# A glimpse into complexity

**Gianfranco Politano**
**Politecnico di Torino, Italy**
Department of Control and Computer Engineering
http://www.testgroup.polito.it/ - **Systems Biology Group**

gianfranco.politano@polito.it

# Goal of this lecture

To comment, from an engineer point of view, the research methodology used in Life Sciences.

# What do engineers do?

- Engineers are problem solvers.
- They study high level math and science and use those (sometimes with their creativity and imagination), to isolate problems, analyze them, address them and come up with practical ways to change things so they perform better.

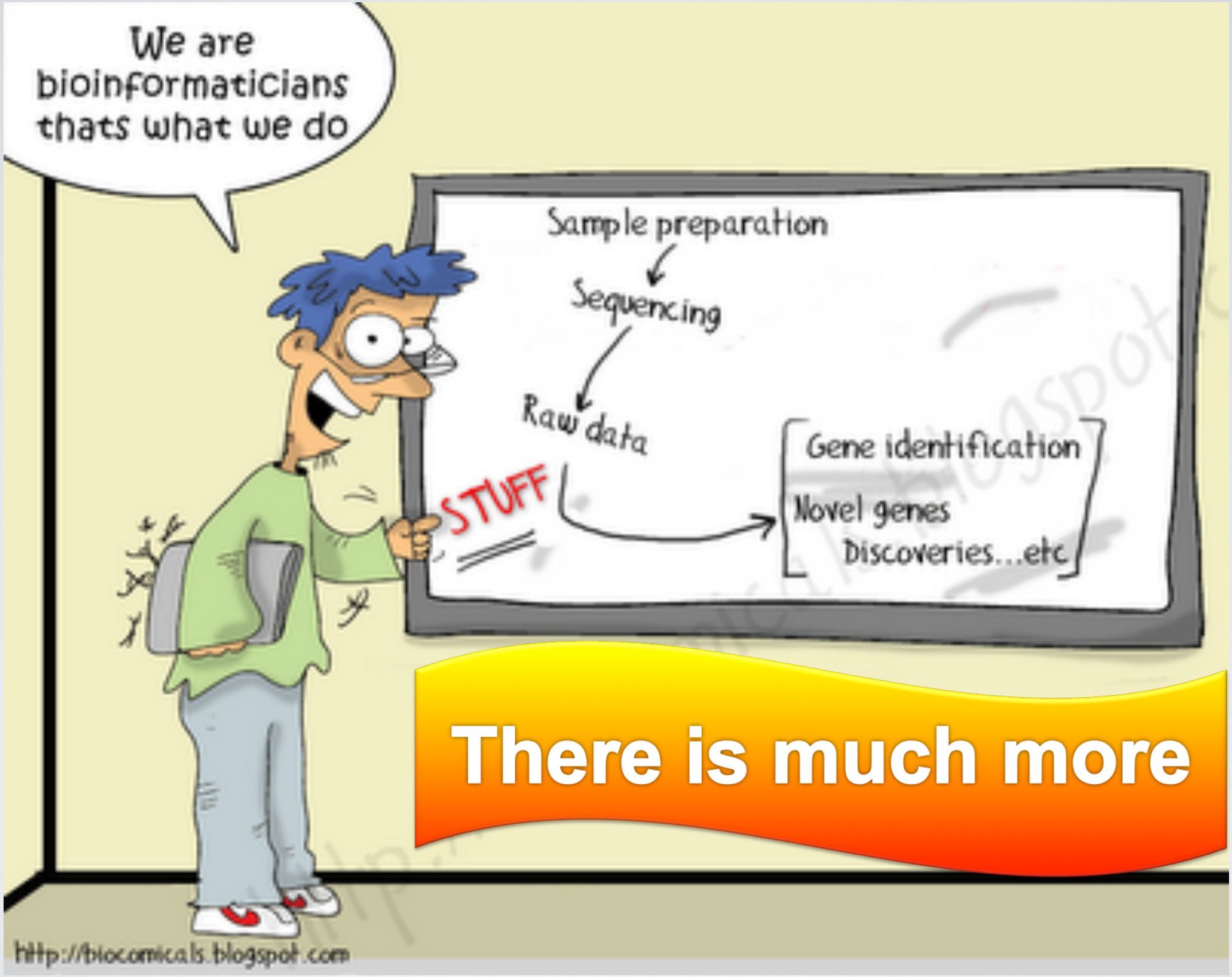
Engineers are trained to be '**big picture**' thinkers.

# First observation

- The research methodology used in Life Sciences is in many aspects different from an "engineered oriented" one.

- Biologists feel comfortable with **UNCERTAINTY**.

- Engineers always strive for **PROOFS**


We are bioinformaticians thats what we do
Sample preparation
Sequencing
Raw data
STUFF
Gene identification
Novel genes
Discoveries...etc
There is much more
http://biocomicals.blogspot.com

# Biology for an engineer

- Biology is **Reverse Engineering**

- "Reverse engineering is the process of **discovering the functional principles of a device**, object, or system **through analysis of its structure, function, and operation**".

# Biology for an engineer

- The most important concept to understand about reverse engineering is that:

designing a system and reverse engineer it are two opposite tasks whose complexity may differ in orders of magnitude.

# Example: sampling problem

- ***Goal***: to reconstruct the dynamics of a signal
- ***Problem***: which sampling frequency do I have to use?
- ***Theorem***: in signal processing a theorem says that the sampling frequency has to be at least the DOUBLE of the signal frequency.

# Translation

- ***Goal***: to figure out the plot of a movie by seeing the fewest possible screenshots of it.

- ***Problem***: How often should we grab a screenshot (the sampling rate) to be able to reliably reconstruct the plot of the movie?

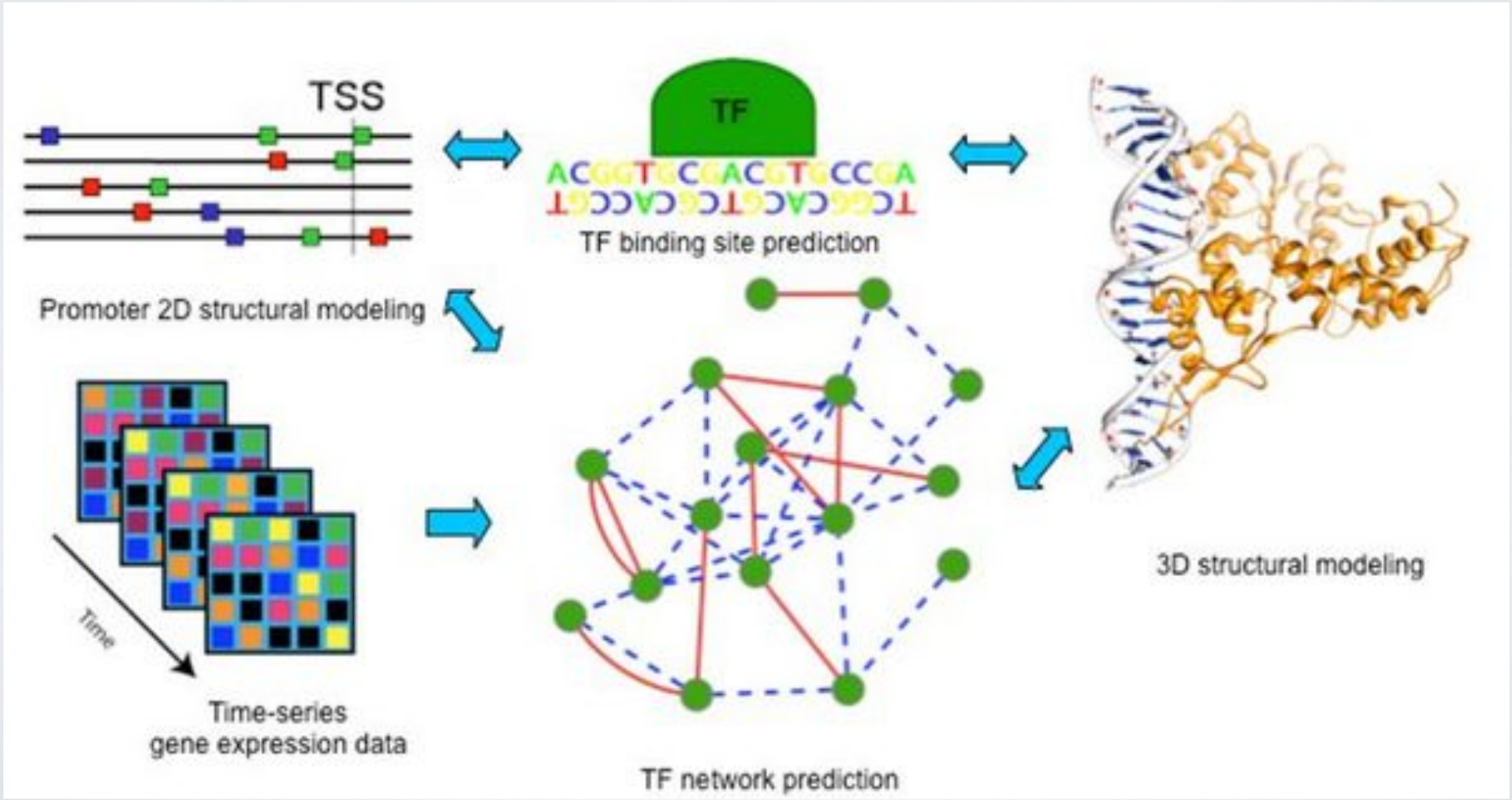
- It depends on the movie dynamics: **how often (seconds/minutes/hours) something important happens?**



- assume that two consecutive significant events never take place less than 10 minutes apart.....
- ....then the sampling theorem law tells us that we must get a screenshot at least every five minutes to be able to infer anything reliable about the plot.

# Problem for biologists

- Time-series are used to reconstruct regulatory networks from gene expression data



# Problem for biologists

- If we can take a screenshot of gene expression every 5 minutes, what I get can be used to reconstruct the network dynamics only if their frequency is in the order of 10 minutes....
- Otherwise I cannot say anything reliable about the original signal.
- So how often, and how long does it takes for a gene to change its expression? What is the gene expression "frequency"?

# Summary

## Biology is Reverse Engineering

The system's complexity should drive the reverse engineering methodology

Let's take this concept a bit further

# Reverse Engineering

Two main approaches:

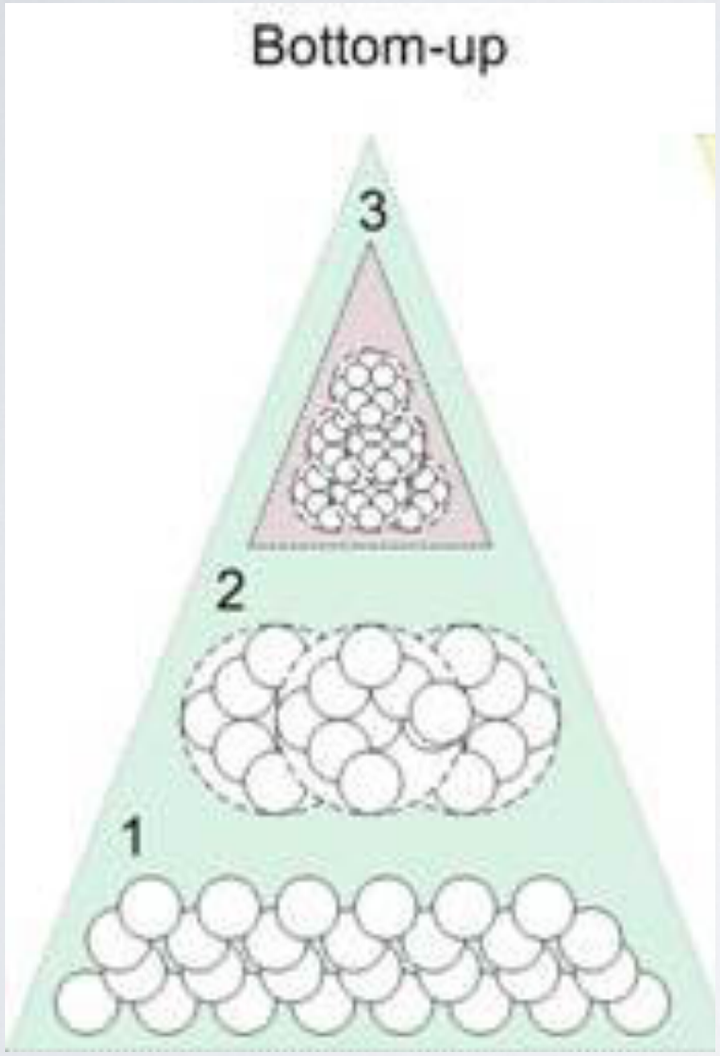
- the **bottom-up** approach, mostly followed by life sciences researchers
- the **top-down** approach, typical of the engineering world


Bottom-up approach

The final goal is, for both, a reliable **model** of the system under investigation.

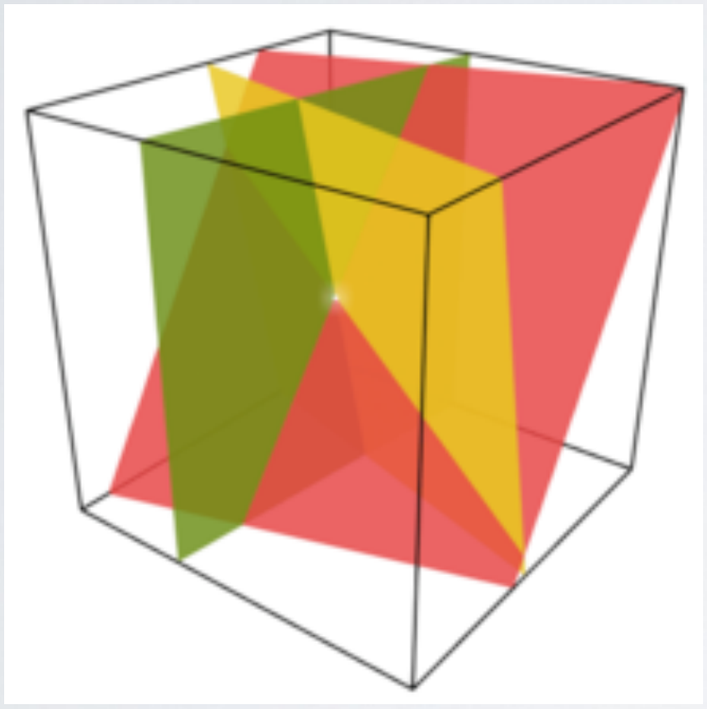
Modeling is a way to encapsulate part of the real world in terms of mathematical relationships.

