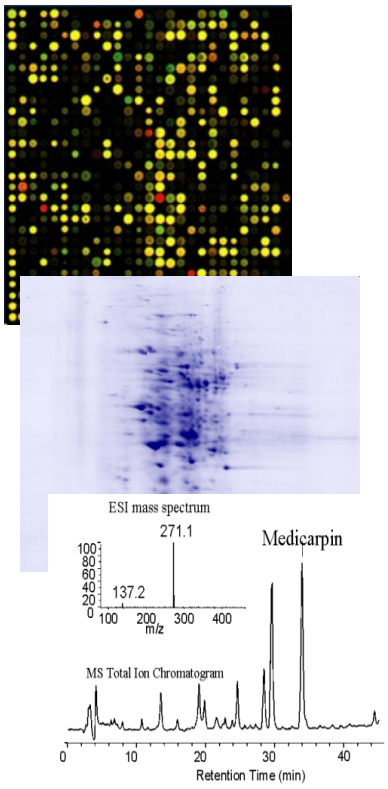
Simulating Systems Genetics for algorithm evaluation

Alberto de la Fuente
alf@crs4.it

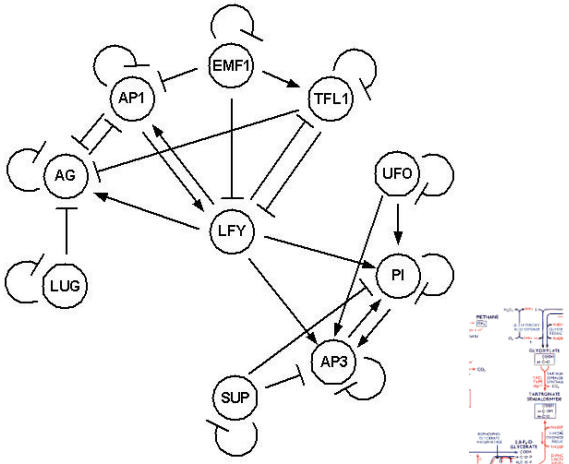
Inferring Regulatory Networks = inverse problem = system identification

"~omics" data

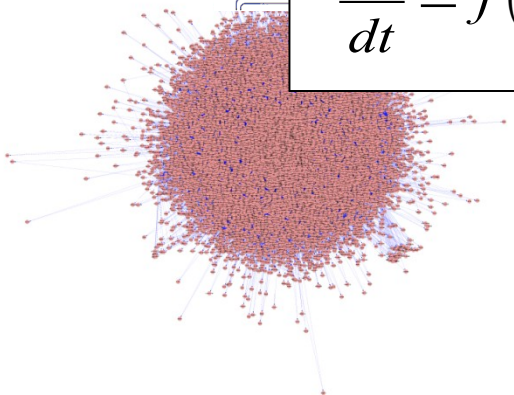


algorithms

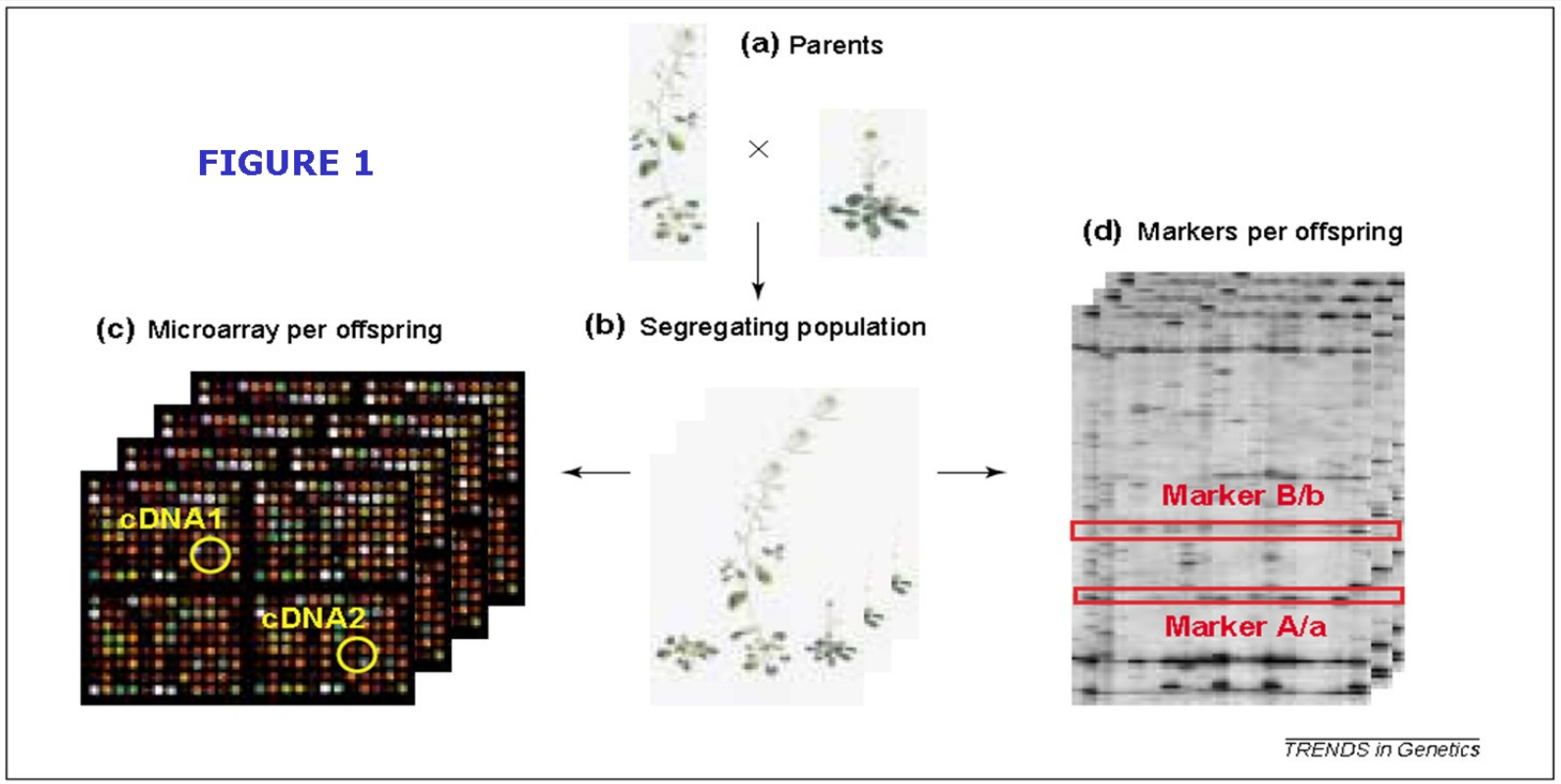
Correlation, partial correlation, regression, linear Ordinary Differential Equations, graphical Gaussian models, perturbation analysis...



