Thesis Topics

Paolo Milazzo

University of Pisa

A.A. 2018-2019

Introduction

The recent developments in biology have produced a huge amount of data about the structure of living matter;

- consider as an example the success of the Human Genome Project

Less is known about the versatile functions that cells and their components show.

In the last few years the scientific interest has started to move from structures to functionalities

The complexity of the cellular processes has stimulated the growth of a new paradigm, that moves from the classical reductionist approach to a system level understending of life

- Such a paradigm is called systems biology

Introduction

A better understanding of the funcitoning of cellular processes may allow:

- a better undertanding of diseases
- the development of more effective drugs
- the development of preventive and early diagnosis techniques

Mathematical and computational modelling may contribute to the study of cellular processes with simulation tools that, based on data from laboratory experiments, could be used to:

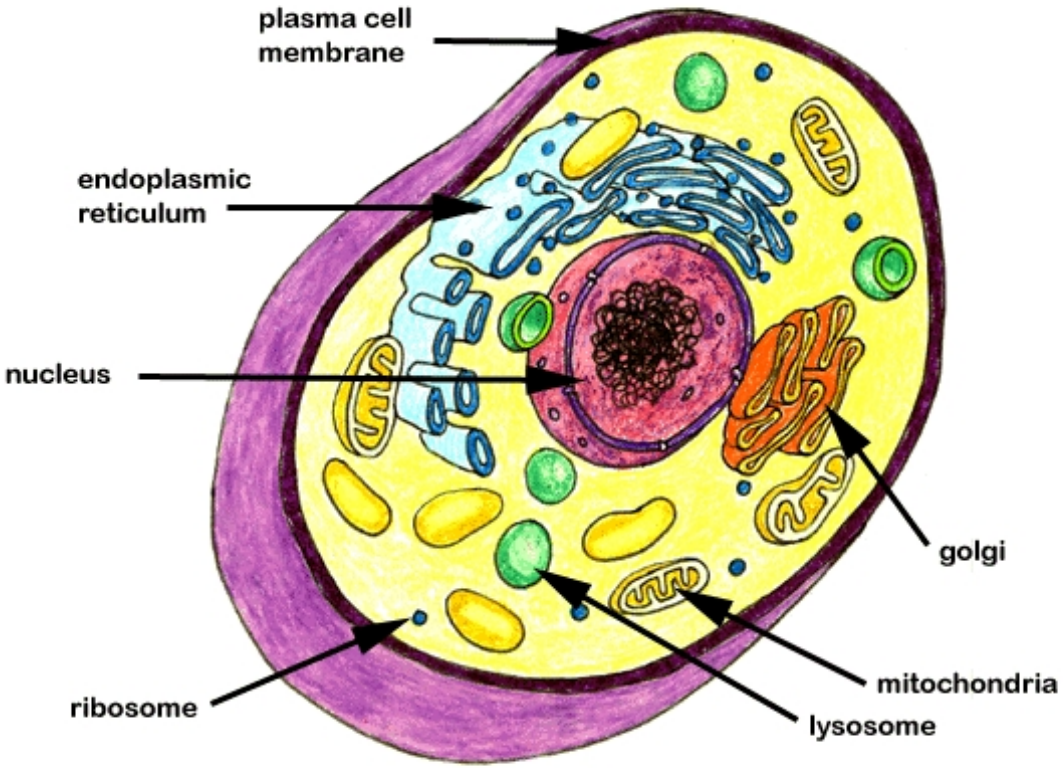
- validate hypotheses
- suggest further experiments
- predict the effect of some treatments

In the future, treatment of diseases will be based on patient-specific therapies

- simulation tools capable to predict the effect of some therapy on a specific patient will be essential

Outline of the talk

Cells: complex systems of interactive components



- Two classifications of cell:
  - prokaryotic
  - eukaryotic
- Main actors:
  - membranes
  - proteins
  - DNA/RNA
  - ions, macromolecules,. . .
- Interaction networks:
  - metabolic pathways
  - signaling pathways
  - gene regulatory networks

