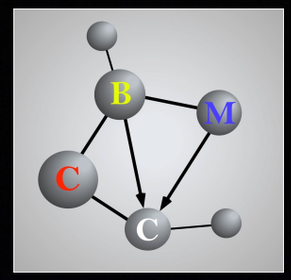
Lauree Magistrali interfacoltà:
LM-9 "Bioinformatica e Biotecnologie Mediche"
LM-18 "Medical Bioinformatics"



Systems Biology

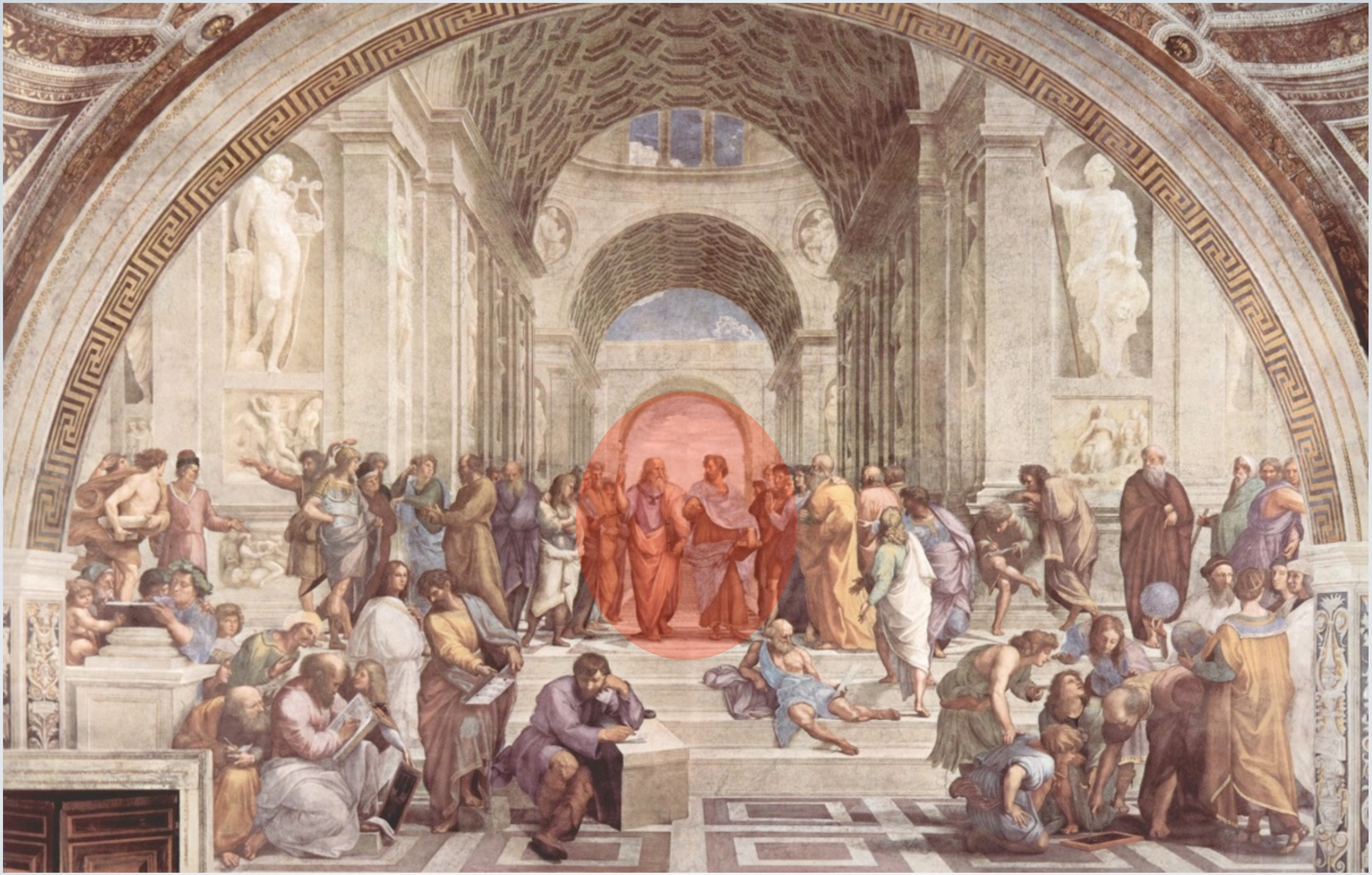
Prof. Carlo Laudanna

Division of General Pathology, Department of Medicine, University of Verona

"Laboratory of Cell Trafficking and Signal Transduction" (LCTST)
&
"The Center for Biomedical Computing" (CBMC)

tel.: 045-8027689
email: carlo.laudanna@univr.it
web: http://dp.univr.it/~laudanna/LCTST

Raffaello Sanzio, ***La Scuola di Atene (1509-1511)***



Raffaello Sanzio: "La Scuola di Atene" (1509-1511)

Central detail, the two main philosophers of ancient times: Plato and Aristoteles.

Plato, painted with the likeness of Leonardo da Vinci, holding in his left hand his work "Timaeus" and pointing to the sky with a finger (indicating the super-celestial, the area beyond the sky, where mental ideas do reside), while Aristoteles holds the "Ethics" and aimed his palm toward the ground, contacting the earthly world and the human will to study the world of nature and being in contact with it.

The painting illustrates the basic principle of Systems Biology, founded on the marriage between "*mathematical abstraction*" (the super-celestial, populating the world of Platonic ideas) and Aristotelian "*empiricism*", based on experimental observation of nature.

It is not a mathematics class!

It is not a bioinformatics class!

It is not a class of nonlinear physical systems!

Systems Biology is an highly integrated, multidisciplinary, discipline that is concerned with studying complex biological systems (**livings**).

The *final* goal is to identify the "emergent properties" of a biological system!

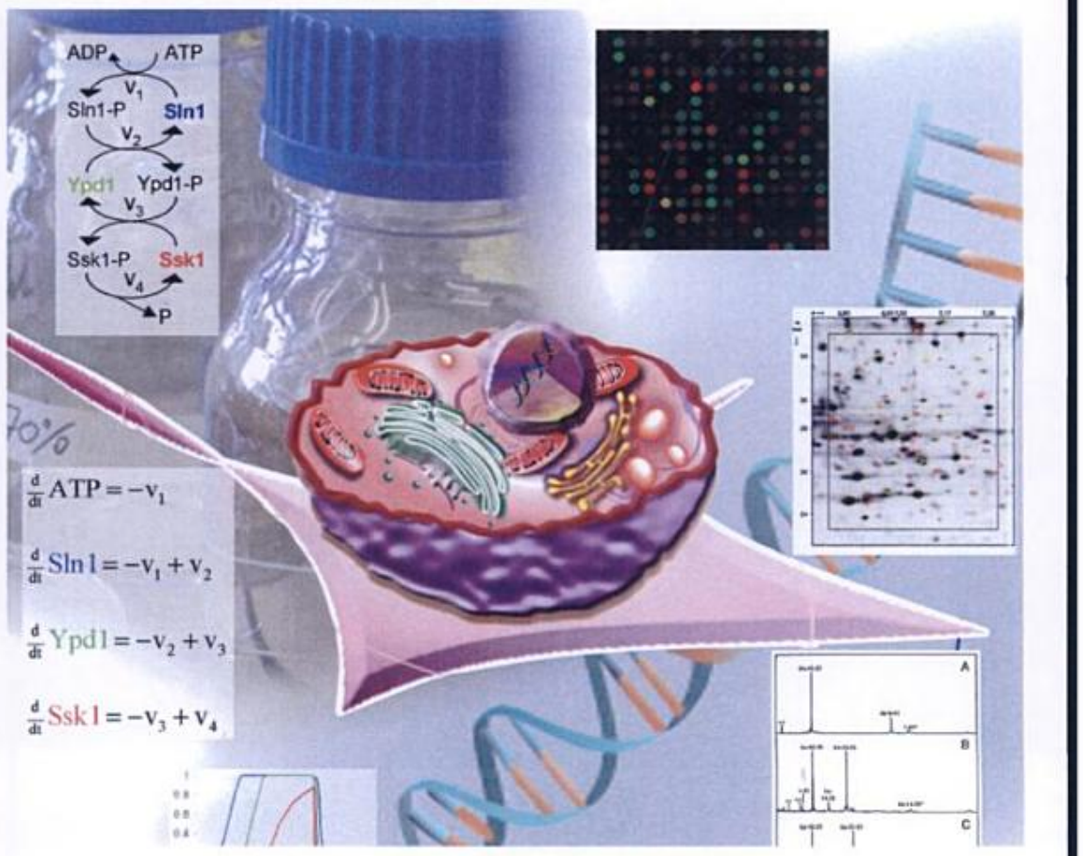
("theoretical biology", "integrative biology", "network biology", "multidimensional biology", "network medicine", "personalised medicine, "precision medicine" etc.")

E.Klipp, R.Herwig, A.Kowald,
C.Wierling, H.Lehrach

WILEY-VCH

Systems Biology in Practice

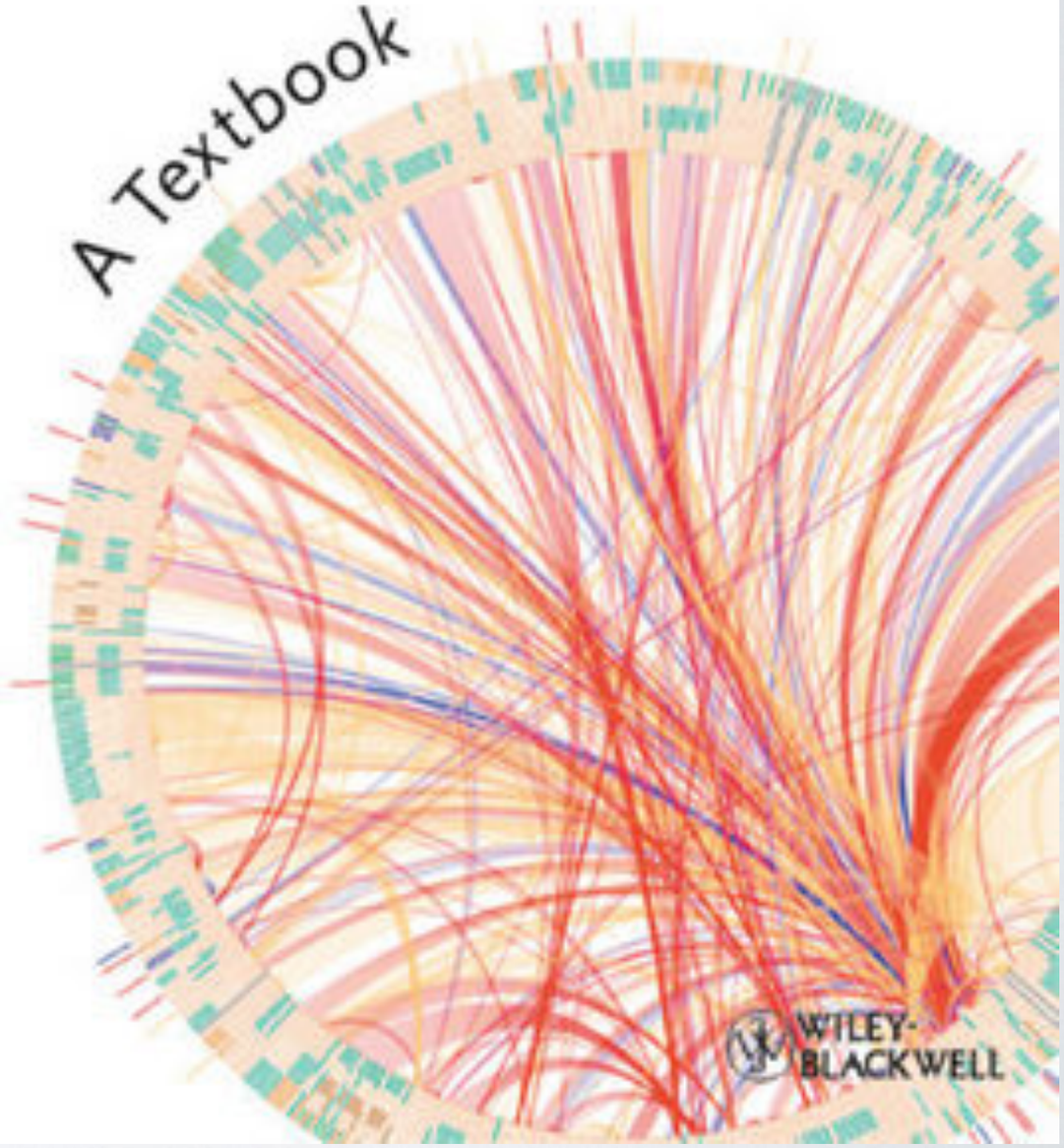
Concepts, Implementation and Application



Systems Biology

Edda Klipp, Wolfram Liebermeister, Christoph Wierling,
Axel Kowald, Hans Lehrach, and Ralf Herwig

A Textbook



WILEY-BLACKWELL

Annu. Rev. Genomics Hum. Genet. 2001. 2:343–72

A New Approach to Decoding Life: Systems Biology

Trey Ideker[1,2], Timothy Galitski[1], and Leroy Hood[1,2,3,4,5]
Institute for Systems Biology[1], Seattle, Washington 98105; Departments of Molecular Biotechnology[2], Immunology[3], Bioengineering[4], and Computer Science and Engineering[5], University of Washington, Seattle, Washington 98195; e-mail: tideker@systemsbiology.org, tgalitski@systemsbiology.org, lhood@systemsbiology.org

Key Words biological information, discovery sciences, genome, proteome

■ **Abstract** Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations. The emergence of systems biology is described, as are several examples of specific systems approaches.

INTRODUCTION

Perhaps the most important consequence of the Human Genome Project is that it is pushing scientists toward a new view of biology—what we call the systems approach. Systems biology does not investigate individual genes or proteins one at a time, as has been the highly successful mode of biology for the past 30 years. Rather, it investigates the behavior and relationships of all of the elements in a particular biological system while it is functioning. These data can then be integrated, graphically displayed, and ultimately modeled computationally. How has the Human Genome Project moved us to this new view? It has done so by catalyzing a new scientific approach to biology, termed discovery science; by defining a genetic parts list of human and many model organisms; by strengthening the view that biology is an informational science; by providing us with powerful new high-throughput tools for systematically perturbing and monitoring biological systems; and by stimulating the creation of new computational methods.

Discovery Science

The Human Genome Project was one of the first modern biological endeavors to practice discovery science. The objective of discovery science is to define all of the elements in a system and to create a database containing that information. For

1527-8204/01/0728-0343$14.00

Review in Advance first posted online on July 6, 2010. (Changes may still occur before final publication online and in print.)

A Decade of Systems Biology

Han-Yu Chuang,[1,2,*] Matan Hofree,[3,*] and Trey Ideker[1–4]

[1]Division of Medical Genetics, Department of Medicine; [2]Bioinformatics Program; [3]Department of Computer Science and Engineering; and [4]Department of Bioengineering, University of California, San Diego, La Jolla, California 92093; email: tideker@ucsd.edu

Annu. Rev. Cell Dev. Biol. 2010. 26:23.1–23.24

The *Annual Review of Cell and Developmental Biology* is online at cellbio.annualreviews.org

This article's doi:
10.1146/annurev-cellbio-100109-104122

1081-0706/10/1110-0001$20.00

*These authors contributed equally to this work.

Key Words

networks, models, high-throughput technology, genome-wide measurements, developmental biology, cell fate, biomarkers, epistasis, genetic interaction, genome-wide association study (GWAS)

Abstract

Systems biology provides a framework for assembling models of biological pathways from systematic measurements. Since the field was first introduced a decade ago, considerable progress has been made in technologies for global cell measurement and in computational analyses of these data to map and model cell function. It has also greatly expanded into the translational sciences, with approaches pioneered in yeast now being applied to elucidate human development and disease. Here, we review the state of the field with a focus on four emerging applications of systems biology that are likely to be of particular importance during the decade to follow: (*a*) pathway-based biomarkers, (*b*) global genetic interaction maps, (*c*) systems approaches to identify disease genes, and (*d*) stem cell systems biology. We also cover recent advances in software tools that allow biologists to explore system-wide models and to formulate new hypotheses. The applications and methods covered in this review provide a set of prime exemplars useful to cell and developmental biologists wishing to apply systems approaches to pathways of interest.

INTERFACE

rsif.royalsocietypublishing.org

Perspective

Cite this article: Ma'ayan A. 2017 Complex systems biology. *J. R. Soc. Interface* **14**: 20170391.
http://dx.doi.org/10.1098/rsif.2017.0391

Received: 28 May 2017
Accepted: 29 August 2017

Subject Category:
Reviews

Subject Areas:
biocomplexity, systems biology, computational biology

Keywords:
complexity, agents, evolution

Author for correspondence:
Avi Ma'ayan
e-mail: avi.maayan@mssm.edu

THE ROYAL SOCIETY
PUBLISHING

Complex systems biology

Avi Ma'ayan

BD2K-LINCS Data Coordination and Integration Center; Mount Sinai Center for Bioinformatics; Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA

AM, 0000-0002-6904-1017

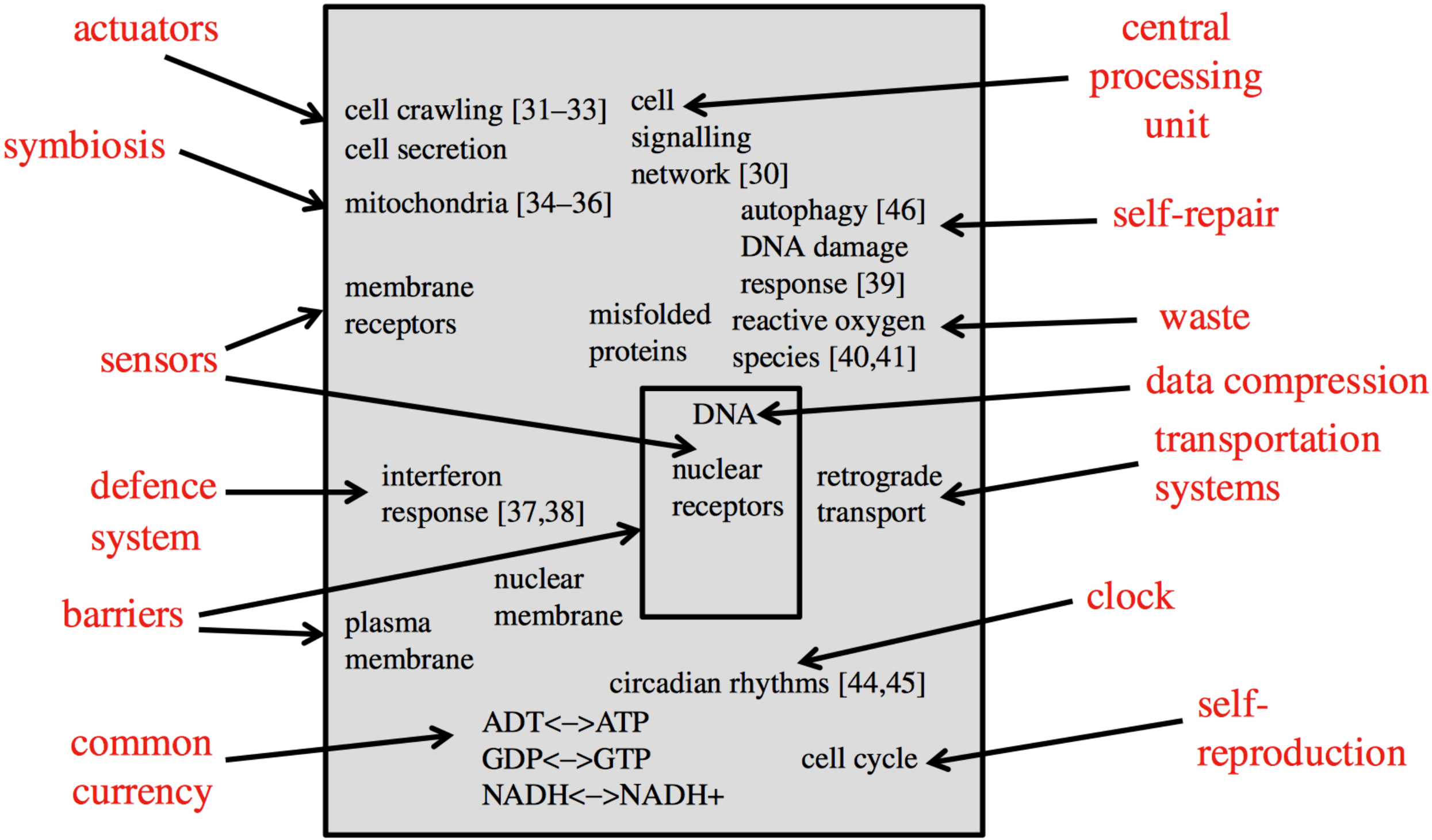
Complex systems theory is concerned with identifying and characterizing common design elements that are observed across diverse natural, technological and social complex systems. Systems biology, a more holistic approach to study molecules and cells in biology, has advanced rapidly in the past two decades. However, not much appreciation has been granted to the realization that the human cell is an exemplary complex system. Here, I outline general design principles identified in many complex systems, and then describe the human cell as a prototypical complex system. Considering concepts of complex systems theory in systems biology can illuminate our overall understanding of normal cell physiology and the alterations that lead to human disease.

1. The science of complex systems theory

Science and technology allow us to understand our environment as well as manipulate it and create new environments and new systems. This led humans to emerge out of nature, and recently to create new complex worlds that highly resemble natural systems [1]. Human-made systems often follow the same design principles governing natural systems. The most important of these design principles is evolution by natural selection [2]. However, human-made systems are not exactly the same as those created by nature. We are gaining an increasing ability to create new complex environments and new machines that perform as well as, or even better than, natural organisms [3]. Man-made complex systems, such as stock markets, or multi-user social online networks, and technologies that can be used to collect and process increasing amounts of data offer us an opportunity to better observe and understand complex systems, natural or man-made. We can increasingly measure the activity of the variables that constitute these systems. This provides a better glimpse at the quantity and connectivity of most variables that control a complex system. When all these variables work together, they make up a system that appears to us as one unit that is alive.

We are beginning to realize that, in general, complex systems, man-made or natural, share many universal design patterns; concepts and principles of design that reappear in diverse, seemingly unrelated systems [4,5]. These design patterns are the essential elements for building successful complex systems that can function, compete, survive, reproduce and evolve for long periods through multiple generations towards increased fitness and overall growth. The science of complex systems theory attempts to gain an understanding about these emerging repeating design principles that reappear in different natural and man-made complex systems and environments [6]. The goal of complex systems science is to define more precisely these properties towards a greater understanding of complex systems as a whole, beyond the understanding of one specific system, or one specific design concept. Better understanding these universal principles will enable us to better digest the rapid changes that occur around us due to technological and social evolution [3]. To study and understand complex systems, when possible, researchers conduct

The cell as complex dynamic system



Topics covered in the "Systems Biology" class

Systems Biology - general concepts - foundations:

1) **Complexity**: definition, origin and nature of complexity in biology
2) The "**emergent properties**" of biological systems: the cellular and molecular circuits
3) Science based on thesis and the deductive method; science based on experimental data the inductive method
4) **Systems Biology**: definition and experimental connotation of Systems Biology
5) Why Systems Biology? The reductionist approach versus the holistic approach
6) The concept of **model**: predict the future in biology?
7) **Static models**: the network abstraction and the topological properties of biological networks
8) **Dynamic models** and biological kinetics

Methods in Systems Biology:

9) High-performance technologies (high throughput methods)
10) Bioinformatics
11) Biological database
12) Software for Systems biology
13) Contexts of Systems Biology: transcriptomics, proteomics, metabolomics, and other "omics".

Systems Biology in practice - applications of Systems Biology to biomedical contexts

14) Networks and diseases
15) The immune system
16) Inflammatory mechanisms
17) Cancer
18) Neurodegenerative diseases
19) Autoimmune diseases
20) Systems pharmacology and drug discovery

Laboratory (12 hours) for the "Systems Biology” class

L1: Mastering Cytoscape

L2: Biological data-bases 1

L3: Biological data-bases 2

L4: Network analysis in adoptive immunity: case study

L5: Metabolic networks in inflammatory diseases: case study

L6: Network analysis in neurodegenerative diseases: case study

Laboratory for the "Systems Biology" class

L1: Mastering Cytoscape:

The file formats, creating networks, extracting sub-networks, generating node information layers, performing topological analysis, applying modular decomposition and intersections tools, exporting data and images, etc.

L2: Biological data-bases 1:

https://www.genenames.org
https://omictools.com
https://www.targetvalidation.org
https://www.ebi.ac.uk
http://www.uniprot.org
https://www.ncbi.nlm.nih.gov
https://www.ncbi.nlm.nih.gov/gene
https://blast.ncbi.nlm.nih.gov/Blast.cgi

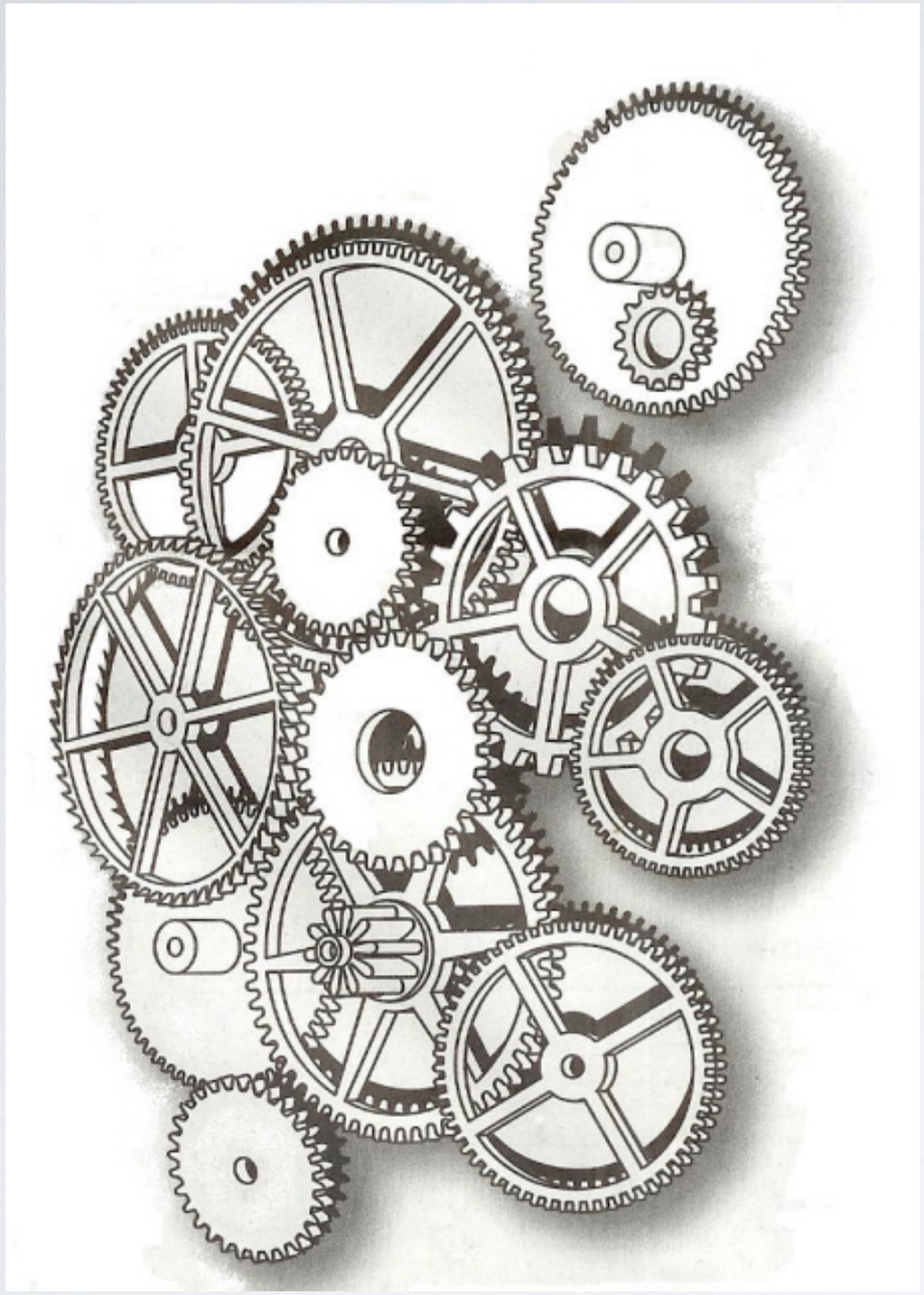
L3: Biological data-bases 2

http://www.geneontology.org
http://amigo.geneontology.org/amigo
https://www.phosphosite.org/homeAction.action
https://string-db.org
https://thebiogrid.org
https://reactome.org
http://www.genome.jp/kegg/
http://disease-connect.org

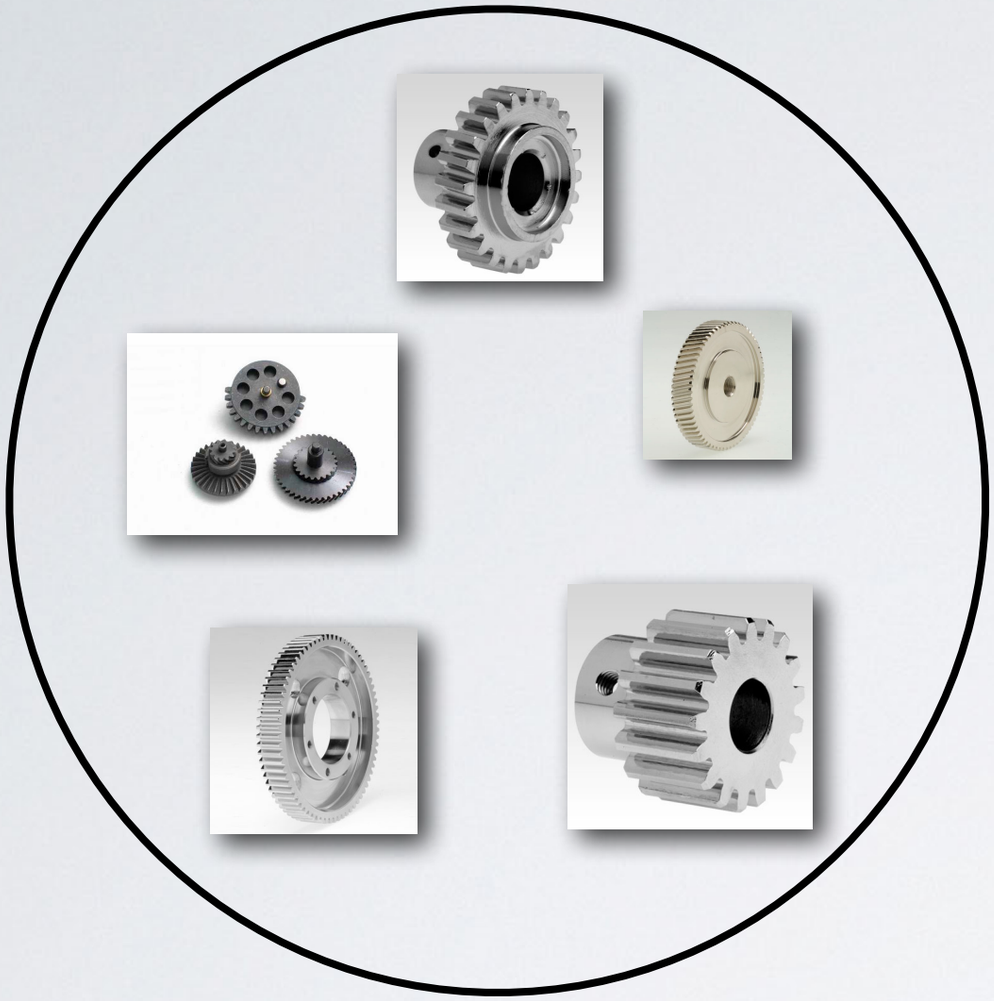
(1)
Complexity: definitions

Gears

Collection of
“sparse” (scattered) gears



Collection of “interacting” gears according to a specific design (architecture)!



Set of objects

↓

NO function

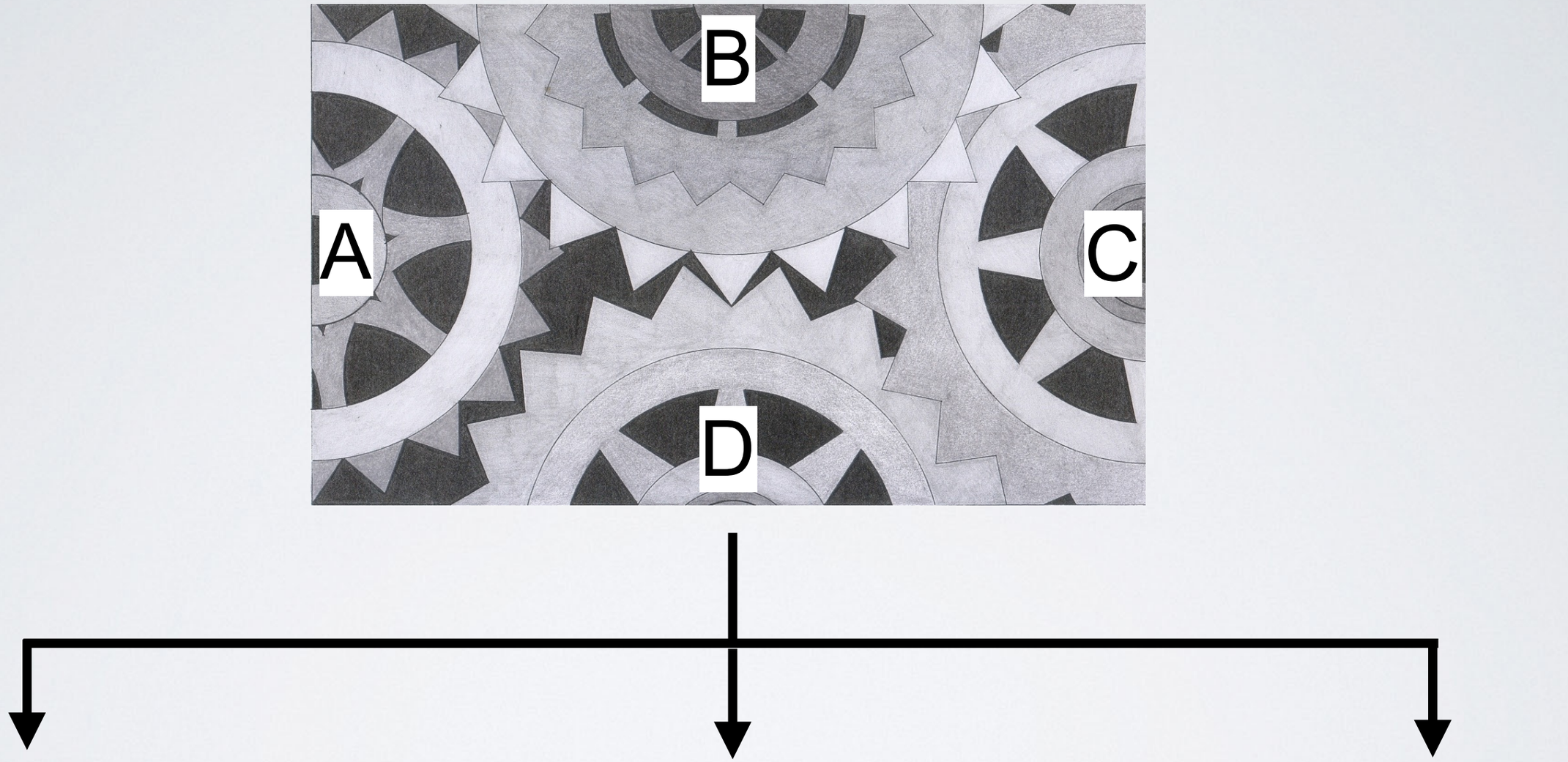
System

↓

Function

(time measuring)

System (mathematical, physical, biological) = set of interacting objects - physically and/or functionally



Simple ⟷ *Complicated* ⟷ Complex

“Simple” System

- It is a set of interacting elements in such a way that the whole of the global behavior is the **simple sum** of the sub-behaviors of the individual components of the set.

- Its behavior is **linear**: a system is linear if responds in a way directly proportional to the stress received. For it applies the principle of ***effect superposition***:

if, upon S1 stress, the system gives the answer R1, and upon S2 stress gives the answer R2, then upon combined stress (S1 + S2) the answer is (R1 + R2) (deterministic predictability).

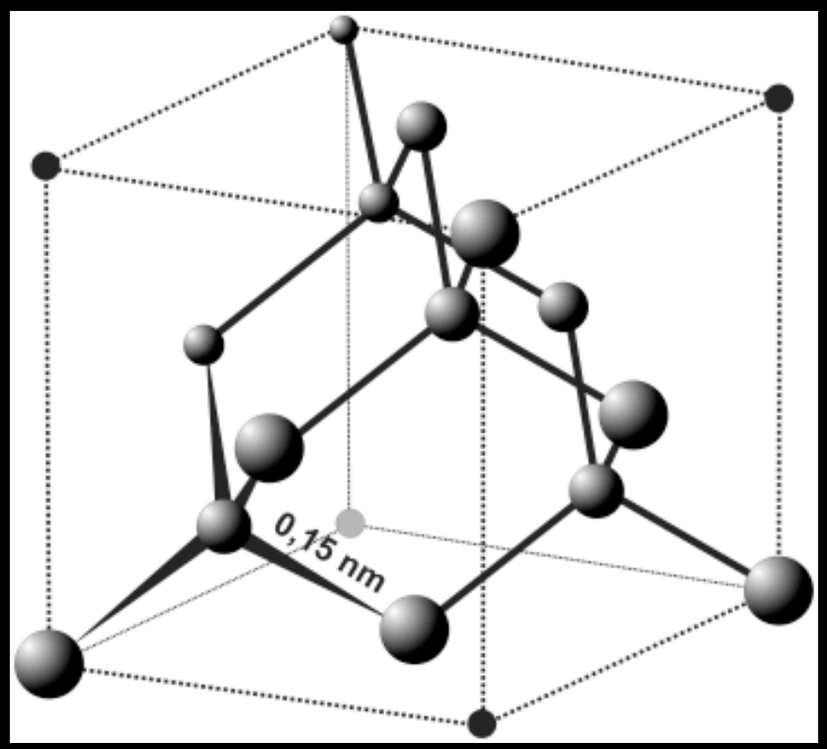
- It requires very little information to be exhaustively described

A system can be "simple"
BUT
at the same time "complicated"

A system is complicated if it is difficult to understand its structure and / or function because it contains a large number of elements, which must be defined one by one

A diamond: "simple" but "complicated"

Reticulum of carbon atoms in a tetrahedral structure (lattice)



Chemical formula: C
Crystal Group: monometric
Crystal system: Cubic

It is quite “laborious” to calculate the fracturing plane

“Complex” system

- It is a set of interacting elements in such a way that its global behavior is NOT derived from the sum of the behaviors of the individual elements of the system

- The behavior is **non-linear:** it is non-linear a system that responds disproportionately to solicitations received (triggering of status changes).

  (S1,R1) (S2,R2) => (S1+S2) (R1+R2) = linear

  (S1,R1) (S2,R2) => (S1+S2) (R1+R2+**R3**) = not linear

- **"Emergent" behavior**: The emergent behavior is a consequence of non-linearity. In a linear system new properties do not appear that are not already present in the individual elements. In a non-linear system, the interacting elements are *functionally* dependent from each other: their combination brings out properties not corresponding to the simple sum of the properties of the individual parts; new properties do appear, called **emergent properties**, absent in the individual, BUT that depend entirely on the "system design”.

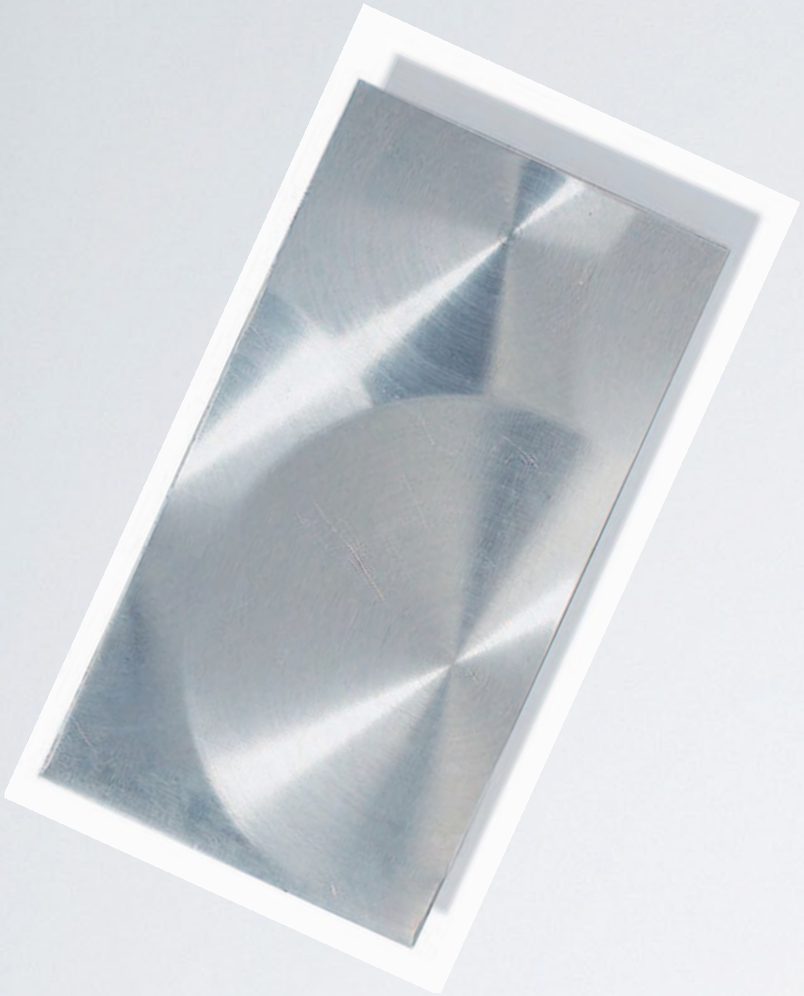
- "Sensitivity to initial conditions" = **butterfly effect**

- Possibility of **chaos** (unpredictable evolution of the system, concept of **space phase**)

- **Dissipation** (apparent violation of the second law of thermodynamics)

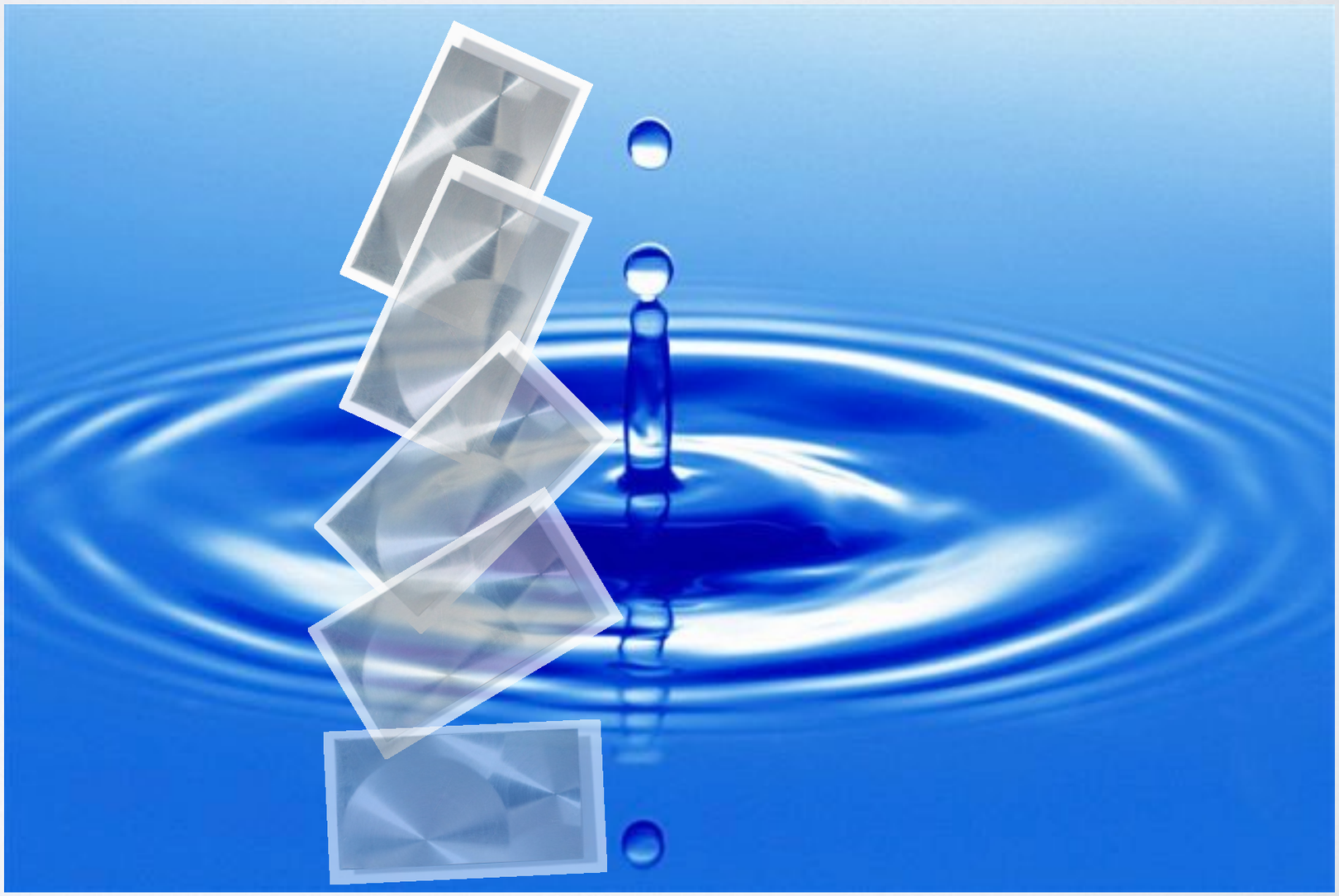
- **Self-organization (biological systems)**

Emergent properties



Single
steel sheet

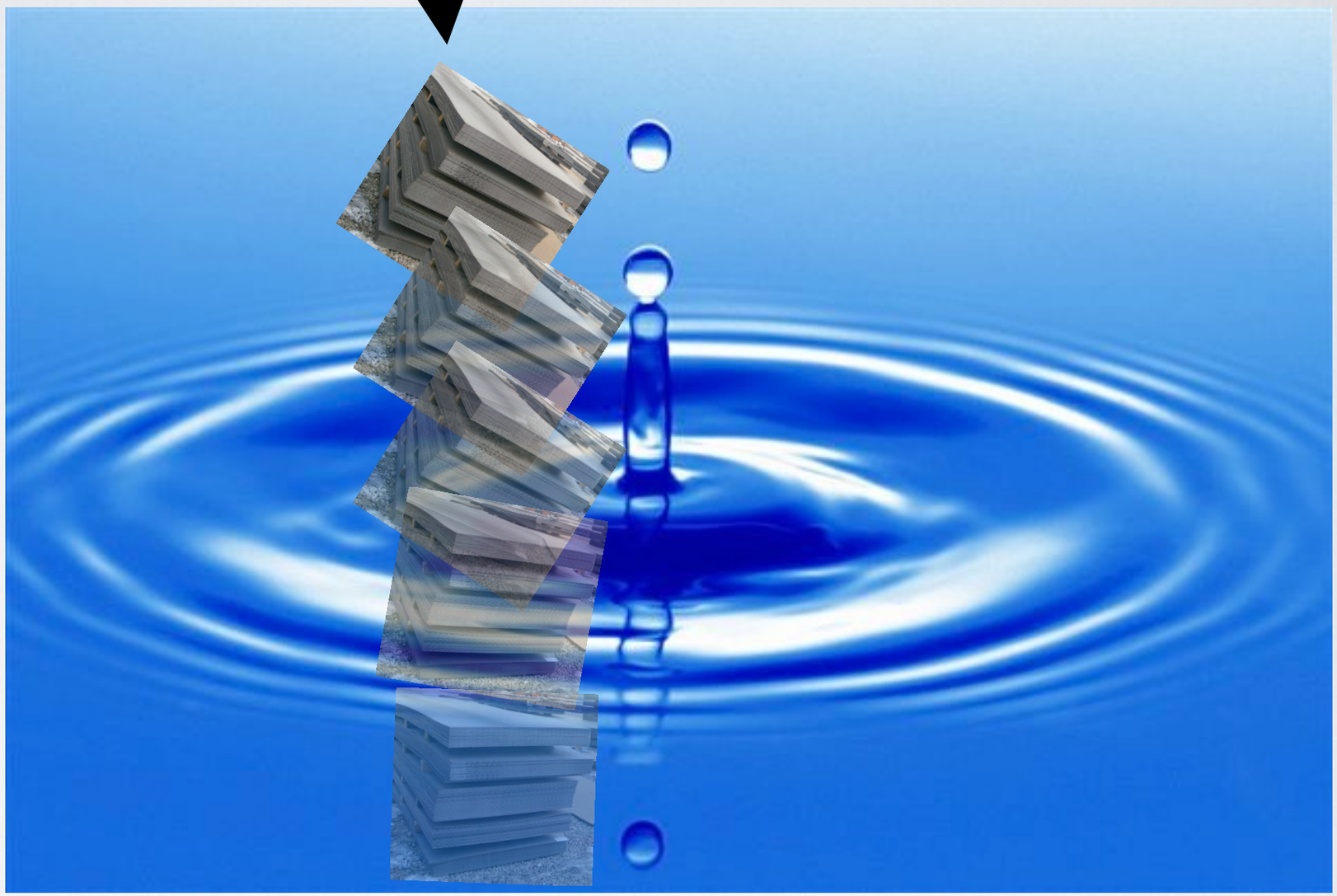
Drop in water



Sinking !

A set of steel sheets

Drop in water



Sinking !

300000 steel sheets + 6000000 steel rivets

=

organization

Steel ship

NO Sinking !

≠

Flotation
NO

Flotation
YES

Why the steel sheets do sink whereas the ship does floats?

The sheets are mounted in such a way that constitute a "system" that surrounds a space with an empty volume (the boat) that move an amount of water with a global mass superior to the whole of the sheets themselves.

