

Biological networks: a flurry of methods and opportunities

Olivier Martin

University of Paris-Sud at Orsay, France

Selection of 2 topics

[I] Computational analysis of model gene networks

(work of S. Ciliberti, O. Martin and A. Wagner, 2007 and 2008)

[II] Counting (stochastically) the size of RNA neutral networks

(work of T. Jorg, O. Martin and A. Wagner, 2008)

Gene regulatory networks

*S. Ciliberti, O.C. Martin and A. Wagner,
PLOS Computational Biology and PNAS 2007*

Issues for the mapping from genotype to phenotype

- (1) structure/topology of the « neutral networks »
- (2) distribution of the robustness (mutational, to noise...)
- (3) effects of selection on robustness
- (4) nature of the trade-off between mutational robustness and phenotypic innovation

Computational tools : some borrowed from RNA
plus use of:

- Permutation and gauge symmetries
- Metropolis sampling of genotype space
- Guided search for connecting paths

The gene network model

Model proposed by Andreas Wagner (Evolution, 1996)

Used since to fit quantitatively *Drosophila* gene expression levels during early development.

Used also for in silico studies by Bergman-Siegel, Li et al., Azevedo, ... to see robustness, canalisation, epistasis, ...

Regulatory network of N transcriptional regulators

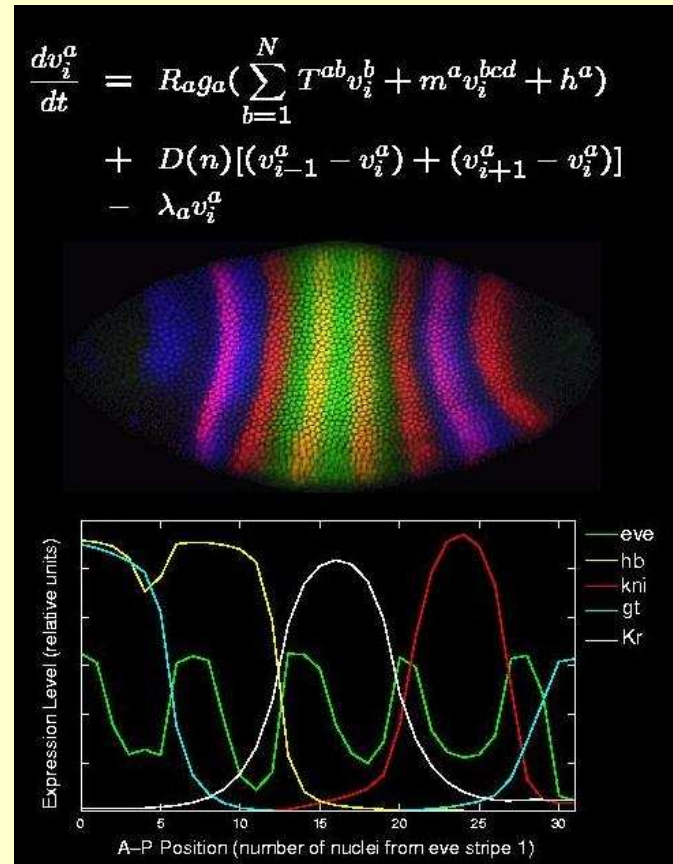
We focus on their expression levels

$$\mathbf{S}(t) = (S_1(t), S_2(t), \dots, S_N(t))$$

at some time t during a developmental or cell-biological process and in one cell or domain of an embryo.

Drosophila development

- Maternal genes (RNAm)
- Segmentation genes
- Homeotic genes



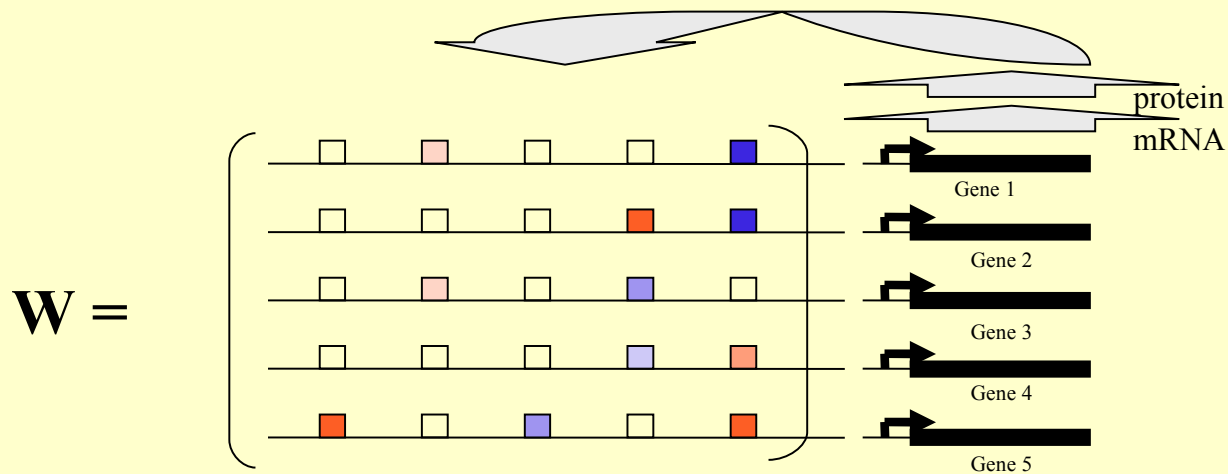
Regulatory dynamics of the model

Discrete time **dynamics** for simplicity:

$$S_i(t+1) = F(W_{i1}S_1(t) + W_{i2}S_2(t) + \dots)$$

where F is a sigmoidal function. Analog of neural network using a *threshold* weighted sum for input to output..

The elements W_{ij} of this matrix indicate the strength of the regulatory influence that gene j has on gene i . This influence can be either activating ($W_{ij} > 0$), repressing ($W_{ij} < 0$) or absent.



Genotypes and phenotypes

The matrix W represents the **genotype** of the network.

The **phenotype** is the (steady-state) expression pattern at long times. For developmental reasons, we think of initial conditions as being given (e.g., maternal RNAs or upstream genes); “proper” development requires that the network reach the right *target pattern* for the S_i .

Strong selection limit:

A genotype is **viable** if and only if it goes to the steady state given by the target pattern. This corresponds to a *0-1 fitness landscape*.

Viable genotypes are rare

If one allows for M interactions (M non-zero entries of W) between N genes, what fraction of the genotypes (regulatory networks) are viable?