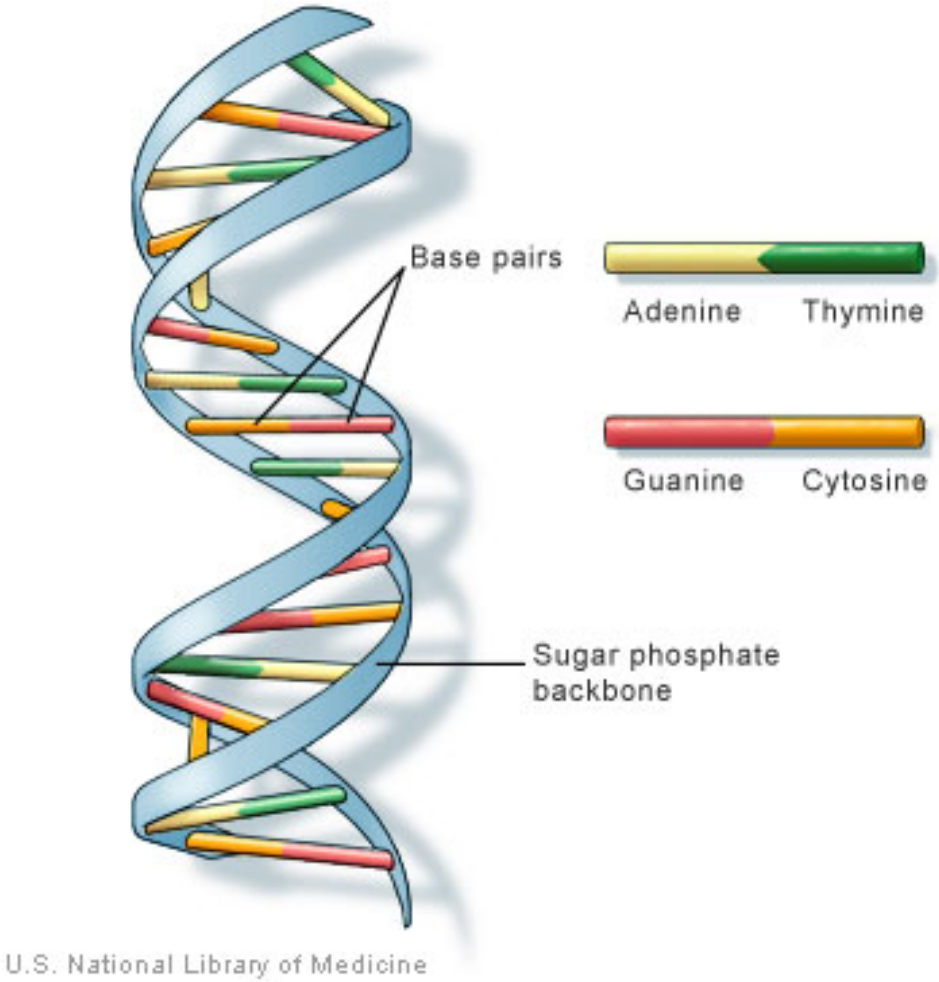
The DNA

The DNA is:

- a molecule
- structured as a string
- over an alphabet of four elements (nucleic acids, bases) denoted A,T,C,G

DNA forms double-stranded helices:

- Base pairing: A–T,C–G
- The *c*omplement of a string is obtained by replacing A with T and C with G, and viceversa
- Two complementary strings form a helic



Proteins

A gene is a substring of the DNA

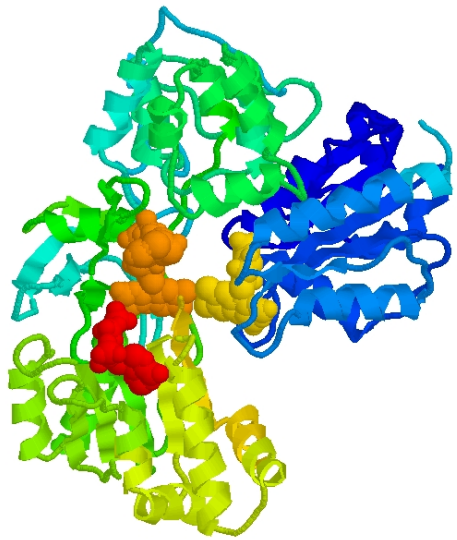
- some genes are the "source code" of proteins

A protein is:

- a molecule
- structured as a string
- over an alphabet of twenty elements (amino acids)

Proteins have complex 3D structures related with their functions:

- Catalysis of chemical reactions (enzymes)
- Transport
- Structure
- .....



The central dogma of Molecular Biology

Schematically, in cells we have this flux of information:

$$DNA \xrightarrow{transcription} RNA \xrightarrow{translation} Protein$$

Where the RNA is a molecule structured as a string over the alphabet A,U,C,G (similar to that of DNA)

- It is essentially a copy of the DNA (this motivates the terminology of transcription)

Both transcription and translation can be regulated in order to synthesize proteins only when necessary

GENE REGULATORY NETWORKS

Example of gene regulation network: the *lac* operon

E. coli is a bacterium often present in the intestine of many animals. It is one of the most completely studied of all living things.

As most bacteria, E.coli is often exposed to a constantly changing physical and chemical environment, and reacts to changes in its environment through changes in the kinds of enzymes it produces.

In order to save energy, bacteria do not synthesize degradative enzymes unless the substrates (e.g. lactose) for these enzymes are present in the environment.

This result is obtained by controlling the transcription of some genes into the corresponding enzymes.

Example of gene regulation network: the *lac* operon

Two enzymes are mainly involved in lactose degradation:

- the *lactose permease*, which is incorporated in the membrane of the bacterium and actively transports the sugar into the cell,
- and the *beta galactosidase*, which splits lactose into glucose and galactose.

The bacterium produces also the *transacetylase* enzyme, whose role in the lactose degradation is marginal.

The sequence of genes in the DNA of E. coli which produces the described enzymes, is known as the *lactose operon*.

