MLSB 07 Evry, September 24th

Artificial Regulatory Network Evolution

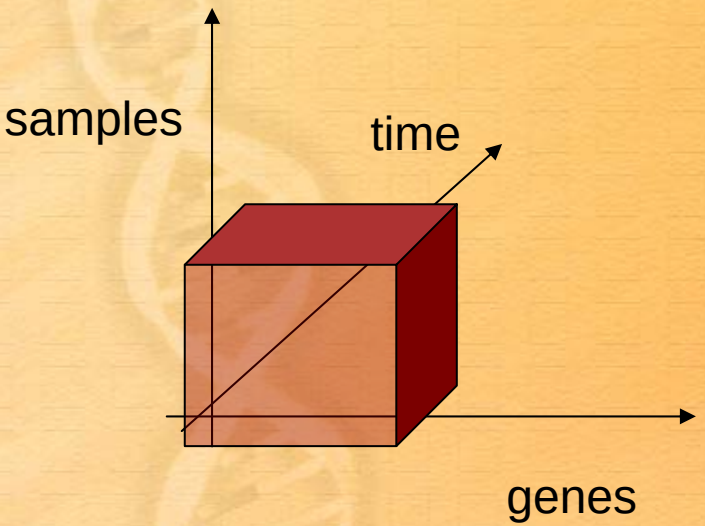
Yolanda Sanchez-Dehesa[1] , Loïc Cerf[1], José-Maria Peña[2], Jean-François Boulicaut[1] and Guillaume Beslon[1]

[1]: LIRIS Laboratory, INSA-Lyon, France

[2]: DATSI , Facultad de Informatica, Universidad Politécnica de Madrid, Spain

Context: from data to knowledge

- Large kinetic transcriptome data sets are announced



- Genetic Network (GN) inference
- We need to design NOW the related data mining algorithms

Problems

- Just a few real data sets are available
- Today, benchmarking is performed on:
  - Randomly generated data
  - Synthetic data w.r.t. models from other fields
  - Data from GN generators biased by topology

Approach

- Can we use simulation to build biologically plausible GNs and thus more relevant kinetic data sets?
- GNs are built by an evolutionary process
- We propose to use artificial evolution to generate plausible GNs

Biologically plausible GN

- To obtain plausible GNs we must respect biological bases of network evolution:
  - GNs are derived from a genome sequence and a proteome component
  - Mutation of the genetic sequence
  - Selection on the phenotype