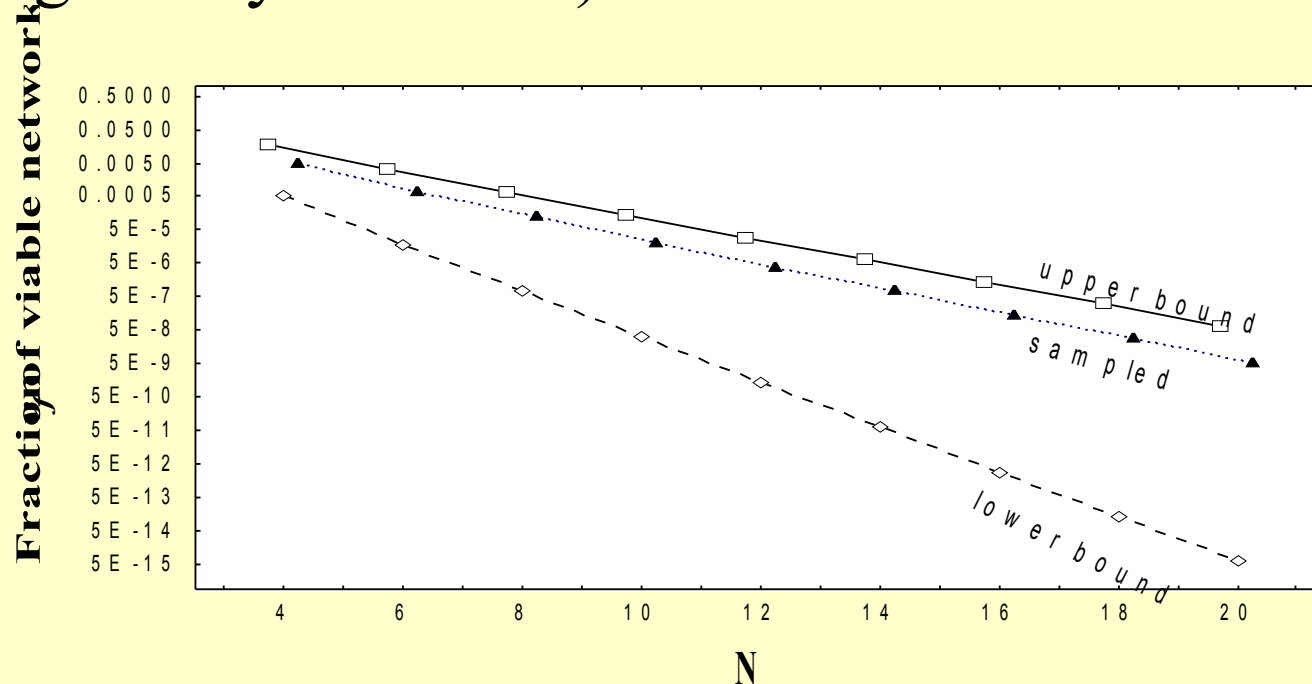
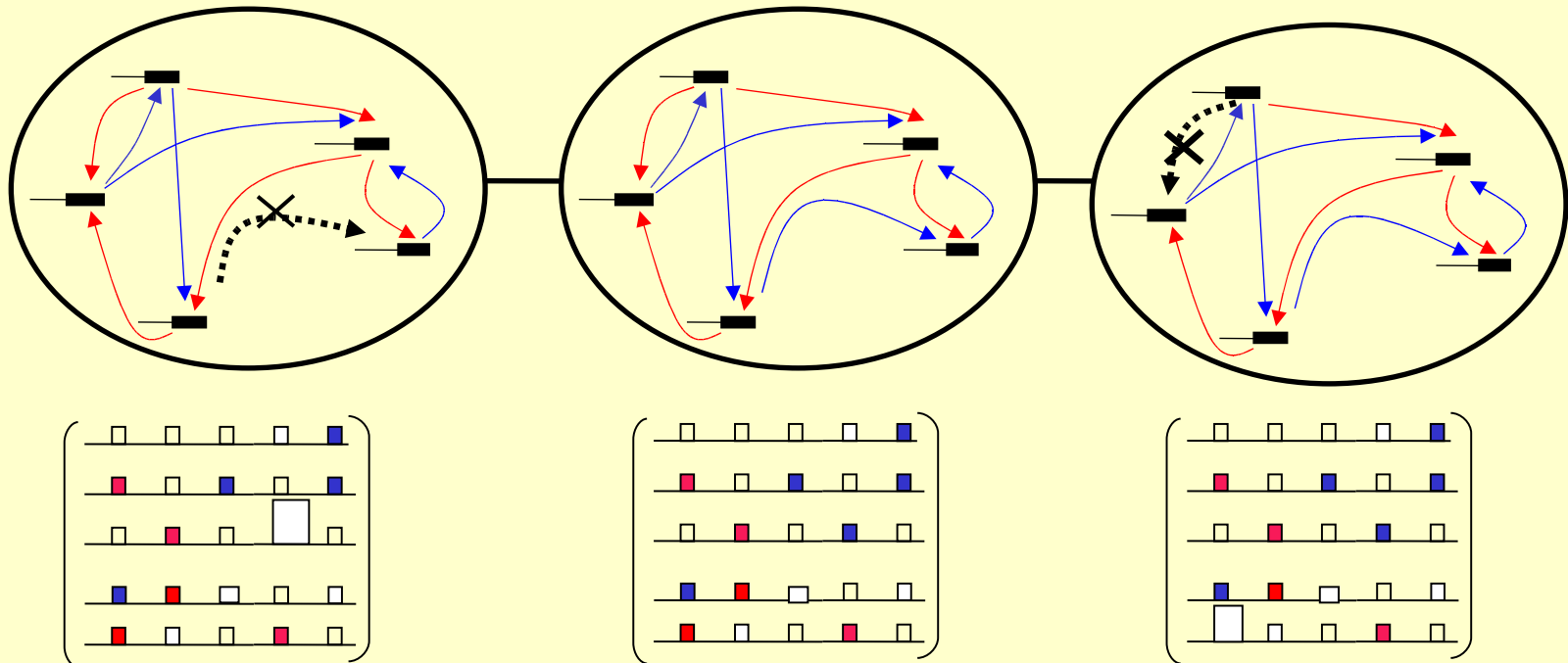


Illustration when $M = 0.25 N^2$

Building a topology for the fitness landscape

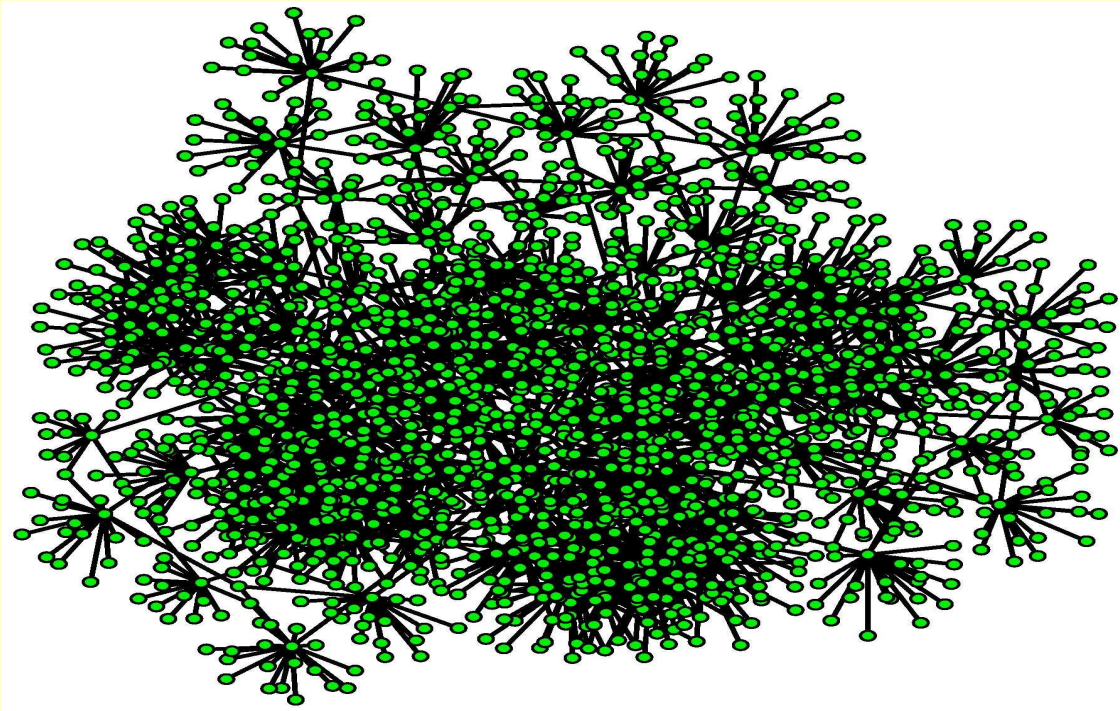
We say that two genotypes (W matrices) are nearest neighbors if they differ in just one regulatory interaction



Questions regarding the associated fitness landscape

Is the set of viable genotypes *connected*?

Can *robustness* evolve through gradual evolution under darwinian selection for viability (fitness) only?



Set of viable genotypes in the near neighborhood of one network

Computational challenges

- (1) Generate a genotype of given phenotype (oriented search)
- (2) Sample *uniformly* genotypes of a given phenotype: use symmetries to reduce exponentially the space size
- (3) Determine the connectivity of the neutral network: do guided search to go from one random genotype to another
- (4) Sample uniformly a connected component of the neutral network: use random walks
- (5) Sample uniformly the surface of a “ball” around a point: use Metropolis with asymmetric rates
- (6) Get the infinite population limit of a population under Darwinian selection: use variance reduction and $1/N$ extrapolation

The viable genotypes form a connected
« metagraph » (neutral network)

**Very few viable networks are *not* in the
giant connected component, and the few
such networks are usually isolated.**

Example: For $M=0.25 N^2$, the fraction of viable networks
not belonging to the giant component is:

$$2.3 \times 10^{-3} \text{ at } N=8$$

$$1.7 \times 10^{-3} \text{ at } N=12$$

$$1.4 \times 10^{-3} \text{ at } N=20$$

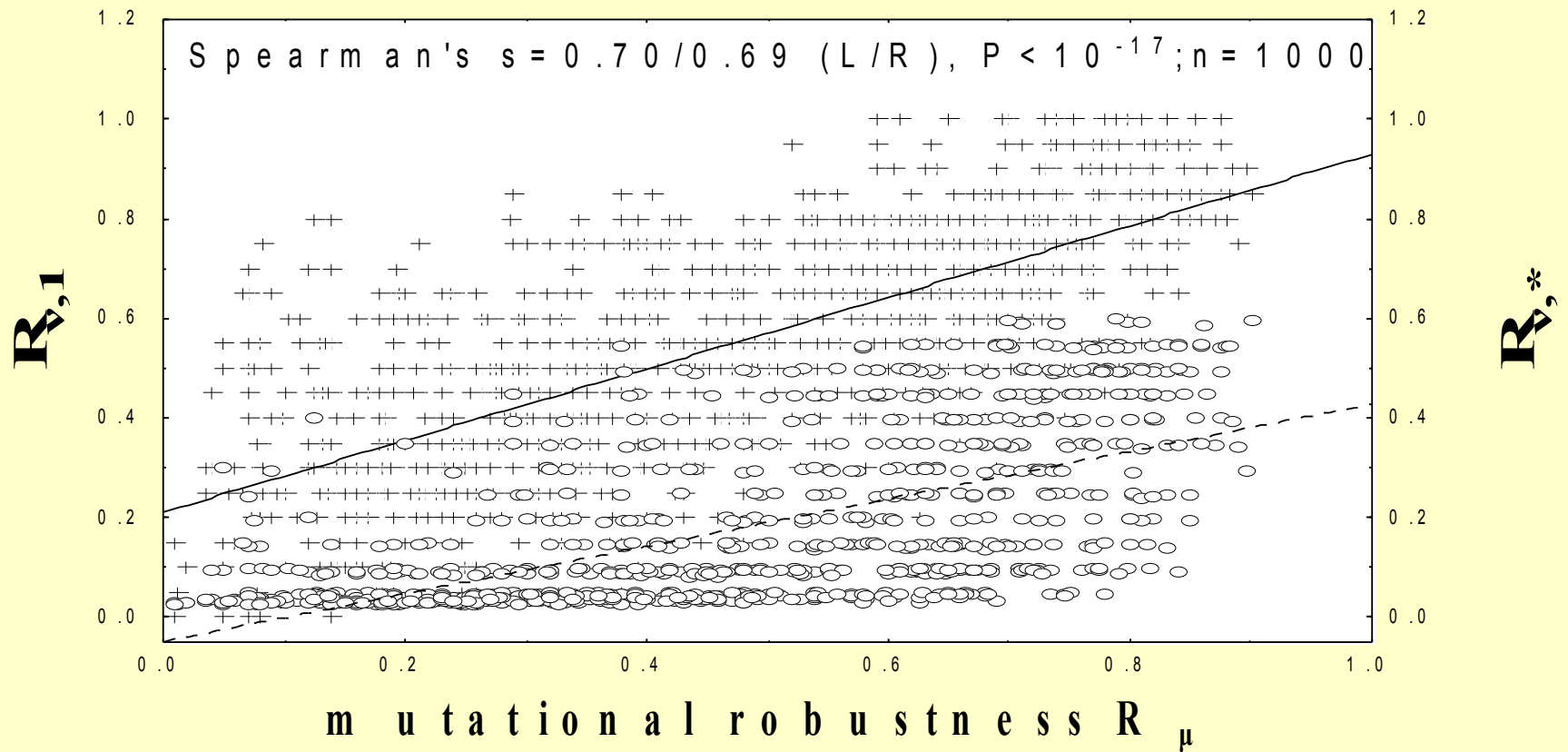
Measures of Robustness

Robustness to noise: the target expression is reached even in the presence of noise in the trajectory or in the initial conditions.

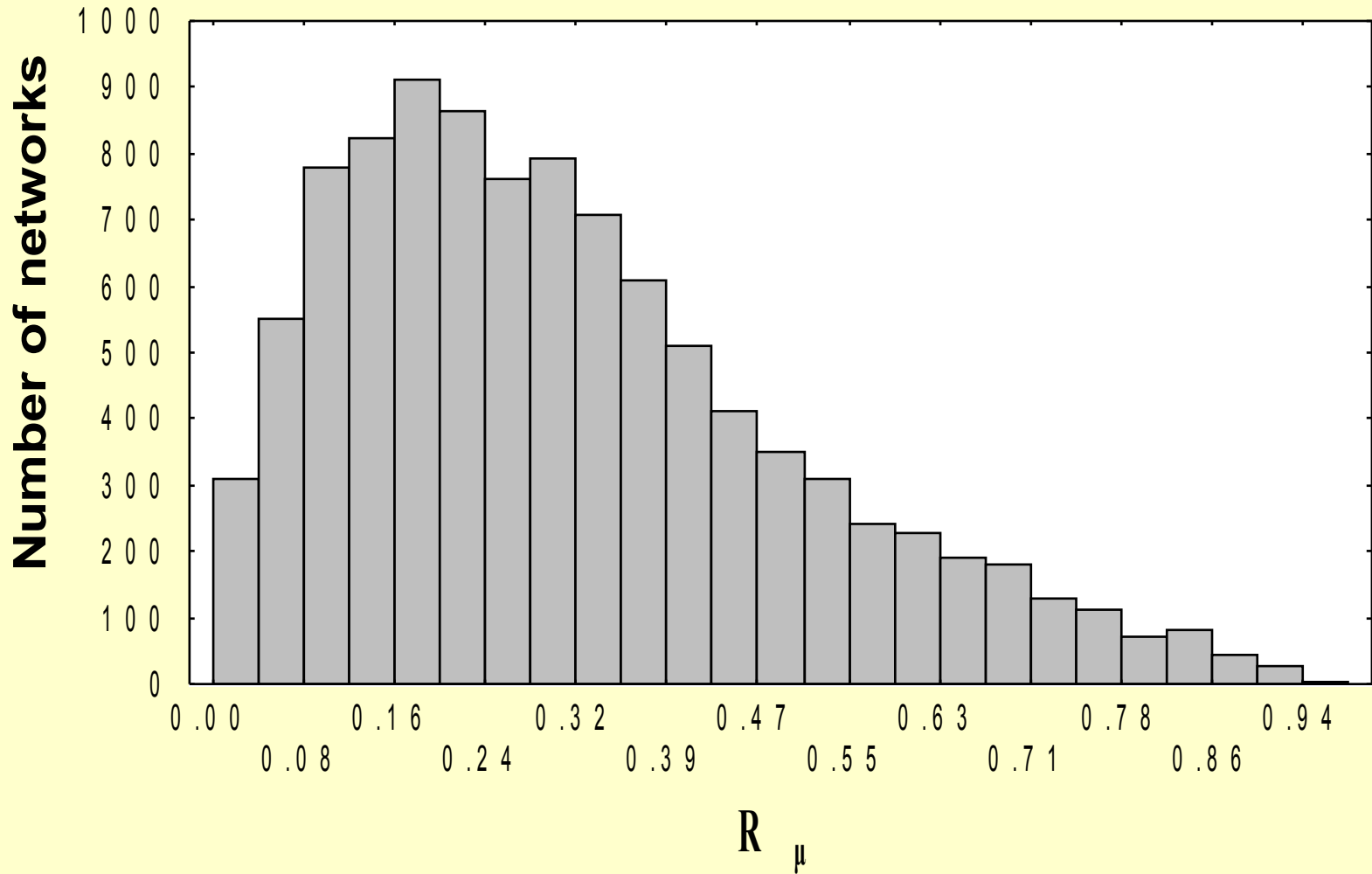
Mutational robustness: survival probability of offspring given that they are mutants.

Using computational techniques allowing one to sample uniformly the space of all viable networks, we have shown that:

- The different measures of robustness are strongly correlated
- Robustness has a broad distribution, even if N and M are large



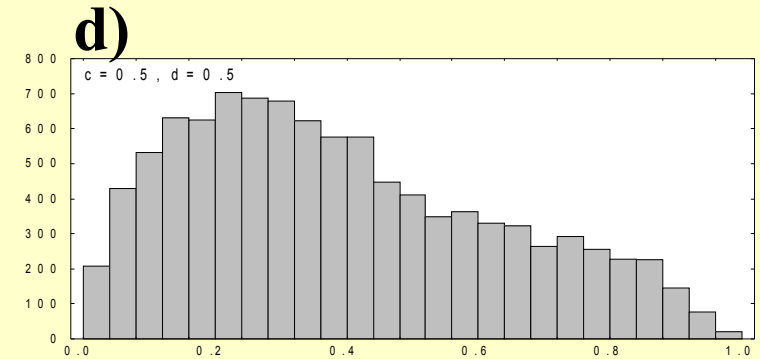
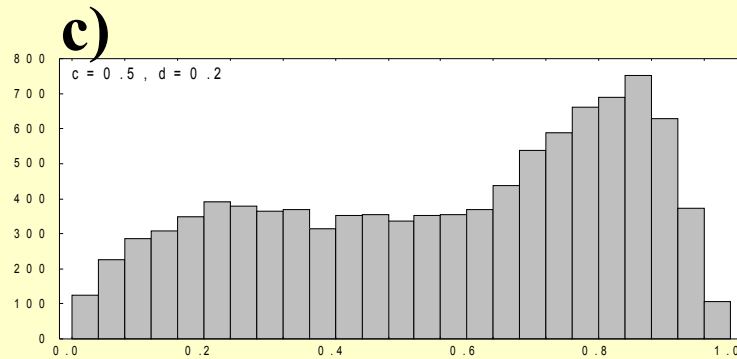
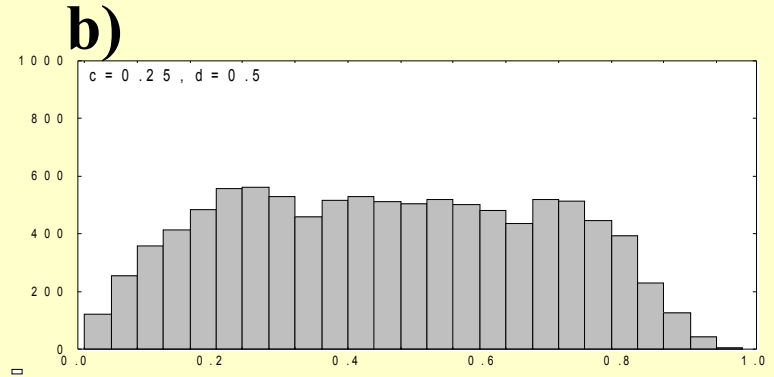
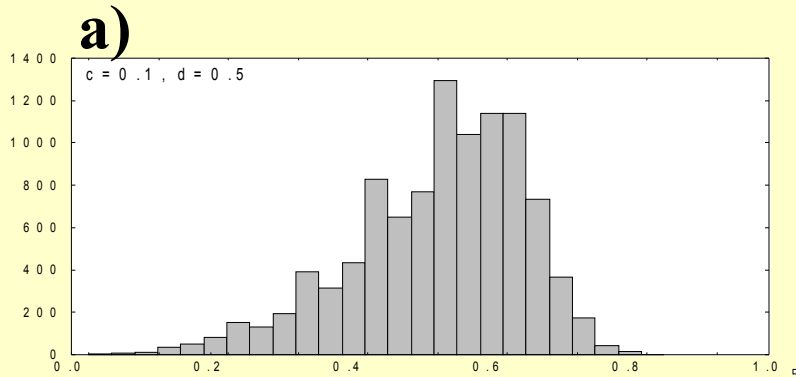
The mutational robustness and our measures of robustness to noise have a strong positive association.