The lactose operon consists of six genes:

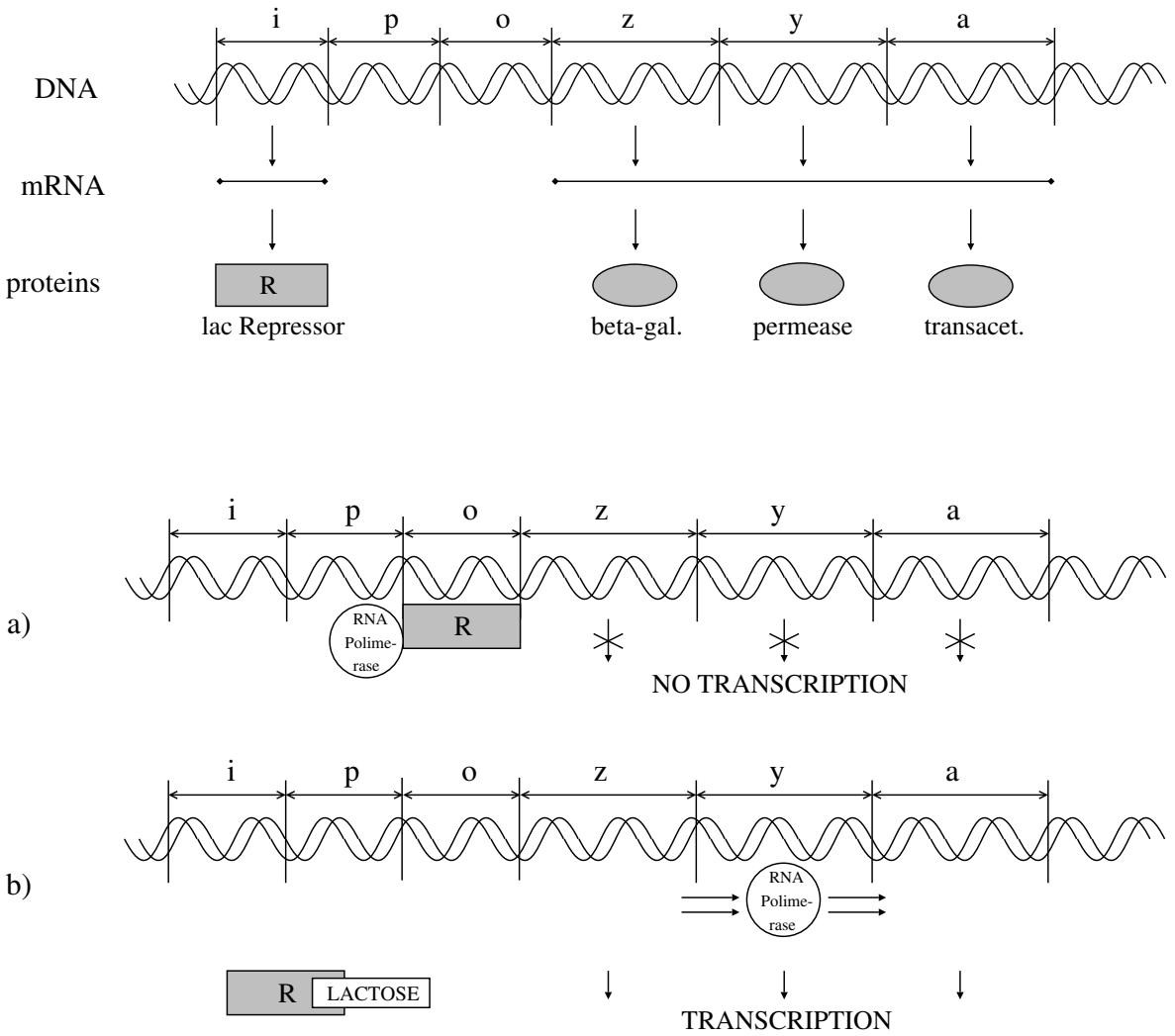
- The first three genes of the operon (i, p and o) regulate the production of the enzymes,
- the last three (z, y and a), called *structural genes*, are transcribed (when allowed) into the mRNA for beta galactosidase, lactose permease and transacetylase, respectively.

Example of gene regulation network: the *lac* operon

The regulation process is as follows:

- Gene i encodes the *lac Repressor*, which, in the absence of lactose, binds to gene o (the *operator*).
- Transcription of structural genes into mRNA is performed by the RNA polymerase enzyme, which usually binds to gene p (the *promoter*) and scans the operon from left to right by transcribing the three structural genes z, y and a into a single mRNA fragment.
- When the lac Repressor is bound to gene o, it becomes an obstacle for the RNA polymerase, and the transcription of the structural genes is not performed.
- On the other hand, when lactose is present inside the bacterium, it binds to the Repressor and this cannot stop anymore the activity of the RNA polymerase. In this case the transcription is performed and the three enzymes for lactose degradation are synthesized.

