

Global structural properties of real metabolic networks are a consequence of simple biochemical and functional constraints

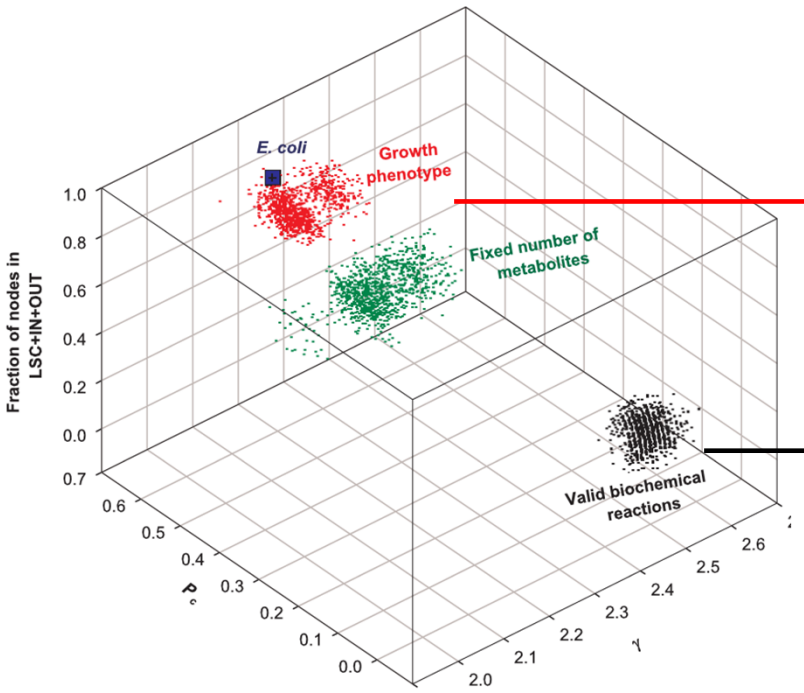


Biological function is the main driver of large-scale structure in metabolic networks.

Conjectured by:
A. Wagner (2007) and
B. Papp *et al.* (2008)

Reference: Samal and Martin, PLoS ONE (2011)

Additional constraints reduce the space of possible networks



Ensemble *uRm-V10*

Ensemble *uRm-V5*

Ensemble *uRm-V1*

Ensemble *uRm*

Ensemble *Rm*

Ensemble *R*

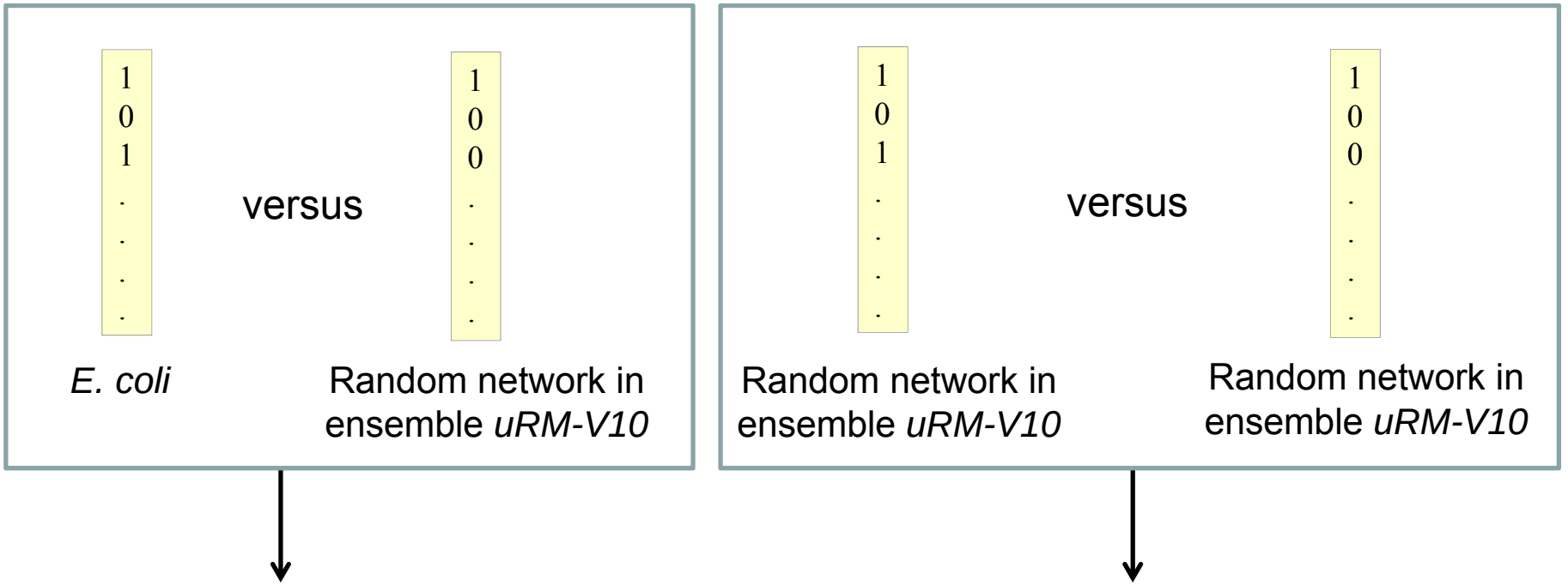
Embedded sets like Russian dolls

Including the viability constraint on the first chemical environment leads to a reduction by at least a factor 10^{50}