The mutational robustness has a broad distribution

The pattern varies with the number N of genes and M of interactions but the mutational robustness always has a broad distribution:

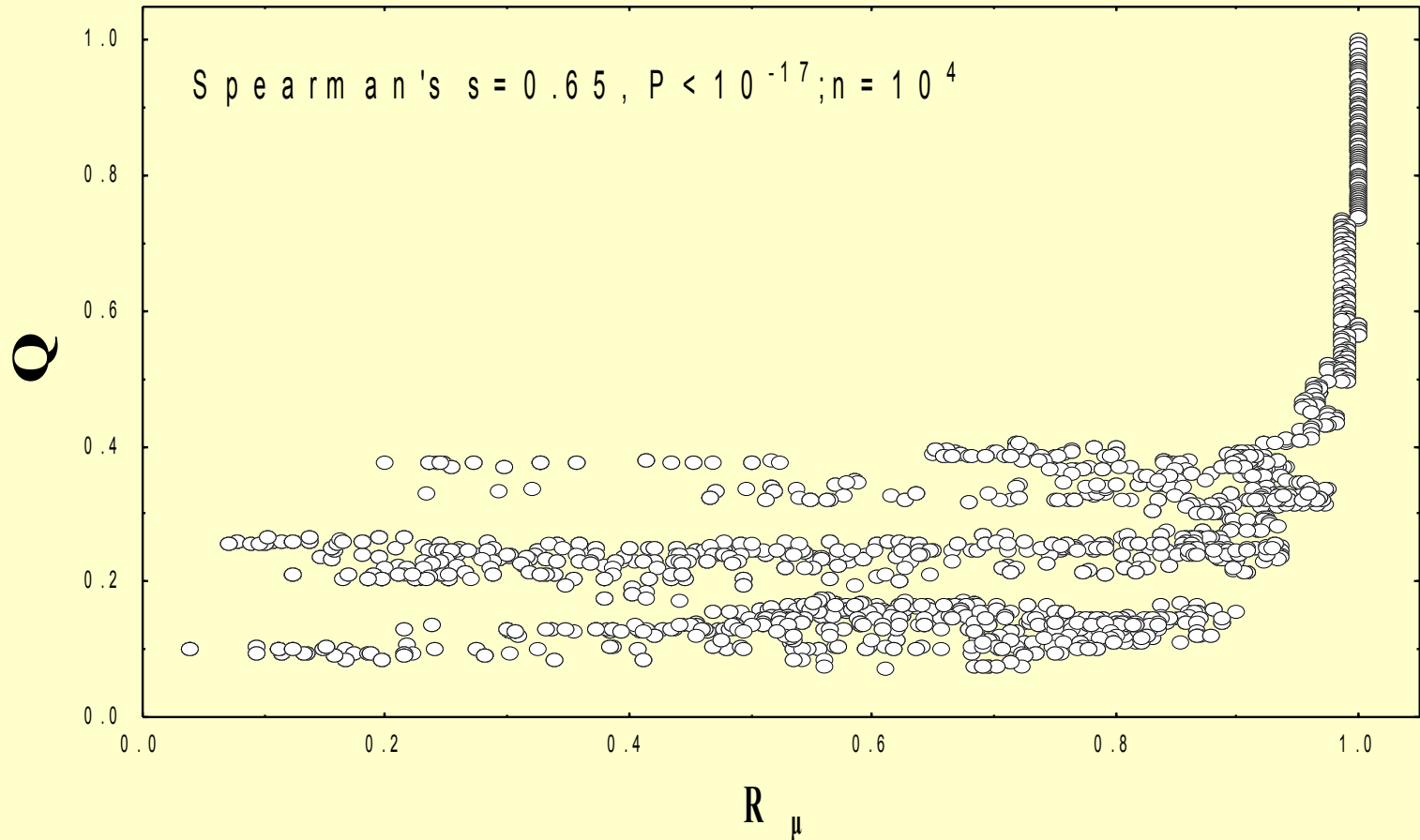
Number of networks



R_μ

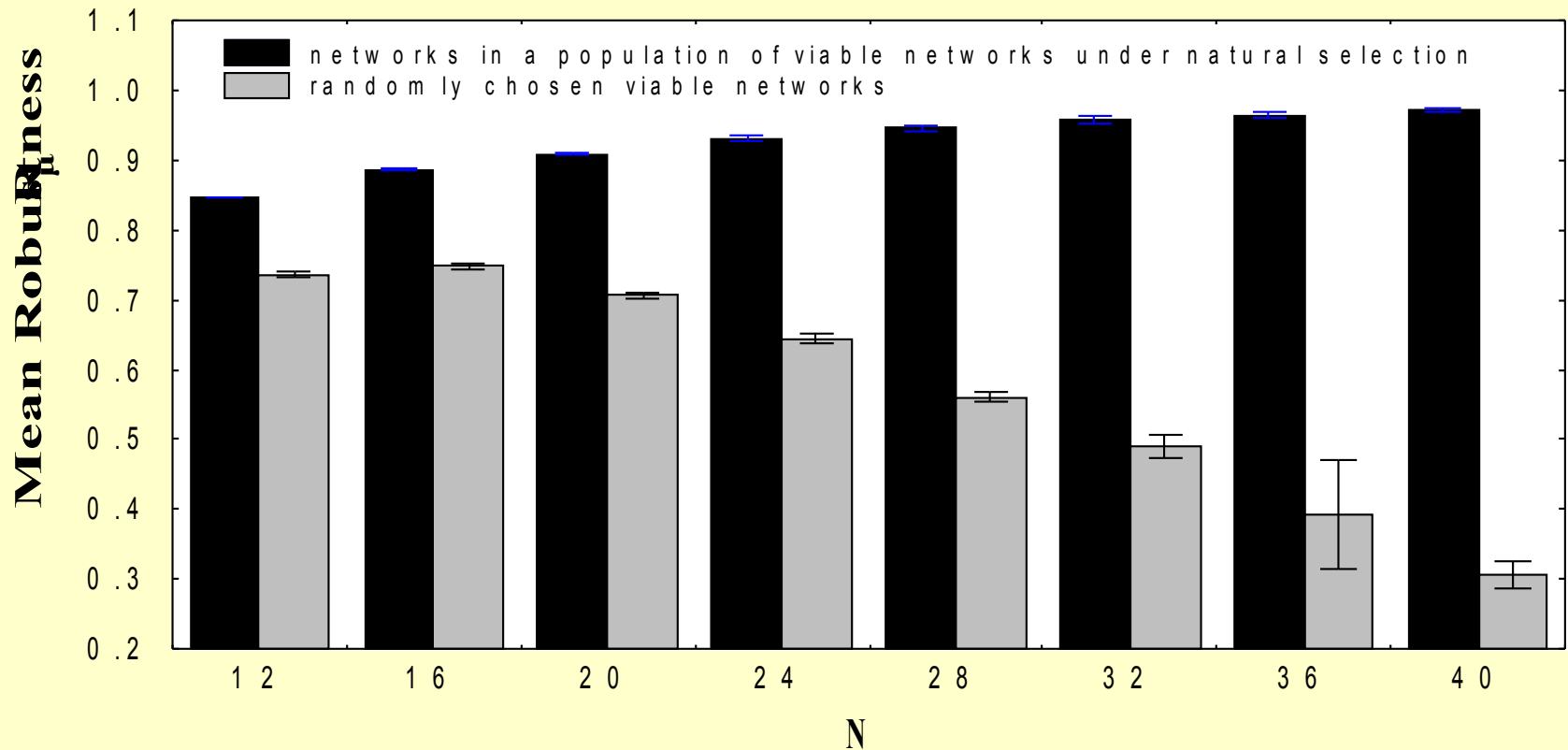
What makes a regulatory network robust?