# Bottom-up



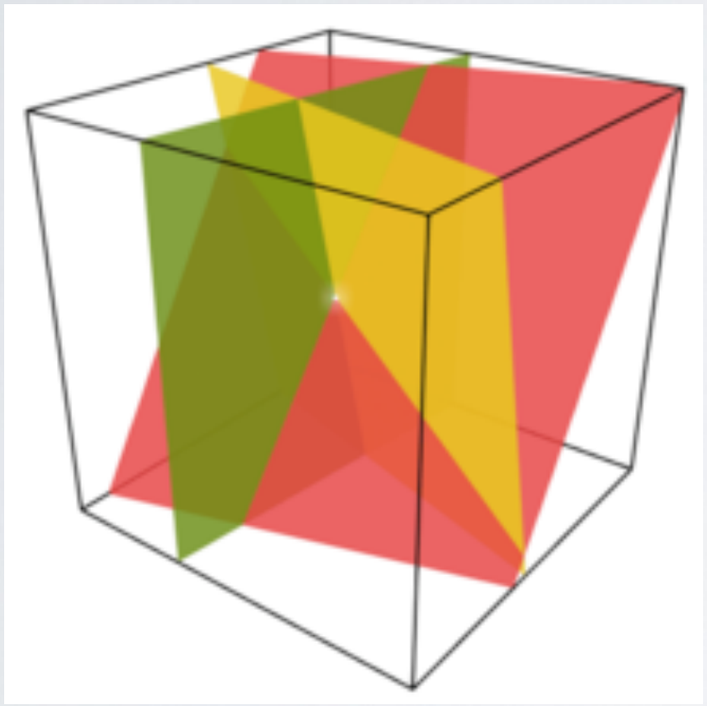

- The **variables believed as important** are selected to determine the state of a system (e.g. a set of genes)
- Their relationships is inferred **from a small set of observations** (e.g. gene expression experiments) (= bottom)
- Local observations are "**merged**" to create a higher-level model (=up)
- This approach works very well with **linear systems.**

# Linear Systems



- We can consider the effect of each "variable" separately, because the sum of their effects equals the effect of their sum.

# Linear Systems



- Linear systems are easy to understand also for non-mathematicians and are also easy to visualize.
- For this reason a large part of the life science world (and the medical one in particular) still reasons in linear terms.
- But **only a few biological systems are truly linear**.

# Complex Systems

- Complex (or non-linear) systems are much more difficult to understand or visualize.
- Complex systems consist of many **diverse and autonomous but interrelated and interdependent components** or parts linked through many (usually dense) interconnections.

# Complex Systems

- Complex systems cannot be described by a single rule and they exhibit **properties that emerge from the interaction of their parts** and which cannot be predicted only from the properties of the parts.

- Complex Systems' dynamics heavily depend on **initial conditions and perturbations** (the butterfly effect…..)

# Complex Systems - examples

- In a soccer team, the players are the parts of the system; the rules of the game are the interconnections; the purpose of the system is both to play and to win the game.

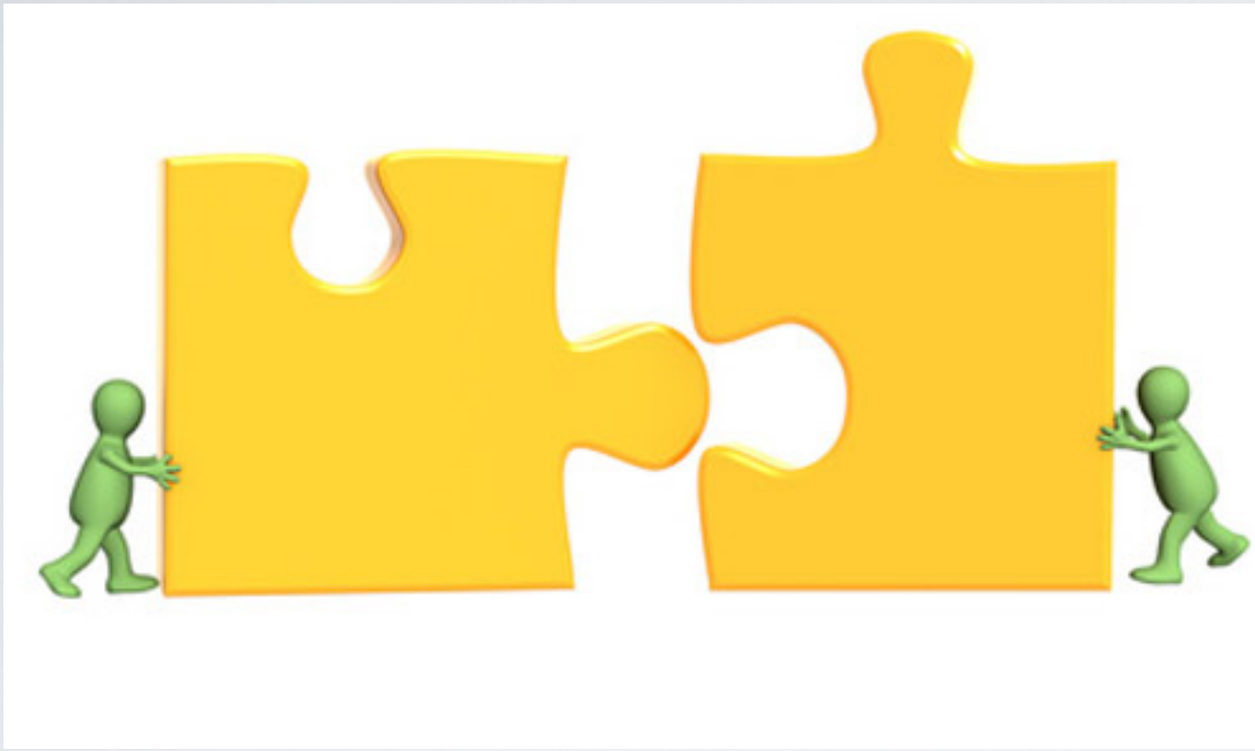
- What emerges is the game itself. Imagine if you isolated one soccer player; he might be able to practice shooting or dribbling, but a game of soccer would never be evident.

- The concept of "game" only emerges from the interactions of the parts of the system

# Complex Systems - examples

- A university
- A beehive
- A cell
- The human body as a collection of cells
- .....

# Problems in Reverse Engineering

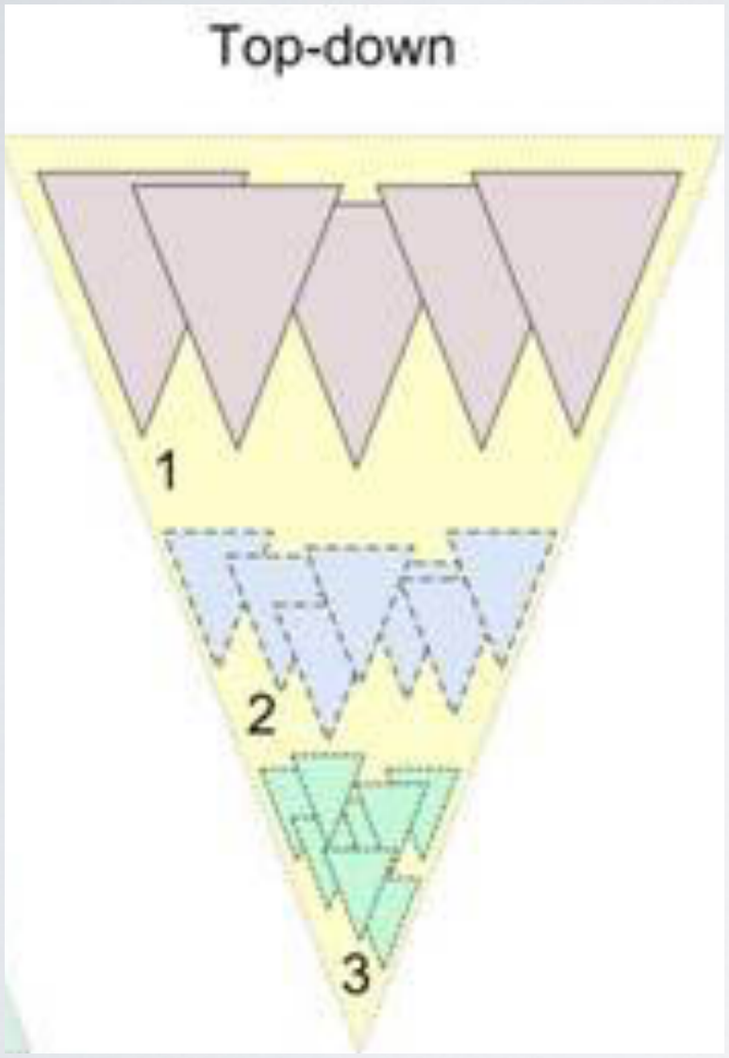


To merge multiple observation for obtaining a larger model (bottom-UP) you need to make all things work together, but if the system is complex …

- ✓ even if a bottom-up approach will allow to get **early insights** on the system behavior …
- ✓ it is likely that some **key *trait d'union*** are **missed** because of the **specificity of the observations**, or, worst, because the **complexity itself hides the basic mechanics**.

So, for Complex Systems, the way to go is TOP-DOWN…

# Top-down




- The **variables** believed as important are **unknown** …
- Their **relationships** is **unknown** …
- A **high-level hypothesis** is formulated, specifying but not detailing any first-level subsystems.
- **Each subsystem is then refined** in yet greater detail, sometimes in many additional subsystem levels

# Problems

- Top-down works on **abstractions and inferences**, so the reached conclusions often are general enough to try to explain the overall mechanics but **their basis often lie on computational assumptions** (i.e., target of miRNAs, or protein interactions) **which may be not right at all**.

# Networks

- Networks allow to easily model the interconnections between the different components of a complex system.
- The study of networks allows to understand properties otherwise invisible

# Biological Networks

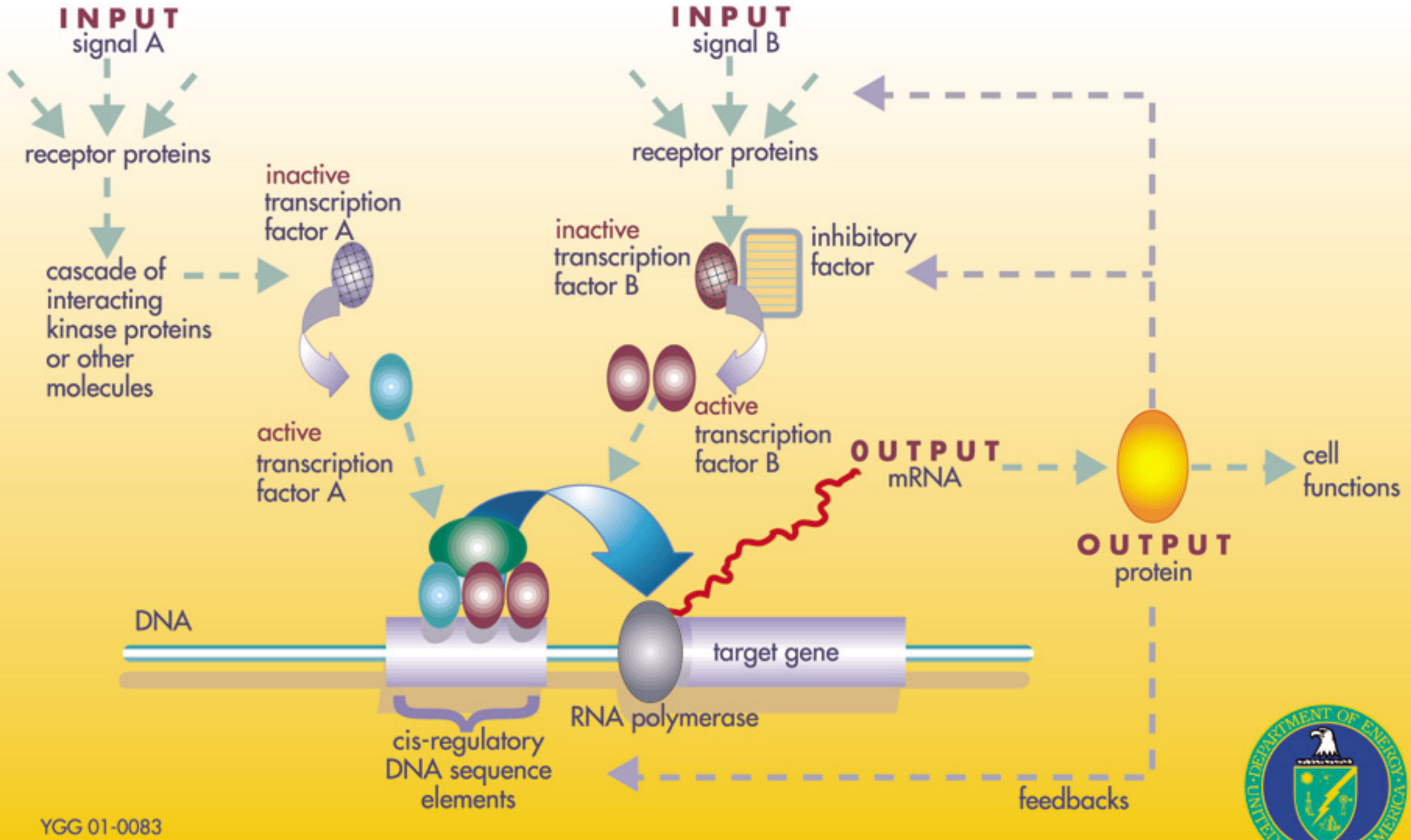
- Metabolic network: dynamic networks of "known structure". Flux Balance Analysis