Reference: Samal *et al*, BMC Systems Biology (2010)

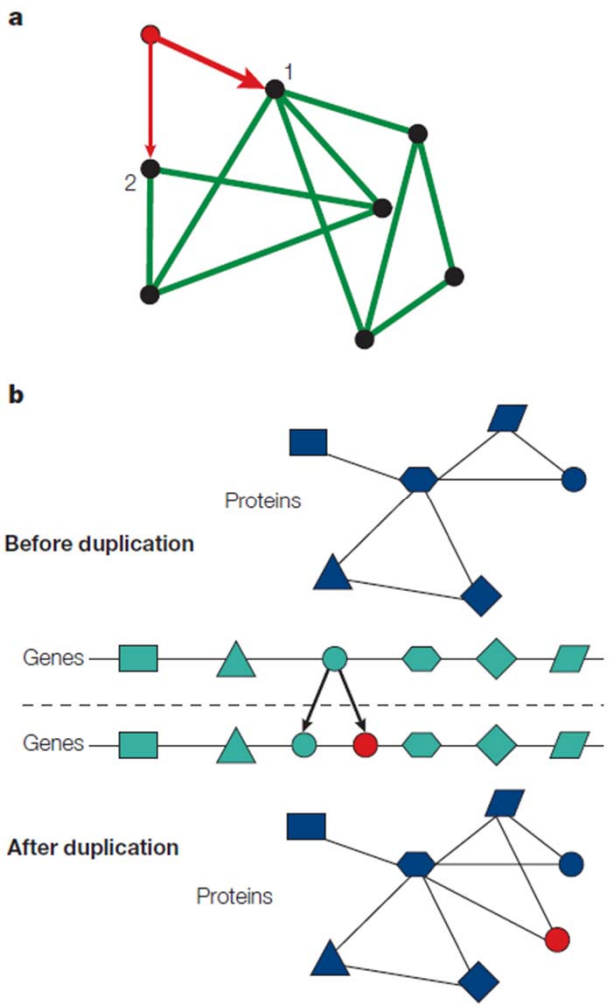
High level of genetic diversity in our randomized ensembles

Any two random networks in our most constrained ensemble *uRM-v10* differ in ~ 60% of their reactions.



Hamming distance between the two networks is ~ 60% of the maximum possible between two bit strings.

Origins of Power law degree distribution: gene duplication and divergence



In Barabasi and Albert model, scale-free networks arise through two basic mechanism: growth and preferential attachment.

It has been suggested that a combination of gene duplication and divergence can explain power laws observed in biological networks.

Say, genes are chosen at random for duplication and the duplicated protein has the same interacting partners in protein interaction network as the original protein.

Then the high degree proteins over time will gain more interacting partners by chance.

This can explain the origin of power laws in protein interaction networks and the line of argument has been extended to explain power laws in metabolic networks.

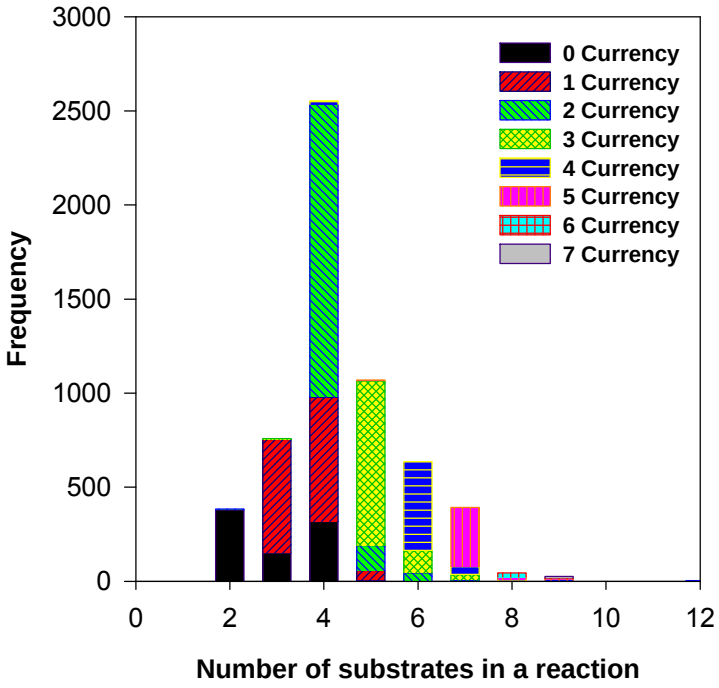
Reference: Barabasi and Oltvai (2004)

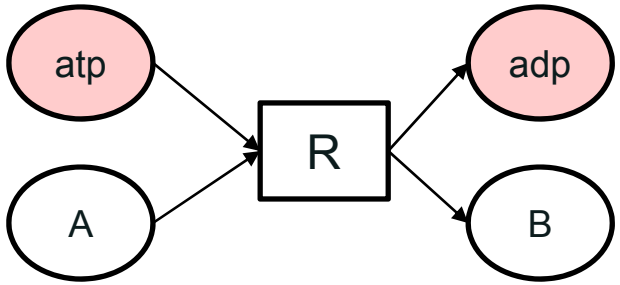
Reaction Degree Distribution

The nature of reaction degree distribution is very different from the metabolite degree distribution which follows a power law.