$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \mathbf{k})$$



Genetical Genomics



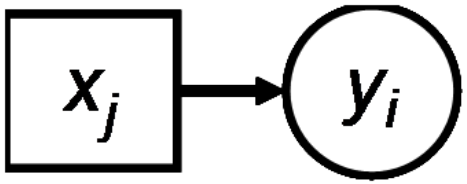
Jansen, R.C., and Nap, J.P. (2001) Trends Genet. 17, 388-391

Inferring Networks from SG data

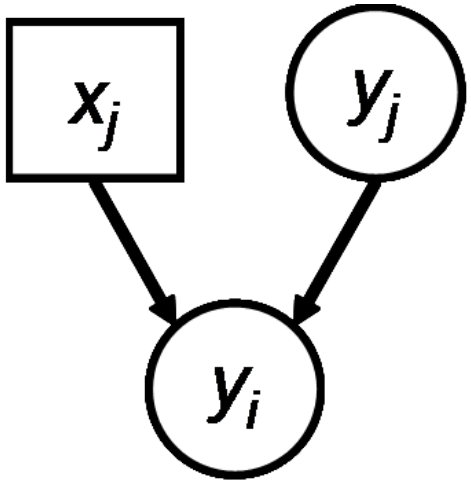
Gene Network inference requires many perturbations

Experimental perturbations are difficult and costly

Use of naturally occurring genetic variations (perturbations)



$$y_{in} = b_0 + b_1 x_{jn} + \varepsilon_{in}$$



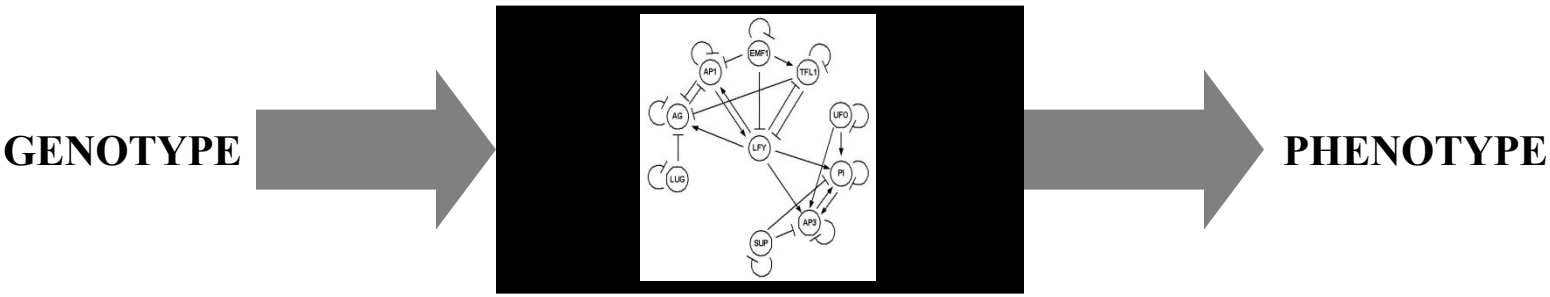
$$y_{in} = b_0 + b_1 y_{jn} + b_2 x_{jn} + \varepsilon_{in}$$

x = genotype data (e.g. SNPs)
y = gene expression 'phenotypes'

Systems Genetics (SG)

Statistical Genetics

Systems Genetics



How to evaluate the results of analysis?