- Protein interaction network: static networks. Topological/ structural measures

- Gene regulatory network (GRN): dynamic networks of "unknown structure". Simulations, equilibrium states, …

GENOMES to LIFE

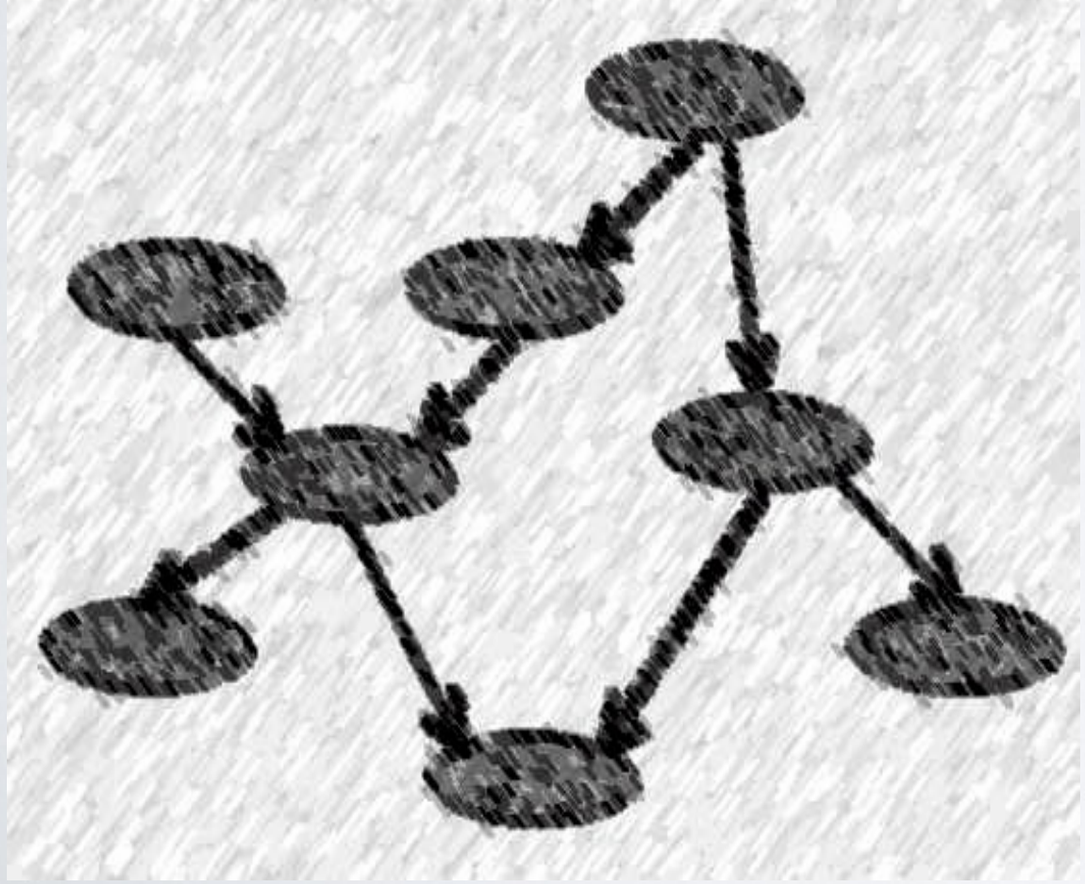
# A GENE REGULATORY NETWORK




Sources: http://www.ornl.gov/sci/techresources/Human_Genome/graphics/slides/images/REGNET.jpg

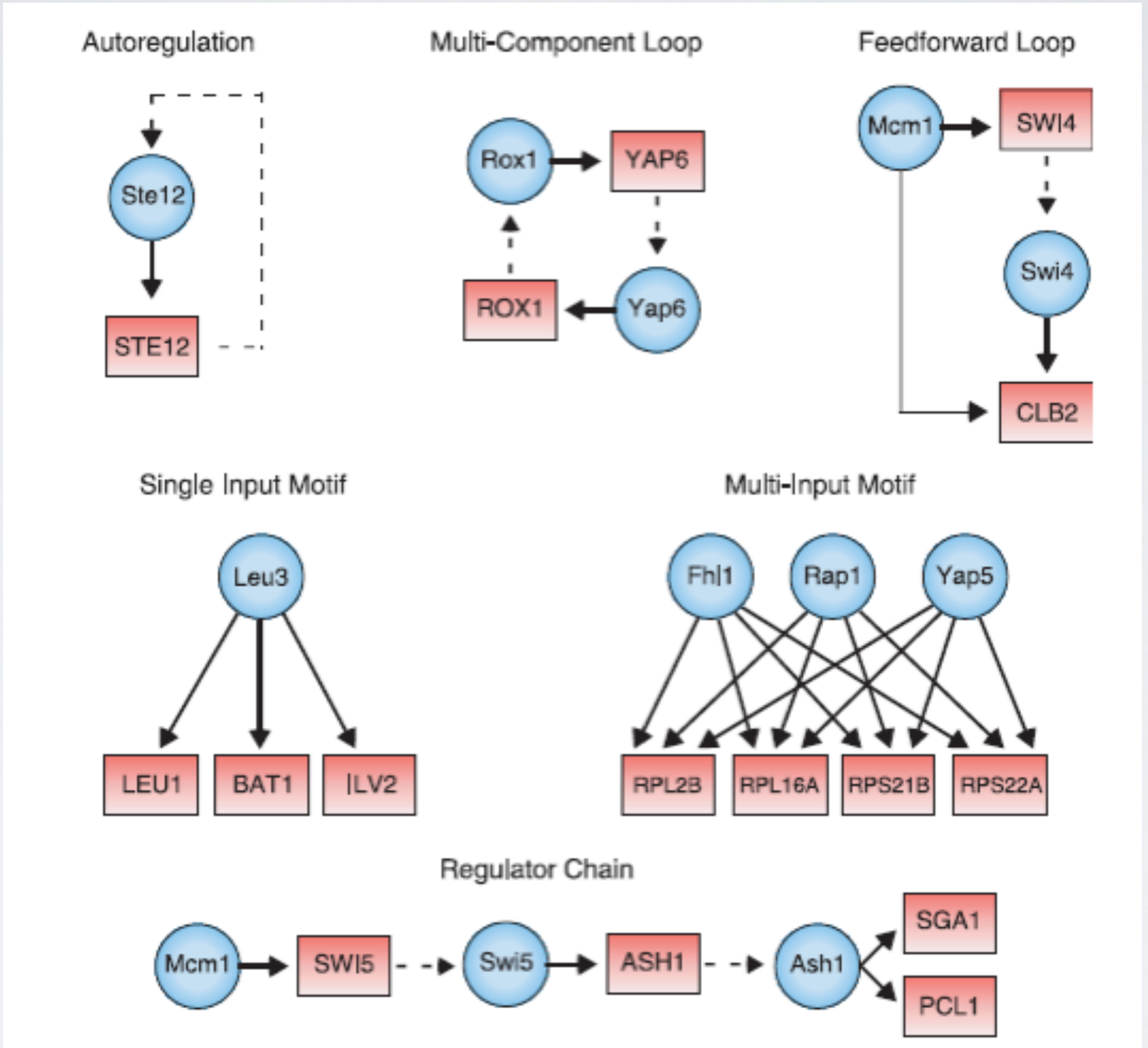
# Simplified Representation of GRN

- A gene regulatory network can be represented by a directed graph;



- Node represents a gene;
- Directed edge stands for the modulation (regulation) of one node by another:
  - e.g. arrow from gene X to gene Y means gene X affects expression of gene Y

# Motifs



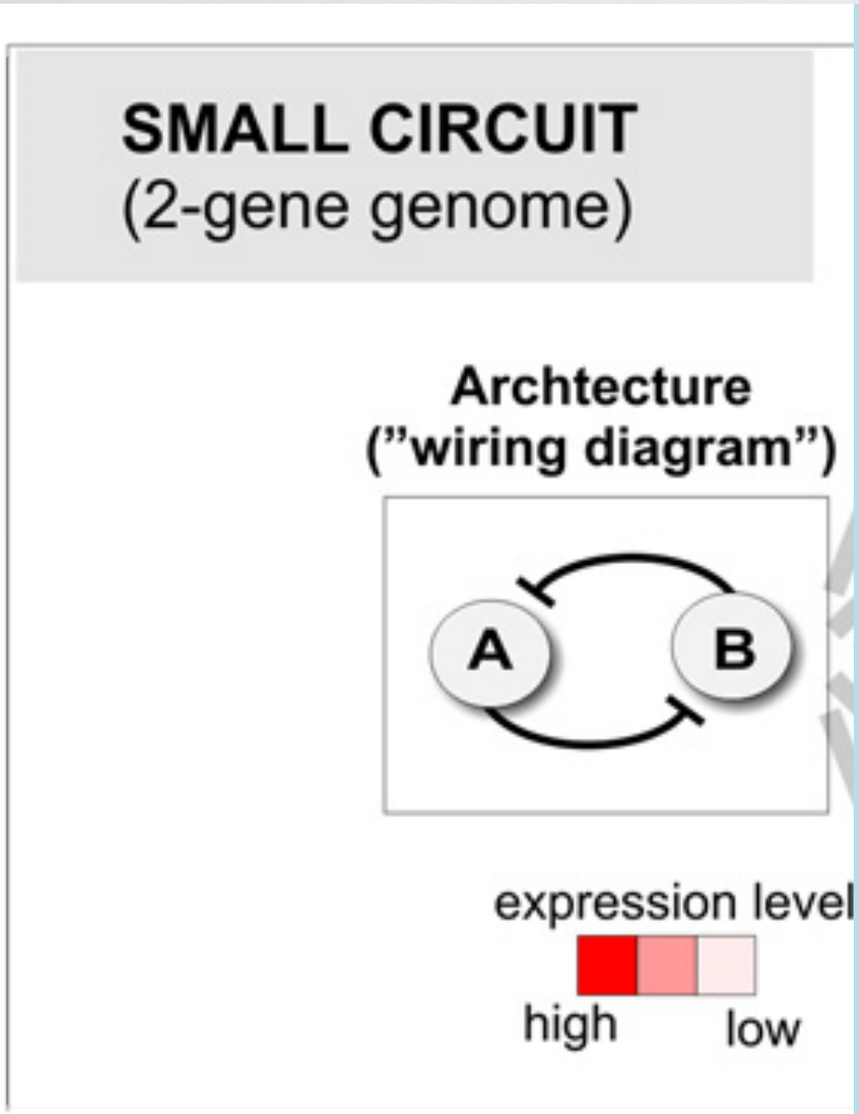

- Network motifs are the simplest units of network architecture.
- They be used to assemble a transcriptional regulatory network.

# Why Study GRN?

- Genes are not independent;
  - They regulate each other and act collectively;
  - This collective behavior can be observed using microarray;
- Some genes control the response of the cell to changes in the environment by regulating other genes;
- Potential discovery of triggering mechanism and treatments for disease;

# Law of attractors

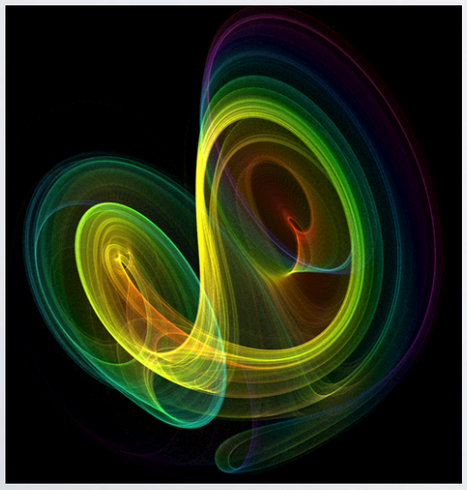
Purely theoretical approach

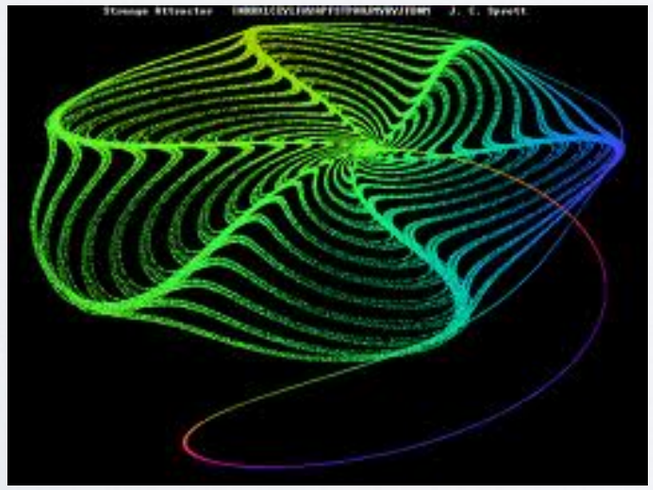


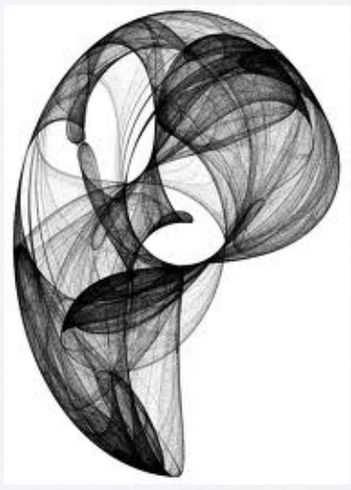

Huang et al., Semin Cell Dev Biol. 2009 Sep;20(7):869-76.