*Moved water generates a "**pushing force**" from the bottom upwards equal to the weight (mass = force) of the displaced liquid volume" which exerts a floating force (Archimedes' principle)*

The ship is a **system**

Floating is an exclusive property of the structure of the system = emergent property of the system (isolated metal sheets do sink)

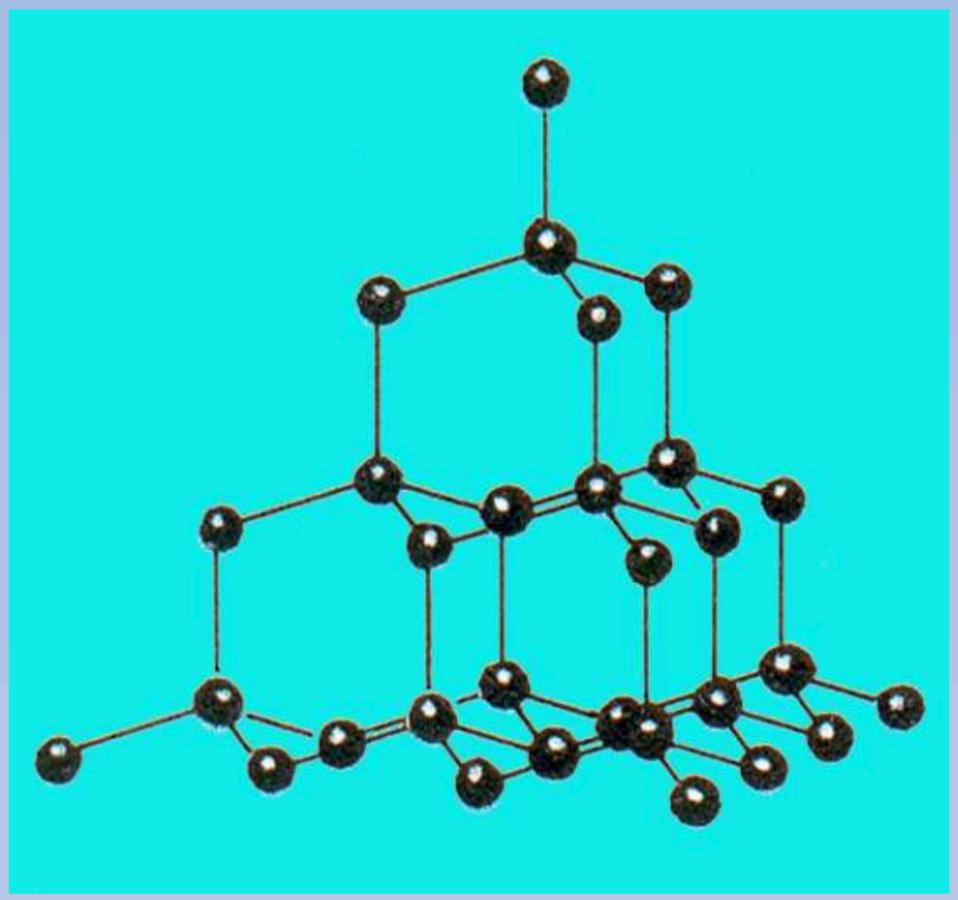
Archimede's principle establishes the mathematical rule that governs the emergent property, and allows its quantification

$$F_A = \rho_{flu}\, g\, V$$

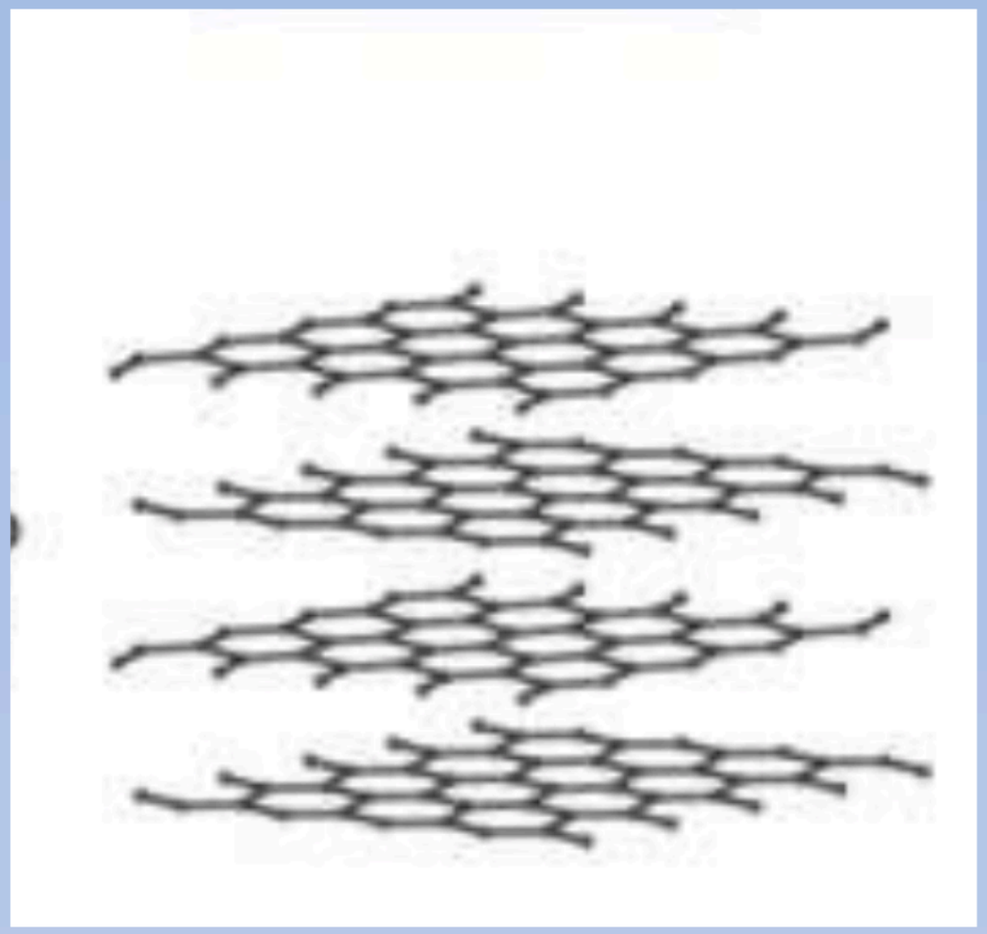
$\rho_{flu\ water} = 1$

$\rho_{flu\ Hg} = 13.6$

Diamond



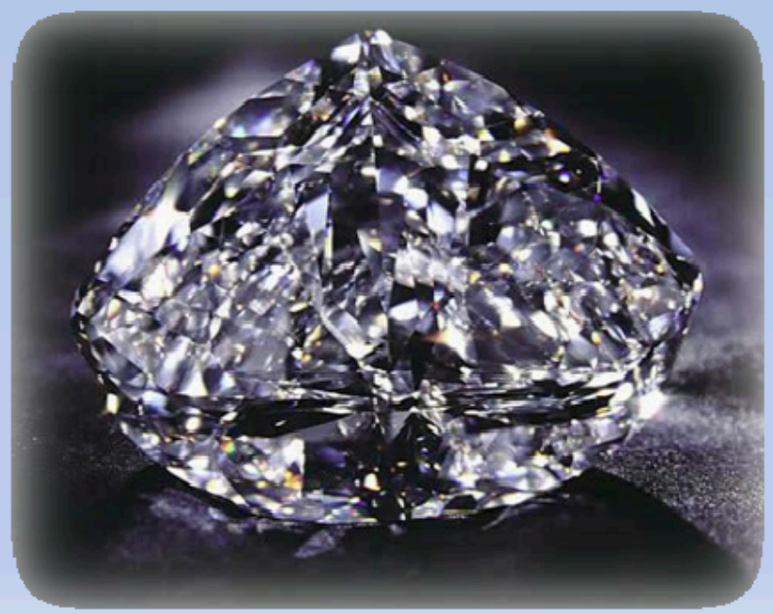
Graphite



Diamond

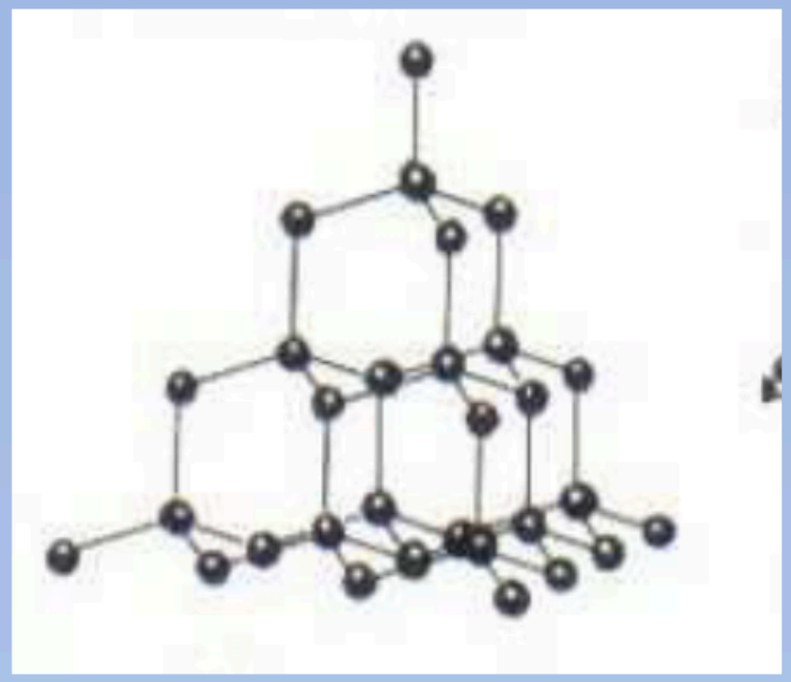
Graphite

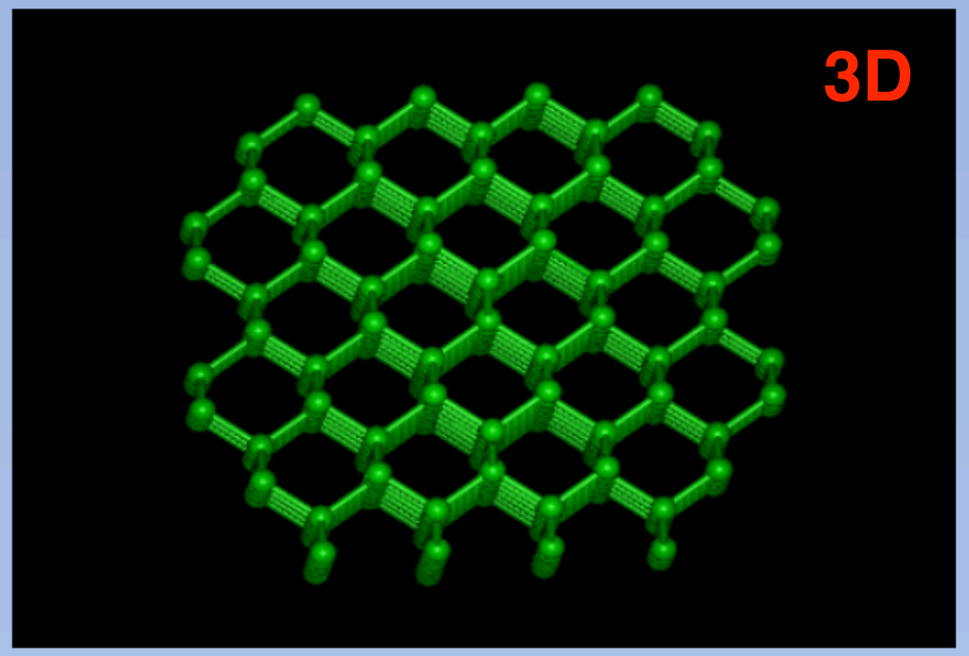
Two polymorphic structures of carbon



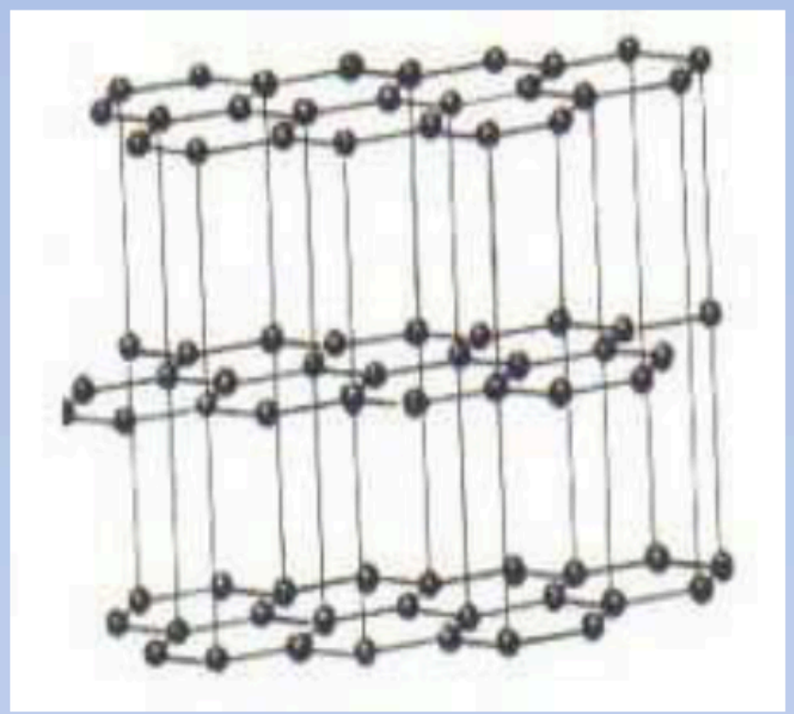
Prof.ssa Annabella De Vito
ITG "Rondani" Parma

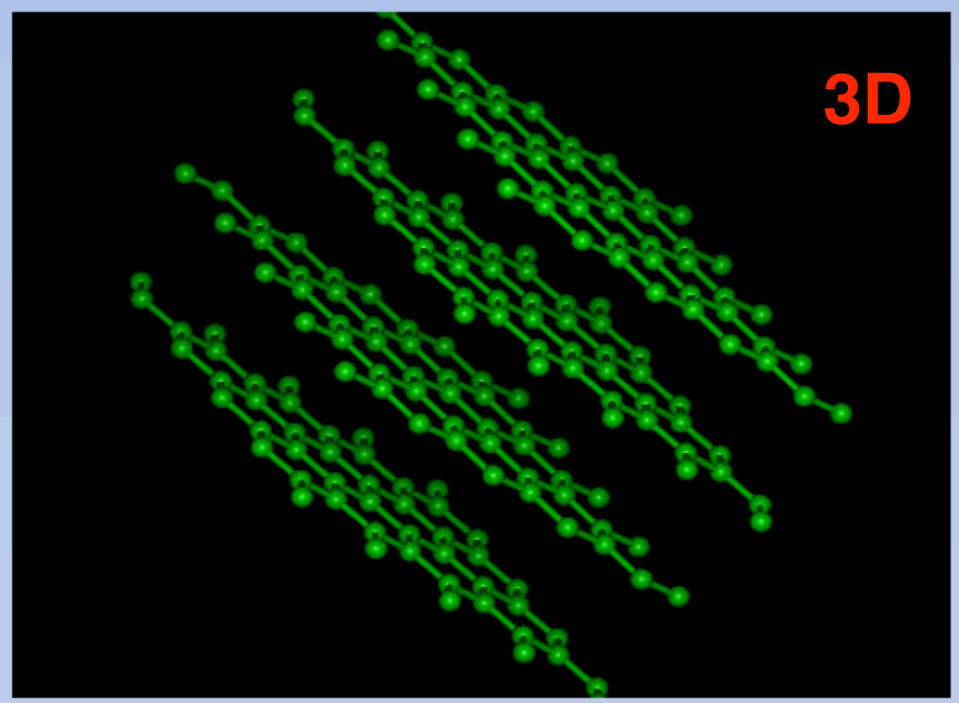
Crystal abitus - DIAMOND





Crystal abitus - GRAPHYTE





Graphite is composed of layers of carbon atoms arranged in **hexagonal structures**.

Diamond is composed of layers of carbon atoms arranged in the **cubic structures**.

The **different crystal structures** of diamond and graphite are the unique determinants of the huge differences between the properties of these two materials.

The two different polymorphic structures have the **same chemical** behaviour (ignition), BUT **totally different physical properties** (*cleavage, hardness, transparency, melting point*). These physical properties only derives (emerge) from atoms organisation.

They are ***EMERGENT PROPERTIES of the system!***

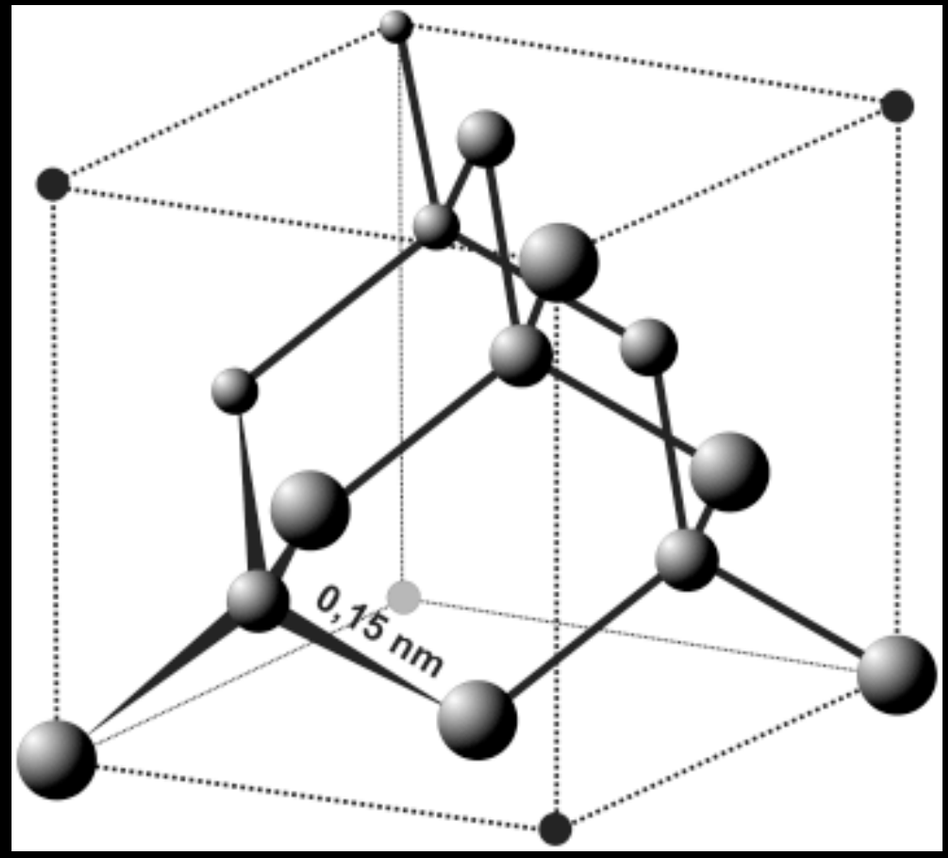
The diamond does show emergent properties

hardness

melting point

transparency

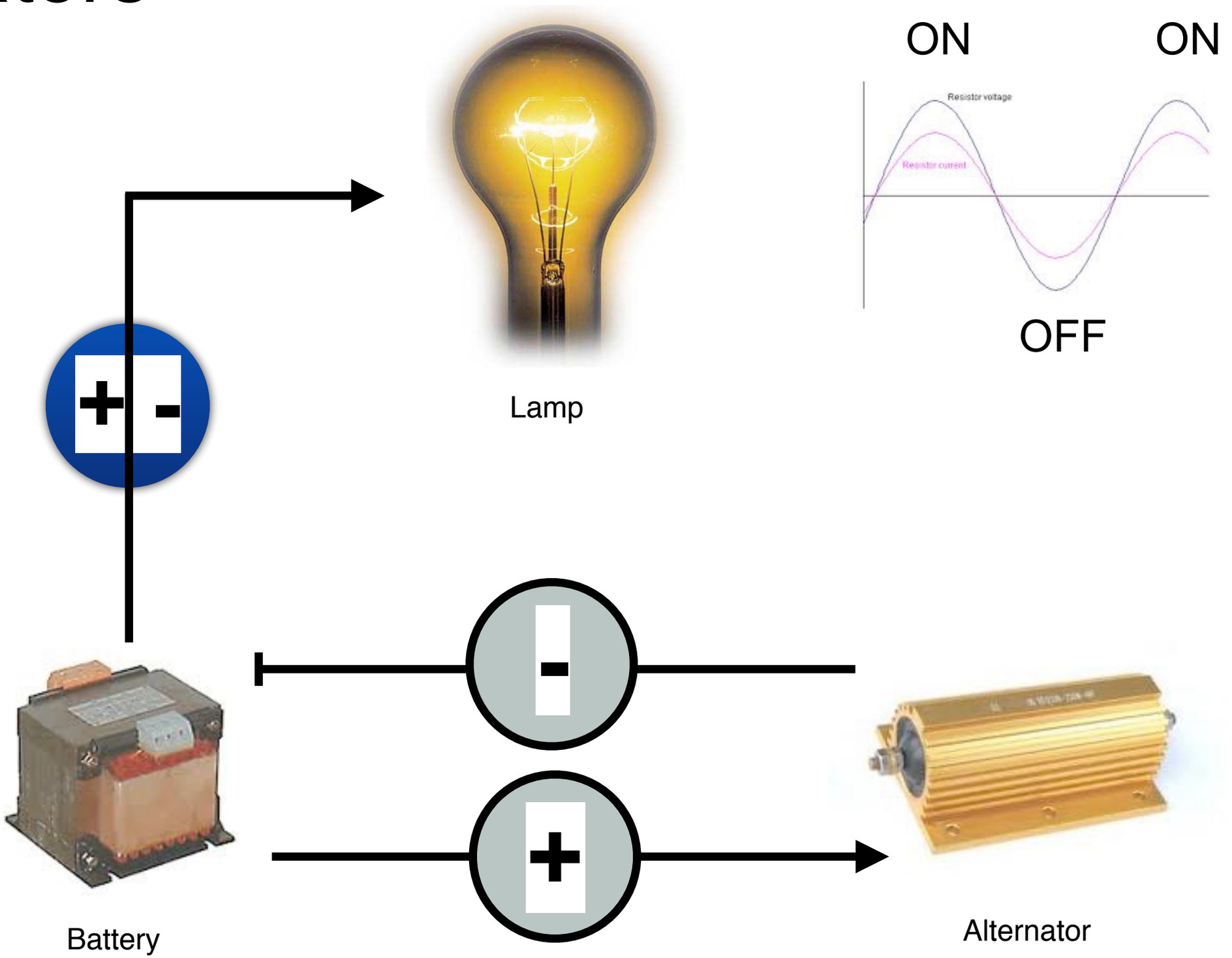
hardness



Crystal system:

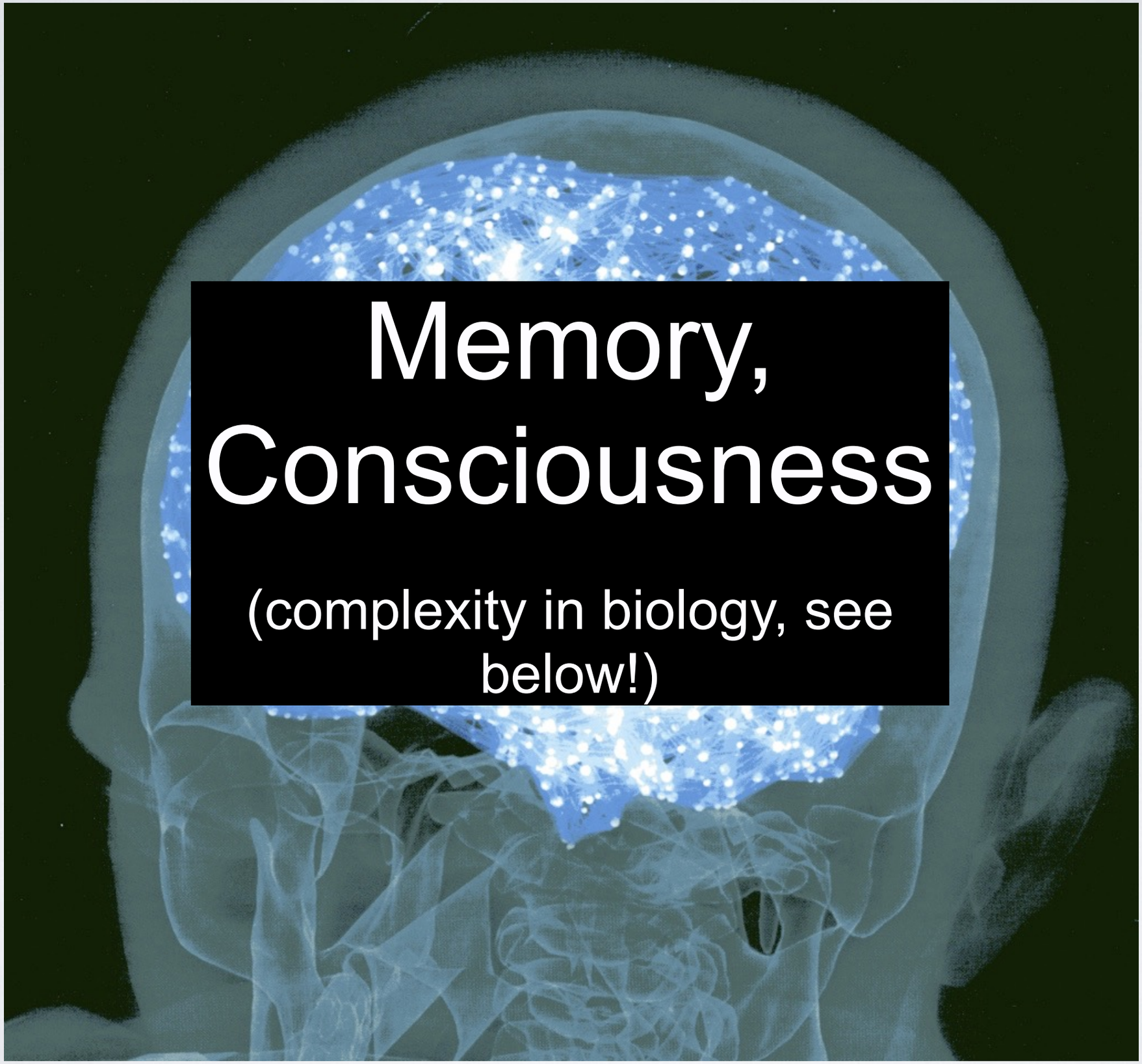
Cubic

Oscillators

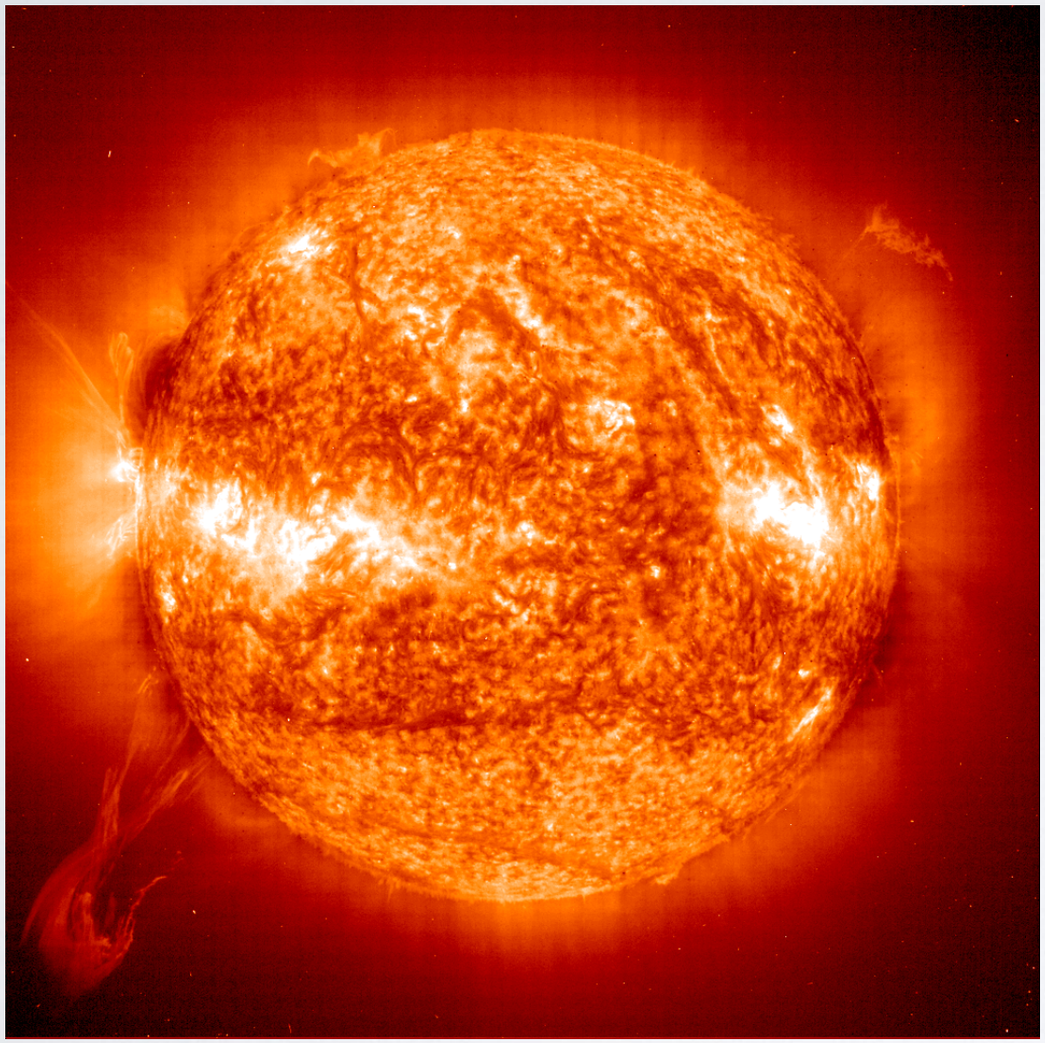


Frequency, amplitude and wavelength are **emergent properties**

Brain = 10^{11} neurons; 10^{14} connexions !



(1)
Complexity in the living world: origin and nature of complexity in biology



Diameter ~ $1{,}4 \times 10^6$ km

Length ~ 10 cm

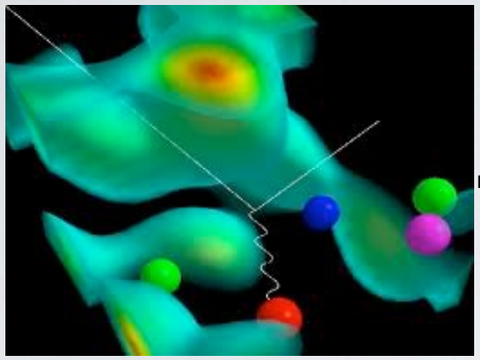
$14 \times 10^9 <$

"A toad is much more complex than a star"

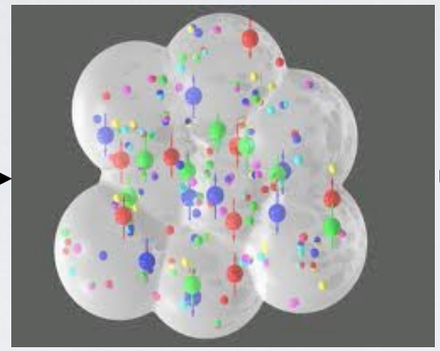
Martin Rees
Astrophysicist

Complexity ≠ Dimension

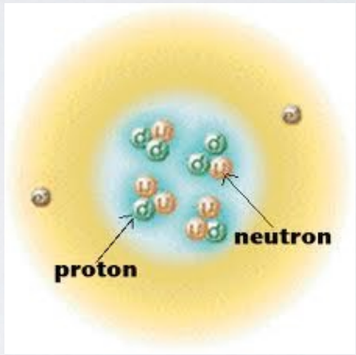
Origin of complexity



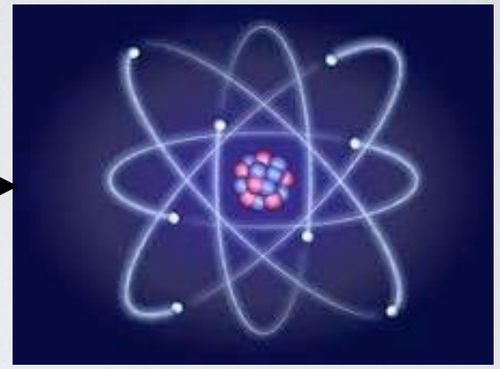
quarks (6)



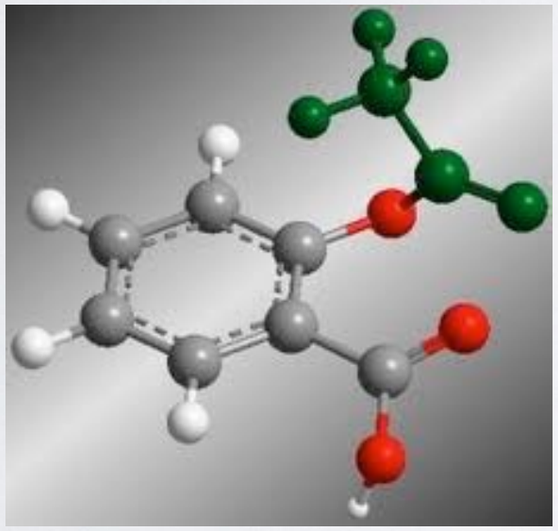
protons-neutrons



atomic nuclei

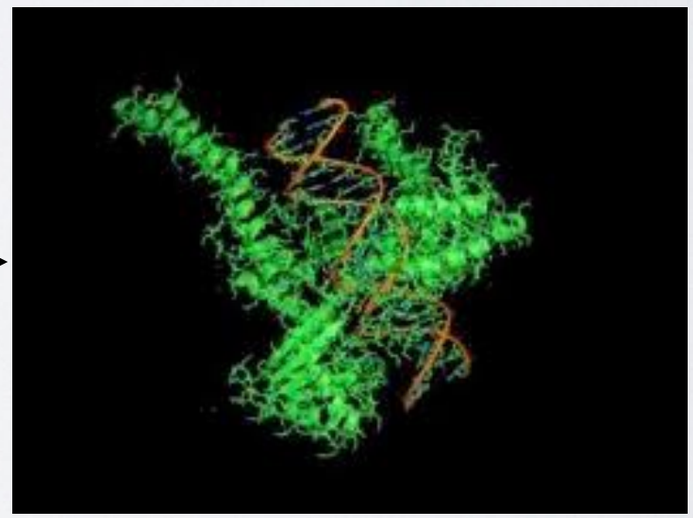


atoms

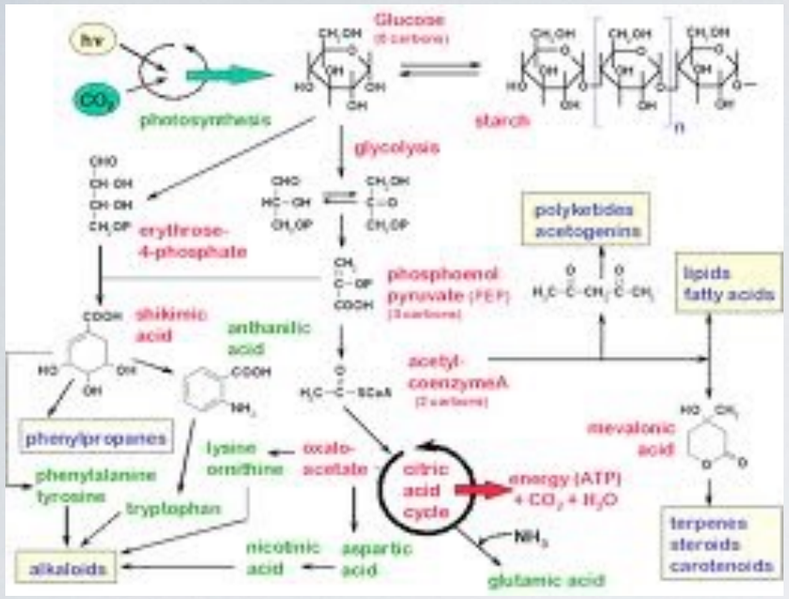


basic
organic molecules

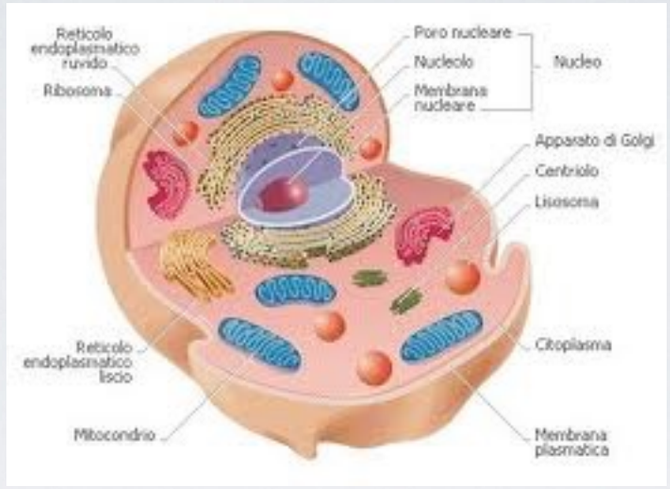
DNA



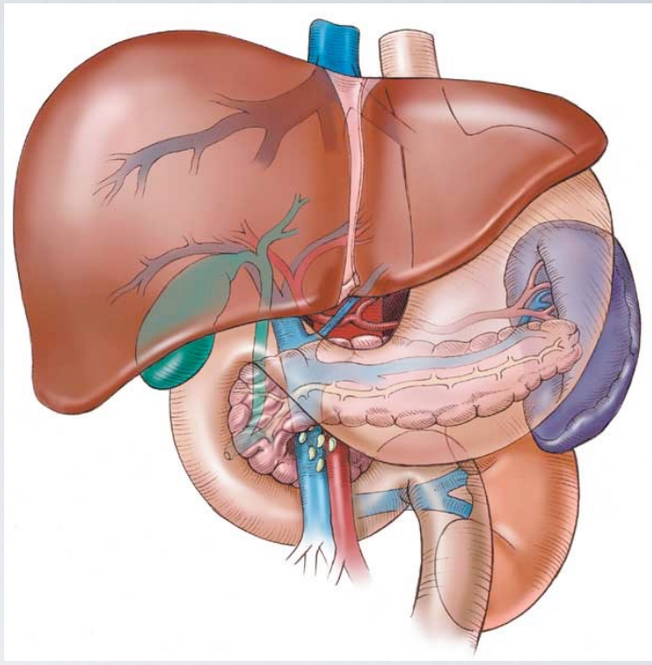
proteins



Molecular circuits



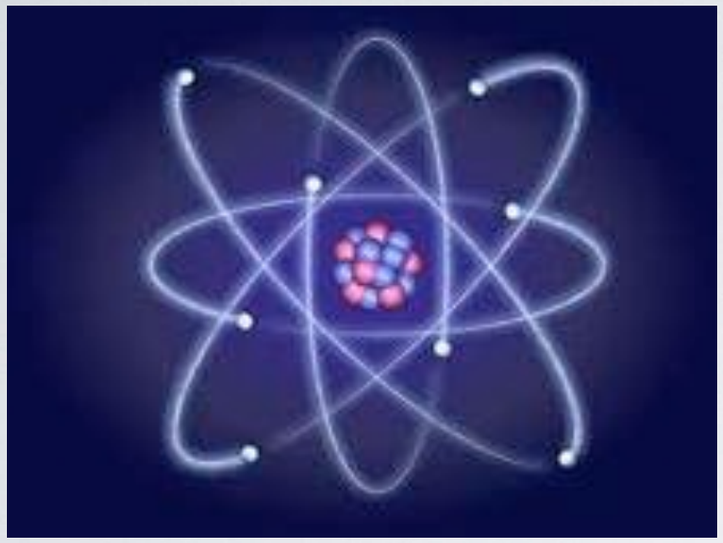
Cells



Organs

Living being

Carbon chemistry = LIFE! BUT …. not just carbon!

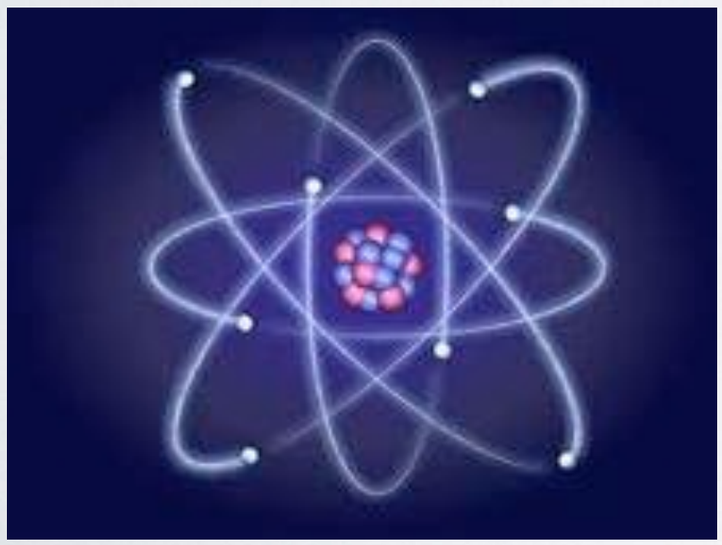


10^n carbon atoms

Simple organization

Diamond

?



10^n carbon atoms

Complex organization

Living being

Biology as a supreme manifestation of <u>complex organisation</u>

3 sub-atomic particles
(protons, neutrons, electrons)

↓

4 elements (atoms) fundamental to biology
(C, H, O, N)
(P, Fe, Ca, S, Se, Cu, Zn = oligo elements)

↓

4 molecules fundamental to biology
(lipids, carbohydrates, nucleic acids, proteins)

↓

Lipids	Carbohydrates	Nucleic acids	Proteins
5 main types	3 main types	2 main types	1 main type but more than 10^5 different sequences
Cholesterol Triglycerides Fatty acids Phospholipids Glycolipids	Monosaccharides Disaccharides Polysaccharides	DNA RNA	

DNA

(homo sapiens)

A-C-G-T

(3x10^9 bases)

64 codons

(10^9 potential)

~23000 coding genes

Proteine

20 (23?) Amino acids

~100000 proteins

6 quarks

Combinations → 2 adrons (protons, neutrons)

Combinations → 1 lepton (elettrone)

Combinations → 4 atoms (C, H, O, N)

Combinations → 4 bases (A, C, G, T)

Combinations

Combinations

Combinations

Combinatorial rules

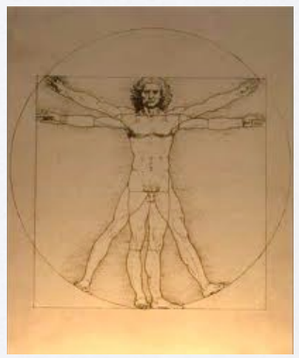
100000 proteins

20 amino acids

~ 23000 genes

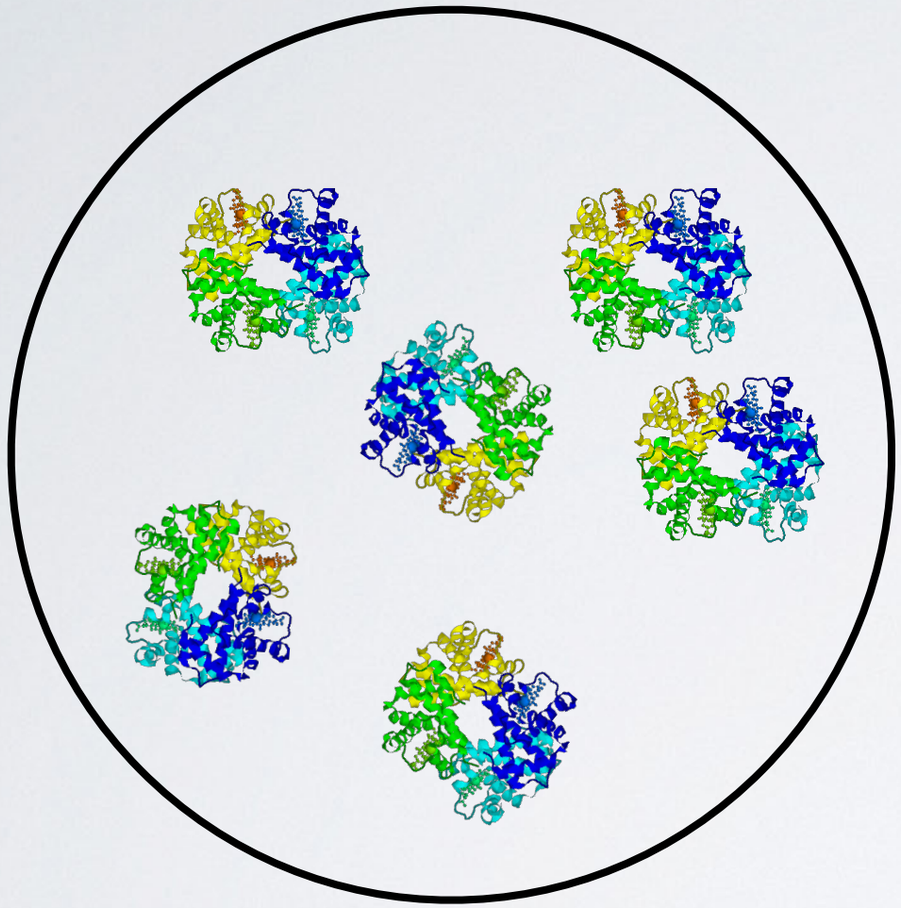
Is this enough ?

?



10^5 proteins sequences

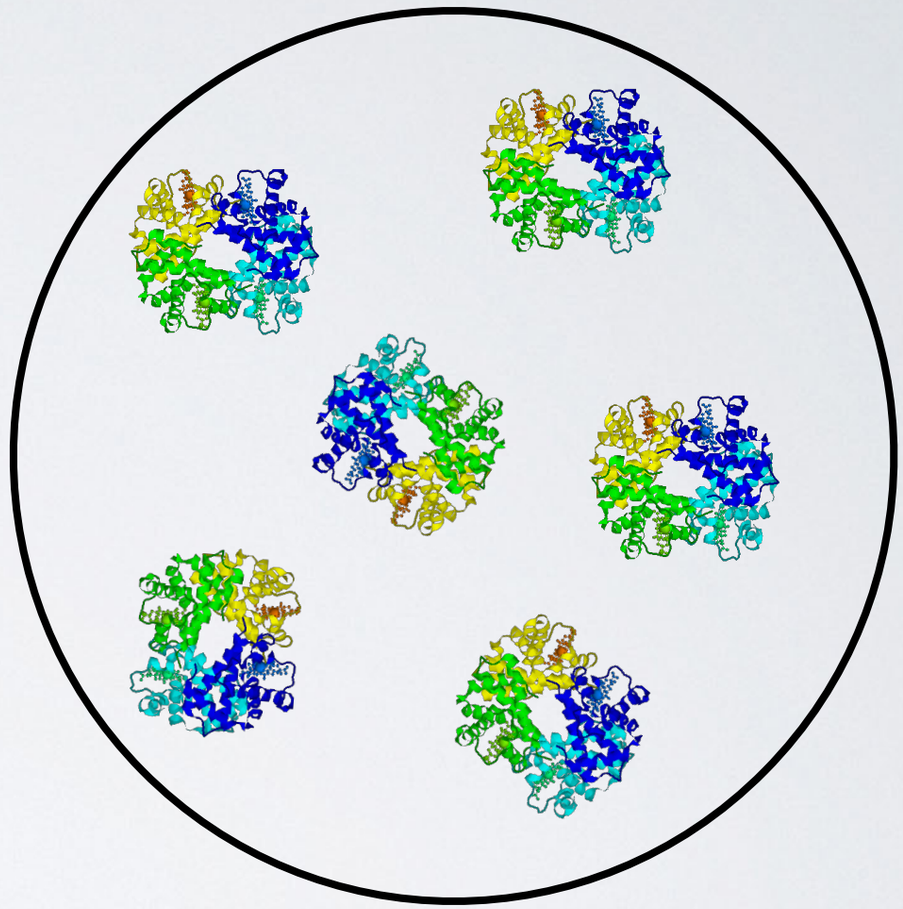
NO
Combinatorial interaction



NO

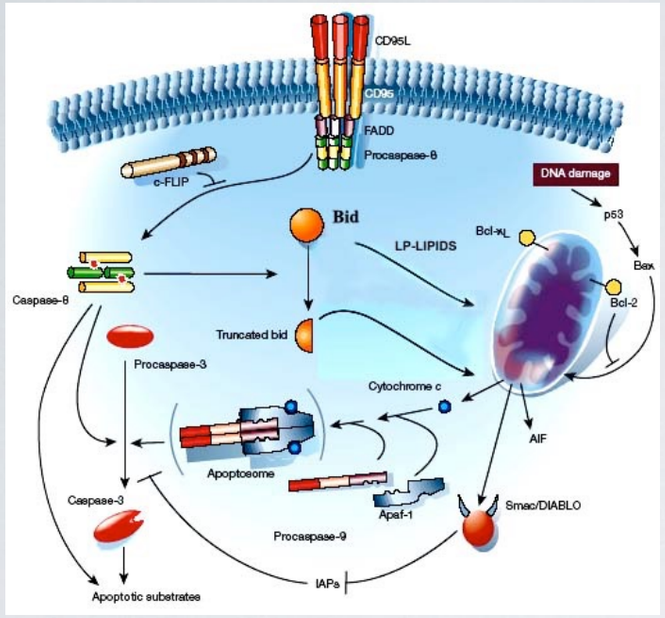
Homo sapiens sapiens

YES
Further level of
combinatorial interaction

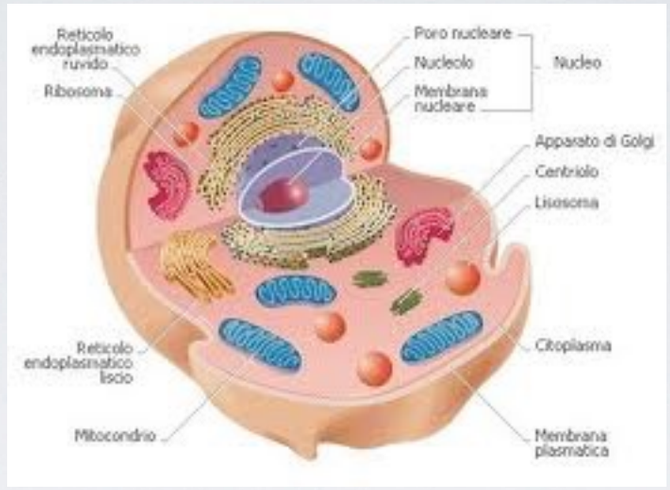


YES

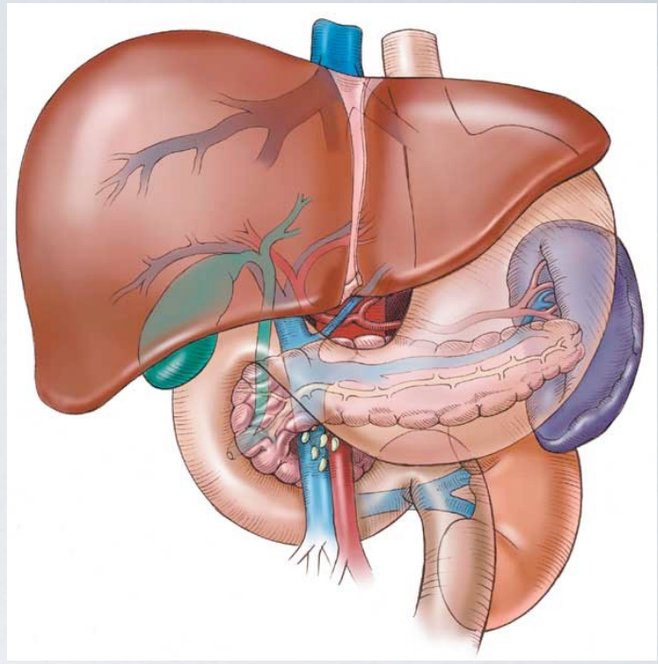
Homo Sapiens sapiens



Molecular circuits

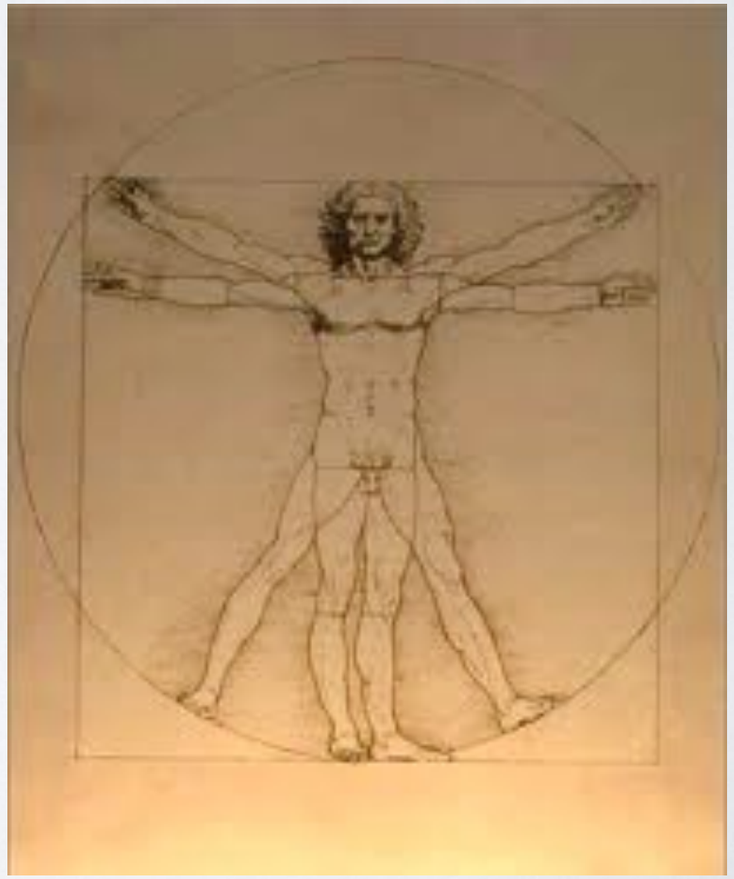


Cells



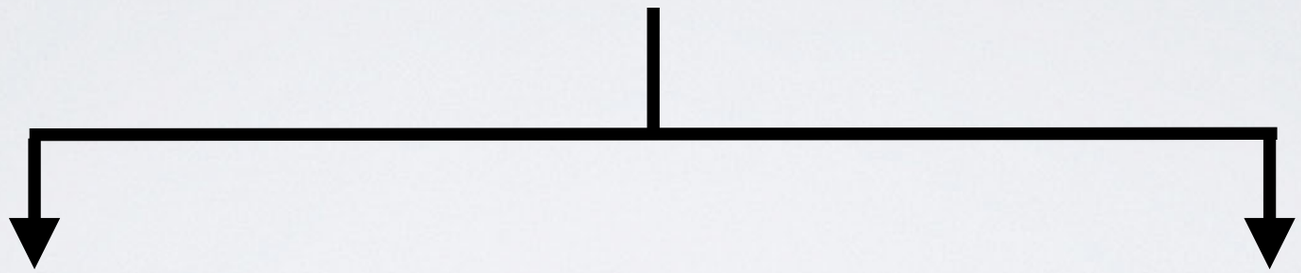
Organs

In biology:
Everything is interaction!
Everything is combination!
Everything is a process!
(may be not only in biology)

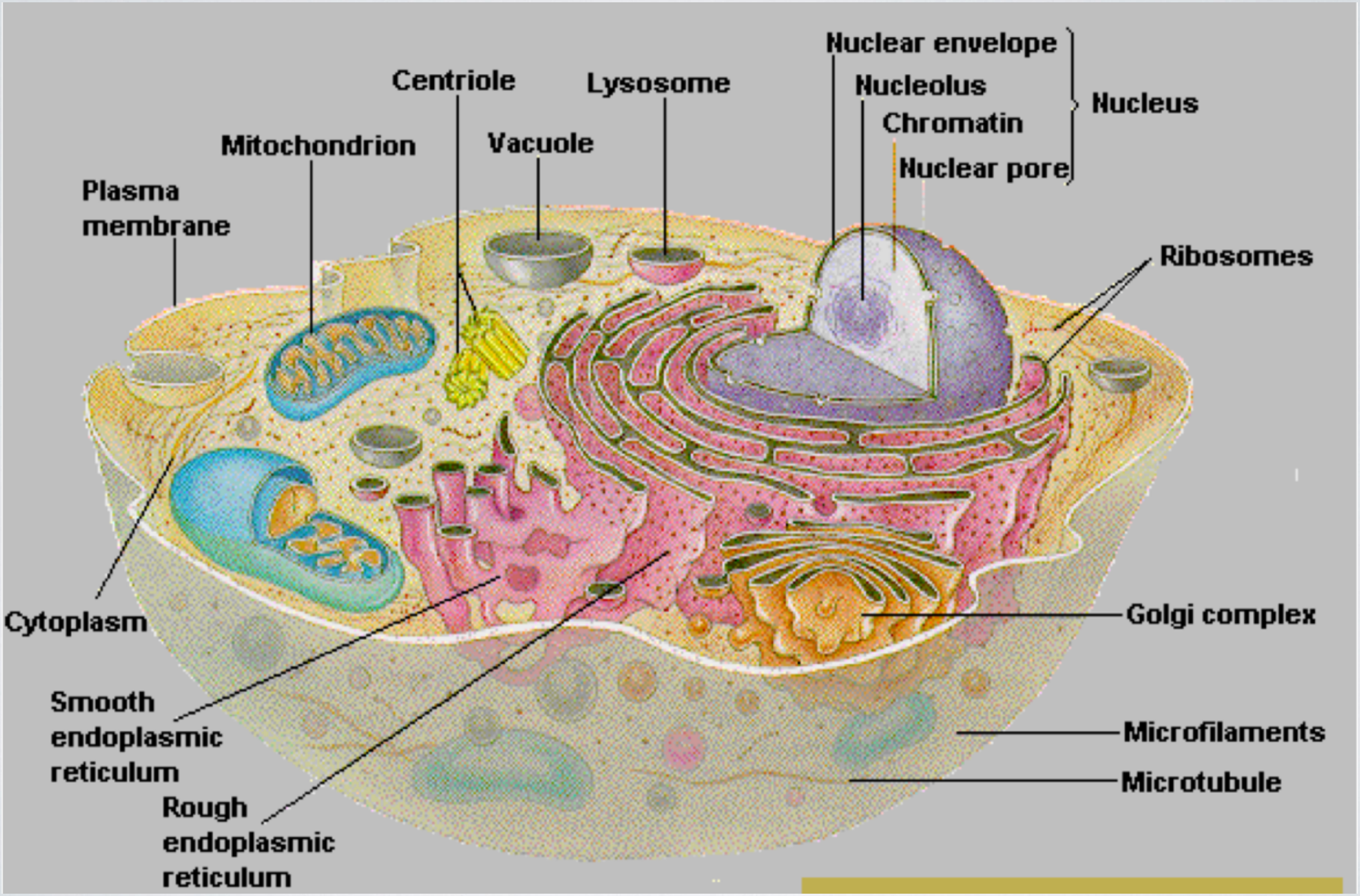


The “**magic word**” in biology:
combinatoriality’

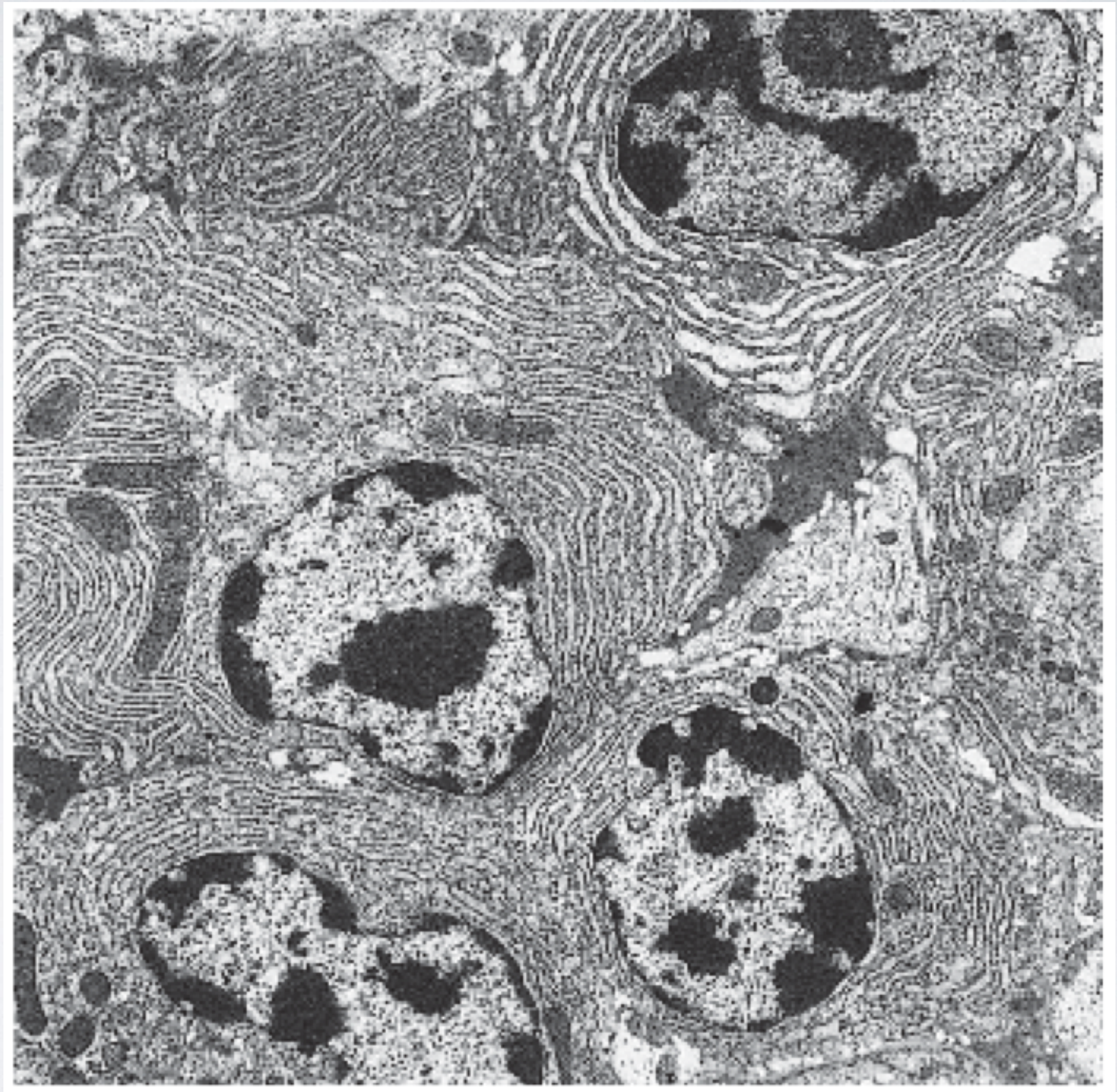
2 types of combinatoriality’