Example of gene regulation network: the *lac* operon



The cell cycle

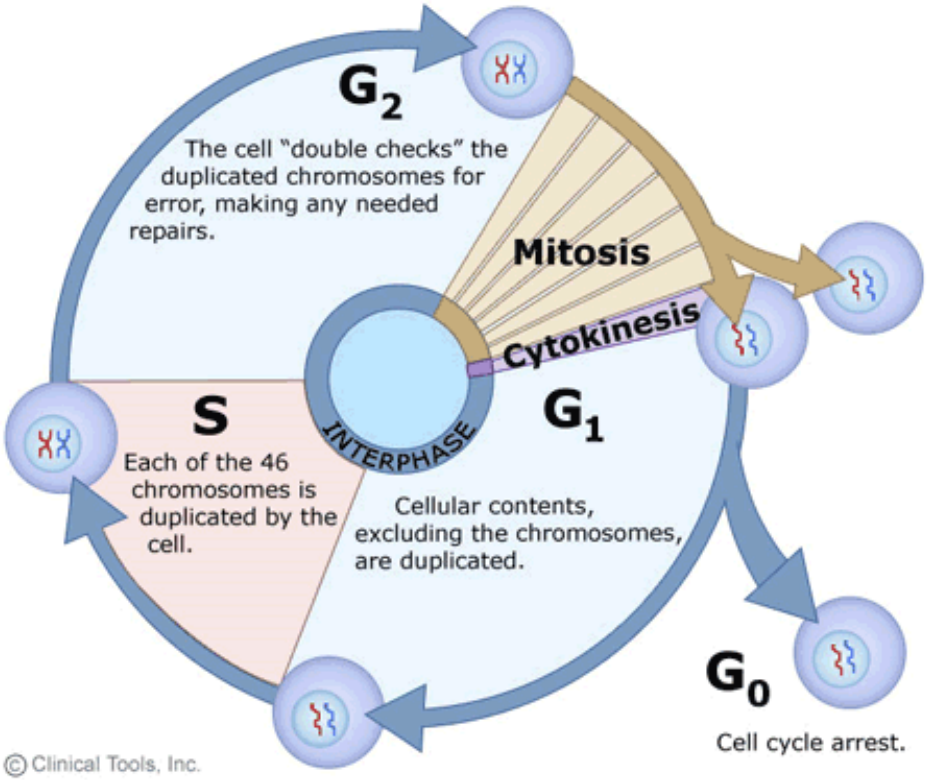
The cell cycle is a series of sequential events leading to cell replication via cell division.

It consists of four phases: G_1, S, G_2 and M.

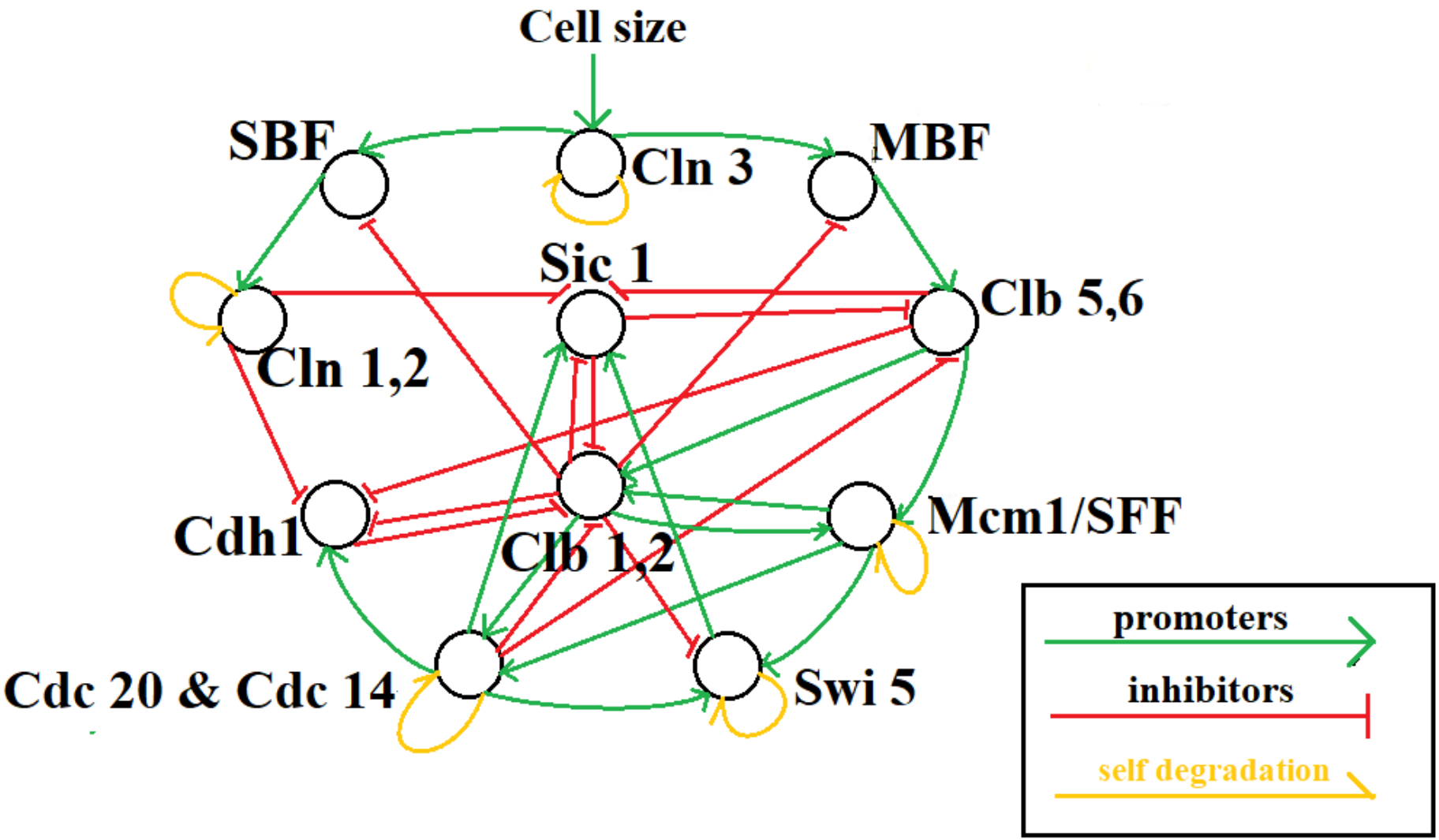
- G_1 and G_2 are gap phases in which the cell prepares itself to enter phases S and M, respectively
- S is a synthesis phase, in which DNA is replicated
- M is a mitosis phase, in which the cell segregates the duplicated sets of chromosomes between daughter cells and then divides.