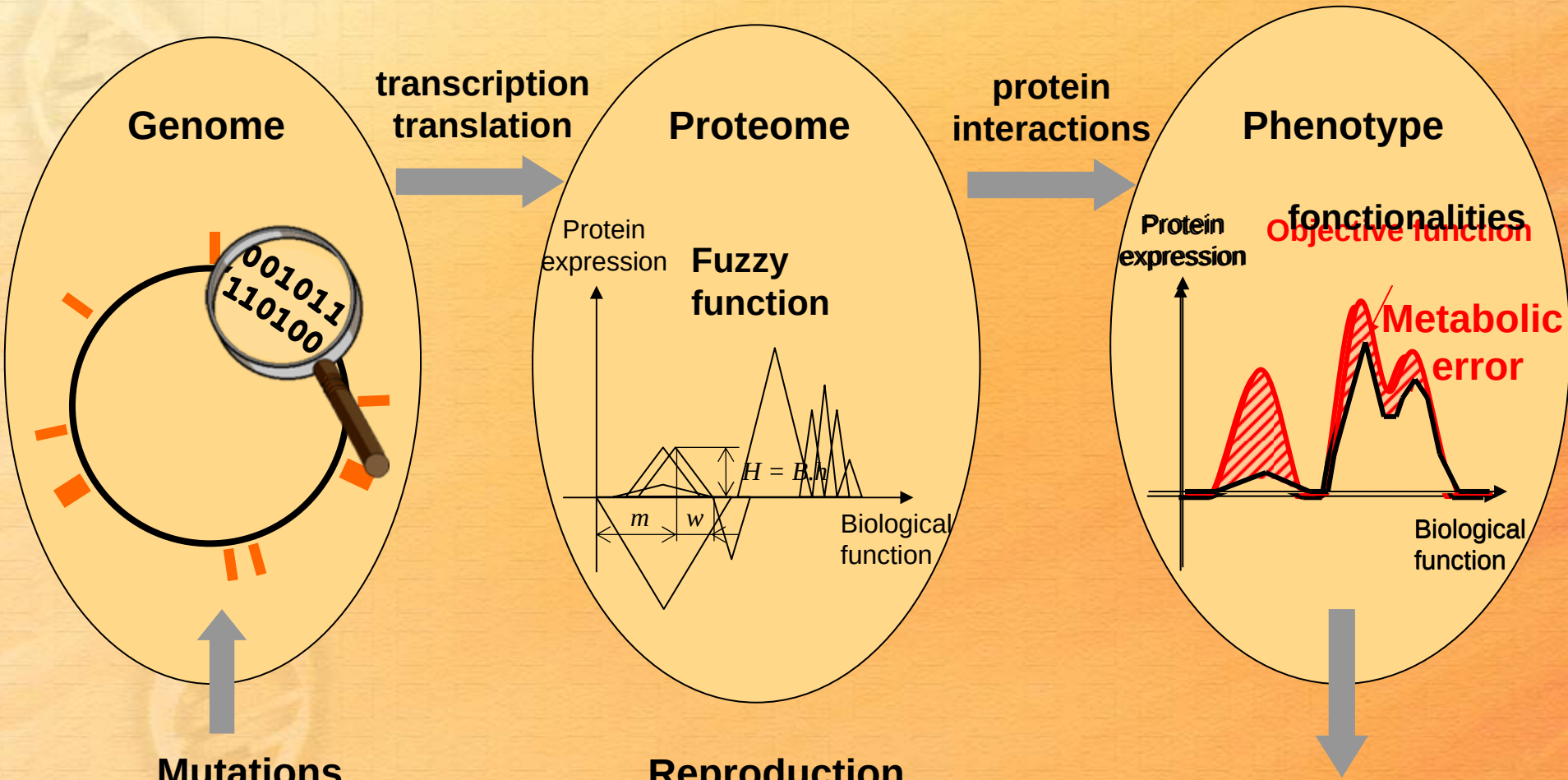
We have developed the RAevol Model

Based on the Aevol* Model

- Studying robustness and evolvability in artificial organisms:
  - Artificial genome, non-coding sequences, variable number of genes
  - Genome: circular double-strand binary string
  - Mutation/selection process

* C. Knibbe PhD, INSA-Lyon, October 2006

Ævol : Artificial Evolution



From Aevol to RAevol

- Interesting properties of Aevol to understand genome evolution:
  - See **C. Knibbe**, A long-term evolutionary pressure on the amount of non-coding DNA (2007). *Molecular Biology and Evolution*, in press. doi: 10.1093/molbev/msm165
- We need to add a regulatory process

  → RAevol

RÆvol : Artificial Evolution

Regulation

The phenotype becomes a dynamic function …

Genome

transcription
translation

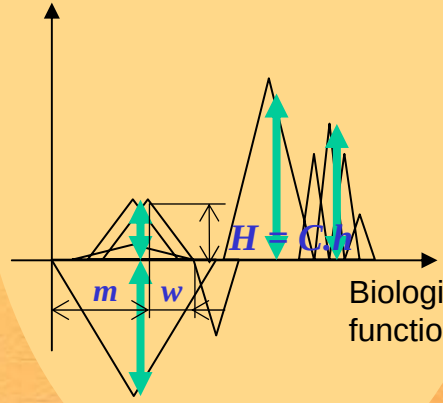
Proteome

protein interactions

Phenotype

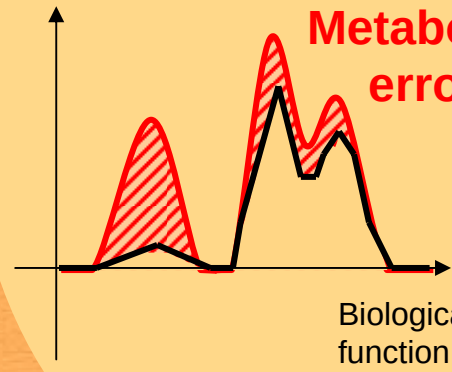
Expression level

Fuzzy logic function



Expression level

Metabolic error



$$\frac{\delta a_i}{\delta t} = basal\_level(1 + (e_i - in_i) \cdot c_i) - \phi$$
$$e_i = \frac{1}{N} \sum_j c_j e^{\beta(u_j^+ - u_{max}^+)}$$
$$in_i = \frac{1}{N} \sum_j c_j e^{\beta(u_j^- - u_{max}^-)}$$