The degree of a reaction is the number of metabolites that participate in it.

The reaction degree distribution is bell shaped with typical reaction in the network involving 4 metabolites.



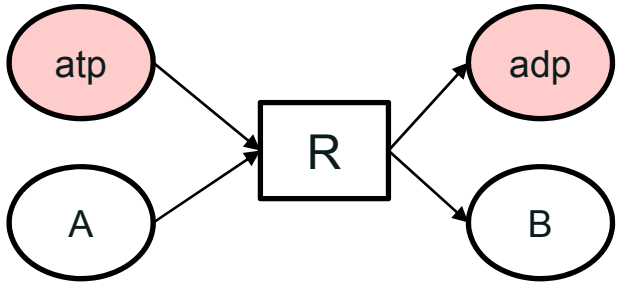


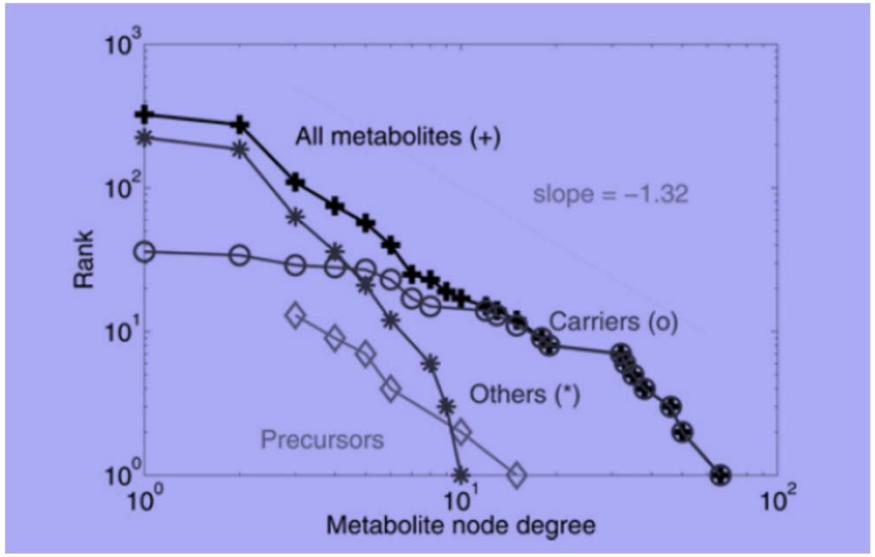
In each reaction, we have counted the number of currency and other metabolites.

Most reactions involve 4 metabolites of which 2 are currency metabolites and 2 are other metabolites.

Scale-free versus Scale-rich network: Origin of Power laws in metabolic networks

Tanaka and Doyle have suggested a classification of metabolites into three categories based on their biochemical roles
(a) ‘Carriers’ (very high degree)
(b) ‘Precursors’ (intermediate degree)
(c) ‘Others’ (low degree)