Q is a « quality » factor which measures the synergy of the $W_{ij} S_j$



The mutational robustness and our measure Q have a strong positive association

Darwinian *selection for fitness* leads to regulatory networks of *high robustness*



Under darwinian selection, the mutational robustness increases with N while random viable networks have a mean robustness that goes to zero!

Phenotypic innovations in this model of gene networks

Mutants lead to new phenotypes, that is the dynamics lead to **other steady state** gene expression patterns.

N.B.: most genotypes in fact lead to non steady state behavior.

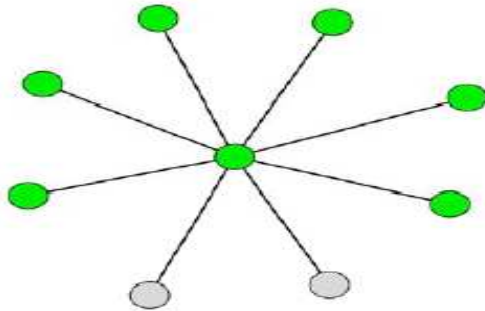
What is the organisation of these innovations in the genotype space?

Does innovation arise at a high level as one drifts on the metagraph?

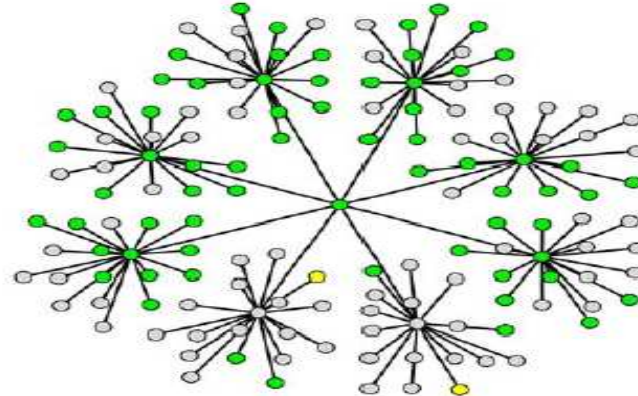
Does mutational robustness hinder discovery of new phenotypes?

Structure in the neighborhood of a viable genotype

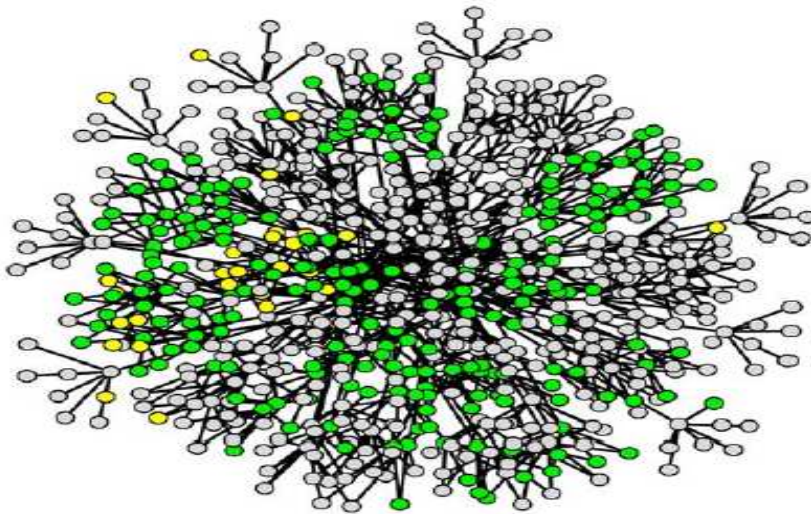
a) 1-neighbors



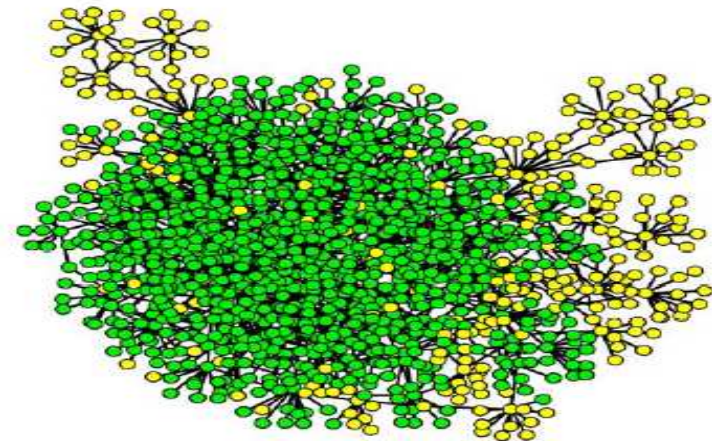
b) 2-neighbors



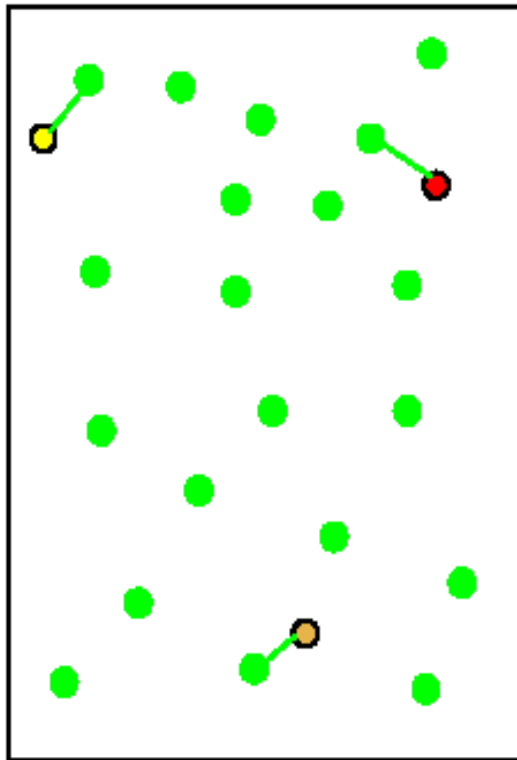
c) 3-neighbors



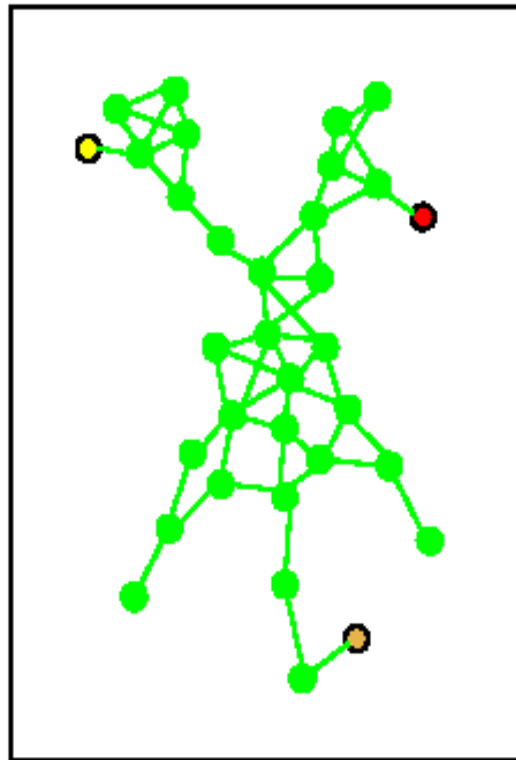
d) 4-neighbors
(fixed points only)



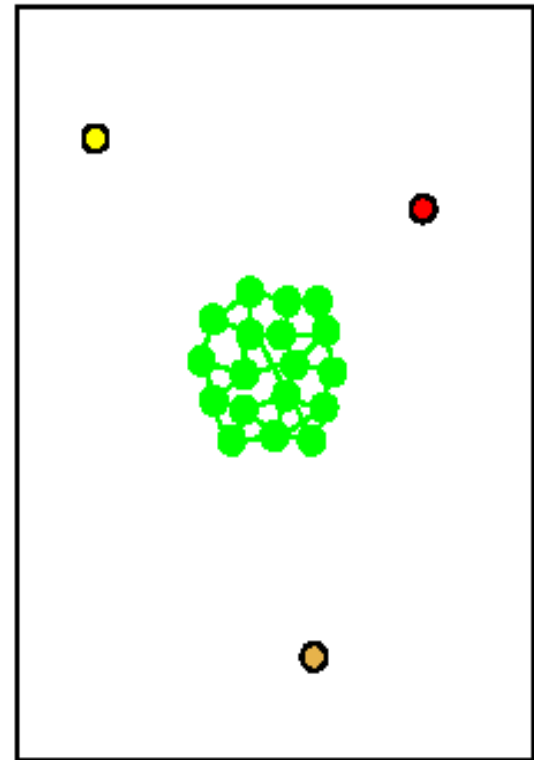
Metagraph *topology*: the tradeoff between robustness and innovation



Low Robustness
Low Innovation



High Robustness
High Innovation



High Robustness
Low Innovation

Conclusions for these networks

- RNA, gene and protein networks seem to all have wide-spanning neutral networks
- The nodes of these networks have a broad (non self averaging) distribution of mutational robustness
- The organization of the networks allows for an ideal tradeoff between robustness and evolutionary innovation
- Can one keep and show these nice properties in biologically more realistic models?
- We saw no evidence for modularity; can genetic recombination induce such structuring of the genotypes under natural selection?