# Attractors

- Stable states are called **ATTRACTORS** and can be:
  - ✓ **point attractors**: one stable state
  - ✓ **periodic attractors**: the systems remains within a limited set of states (e.g. cellular functions, cellular cycle, …)
  - ✓ **strange attractors**: the set of stable states is not well defined……but it exists







- Now let's imaging the system under study is not made of 2 genes, but of 30.000 ...

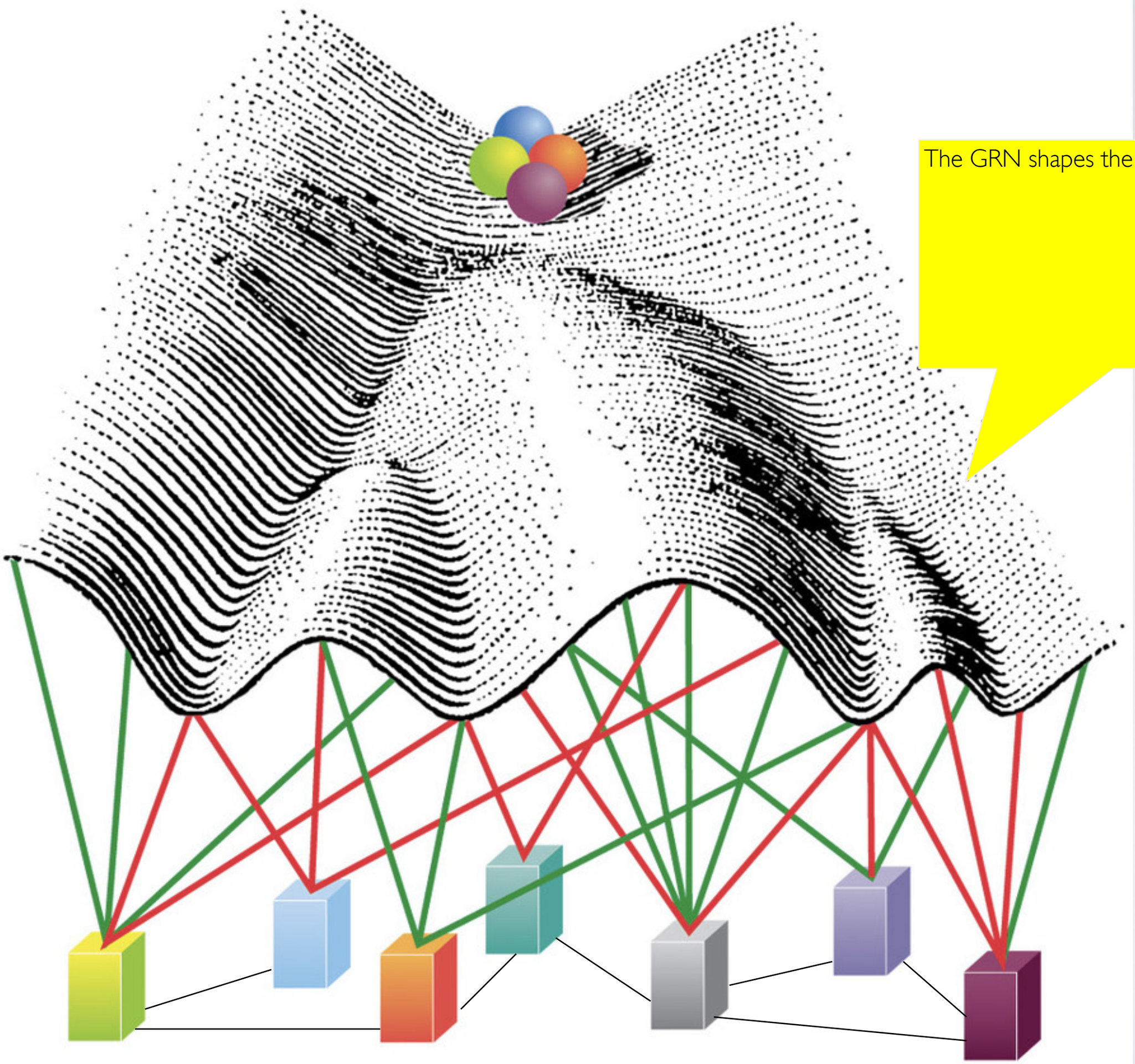
- Not easy to visualize a state space in 30.000 dimensions....

# Epigenetic landscape

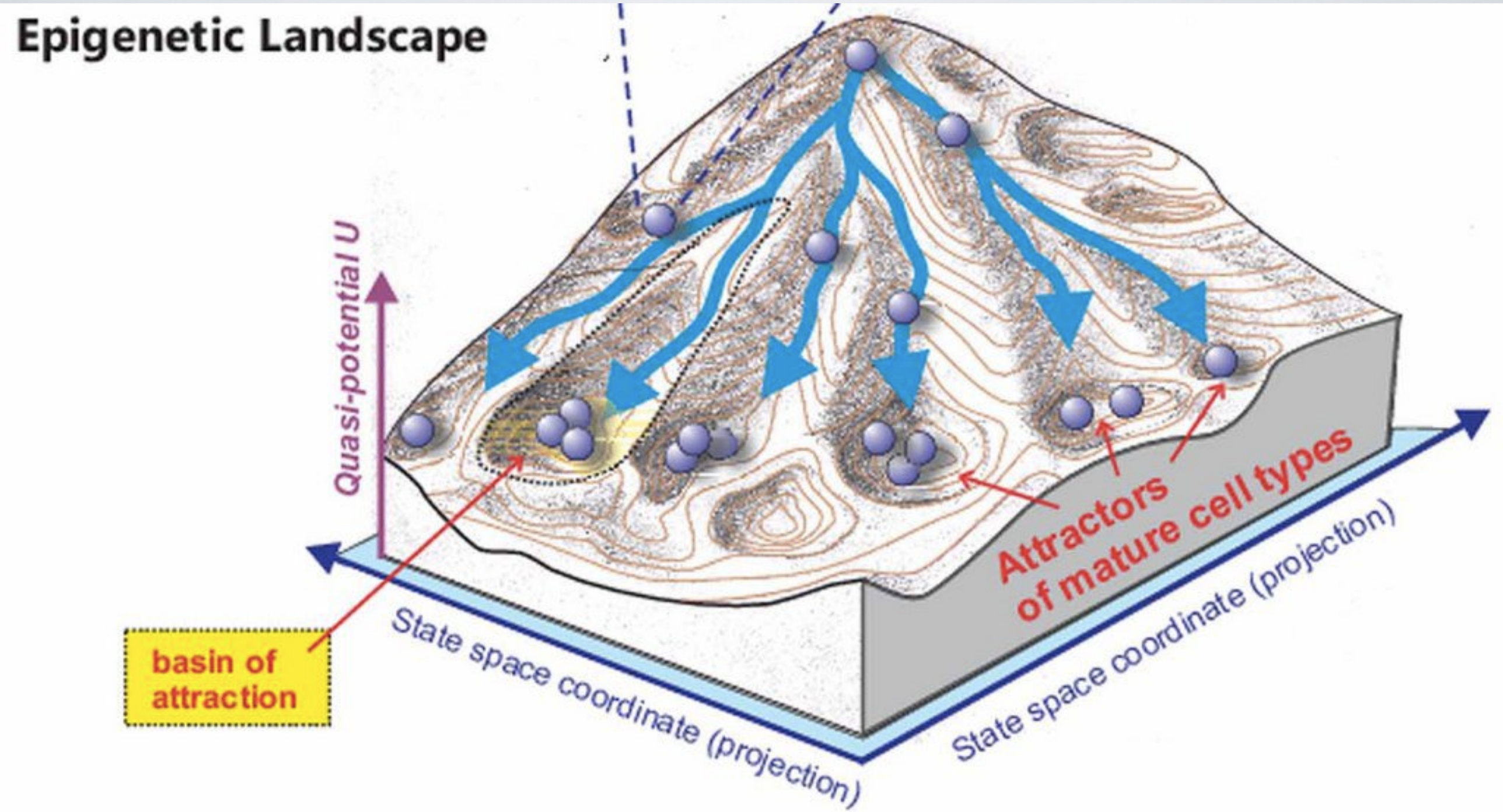
A qualitative "conceptual" representation of the possible states has been named **epigenetic landscape** by Conrad Hal Waddington in 1932.

In this context "EPIGENETIC" has a more holistic meaning than the "traditional" one, linked only with chemical modifications of the DNA (methilation, ...).

The GRN shapes the epigenetic landscape



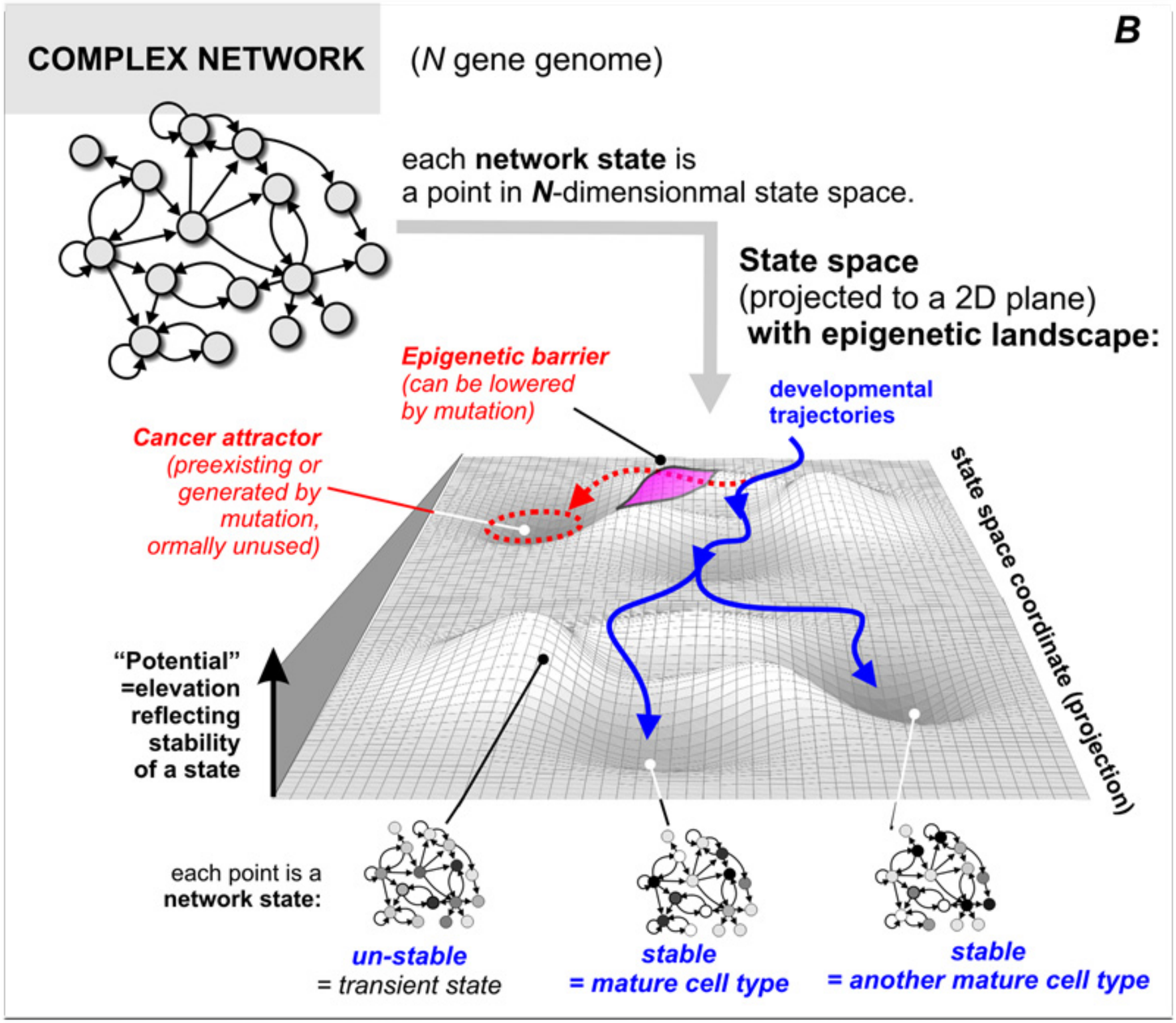
# Differentiation



# Imagine ...

- The number of variables is so huge that we can easily picture parts of the landscape that look (to us) almost identical, but maybe differ in small details.

# Cancer attractors




Some attractors (byproducts of the complex dynamics of the GRN), most likely represent abnormal, possible but usually unreachable gene expression patterns.

Huang et al., Semin Cell Dev Biol. 2009 Sep;20(7):869-76.

# Challenges

- Can we "reconstruct" the epigenetic landscape? (which correspond to Reverse Engineering the network)

- Are Microarrays the "picture" of (a part of) the attractors?

- Network dynamics simulation (up to 100 nodes....):

  ✓ with Boolean Networks we have $2^N$ states (where N is the number of nodes)

  ✓ ... but genes are not boolean ...

  ✓ divide and conquer approaches?

# Challenges

- Does it make sense to study networks WITHOUT miRNAs, ceRNAs, etc... ?