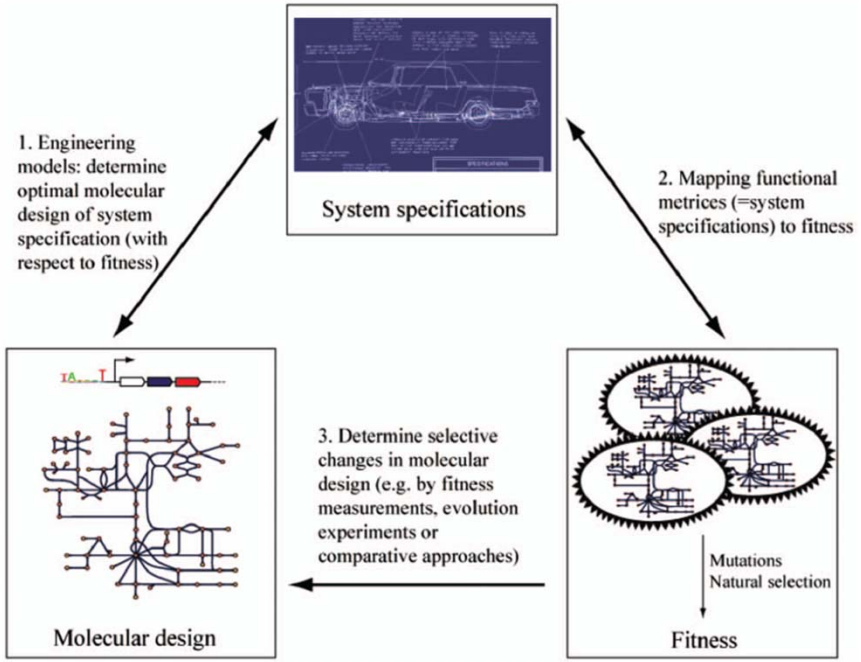


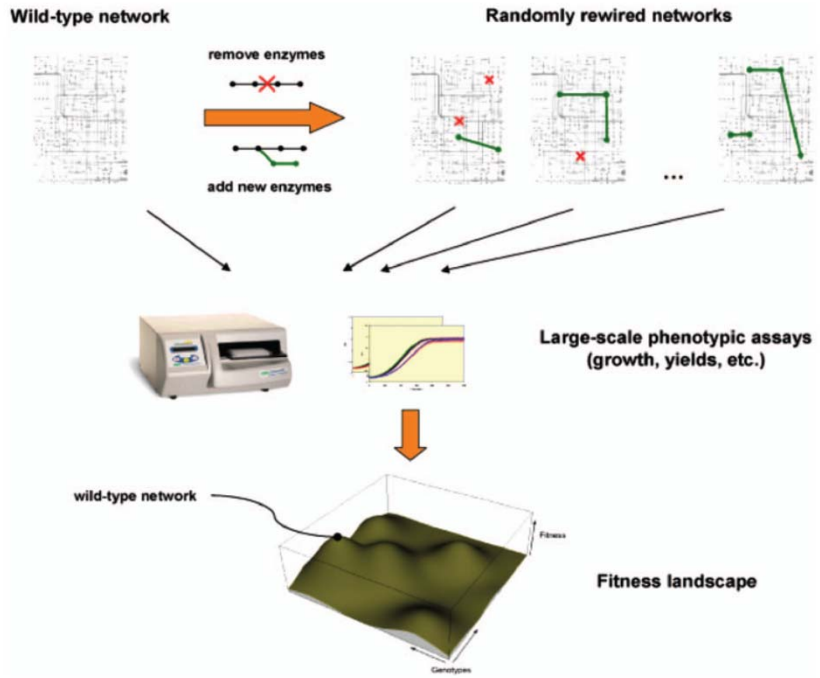
Gene duplication and divergence at the level of protein interaction networks can lead to observed power laws in metabolic networks.

However, the presence of ubiquitous (high degree) ‘currency’ metabolites along with low degree ‘other’ metabolites in each reaction can alternatively explain the observed power laws (or at least fat tail) in the metabolite degree distribution.

Reference: Tanaka (2005); Tanaka, Csete, and Doyle (2008)

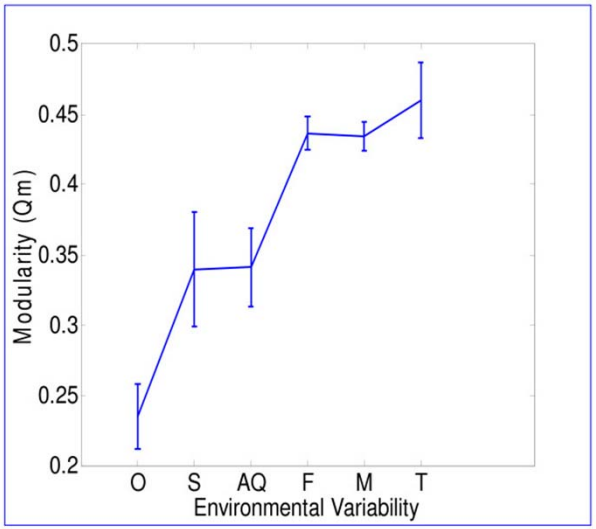
MCMC sampling method: A computational framework to address questions in evolutionary systems biology

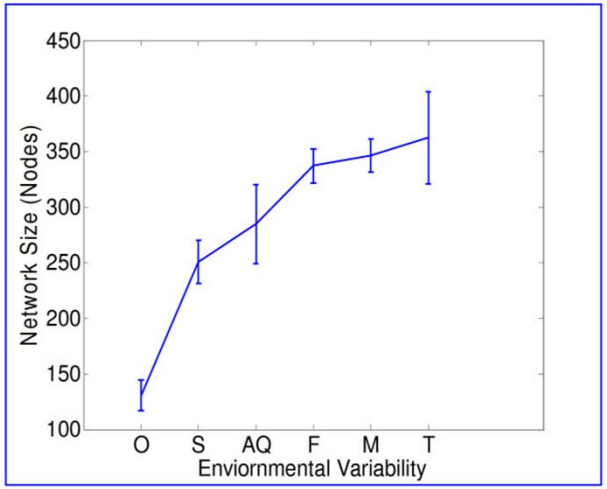




Reference: Papp et al (2008)

Empirical studies have shown that generalist prokaryotes have more modular metabolism than specialists





References:
Parter et al (2007); Kreimer et al (2008)

Empirical studies by the groups of Uri Alon and Eytan Ruppin have shown that the metabolic networks of generalist prokaryotes (with the ability to live in many environments) are more modular than specialist prokaryotes.
These observations led to the suggestion that environmental variability increases modularity in metabolic networks.