The duration of the cell cycle depends on the type of cell (e.g a human normal cell takes approximately 24 hours to perform a cycle).

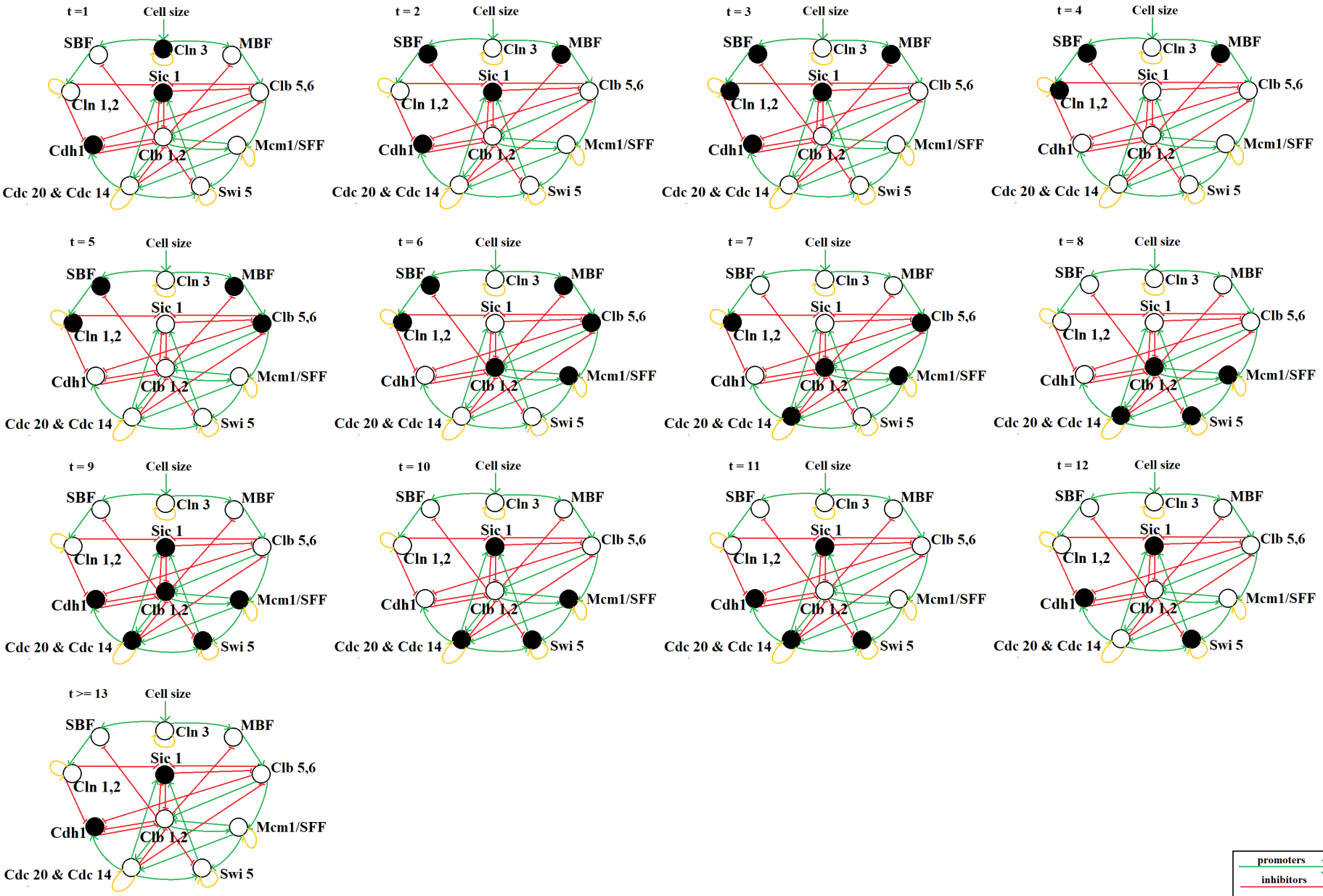
The cell cycle (model)



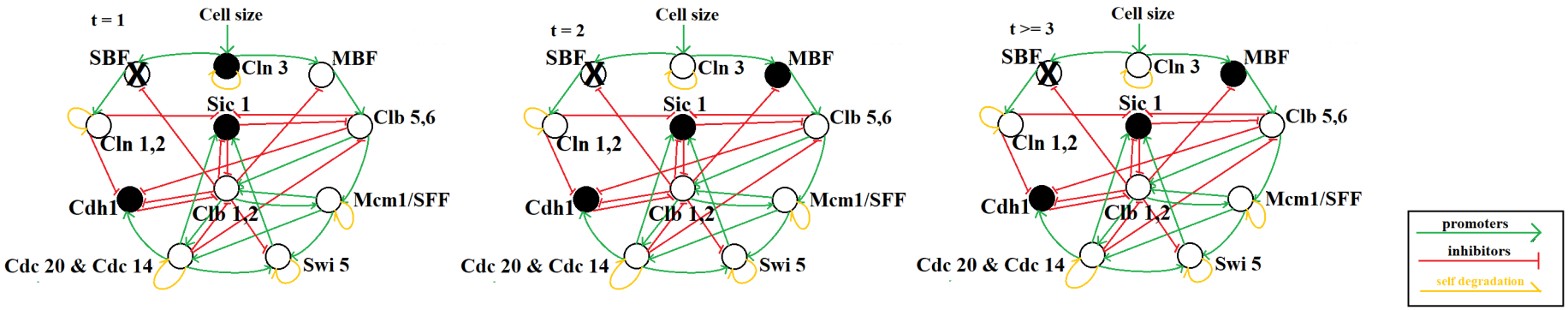
The cell cycle (dynamics)