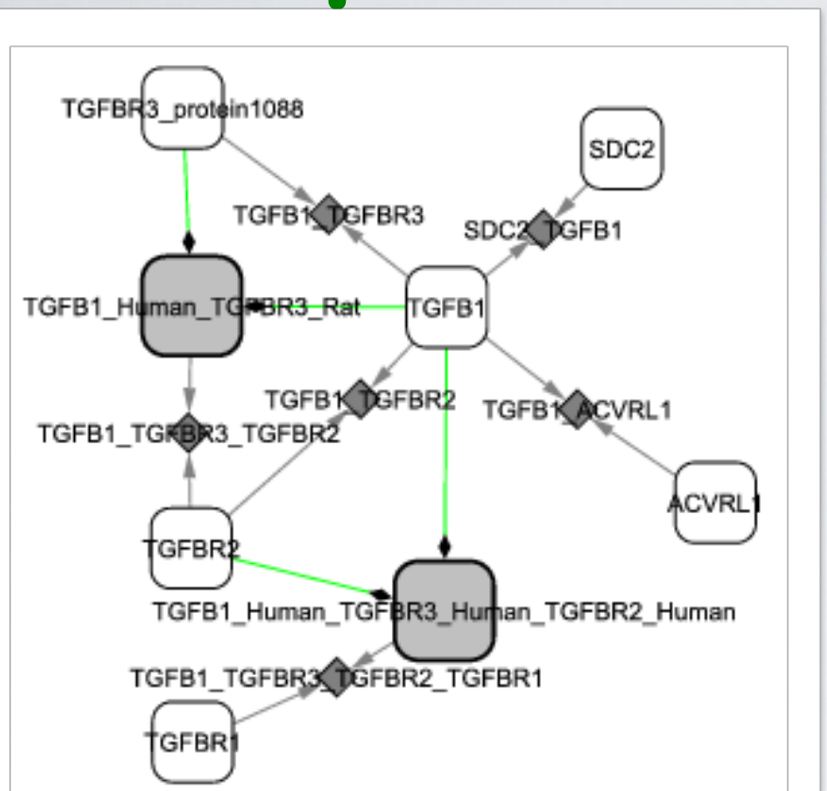
# An example: the Pathway Protection Loops

- ***Hypothesis***: some **miRNA** play a **PROTECTIVE role** of the **pathway** that express them

1

Pathway (intragenic) miRNA

2




Biological Pathway


Pathway Antagonist Transcription Factor (PATF)

3

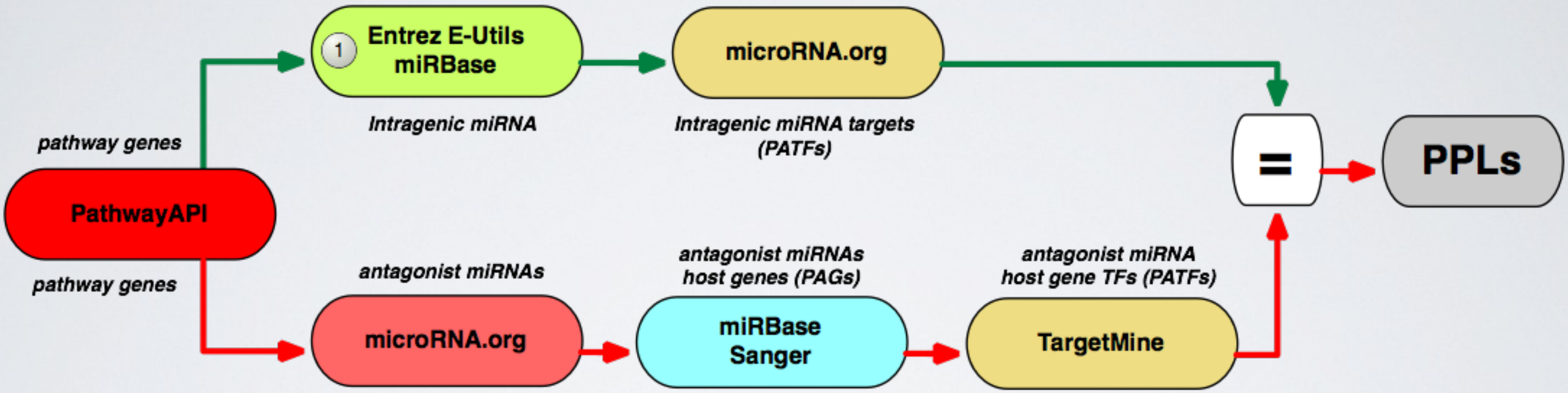
Pathway Antagonist Gene (PAG)

4

Antagonist miRNA

5

# The pipeline



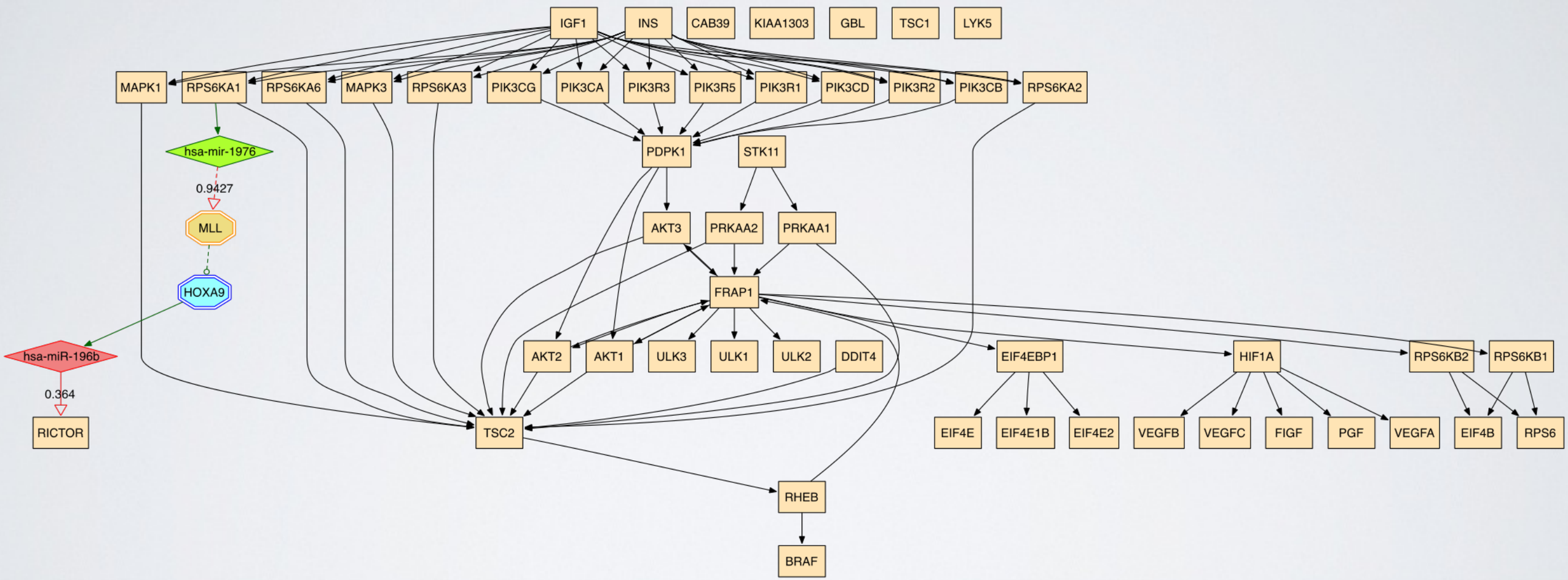
# Results

- 3 sets of pathways analyzed: metabolic, non-metabolic, and random

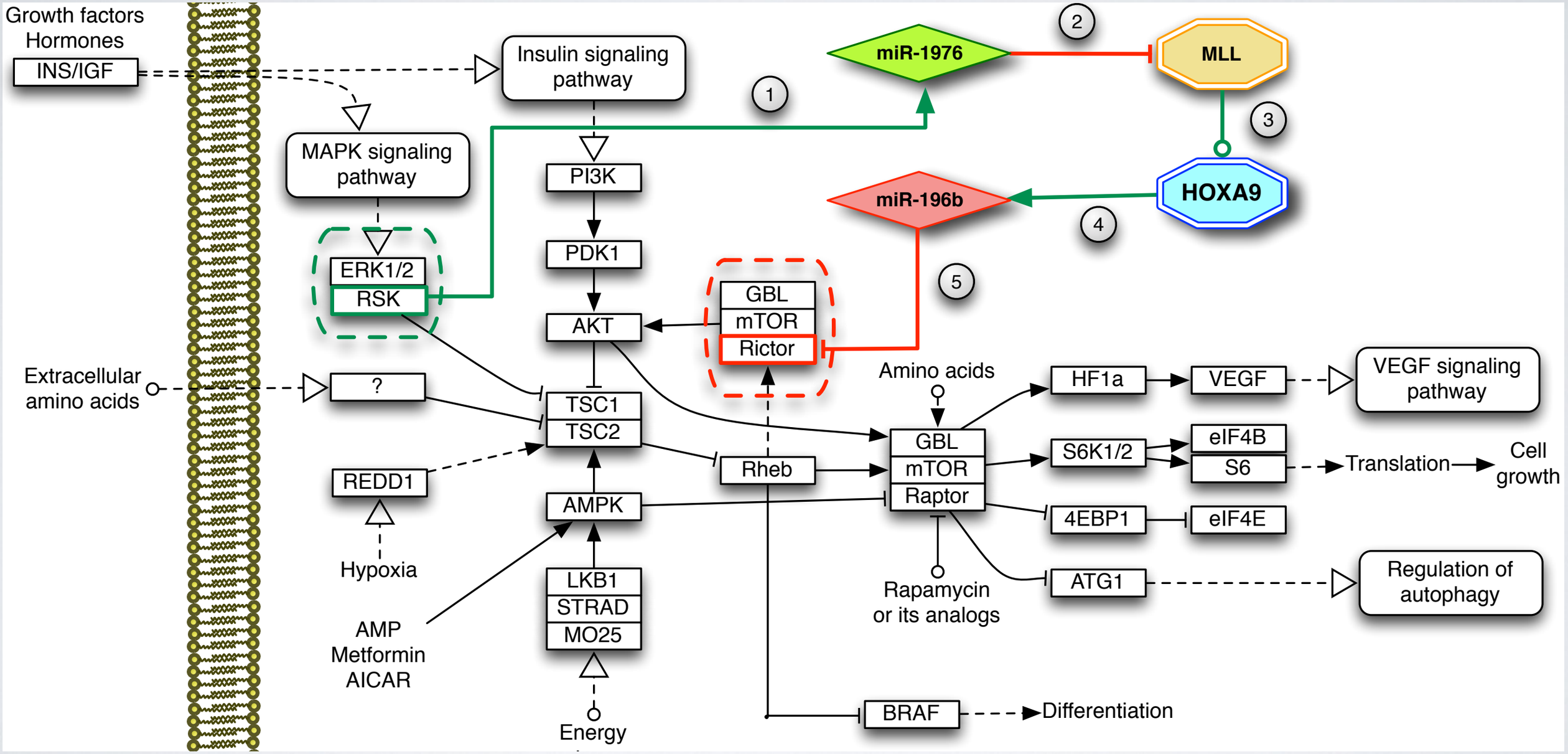
- **PPLs** appear in about **55% of non-metabolic pathways** while they appear only in about **9% of metabolic pathways**.

- From the statistical analysis, we can conclude that the presence of **PPLs** in the set of considered KEGGs **is not a random event.**

# Results


IGF1
INS
CAB39
KIAA1303
GBL
TSC1
LYK5
MAPK1
RPS6KA1
RPS6KA6
MAPK3
RPS6KA3
PIK3CG
PIK3CA
PIK3R3
PIK3R5
PIK3R1
PIK3CD
PIK3R2
PIK3CB
RPS6KA2
hsa-mir-1976
0.9427
MLL
HOXA9
hsa-miR-196b
0.364
RICTOR
PDPK1
STK11
AKT3
PRKAA2
PRKAA1
FRAP1
AKT2
AKT1
ULK3
ULK1
ULK2
DDIT4
TSC2
RHEB
BRAF
EIF4EBP1
HIF1A
RPS6KB2
RPS6KB1
EIF4E
EIF4E1B
EIF4E2
VEGFB
VEGFC
FIGF
PGF
VEGFA
EIF4B
RPS6

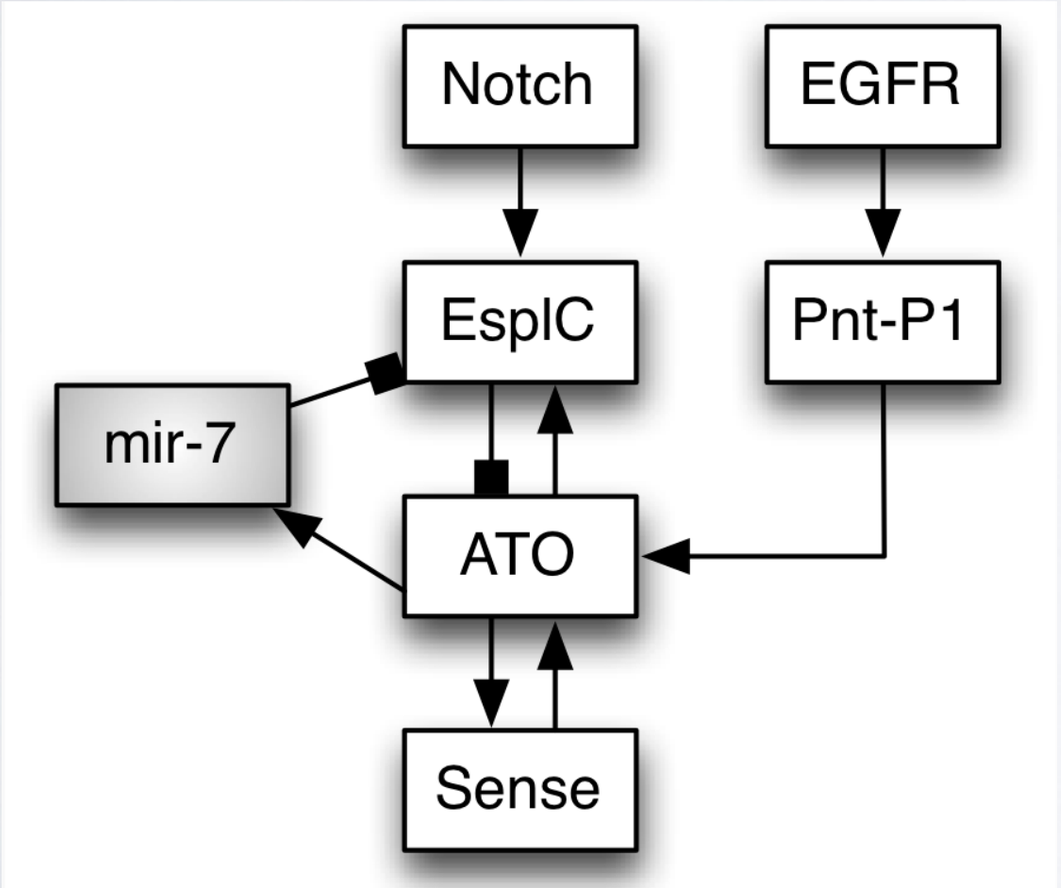
# Results


Growth factors
Hormones
INS/IGF
Insulin signaling pathway
MAPK signaling pathway
ERK1/2
RSK
PI3K
PDK1
AKT
GBL
mTOR
Rictor
miR-1976
MLL
HOXA9
miR-196b
1
2
3
4
5
Extracellular amino acids
?
TSC1
TSC2
REDD1
Hypoxia
AMPK
LKB1
STRAD
MO25
AMP
Metformin
AICAR
Energy
Rheb
Amino acids
GBL
mTOR
Raptor
Rapamycin
or its analogs
HF1a
VEGF
VEGF signaling pathway
S6K1/2
eIF4B
S6
Translation
Cell growth
4EBP1
eIF4E
ATG1
Regulation of autophagy
BRAF
Differentiation

# An example: Boolean Networks to Model Post-Transcriptional Regulation

**mir-7** belongs to an incoherent feed-forward loop.

This motif leads to an accelerated and transient pulse to downstream genes expression.