However, note that the result from these studies is not conclusive given the size of the metabolic networks also increases with environmental variability.

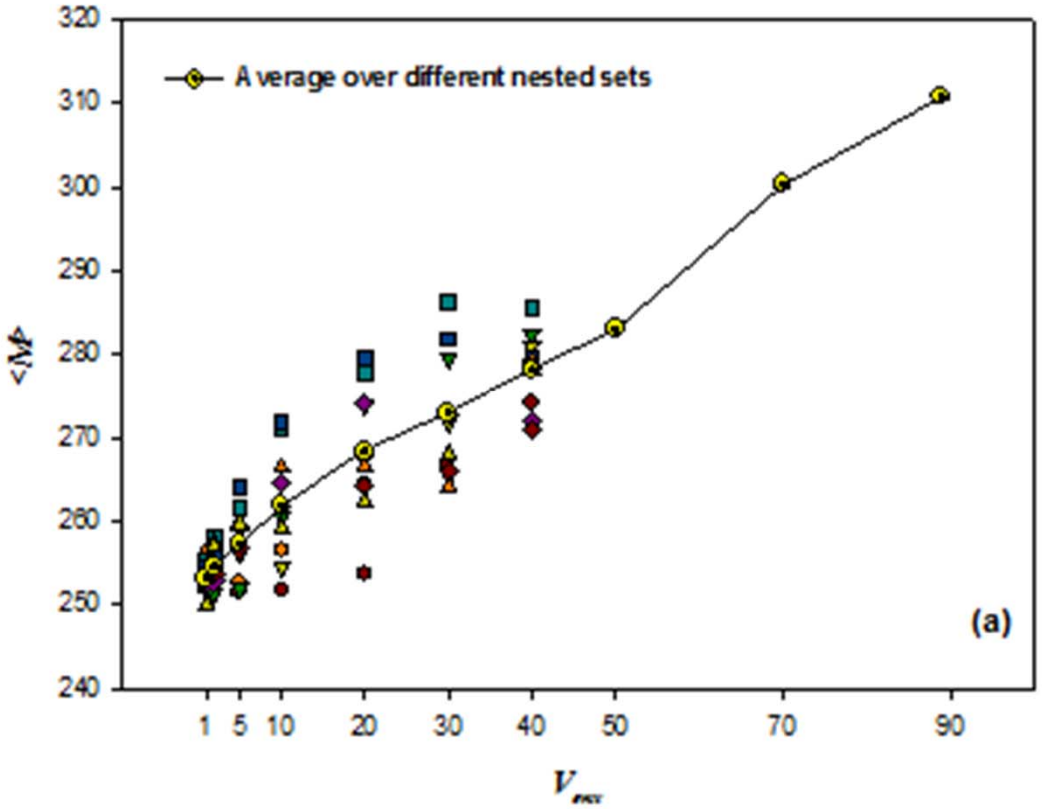
We decided to address this question using our MCMC sampling of random viable metabolic networks that have not been subject to unknown selection pressures unlike metabolism of real organisms.

Sampling of networks with different environmental versatility V_{env}

- MCMC sampling method can be used to sample random viable metabolic networks with a given phenotype.
- The desired phenotype is viability on a given set of environments.
- If the desired phenotype consists of V_{env} different environments, we designate the **Environmental Versatility Index** of sampled networks to be $\boldsymbol{V_{env}}$.
- V_{env} =1 refers to networks viable in 1 environment, V_{env} =2 refers to networks viable in 2 environments, and so on.

Versatility V_{env}	**Number of Sampled networks**
1	1000
2	1000
5	1000
10	1000
.	.
.	.
.	.

Modularity increases with environmental versatility



The Environmental Versatility Index V_{env} denotes the number of distinct environments in which a genotype is viable.

The modularity index *M* for a genotype gives the number of reactions contained in the FCSs (modules) of that genotype.

Reference: Samal, Wagner*, Martin*, BMC Systems Biology (2011)

Two scenarios for the evolution of modularity

- Modularity might result from directional selection favouring change in one trait while stabilizing selection maintains other traits unchanged.

AMER. ZOOL., 36:36–43 (1996)

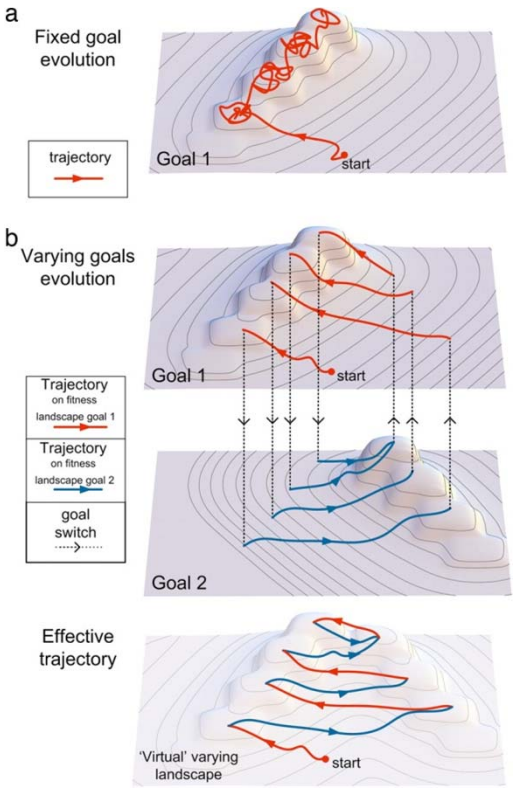
Homologues, Natural Kinds and the Evolution of Modularity[1]