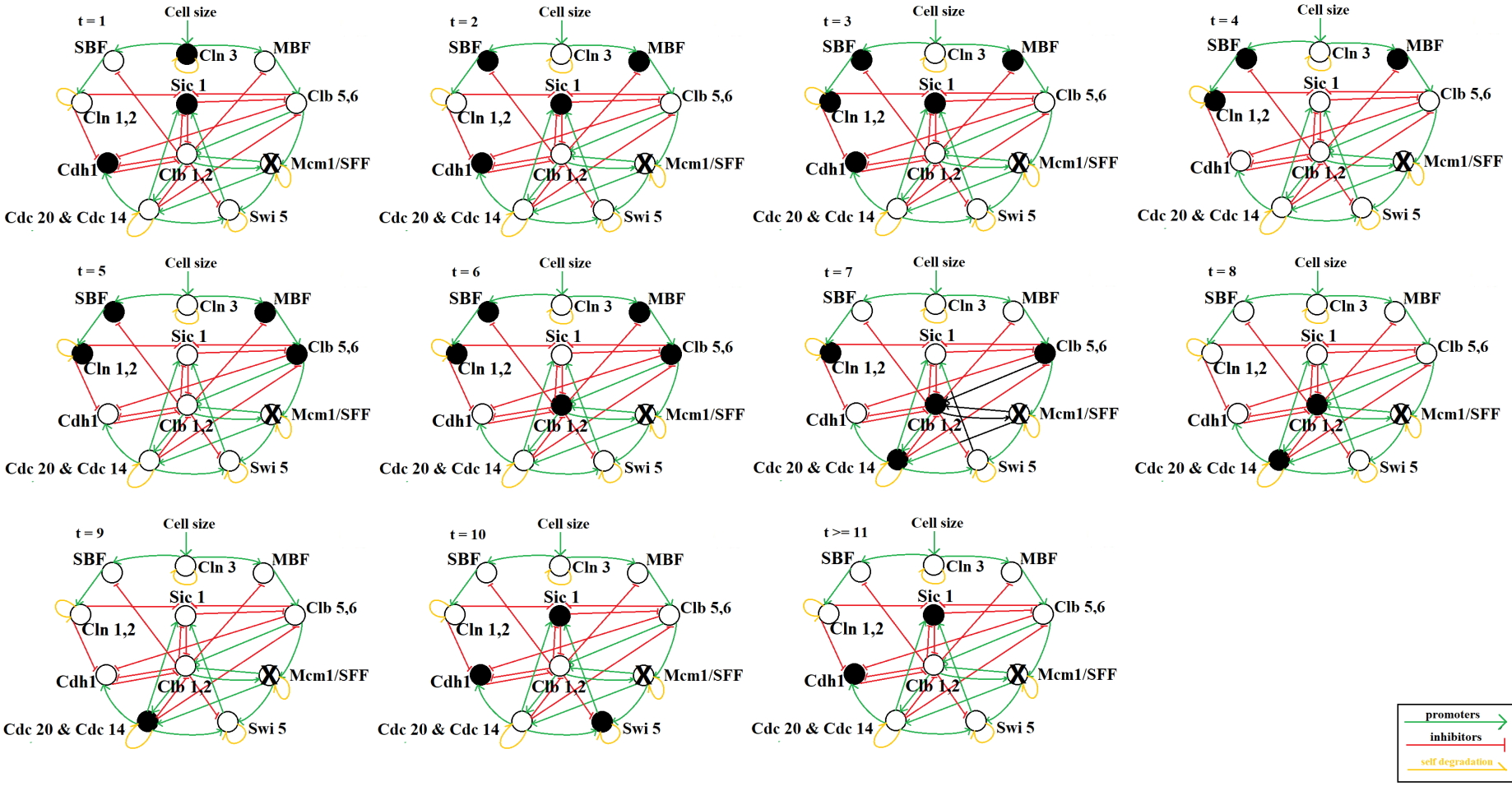
The cell cycle (dynamics - SBF K.O.)



The cell cycle (dynamics - Mcm1/SFF K.O.)



The cell cycle

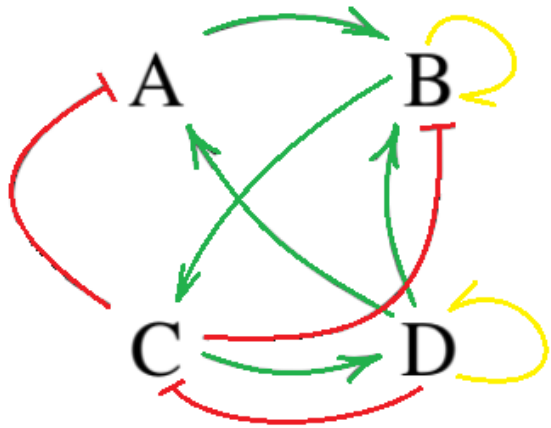


THESIS TOPIC 1

Modeling and analysis of gene regulation networks

(collaborators: R. Barbuti, R. Gori, F. Levi)