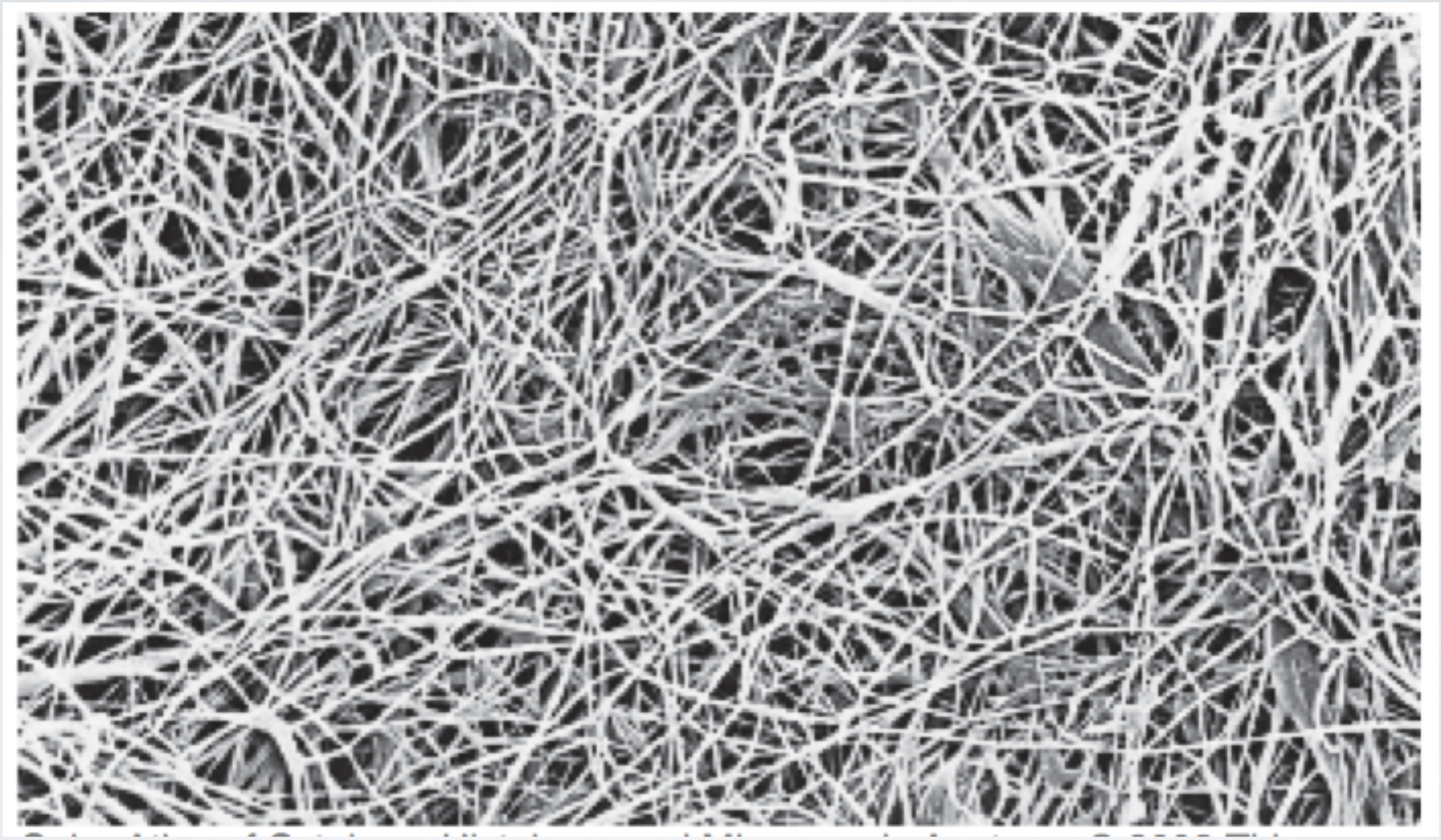
Structural combinatoriality



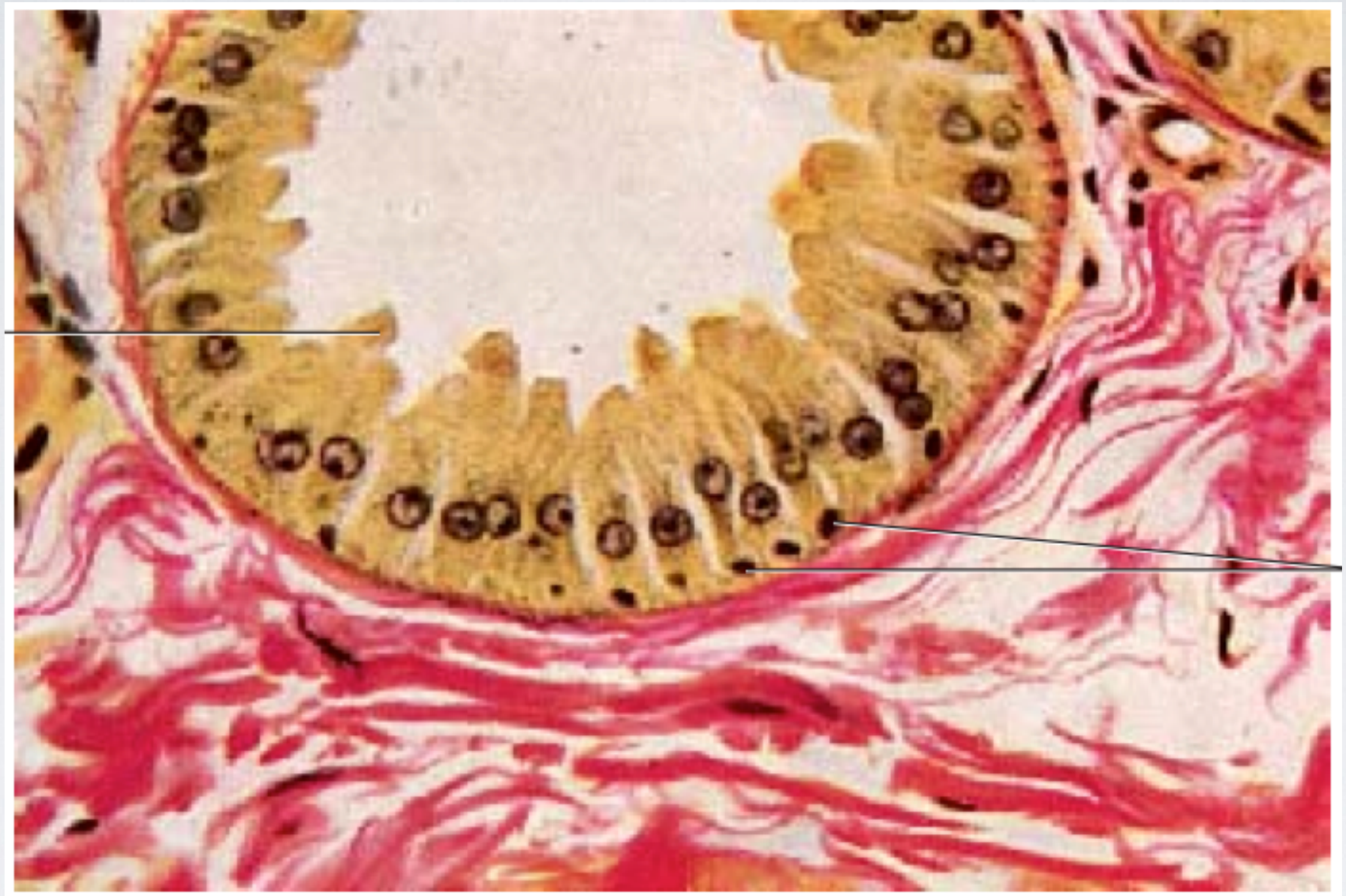
Sub-cellular organuli



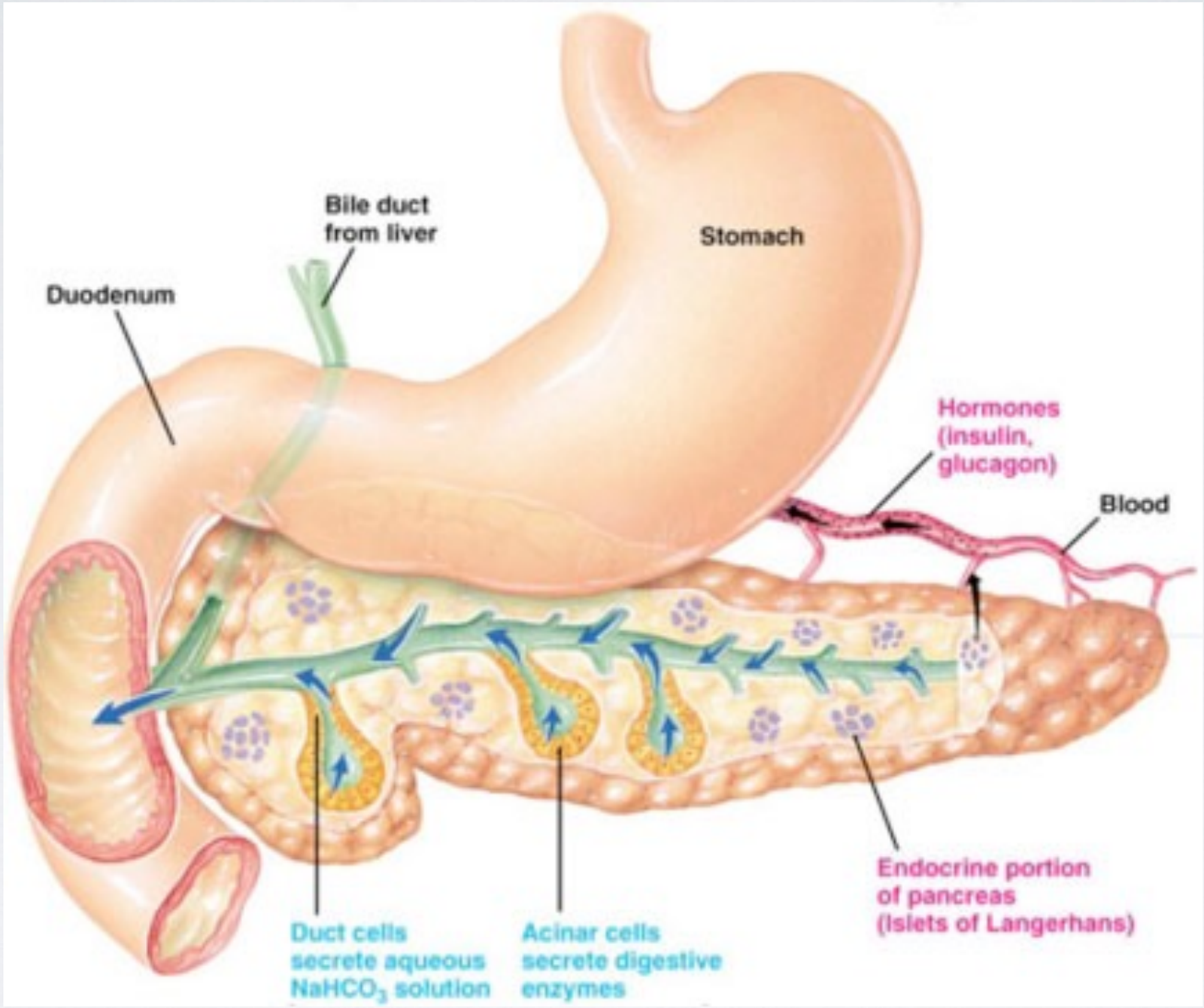
Collage fibers (connective tissue)



Exocrine glandular acini



Pancreas


Bile duct
from liver
Stomach
Duodenum
Hormones
(insulin,
glucagon)
Blood
Endocrine portion
of pancreas
(Islets of Langerhans)
Duct cells
secrete aqueous
$NaHCO_3$ solution
Acinar cells
secrete digestive
enzymes

Circuital combinatoriality (dynamics)

Elements (cells, molecules) with different function reciprocally interact in specific and changeable way over time and / or space (dynamics) to provide appropriate feed-back to environmental changes (Processing of matter-energy and/or information)

Cellular circuits:

- neurons
- leukocytes

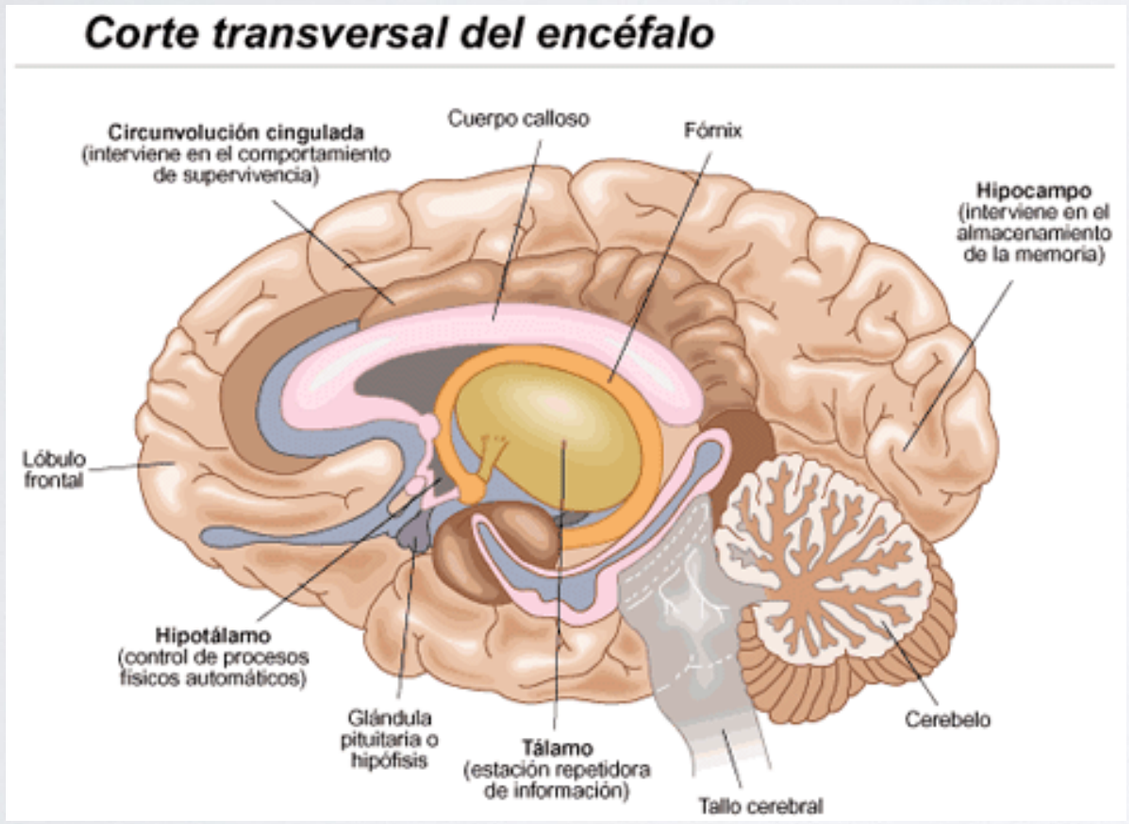
Molecular circuits:

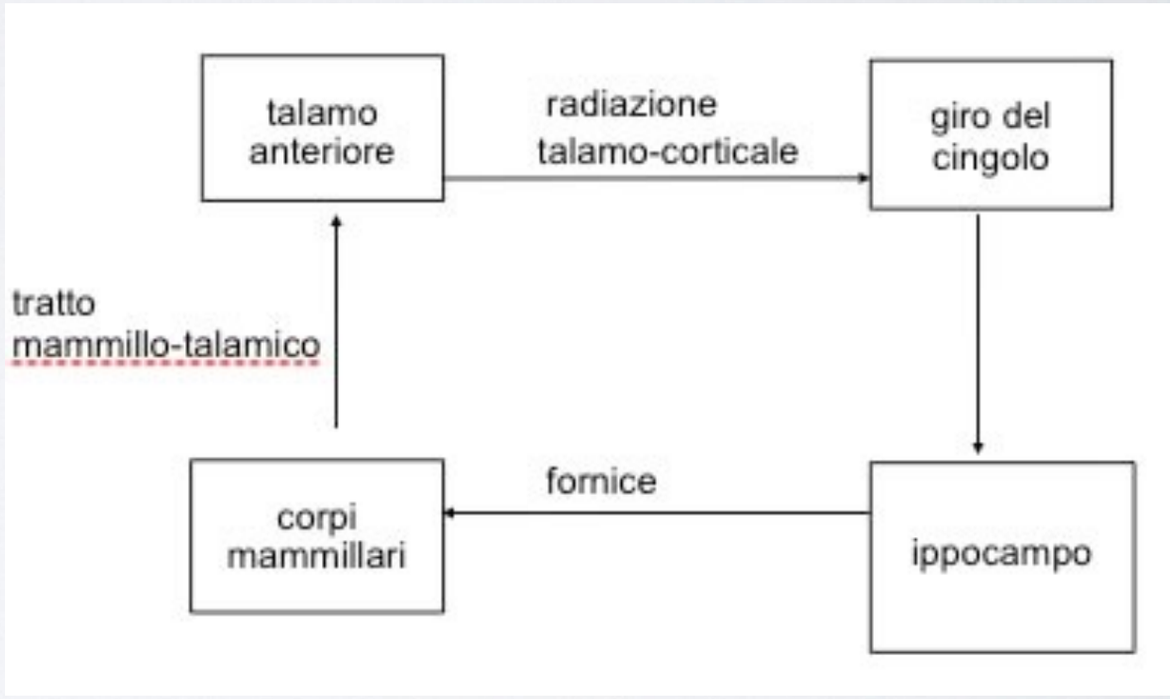
- hormones
- metabolism
- signal transduction

Basically, biology is a “net of nets” (circuits) generating “emergent properties”.
Modelling, discovering and understanding them is the main goal of Systems Biology.

(2)
Emergent properties in biological systems: cellular and molecular circuits

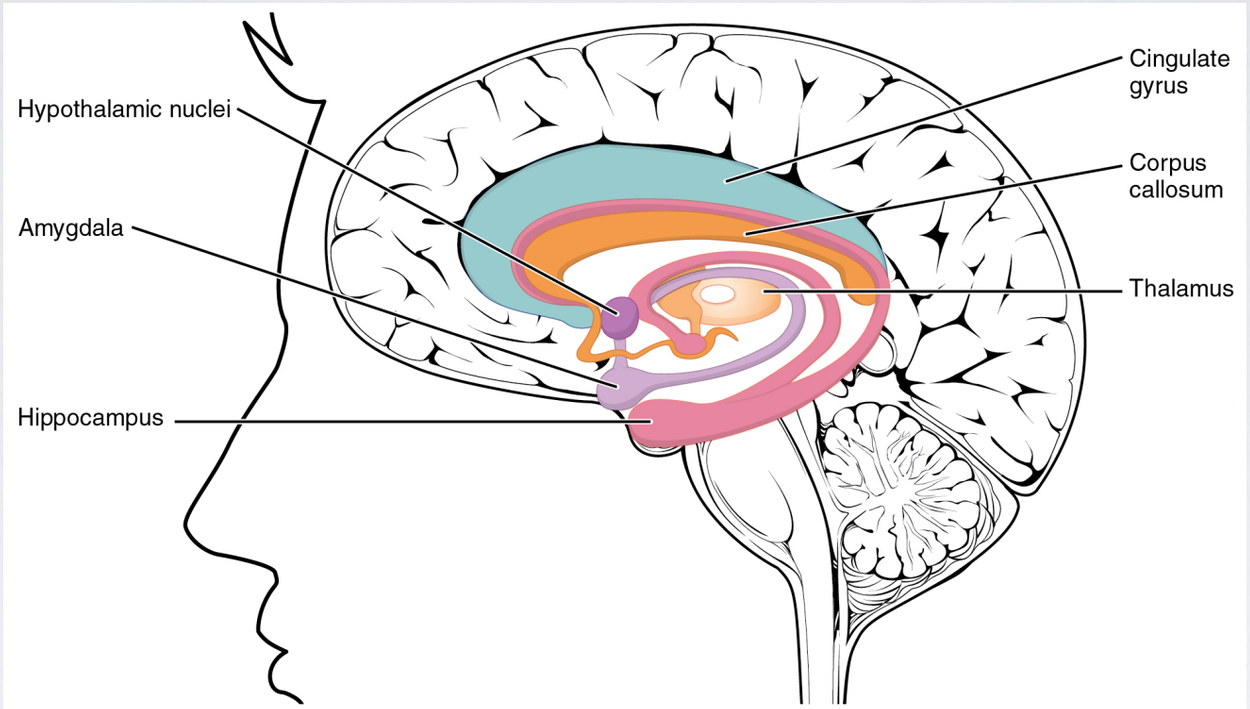
Papez circuit: bi-stability and the emergence of memory





Reverberating cellular circuit

The limbic system



The **limbic system** is a set of brain structures located on both sides of the thalamus, immediately beneath the cerebrum.

The limbic system supports a variety of functions including emotion, behavior, motivation, long-term memory, and olfaction.

It is not a separate system but is a *collection of interconnected structures, organised in a peculiar architecture.*

Cortical areas:

- Limbic lobe
- Orbitofrontal cortex, a region in the frontal lobe involved in the process of decision-making.
- Piriform cortex, part of the olfactory system.
- Entorhinal cortex, related with memory and associative components.
- Hippocampus and associated structures, which play a central role in the consolidation of new memories.
- Fornix, a white matter structure connecting the hippocampus with other brain structures, particularly the mammillary bodies and septal nuclei

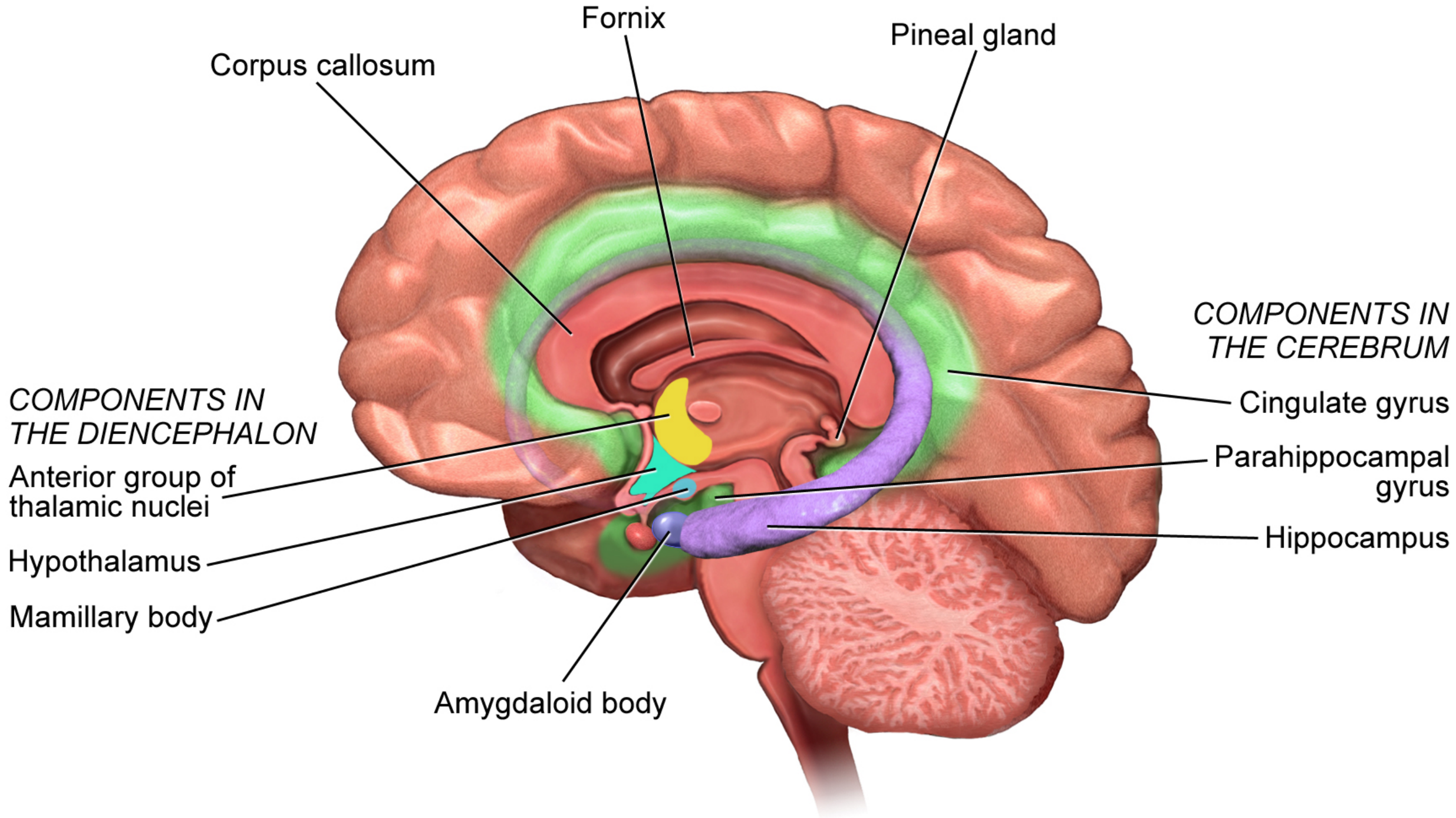
Subcortical areas:

- Septal nuclei, a set of structures that lie in front of the lamina terminalis, considered a pleasure zone.
- Amygdala, located deep within the temporal lobes and related with a number of emotional processes.
- Nucleus accumbens: involved in reward, pleasure, and addiction.

Diencephalic structures:

- Hypothalamus: a center for the limbic system, connected with the frontal lobes, septal nuclei and the brain stem reticular formation via the medial forebrain bundle, with the hippocampus via the fornix, and with the thalamus via the mammillothalamic fasciculus. It regulates a great number of autonomic processes.
- Mammillary bodies, part of the hypothalamus that receives signals from the hippocampus via the fornix and projects them to the thalamus.
- Anterior nuclei of thalamus receive input from the mammillary bodies. Involved in memory processing.

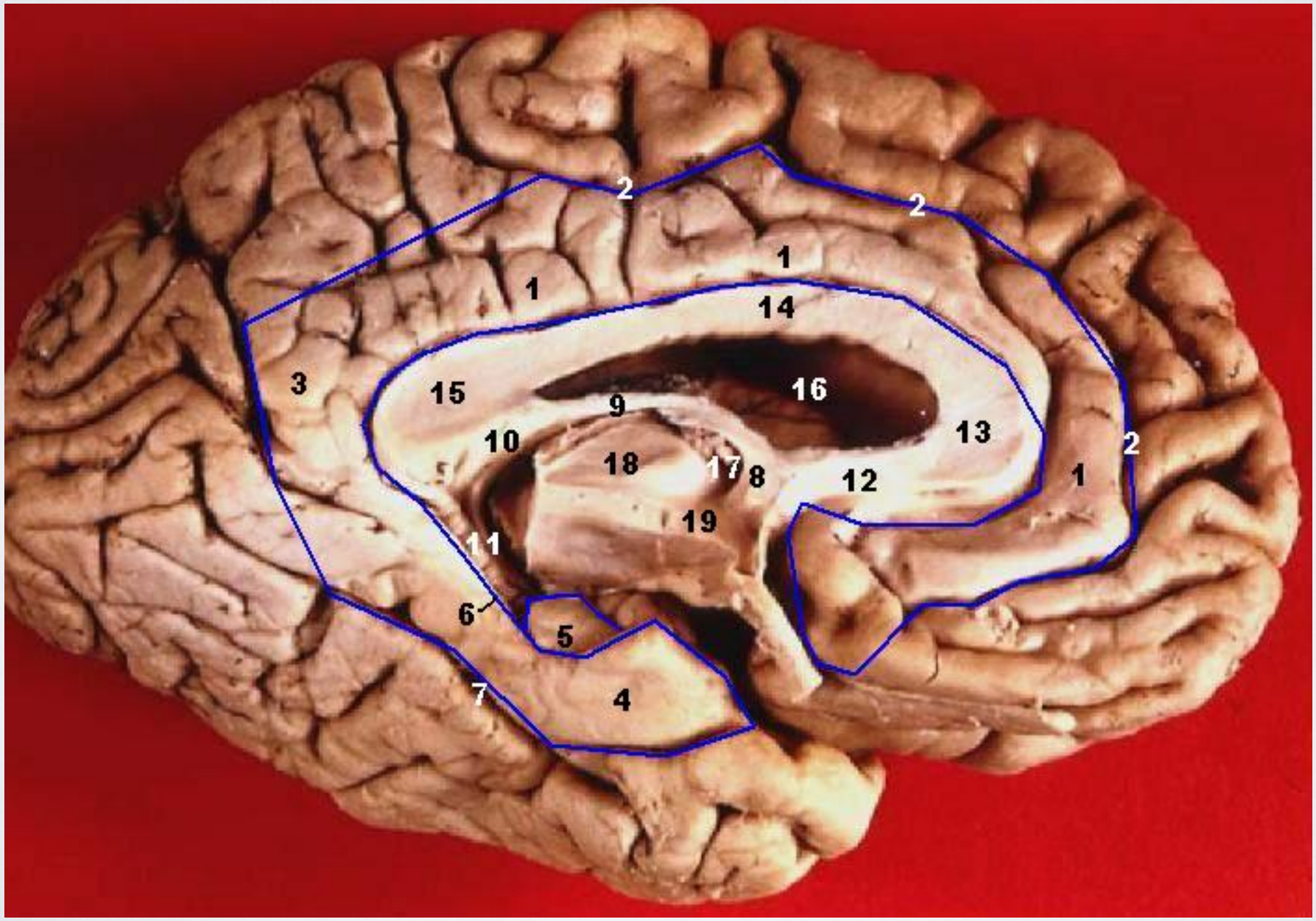
The Limbic System

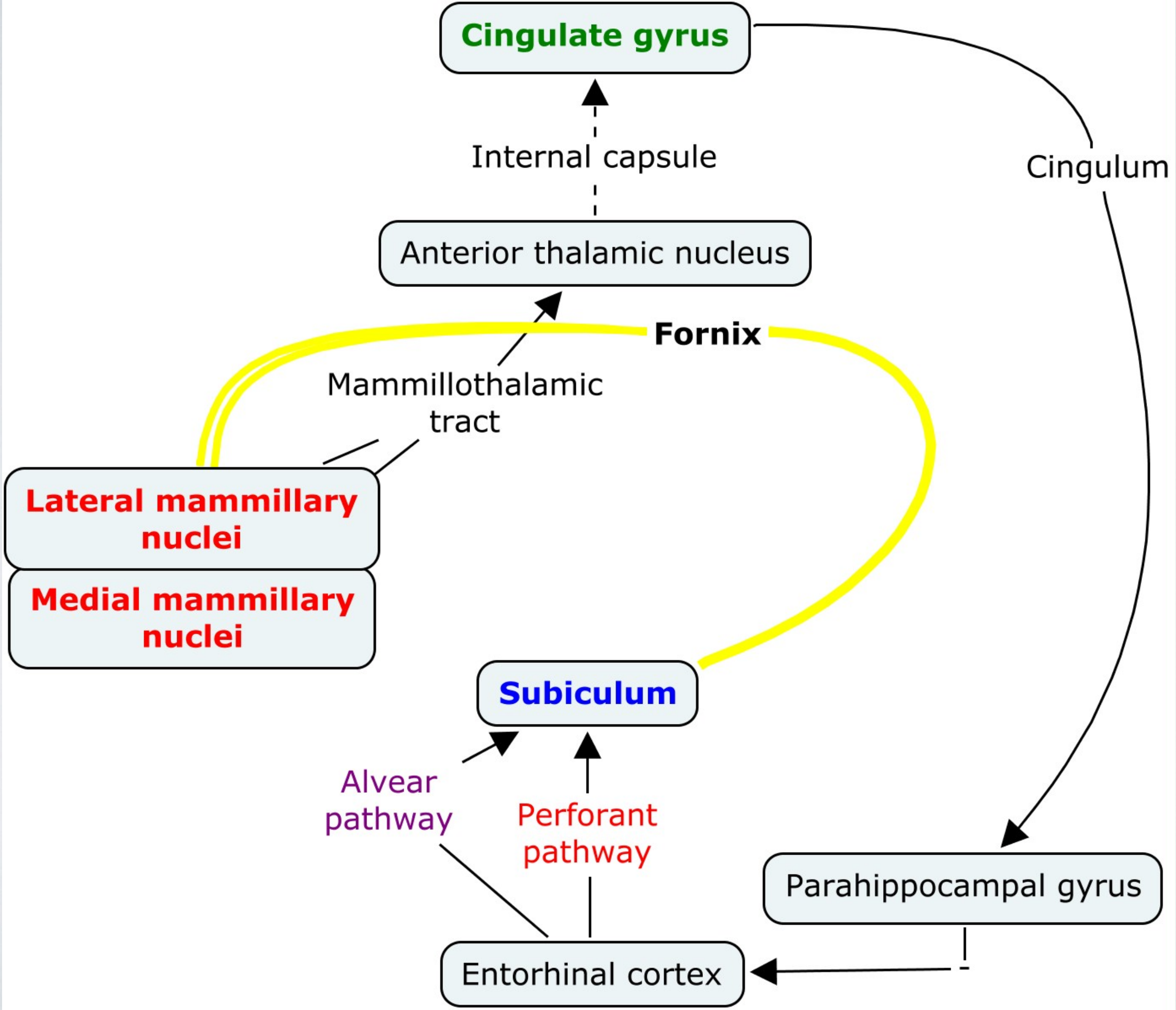


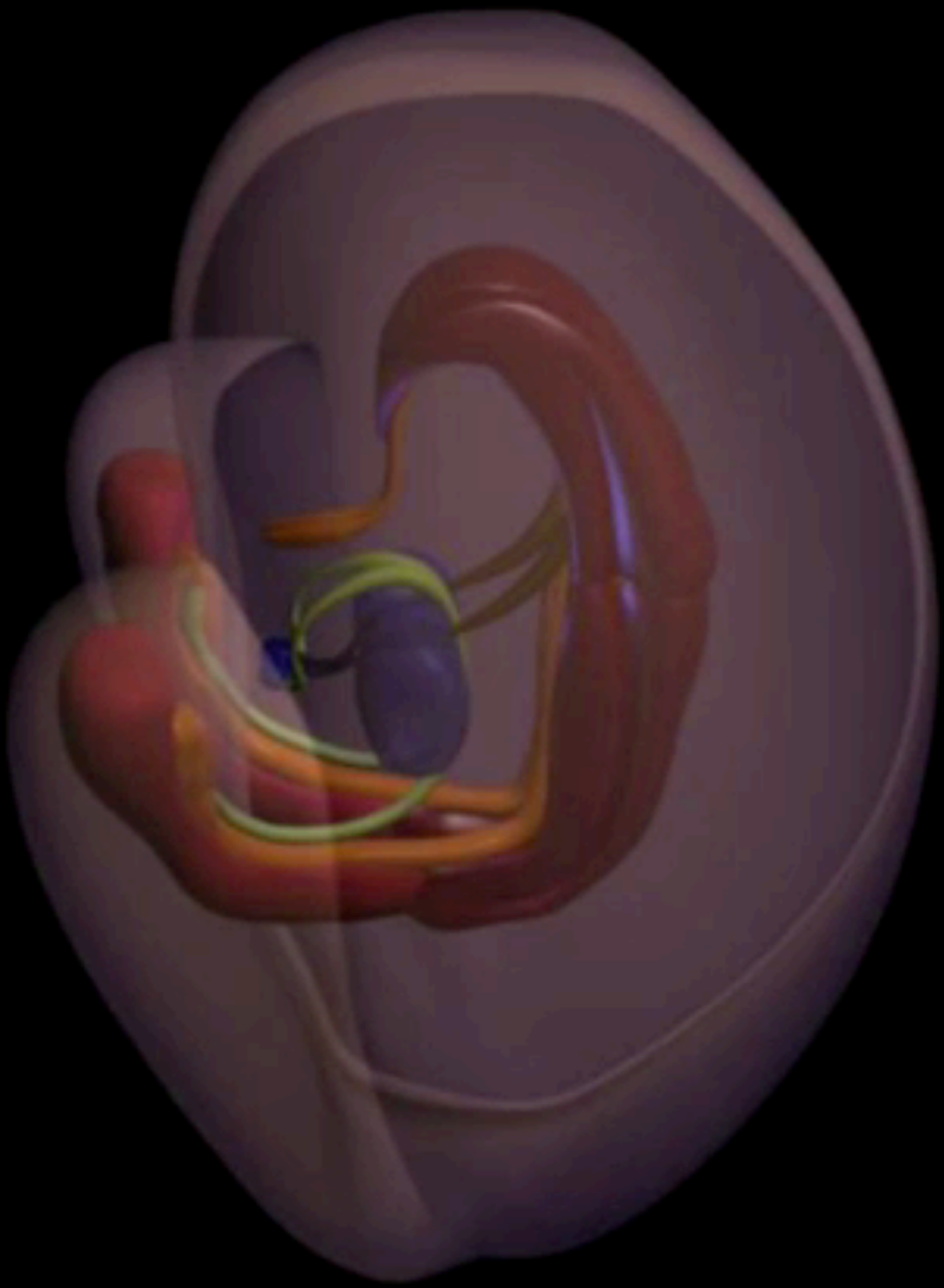
Central nervous system memorisation:
The Papez circuit

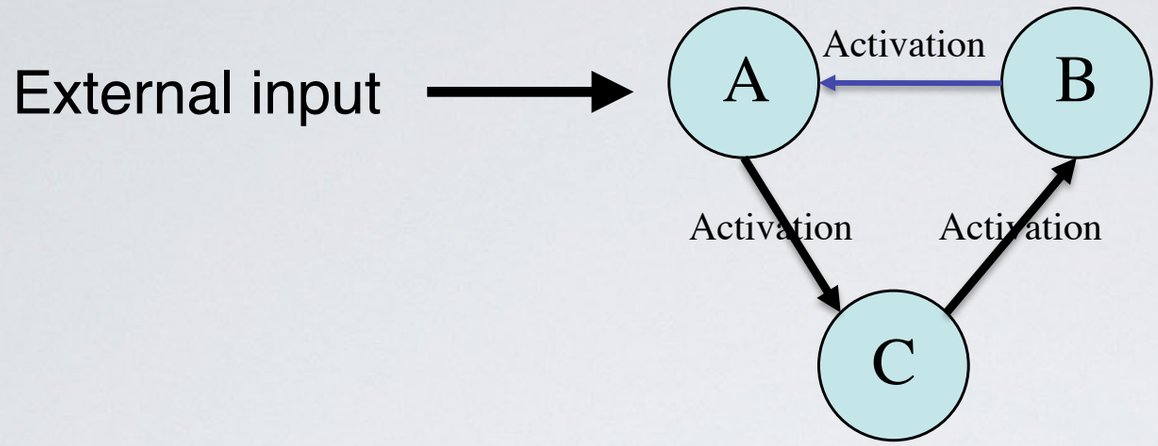
The **Papez circuit** (or **medial limbic circuit**) is a neural circuit for the control of emotional expression. It a part of limbic system.

It has a particularly relevant role in **memory functions.**

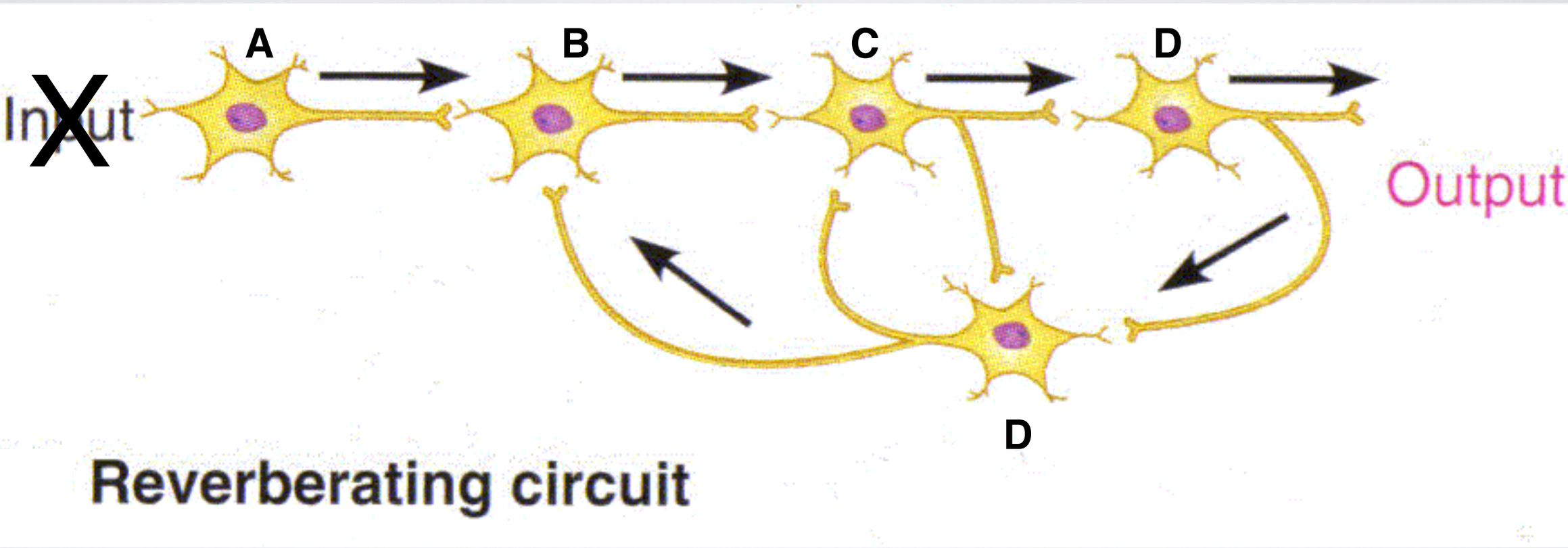



Cingulate gyrus
Internal capsule
Cingulum
Anterior thalamic nucleus
Fornix
Mammillothalamic tract
Lateral mammillary nuclei
Medial mammillary nuclei
Subiculum
Alvear pathway
Perforant pathway
Parahippocampal gyrus
Entorhinal cortex





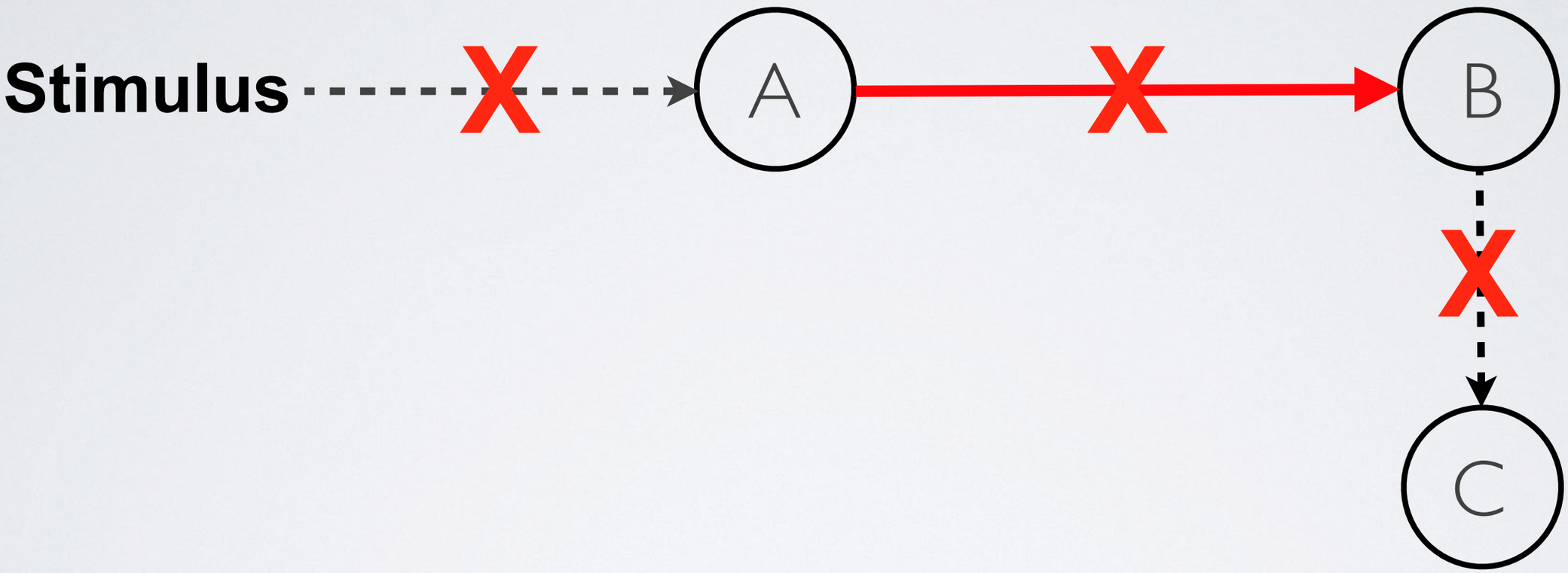
Information recirculates
even if the external
input is OFF



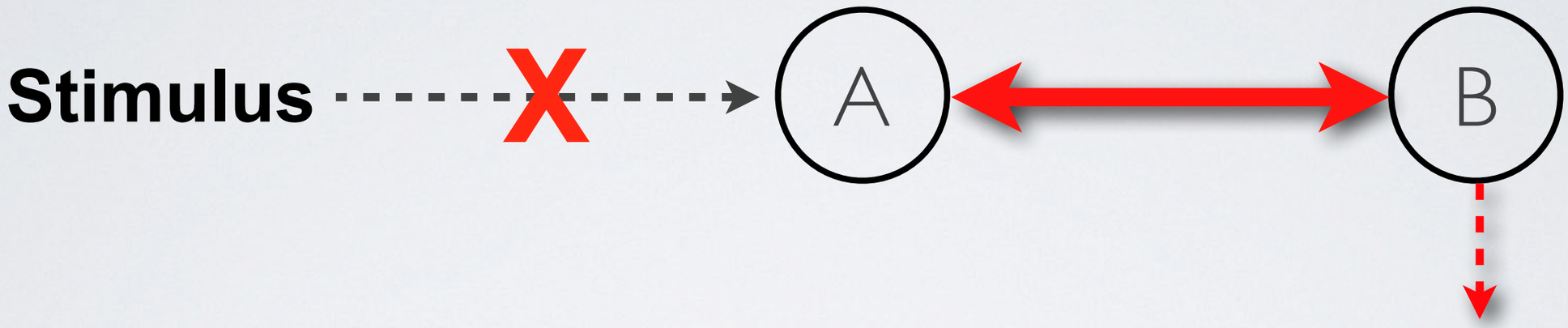
The Papez circuit is a reverberating circuit devoted to signal memorisation and information storage.

Memory is a **property** of CNS architecture

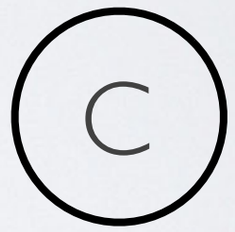
A non-reverberating circuits does not trigger memorisation



A reverberating circuit generates bi-stability and triggers the emergence of **memorisation**



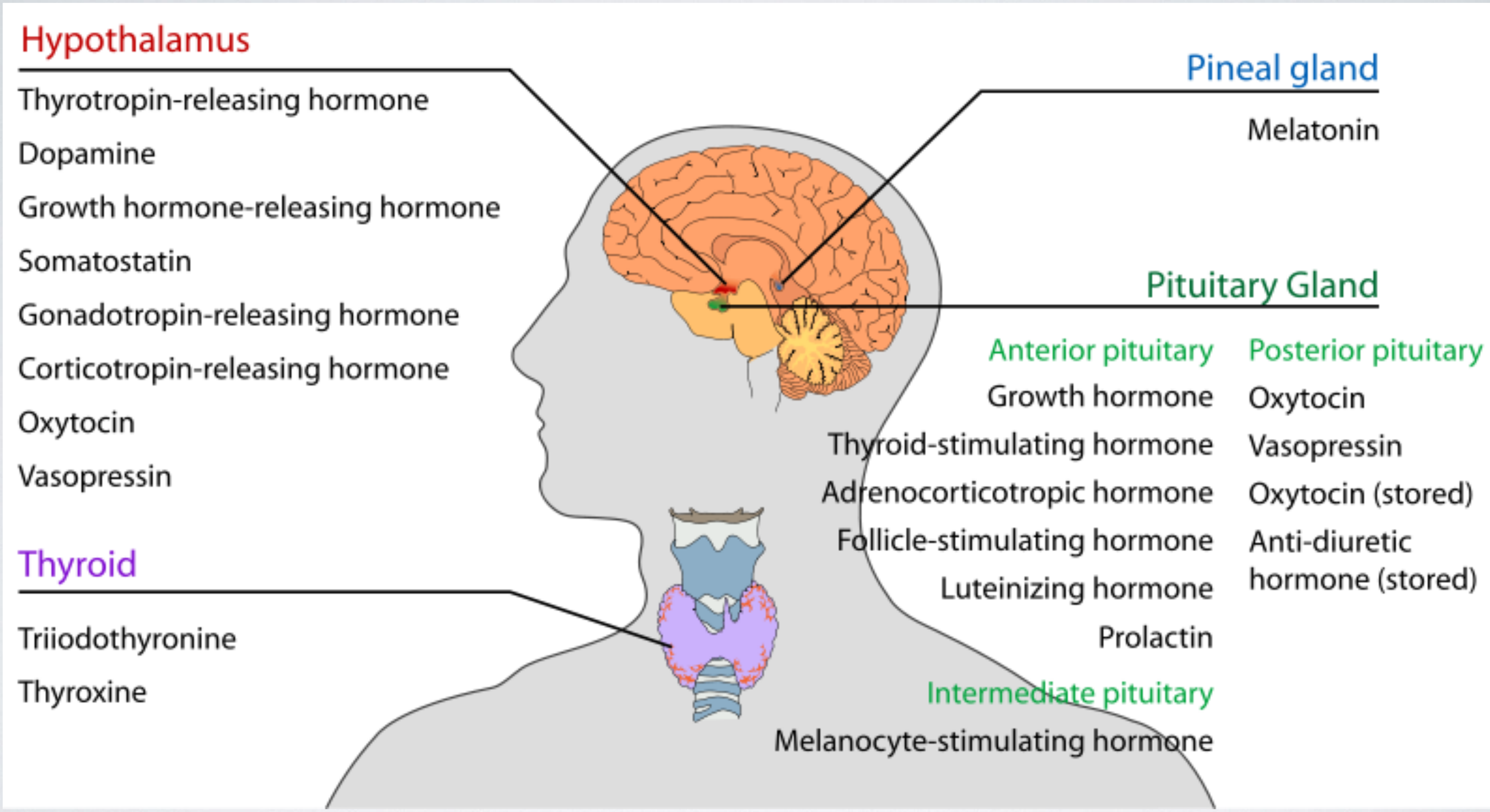
The system records the stimulus



Memory is an emergent property!