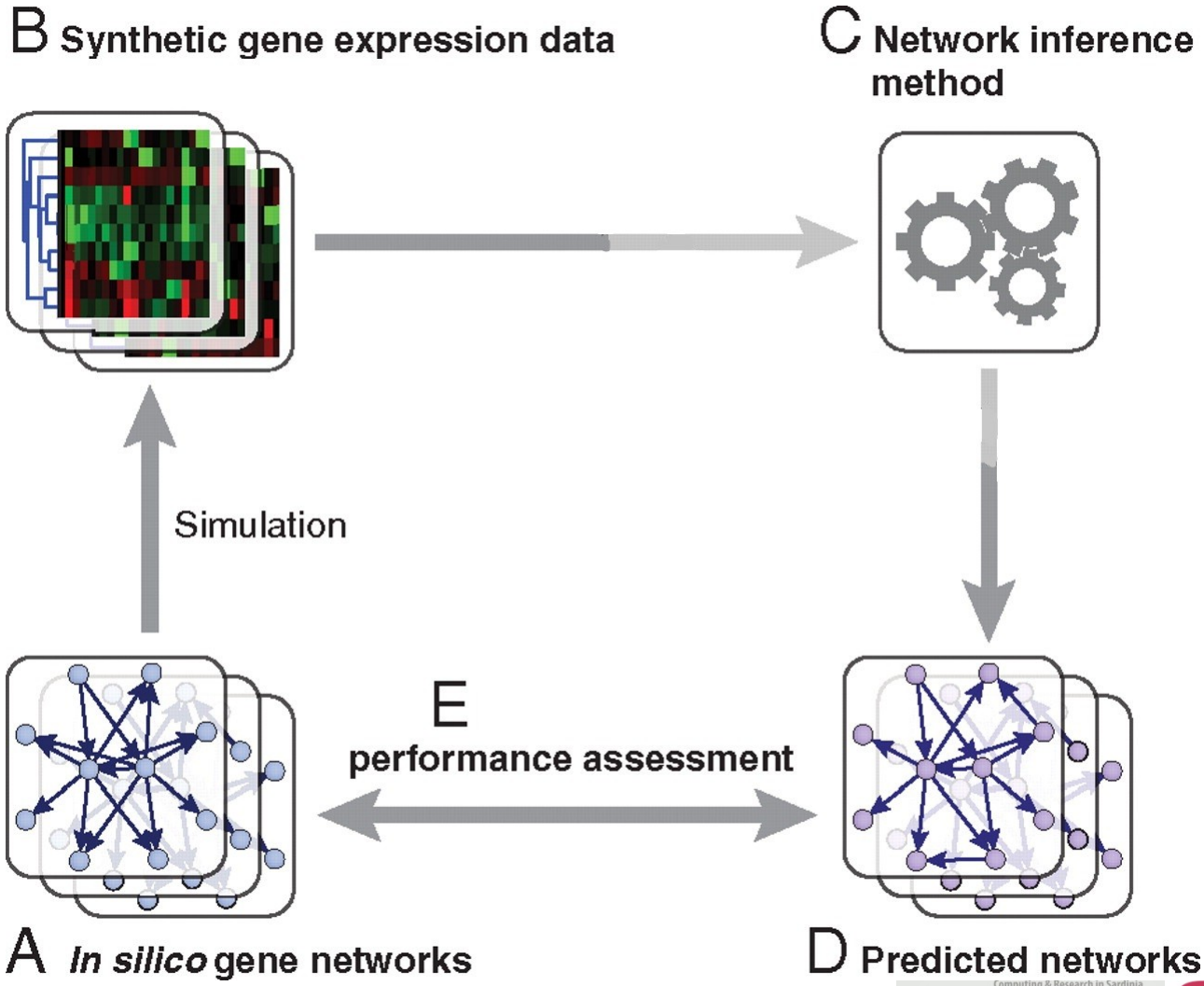
- Ideally, to perform experiments to validate the generated hypotheses

- Interpretation based on prior knowledge
  - Gene Ontology
  - ‘Known’ networks → Bronze Standards

Need for Gold Standard benchmarks!

- **Simulated data**

How to evaluate the results of analysis?