We have proposed a translation of gene regulation networks into Reaction Systems



$\Downarrow$

$(\{D\},\{C\},\{A\})$ $(\{C\},\emptyset,\{D\})$ $(\{A,D\},\emptyset,\{B\})$ $(\{A\},\{C\},\{A\})$ (
$(\{B\},\{D\},\{C\})$ $(\{A\},\{C\},\{B\})$ $(\{D\},\{C\},\{B\})$ $(\{A,D\},\{\},\{A\})$ (

THESIS TOPIC 1

Modeling and analysis of gene regulation networks

(collaborators: R. Barbuti, R. Gori, F. Levi)

The obtained Reaction Systems can be used to

- simulate the gene regulation network
- perform causality analyses
  - given an observed "final" configuration of a network, which could have been the possible "initial" configurations?
  - we have proposed formula based predictors to answer this question
  - a formula based predictor is a logic formula describing all the possibile initial configurations
- This methodology requires further developments....

THESIS TOPIC 1

Modeling and analysis of gene regulation networks

(collaborators: R. Barbuti, R. Gori, F. Levi)

References:

- R. Barbuti, P. Bove, R. Gori, F. Levi, P. Milazzo
  Simulating gene regulatory networks using reaction systems
  Proc. of the 27th International Workshop on Concurrency, Specification and Programming, CS&P 2018, pages 119-132
  `http://ceur-ws.org/Vol-2240/paper11.pdf`
- R. Barbuti, R. Gori, F. Levi, F., P. Milazzo
  Investigating dynamic causalities in reaction systems
  Theoretical Computer Science, n. 623, pages 114-145
  `https://doi.org/10.1007/978-3-319-54072-6_3`

PATHWAYS AND PROTEIN INTERACTION NETWORKS

Example of signalling pathway: the EGF pathway

A classical example of biological system is the EGF signal transduction pathway.

If EGF proteins are present in the environment of a cell, they must be interpreted as a signal from the environment meaning that new cells are needed.

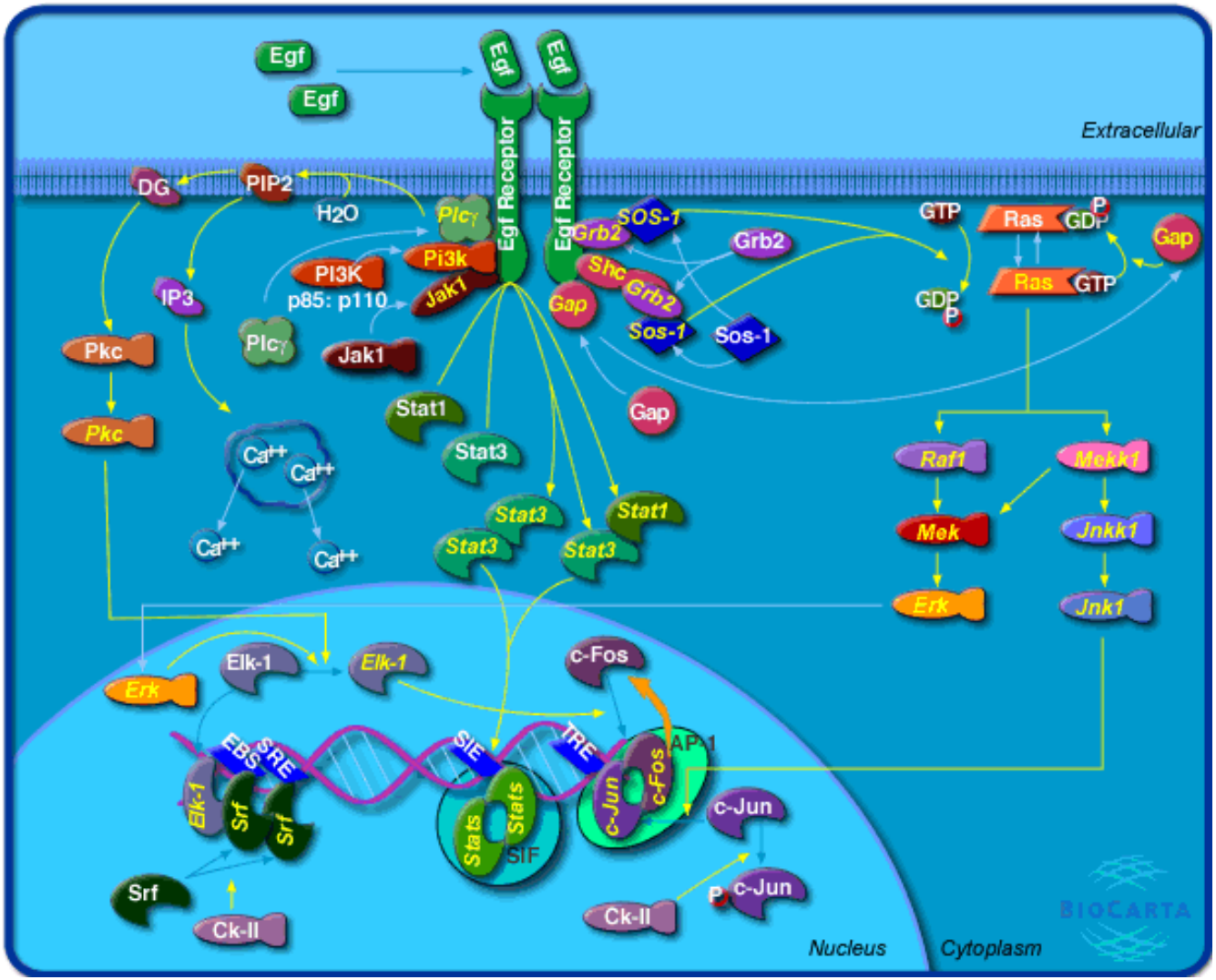
A cell recognizes the EGF signal from the environment because it has on its membrane some EGF receptor proteins (EGFR), which are transmembrane proteins (they have some intra–cellular and some extra–cellular domains).