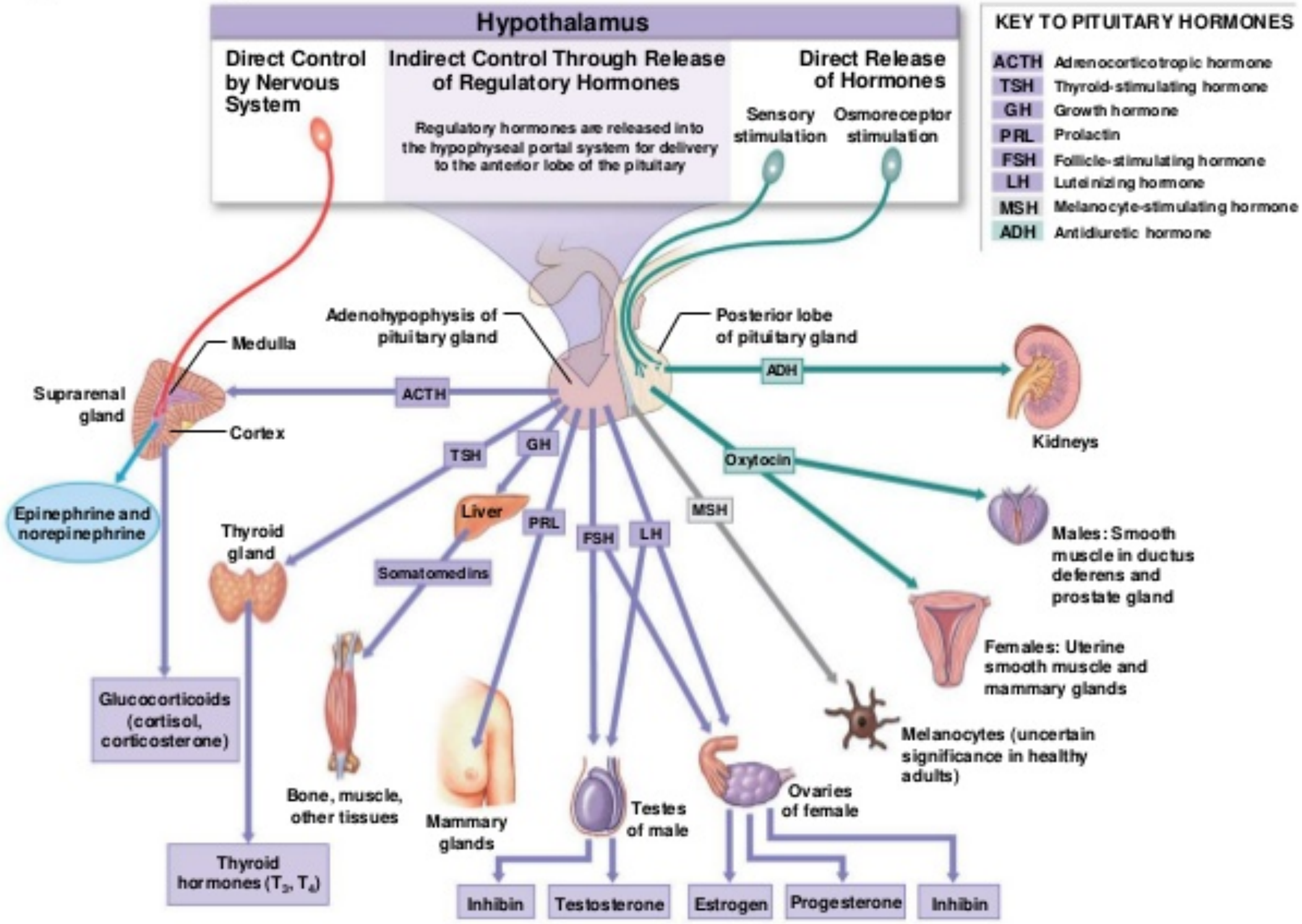
Feedback mechanisms and the oscillators

The pituitary and the hormonal oscillators



The pituitary and the hormonal oscillators

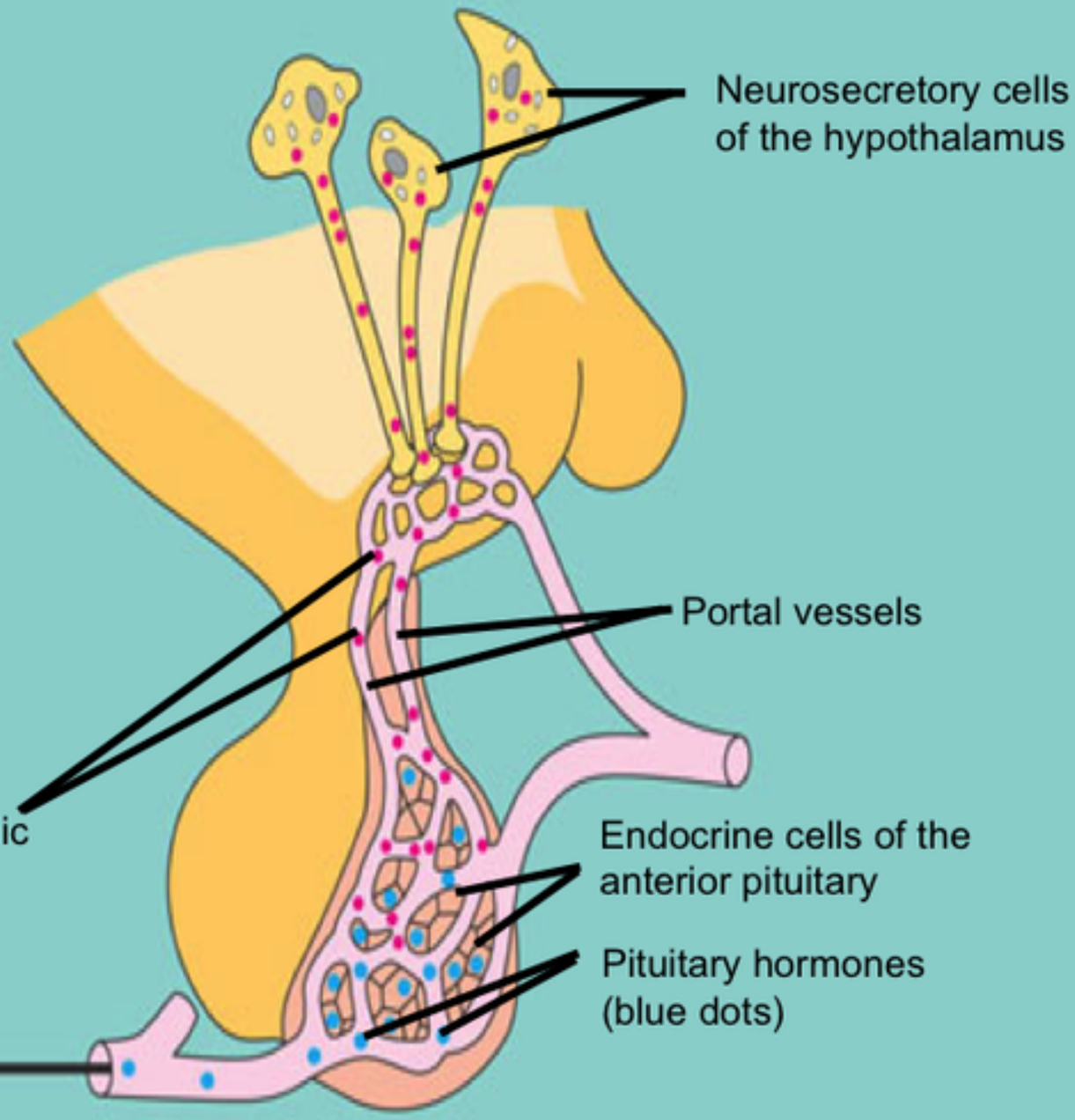
Figure 19.4 Pituitary Hormones and Their Targets



Tropic Effects Only
FSH, follicle-stimulating hormone
LH, luteinizing hormone
TSH, thyroid-stimulating hormone
ACTH, adrenocorticotropic hormone

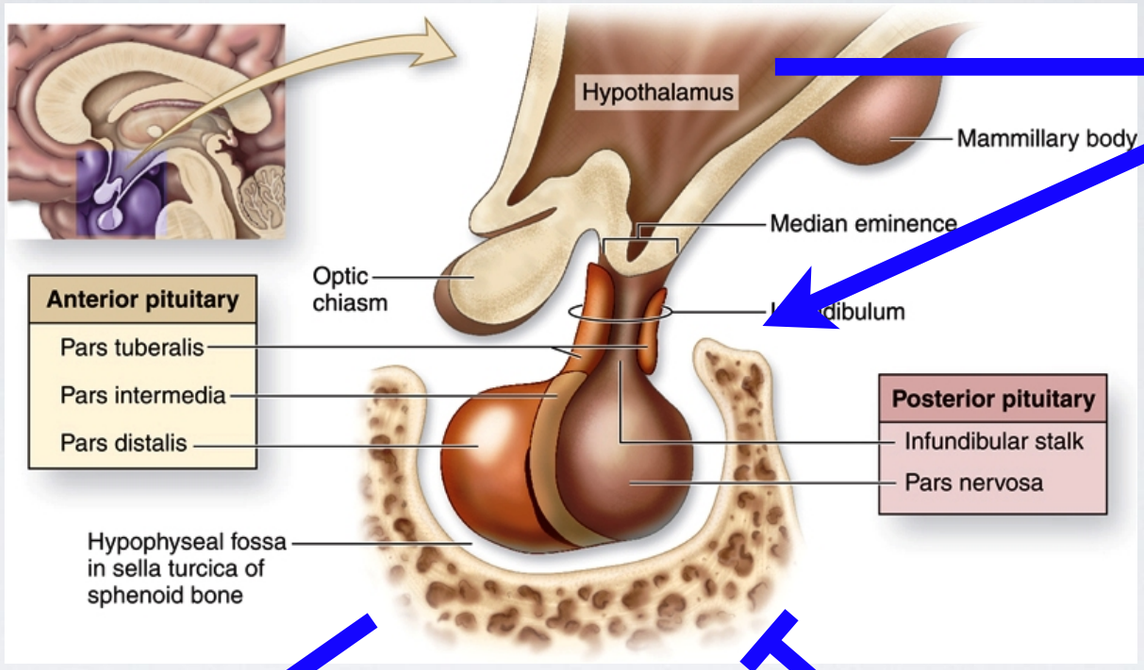
Nontropic Effects Only
Prolactin
MSH, melanocyte-stimulating hormone
Endorphin

Nontropic and Tropic Effects
Growth hormone



HORMONE	FSH and LH	TSH	ACTH	Prolactin	MSH	Endorphin	Growth hormone
TARGET	Testes or ovaries	Thyroid	Adrenal cortex	Mammary glands	Melanocytes	Pain receptors in the brain	Liver, Bones

Thyroid circuit

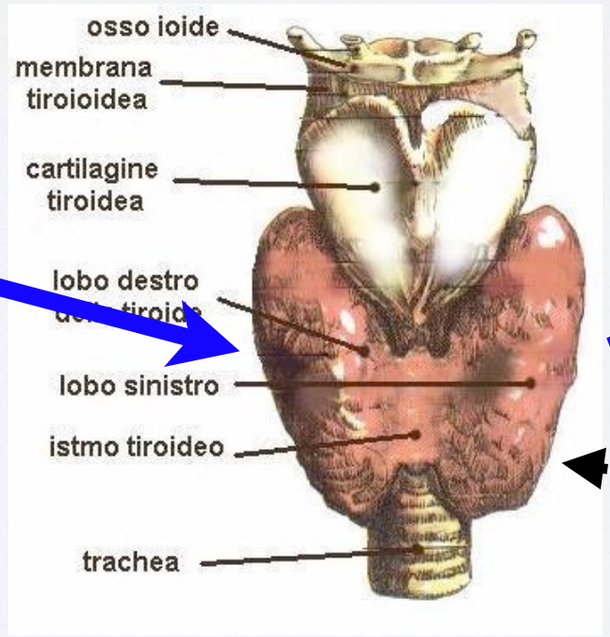


TRH

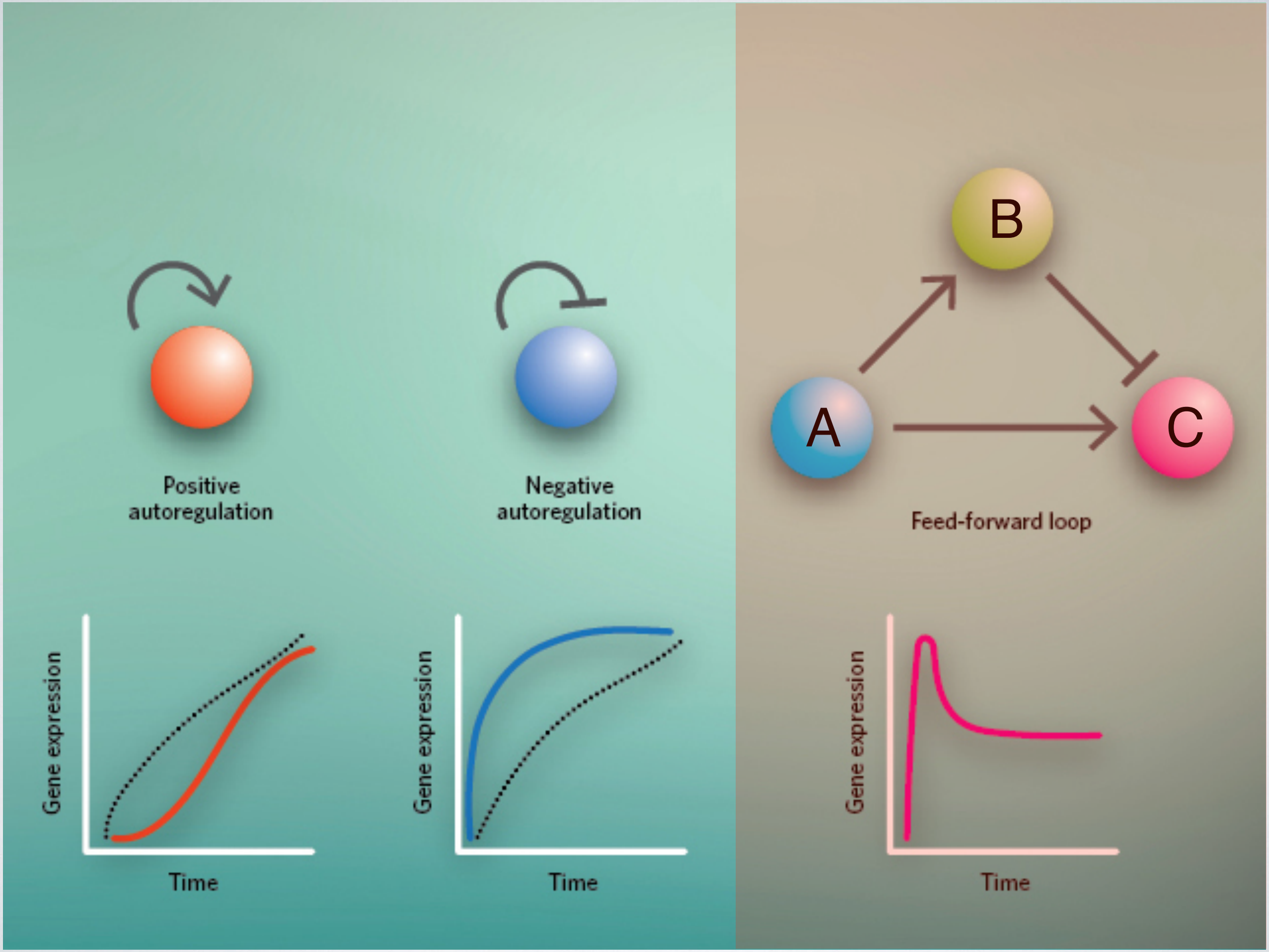
TSH

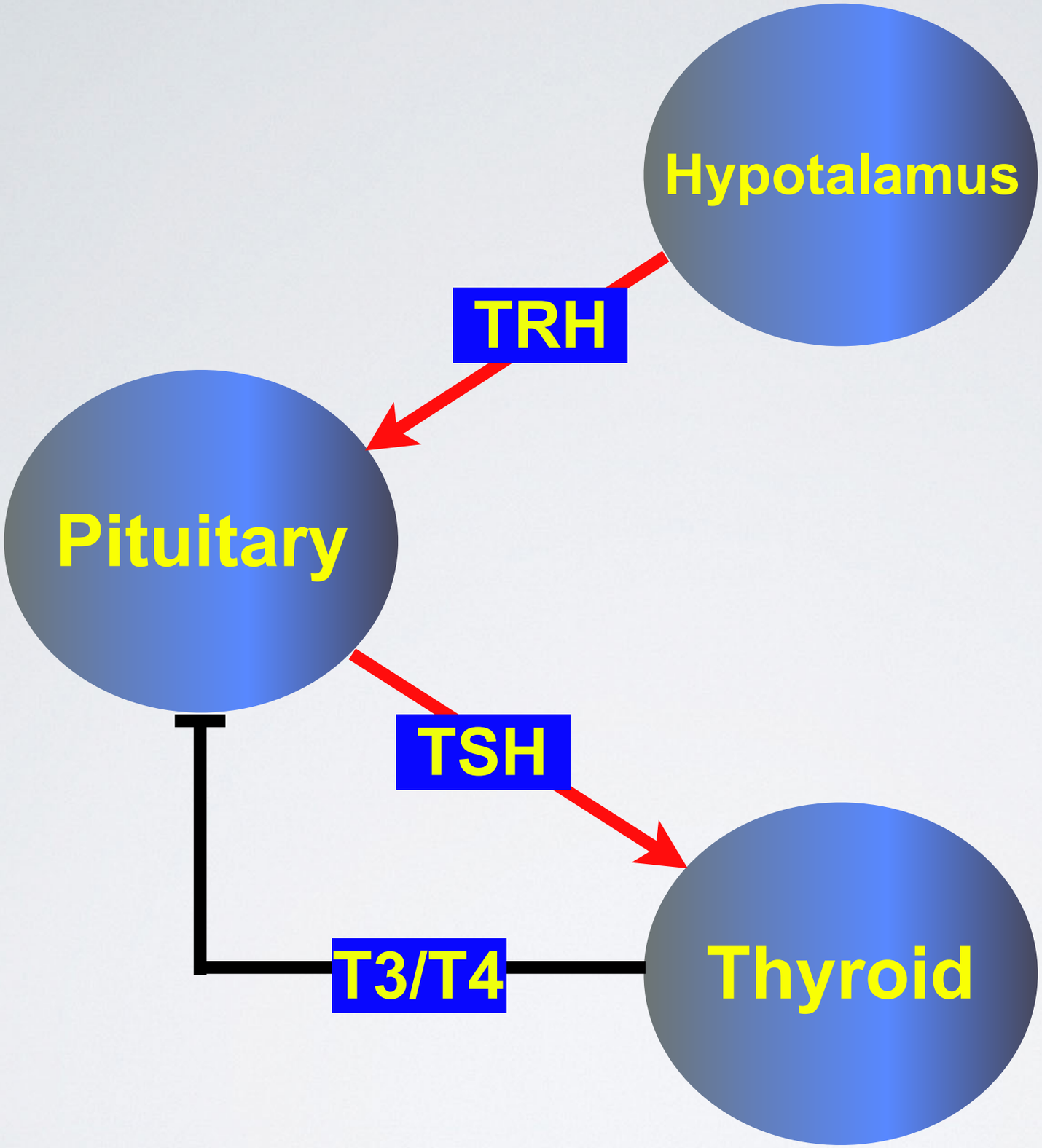
T3/T4

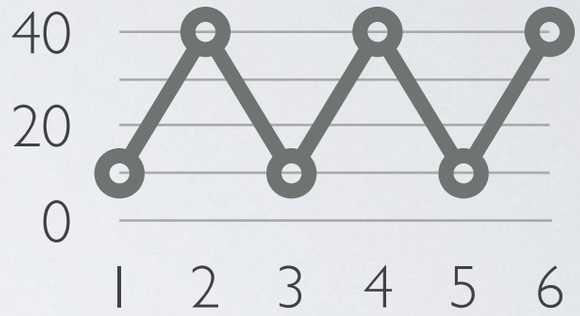
Negative feedback

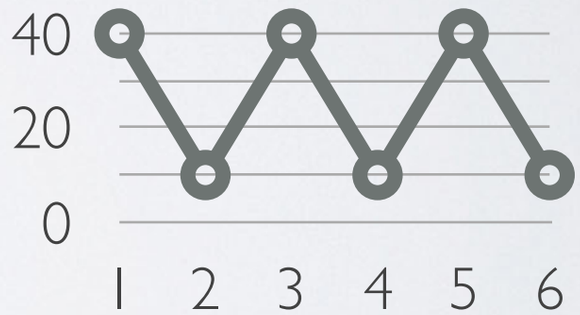


Thyroid


Positive autoregulation
Gene expression
Time
Negative autoregulation
Gene expression
Time
B
A
C
Feed-forward loop
Gene expression
Time

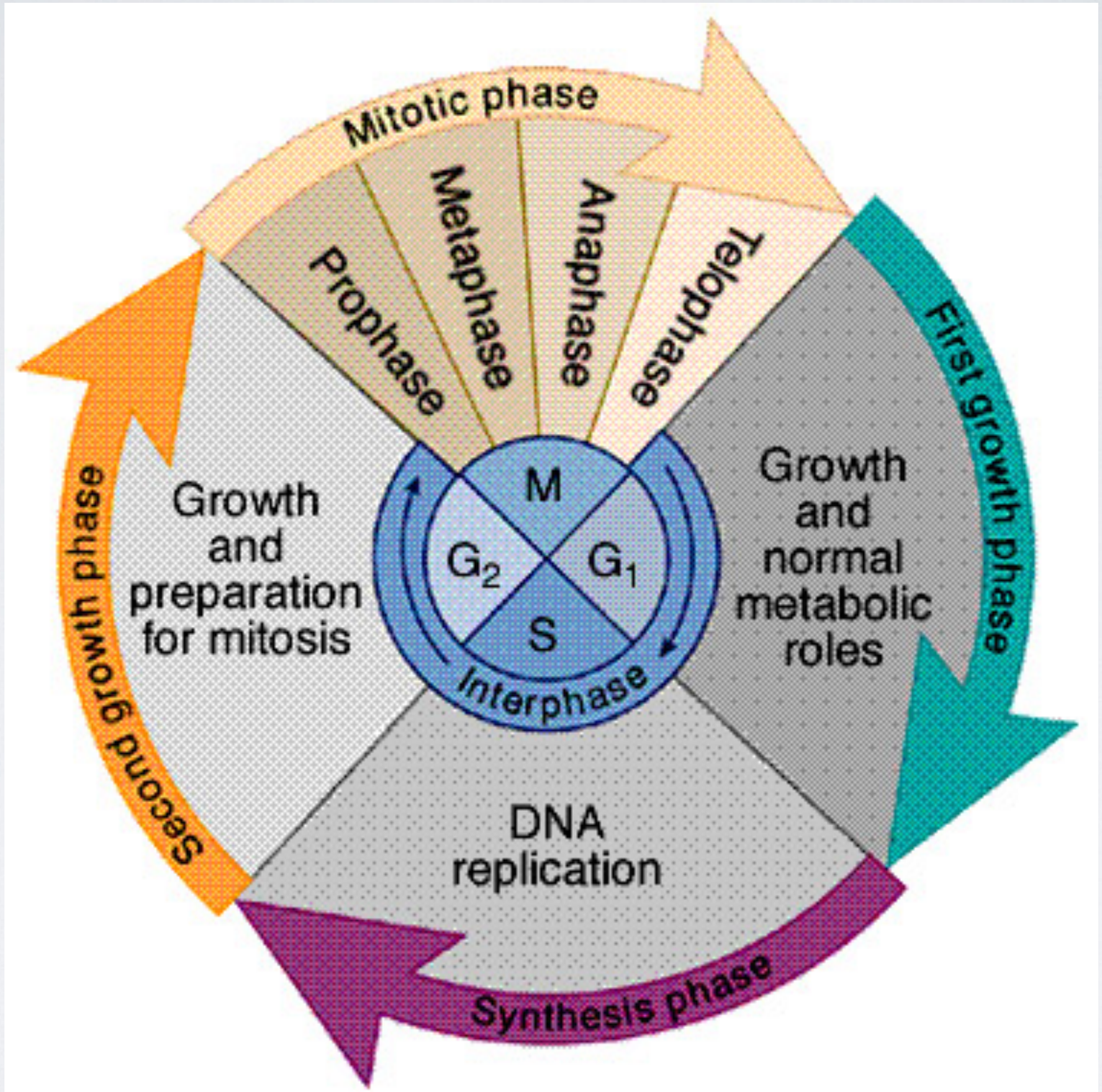






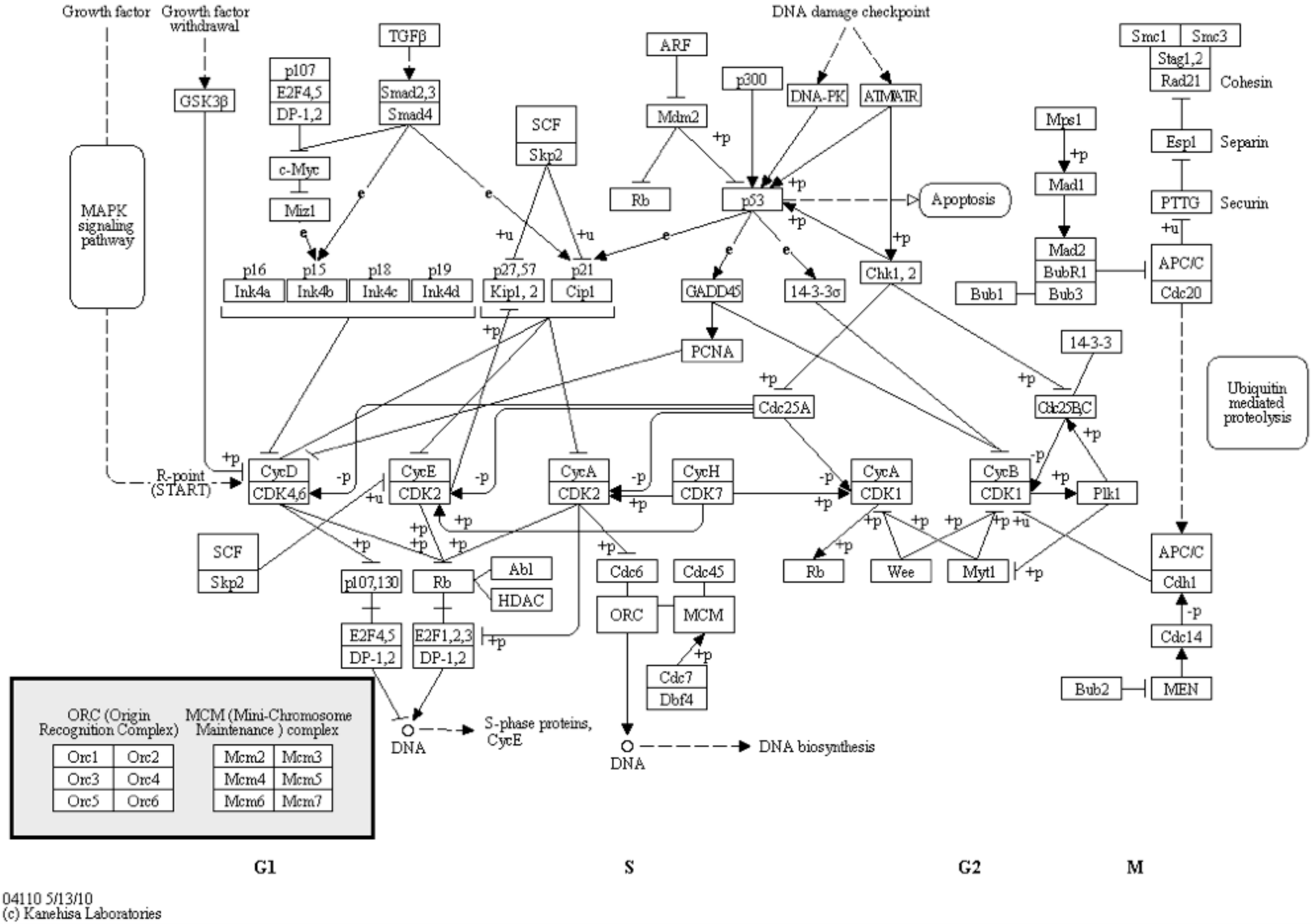
The synthesis (and the blood levels) of hormones fluctuates:
hormonal oscillation is an emergent property

The cell cycle: a molecular oscillator



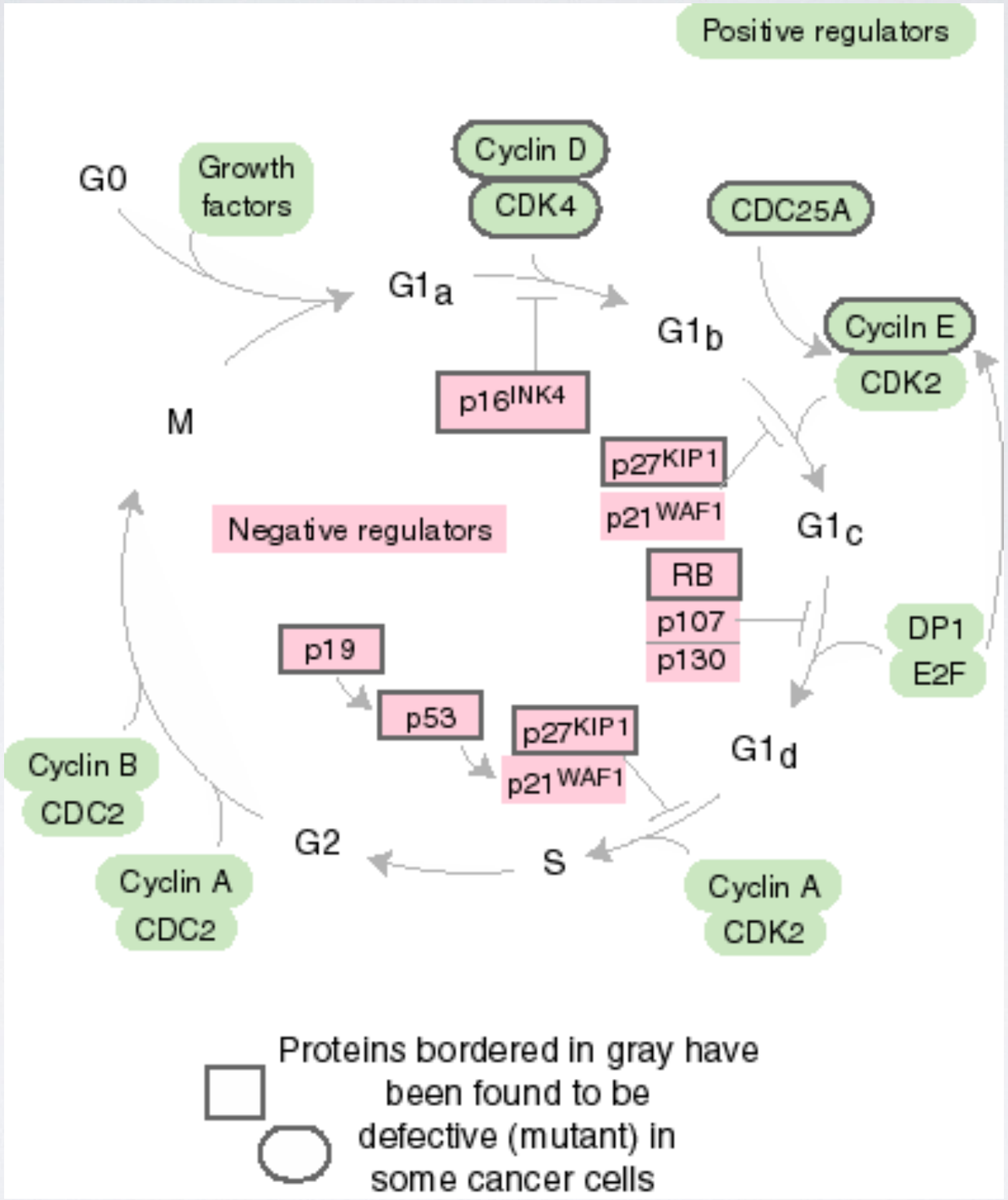
Period ~ 12 hours

CELL CYCLE

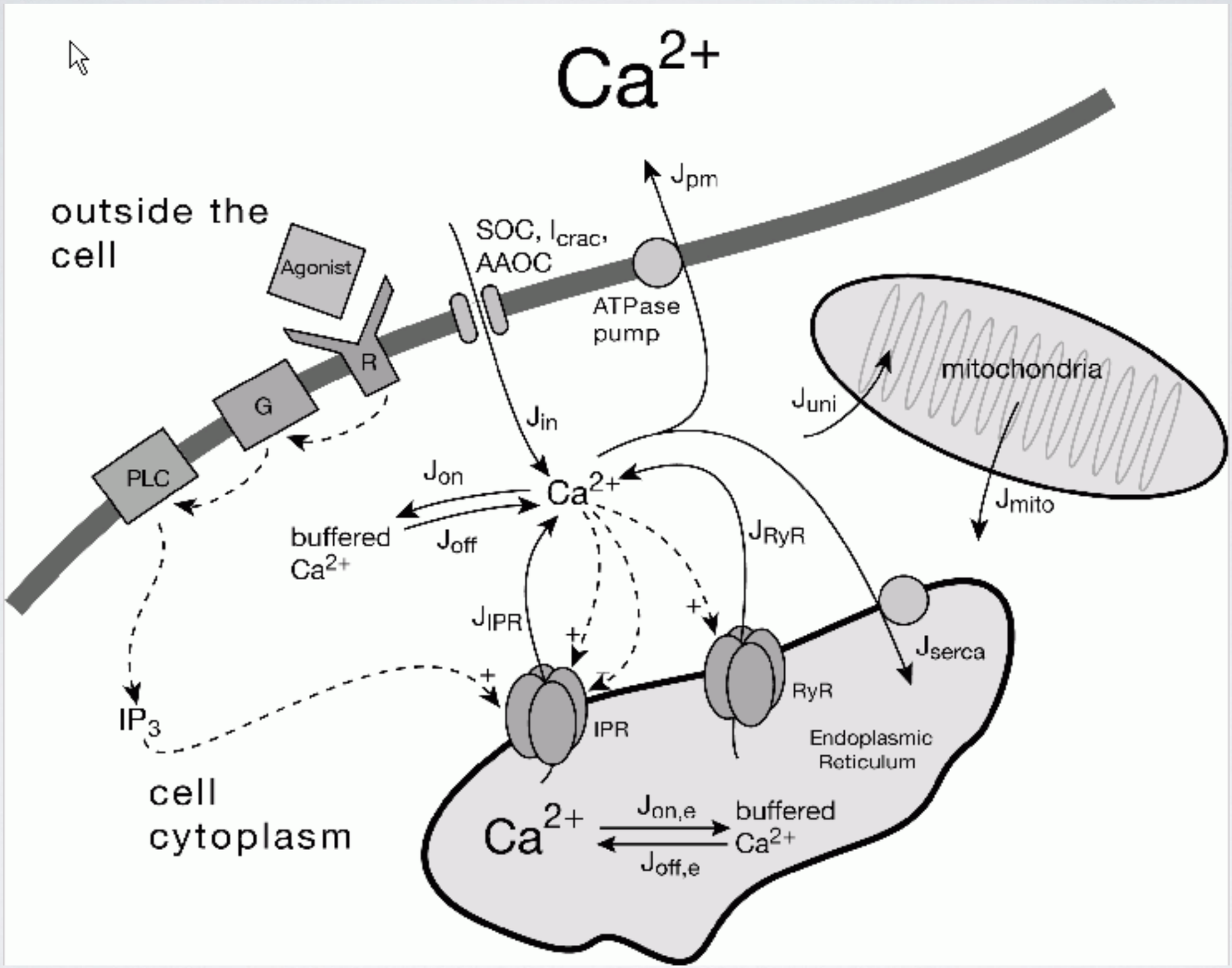


04110 5/13/10

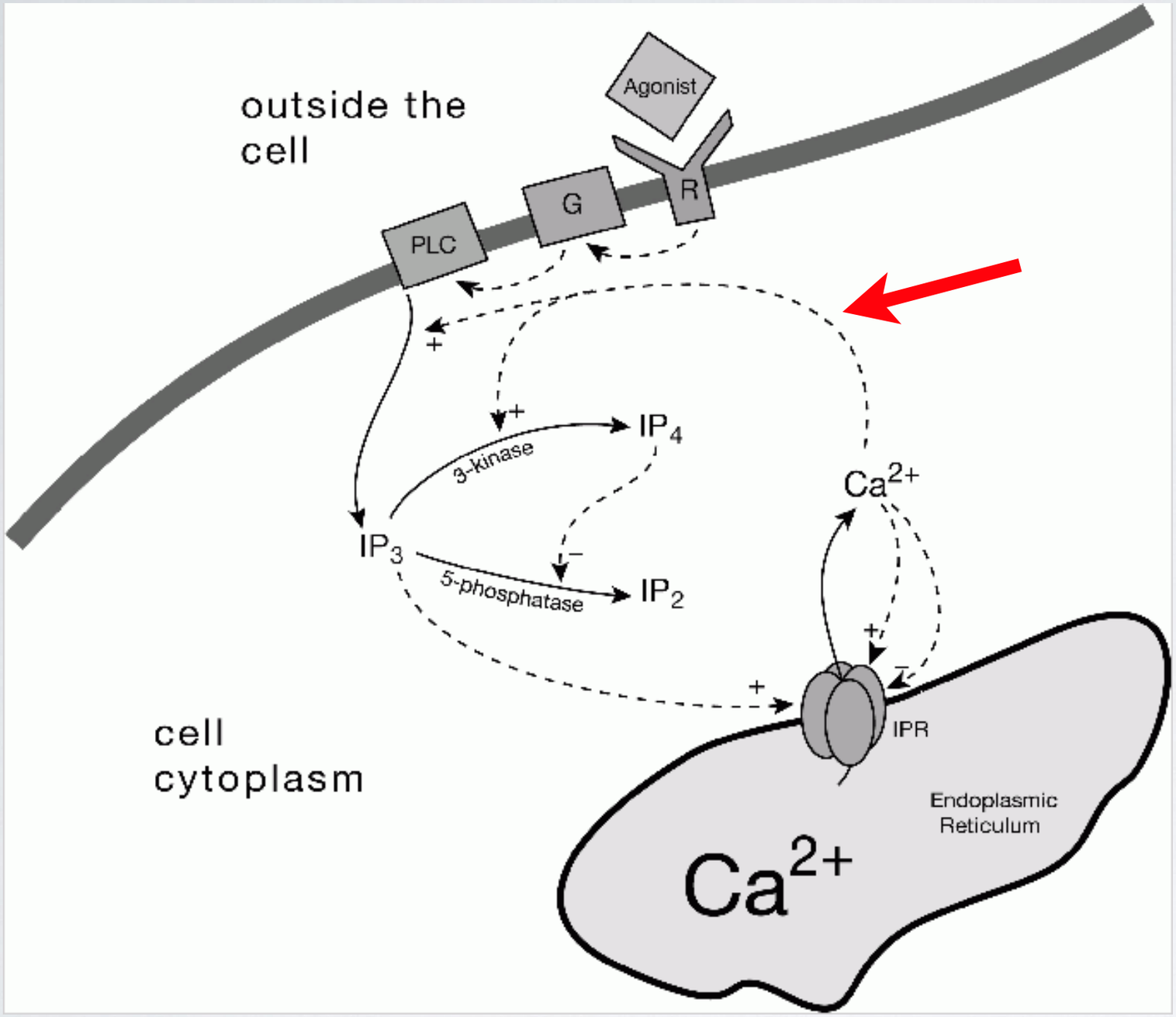
cell cycle: cyclins oscillate

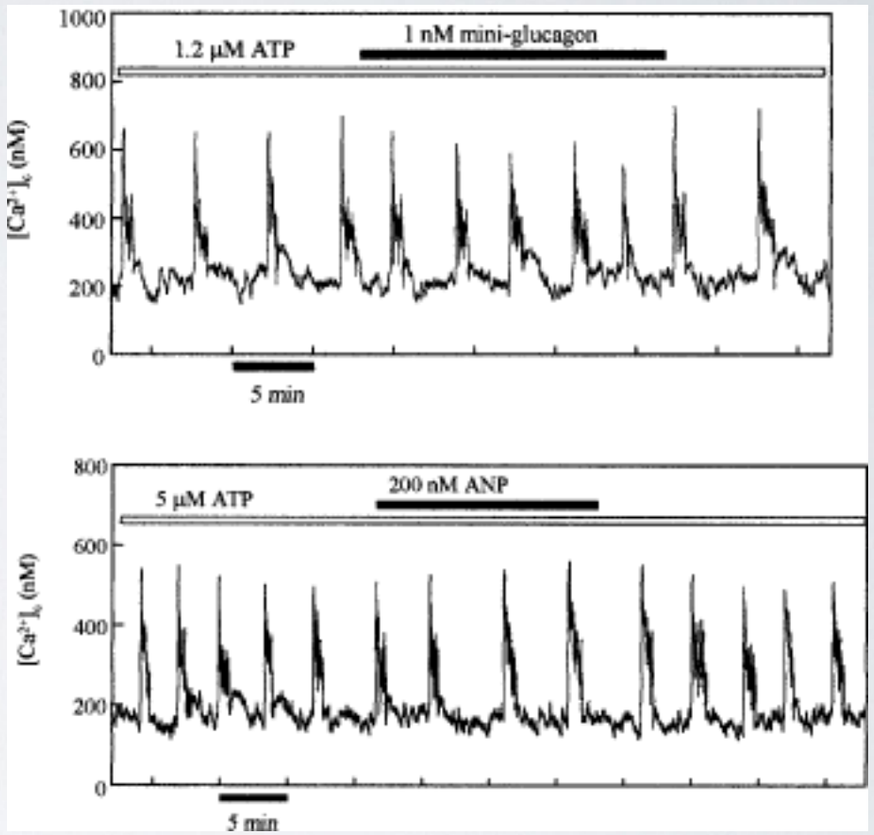


The oscillation of intracellular calcium



Amplitude + frequency





Amplitude + frequency

The frequencies of calcium oscillations are optimized for efficient calcium-mediated activation of Ras and the ERK/MAPK cascade

Sabine Kupzig*, Simon A. Walker†, and Peter J. Cullen*‡

*Henry Wellcome Integrated Signalling Laboratories, Department of Biochemistry, School of Medical Sciences, University of Bristol, Bristol BS8 1TD, United Kingdom; and †Signalling Programme, The Babraham Institute, Babraham Research Campus, Cambridge CB2 4AT, United Kingdom

Edited by Solomon H. Snyder, The Johns Hopkins University School of Medicine, Baltimore, MD, and approved March 10, 2005 (received for review December 22, 2004)

Ras proteins are binary switches that, by cycling through inactive GDP- and active GTP-bound conformations, regulate multiple cellular signaling pathways, including those that control growth and differentiation. For some time, it has been known that receptor-mediated increases in the concentration of intracellular free calcium ($[Ca^{2+}]_i$) can modulate Ras activation. Increases in $[Ca^{2+}]_i$ often occur as repetitive Ca^{2+} spikes or oscillations. Induced by electrical or receptor stimuli, these repetitive Ca^{2+} oscillations increase in frequency with the amplitude of receptor stimuli, a phenomenon critical for the induction of selective cellular functions. Here, we show that Ca^{2+} oscillations are optimized for Ca^{2+}-mediated activation of Ras and signaling through the extracellular signal-regulated kinase (ERK)/mitogen-activated protein kinase (MAPK) cascade. We present additional evidence that Ca^{2+} oscillations reduce the effective Ca^{2+} threshold for the activation of Ras and that the oscillatory frequency is optimized for activation of Ras and the ERK/MAPK pathway. Our results describe a hitherto unrecognized link between complex Ca^{2+} signals and the modulation of the Ras/ERK/MAPK signaling cascade.

RASAL | CAPRI | GRF | GDP

For a wide variety of cell surface receptors, activation leads to an increase in the concentration of intracellular free calcium ($[Ca^{2+}]_i$) (1–3). Once induced, the elevation in $[Ca^{2+}]_i$ is responsible for controlling a diverse array of cellular processes, including secretion, contraction, learning, and proliferation (3). Understanding how receptor-mediated increases in $[Ca^{2+}]_i$ are capable of modulating so many physiological processes is one of the major challenges in the study of Ca^{2+} signaling. It appears that such control is achieved through a complex relationship between the amplitude and spatiotemporal patterning of the Ca^{2+} signal and its resultant ability to couple to an extensive molecular repertoire of Ca^{2+}-sensing proteins (3).

Receptor-mediated increases in $[Ca^{2+}]_i$ are often observed as repetitive Ca^{2+} spikes or oscillations that increase their frequency with the amplitude of the receptor stimuli (refs. 4 and 5; reviewed in ref. 3). These frequency-encoded signals appear to be critical for the induction of selective cellular functions (3). For example, the frequency of receptor-mediated Ca^{2+} oscillations determines the efficiency of gene expression driven by the transcription factors NF-AT, OAP, and NF-κB (6–8) and mitochondrial ATP production (9). To decode the information contained within Ca^{2+} oscillations, cells have evolved a number of frequency-modulated decoders. Such proteins include calmodulin (10), protein kinase C (11–15), calpain (16), calmodulin-dependent protein kinase II (17, 18), and the Ras GTPase-activating protein RASAL (19).

Ras proteins are binary molecular switches that regulate multiple signaling pathways, including those controlling growth and differentiation, through an ability to cycle between inactive GDP- and active GTP-bound conformations (20–23). The magnitude and duration of Ras signaling is controlled by two classes of proteins: Guanine nucleotide exchange factors modulate Ras activation by enhancing the exchange of GDP for GTP, and GTPase-activating proteins regulate inactivation by increasing the intrinsic Ras GTPase activity (20–23). Although it has been known for some time that increases in $[Ca^{2+}]_i$ can modulate Ras activation (for example, Ca^{2+} influx through voltage-operated ion channels or release from internal stores can activate Ras in neuronal cells) (24), only recently have molecular entities been described that allow for this coupling (reviewed in ref. 25).

Two families of Ras guanine nucleotide exchange factors (GEFs), RasGRFs (26–29) and RasGRPs (30–36), the latter also being known as CalDAG-GEFs, are modulated by increases in $[Ca^{2+}]_i$. For RasGRFs, this modulation occurs indirectly through association with Ca^{2+}/calmodulin, whereas for RasGRPs, a more direct control is achieved through association of Ca^{2+} with atypical EF hands (25). In addition to stimulating Ras activation, increases in $[Ca^{2+}]_i$ also mediate Ras inactivation through the Ca^{2+}-triggered RasGTPase-activating proteins (RasGAPs) RASAL and CAPRI (19, 37). These proteins are cytosolic, inactive RasGAPs that, upon a receptor-mediated elevation in $[Ca^{2+}]_i$, undergo a rapid, C_2 domain-dependent association with the plasma membrane, an association that leads to an increase in their RasGAP activity (19, 37). Unlike CAPRI, which undergoes a transient association with the plasma membrane and does not sense receptor-mediated Ca^{2+} oscillations, the plasma membrane association of RASAL occurs in an oscillatory manner (19). This oscillatory association occurs in synchrony with underlying receptor-mediated Ca^{2+} oscillations and is frequency-modulated such that, upon increasing the amplitude of receptor stimuli, the frequency of RASAL membrane association is enhanced (19). CAPRI and RASAL therefore constitute molecular entities that can sense the amplitude and frequency, respectively, of complex Ca^{2+} signals, decoding these distinct temporal signals through a modulation of plasma-membrane-associated Ras.

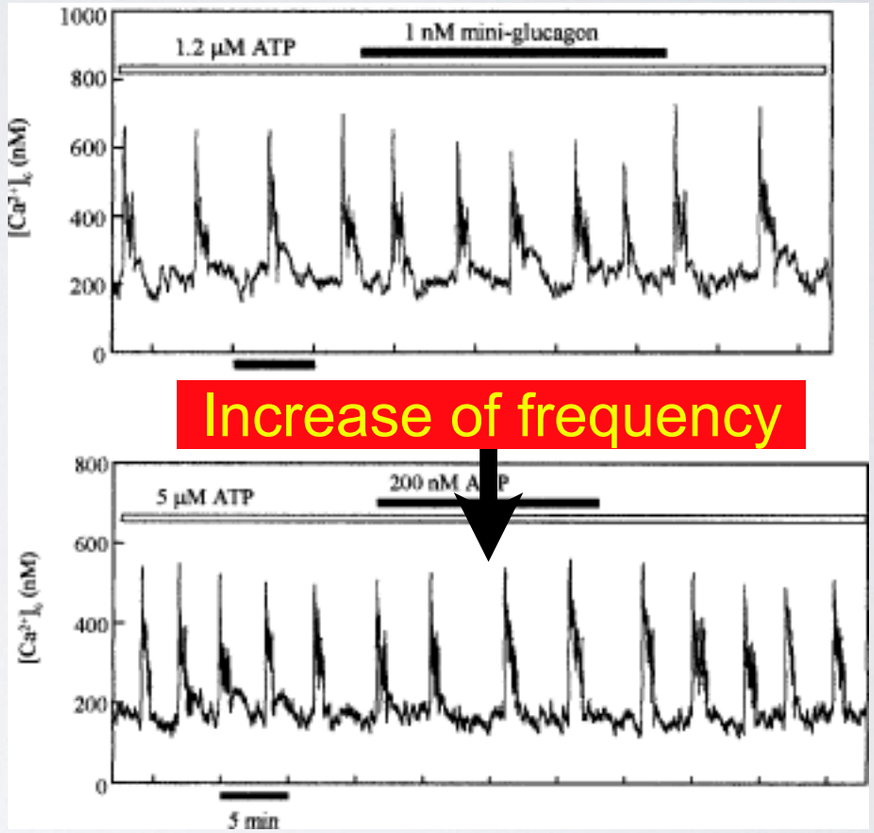
The characterization of such distinct Ca^{2+} sensors, tuned to detect different temporal Ca^{2+} signals, has raised the issue of whether the temporal dynamics of receptor-mediated Ca^{2+} oscillations are optimized for efficient Ca^{2+}-mediated activation of Ras and downstream Ras-dependent signaling (25, 38). Here, we have addressed this issue, presenting data showing that the temporal dynamics of Ca^{2+} signals are indeed optimized for activation of Ras and the downstream extracellular signal-

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviations: $[Ca^{2+}]_i$, concentration of intracellular free calcium; EGF, epithelial growth factor; ERK, extracellular signal-regulated kinase; MAPK, mitogen-activated protein kinase.

‡To whom correspondence should be addressed. E-mail: pete.cullen@bris.ac.uk.



HRas-MAPK pathway

The H-Ras-MAPK pathway is not regulated by the amount (concentration) of absolute Ca_i released (amplitude)

BUT

by the <u>frequency of the oscillations</u> (time flow)

An emergent property that directly regulates a biological function

It is not the function of a single molecule that regulates.

Instead, the ensemble of interacting molecules, appropriately combined, generates circuits generating <u>emergent functions</u>.

Complex systems and dissipative behaviour

A dissipative structure is characterised by the "**spontaneous**", *self-organised*, appearance of symmetry breaking (anisotropy) and the formation of complex structures where interacting particles exhibit long range correlations.

Examples in everyday life include convection, turbulent flow, cyclones, hurricanes and … **living organisms.**

Two **fundamental** properties of complex biological systems

(1)
Biological systems are
dissipative structures
(Ilya Prigogine,
Nobel Laureate for chemistry 1977)

Thermodynamically open system working *far from thermodynamic equilibrium* and constantly exchanging with the environment:
energy, matter, information and / or entropy

"Apparent" violation of 2º thermodynamics law

(2)
Biological systems show <u>self-organisation</u>

Biological systems are characterised by the **spontaneous** generation of **anisotropy**, that is of ordered structures with "increasing" complexity.

Biological systems, when traversed by increasing flows of energy and matter, <u>instead of accumulating entropy</u>, do evolve and, passing through phases of instability, do <u>increase the complexity of their structure</u> (that is increase the order) and reduce their entropy.
Losing entropy to generate complexity (patterns).

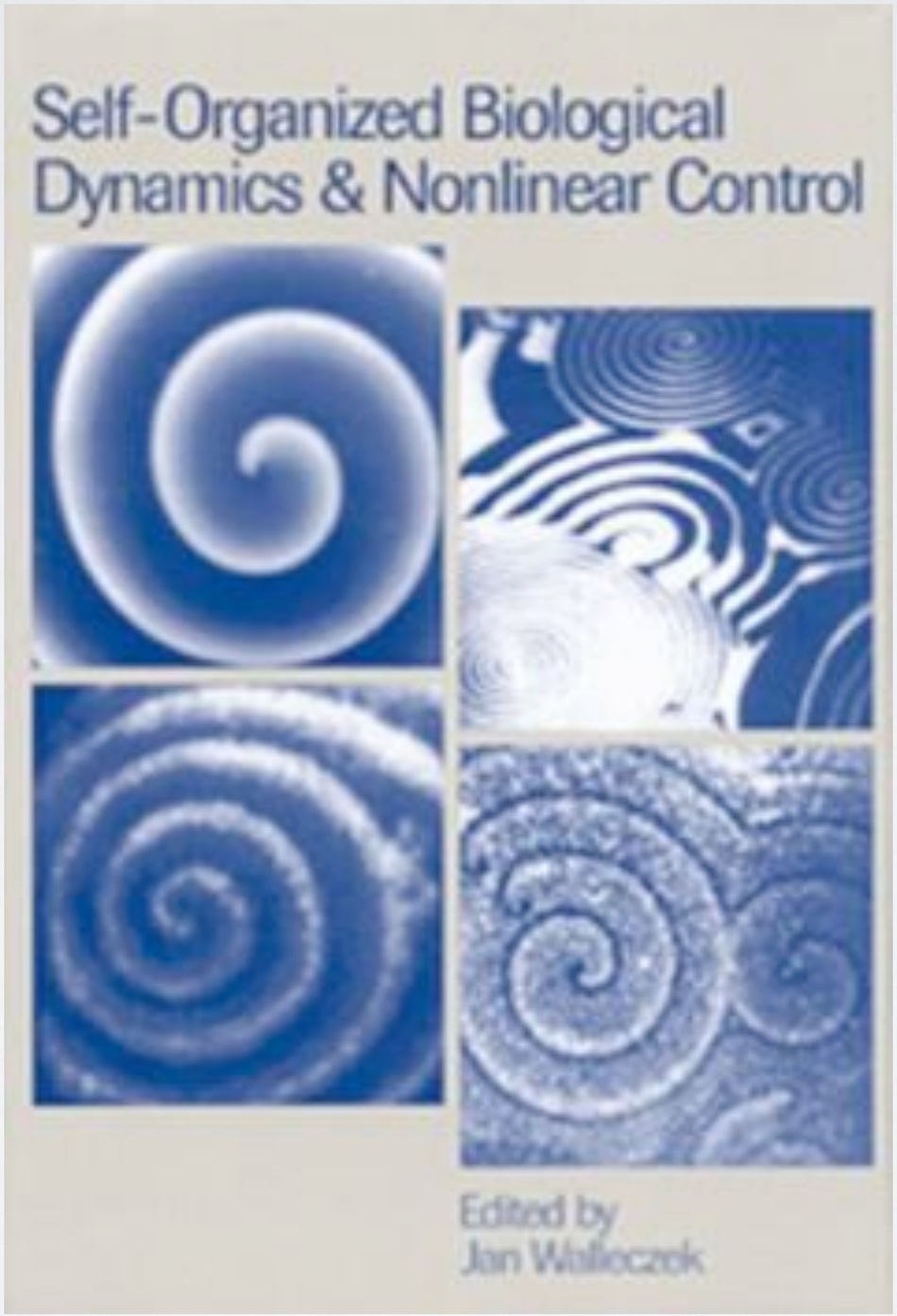
Complexity = Organisation = Order

Emergence of patterns (order, structures)

JAN WALLECZEK

Self-Organized Biological Dynamics and Nonlinear Control:

Toward Understanding Complexity, Chaos and Emergent Function in Living Systems



How do we study biological complex systems?

(3)
Theses-based science: the deductive method (Aristotele)

The deductive method (literally "led by") is the rational process that derives a particular conclusion from more general assumption (Universal), within which that conclusion appears to be implied.

The deductive method starts first from (undemonstrated) postulates and general principles and, through a series of rigorous logical concatenation, proceed towards more particular determinations related to reality, observable and measurable.

eg. Aristotelian syllogism

1_ A = B
2_ B = C
3_ A = C

Mathematics

Theoretical physics (?)

The deductive method always starts from a postulate or from an axiom or a group of axioms, that is, from an absolute truth that does not need to be verified, from which it deduces, through a reasoning, the particular facts.

Therefore the validity of what is deducted collapses if it is proved that the starting affirmation is false or arbitrary.

In this way, precisely collapses the original assumption on which the entire argument (logical process) itself was founded.

(3)
Science based on experimental data: the inductive method (Galileo Galilei)

The inductive method (literally “conduct in") (reverse process of deduction) implies that the thought is funded on experience.

The sensitive data of experience (observation of Nature) are “induced”, de facto introduced, into the intellect, which then draws universal and abstract laws.

Also called “a posteriori” as the conceptual analysis of the reality is only possible upon the experience.

All experimental science

Deductive Method

↓

Universal

↓

Particular

Inductive Method

↓

Particular

↓

Universal

↔

How does Science really work?

(Popper, Khun)

2 SENSITIVE DATA

induction ⟷ *deduction*

1 OBSERVATION

3 THEORY

4 EXPERIMENT

Reductionism

Holism

Never demonstrated
only
"corroborated" by data

(4)
Systems Biology: definition and experimental connotation of Systems Biology

Definition: Systems biology is a discipline of life science that studies living organisms as complex systems that evolve over time, ie from the point of view of dynamic interaction of the parts of which they are composed. The final goal is to identify the emergent properties driving life.

A systems biology study is performed by integrating the results of different high-performance experiments (high throughput molecular state data) with analysis methods derived from dynamical systems theory, mathematics and bioinformatics.

Systems biology starts from the detailed knowledge of genomic, proteomic and / or metabolic molecular states of a living organism, by means of advanced multiplexed, high-performance, genomics, proteomics and metabolomics techniques, to determine dynamic molecular changes derived from a perturbation of the system.

The obtained experimental data are then processed using approaches of systems theory, bioinformatics and mathematical-statistical, with the goal of creating a predictive model of the functioning of biological systems.

Differently of molecular biology, which focuses on individual biological macromolecules, such as nucleic acids and proteins, and of traditional physiology, which studies biological systems not necessarily from the molecular point of view, systems biology studies the nature of the dynamic interactions between the various molecules that form, over time, a functioning biological system.

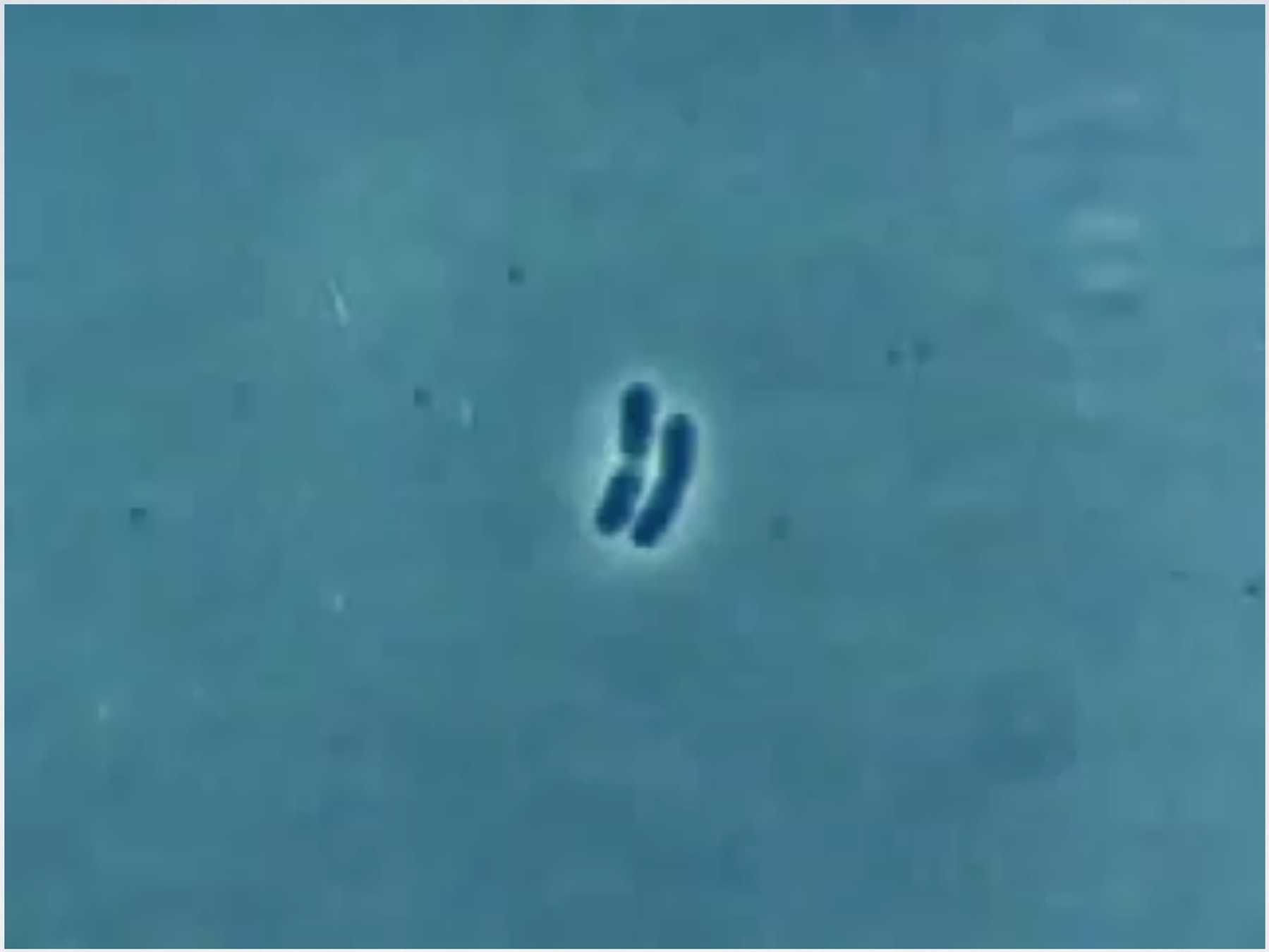
The ultimate goal is to construct a mathematical model that *allows the simulation* of the biological phenomenon computer in order to identify the **"emergent properties"** of the system.