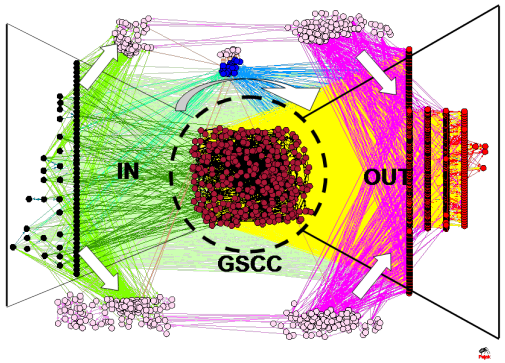


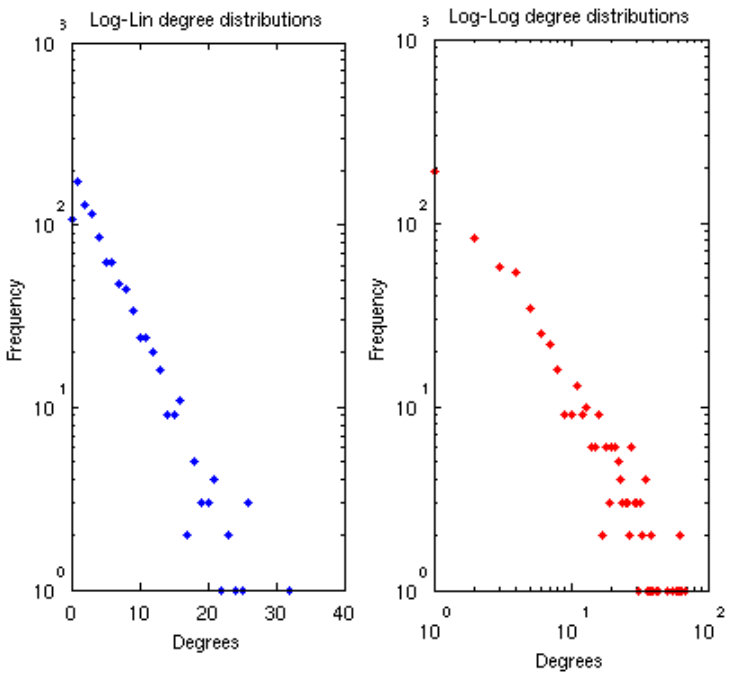
SysGenSIM: Simulating Systems Genetics

The STATSEQ benchmark

- Goal:
  - Unbiased evaluation of network inference algorithms
- Main questions:
  - Which algorithms are most effective?
  - How does network size affect the ability to infer networks?
  - How does genetic linkage between affect the ability to infer networks?
  - How does 'heritability' of transcripts affect the ability to infer networks?
  - How does population size affect the ability to infer networks?

The STATSEQ benchmark: Networks





- 3 networks of 100 genes
- 3 networks of 1000 genes
- 3 networks of 5000 genes

- Average node degree = 6 (3 inputs + 3 outputs)

- Topology: 'scale-free out degree and exponential in-degree distributions' were generated
  - An out-degree sequence was generated by sampling from a power-law
  - An in-degree sequence was generated by sampling from an exponential distribution
  - Nodes were connected satisfying their assigned in and out degree

- Signs (activation or repression) were assigned uniformly 50:50 chance

The STATSEQ benchmark: Gene expression data simulation

Cis-effect: a polymorphism in promotor region of gene G_g affects the basal transcription rate.

Z_{cg} for allele 'A' ≠ Z_{cg} allele 'a'.

Trans-effect: a polymorphism in coding region of gene G_k affects the way it affects its targets.

Z_{tk} for allele 'B' ≠ Z_{tk} allele 'b'.

$$\frac{dG_g}{dt} = v_{transcription_{G_g}} - v_{degradation_{G_g}} = Z_g^c \cdot V_g \cdot \theta_g^{syn} \cdot \prod_{k \in R_g} \left(1 + A_{gk} \frac{G_k^{h_{gk}}}{G_k^{h_{gk}} + (K_{gk} / Z_k^t)^{h_{gk}}} \right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$

All parameters set to 1 for simplicity, except:

Z ~ Uniform[0.5, 0.8] for one allele and 1 for the other allele
Hill cooperativity coefficient ~ Gamma[1,1.67]

Transcription biological variance ~ Gaussian[1, x]
Degradation biological variance ~ Gaussian[1, x]

The STATSEQ benchmark: Gene expression data simulation

Cis-effect: a polymorphism in promotor region of gene G_g affects the basal transcription rate.

Z_{cg} for allele 'A' ≠ Z_{cg} allele 'a'.

Trans-effect: a polymorphism in coding region of gene G_k affects the way it affects its targets.

Z_{tk} for allele 'B' ≠ Z_{tk} allele 'b'.