GÜNTER P. WAGNER
Center for Computational Ecology, Department of Biology, Yale University

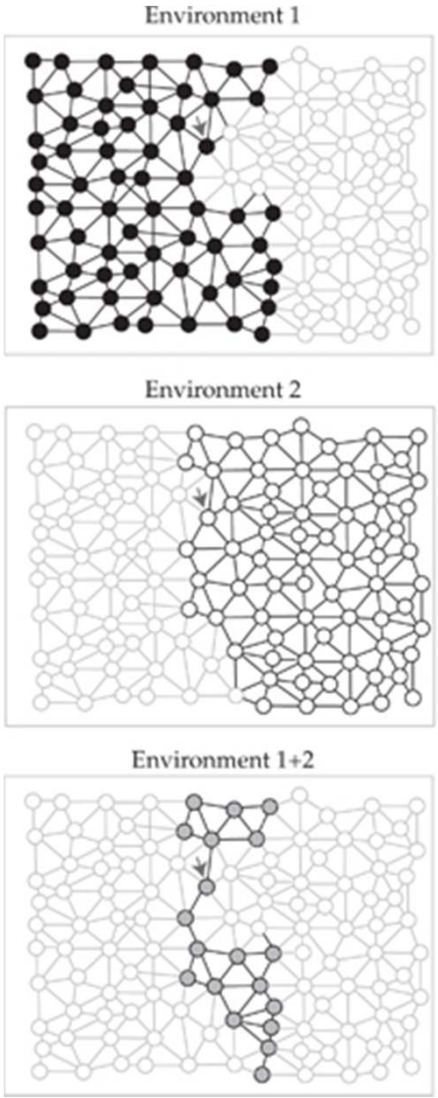
- Modular fluctuations in evolutionary goals can be sufficient to produce and maintain modularity. In this scenario, modularity will be lost once there are no fluctuations in the evolutionary goals.

Functional constraints lead to emergence of modularity

Uri Alon's Picture



Our Picture



- Our results suggest a scenario which can explain the emergence of modularity in both non-fluctuating and fluctuating environment scenario.
- The main requirement for emergence is an increase in the number of functions that a network performs.

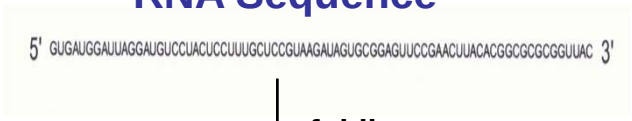
The intersection of genotype spaces for two different environments contains more modular networks. Modularity is ultimately a property of the genotype-phenotype map.

Genotype networks: A many-to-one genotype to phenotype map

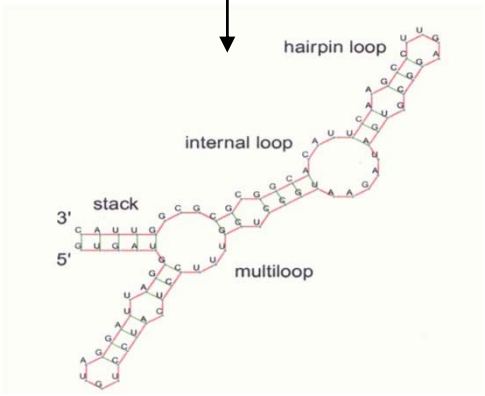
Genotype

Phenotype

RNA Sequence



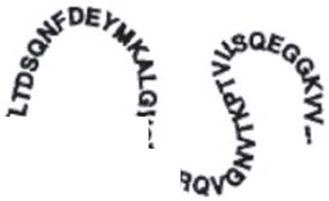
folding



Secondary structure of pairings

Reference: Fontana & Schuster (1998)

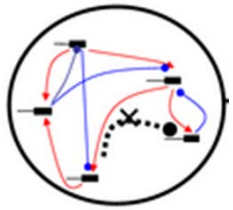
Protein Sequence



Three dimensional structure

Reference: Lipman & Wilbur (1991)

Gene network



Gene expression levels

Reference: Ciliberti, Martin & Wagner (2007)

In the context of RNA, genotype networks are commonly referred to as Neutral networks.

We have studied the properties of the metabolic genotype-phenotype map using our framework in detail.