Neutral network sizes of biological RNA molecules can be computed and are atypically large

T. Jörg¹ O. C. Martin¹ A. Wagner²

¹LPTMS, Université de Paris-Sud, France

²Department of Biochemistry, University of Zürich, Switzerland

Gene regulatory networks: Dynamics, spatial organization and
inference, Torino, April 23-24, 2008



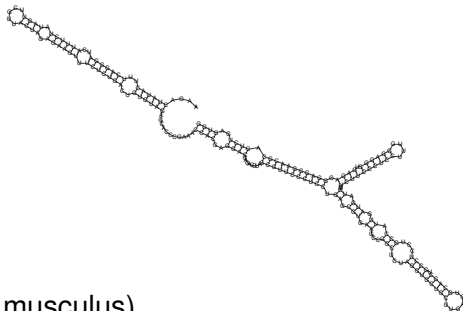
- 1 Introduction
 - RNA secondary structures
 - Neutral networks and evolution
- 2 A Monte Carlo approach to estimate the neutral network size
 - Why is it difficult to estimate the neutral network size?
 - Monte Carlo approach
 - Ranking structures
- 3 Results
 - Does it work?
 - Neutral networks of biological RNAs
- 4 Summary

Structural elements of RNA

- RNA sequence G built from 4-letter alphabet A,C,G,U

AAGACUAUACUUUCAGGGAUCAUUUCUAUAGUUCGUUACUAGAGAAGUUU
CUCUGACUGUGUAGAGCACCCGAAACCACGAGGACGAGACGUAGCGUCC
CUCCUGAGCGUGAAGCCGGCUCUAGGUGCUGCUUGACUGCAGCUGCCUCC
UGCCAUUGAUGAUCGUUCUUCCCUCCUUUGGGAGGGUGAGAGGGAGGGAA
CGCAGUCUGAGUGG

- Maps into well-defined secondary structure S



snoRNA (mus musculus)

Neutral networks

- Sequence 1

ACCGUUCGAAGCGAGUUACCUCCGU
GUACAUCAUUUACUGGCCUUAGGUC
CAAUUGCGGACUGCGAUCAACUGUA
GCUAAUACCGACAGACGCUCUUACA

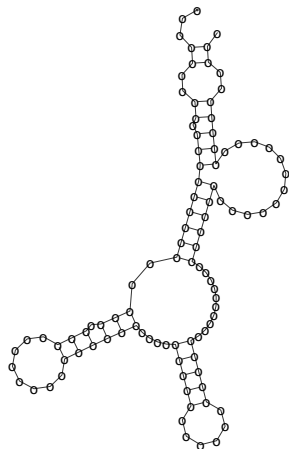
- Sequence 2

UGAAGGGGGUGCAGGGAUUAUAGUA
UGCAAUCAACAGCAACUCGCAGAUG
GGCCCGACCAUUAUUCAGACAGGAU
CCCCAAGGUCUCAAGUGCCUUUCUG

- Sequence 3

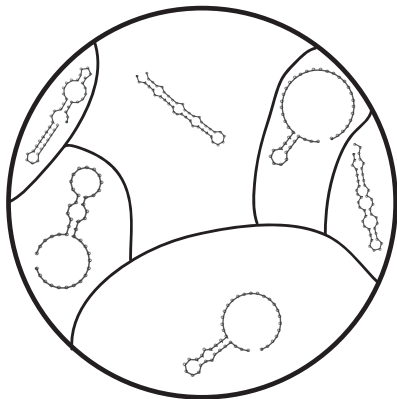
CCUGCCCCGCUGAGGCCUUCGUGAG
CUUUUCCCGUUGAGUCAGAUUGGGC
GCAUUGACGCUCUACCUAGUAAAAG
GCCGAUACCAUCCUUCACGCCGCG

- Sequence ...



snRNA (homo sapiens)

Neutral networks and evolution



Definitions

Sequence = Genotype

Structure = Phenotype \sim Function

Some facts

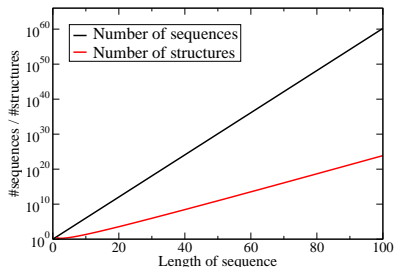
- 1968 Kimura formalizes the concept of neutral (molecular) evolution.
- Mid-nineties: Schuster et *al.* start systematic study.
- Huge differences in size of different neutral networks.
- (Mostly) connected through single-base mutations.
- Nontrivial structure.
- Related to fundamental concepts in evolution: Robustness and evolvability.

Why is it difficult to estimate the neutral network size?

Number of sequences – Number of structures

$$N_G = 4^L$$

$$N_S = 1.4848 \times L^{-\frac{3}{2}} (1.84892)^L$$



Example: Hammerhead ribozyme

Number of sequences = $4^{54} =$

3.2×10^{32}

Size of neutral network = 8.0×10^{22}

Number of compatible sequences =

1.9×10^{25}

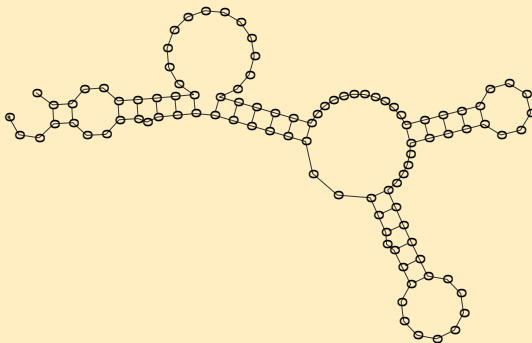
Compatible sets

Allowed pairings: A-U, C-G, G-C, G-U, U-A, U-G

Monte Carlo approach: Setting the stage (I)

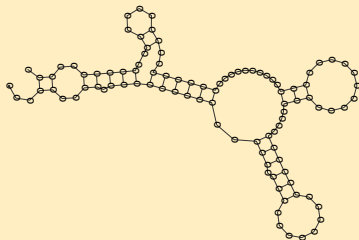
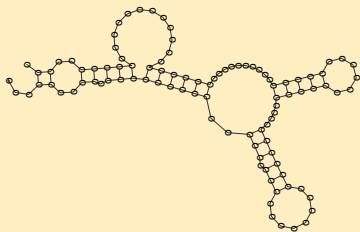
Bracket notation

...((..((..(((.....))))))....((((.....)))).....)))).



Monte Carlo approach: Setting the stage (II)

Distance from target structure: Base pair distance $d(S_0, S)$



...(((.(.(((.((((((((((..(((.((((.....))))))....((((.....)))).....))))))....))))))..)).
...(((.(.(((.((((((((((..(((.((((.....))))))....((((.....)))).....))))))..((.....)..))))))..)).

Minimal number of base pairs that have to be opened and closed to transform one structure in bracket notation into the other.

Monte Carlo approach: The building blocks (I)

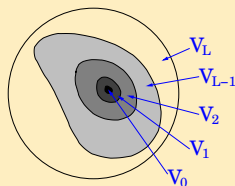
Splitting the space of sequences

- Consider the subspaces V_n of all sequences G having at most distance n from the target structure S_0 .

- $V_n \subset V_{n+1}$.

- V_0 is the neutral network.

- $|V_0| = \frac{|V_0|}{|V_1|} \times \frac{|V_1|}{|V_2|} \times \dots \times \frac{|V_{L-1}|}{|V_L|} \times |V_L|$
 $= |V_L| \prod_{i=1}^L r_i$ with $r_i = \frac{|V_{i-1}|}{|V_i|}$ and $|V_L| = 4^L$.



The main idea

Instead of estimating $\frac{|V_0|}{|V_L|}$ directly, we estimate the ratios $r_i = \frac{|V_{i-1}|}{|V_i|}$.

Monte Carlo approach: The building blocks (II)

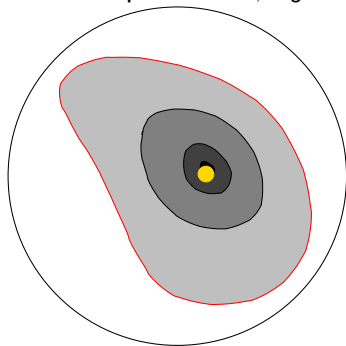
Initialisation

Using inverse fold we find $L + 1$ sequences G_i ($i \in \{0, \dots, L\}$) on the neutral network of the target structure S_0 .