Systems biology as a result of the procedures and principles applied is often identified as a "discovery science", as opposed to the traditional "hypotesis-driven science"

(5)
Why Systems Biology?
Reductionism versus holism

Biology studies that part of the physical world characterised by *being alive*!

Bacteria duplication

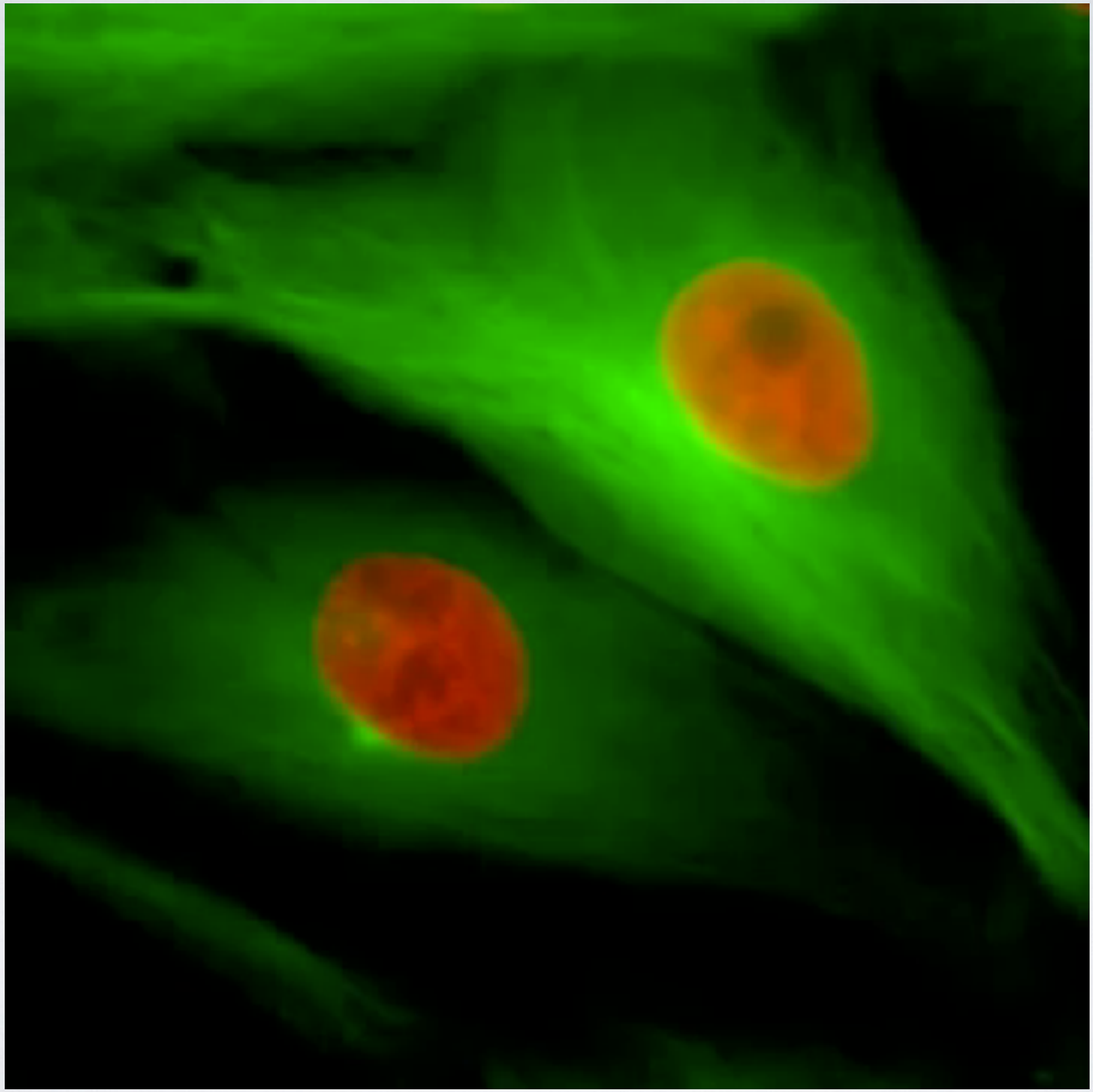


20 min

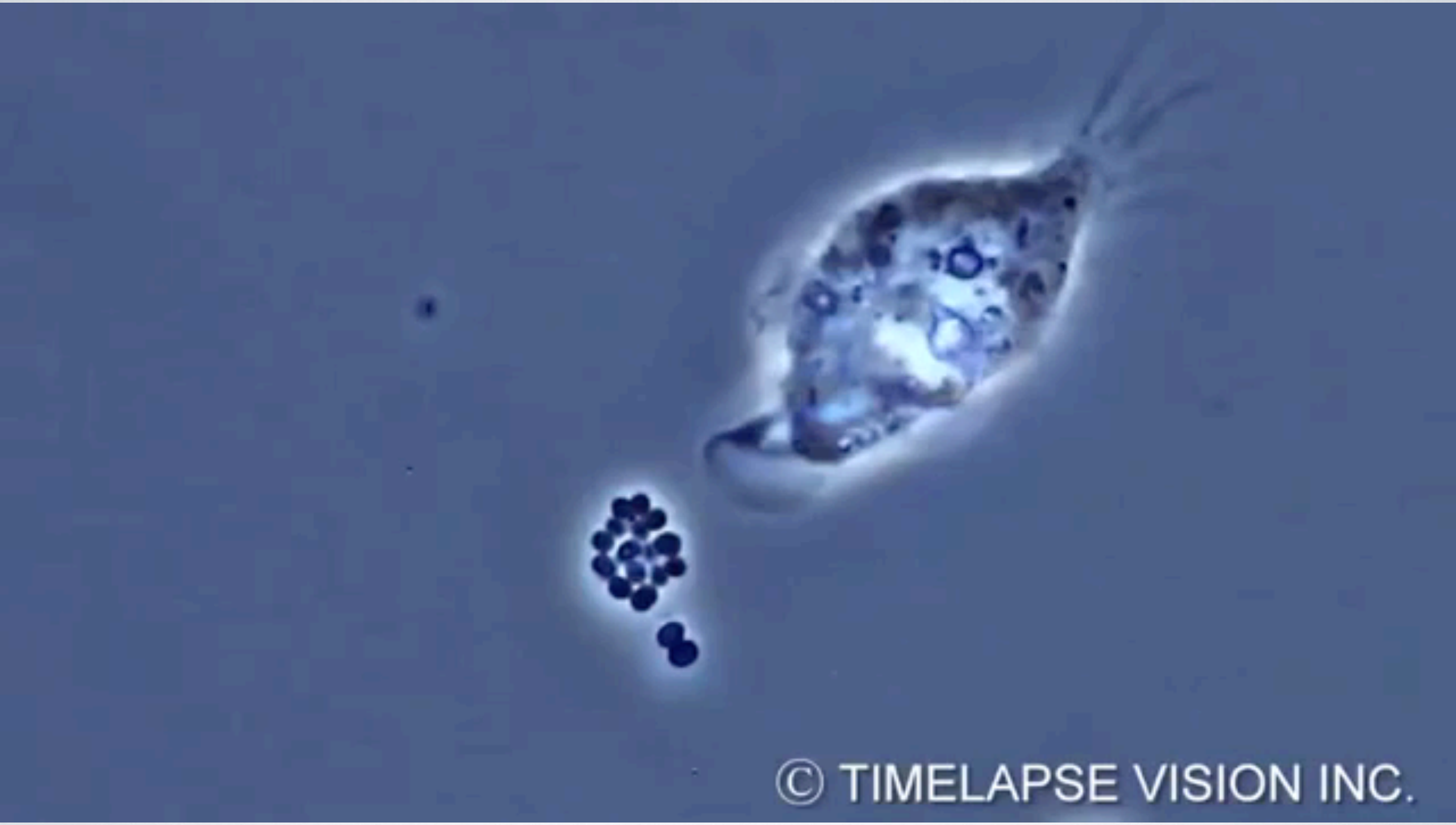
24 ore = $5x10^{21}$

Eukaryotic cell cycle

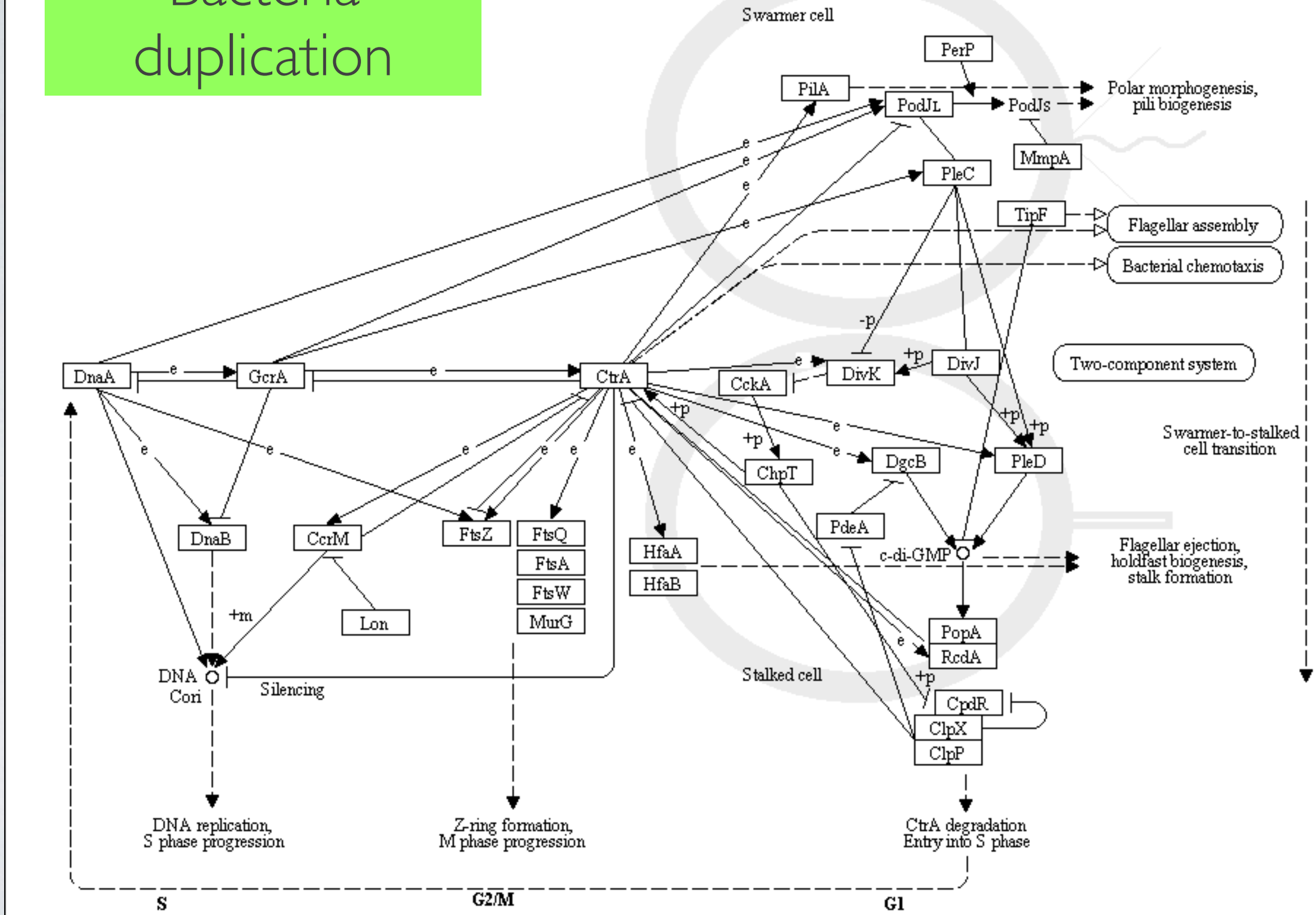
12-14 h



Leukocyte chemotaxis

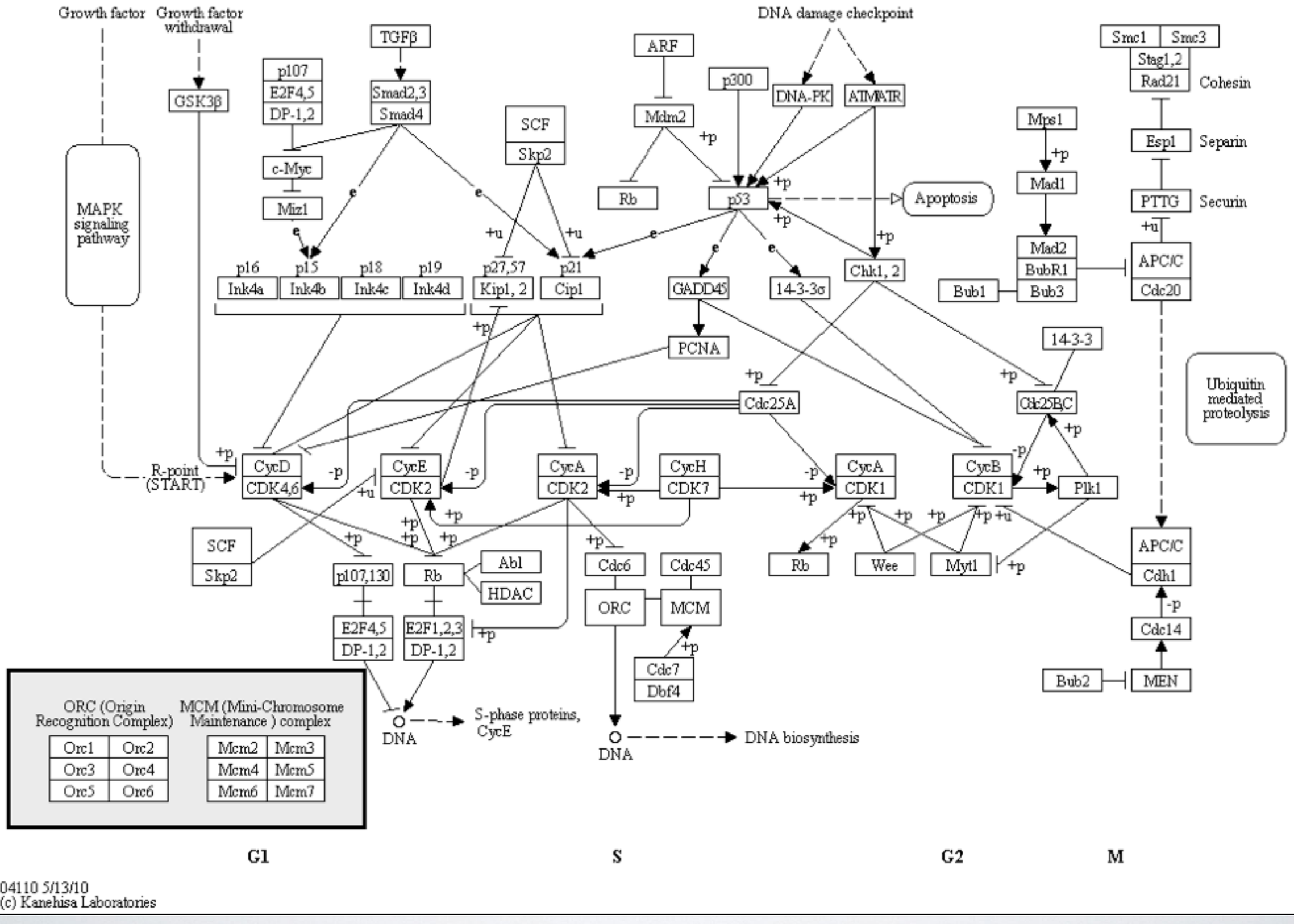


Bacteria duplication

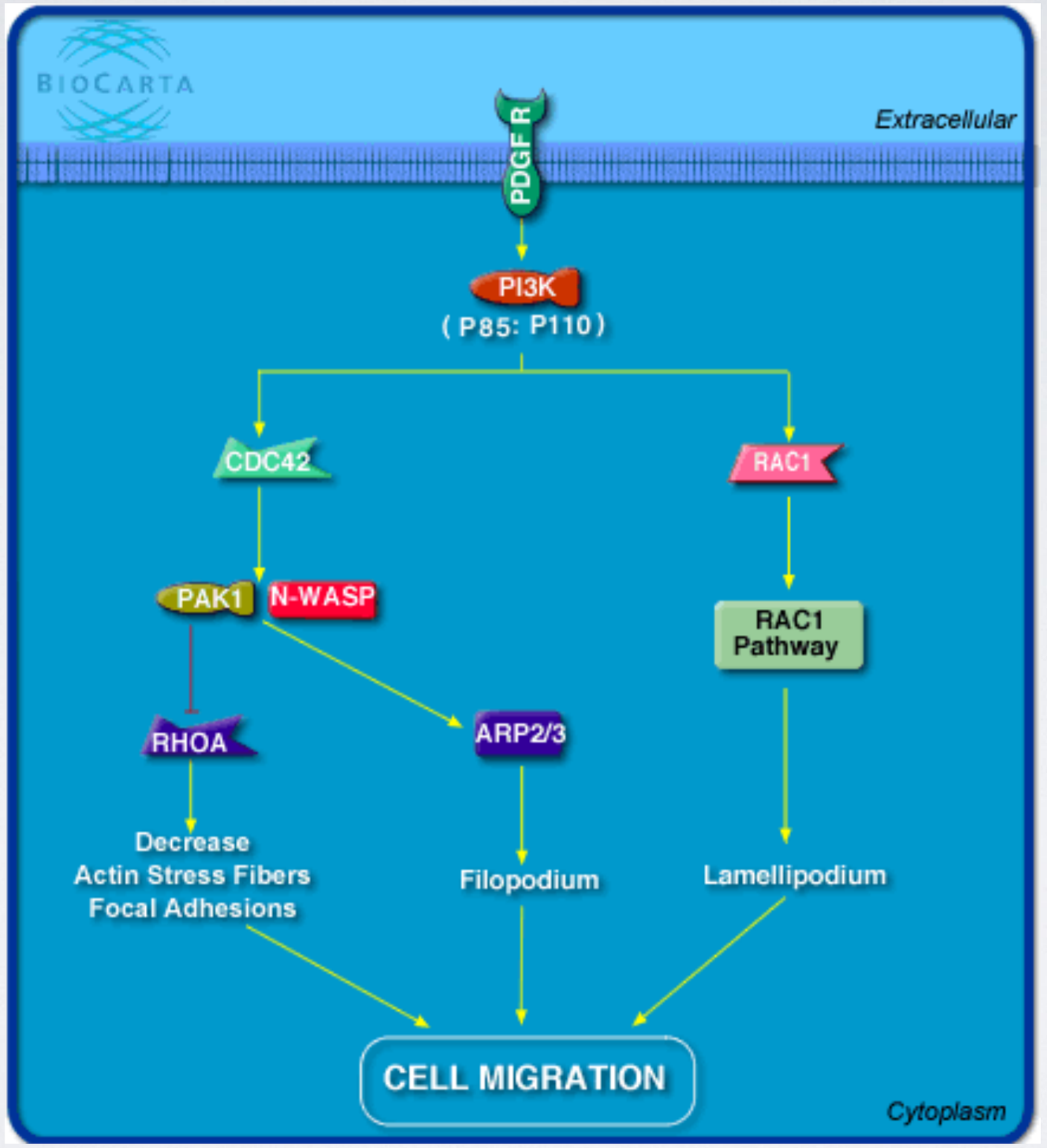


04112 4/13/10

Eukaryotic cell cycle



Leukocyte chemotaxis



Bacteria duplication	**Eukaryotic cell cycle**	**Leukocyte chemotaxis**
~ 60 molecole	~ 110 molecole	~ 50 molecole
Es.	**Es.**	**Es.**
FtsQ	**CDC25A**	**Rac1**

Are they alive?

Definition of life
(I)

An open thermodynamic system, capable of supporting themselves in an state of energetic stationary imbalance, and able to direct a series of chemical reactions towards the synthesis of itself

Definition of life
(2)

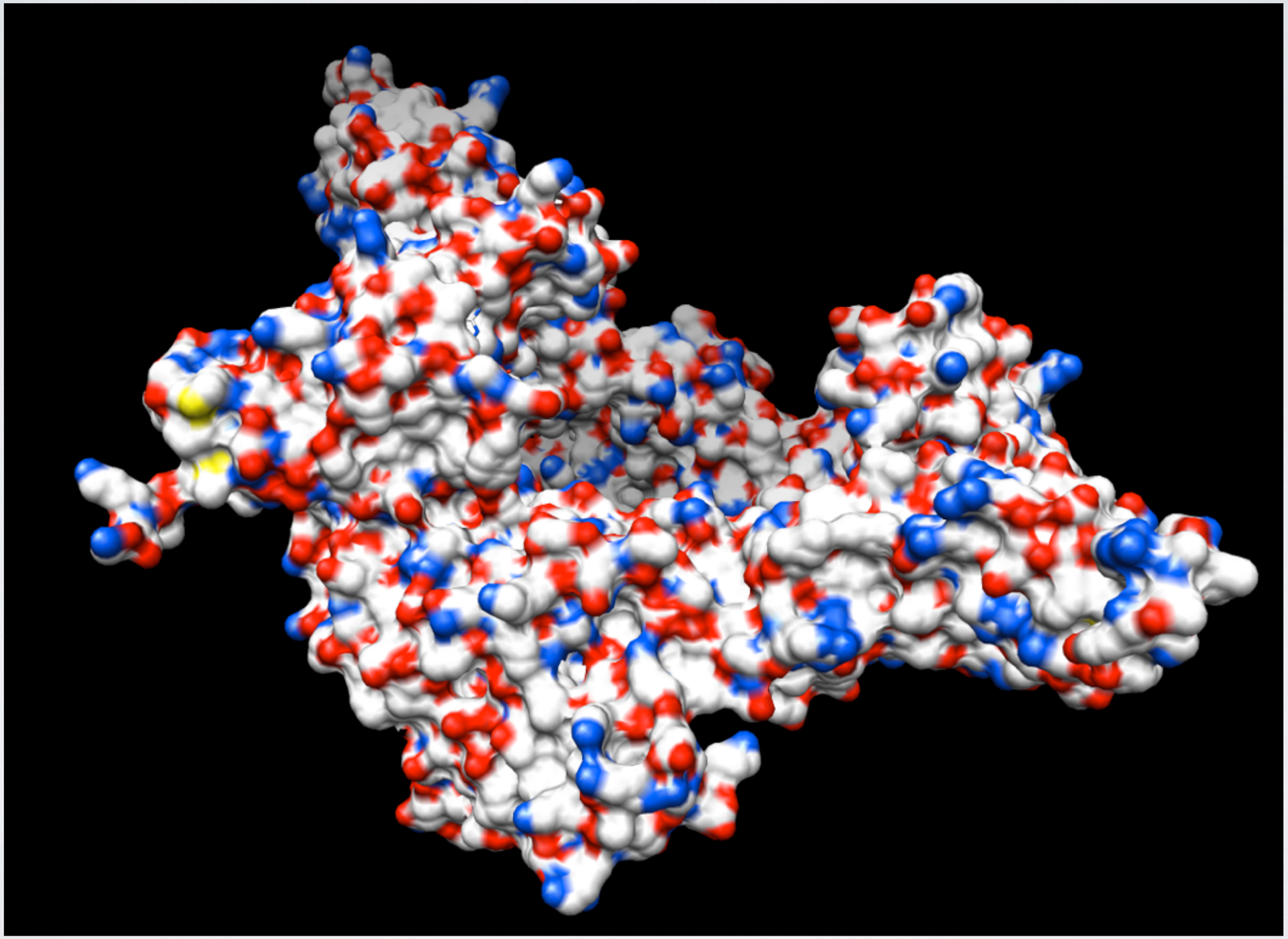
An autonomous physical entity that:

1) replicates itself

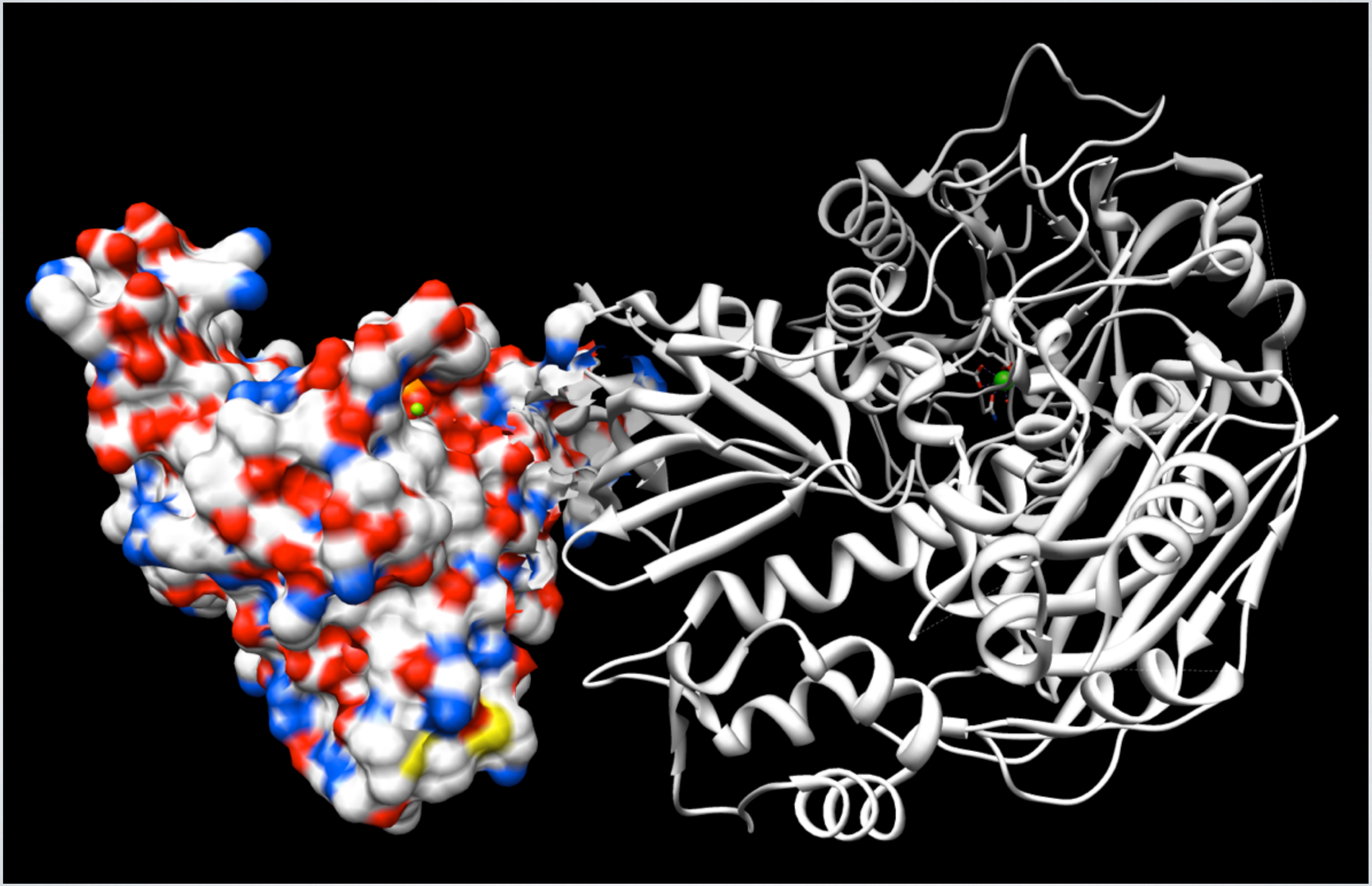
2) mutates

3) replicates the mutations

Is Rac1 alive?



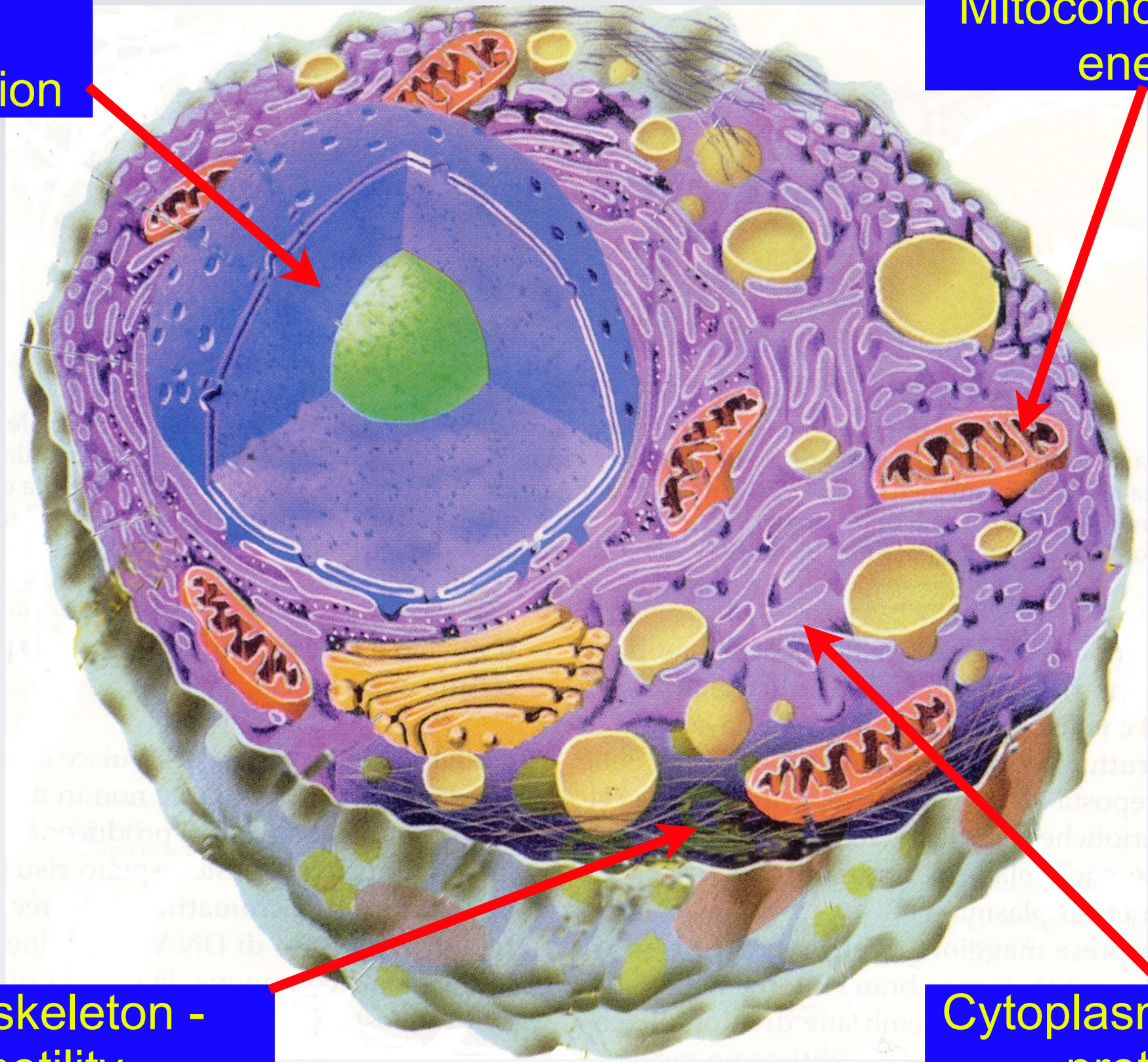
Is Rac1 associated to PLC-beta alive?



Surely, a cell is alive!

Nucleus - DNA Information

Mitocondria - ATP energy



Cytoskeleton - motility

Cytoplasm - mRNA proteins

The cell is "alive" because it's a highly complex and fully integrated physical object which

We need a scientific approach allowing dealing both with the molecular details as well as with the global dimension of living beings.

The new approach must allow modelling biological systems to generate predictive models of cell and organ functions

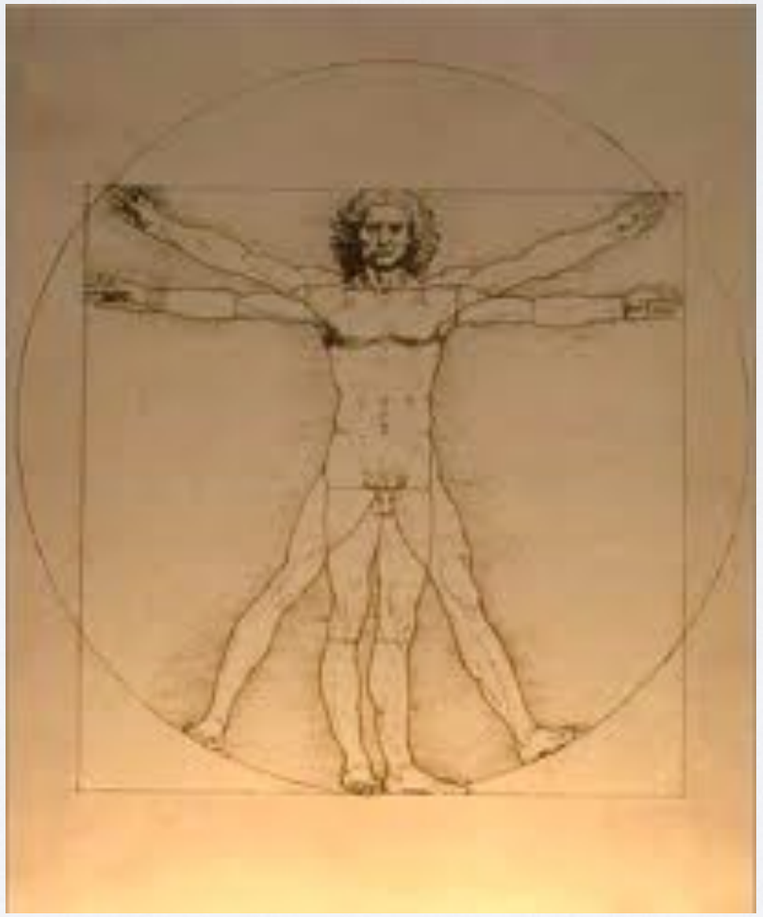
Life is an emergent property of a physical system characterised by <u>extreme</u> **combinatoriality**

How science deals with the study of life

Holism

Quantification

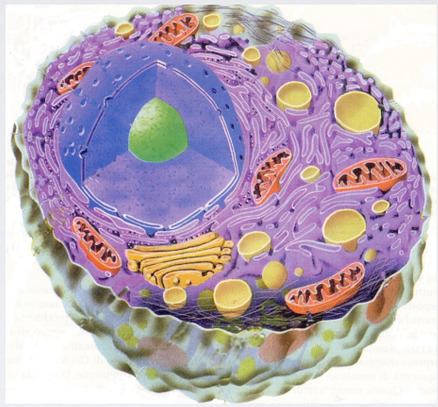
Functions



Model

No molecular dimension!

Reductionism



1 function

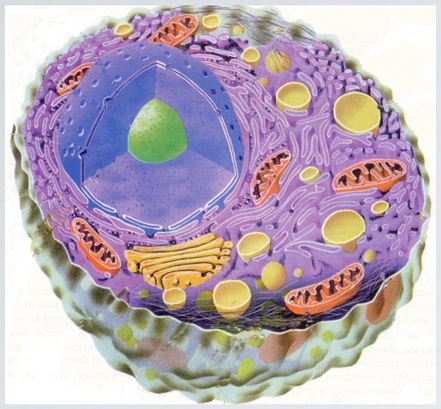
1 molecule

1 cause-effect relationship

Intuitive model

No global dimension!

Holistic reductionism = Systems biology



Functions

Global ***"omics" asset***
(high throughput methods)

Global cause-effect relationship

Computational analysis

Predictive mathematical modeling

System properties
("emergent properties")

	System analysis	Molecular level information	Quantitative analysis	Mathematical model	Emergent properties
Holism (physiology)	YES	NO	YES	YES	YES
Reductionism (molecular biology)	NO	YES	SI/NO	NO	NO
Systems biology	YES	YES	YES	YES	YES

"Cum granu salis"

Overall, Systems Biology provides a conceptual framework for common issues such as, network biology, network medicine, multidimensional integrative medicine and *in silico* drug testing.

In modern medicine, such a conceptual framework is nowadays synthesised in the advanced vision of personalized medicine or precision medicine.

The ultimate goal and 'the creation of a "computable" model.

(6)
The model: predicting the future in biology?

Systems biology aims at creating *molecular functional models* of the living phenomena.

But what is a model?

"A model is a selective abstraction of reality”

(Albert Einstein)

Abstraction: definition

The term abstraction indicates the mental process by which you replace a set of distinct objects with a more general concept, describing the objects based on their common properties. For example, starting from the set of all living beings, you can derive the generic concept of animal based on the characteristics shared by all animals with respect to the plants.

Abstraction: object vs. numbers

Few apples Qualitative determination Many apples

3 apples Quantitative determination 15 apples

3 apple + 15 apples = 18 apples

3 + 15 = 18

Numbers are quantitative abstractions of reality

Modelling in science

In science, a model is a representation of an object or a phenomenon, which corresponds to the modelled entity in way that it reproduces the characteristics or basic behaviours of the entity itself.

If correct, modelling allows ***predicting the future*** behaviour of the studied object.

If fully “rigorous”, the model is not 'minimally influenced by the expectations or by the subjective interpretation of the observer; thus, the observation and scientific data, underpinning the formulation of theoretical models, are totally “invariant” with respect to the observer.

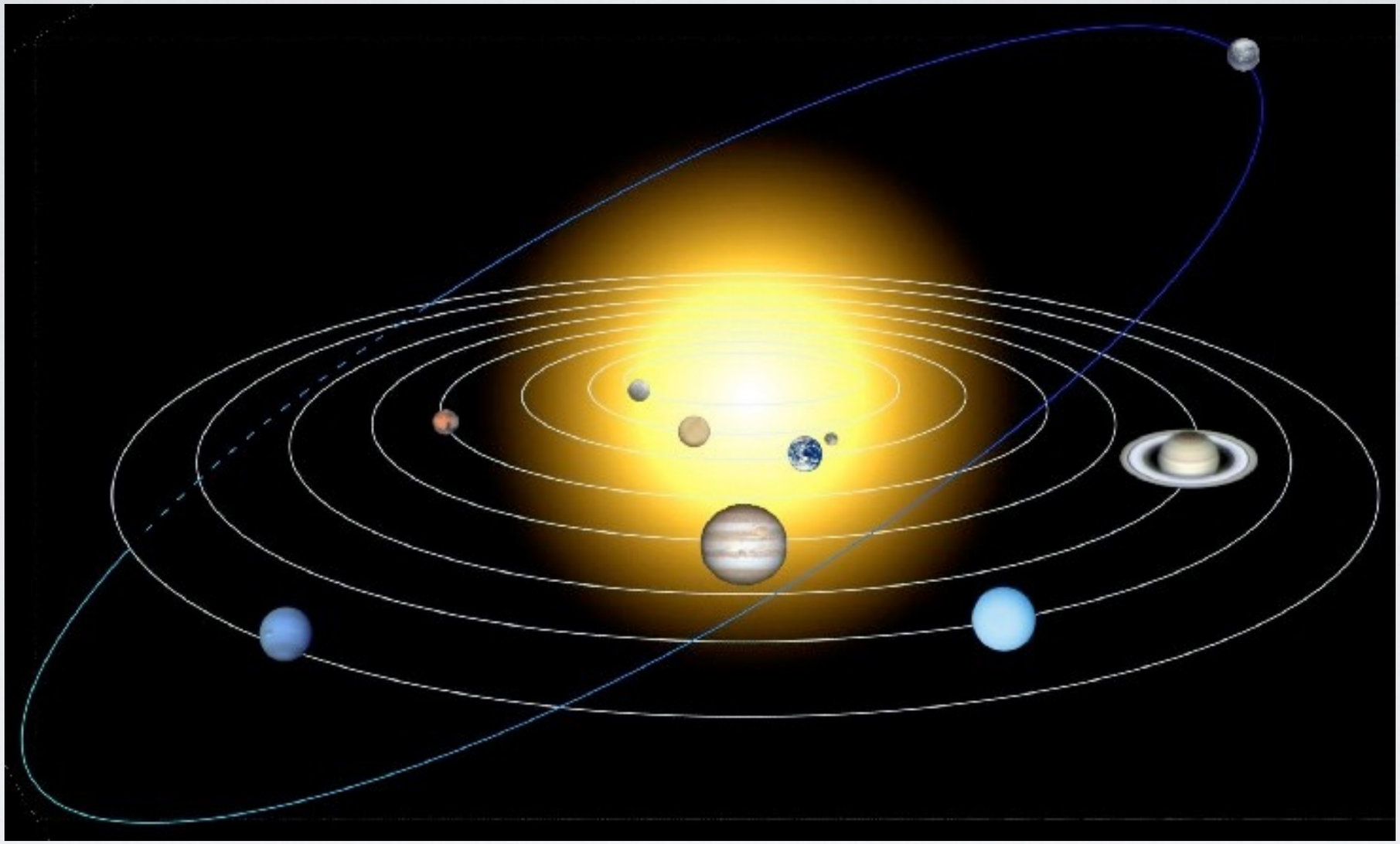
In physics, a model is defined as a conceptual representation (often a simplification) of the real world or a part thereof, capable of explaining and/or predicting its functioning.

In physics, the models based on conceptual assumptions are translated into mathematical formulas.

In mathematics, a model is constructed using the language and tools of mathematics.

Its purpose is to represent as accurately as possible a certain object, a real phenomenon or a physical process from quantitative point of view.

The ultimate goal of mathematical modelling is to be able to study mechanisms and aspects of physical phenomena <u>where the object modelled *is not directly accessible*</u>, so as to predict its future behaviour.



Solar system

Gravitation laws

Predicting planet motion

Modelling in systems biology

"A model is a selective abstraction of reality"

In systems biology, the modelling phase, which follows the experimental phase, includes all theoretical and applied mathematical treatments and computational techniques used to **abstract, represent and simulate** the behaviour of complex cellular and/or molecular systems, in order to construct mathematical models capable of identifying the emergent properties of complex biological systems and to predict their future behaviour.

Static models **Dynamic models**

Static models: the graph-network abstraction and the topological properties of biological networks

Examples of complex biological systems possibly abstracted as physical networks:

- Central nervous system
- Immune system
- Metabolism
- Endocrine system
- Signaling mechanisms
- Ecosystems
- WWW

Central Nervous System =
set of physically connected neurons

Endocrine System =
set of functionally related cells and hormones

Immune system =
collection of cells and molecules physically and functionally connected

Signal transduction system =
set of intracellular molecules physically and/or functionally linked devoted to information processing

Metabolism =
set of intracellular molecules physically and/or functionally linked devoted to mass/energy processing

Ecosystems =
sets of predator-prey couples leading to speciation on the basis of Darwinian selection

World Wide Web =
collection of connected computers processing information

What do they have in common?

Generation of *architectural* system-level properties

↓

Display global cause-effect relationships
(possible triggering of butterfly effects)

↓

Analysis needs creation of topological mathematical model

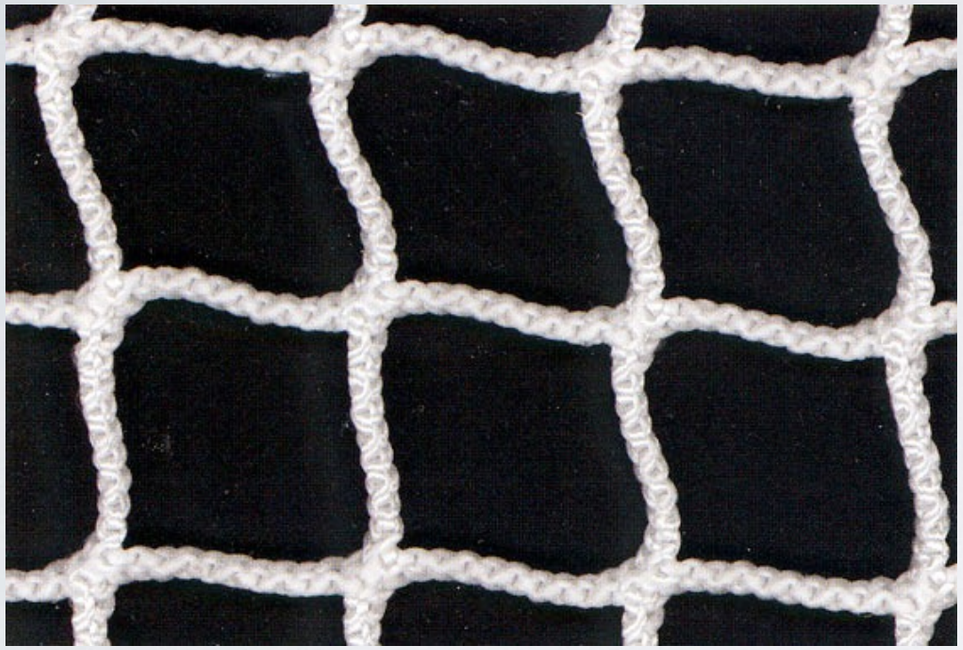
To build a static mathematical model we need the right abstraction!

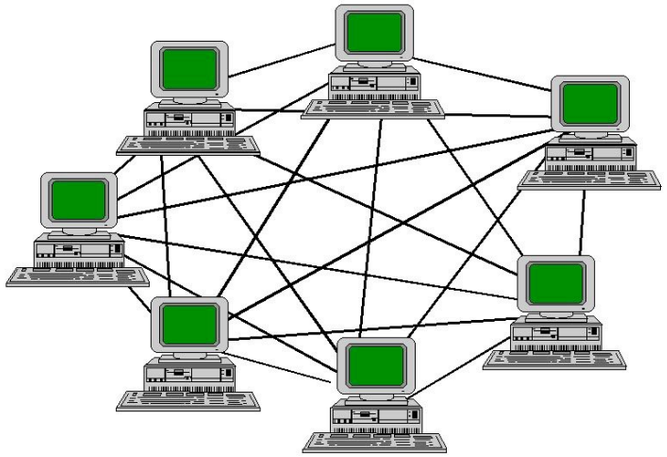
Abstraction =

mental process by which you replace a set of distinct objects with a more general concept, describing the objects based on their <u>common</u> properties

The abstraction must allow to recapitulate the essence of the phenomenon: it must “train the complexity”.

Starting from “real” physical networks





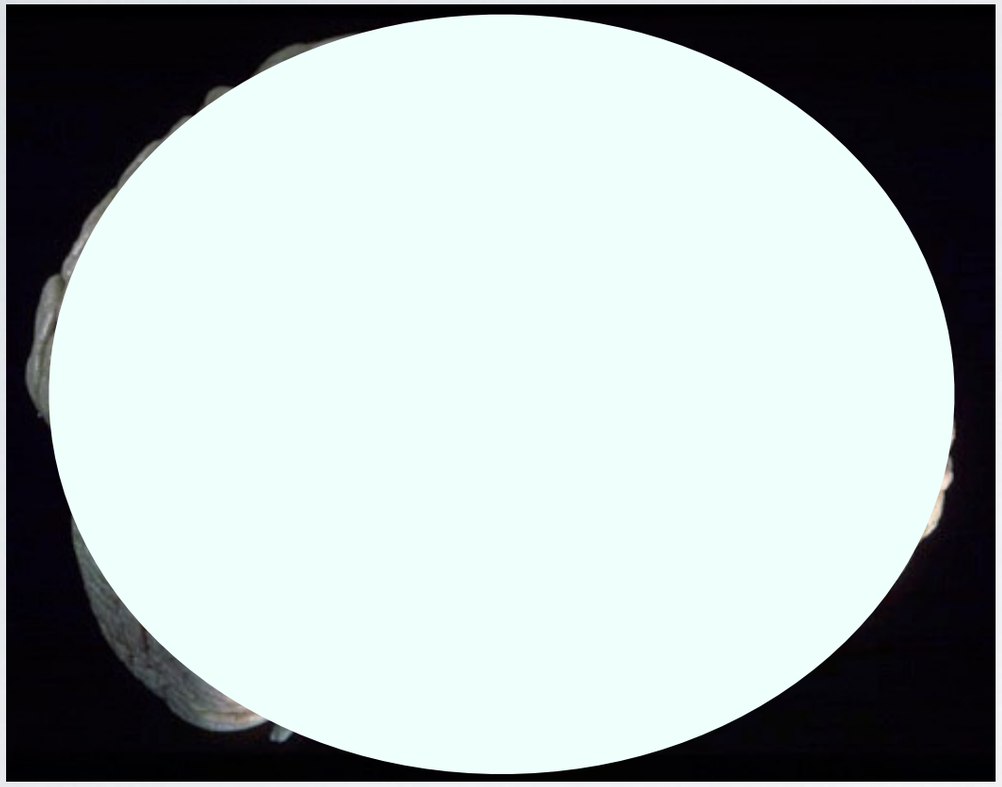
Instinctively ...

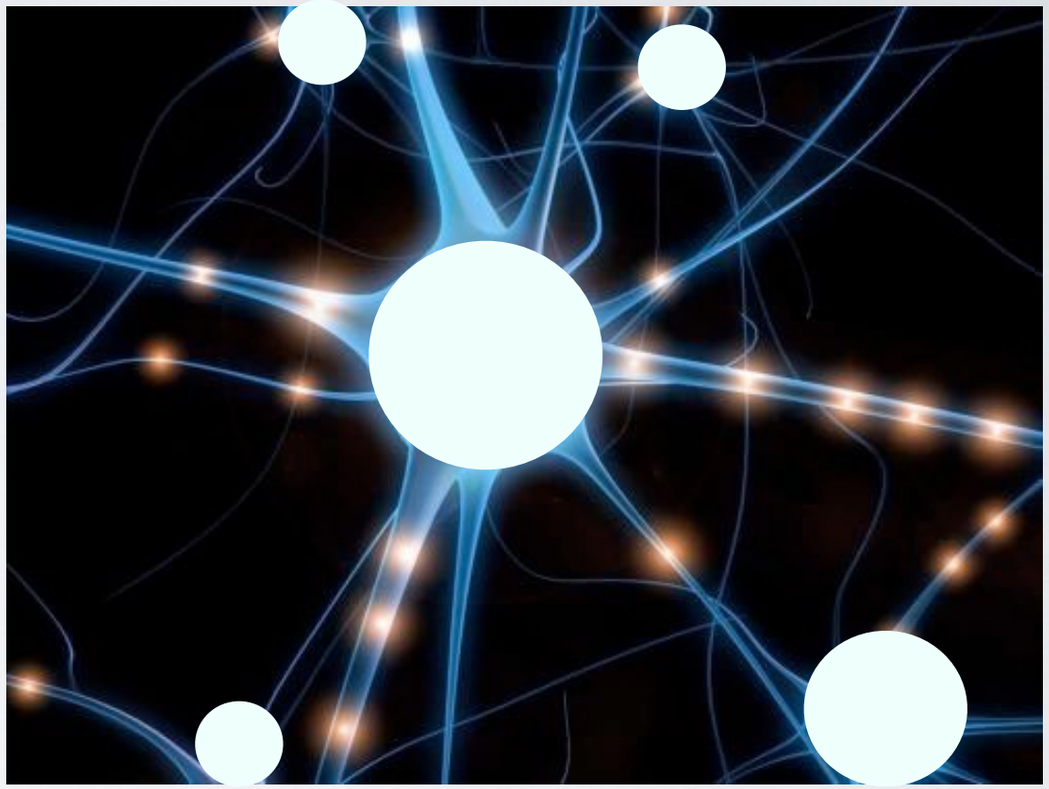
our mind develops a conceptual image

which allows to represent interacting objects as a network of relationships

How to abstract a conceptual image of a network of relationships?

Example:
how to abstract the central nervous system?





Real objects are represented
as 2D geometric shapes

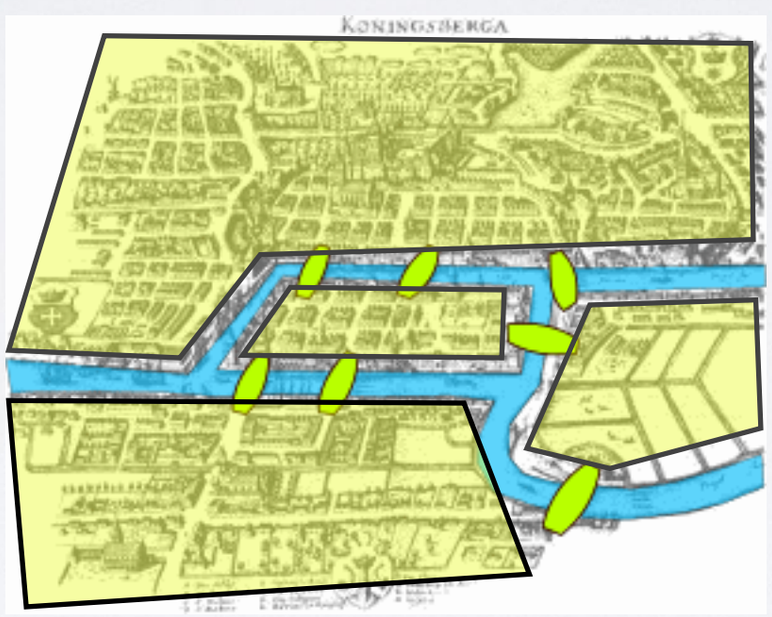
Symbolic representation of a set of objects on a 2D plane

Is it satisfactory?

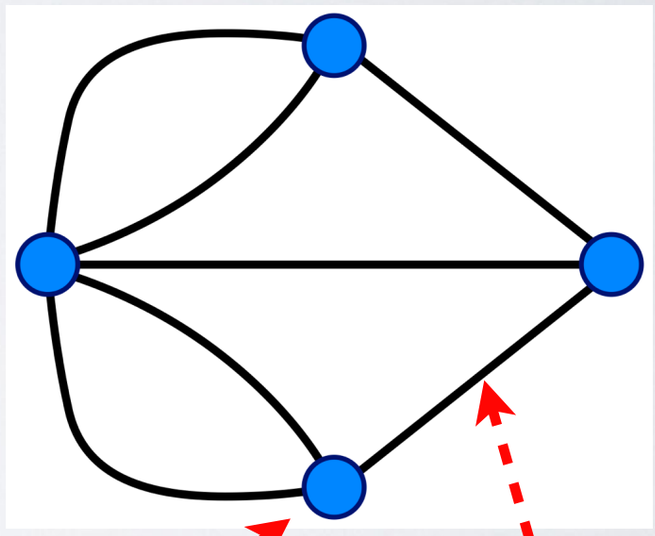
We must represent not only the objects
BUT
also, and especially, their relationships!