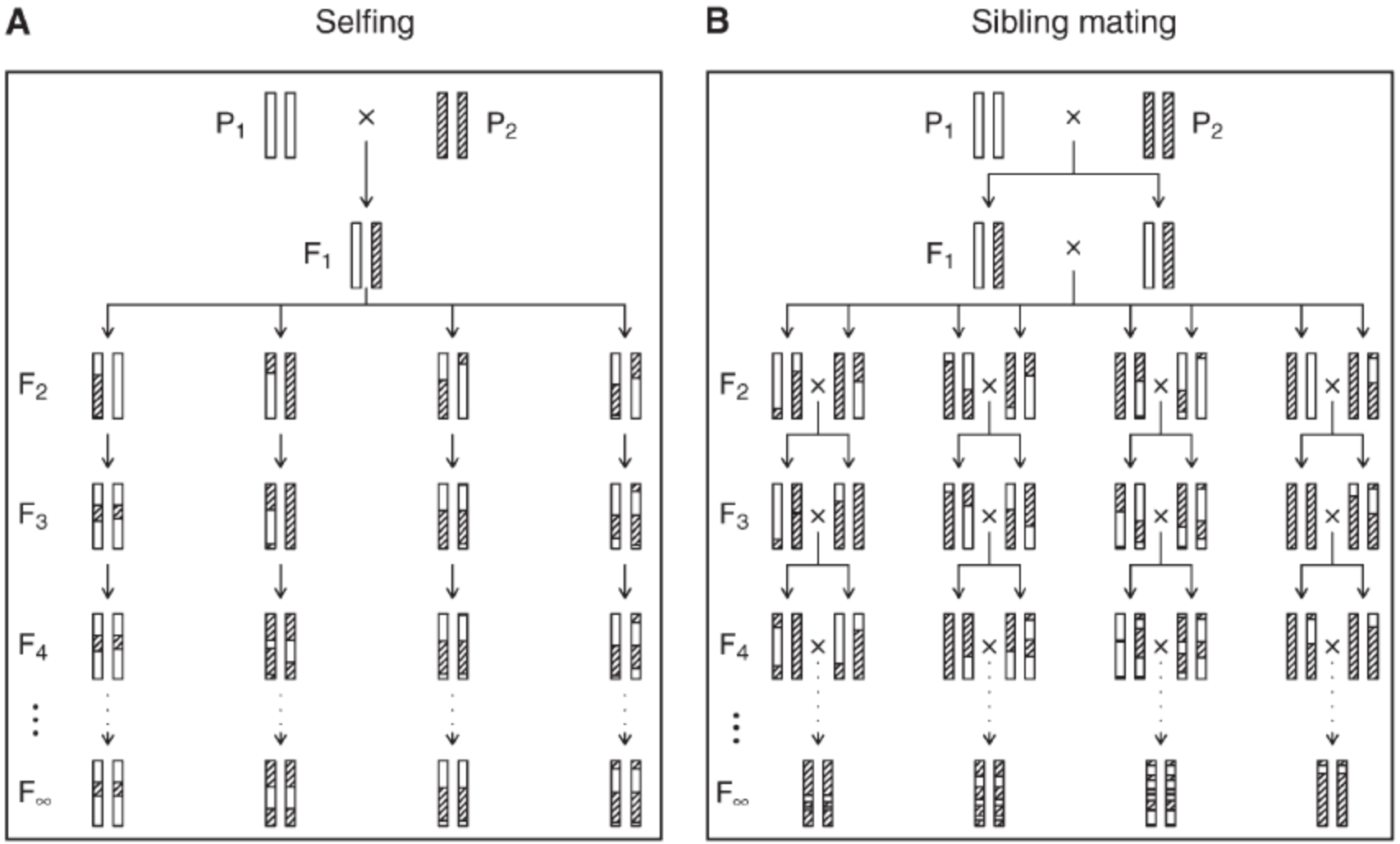
$$\frac{dG_g}{dt} = v_{transcription_{G_g}} - v_{degradation_{G_g}} = Z_g^c \cdot V_g \cdot \theta_g^{syn} \cdot \prod_{k \in R_g} \left(1 + A_{gk} \frac{G_k^{h_{gk}}}{G_k^{h_{gk}} + (K_{gk} / Z_k^t)^{h_{gk}}} \right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$

- Steady state gene-expression levels for all individuals were calculated after adjusting the Zs to the 'genotype data' and sampling the thetas.
  - Population size = [300, 900]

- Experimental noise was multiplied with the steady state values.
  - Gaussian[1,0.1]

	RIL1	RIL2	RIL3	RIL4
Gene 1	0.7650440	0.8454307	0.6584363	0.8969617
Gene2	0.2472525	0.1686617	0.0707179	0.2813917
Gene3	0.1324578	0.0878289	0.1704999	0.4884665
Gene4	0.0988393	0.0334034	0.0774731	0.3013401
Gene5	0.1537819	0.1337609	0.2049485	0.2143432
Gene6	0.9966201	0.8622003	0.5491228	0.6797666
Gene7	0.5078467	0.5678145	0.4306351	0.3840879

The STATSEQ benchmark: RIL population



K. W. Broman (2005) Genetics 169: 1133–1146

The STATSEQ benchmark: Chromosome

Cis-effect: a polymorphism in promotor region of gene G_g affects the basal transcription rate.