Example of signalling pathway: the EGF pathway

The signalling pathway is as follows:

- One of the extra–cellular domains binds to one EGF protein in the environment, forming a signal–receptor complex on the membrane.
- This causes a conformational change on the receptor protein that enables it to bind to another one signal–receptor complex.
- The formation of the binding of the two signal–receptor complexes (called dimerization) causes the phosphorylation (addition of some phosphate groups PO_4) of some intra–cellular domains of the dimer.
- This causes the internal domains of the dimer to be recognized by proteins that are in the cytoplasm, which bind to the dimer, enabling a chain of protein–protein interactions inside the cell.
- This chain of interactions finally activate some proteins which bind to the DNA and stimulate synthesis of proteins for cell proliferation.

Example of signalling pathway: the EGF pathway



Modeling pathways

A pathway is actually a set of (bio)chemical reactions:

$$Egf + R \xrightarrow{k_1} ER$$
$$2ER \xrightarrow{k_2} ERdim$$
$$ERdim \xrightarrow{k_3} ERdimP$$
$$ERdimP + Grb2 \xrightarrow{k_4} EGrb2$$
$$EGrb2 + SOS \xrightarrow{k_4} EGrb2SOS$$
$$EGrb2SOS + RasGDP \xrightarrow{k_5} EGrb2SOS + RasGTP$$
$$\vdots$$

- Simulation techniques of chemical reactions can be used to study the dynamics of cell pathways.
- However, these sets of chemical reactions can be huge and simulation can take long times

THESIS TOPIC 2

Machine learning methods to predict dynamical properties of cell pathways

(collaborator: A. Micheli)

Very roughly speaking:

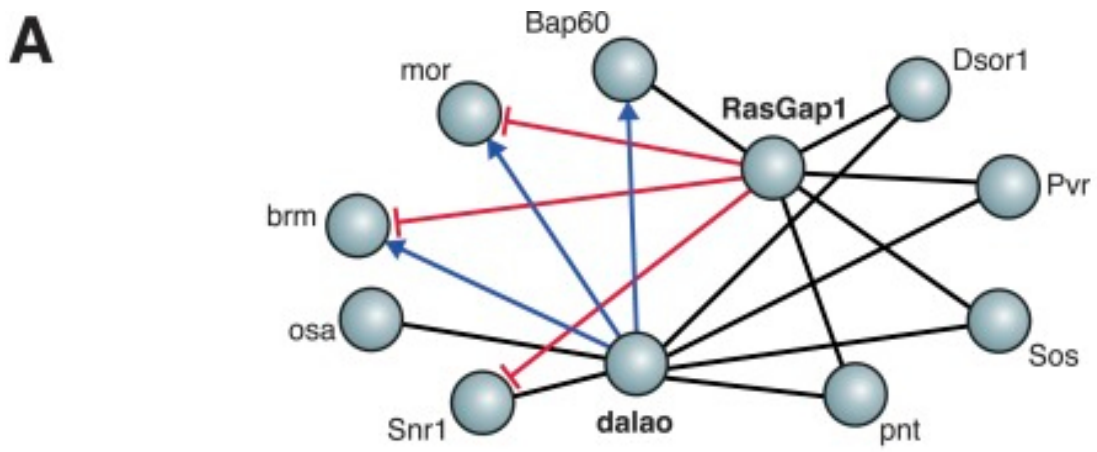
- We are constructing a dataset of simulation results of cell pathways
- We consider the Petri Net representation (i.e. a graph) of the chemical reactions constituting each pathway
- For each pathway we perform a number of simulations
- Machine learning methods for graphs could be applied to learn (specific properties) of the simulation results
- The obtained model could be used to predict dynamical properties of new pathways without running simulations

Protein interaction networks

The problem of the size and complexity of pathway models is often solved by considering a more abstract representation of pathways as protein interaction networks

A protein interaction network is a graph in which nodes are proteins and edges represent the existence of a reaction in some pathway in which both the two connected proteins are (somehow) involved

Edges of different types may represent different types of interactions between proteins



THESIS TOPIC 3

Analysis of protein interaction networks for biomedical applications

(collaborator: C. Priami)

Several analysis techniques exist for protein interaction networks aimed at investigating relationships between proteins

In the biomedical context it is often interesting to understand:

- which proteins could be influenced by a disfunction of some other proteins (a disease)?
- which proteins should be addressed by a new drug to be developed (target identification)?
- which proteins could be influenced by a new drug (toxicity prediction)?
- which existing drug could be used to treat a new disease (drug repurposing)?

We would like to develop new methodologies based on protein interaction networks and investigate new application cases, in particular in the context of drug repurposing