Euler (1707 - 1783)

Konigsberg bridges



The abstraction:
the graph

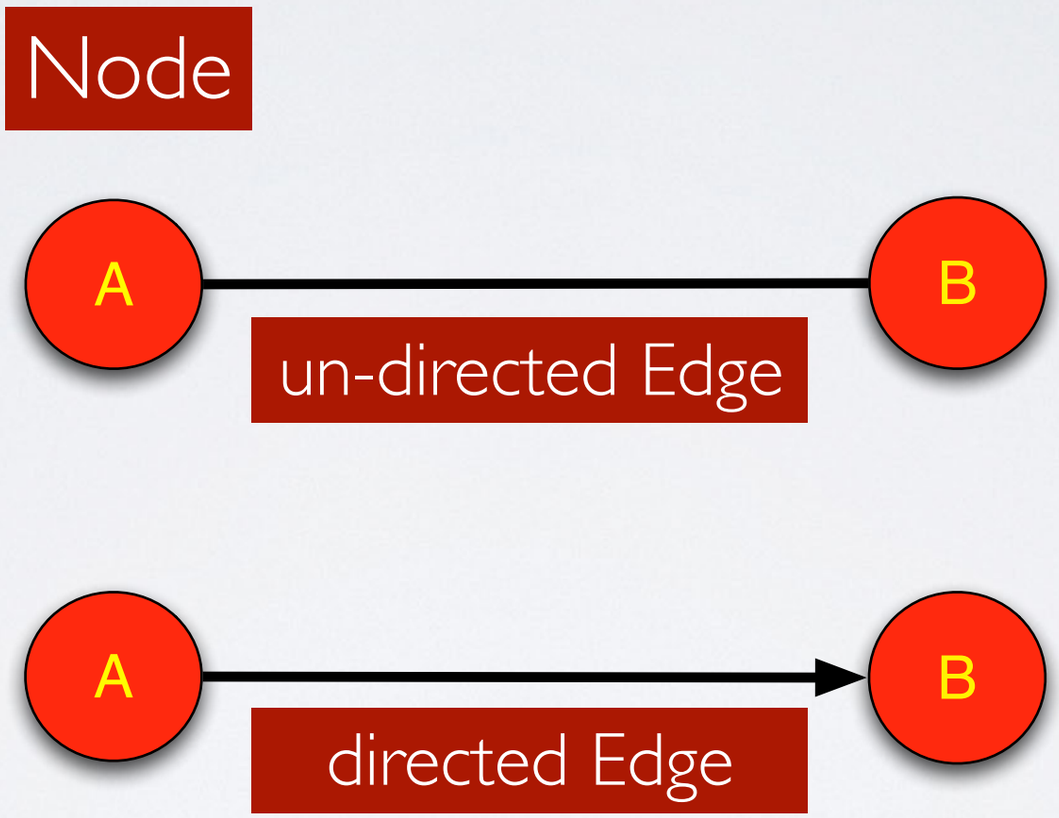


Konigsberg area

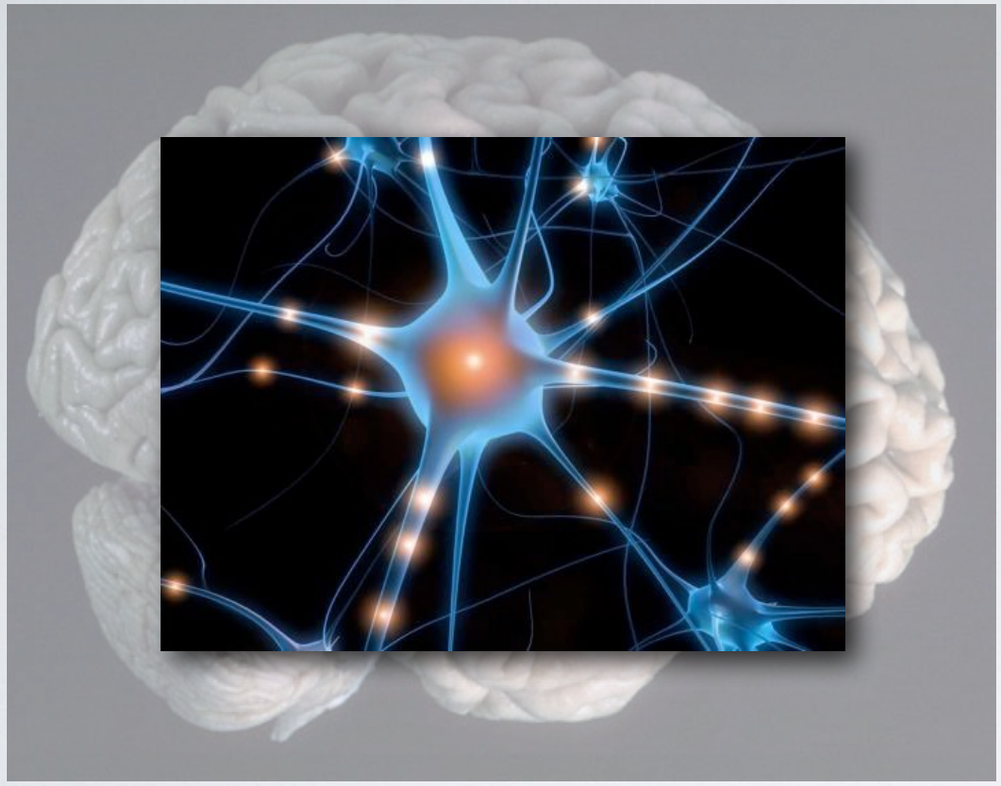
Konigsberg bridge

Euler's abstraction

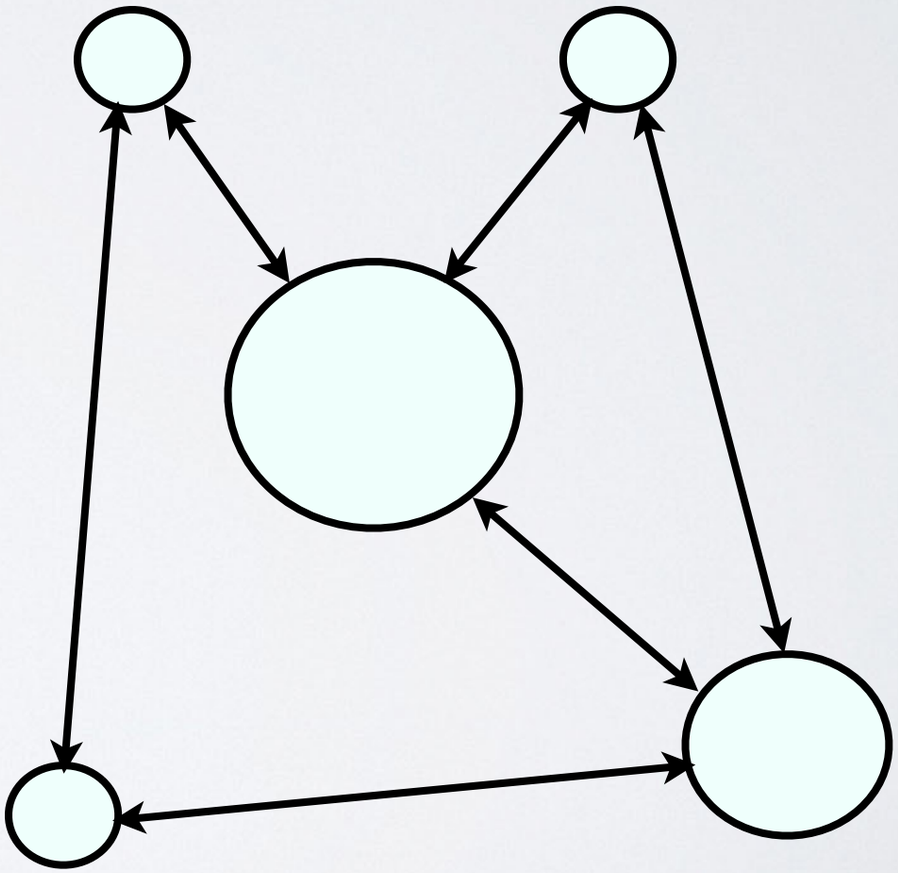
Fundamental BINARY elements in graphs



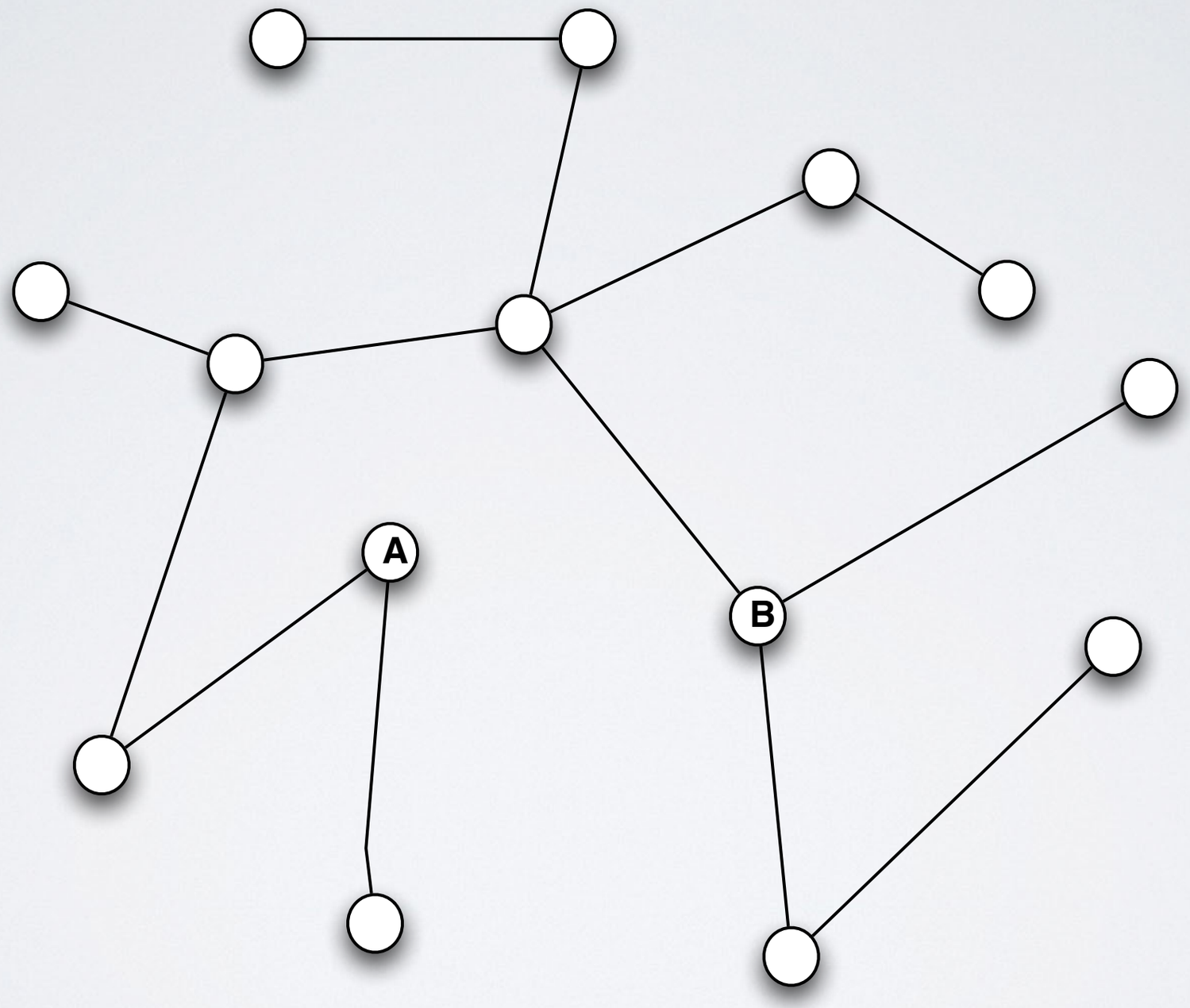
"Abstract" representation of any type of complex system!



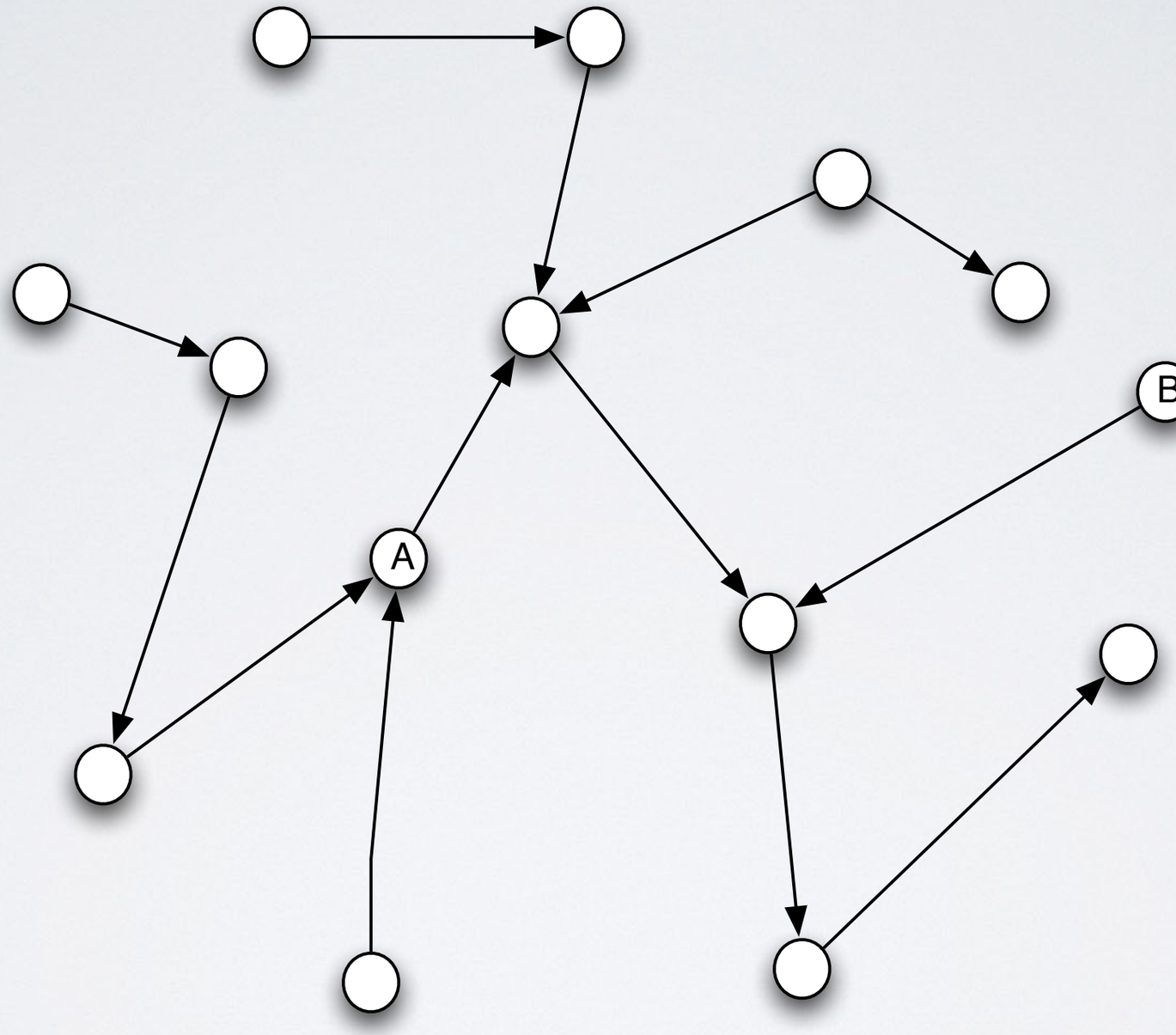
OK!



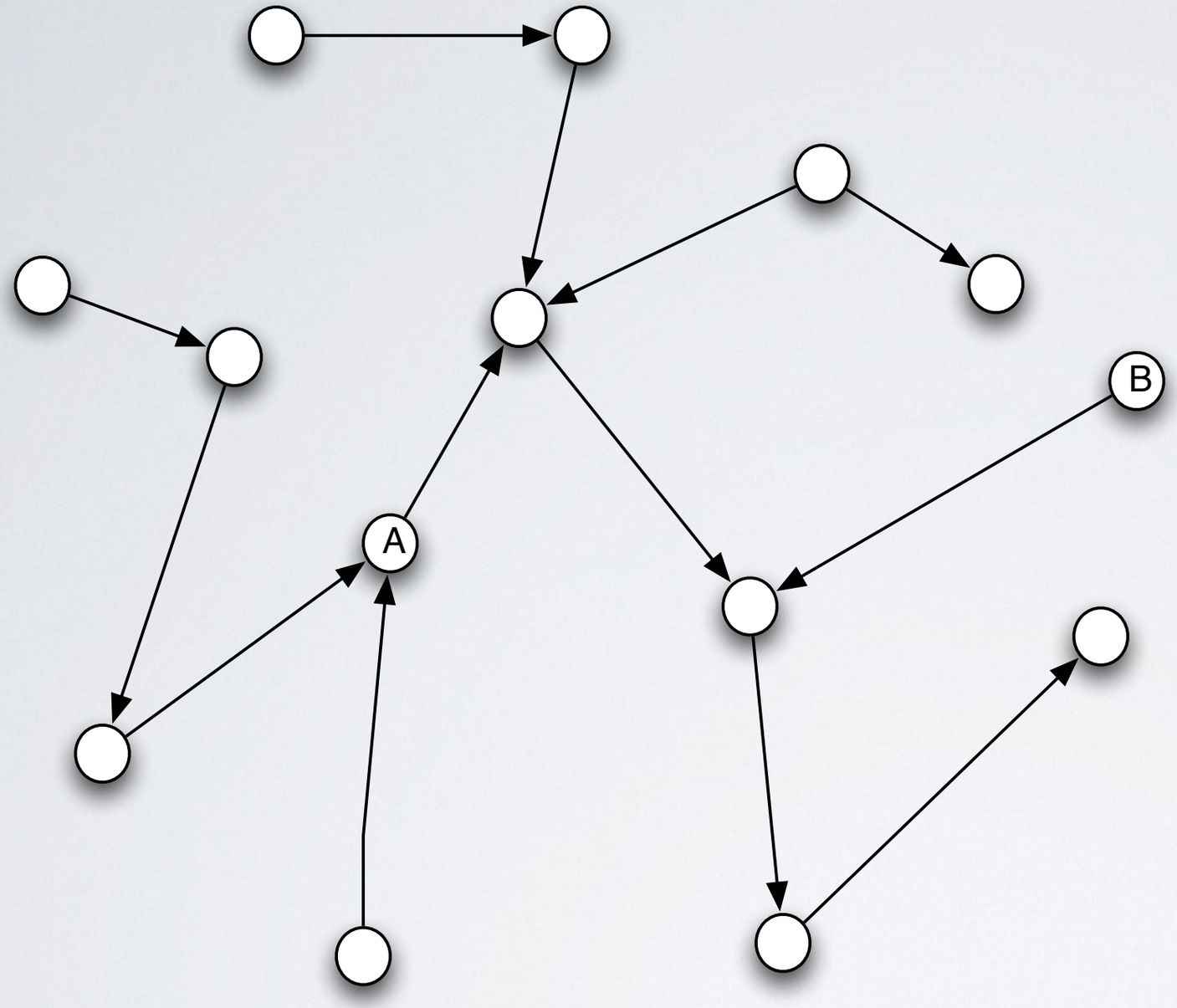
“Undirected” graph



“Directed” graph



Euler’s abstraction: the graph theory



- ➡ atom,
- ➡ molecule,
- ➡ cell,
- ➡ living being,
- ➡ ecosystem,
- ➡ bank,
- ➡ computer,
- ➡ *etc*.

Why to abstract a complex system as a graph is so useful?

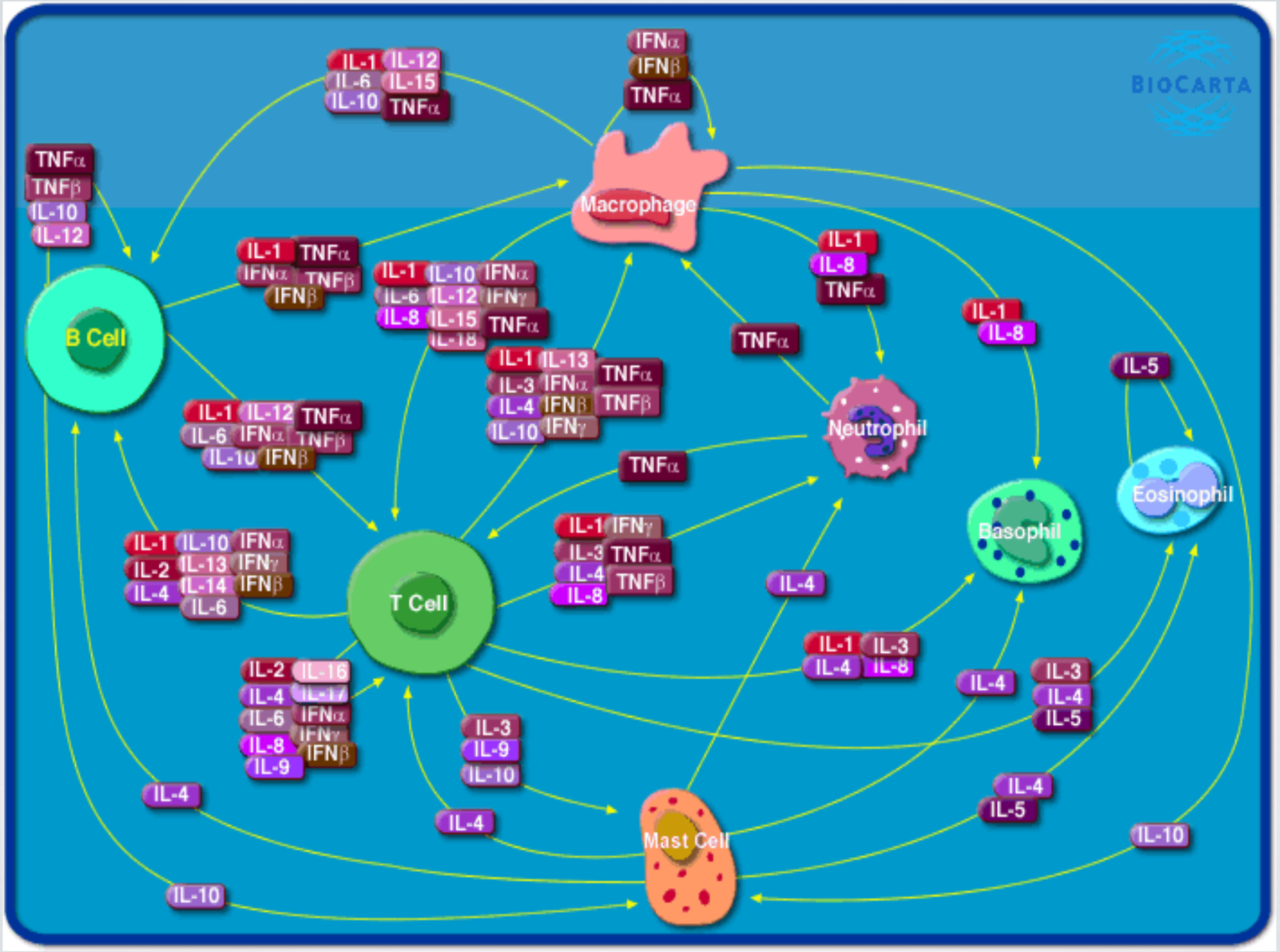
Concept universalization

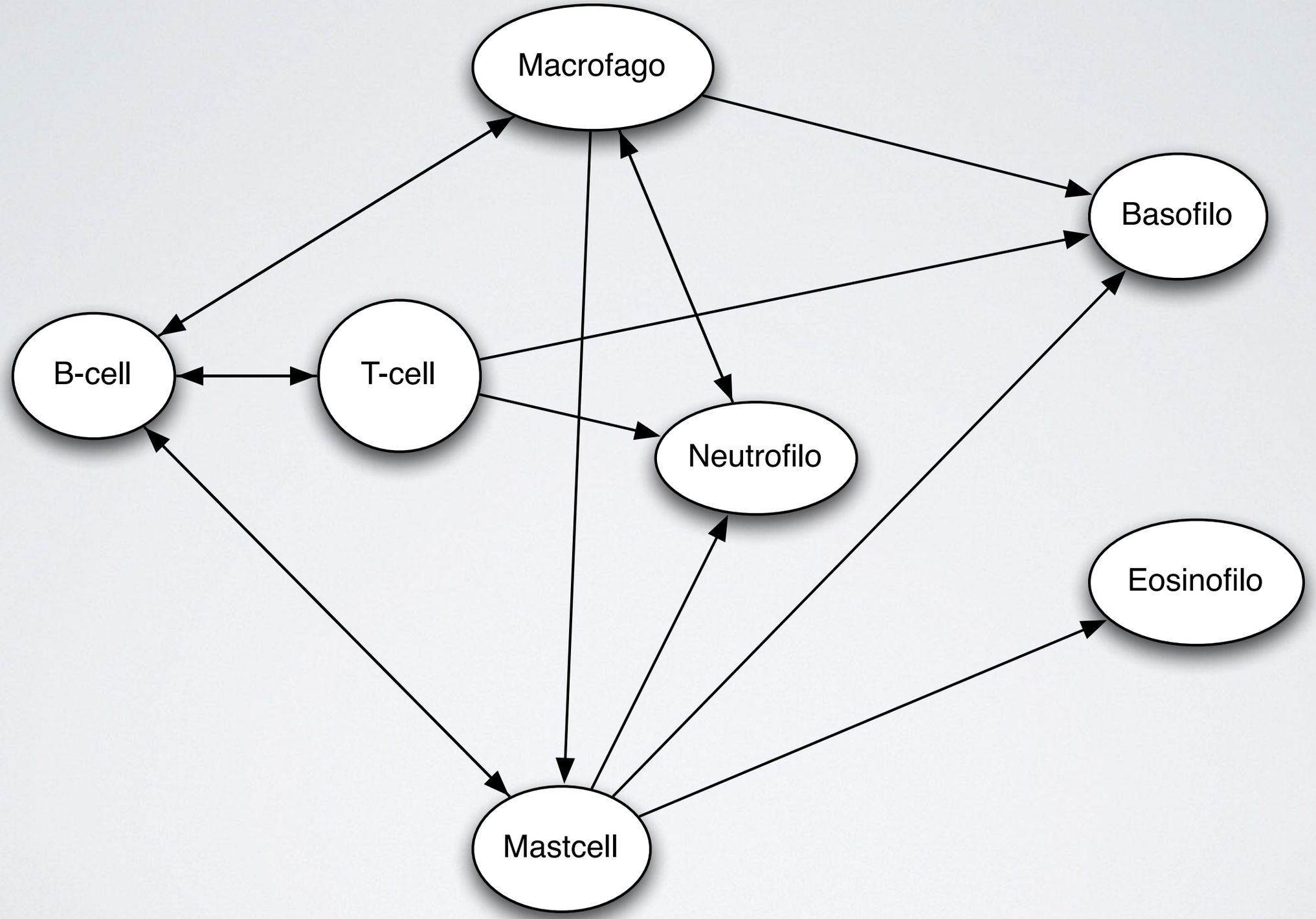
- Complexity simplification (training): a network of physical objects becomes a graph on 2D plane (planar geometry)

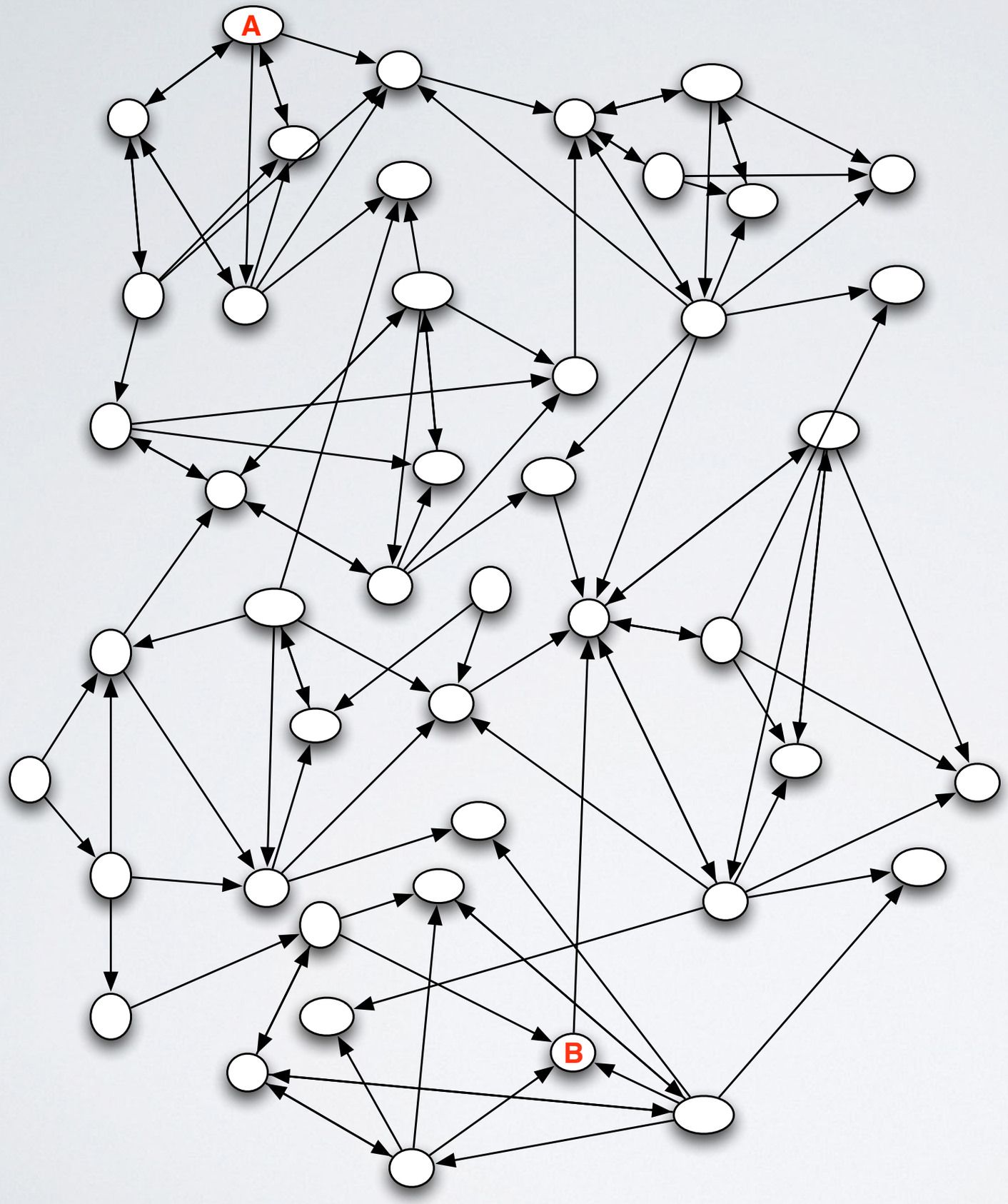
Quantification of network structures:

- static models building
- quantification of architectural properties
- functional prediction

Network of cells and cytokines in the immune system




Macrofago
Basofilo
B-cell
T-cell
Neutrofilo
Eosinofilo
Mastcell

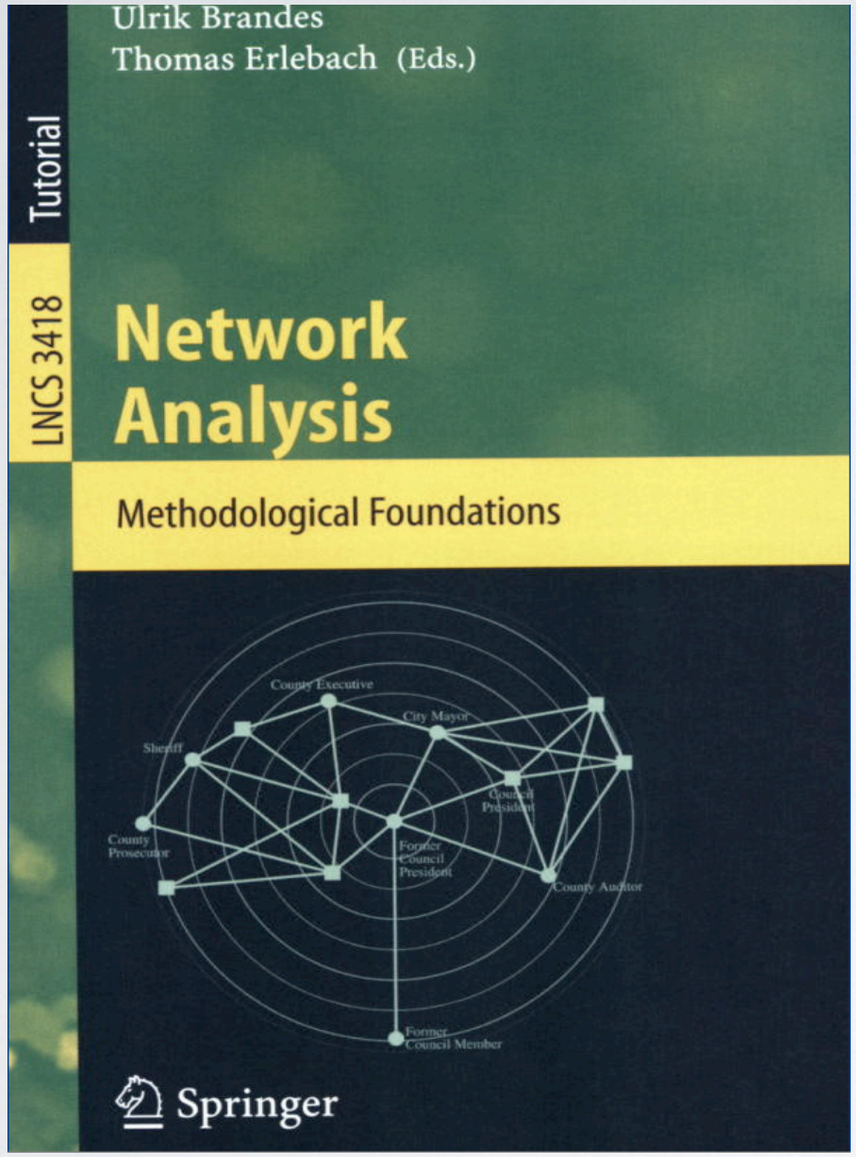


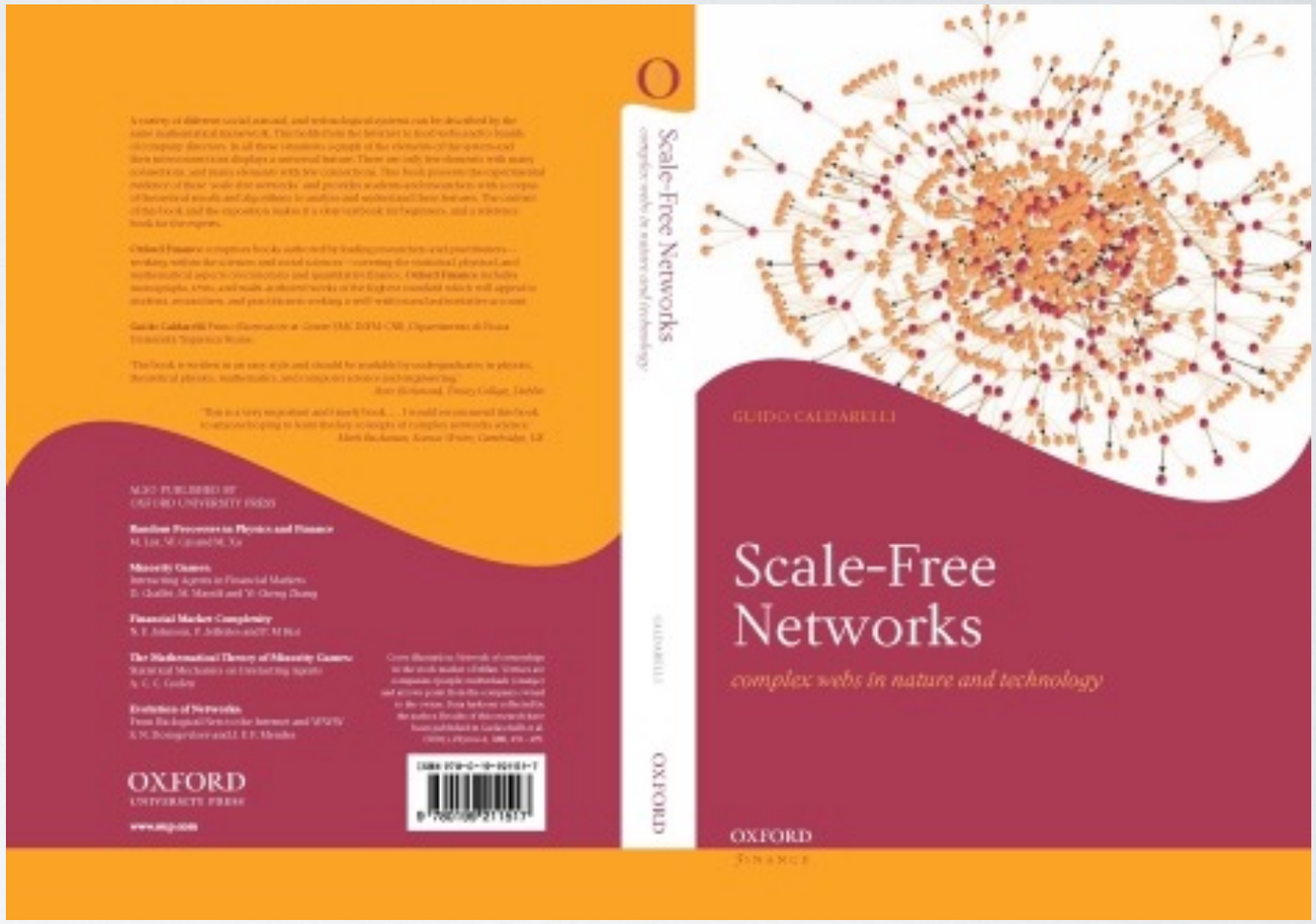
And now?
What to do?

Graph theory: the "TOPOLOGICAL" analysis of networks

Topology or study of the places (from greek τοπος = place, and λογος = study) is one of the most important branches of modern mathematics and physics.

It concerns the study of the properties of *figures and forms* that do not change when a deformation is applied *without* any "tear", "overlapping" or "gluing".


Ulrik Brandes
Thomas Erlebach (Eds.)
Tutorial
LNCS 3418
Network Analysis
Methodological Foundations
County Executive
City Mayor
Sheriff
Council President
County Prosecutor
Former Council President
County Auditor
Former Council Member
Springer


Scale-Free Networks
complex webs in nature and technology
GUIDO CALDARELLI
OXFORD
OXFORD
UNIVERSITY PRESS
www.oup.com
Scale-Free Networks
complex webs in nature and technology
OXFORD
SCIENCE

NATURE REVIEWS GENETICS,
VOLUME 5, FEBRUARY 2004, pp. 101-113

NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION

*Albert-László Barabási** & *Zoltán N. Oltvai*‡

A key aim of postgenomic biomedical research is to systematically catalogue all molecules and their interactions within a living cell. There is a clear need to understand how these molecules and the interactions between them determine the function of this enormously complex machinery, both in isolation and when surrounded by other cells. Rapid advances in network biology indicate that cellular networks are governed by universal laws and offer a new conceptual framework that could potentially revolutionize our view of biology and disease pathologies in the twenty-first century.

PROTEIN CHIPS
Similar to cDNA microarrays, this evolving technology involves arraying a genomic set of proteins on a solid surface without denaturing them. The proteins are arrayed at a high enough density for the detection of activity, binding to lipids and so on.

**Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA.*
‡*Department of Pathology, Northwestern University, Chicago, Illinois 60611, USA.*
e-mails: alb@nd.edu; zno008@northwestern.edu
doi:10.1038/nrg1272

Reductionism, which has dominated biological research for over a century, has provided a wealth of knowledge about individual cellular components and their functions. Despite its enormous success, it is increasingly clear that a discrete biological function can only rarely be attributed to an individual molecule. Instead, most biological characteristics arise from complex interactions between the cell's numerous constituents, such as proteins, DNA, RNA and small molecules[1–8]. Therefore, a key challenge for biology in the twenty-first century is to understand the structure and the dynamics of the complex intercellular web of interactions that contribute to the structure and function of a living cell.

The development of high-throughput data-collection techniques, as epitomized by the widespread use of microarrays, allows for the simultaneous interrogation of the status of a cell's components at any given time. In turn, new technology platforms, such as PROTEIN CHIPS or semi-automated YEAST TWO-HYBRID SCREENS, help to determine how and when these molecules interact with each other. Various types of interaction webs, or networks, (including protein–protein interaction, metabolic, signalling and transcription-regulatory networks) emerge from the sum of these interactions. None of these networks are independent, instead they form a 'network of networks' that is responsible for the behaviour of the cell. A major challenge of contemporary biology is to embark on an integrated theoretical and experimental programme to map out, understand and model in quantifiable terms the topological and dynamic properties of the various networks that control the behaviour of the cell.

Help along the way is provided by the rapidly developing theory of complex networks that, in the past few years, has made advances towards uncovering the organizing principles that govern the formation and evolution of various complex technological and social networks[9–12]. This research is already making an impact on cell biology. It has led to the realization that the architectural features of molecular interaction networks within a cell are shared to a large degree by other complex systems, such as the Internet, computer chips and society. This unexpected universality indicates that similar laws may govern most complex networks in nature, which allows the expertise from large and well-mapped non-biological systems to be used to characterize the intricate interwoven relationships that govern cellular functions.

In this review, we show that the quantifiable tools of network theory offer unforeseen possibilities to understand the cell's internal organization and evolution, fundamentally altering our view of cell biology. The emerging results are forcing the realization that, notwithstanding the importance of individual molecules, cellular function is a contextual attribute of strict and quantifiable patterns of interactions between the myriad of cellular constituents. Although uncovering the generic organizing principles of cellular networks

Cell

Leading Edge
Review

Interactome Networks and Human Disease

Marc Vidal,[1,2,*] Michael E. Cusick,[1,2] and Albert-László Barabási[1,3,4,*]
[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[3]Center for Complex Network Research (CCNR) and Departments of Physics, Biology and Computer Science, Northeastern University, Boston, MA 02115, USA
[4]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: marc_vidal@dfci.harvard.edu (M.V.), alb@neu.edu (A.-L.B.)
DOI 10.1016/j.cell.2011.02.016

Complex biological systems and cellular networks may underlie most genotype to phenotype relationships. Here, we review basic concepts in network biology, discussing different types of interactome networks and the insights that can come from analyzing them. We elaborate on why interactome networks are important to consider in biology, how they can be mapped and integrated with each other, what global properties are starting to emerge from interactome network models, and how these properties may relate to human disease.

Introduction

Since the advent of molecular biology, considerable progress has been made in the quest to understand the mechanisms that underlie human disease, particularly for genetically inherited disorders. Genotype-phenotype relationships, as summarized in the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al., 2009), include mutations in more than 3000 human genes known to be associated with one or more of over 2000 human disorders. This is a truly astounding number of genotype-phenotype relationships considering that a mere three decades have passed since the initial description of Restriction Fragment Length Polymorphisms (RFLPs) as molecular markers to map genetic loci of interest (Botstein et al., 1980), only two decades since the announcement of the first positional cloning experiments of disease-associated genes using RFLPs (Amberger et al., 2009), and just one decade since the release of the first reference sequences of the human genome (Lander et al., 2001; Venter et al., 2001). For complex traits, the information gathered by recent genome-wide association studies suggests high-confidence genotype-phenotype associations between close to 1000 genomic loci and one or more of over one hundred diseases, including diabetes, obesity, Crohn's disease, and hypertension (Altshuler et al., 2008). The discovery of genomic variations involved in cancer, inherited in the germline or acquired somatically, is equally striking, with hundreds of human genes found linked to cancer (Stratton et al., 2009). In light of new powerful technological developments such as next-generation sequencing, it is easily imaginable that a catalog of nearly all human genomic variations, whether deleterious, advantageous, or neutral, will be available within our lifetime.

Despite the natural excitement emerging from such a huge body of information, daunting challenges remain. Practically, the genomic revolution has, thus far, seldom translated directly into the development of new therapeutic strategies, and the mechanisms underlying genotype-phenotype relationships remain only partially explained. Assuming that, with time, most human genotypic variations will be described together with phenotypic associations, there would still be major problems to fully understand and model human genetic variations and their impact on diseases.

To understand why, consider the "one-gene/one-enzyme/one-function" concept originally framed by Beadle and Tatum (Beadle and Tatum, 1941), which holds that simple, linear connections are expected between the genotype of an organism and its phenotype. But the reality is that most genotype-phenotype relationships arise from a much higher underlying complexity. Combinations of identical genotypes and nearly identical environments do not always give rise to identical phenotypes. The very coining of the words "genotype" and "phenotype" by Johannsen more than a century ago derived from observations that inbred isogenic lines of bean plants grown in well-controlled environments give rise to pods of different size (Johannsen, 1909). Identical twins, although strikingly similar, nevertheless often exhibit many differences (Raser and O'Shea, 2005). Likewise, genotypically indistinguishable bacterial or yeast cells grown side by side can express different subsets of transcripts and gene products at any given moment (Elowitz et al., 2002; Blake et al., 2003; Taniguchi et al., 2010). Even straightforward Mendelian traits are not immune to complex genotype-phenotype relationships. Incomplete penetrance, variable expressivity, differences in age of onset, and modifier mutations are more frequent than generally appreciated (Perlis et al., 2010).

We, along with others, argue that the way beyond these challenges is to decipher the properties of biological systems, and in particular, those of molecular networks taking place within cells. As is becoming increasingly clear, biological systems and cellular networks are governed by specific laws and principles, the understanding of which will be essential for a deeper comprehension of biology (Nurse, 2003; Vidal, 2009).

Accordingly, our goal is to review key aspects of how complex systems operate inside cells. Particularly, we will review how by interacting with each other, genes and their products form complex networks within cells. Empirically determining and modeling cellular networks for a few model organisms and for

Network medicine: a network-based approach to human disease

Albert-László Barabási[*‡§], *Natali Gulbahce*[*‡||] *and Joseph Loscalzo*[§]

Abstract | Given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intercellular network that links tissue and organ systems. The emerging tools of network medicine offer a platform to explore systematically not only the molecular complexity of a particular disease, leading to the identification of disease modules and pathways, but also the molecular relationships among apparently distinct (patho)phenotypes. Advances in this direction are essential for identifying new disease genes, for uncovering the biological significance of disease-associated mutations identified by genome-wide association studies and full-genome sequencing, and for identifying drug targets and biomarkers for complex diseases.

*Center for Complex Networks Research and Department of Physics, Northeastern University, 110 Forsyth Street, 111 Dana Research Center, Boston, Massachusetts 02115, USA.
‡Center for Cancer Systems Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA.
§Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, Massachusetts 02115, USA.
||Department of Cellular and Molecular Pharmacology, University of California, 1700 4th Street, Byers Hall 309, Box 2530, San Francisco, California 94158, USA.
Correspondence to A.-L.B.
e-mail: alb@neu.edu
doi:10.1038/nrg2918

Most cellular components exert their functions through interactions with other cellular components, which can be located either in the same cell or across cells, and even across organs. In humans, the potential complexity of the resulting network — the human interactome — is daunting: with ~25,000 protein-coding genes, ~1,000 metabolites and an undefined number of distinct proteins[1] and functional RNA molecules, the number of cellular components that serve as the nodes of the interactome easily exceeds 100,000. The number of functionally relevant interactions between the components of this network, representing the links of the interactome, is expected to be much larger[2].

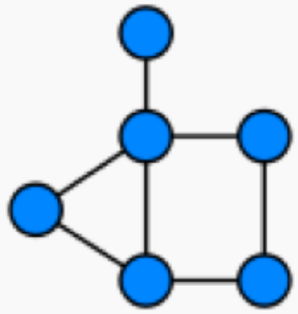
This inter- and intracellular interconnectivity implies that the impact of a specific genetic abnormality is not restricted to the activity of the gene product that carries it, but can spread along the links of the network and alter the activity of gene products that otherwise carry no defects. Therefore, an understanding of a gene's network context is essential in determining the phenotypic impact of defects that affect it[3,4]. Following on from this principle, a key hypothesis underlying this Review is that a disease phenotype is rarely a consequence of an abnormality in a single effector gene product, but reflects various pathobiological processes that interact in a complex network. A corollary of this widely held hypothesis is that the interdependencies among a cell's molecular components lead to deep functional, molecular and causal relationships among apparently distinct phenotypes.

Network-based approaches to human disease have multiple potential biological and clinical applications. A better understanding of the effects of cellular interconnectedness on disease progression may lead to the identification of disease genes and disease pathways, which, in turn, may offer better targets for drug development. These advances may also lead to better and more accurate biomarkers to monitor the functional integrity of networks that are perturbed by diseases as well as to better disease classification. Here we present an overview of the organizing principles that govern cellular networks and the implications of these principles for understanding disease. These principles and the tools and methodologies that are derived from them are facilitating the emergence of a body of knowledge that is increasingly referred to as network medicine[5–7].

The human interactome

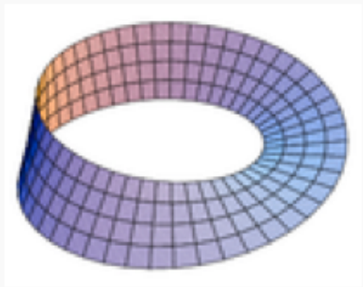
Although much of our understanding of cellular networks is derived from model organisms, the past decade has seen an exceptional growth in human-specific molecular interaction data[8]. Most attention has been directed towards molecular networks, including protein interaction networks, whose nodes are proteins that are linked to each other by physical (binding) interactions[9,10]; metabolic networks, whose nodes are metabolites that are linked if they participate in the same biochemical reactions[11–13]; regulatory networks, whose directed links represent either regulatory relationships between a transcription factor and a gene[14], or post-translational

Topological objects in *n* dimensions

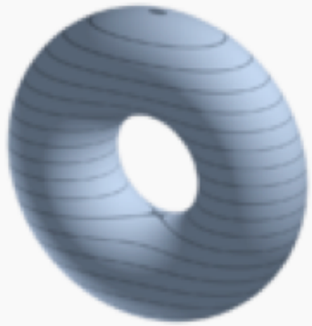


Grafo

Nodo

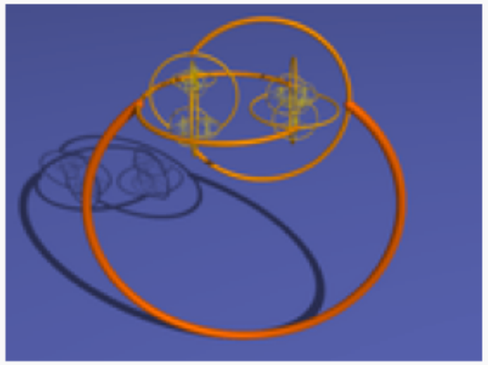


Nastro di Möbius

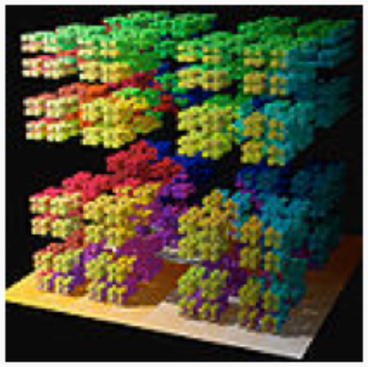


Toro

Fibrazione di Hopf



Sfera cornuta di Alexander

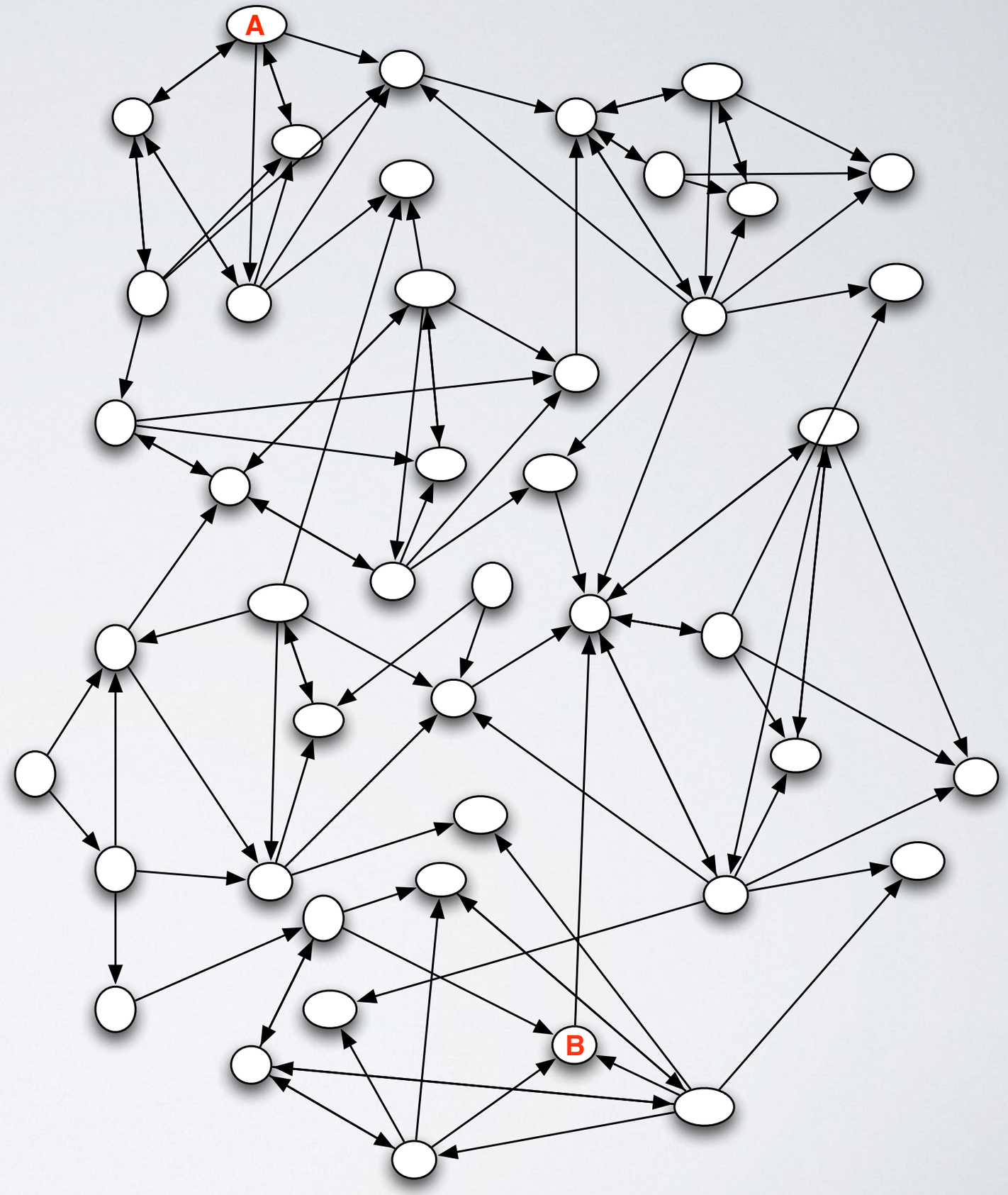


Insieme di Cantor

3-varietà

The topological analysis of the graphs allows ***quantifying*** the architectural properties, global or local, of ***natural networks***.

Creating "static models" of real networks



Since 'a graph” is an abstraction of a real physical network, the properties' of the graph do correspond to the properties of a real physical network.

When we characterize the properties of the graph, we, de facto, characterize the properties of the “abstracted” physical network

Full abstractive correspondence

A GRAPH is just a GRAPH !!

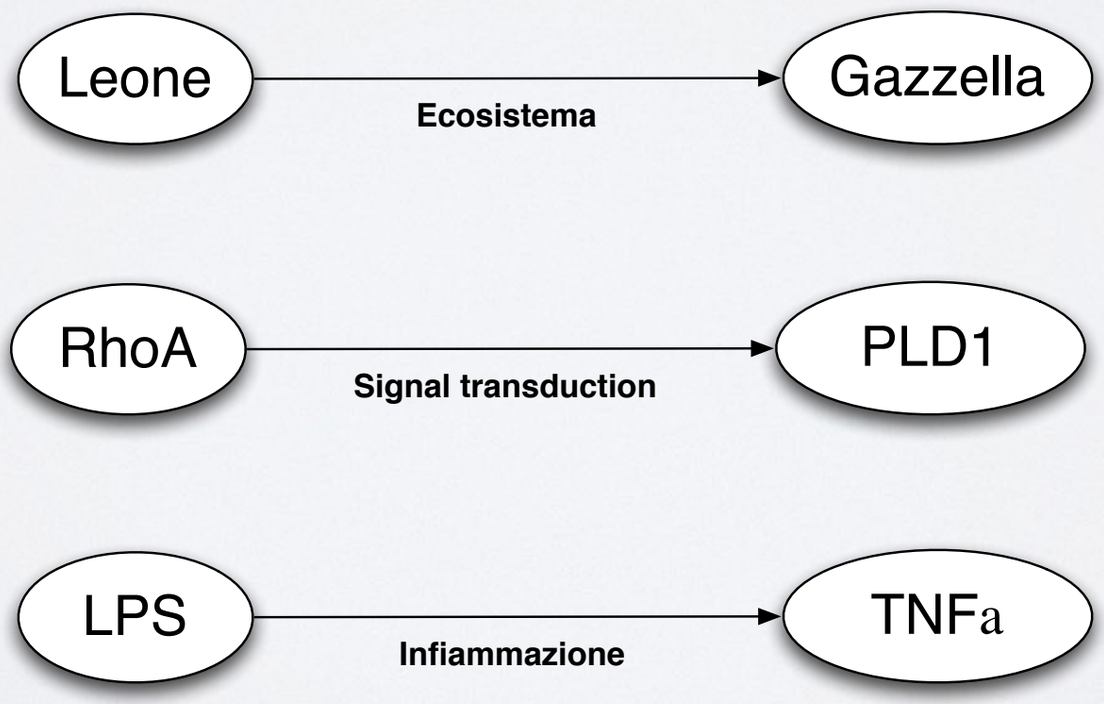
The is, just a drawing on a 2D plane!