Mutations

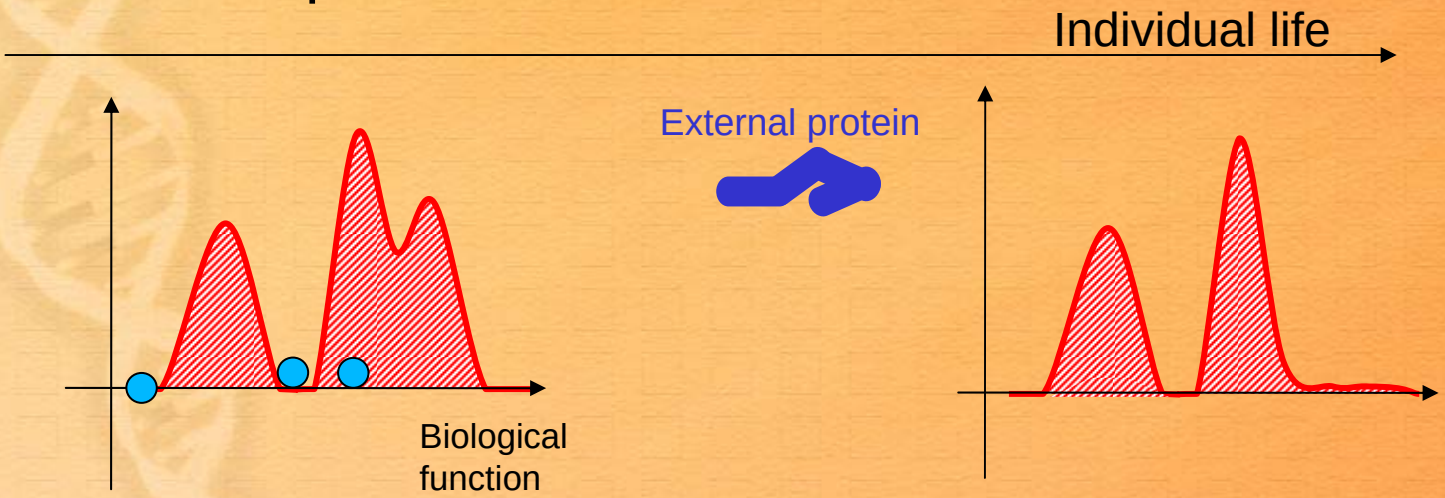
Reproduction

Selection

Banzhaf03:On Evolutionary Design, Embodiment and Artificial Regulatory Networks

Experimental setup

- **Simulations:** 1000 individuals, mutation rate 1.10^{-5} , 15000 generations
- Organisms must perform 3 metabolic functions
- The incoming of an external signal (protein) triggers an inhibition process