Reference: Samal *et al*, BMC Systems Biology (2010)

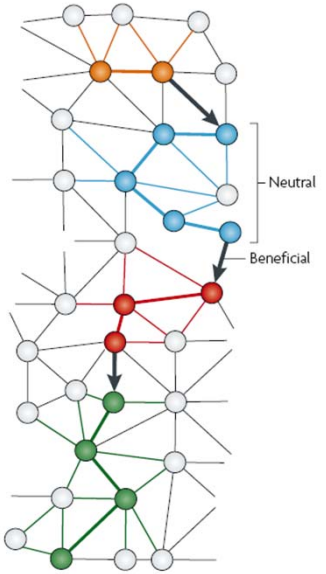
Implications: Origins of Evolutionary Innovations

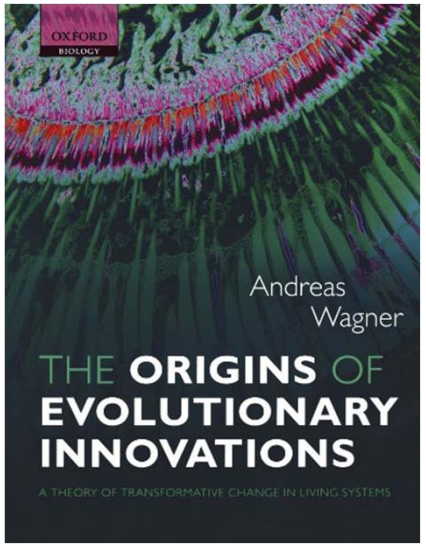
- "*How do organisms maintain existing phenotype while exploring for new and better adapted phenotypes?*"

- It is important to realize that Darwin's theory explains the *survival of the fittest* but does not explain the *arrival of the fittest*.

- Our MCMC simulations show that: "Starting with the reference genotype, one can evolve to very different genotypes with gradual small changes while maintaining the existing phenotype".

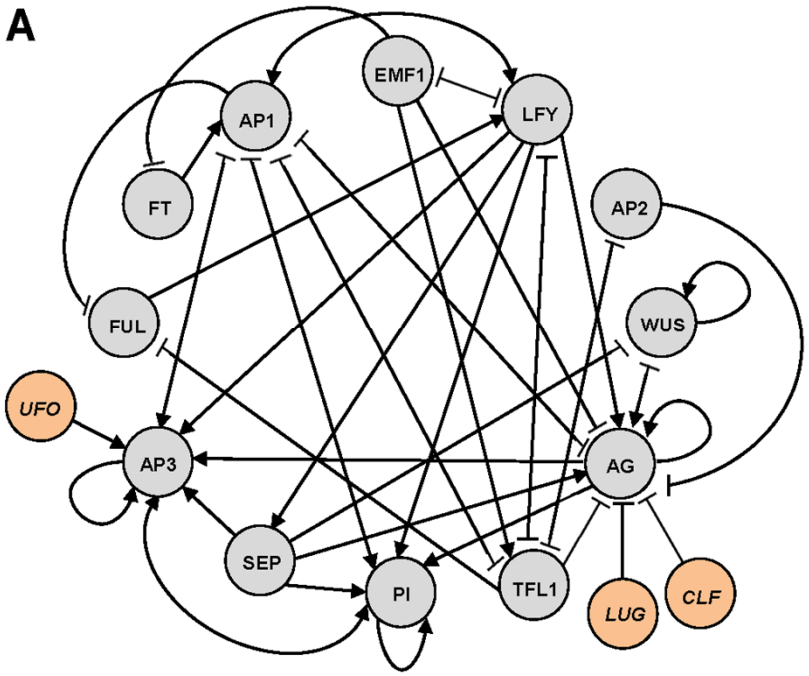
- Neighborhoods of different genotypes with a given phenotype can have access to very different novel phenotypes.





Floral Organ Specification (FOS) gene regulatory network

A



B

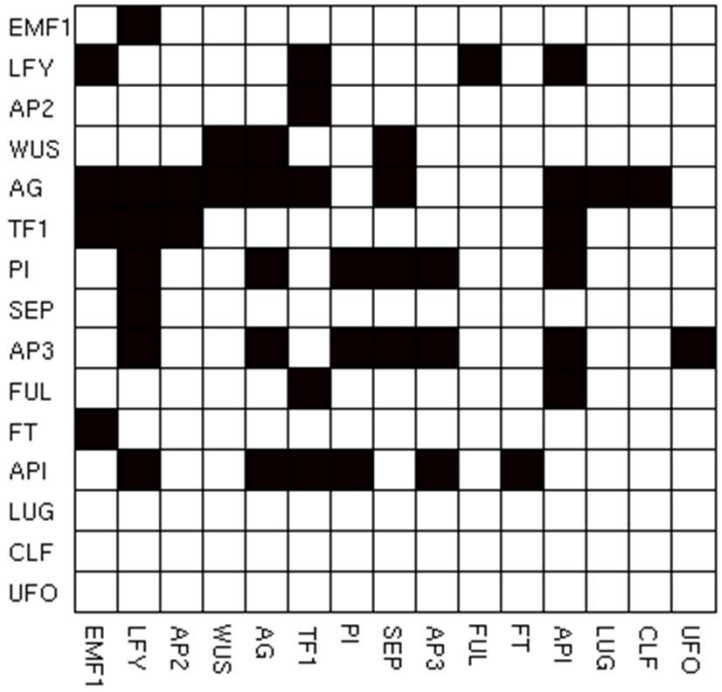
	EMF1	LFY	AP2	WUS	AG	TFL1	PI	SEP	AP3	FUL	FT	AP1	*LUG*	*CLF*	*UFO*
Inflorescence I1	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0
Inflorescence I2	1	0	0	0	0	1	0	0	0	0	0	0	1	1	1
Inflorescence I3	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0
Inflorescence I4	1	0	0	1	0	1	0	0	0	0	0	0	1	1	1
Sepal	0	1	1	0	0	0	0	1	0	0	1	1	1	1	0
Petal p1	0	1	1	0	0	0	1	1	1	0	1	1	1	1	1
Petal p2	0	1	1	0	0	0	1	1	1	0	1	1	1	1	0
Stamen st1	0	1	1	0	1	0	1	1	1	1	1	0	1	1	1
Stamen st2	0	1	1	0	1	0	1	1	1	1	1	0	1	1	0
Carpel	0	1	1	0	1	0	1	1	0	1	1	0	1	1	0