The GRAPH mathematical topological properties

assume a *specific contextual meaning*

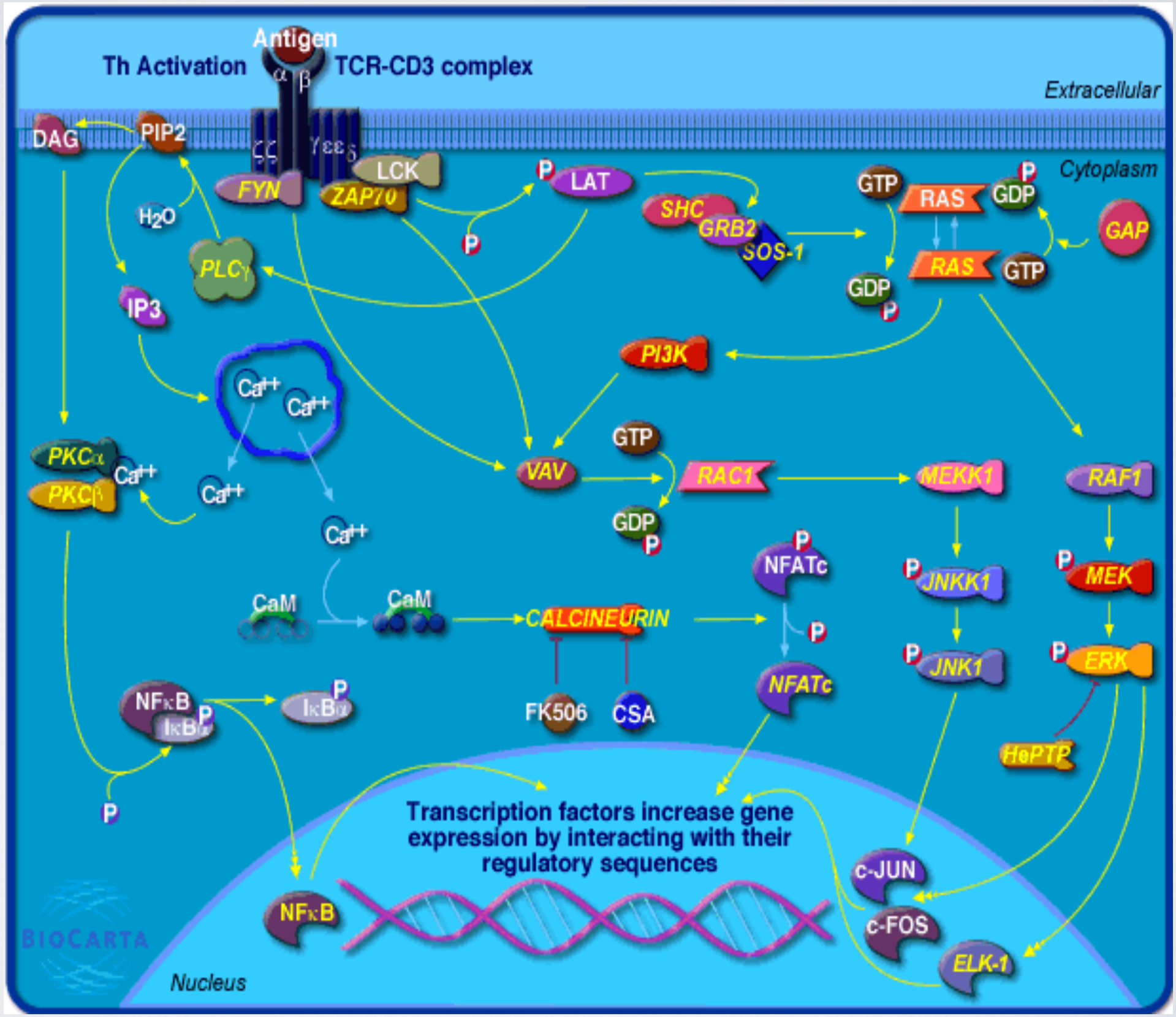
that depends on the "physical" nature of the

network that it abstracts.

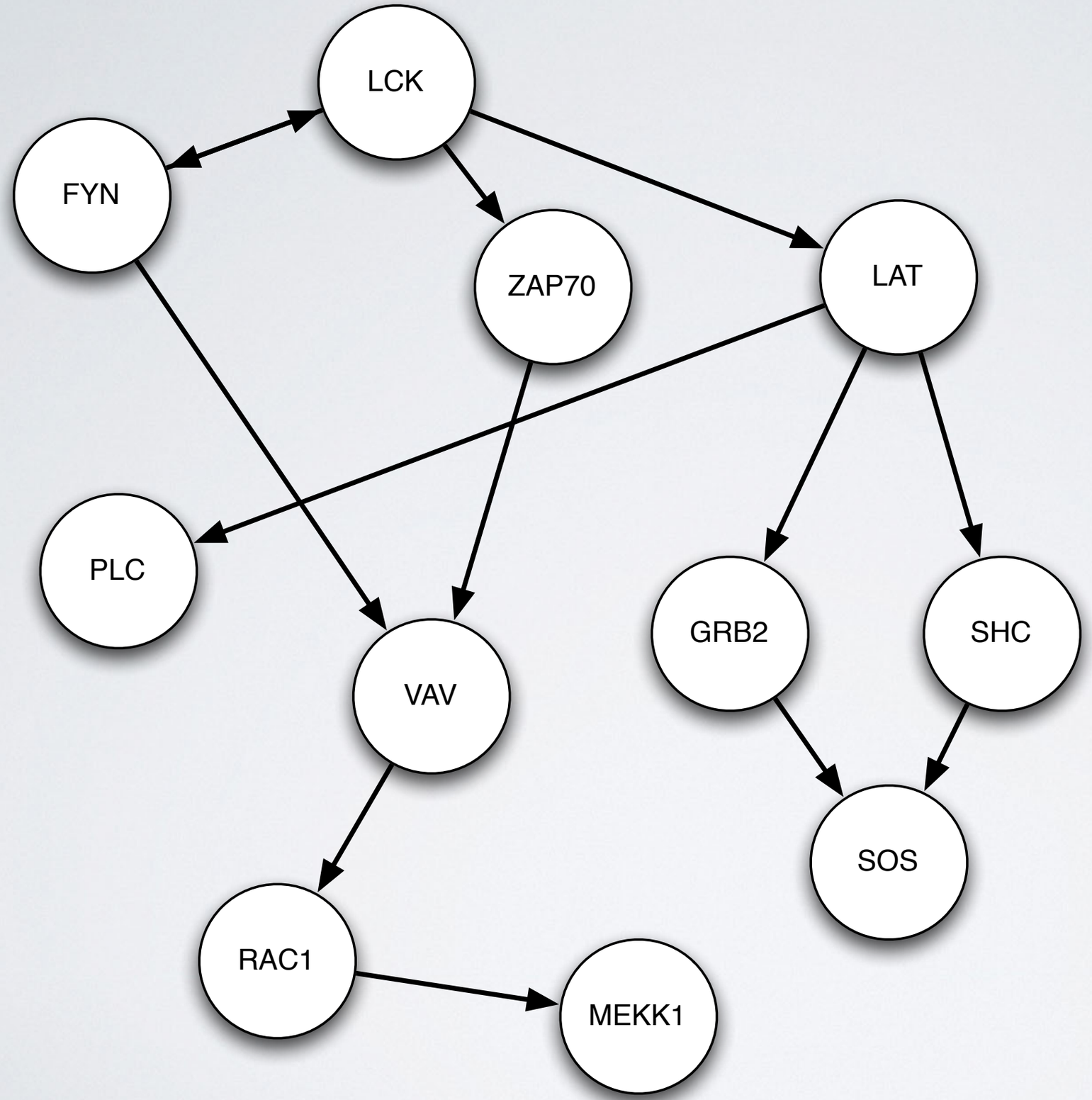


What does it mean to "analyze" a biological network?

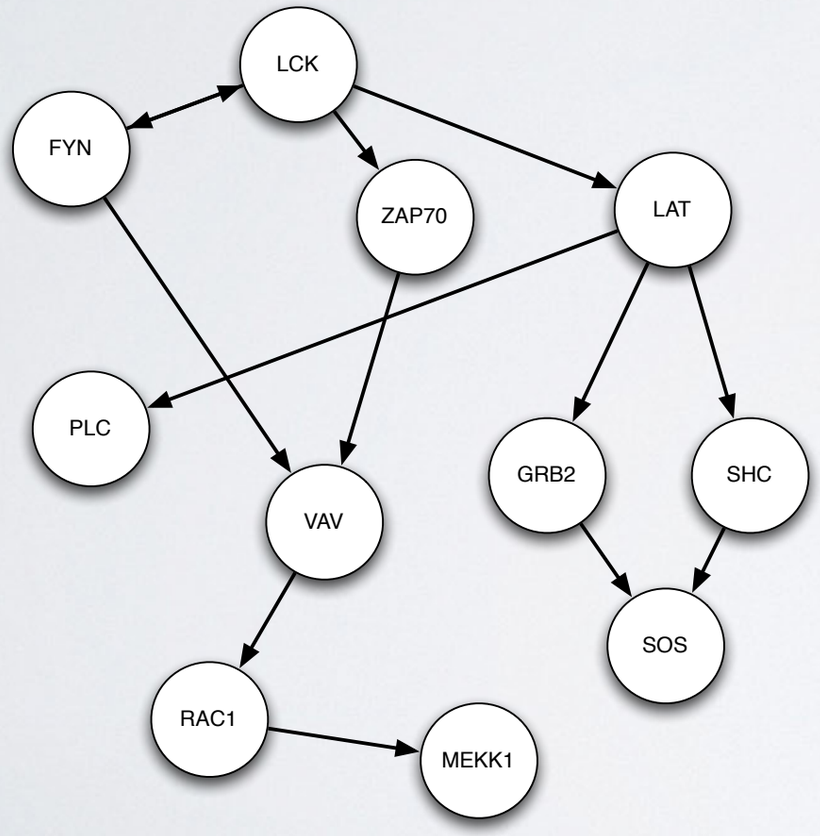
Es.
signal
transduction
by TcR



Step (1) Transforming the physical network into a graph



Step (2)
"Low level" questions
(close to the graph)



- **Which node has more connections?**
- **Which node brings more nodes in connection?**
- **Are there dense aggregates of nodes?**
- **Is the graph fully connected?**
- **Are there nodes with greater <u>importance</u>?**
- **Is there a path that passes through all nodes (Hamiltonian path)?**

Step (3)

“High level” questions

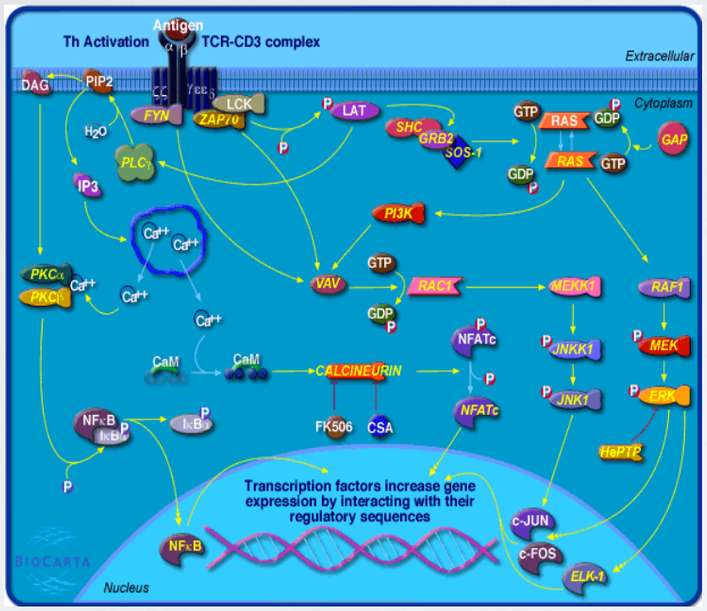
(close to the real network)

- Is the system 'robust? (= tolerant to interferences; f*ault-tolerant*)
- What are the possible regulators / effectors of a molecule?
- Are there cell-type-specific regulatory mechanisms?
- What are the regulatory mechanisms of a certain cellular events?

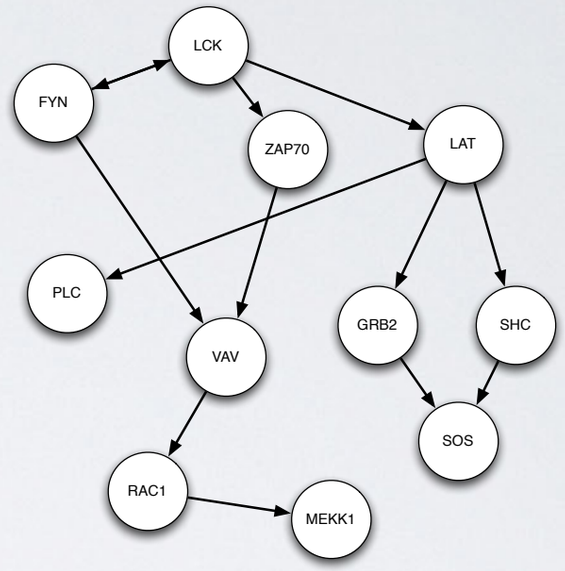
- What are the side effects of a drug?
- What are the mechanisms of action of a drug?
- What are the general molecular alterations during septic shock?
- What the molecular mechanism of resistance to therapy in AML?

Physical networks are "quantified" in relation to the reference degree

(Quantitative parameters in the system of real and rational numbers)



=



(A): ELEMENTS = Centrality Indexes (at least 17) = analysis of individual nodes

(B): GRUPPS = Identification of "node aggregates" (cluster, modules, motifs)

(C): NETWORKS = Global statistical properties

A): ELEMENTS = Centrality indexes

"What are the most important nodes?"
Topologically = Functionally

(B): GROUPS = Identifications of aggregates

"Are there preferential aggregates of nodes?"
"Which nodes preferentially participates to regulatory clusters?"

(C): NETWORKS = Global statistical properties

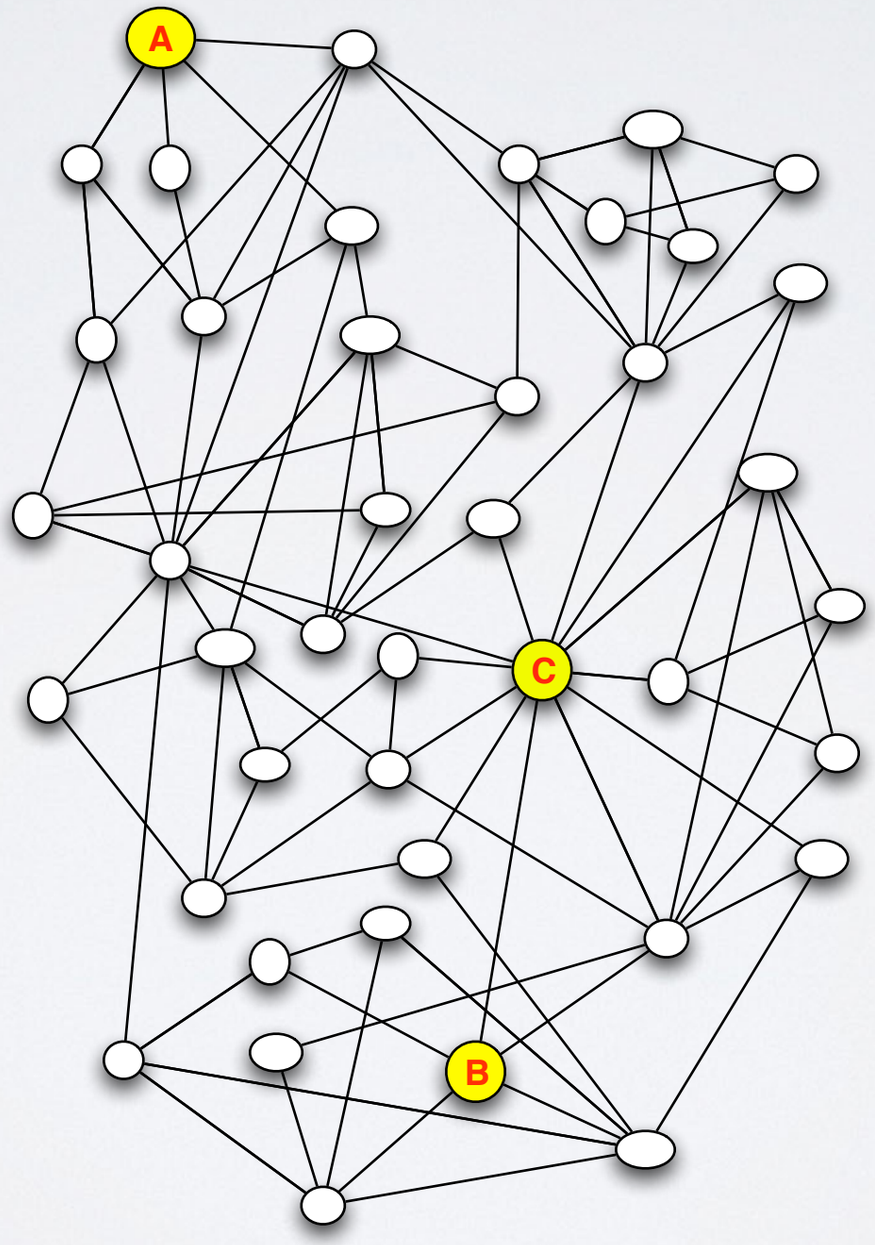
"Is the network stable"
"Is the network robust to interferences?"

"Node degree" *k* = numero di connessioni

A = *k* 4 B = *k* 5 C = *k* 12

"Average degree" <*k*> = numero medio di connessioni/nodo

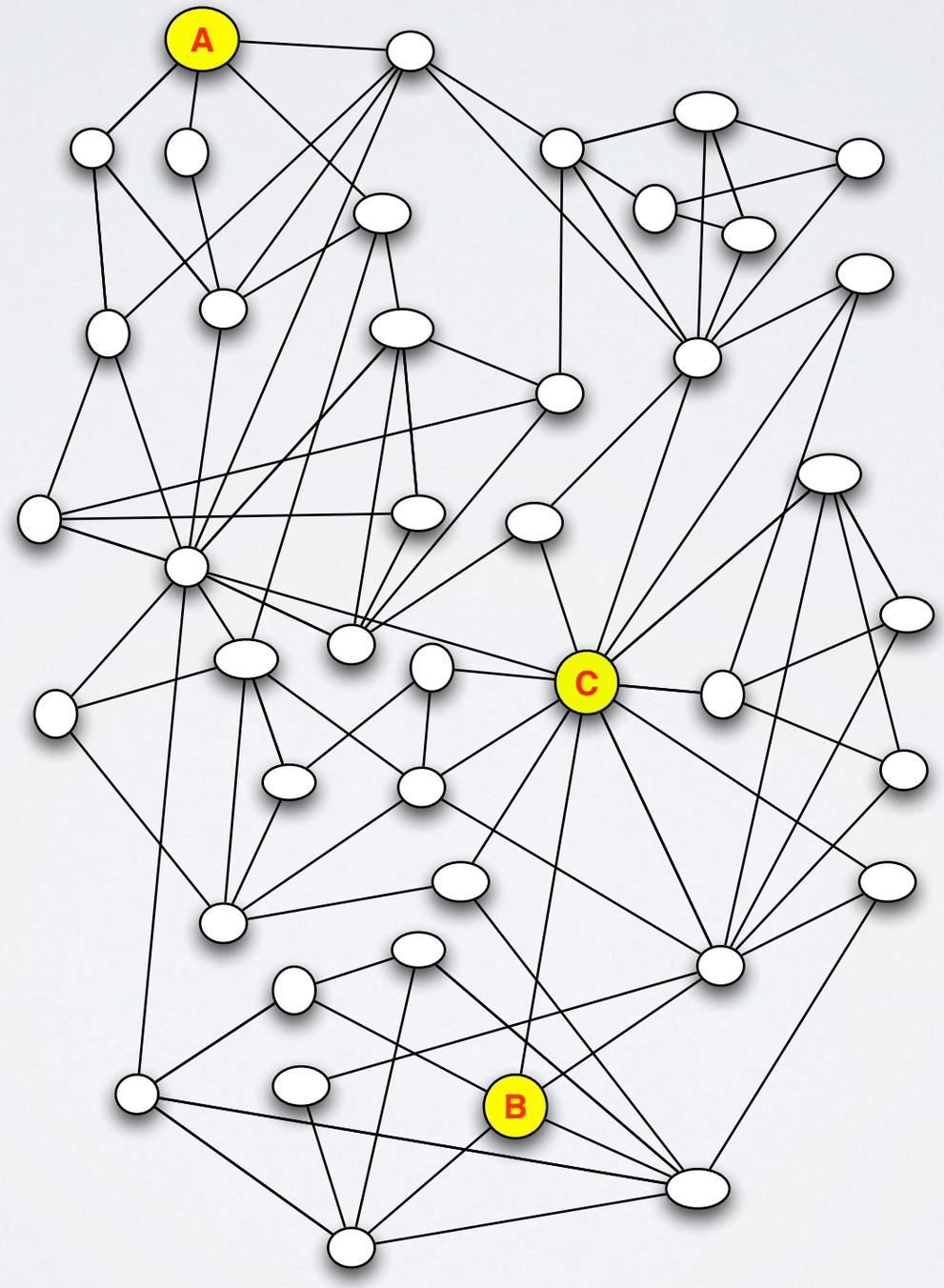
*k*4 + *k*5 + *k*12 = <k> 7



"Node degree distribution" *P(k)* = probabilità che un nodo abbia esattamente *k* connessioni

Probabilità *k* = fra 3 e 5 molto alta (la maggioranza dei nodi)

Probabilità *k* = ≥ 12 molto bassa (solo C)

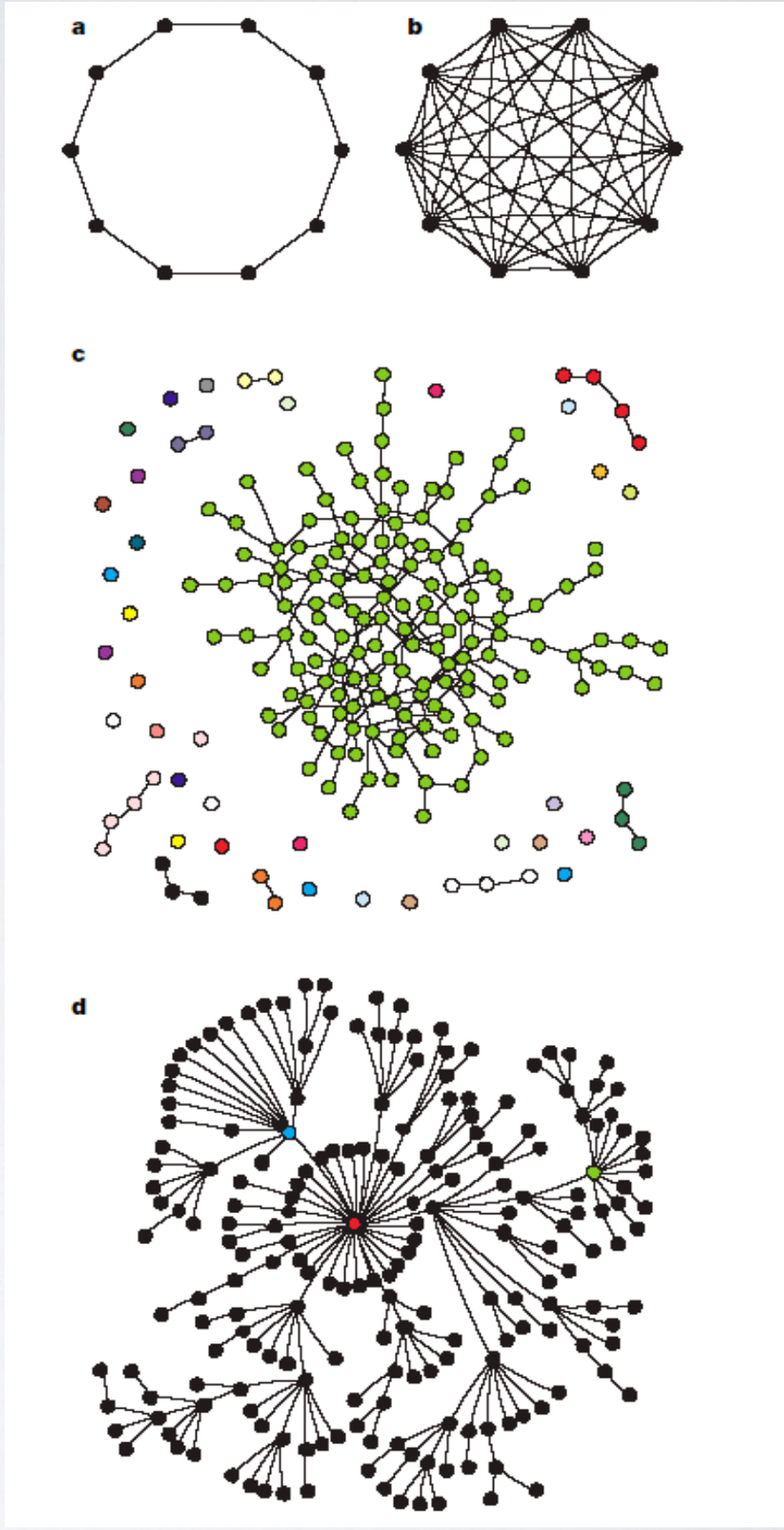


The degree and the calculation of its distribution P(*k*) allows for a global network classification

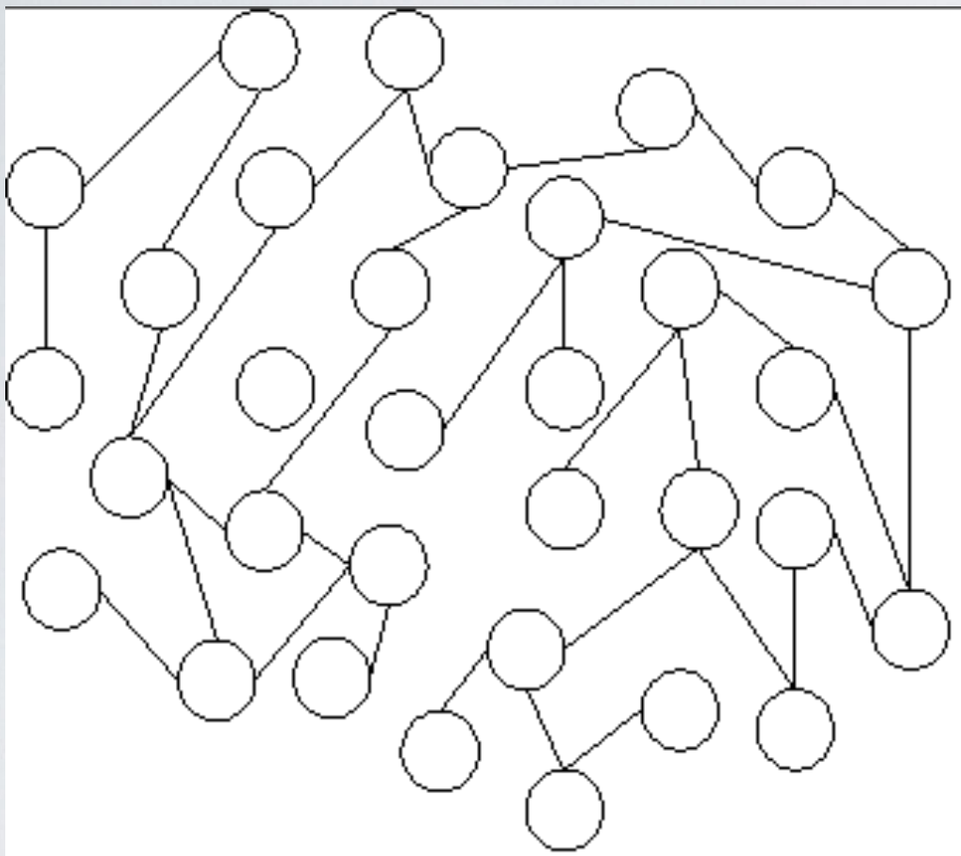
Regular network

Random network
(Erdös-Renyi)

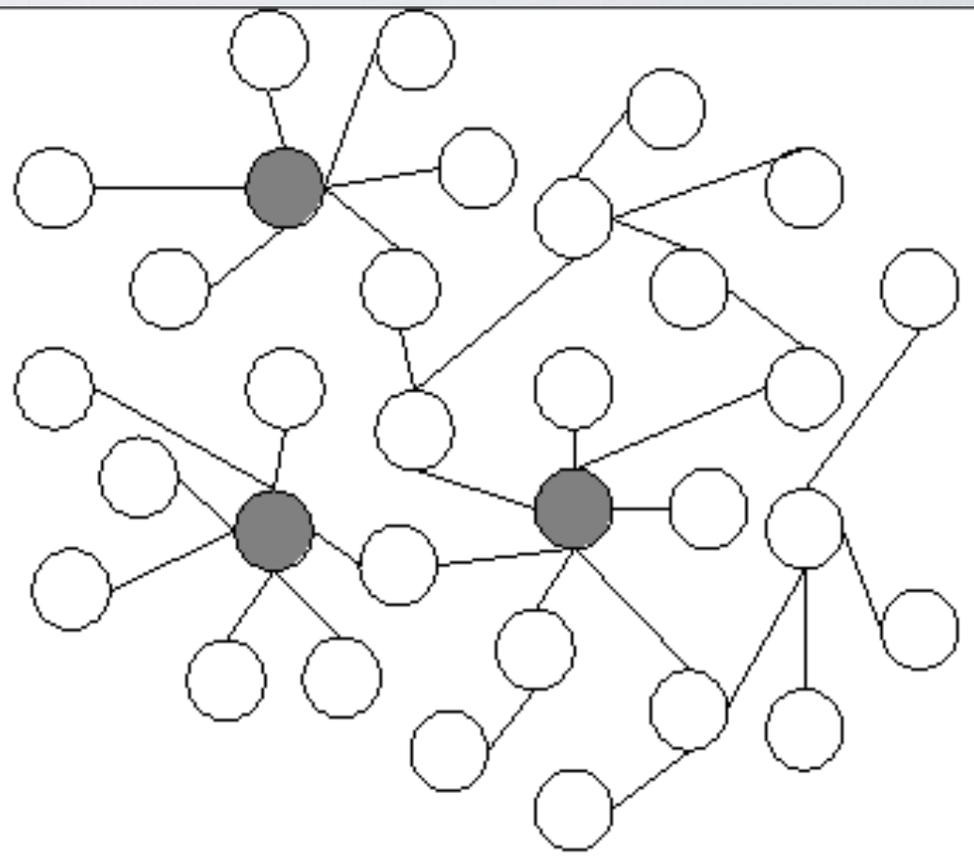
Scale-free network
(Barabasi)



Architectural invariance



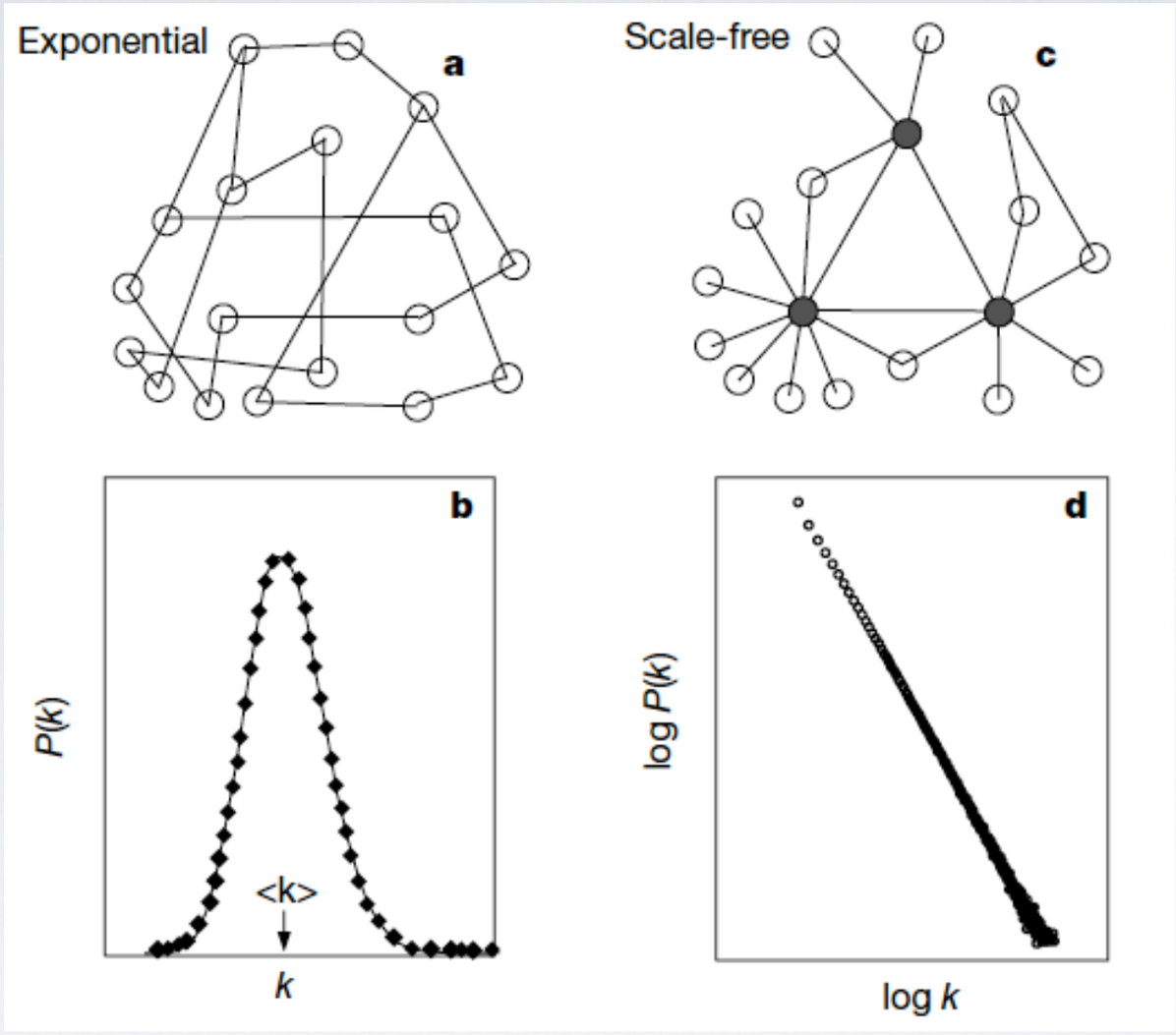
(a) Random network



(b) Scale-free network

Random network

Scale-free network



Overall very homogenous

Node connectivity follows the Poisson distribution

The probability to find highly connected nodes (with many edges) decays exponentially
Normally k = 2 or 3

Overall very heterogeneous

Topology is dominated by few highly interconnected node, (HUBs) with all other nodes minimally connected

The probability to find highly connected nodes (with many edges) follows a logarithmic trend (power law)

A "**party hub**" is a node that interacts in a functional module with a majority of its neighbours (co-expression).

Interactions are constitutive

A "**date hub**" is a node that does not interact in any functional module with a majority of its neighbours (NO co-expression).

Interactions depends on the functional context.

Differentiating Party and Date Hubs in Protein Interaction Networks using Semantic Similarity Measures

Edward Casey Kenley
Dept. of Computer Science
Baylor University
Waco, Texas, 76798
casey_kenley@baylor.edu

Lyles Kirk
Dept. of Computer Science
Baylor University
Waco, Texas, 76798
lyles_kirk@baylor.edu

Young-Rae Cho
Dept. of Computer Science
Baylor University
Waco, Texas, 76798
young-rae_cho@baylor.edu

ABSTRACT

Protein-protein interactions are fundamental to the biological processes within a cell. In the scale-free, small-world network typically modeled by protein interaction networks, hubs play a key role in maintaining the network structure. From the biological perspective, hubs are expected to be functionally essential proteins, participating in critical interactions of biological processes. Hubs can be classified into two different categories, party hubs (intra-module hubs) and date hubs (intermodule hubs), which vary in the timing and place of their associations with their interacting partners. This paper introduces a novel measure for identifying and differentiating party and date hubs in a protein interaction network. Our approach is based on the semantic similarity measure integrated with Gene Ontology data. Combined with the centrality measures of degree, betweenness, and closeness, we demonstrate that this measure detects potential party hubs and date hubs that match the confirmed party and date hubs with high accuracy.

Keywords

protein-protein interactions, protein interaction network, hubs, Gene Ontology, semantic similarity

1. INTRODUCTION

Network topology often hides behavior that is not always apparent when studying the network's entities individually. By decomposing the network with data mining techniques, such as frequent pattern mining and clustering, the hidden complexities of the graph can be unraveled. Sometimes even simple features of the network, such as node connectedness, are informative. Detecting high-degree hub nodes will usually reveal structurally critical nodes that hold the network together. The most pronounced example comes from the networks of living organisms.

Protein-protein interactions (PPIs) are fundamental to biological processes within a cell. A protein interaction network can be established from the PPI dataset. The hub in PPI networks is defined as a protein that interacts with many other proteins. In the scale-free, small-world network typically modeled by PPI networks, the hubs play a key role in maintaining the network structure. The hubs communicate with non-hubs efficiently by close links. They also connect with each other and allow information to travel freely throughout the network.

Previous studies [6, 5] have categorized hubs in PPI networks into two distinct classes: party hubs and date hubs. A party hub interacts with all its neighbors simultaneously at the same place. These hubs serve as central communication points within functional modules. A functional module represents a group of proteins that participate in the same biological process. Date hubs function oppositely to party hubs. A date hub interacts with its neighbors at different times and different place. These hubs function as messengers between different functional modules throughout the PPI network.

Identifying potential hubs in PPI networks can reveal functionally essential proteins that are conserved in evolution. Significant correlation between connectivity and essentiality has been investigated [7, 2]. Since hubs represent essential proteins, they can be determined by lethality of a cell. Removing a hub protein has a higher chance to be lethal than removing any non-hub protein. Because the party hubs facilitate communication within functional modules, removing one of the party hubs will likely cause a particular biological process to stop functioning. On the other hand, since the date hubs are typically located between functional modules via interconnections, removing a date hub can partition the PPI network into several sub-graphs. This can lead to malfunctioning cooperation between functional modules. In addition, from the network structure perspective, removing a date hub causes more damage to the connectivity of PPI networks than removing a party hub.

Hubs have been classified by gene co-expression experiments. Han et al. [6] evaluated a Pearson correlation coefficient (PCC) between a protein's mRNA expression levels and the simultaneous expression levels of its neighbors. Proteins designated as hubs have sufficiently high degree and their labels are set as party hubs or date hubs depending on which side of a threshold its PCC score fell on. Han et al. defined that a hub protein in a PPI network has degree greater than five and the threshold of PCC is set to one-half. Hubs with PCC scores greater than one-half are classified into party hubs whereas hubs with PCC scores less

Example of scale-free network with Cytoscape

The generation of a scale-free network is (supposedly)

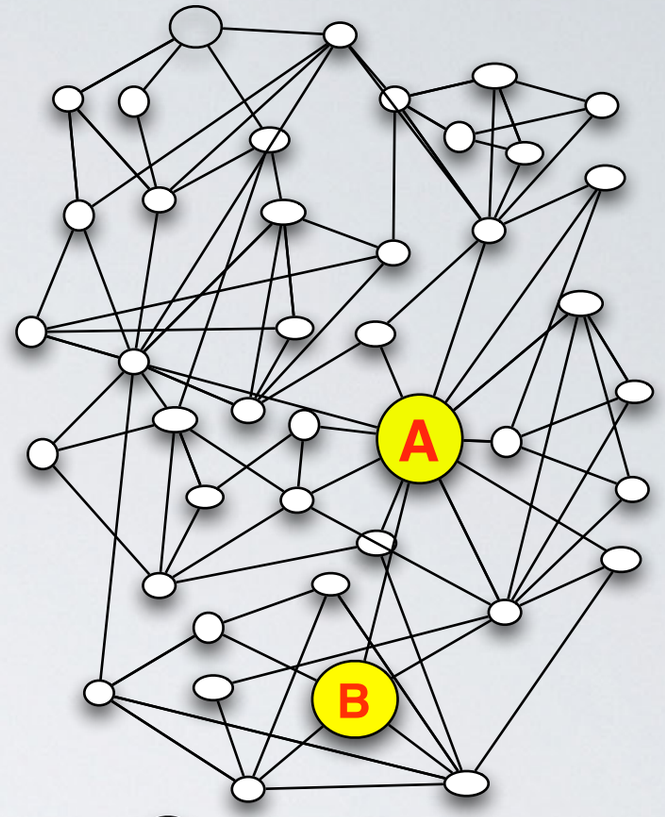
“simple”!

When a node has to establish a new connection, it is more likely to generate a connection with a node that already has many links, bringing them to an exponential growth with the increase in the number of network connections. Overall, it is a kind of situation: *the rich get richer while the poor get poorer* (in proportion). Nodes “attracting" more connections are the network *hubs.*

Scale-free networks are "robust"
(robustness = emergent property)

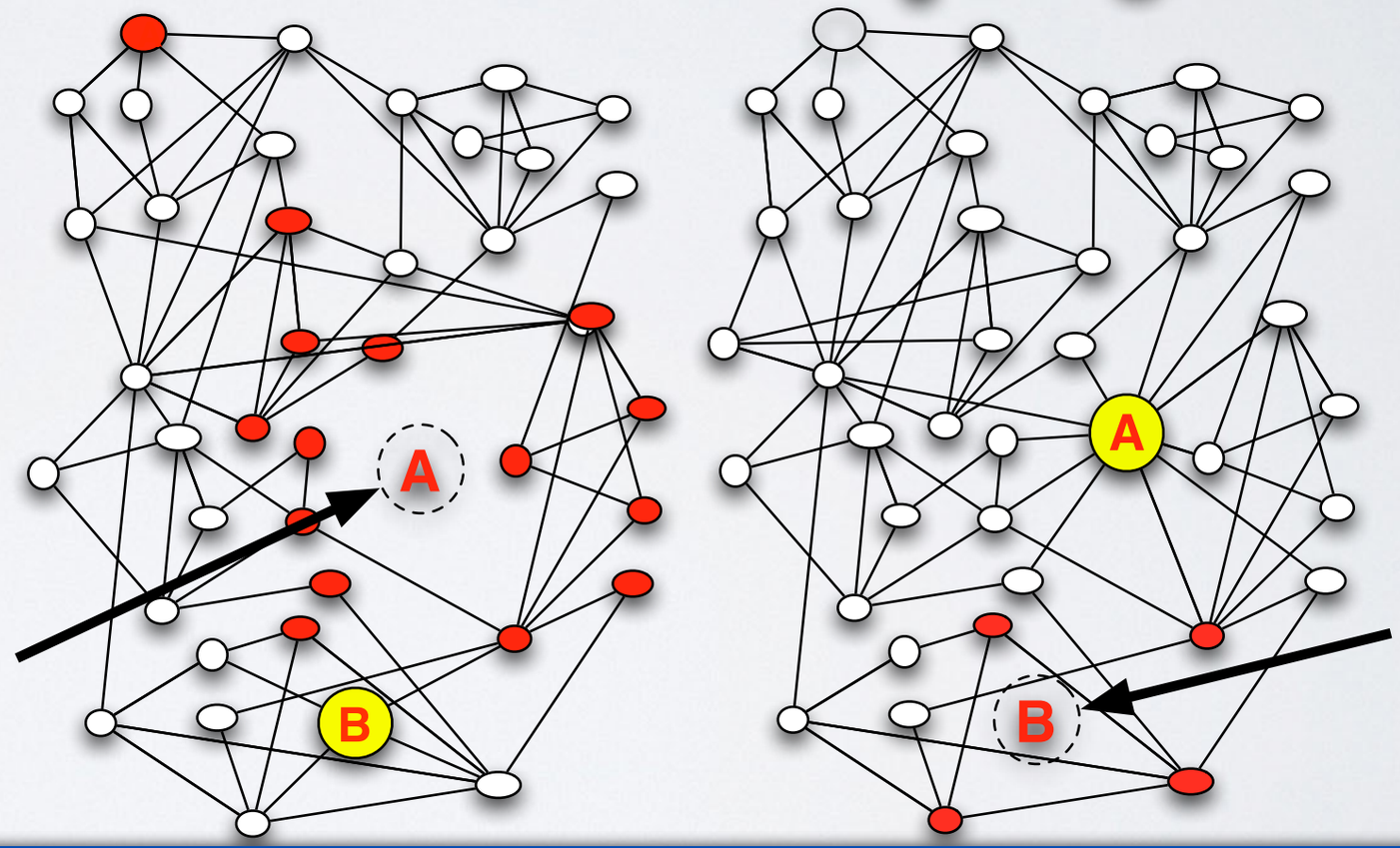
A = *k* 12

B = *k* 5



Robustness measures the *resilience* of a system to "casual" interference

Es. oncogenes are hubs



Biological networks "often" are "**scale-free**"!

$P(k) \sim k^{-\gamma}$ = Robustness

REVIEW www.rsc.org/molecularbiosystems | Molecular BioSystems

The powerful law of the power law and other myths in network biology†

Gipsi Lima-Mendez* and Jacques van Helden*

Received 5th May 2009, Accepted 12th August 2009
First published as an Advance Article on the web 2nd October 2009
DOI: 10.1039/b908681a

For almost 10 years, topological analysis of different large-scale biological networks (metabolic reactions, protein interactions, transcriptional regulation) has been highlighting some recurrent properties: power law distribution of degree, scale-freeness, small world, which have been proposed to confer functional advantages such as robustness to environmental changes and tolerance to random mutations. Stochastic generative models inspired different scenarios to explain the growth of interaction networks during evolution. The power law and the associated properties appeared so ubiquitous in complex networks that they were qualified as "universal laws". However, these properties are no longer observed when the data are subjected to statistical tests: in most cases, the data do not fit the expected theoretical models, and the cases of good fitting merely result from sampling artefacts or improper data representation. The field of network biology seems to be founded on a series of myths, *i.e.* widely believed but false ideas. The weaknesses of these foundations should however not be considered as a failure for the entire domain. Network analysis provides a powerful frame for understanding the function and evolution of biological processes, provided it is brought to an appropriate level of description, by focussing on smaller functional modules and establishing the link between their topological properties and their dynamical behaviour.

Myth 1: the degree distribution of biological networks follows a power law

Myth 2: Biological networks are scale-free

Myth 3: the metabolic network is a small world

Myth 4: small worlds are tolerant to random deletions, but vulnerable to targeted attacks

Myth 5: biological networks grow by preferential attachment

Consistent differences between transcription, signalling and metabolic networks

A more conservative conclusion:

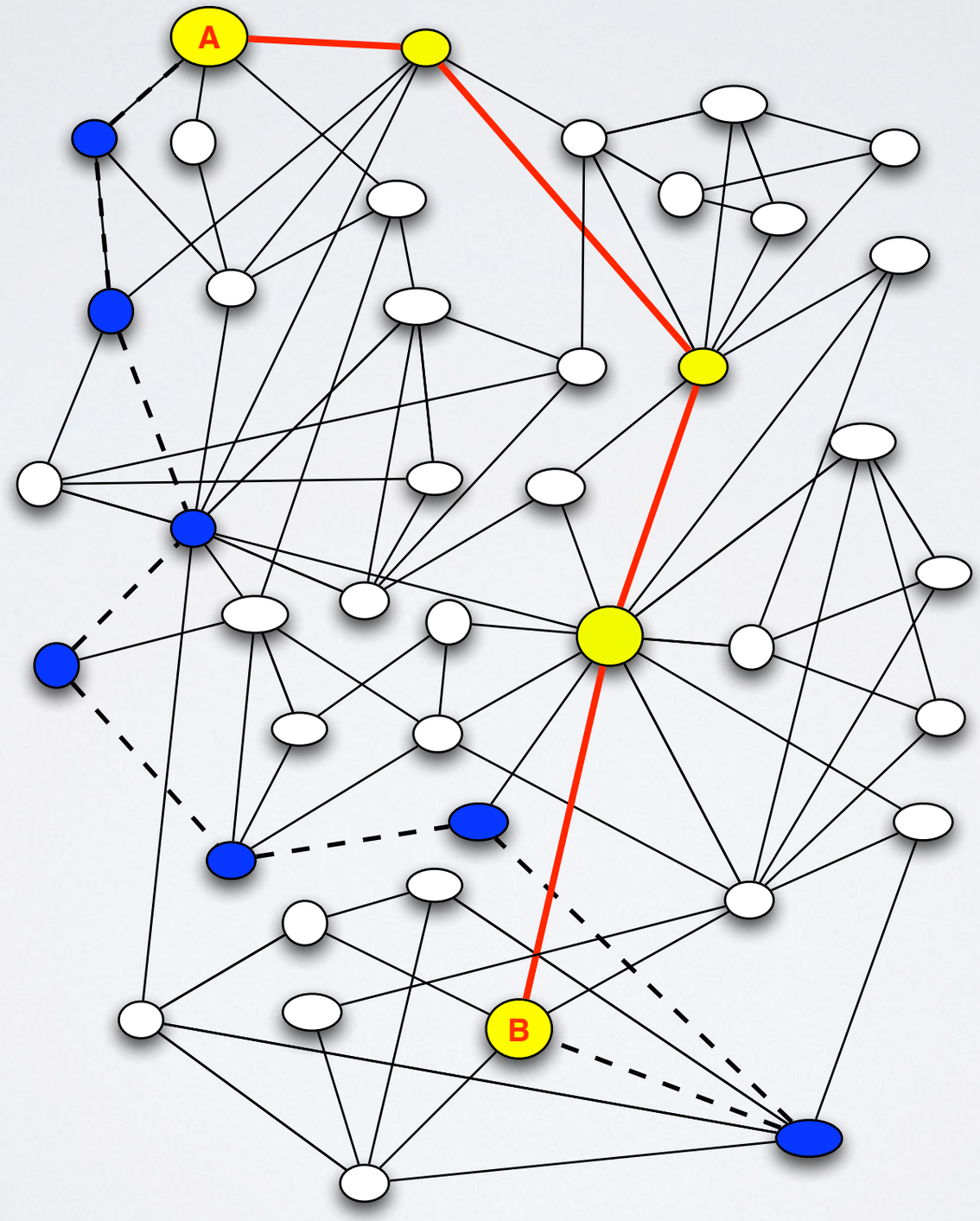
Power-law distribution, scale-free structure, robustness, preferential attachment and so on should not ne considered "universal", absolute properties of "all" biological networks,

but

type-, dimension and context-specific properties.

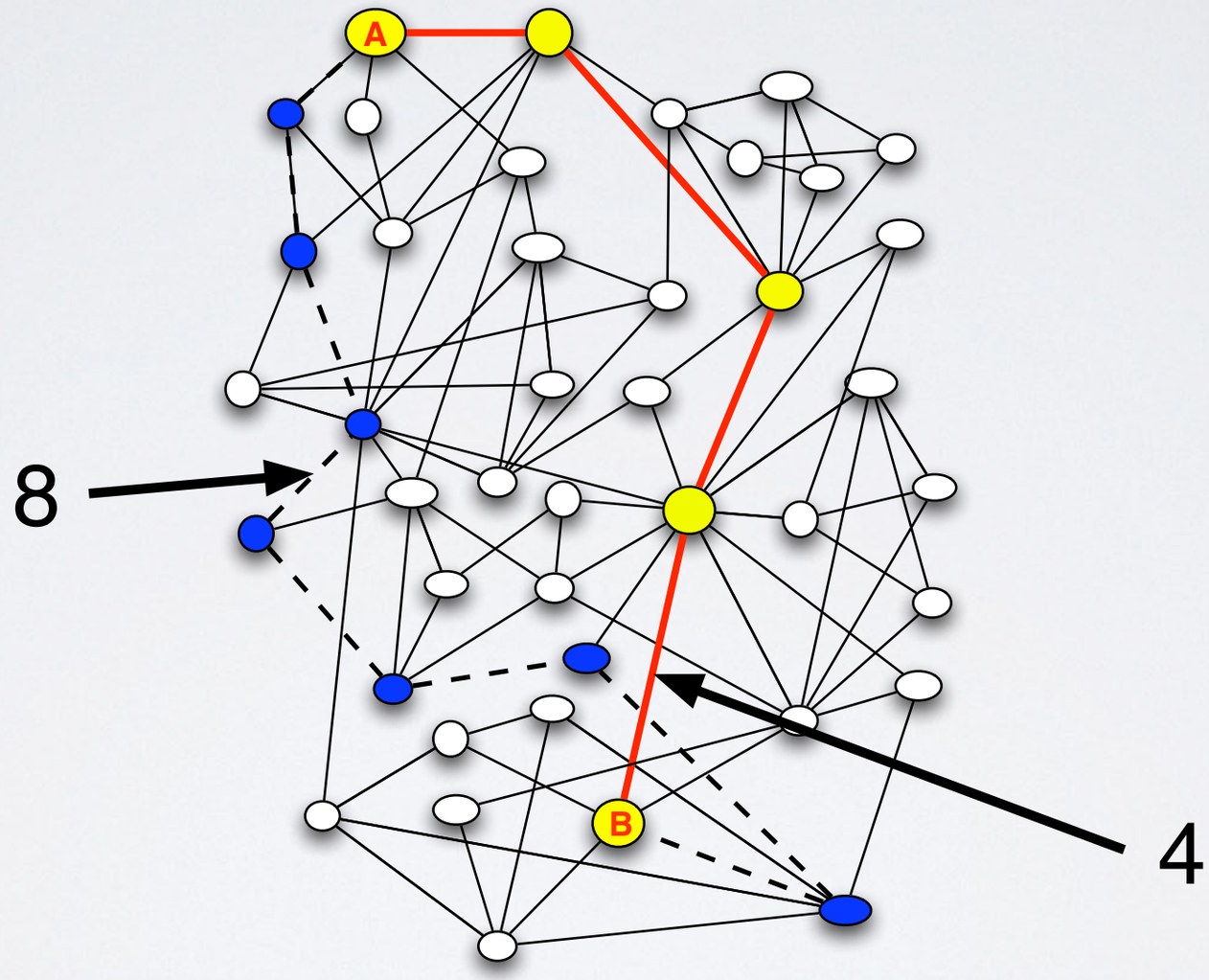
"Shortest path" = The minimal number of edges between two nodes

Shortest path A-B = 4



“Average shortest path” =
The“average” minimal path between all network nodes.

It measures the “**navigability**” of a network; that is the easiness of moving between nodes.