The mutation step

- For each sequence G_i we try a random point mutation restricted to the compatible set of the target structure S_0 , i.e. $G_i \rightarrow G'_i$.
- If $d(S_0, S(G'_i)) \leq i$ we accept the mutation, else we reject it and continue with the old G_i .

Example: $L = 4, G_3$



Monte Carlo approach: The building blocks (II)

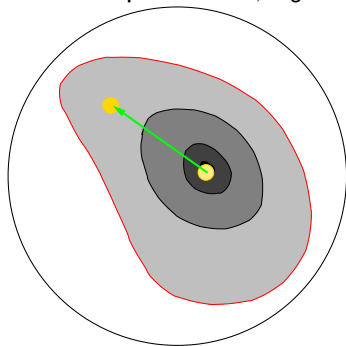
Initialisation

Using inverse fold we find $L + 1$ sequences G_i ($i \in \{0, \dots, L\}$) on the neutral network of the target structure S_0 .

The mutation step

- For each sequence G_i we try a random point mutation restricted to the compatible set of the target structure S_0 , i.e. $G_i \rightarrow G'_i$.
- If $d(S_0, S(G'_i)) \leq i$ we accept the mutation, else we reject it and continue with the old G_i .

Example: $L = 4, G_3$



Monte Carlo approach: The building blocks (II)

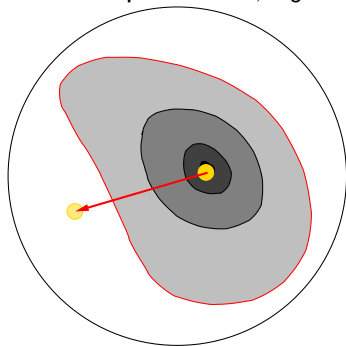
Initialisation

Using inverse fold we find $L + 1$ sequences G_i ($i \in \{0, \dots, L\}$) on the neutral network of the target structure S_0 .

The mutation step

- For each sequence G_i we try a random point mutation restricted to the compatible set of the target structure S_0 , i.e. $G_i \rightarrow G'_i$.
- If $d(S_0, S(G'_i)) \leq i$ we accept the mutation, else we reject it and continue with the old G_i .

Example: $L = 4, G_3$



Monte Carlo approach: The building blocks (II)

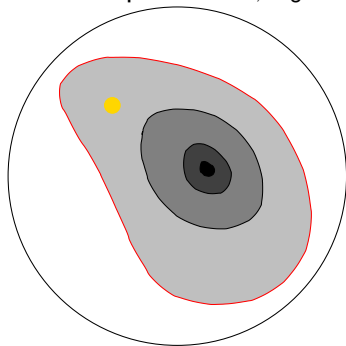
Initialisation

Using inverse fold we find $L + 1$ sequences G_i ($i \in \{0, \dots, L\}$) on the neutral network of the target structure S_0 .

The mutation step

- For each sequence G_i we try a random point mutation restricted to the compatible set of the target structure S_0 , i.e. $G_i \rightarrow G'_i$.
- If $d(S_0, S(G'_i)) \leq i$ we accept the mutation, else we reject it and continue with the old G_i .

Example: $L = 4, G_3$



Monte Carlo approach: The building blocks (II)

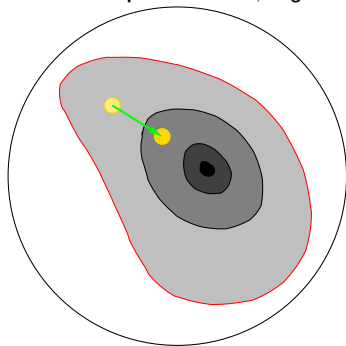
Initialisation

Using inverse fold we find $L + 1$ sequences G_i ($i \in \{0, \dots, L\}$) on the neutral network of the target structure S_0 .

The mutation step

- For each sequence G_i we try a random point mutation restricted to the compatible set of the target structure S_0 , i.e. $G_i \rightarrow G'_i$.
- If $d(S_0, S(G'_i)) \leq i$ we accept the mutation, else we reject it and continue with the old G_i .

Example: $L = 4, G_3$



Monte Carlo approach: The building blocks (II)

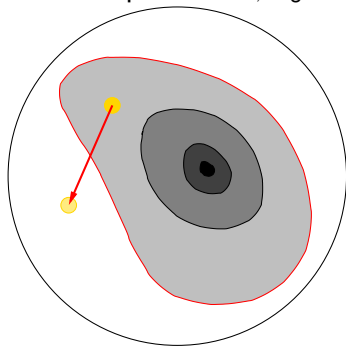
Initialisation

Using inverse fold we find $L + 1$ sequences G_i ($i \in \{0, \dots, L\}$) on the neutral network of the target structure S_0 .

The mutation step

- For each sequence G_i we try a random point mutation restricted to the compatible set of the target structure S_0 , i.e. $G_i \rightarrow G'_i$.
- If $d(S_0, S(G'_i)) \leq i$ we accept the mutation, else we reject it and continue with the old G_i .

Example: $L = 4, G_3$

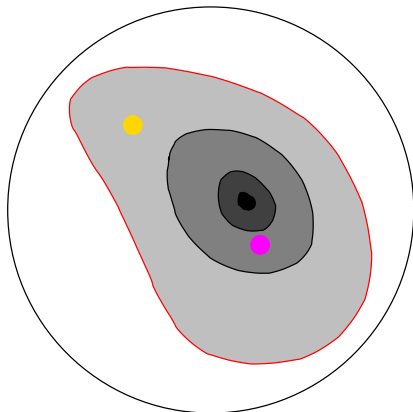


Monte Carlo approach: The building blocks (III)

The exchange step

- After a mutation step on all sequences G_i we try to swap adjacent sequences G_i and G_{i+1} .
- We accept the swap move if the distance of both sequences is smaller or equal to i , else we reject it.

Example: $L = 4$, G_3 and G_4

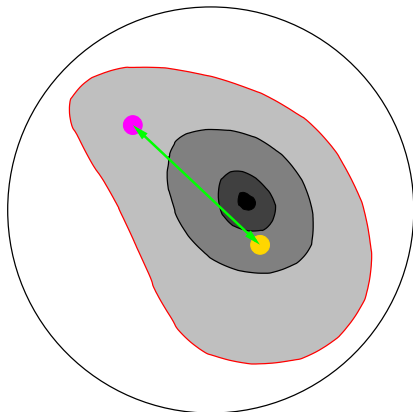


Monte Carlo approach: The building blocks (III)

The exchange step

- After a mutation step on all sequences G_i we try to swap adjacent sequences G_i and G_{i+1} .
- We accept the swap move if the distance of both sequences is smaller or equal to i , else we reject it.

Example: $L = 4$, G_3 and G_4

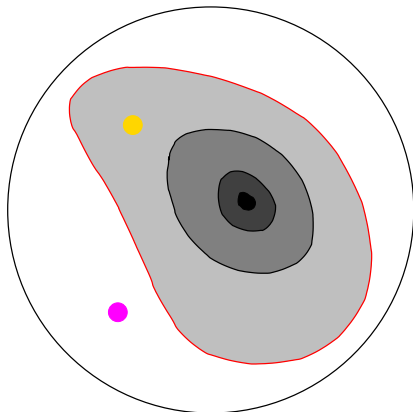


Monte Carlo approach: The building blocks (III)

The exchange step

- After a mutation step on all sequences G_i we try to swap adjacent sequences G_i and G_{i+1} .
- We accept the swap move if the distance of both sequences is smaller or equal to i , else we reject it.

Example: $L = 4$, G_3 and G_4

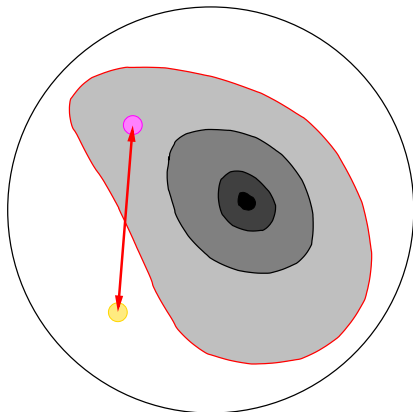


Monte Carlo approach: The building blocks (III)

The exchange step

- After a mutation step on all sequences G_i we try to swap adjacent sequences G_i and G_{i+1} .
- We accept the swap move if the distance of both sequences is smaller or equal to i , else we reject it.