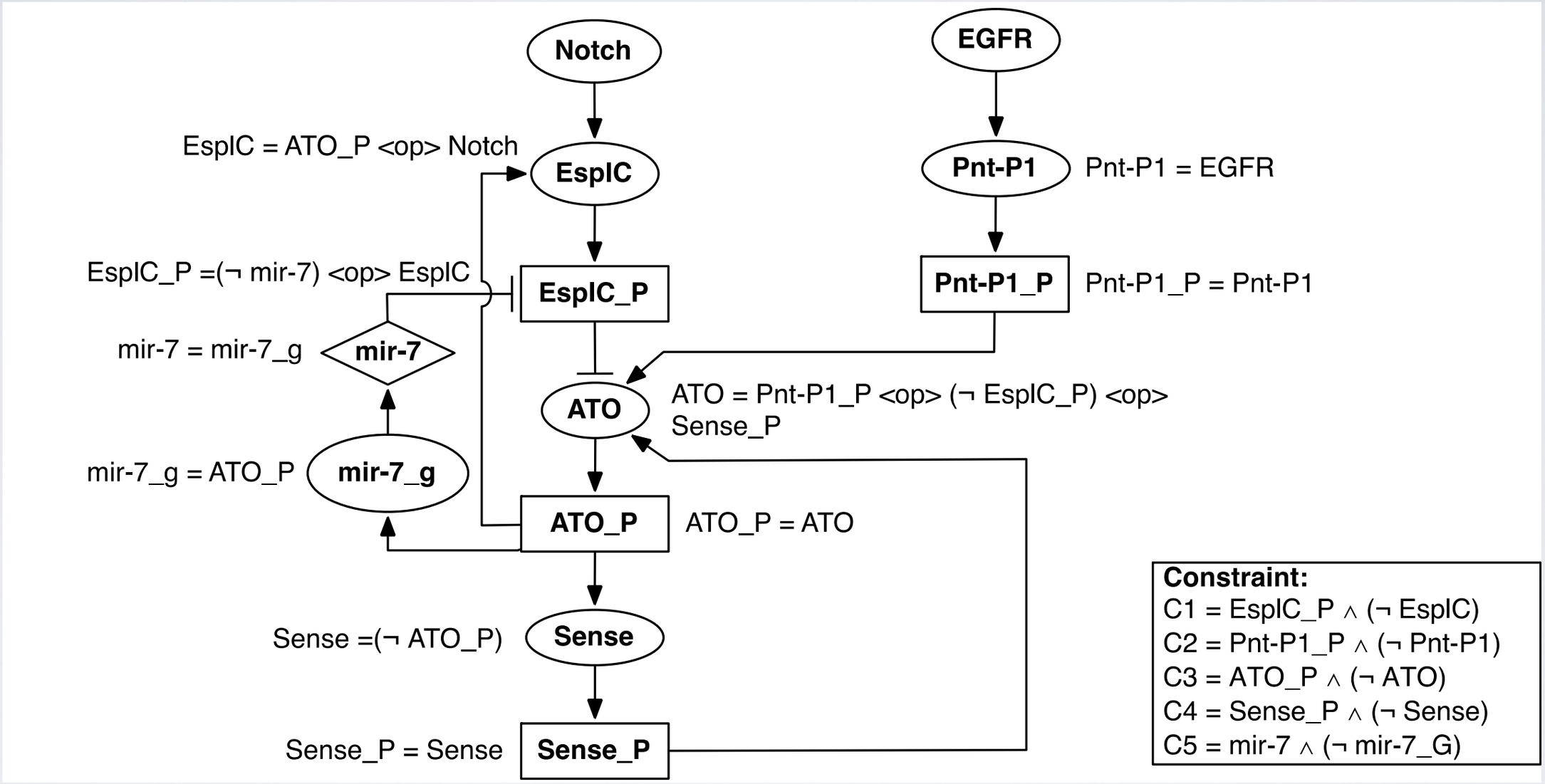



Sensory organ precursor (SOP) determination network

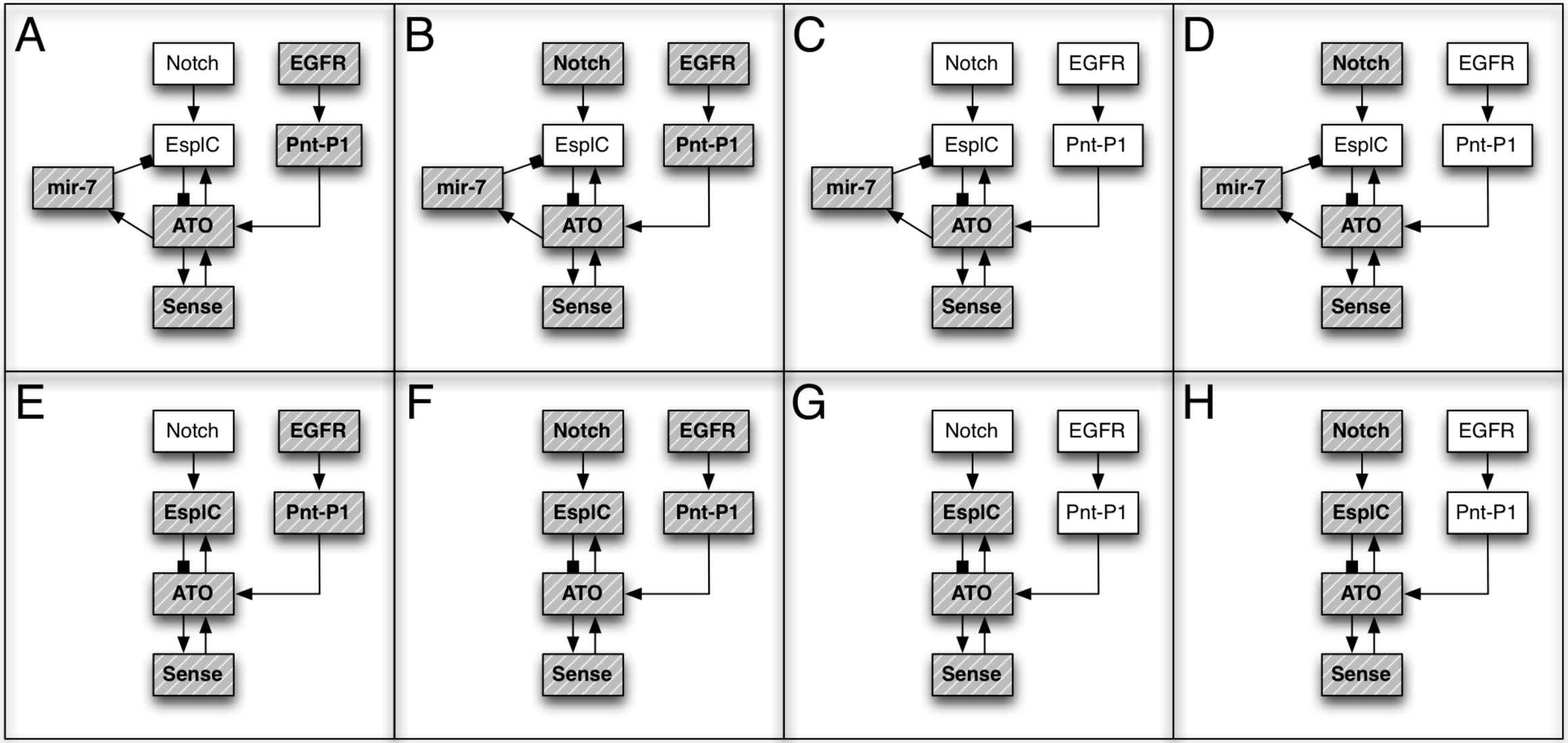
In the resulting network:

- fluctuating peaks of ATO would result in transient pulses of ATO repression by EsplC
- sustained increase of ATO would result in sustained repression of EsplC by miR-7 and a corresponding stabilization of ATO.

# Results


Notch
EsplC = ATO_P <op> Notch
EsplC
EsplC_P =(¬ mir-7) <op> EsplC
EsplC_P
mir-7 = mir-7_g
mir-7
EGFR
Pnt-P1
Pnt-P1 = EGFR
Pnt-P1_P
Pnt-P1_P = Pnt-P1
ATO
ATO = Pnt-P1_P <op> (¬ EsplC_P) <op> Sense_P
mir-7_g = ATO_P
mir-7_g
ATO_P
ATO_P = ATO
Sense =(¬ ATO_P)
Sense
Sense_P = Sense
Sense_P
Constraint:
C1 = EsplC_P ∧ (¬ EsplC)
C2 = Pnt-P1_P ∧ (¬ Pnt-P1)
C3 = ATO_P ∧ (¬ ATO)
C4 = Sense_P ∧ (¬ Sense)
C5 = mir-7 ∧ (¬ mir-7_G)

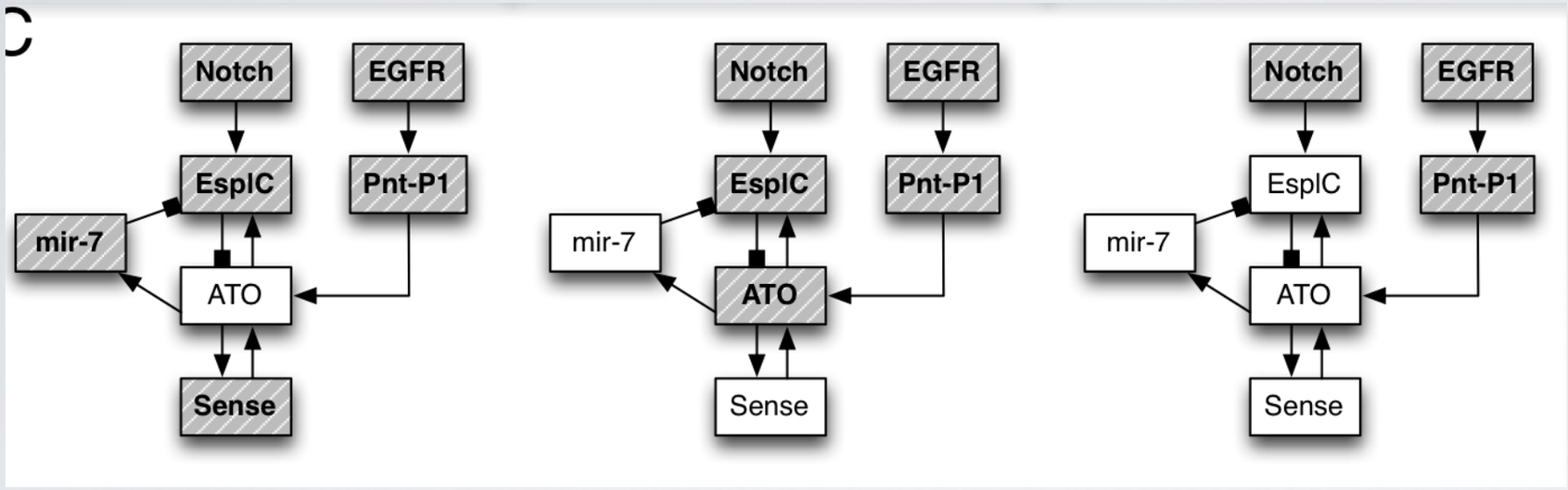
# Results




Increase of ATO would result in sustained repression of EsplC

by miR-7 and a corresponding stabilization of ATO.