Z_{cg} for allele 'A' ≠ Z_{cg} allele 'a'.

Trans-effect: a polymorphism in coding region of gene G_k affects the way it affects its targets.

Z_{tk} for allele 'B' ≠ Z_{tk} allele 'b'.

$$\frac{dG_g}{dt} = v_{transcription_{G_g}} - v_{degradation_{G_g}} = Z_g^c \cdot V_g \cdot \theta_g^{syn} \cdot \prod_{k \in R_g} \left(1 + A_{gk} \frac{G_k^{h_{gk}}}{G_k^{h_{gk}} + (K_{gk} / Z_k^t)^{h_{gk}}} \right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$

- 100 gene-networks → 5 chromosomes with each N(20, 2) genes/markers
- 1000 gene-networks → 25 chromosomes with each N(40, 2) genes/markers
- 5000 gene-networks → 25 chromosomes with each N(200, 2) genes/markers

•Each gene has a polymorphism, either in the gene's promoter region (leading to a 'cis effect' on its own expression rate) or in the gene's coding region (leading to 'trans effects' on its targets): **probability ¼ vs ¾.**
(T,T,T,C,C,T,T,C,T,T……,T,T)

- The molecular effect of each locus is decided:
  either the 0 allele → Z=1, or the 1 allele → Z=1, decided by 'flip of coin'.
  (0, 0,1,0,0,1,0,1,1,1……,0,1)

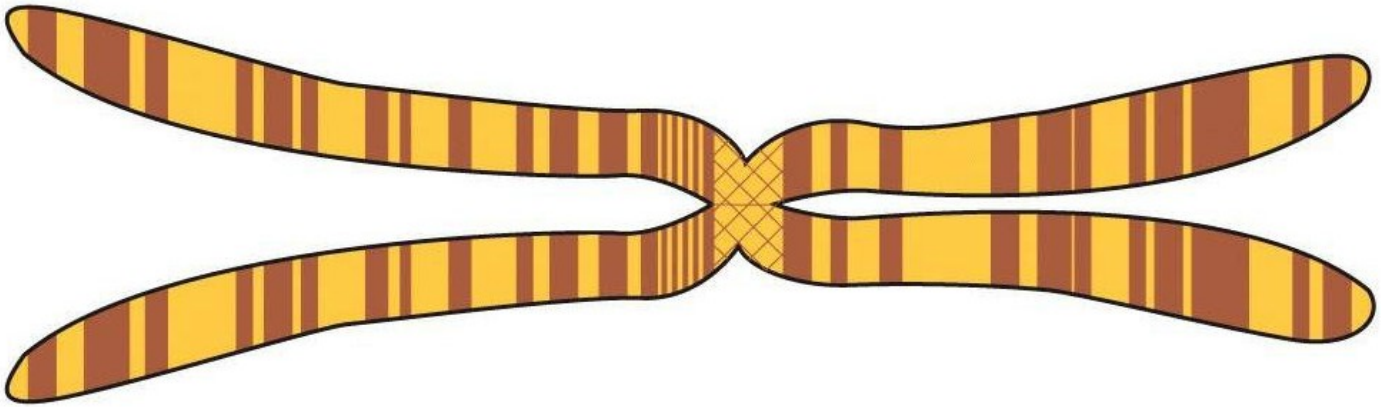
The STATSEQ benchmark: Genetic map

- Markers/genes are placed on the chromosomes, at distances ***d*** sampled from:

  N(1 centiMorgan, 0.2 cM), leading to STRONG linkage

  or

  N(5 cM, 1 cM), leading to WEAK linkage



The STATSEQ benchmark: Genotype data simulation

a. Generate the marker distances d by sampling the values according to the selected distribution (i.e. small or large).
b. Set the gene positions at markers, i.e. the genetic distances between genes are the same between markers.
c. Convert[10] the distances d in recombination rates r.
d. Transform[11] the recombination rates r in the probabilities p for which a genotype X_i corresponding to gene i is equal to the genotype X_{i-1} of gene X_{i-1}.
e. Generate the genotype vector $\mathbf{X}^h$ as:
   i. Randomly set X_1 to 0 or 1 with the same probability.
   ii. Set the following genotype binary values by rounding to the nearest integer $X_k = [(2X_{k-1} - 1)(p_k - 0.5) + \rho]$, where ρ is sampled from the $[0, 1]$ uniform distribution.

[10] According to Haldane: $r = 0.5(1 - e^{-0.02d})$, or to Kosambi: $r = 0.5(e^{0.04d} - 1)/(e^{0.04d} + 1)$.

[11] According to the *selfing* inbred line cross: $p = 1/(1 + 2r)$, or to the *sibling-mating*: $p = (1 + 2r)/(1 + 6r)$.