Example: $L = 4$, G_3 and G_4



Monte Carlo approach: The building blocks (III)

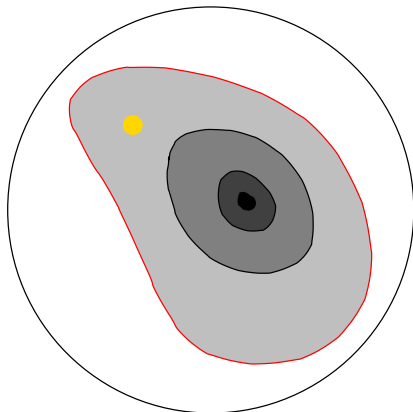
The exchange step

- After a mutation step on all sequences G_i we try to swap adjacent sequences G_i and G_{i+1} .
- We accept the swap move if the distance of both sequences is smaller or equal to i , else we reject it.

Allows to measure

$$r_i = \frac{|V_{i-1}|}{|V_i|} \text{ in a simple way}$$

Example: $L = 4, G_3$

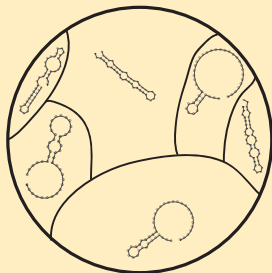


Ranking structures

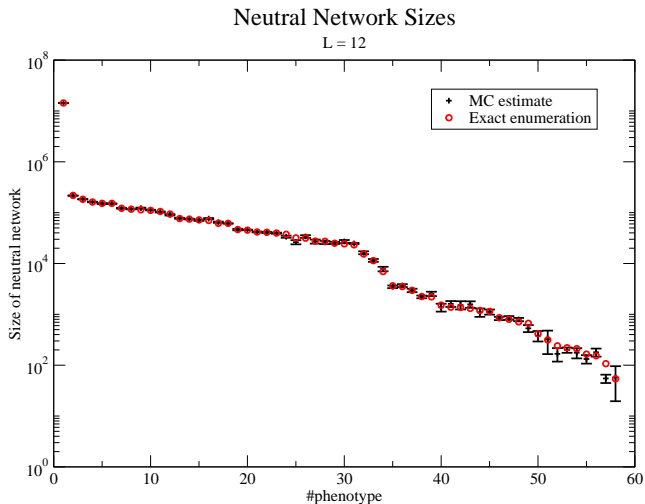
How to estimate the (relative) rank R of a structure S ?

- Choose a sample of M random genotypes of length L .
- Determine their secondary structure S_i ($i \in \{1, \dots, M\}$).
- Determine the size of their neutral networks N_{S_i} .
- A structure S_i is chosen with probability proportional to the size of its neutral network $N_{S_i} \rightarrow$ **Biased sampling**.
- We can correct for this bias using

$$P(S) = \frac{R(S)}{N_{\text{structures}}} = \frac{\sum_{\{i | N_{S_i} > N_S\}} \frac{1}{N_{S_i}}}{\sum_{i=1}^M \frac{1}{N_{S_i}}}$$

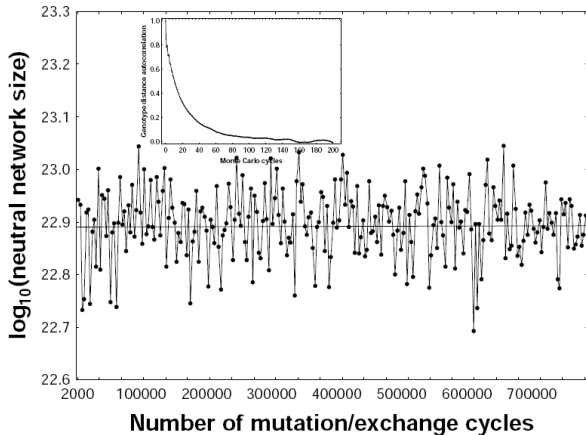
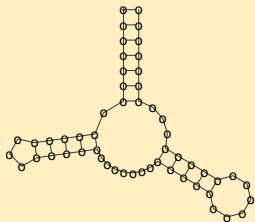


Comparison with exact enumeration



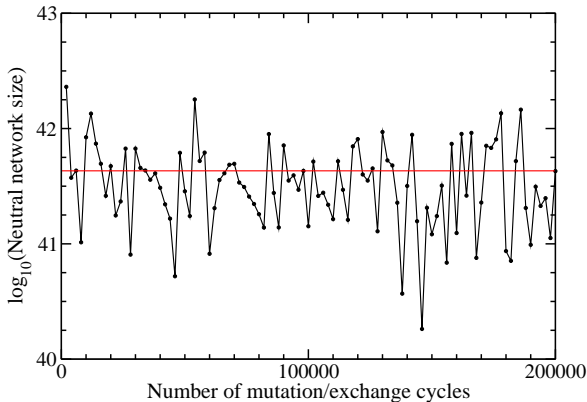
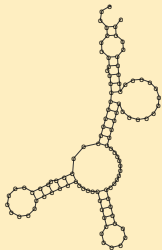
Does it work? (II)

Hammerhead
Ribozyme $L = 54$



Does it work? (III)

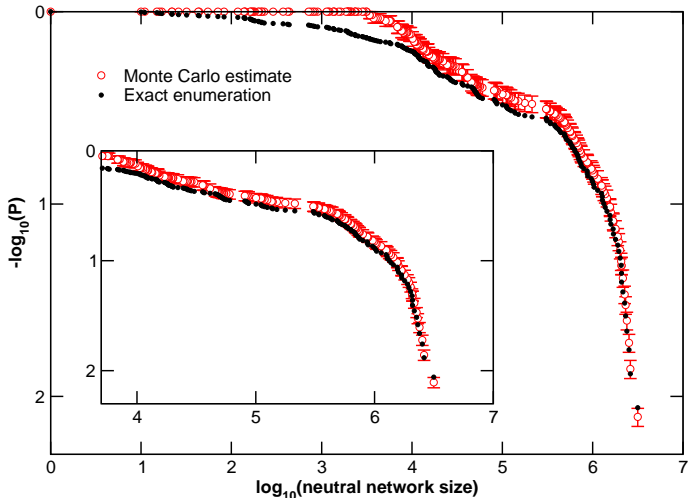
snRNA $L = 100$
(homo sapiens)



Number of compatible sequences = 9.6×10^{49} .
Random search needs at least 10^{10} sequences.

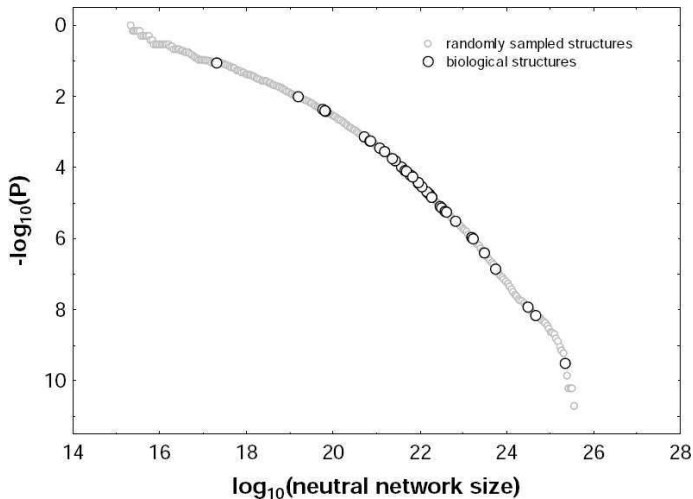
Does it work? (IV)

Comparison with exact enumeration ($L = 14$)



Neutral networks of biological RNAs

Relative rank (p-value) of biological vs random structures
($L = 50$)



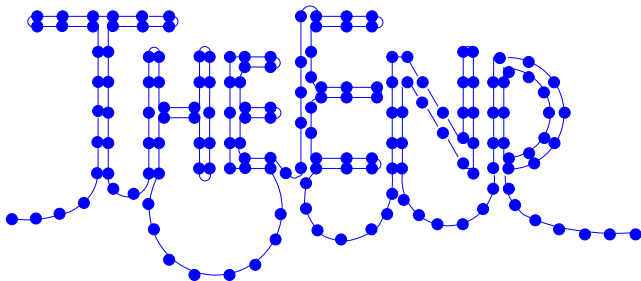
Summary

- The size of neutral networks of biological RNA molecules can be estimated using Monte Carlo techniques.
- Neutral network sizes of biological RNA molecules are atypically large.

Acknowledgements: GENNETEC

Summary

- The size of neutral networks of biological RNA molecules can be estimated using Monte Carlo techniques.
- Neutral network sizes of biological RNA molecules are atypically large.



Acknowledgements: GENNETEC

Outlook

- Inference problems are ubiquitous and even simple statistical questions can lead to computationally subtle challenges
- Data for biological networks are on the rise and methods have not yet reached a high level of maturity (so it is a good time to participate in the field)
- Techniques from other fields (combinatorial optimization, statistical physics...) can help: don't reinvent the wheel
- With the very recent (re)sequencing technologies, loads of data will beg for analysis: a good combination of modeling, statistics and bioinformatics will allow extraction of very interesting biological results
- One can expect that in the near future, relevant computational tools for biological networks will become « turn-key » and easily shared within the community.