# Results



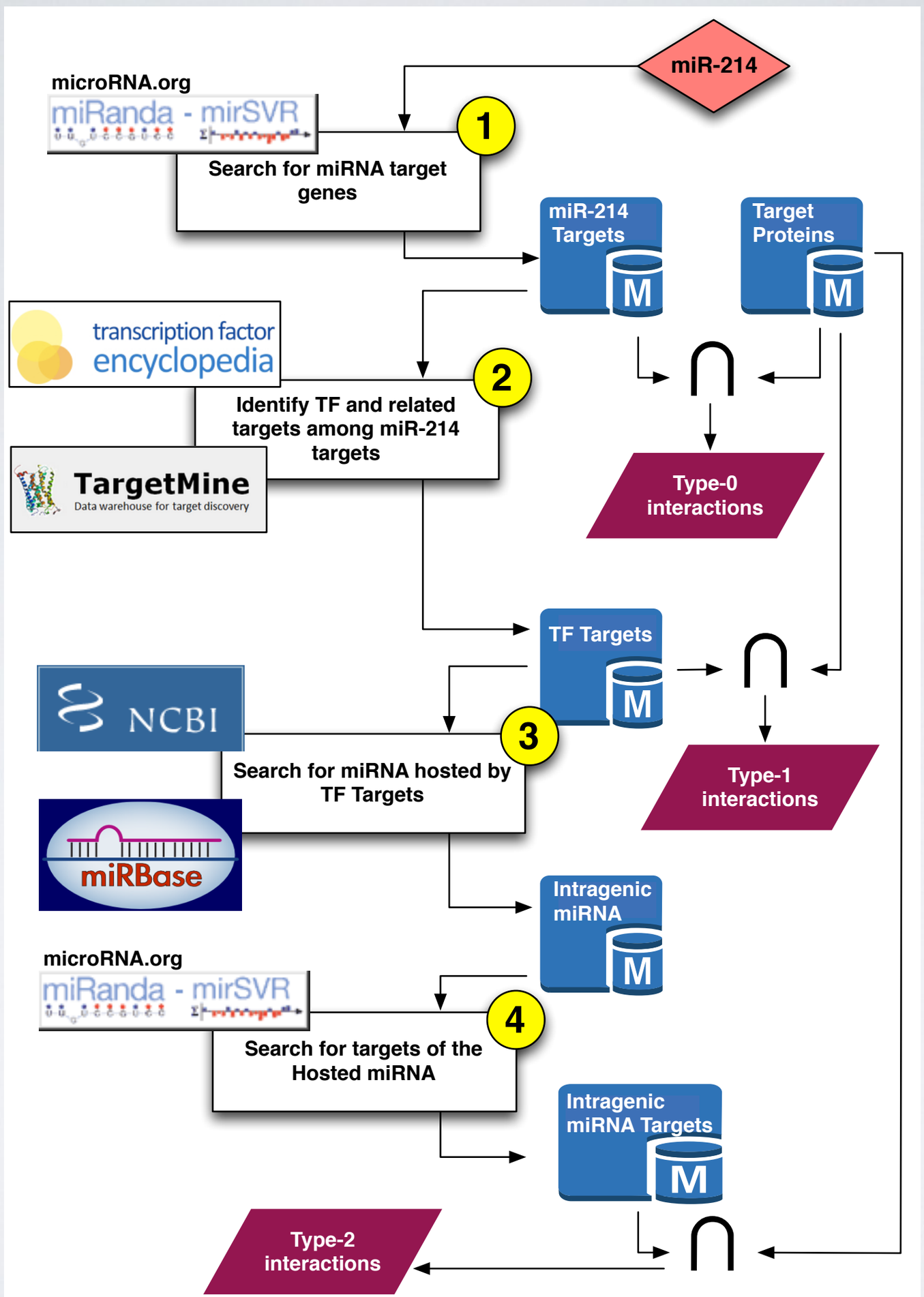


Fluctuating peaks of ATO would result in transient pulses of ATO repression by EsplC

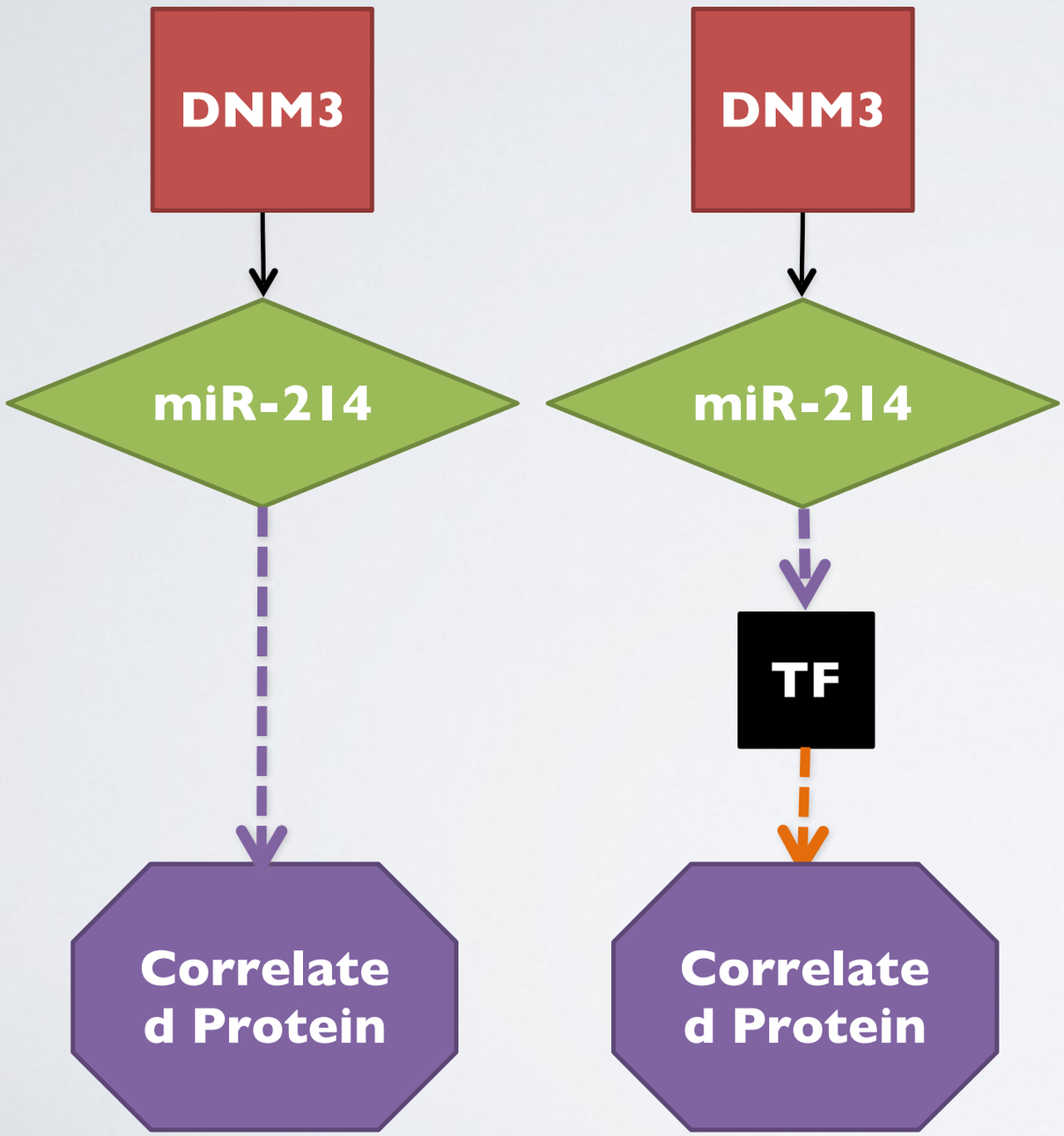
# An Example: Regulators of miR-214 in melanoma progression

1. Starting from 1 miRNA and a list of (73) Correlated Proteins, is it possible to search for regulatory modules?


2. Is it possible to generalize the approach?
    1. Investigate GRNs identifying regulatory modules
    2. to retrieve/merge/manage (post-)transcriptional regulations from public DBs

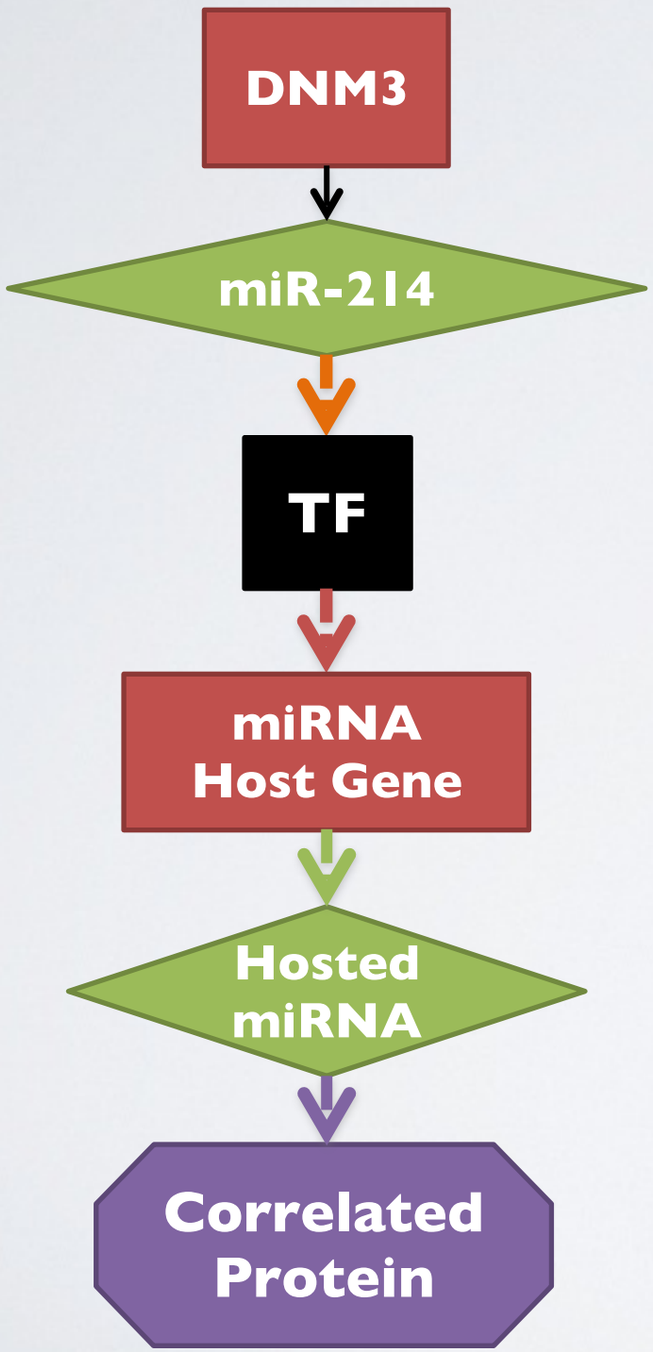

miR-214
microRNA.org
miRanda - mirSVR
1
Search for miRNA target genes
miR-214 Targets
M
Target Proteins
M
transcription factor encyclopedia
2
Identify TF and related targets among miR-214 targets
TargetMine
Data warehouse for target discovery
Type-0 interactions
TF Targets
M
NCBI
3
Search for miRNA hosted by TF Targets
miRBase
Type-1 interactions
Intragenic miRNA
M
microRNA.org
miRanda - mirSVR
4
Search for targets of the Hosted miRNA
Intragenic miRNA Targets
M
Type-2 interactions

# Results



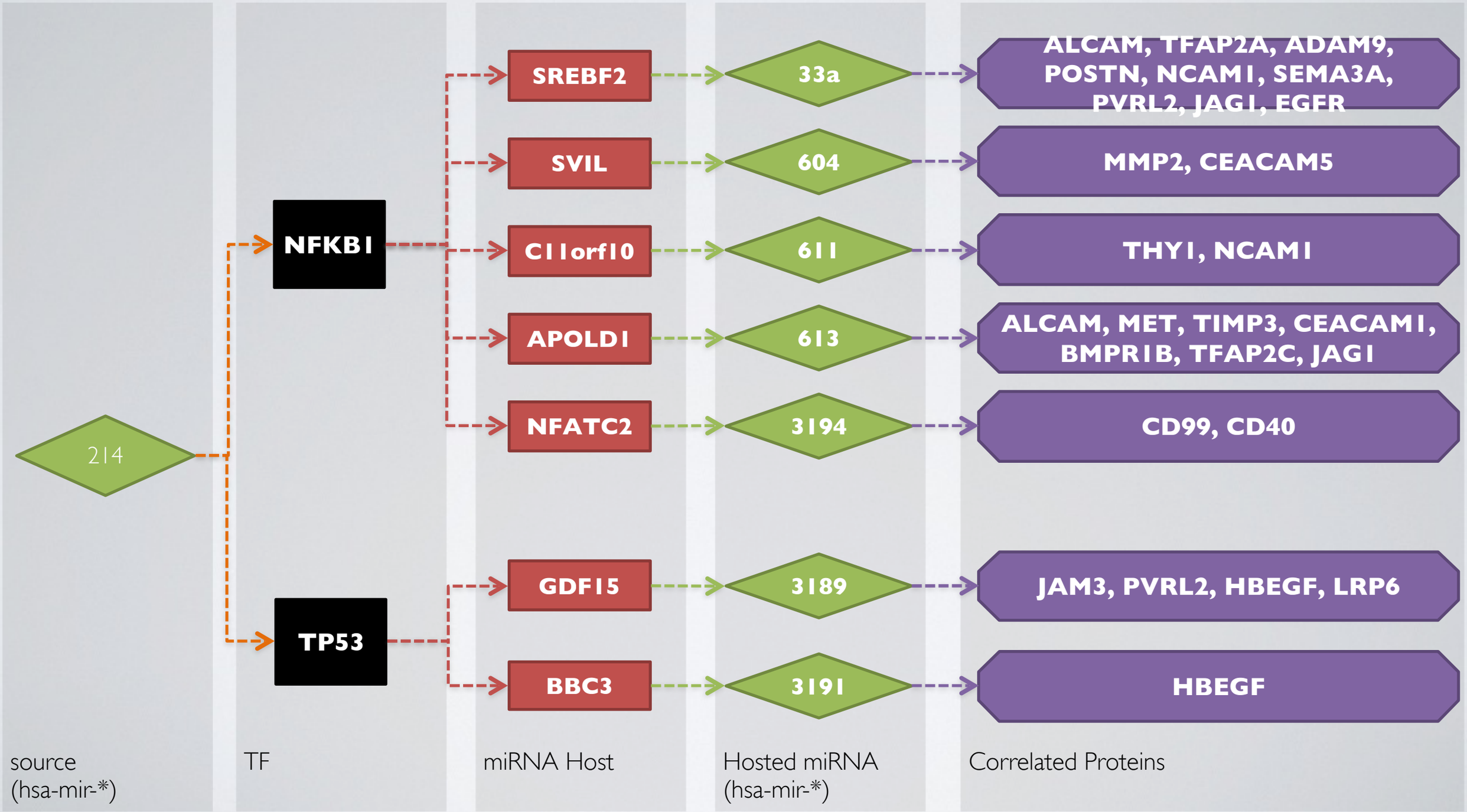

- The computational analysis didn't show Type-0 or Type-1 interactions
- It doesn't imply that they do not exist
- It does imply that there is no evidence of their presence in the available databases