	RIL1	RIL2	RIL3	RIL4
Gene 1	0	0	1	1
Gene2	0	0	1	1
Gene3	0	0	1	1
Gene4	0	0	1	1
Gene5	0	0	1	1
Gene6	0	1	1	1
Gene7	0	0	1	1

The STATSEQ benchmark: Heritability

Cis-effect: a polymorphism in promotor region of gene G_g affects the basal transcription rate.

Z_{cg} for allele 'A' ≠ Z_{cg} allele 'a'.

Trans-effect: a polymorphism in coding region of gene G_k affects the way it affects its targets.

Z_{tk} for allele 'B' ≠ Z_{tk} allele 'b'.

$$\frac{dG_g}{dt} = v_{transcription_{G_g}} - v_{degradation_{G_g}} = Z_g^c \cdot V_g \cdot \theta_g^{syn} \cdot \prod_{k \in R_g} \left(1 + A_{gk} \frac{G_k^{h_{gk}}}{G_k^{h_{gk}} + (K_{gk} / Z_k^t)^{h_{gk}}} \right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$

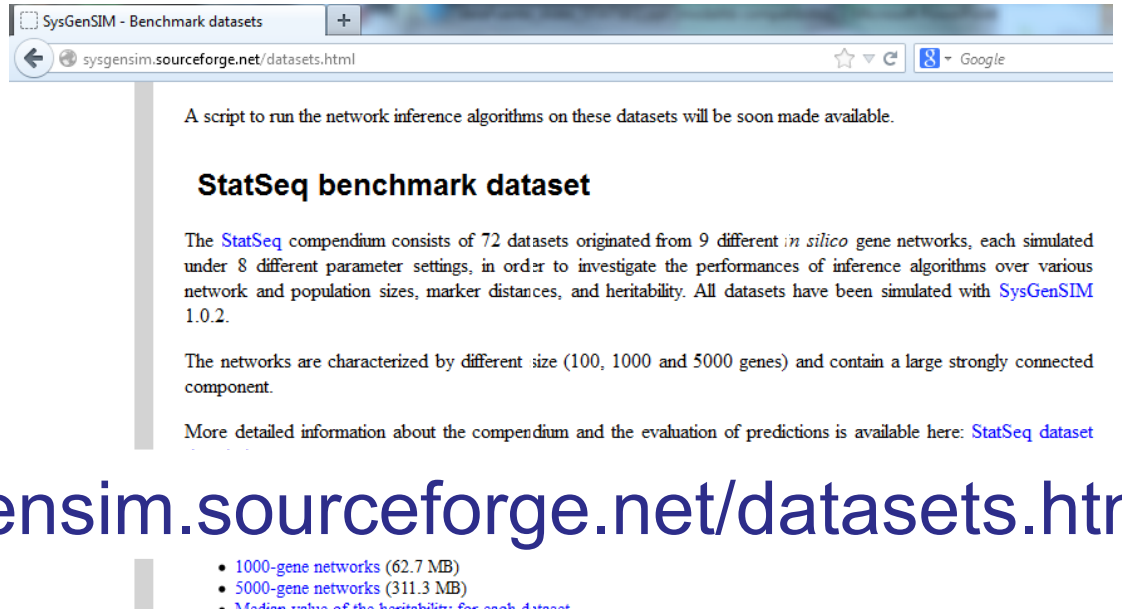
Transcription biological variance (thetas) are sampled either from

N[1, 0.1] (small biological variance, leading to HIGH heritability, median ≈ 75%)
or
N[1, 0.25] (large biological variance, leading to LOW heritability, median ≈ 35%)

$$H_i = \frac{variance\ of\ gene\ i\ expression\ with\ only\ genetic\ variability}{variance\ of\ gene\ i\ expression\ with\ both\ genetic\ variability\ and\ biological\ variability\ (\theta)}$$

The STATSEQ benchmark

72 datasets =
3 network sizes ×
3 replicates ×
2 population sizes ×
2 heritability levels ×
2 linkage strengths



http://sysgensim.sourceforge.net/datasets.html

The STATSEQ benchmark

- Coming up (today and tomorrow):
  - Presentations on applications of algorithms on the benchmark
    - 8 research groups (from France, Belgium, Germany, Netherlands, Italy and the USA) were given the benchmark and the networks' Gold standard to evaluate and optimize their algorithms.

The STATSEQ benchmark

- Final questions:
  - Which algorithms are most effective?
  - How does network size affect the ability to infer networks?
  - How does genetic linkage between affect the ability to infer networks?
  - How does 'heritability' of transcripts affect the ability to infer networks?
  - How does population size affect the ability to infer networks?
  - **How realistic is the simulated benchmark??**

Simulating Systems Genetics with SysGenSIM

Andrea Pinna[1], Nicola Soranzo[1], Ina Hoeschele[2,3], Alberto de la Fuente[1]

[1] Center for Advanced Studies, Research and Development (CRS4) Bioinformatica, Pula, Italy
[2] Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA
[3] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA

CRS4 IDEAS BECOME LIFE

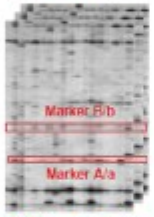
CRS4 bioinformatica

Overview

The central goal of **systems biology** is to gain a predictive, system-level **understanding of biological networks**. This entails **inferring causal networks** from observations on a perturbed biological system. An ideal experimental design for causal inference is randomized, **multifactorial** perturbation.