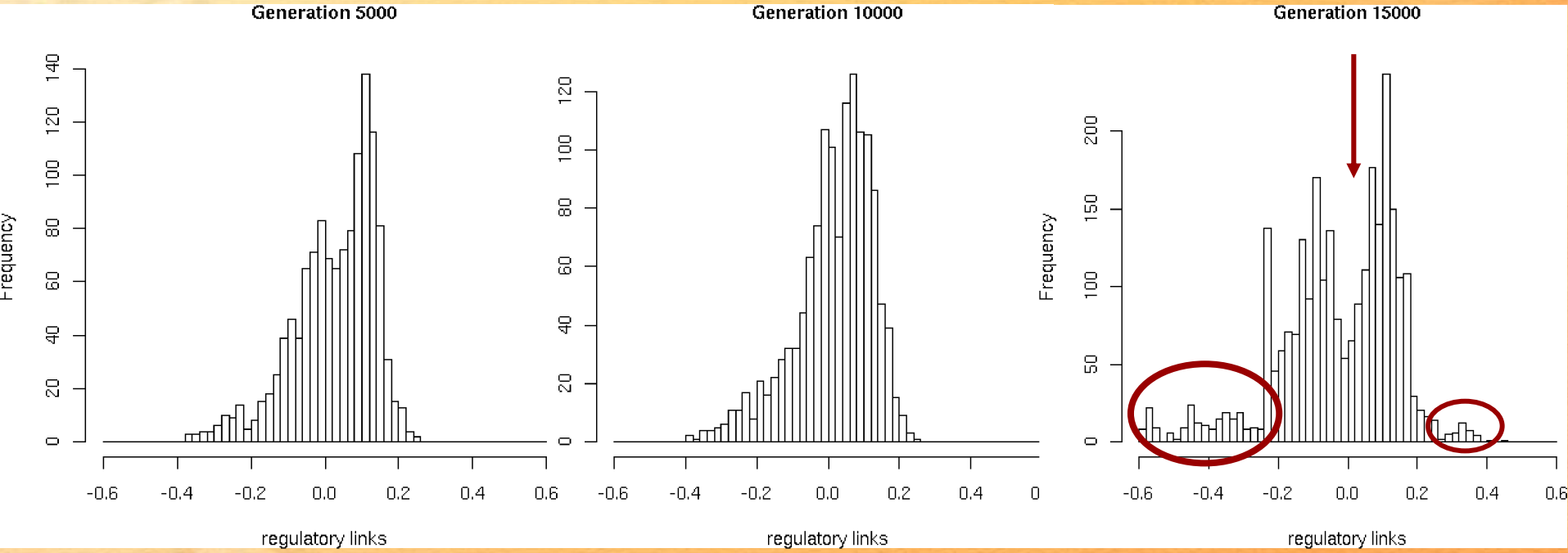
First results

Generation	0	1000	5000	10000	15000
Metabolic error	0.1377	0.0616	0.0206	0.0171	0.0161
Genome size (in kilobase pairs)	5.0	21.3	15.3	12.3	12.8
Nb. of nodes in the metabolic network	1	24	34	36	37
Nb. of nodes in the genetic network	3	24	35	37	45
Nb. of transcription factors	2	0	1	1	8
Nb. of links	3	576	1223	1332	2601
positive links	2	508	771	827	1276
negative links	1	68	452	505	1325

- The metabolic network mainly grows during the 5000 first generations → GN grows likely
- Transcription factors appear after 10000 generations
  → GN grows independently from metabolism

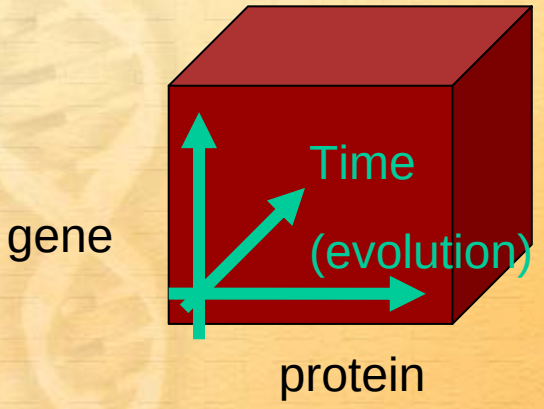
First results

Regulatory Links Values

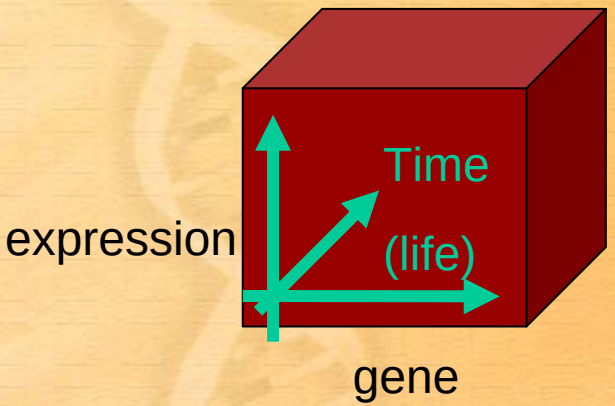


- First phase: quasi-normal distribution
- Second phase: multimodal distribution, strong links (mainly inhibitory) …

Conclusion and perspectives



- RAevol generates plausible GNs (protein-gene expression levels) along evolution



- Studying the generation of kinetic transcriptome data sets is ongoing

- Towards more realistic benchmarks for data mining algorithms

Open issues

- Systematic experiments
  - → effect of mutation rates
  - → effect of environment stability
- Study the network topology
  - →Compare the network topology with real organisms…
  - →Do frequent motifs/modules appear in the network ?

The Aevol Model

- **Interesting properties** of the Aevol Model:
  - Transcription/translation process → Different RNA production levels
  - Explicit (abstract) proteome → interactions between proteins and genetic sequence
  - Variable gene number → Variable network size
  - Complex mutational process (mutations, InDel, rearrangements, …) → Different topology emergence