# Results

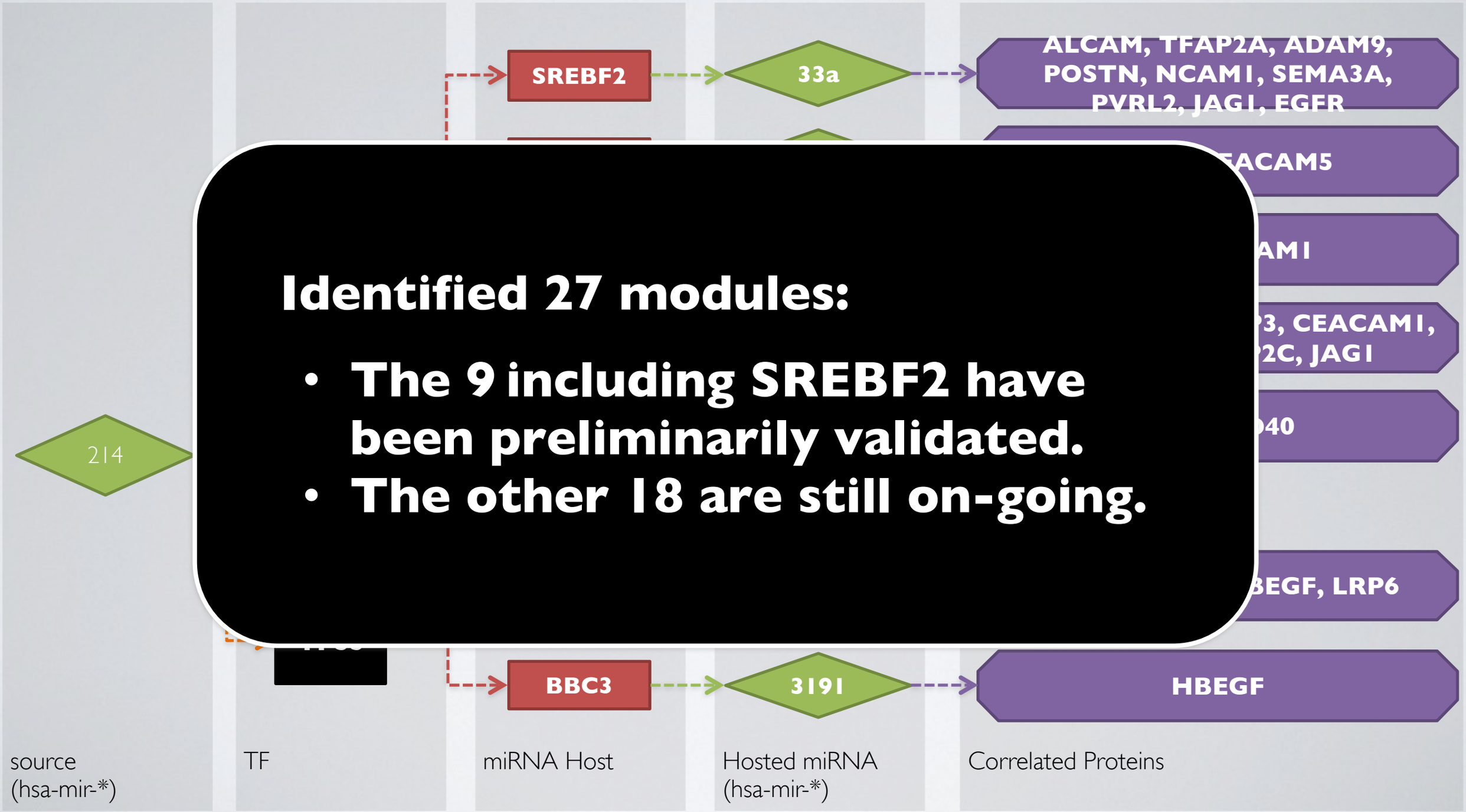



- No SIGN prediction for the resulting differential expression underlined by computational analysis
  - TargetMine does not include them.
  - TF Encyclopedia: no programmatically available.

# Results


NFKB1
TP53
214
SREBF2
SVIL
C11orf10
APOLD1
NFATC2
GDF15
BBC3
33a
604
611
613
3194
3189
3191
ALCAM, TFAP2A, ADAM9, POSTN, NCAM1, SEMA3A, PVRL2, JAG1, EGFR
MMP2, CEACAM5
THY1, NCAM1
ALCAM, MET, TIMP3, CEACAM1, BMPR1B, TFAP2C, JAG1
CD99, CD40
JAM3, PVRL2, HBEGF, LRP6
HBEGF
source (hsa-mir-*)
TF
miRNA Host
Hosted miRNA (hsa-mir-*)
Correlated Proteins

# Results




**Identified 27 modules:**

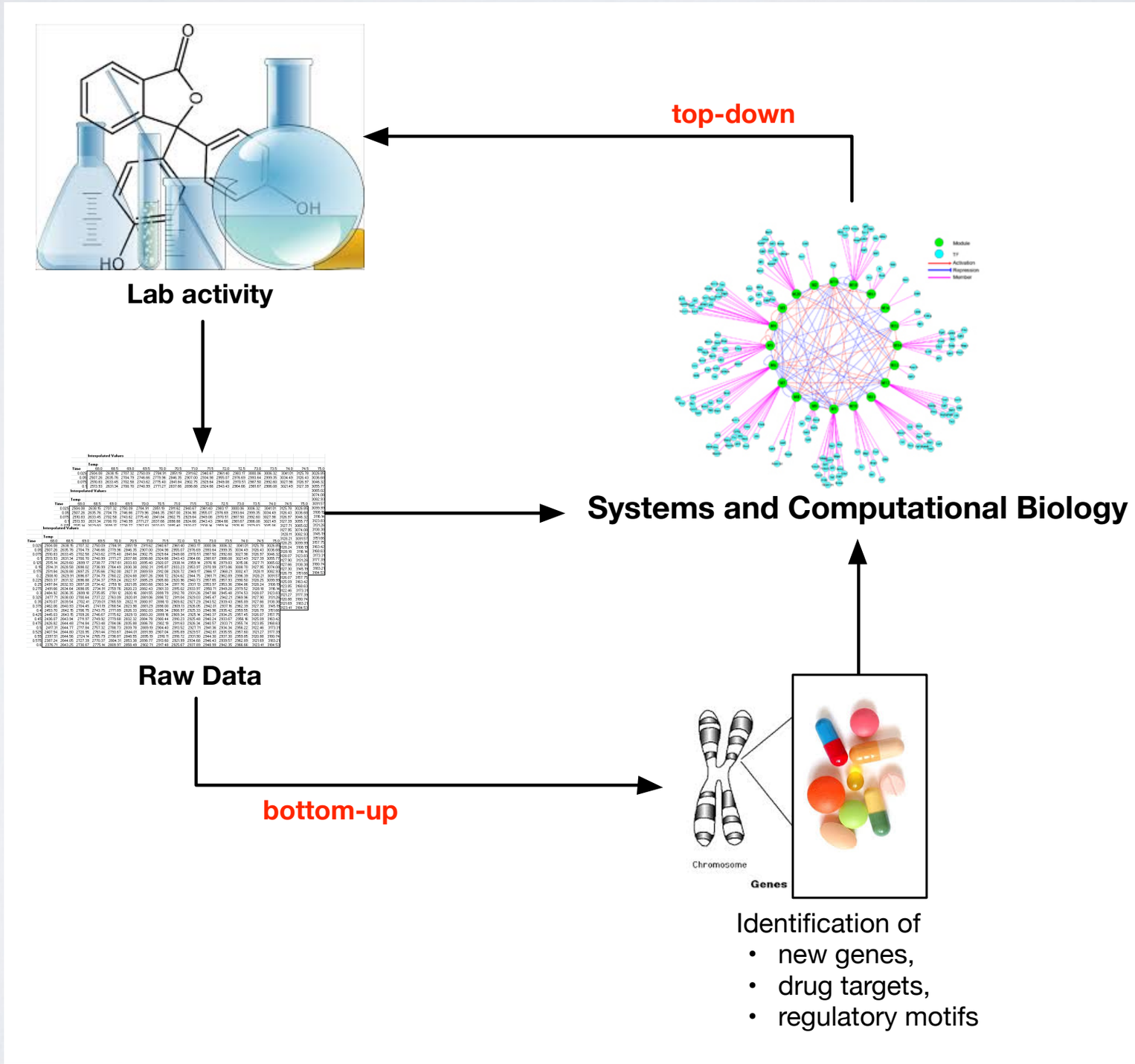
- **The 9 including SREBF2 have been preliminarily validated.**
- **The other 18 are still on-going.**

# Results

- Since miR-33a inhibits the motility of lung cancer cells (Rice et al. 2013), its down-regulation related to miR-214 overexpression could contribute to increase cell motility.

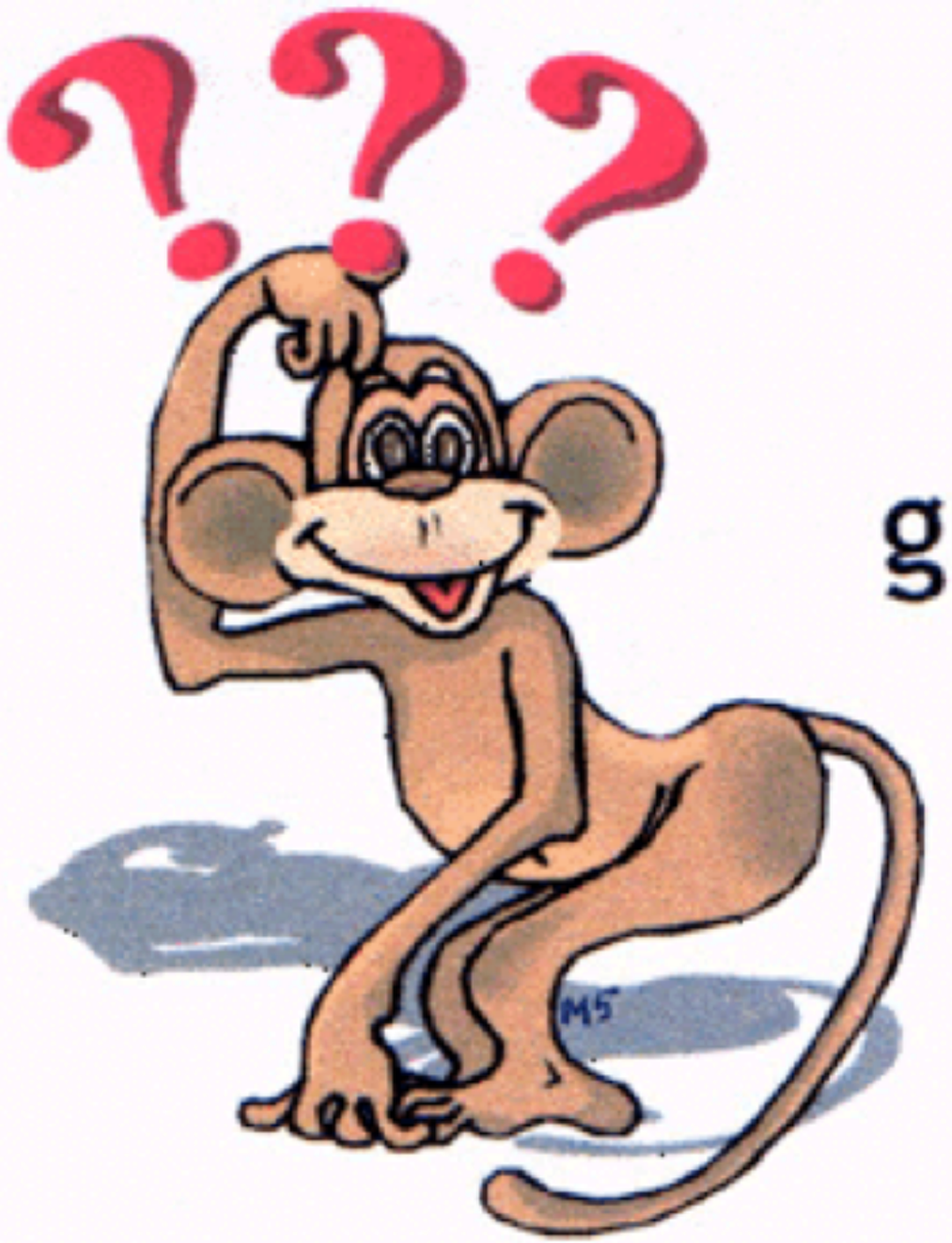
- Since SREBF2 and miR-33a act in concert to cholesterol homeostasis (Najafi-Shoushtari et al., 2010), and considering the lipogenic pathway as a metabolic hallmark of cancer cells, this confirms the potential role of miR-214 in cancer formation and progression.

# The role of **System Biology**



# The role of **System Biology**

- **Systems Biology and Computational Biology** are not only the latest fashion in biology, but a necessary step to overcome the limitations of both methodological approaches and **try to find a "middle ground"**.



Questions
are
guaranteed in
life;
Answers
aren't.