The recognition that the genetic variation in a segregating population represents randomized, multifactorial perturbations gave rise to **Genetical Genomics** and **Systems Genetics**, where a segregating or genetically randomized population is **DNA marker genotyped**, profiled for **phenotypes of interest** and for **genome-wide gene transcription**, and potentially profiled for protein expression, metabolomics, DNA methylation, etc.

- Gene expression levels
- Gene activities



- Genotypes
- Genetic markers linked to genetic polymorphisms
- Mutations
- Genetic perturbations

- DNA sequence
- Gene positions
- Genetic markers

The toolbox

SysGenSIM is an open source **MATLAB toolbox** to simulate **systems genetics** and **single-gene perturbation** experiments. By browsing through few tabs, the user can:

- Generate **large gene networks** (> 10000 genes) with alternative topologies.
- Define the parameters of the genetic (or perturbation) **experiment**.
- Select the **model** parameters.
- Select the **desired output** files and figures.

Gene expression levels and phenotypes are described using **nonlinear differential equations** based on biochemical kinetics. Parameters are adjusted according to **genotypes** of a simulated population. Gene expression and phenotype steady state values are calculated for each genotype. The resulting data can be used for **evaluation purposes**.

The program allows for flexibility and **realistic choices** for parameter settings (e.g. probabilities for **cis-** and **trans- acting** functional polymorphisms, **heritabilities** of expression and phenotype traits).

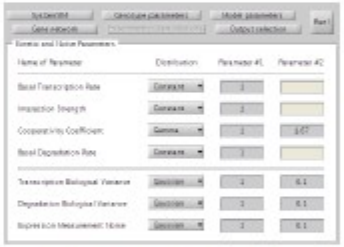
Another relevant feature is the reproduction of single-gene **knock-out**, **knock-down**, and **over-expression** experiments.

Model equation

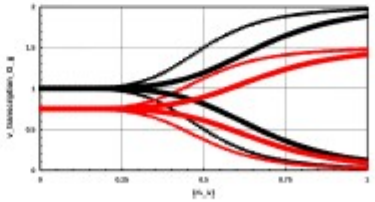
- Steady-state gene expression profiles are simulated with **nonlinear ODEs**.
- Transcription rate shows two features of biochemical kinetics: **saturation** and **cooperativity**.
- RNA decay is assumed as a **first order process**.

$$\frac{\mathrm{d}\,G_g}{\mathrm{d}\,t} = Z_g^{\varepsilon} \cdot V_g \cdot \theta_g^{\eta_g} \cdot \prod_k \left[1 + A_{k,g} \frac{G_k^{h_{k,g}}}{G_k^{h_{k,g}} + (K_{k,g}/Z_k^{t})^{h_{k,g}}} \right] - \lambda_g \cdot \theta_g^{\delta\eta_g} \cdot G_g$$

- G_g is the **mRNA concentration** of gene g.
- V_g is its **basal transcription rate**.
- λ_g is the **degradation rate** constant.
- $K_{k,g}$ is the **interaction strength**.
- $h_{k,g}$ is a **cooperativity** coefficient.
- $A_{k,g}$ is an element of the **adjacency matrix** A.
- $\theta_g^{\eta_g}$ and $\theta_g^{\delta\eta_g}$ represent **non-genetic biological noise** in transcription and degradation rates.
- Z_g^{ε} and Z_k^{t} incorporate the **effects of DNA polymorphisms** in the model.



Values for all **kinetic and noise** parameters can be **sampled** from constant, uniform, Gaussian and Gamma distributions, and such that the estimated **heritabilities** of the expression profiles are close to those found in **real data**.



Transcription rate of the target gene g is plotted as a **function** of the source gene k at **different values** of Z_g^{ε} and Z_k^{t} (other parameters are constant):

- Black squares $Z_g^{\varepsilon} = 1$ and $Z_k^{t} = 1$.
- Red squares $Z_g^{\varepsilon} = 0.75$ and $Z_k^{t} = 1$.
- Black circles $Z_g^{\varepsilon} = 1$ and $Z_k^{t} = 0.75$.
- Red circles $Z_g^{\varepsilon} = 0.75$ and $Z_k^{t} = 0.75$.

Efficient steady state solving

Directed networks generally have a **bow-tie structure** with three major components: a **strongly connected cyclic component** and three **acyclic** parts: in- and out-components and tendrils.

- Steady states for genes in the acyclic