Reference: Alvarez-Buylla et al, The Arabidopsis Book (2010)

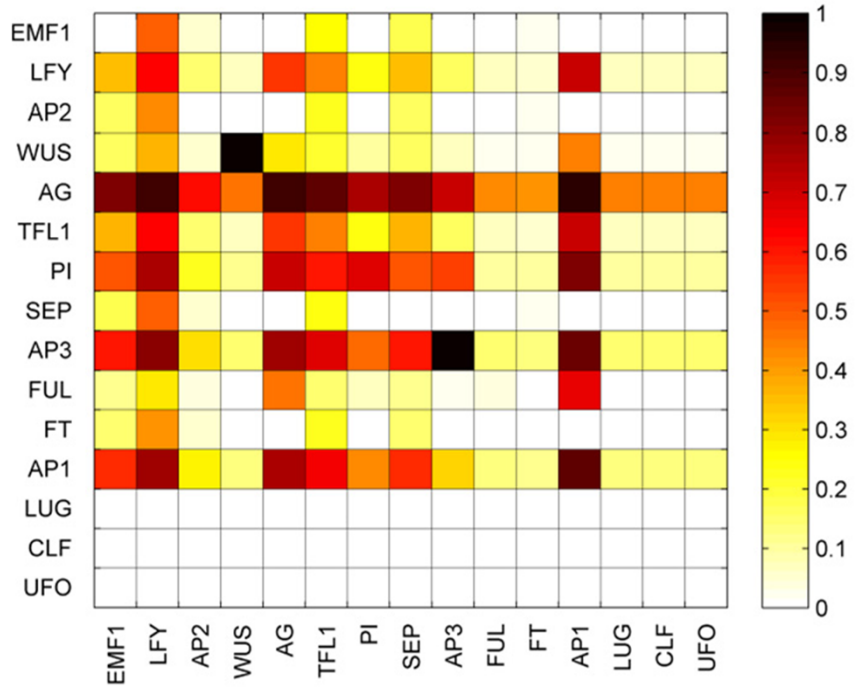
We have studied the Boolean Gene Regulatory Network model for Arabidopsis Floral Organ Specification network containing 15 genes connected by 46 edges.
There are 10 attractors of the network whose expression states specify different cell types (shoot apical meristem, sepal, petal, stamen and carpel).

We have developed a Markov Chain Monte Carlo (MCMC) method similar to the case of metabolic networks to sample the space of functional gene regulatory networks.

Edge usage in real and sampled networks with functional constraints



Arabidopsis network



Sampled networks with phenotype constraints

Reference: Henry, Moneger, Samal* and Martin*, in submission

Summary

- We have proposed a null model based on Markov Chain Monte Carlo (MCMC) sampling to generate benchmark ensembles with desired phenotype for metabolic networks.
- Our realistic benchmark ensembles can be used to distinguish between 'typical' and 'atypical' properties of a network.
- We show that many large-scale structural properties of metabolic networks are by-products of functional or phenotypic constraints.
- Modularity in metabolic networks can arise due to phenotypic constraints of growth in many different environments.
- Our framework can be used to address many questions in evolutionary systems biology.

Acknowledgements

Olivier Martin

Adrien Henry

Andreas Wagner

João Rodrigues

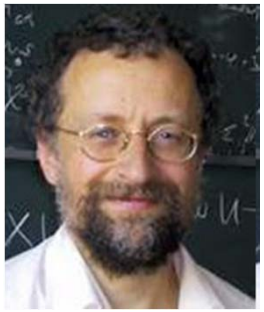
GDRE 513 SysBio

LPTMS

Université Paris Sud

INRA

University of Zurich



Jürgen Jost

Max-Planck-Institut für Mathematik in den Naturwissenschaften

Françoise Monéger

ENS DE LYON

FEBS