“Diameter” = the maximal shortest path between all network nodes
“Average path length” = the “average” distance (not shortest) between all network nodes

Example of shortest path calculation with Cytoscape

"Clustering coefficient" *C* =

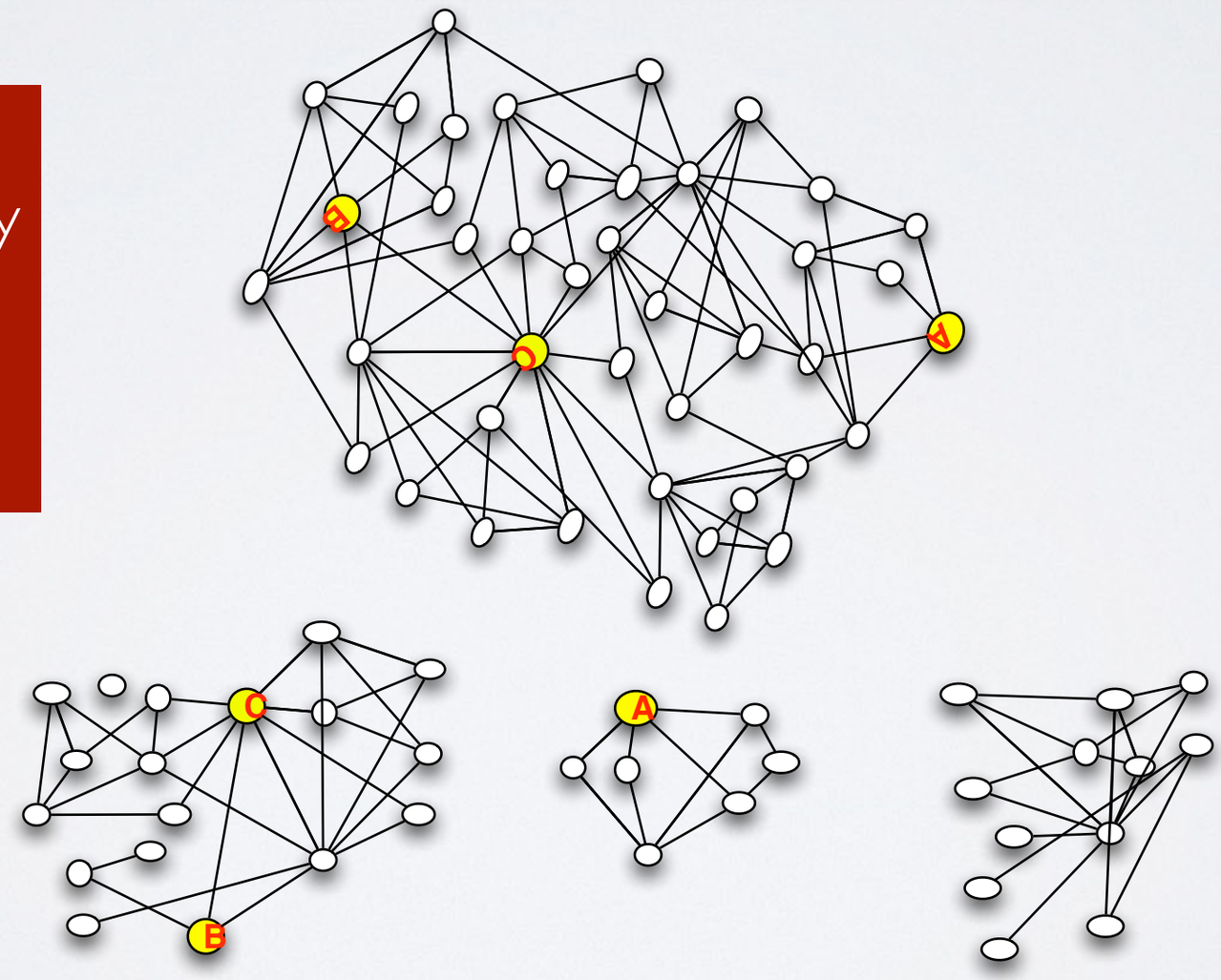
Probability that if A is connected to B, and B is connected to C, then A is connected to C

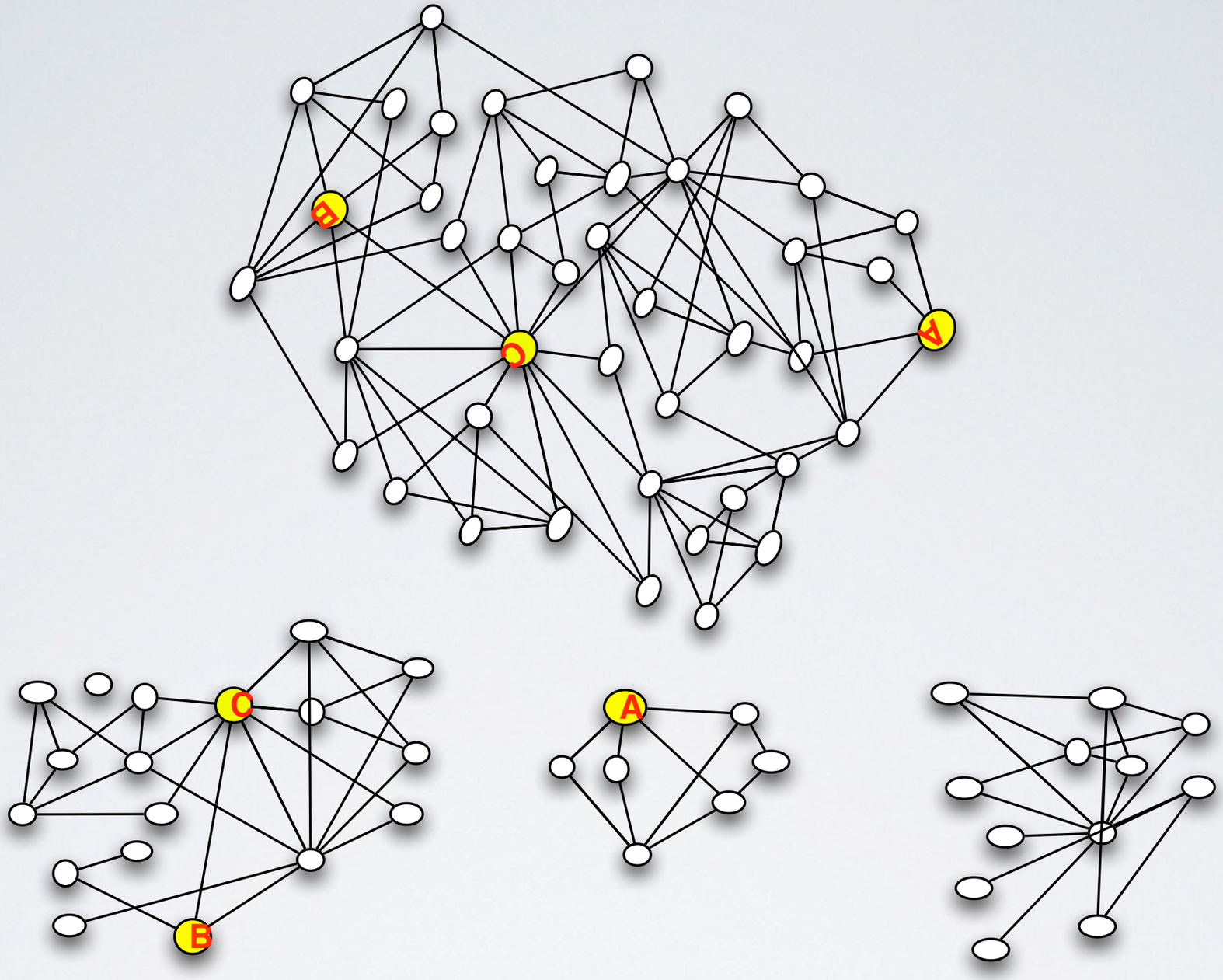
"Average Clustering coefficient" *<C>* =

Average probability of the network to generate clusters

Measure the global tendency of nodes to organise themselves in groups (clusters)

Probability of *cluster* generation by individual nodes and/or overall network





Modular decomposition
es. MCODE

Example of network modular decomposition within Cytoscape

"**Node degree distribution**" ***P(k)***
(scale-free topology)

+

"**Average Clustering coefficient**" ***<C>***
(decomposizione modulare)

=

Hierarchical network analysis

Biological networks can be scale-free, modular and hierarchical

$P(k) \sim k^{-\gamma} \longleftrightarrow C_1 = 2n_1/k(k-1)$

Cl concentrations in the Sajama ice core, and to a number of other pedological and geomorphological features indicative of long-term dry climates (*8*, *11–14*, *18*). This decline in human activity around the Altiplano paleolakes is seen in most caves, with early and late occupations separated by largely sterile mid-Holocene sediments. However, a few sites, including the caves of Tulan-67 and Tulan-68, show that people did not completely disappear from the area. All of the sites of sporadic occupation are located near wetlands in valleys, near large springs, or where lakes turned into wetlands and subsistence resources were locally still available despite a generally arid climate (*7*, *8*, *19*, *20*).

Archaeological data from surrounding areas suggest that the Silencio Arqueológico applies best to the most arid areas of the central Andes, where aridity thresholds for early societies were critical. In contrast, a weaker expression is to be expected in the more humid highlands of northern Chile (north of 20°S, such as Salar Huasco) and Peru (*21*). In northwest Argentina, the Silencio Arqueológico is found in four of the six known caves (*22*) [see review in (*23*)]. It is also found on the coast of Peru in sites that are associated with ephemeral streams (*24*). The southern limit in Chile and northwest Argentina has yet to be explored.

References and Notes

1. T. Dillehay, *Science* **245**, 1436 (1989).
2. D. J. Meltzer *et al.*, *Am. Antiq.* **62**, 659 (1997).
3. T. F. Lynch, C. M. Stevenson, *Quat. Res.* **37**, 117 (1992).
4. D. H. Sandweiss *et al.*, *Science* **281**, 1830 (1998).
5. L. Núñez, M. Grosjean, I. Cartajena, in *Interhemispheric Climate Linkages*, V. Markgraf, Ed. (Academic Press, San Diego, CA 2001), pp. 105–117.
6. M. A. Geyh, M. Grosjean, L. Núñez, U. Schotterer, *Quat. Res.* **52**, 143 (1999).
7. J. L. Betancourt, C. Latorre, J. A. Rech, J. Quade, K. Rylander, *Science* **289**, 1542 (2000).
8. M. Grosjean *et al.*, *Global Planet. Change* **28**, 35 (2001).
9. C. Latorre, J. L. Betancourt, K. A. Rylander, J. Quade, *Geol. Soc. Am. Bull.* **114**, 349 (2002).
10. Charcoal in layers containing triangular points has been ^{14}C dated at Tuina-1, Tuina-5, Tambillo-1, San Lorenzo-1, and Tuyajto-1 between 13,000 and 9000 cal yr B.P. (table S1 and fig. S1).
11. P. A. Baker *et al.*, *Science* **291**, 640 (2001).
12. G. O. Seltzer, S. Cross, P. Baker, R. Dunbar, S. Fritz, *Geology* **26**, 167 (1998).
13. L. G. Thompson *et al.*, *Science* **282**, 1858 (1998).
14. M. Grosjean, *Science* **292**, 2391 (2001).
15. E. P. Tonni, written communication.
16. M. T. Alberdi, written communication.
17. J. Fernandez *et al.*, *Geoarchaeology* **6**, 251 (1991).
18. The histogram of middens is processed from (*9*).
19. M. Grosjean, L. Núñez, I. Cartajena, B. Messerli, *Quat. Res.* **48**, 239 (1997).
20. The term Silencio Arqueológico describes the mid-Holocene collapse of human population at those archaeological sites of the Atacama Desert that are vulnerable to multicentennial or millennial-scale drought. The term Silencio Archaeológico does not conflict with the presence of humans at sites that are not susceptible to climate change, such as in spring and river oases that drain large (Pleistocene) aquifers or at sites where wetlands were created during the arid middle Holocene, such as Tulan-67, Tulan-68, and Laguna Miscanti.
21. M. Aldenderfer, *Science* **241**, 1828 (1988).
22. A mid-Holocene hiatus is found at Inca Cueva 4, Huachichocana 3, Pintocamayoc, and Yavi, whereas occupation continued at the oases of Susques and Quebrada Seca.
23. L. Núñez *et al.*, *Estud. Atacamenos* **17**, 125 (1999).
24. D. H. Sandweiss, K. A. Maasch, D. G. Anderson, *Science* **283**, 499 (1999).

25. Grants from the National Geographic Society (5836-96), the Swiss National Science Foundation (21-57073), and Fondo Nacional de Desarrollo Científico y Tecnológico (1930022) and comments by J. P. Bradbury, B. Meggers, G. Seltzer, and D. Stanford are acknowledged.

Supporting Online Material

www.sciencemag.org/cgi/content/full/298/5594/821/DC1

Figs. S1 to S3

Tables S1 and S2

22 July 2002; accepted 9 September 2002

Network Motifs: Simple Building Blocks of Complex Networks

R. Milo,[1] S. Shen-Orr,[1] S. Itzkovitz,[1] N. Kashtan,[1] D. Chklovskii,[2] U. Alon[1]*

Complex networks are studied across many fields of science. To uncover their structural design principles, we defined "network motifs," patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. We found such motifs in networks from biochemistry, neurobiology, ecology, and engineering. The motifs shared by ecological food webs were distinct from the motifs shared by the genetic networks of *Escherichia coli* and *Saccharomyces cerevisiae* or from those found in the World Wide Web. Similar motifs were found in networks that perform information processing, even though they describe elements as different as biomolecules within a cell and synaptic connections between neurons in *Caenorhabditis elegans*. Motifs may thus define universal classes of networks. This approach may uncover the basic building blocks of most networks.

Many of the complex networks that occur in nature have been shown to share global statistical features (*1–10*). These include the "small world" property (*1–9*) of short paths between any two nodes and highly clustered connections. In addition, in many natural networks, there are a few nodes with many more connections than the average node has. In these types of networks, termed "scale-free networks" (*4, 6*), the fraction of nodes having k edges, $p(k)$, decays as a power law $p(k) \sim k^{-\gamma}$ (where γ is often between 2 and 3). To go beyond these global features would require an understanding of the basic structural elements particular to each class of networks (*9*). To do this, we developed an algorithm for detecting network motifs: recurring, significant patterns of interconnections. A detailed application to a gene regulation network has been presented (*11*). Related methods were used to test hypotheses on social networks (*12, 13*). Here we generalize this approach to virtually any type of connectivity graph and find the striking appearance of

[1]Departments of Physics of Complex Systems and Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel 76100. [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA.

*To whom correspondence should be addressed. E-mail: urialon@weizmann.ac.il

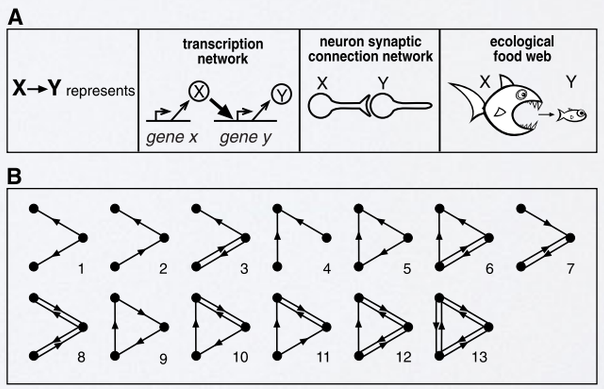


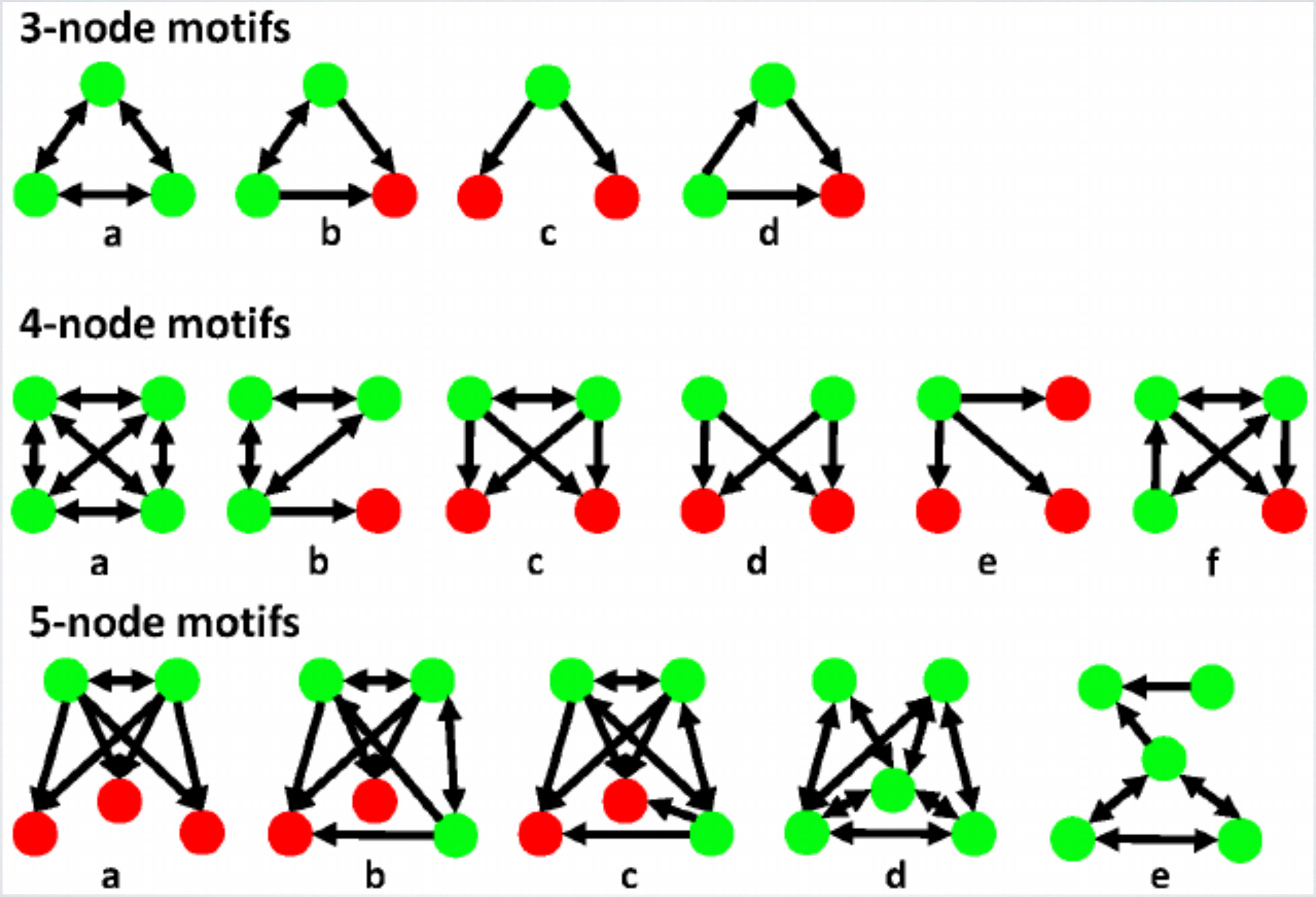
Fig. 1. (**A**) Examples of interactions represented by directed edges between nodes in some of the networks used for the present study. These networks go from the scale of biomolecules (transcription factor protein X binds regulatory DNA regions of a gene to regulate the production rate of protein Y), through cells (neuron X is synaptically connected to neuron Y), to organisms (X feeds on Y). (**B**) All 13 types of three-node connected subgraphs.

Network motifs are sub-graphs that repeat themselves in a specific network or even among various networks.

Motifs are local property of networks, defined as recurrent and statistically significant sub-graphs or patterns.

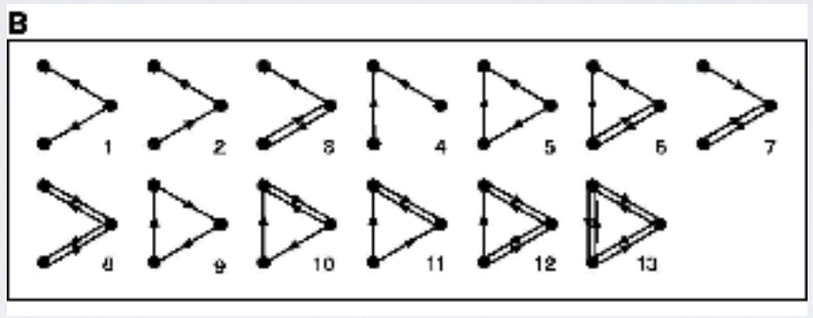
Motifs may reflect functional properties and are used as useful concept to uncover structural design principles of complex networks.

"**Network motifs**" = directed patterns of node connections that appear more frequently in real networks than in random networks artificially generated by computers

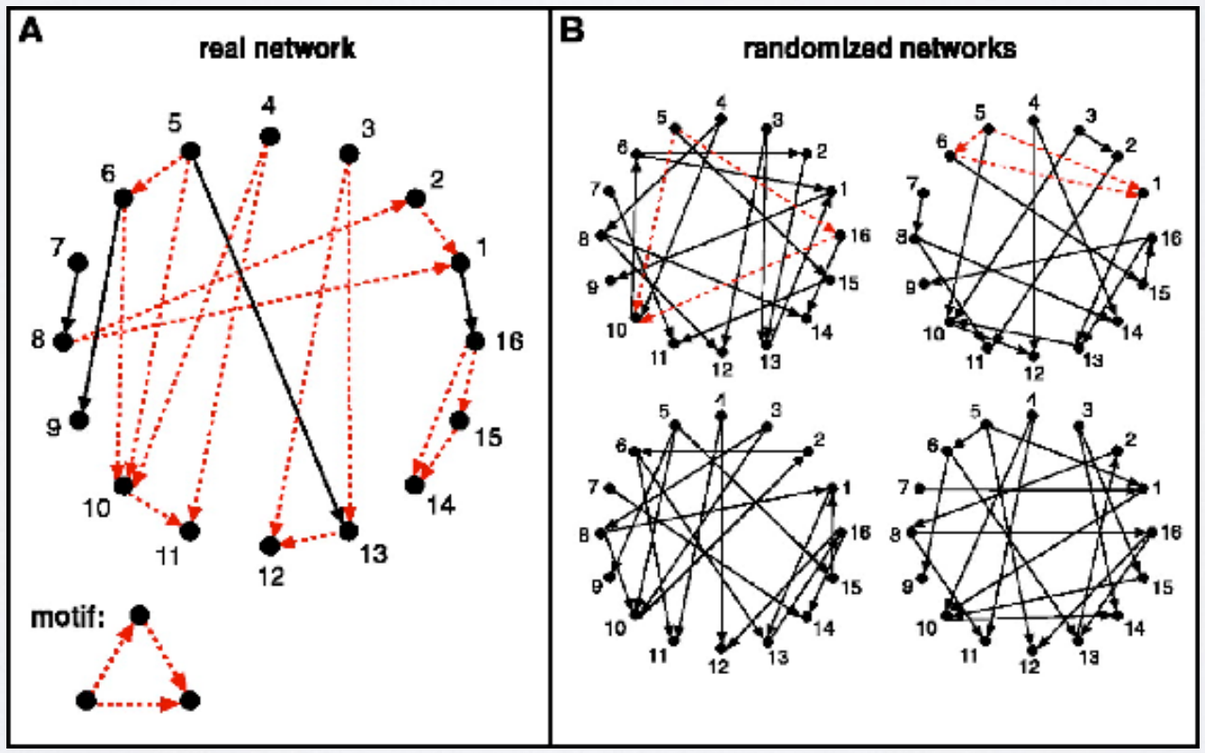


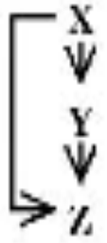
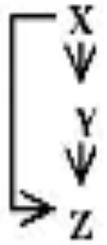
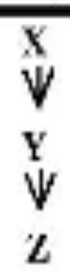
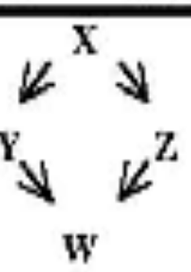
"**Network motifs**": many possible configurations depending on node number

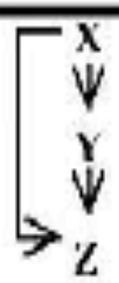
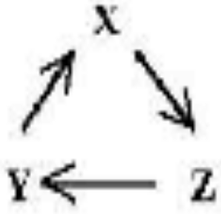
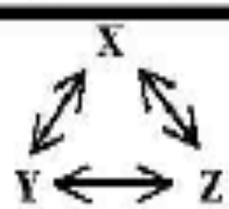
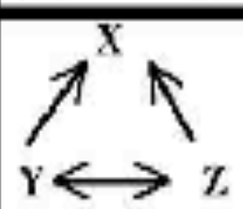
3 nodi = 13 possibili configurazioni



4 nodi = 199 possibili configurazioni

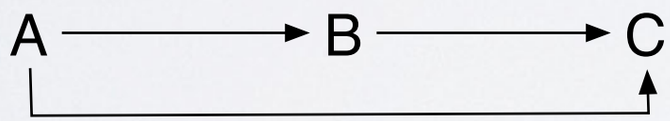


Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
S. cerevisiae[¶]	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans[†]	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			

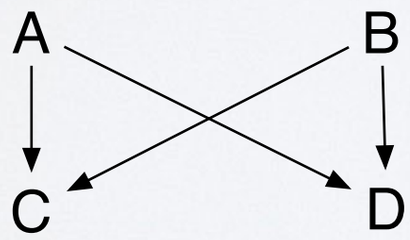
Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Electronic circuits (forward logic chips)			 **Feed-forward loop**			 **Bi-fan**			 **Bi-parallel**		
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)			 **Three-node feedback loop**			 **Bi-fan**			 **Four-node feedback loop**		
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web			 **Feedback with two mutual dyads**			 **Fully connected triad**			 **Uplinked mutual dyad**		
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

In all informational networks (including signalling and gene expression networks) “feed-forward loops” and “bi-fan” motifs are present.

In metabolic networks different motifs are generated



Feed-forward loop



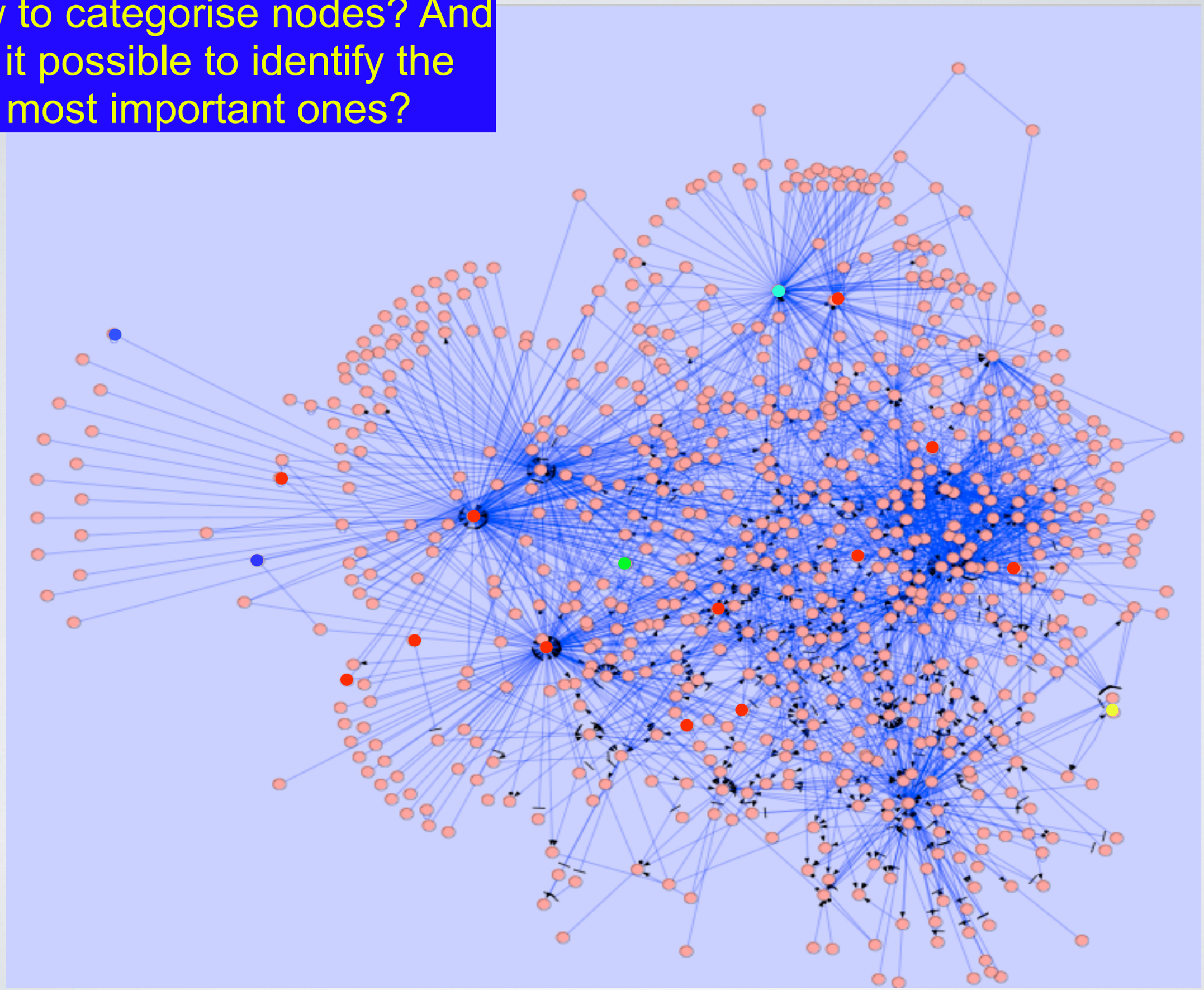
Bi-fan

This suggests that the basic building blocks of natural networks are context specific.

<u>An important question:</u>

Is network evolution and functional diversification controlled by basic mathematical rules?

How to categorise nodes? And is it possible to identify the most important ones?



Indexes of centrality based on the calculation of the shortest path

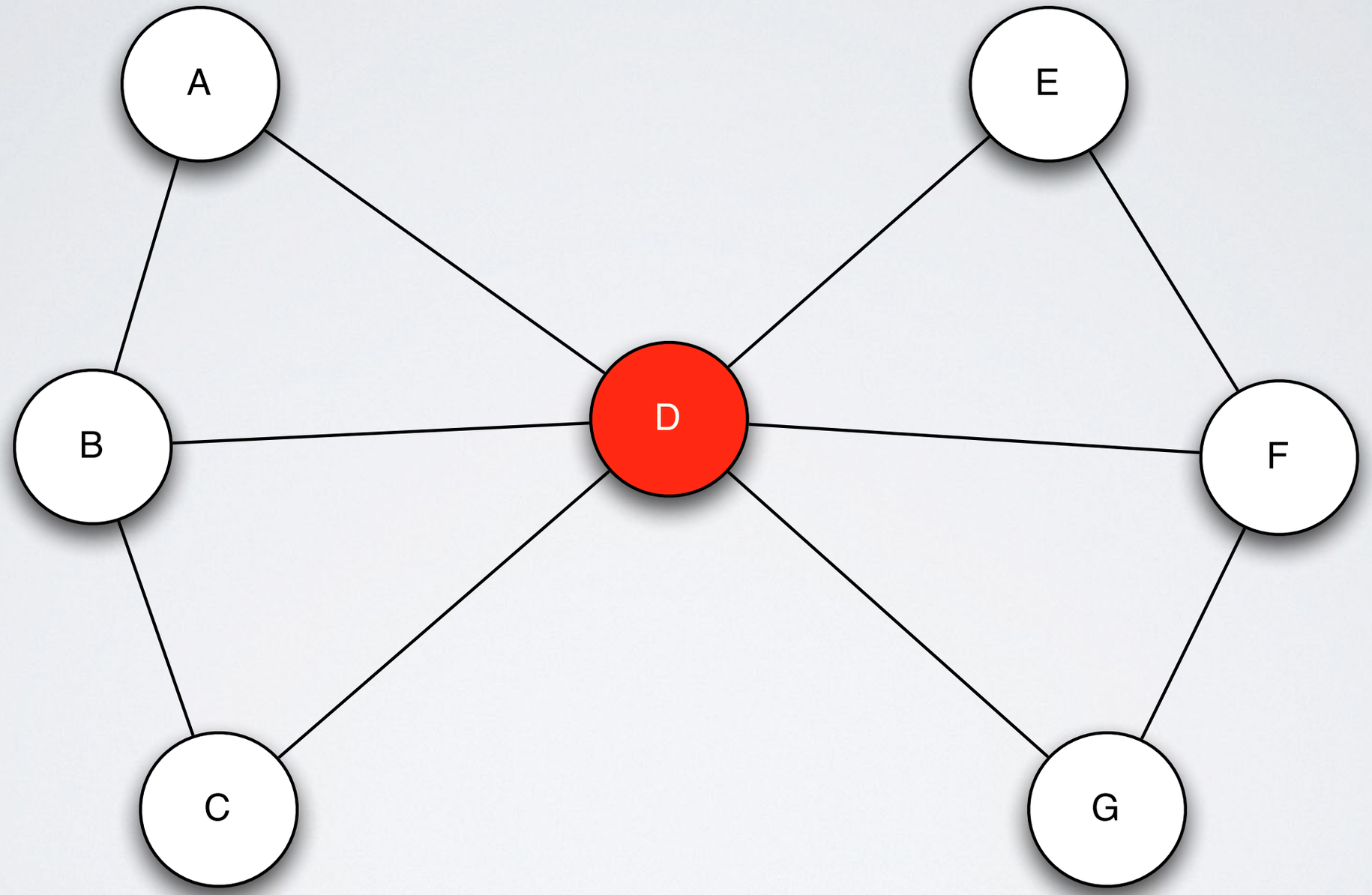
- Betweenness
- Stress
- Centroid
- Radiality
- Eccentricity
- Closeness
- Bridging centrality
- Eigenvector

CentiServer

comprehensive centrality resource and server for centralities calculation

Node betweenness indicates the functional relevance of the node as capable of bringing in communication individual nodes or set of nodes in a network. Thus a high betweenness score indicates the ability of a node to function as a communicator node. For example, if a graph node does abstract a signaling protein, then that protein is able to keep in functional communication other distant proteins, facilitating the informational biochemical flow

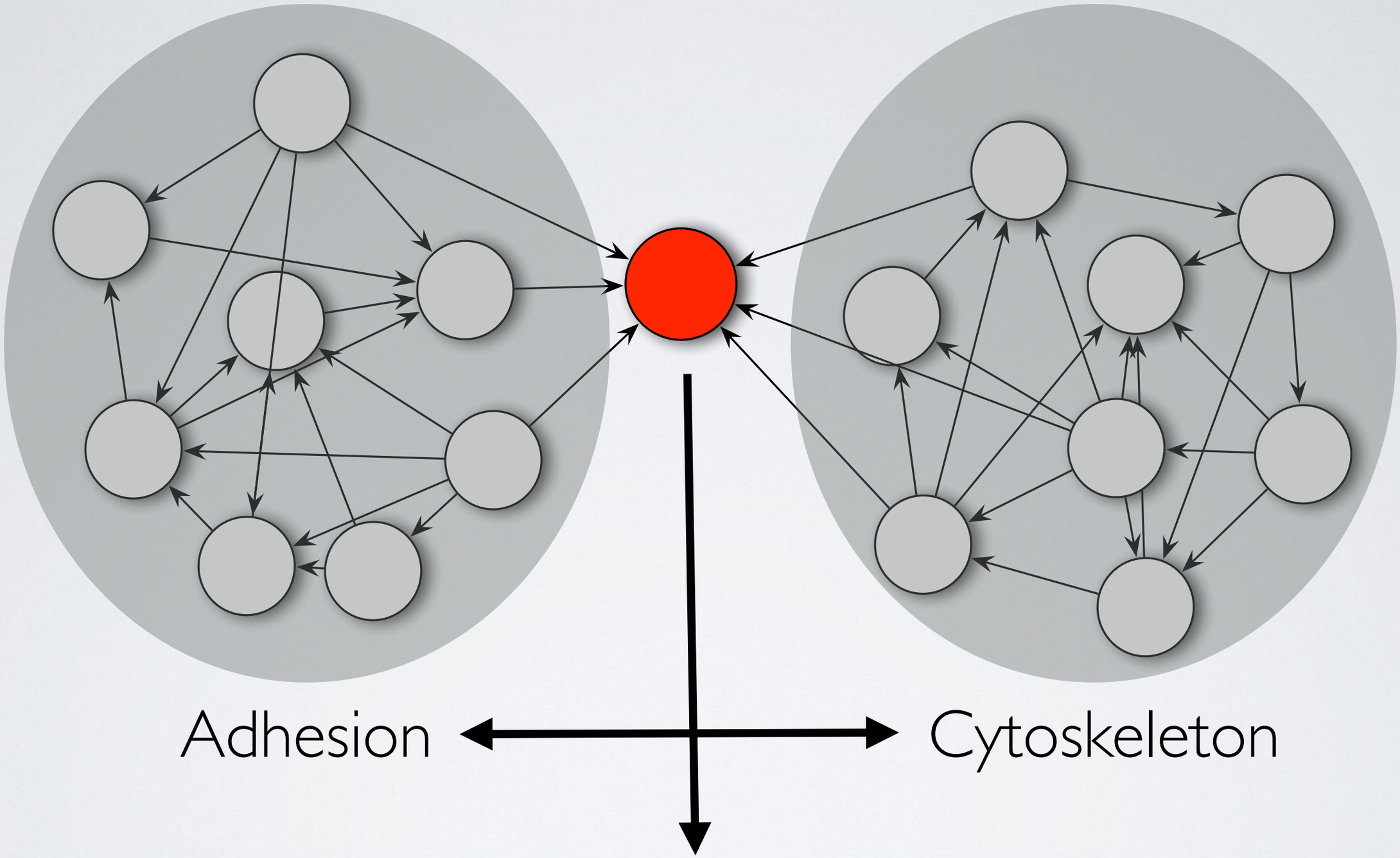
Es. betweenness



Node centroid indicates the "probability " that a node has to organise (bringing along) clusters of nodes or functional modules. A high value indicates the centroid capacity of a node to function *as an organiser of functional modules*.

For example, if the graph abstracts a signaling network, then a protein (node) with high centroid is able to act as coordinator of closely linked proteins (signalosomes), thus greatly facilitating the emergence and coordination of specific biochemical functions.

Es. centroid



Chemotaxis

Example of centrality calculation within Cytoscape

Summary of possible graph quantifications (rather incomplete!)

"Node degree"

"Average degree"

"Node degree distribution"

"Shortest path"

"Average shortest path"

"Diameter"

"Centrality"

"Clustering coefficient"

"Average Clustering coefficient"

Modular decomposition

Network motifs

(3)

"High level questions"

Is the system robust? (= tolerant to interference)

What are the possible regulators / effectors of a molecule?

Are there cell-type-specific regulatory mechanisms?

What are the regulatory mechanisms of a certain cellular events?

What are the side effects of a drug?

What are the mechanisms of action of a drug?

What are the general molecular alterations during septic shock?

What the molecular mechanism of resistance to therapy in B-CLL leukaemia?

Example:
B-CLL
=
B-cell chronic lymphoid leukaemia

Chronic lymphocytic leukaemia (CLL) is the most common leukaemia in the Western world.

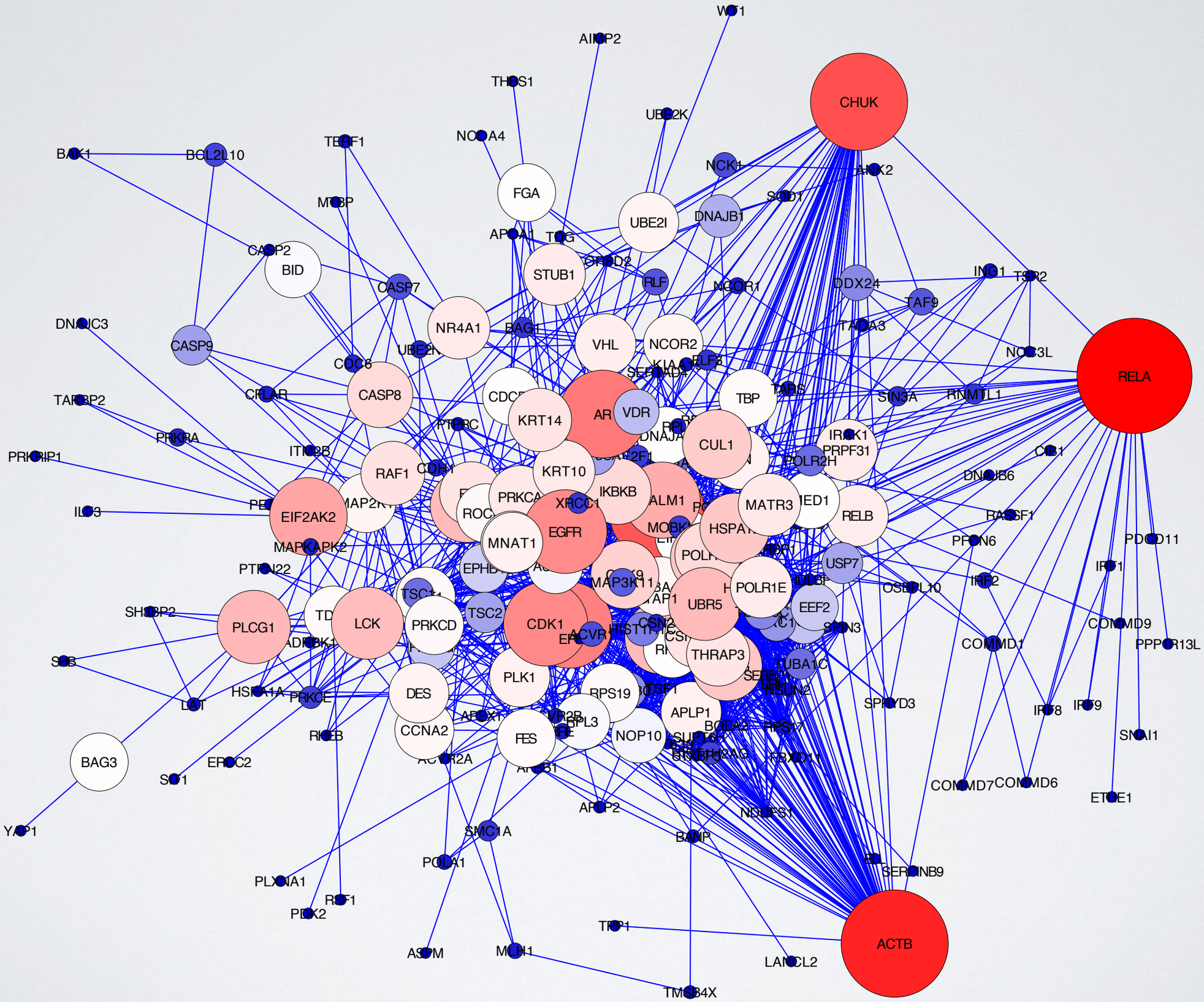
Diagnosis is based on the results of flow cytometric analysis of malignant B cells obtained from peripheral blood, bone marrow, lymph nodes, and other organs.

Genes involved in B-CLL pathogenesis and progression

CDKN2A
KRAS
NRAS
PTEN
SF3B1
AKT1
ATM
BCL2
BAG1
MCL1
BAX
DAPK1
EIF2AK2
MDM2
MMP9
NFKB1
PIK3C2A
PIK3C2B
TCL1A
TERT
TNFSF13
TNFSF13B
TP53
ZAP70
DLEU7
DLEU1
DLEU2
NOTCH1
BIRC3

MYD88
XPO1
ROR1
CXCR4
CXCL12
BCR
LYN
SYK
BTK
PIK3C2A
PIK3C2B
PIK3CD
AKT1
MTOR
PRKCB
PRKCG
PRKCD
PRKCE
PRKCZ
PIM1
PIM2
PIM3
MAP3K14
miR15a
miR16a
miR34b
miR34c
miR29b
miR181b

Example (1): B-CLL-related gene network: most important genes?



The betweenness allows creating "a hierarchical" distribution in the nodes of a graph.

If the graph abstract a signal transduction network controlling a leukaemia, such as B-CLL, then the nodes with the highest betweenness are proteins which have a greater regulatory role in that specific pathology. Thus, they could be critical targets of therapy.

POLR1A	980.5752201218744
BCL2L10	98.1129828243036
CASP8	972.9847961866025
HIST1H2AG	93.81629994245209
PRKCE	91.36638628304313
UBE2K	9.704326174914408
MOBKL1B	89.42005465096672
H2AFX	873.2376959559298
CSNK2A2	850.3599434784832
E2F1	85.79770426816815
RAF1	841.8706710726763
EZR	808.7017530971103
KRT14	805.3953908450881
APLP2	8.632076269691636
OSBPL10	8.461386862034189
THRAP3	796.3531866931876
XRCC1	78.35366944609784
SMC1A	76.96810224719924
RNMTL1	76.27244149308481
BAG1	76.12150503512692

POLR1A	980.5752201
CASP8	972.9847962
H2AFX	873.237696
CSNK2A2	850.3599435
RAF1	841.8706711
EZR	808.7017531
KRT14	805.3953908
THRAP3	796.3531867
BCL2L10	98.11298282
HIST1H2AG	93.81629994
PRKCE	91.36638628
MOBKL1B	89.42005465
E2F1	85.79770427
XRCC1	78.35366945
SMC1A	76.96810225
RNMTL1	76.27244149
BAG1	76.12150504
UBE2K	9.704326175
APLP2	8.63207627
OSBPL10	8.461386862

Example (2) of network topological analysis with Cytoscape

The Context:

B-CLL (B-cell Chronic Lymphocytic Leukaemia)
and
CML (Chronic Myeloid Leukaemia)

The questions:

Are there common driver genes between the two leukemias?
If so, which are most important ones?
And is it possible to identify drugs to block them?

The context:

Signal transduction
+
Human protein-protein interactome (PPI)

Definitions:

Signal transduction = a complex homeostatic molecular system controlling environmental information processing in all living forms (bacteria —> humans); about 5000 genes

Protein-protein interactome (PPI) = the ensemble of all "known" protein-protein binary physical interaction occurring in a species (genetically determined)

(to date: **16501 proteins - 308822 binary interactions;** estimated to be about 400000-500000 binary interactions)

The PPI interactome is the **space** of all possible protein-protein interactions in which all biochemical events of signalling may potentially occur.

Analysis flow-chart

1. FN network of BCR-ABL1 (Chr 9-22; Philadelphia)
2. FN network from B-CLL probe
3. Subnetwork from intersection
4. Calculating centralities
5. Identification of kinases and phosphatases
6. Sub-network inference upon extraction of kinases and phosphatases with high centrality indexes (over the average)

The logic: novel anti-leukemia drugs are mainly kinase inhibitors

1. *FN network of BCR-ABL1*

=

584 nodes (proteins)

14731 binary interactions

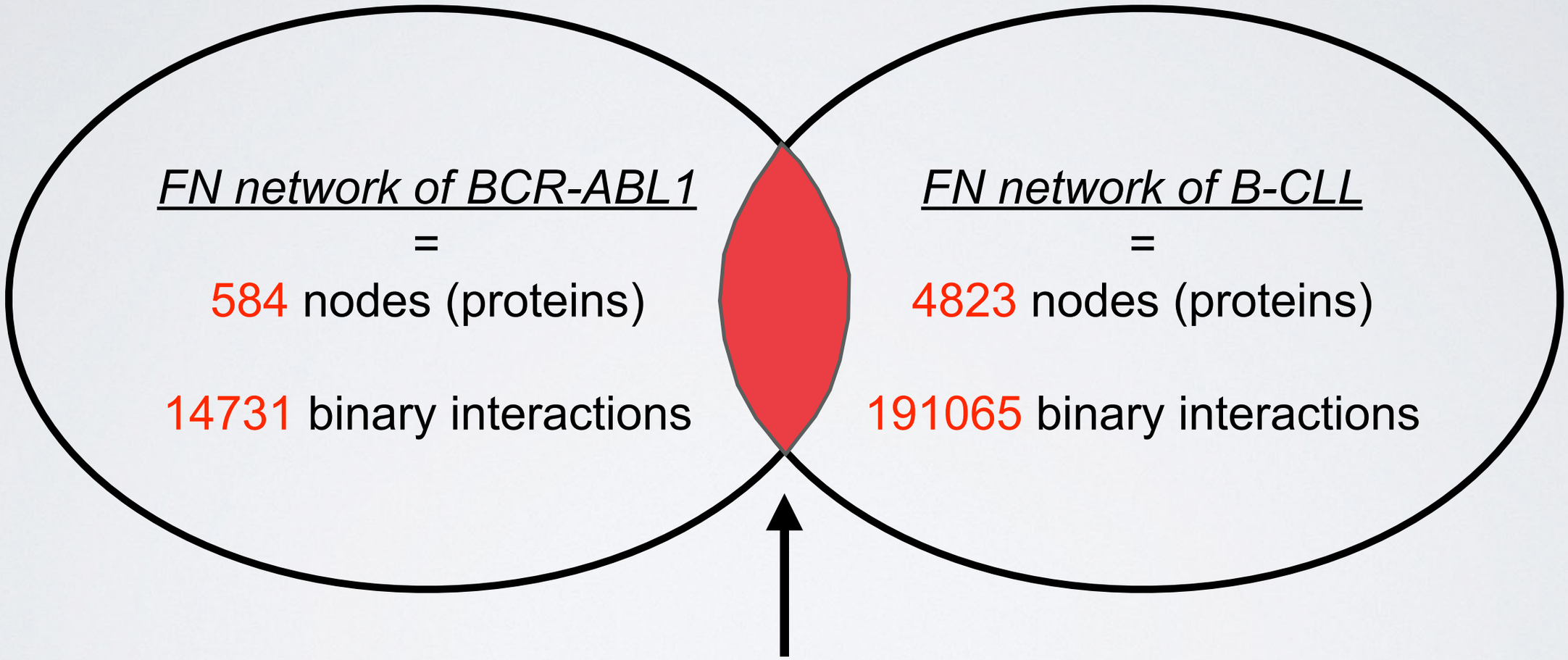
2. *FN network from B-CLL probe*

=

4823 nodes (proteins)

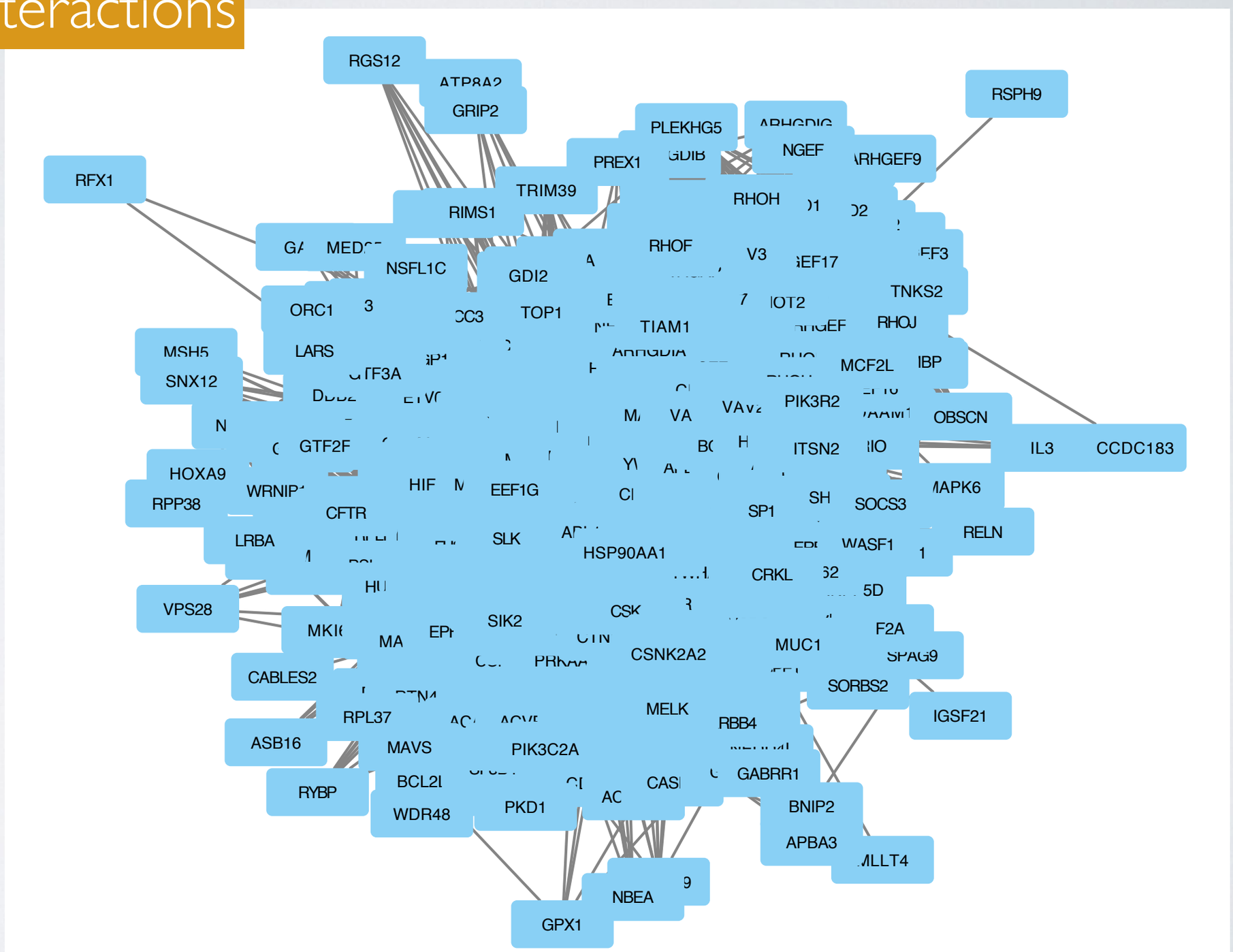
191065 binary interactions

3. Subnetwork from intersection

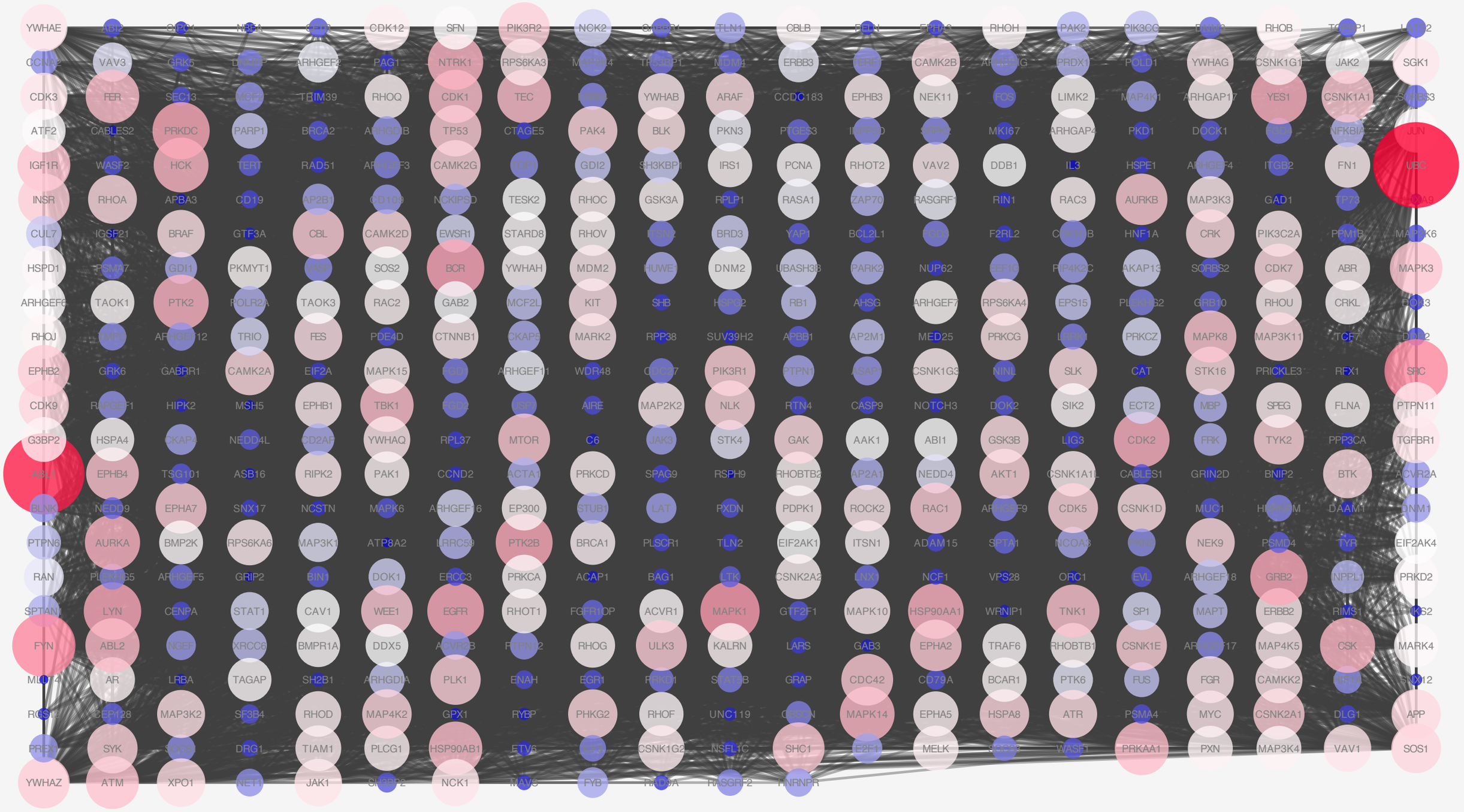


3.Subnetwork from intersection

474 nodes (proteins)
13791 binary interactions

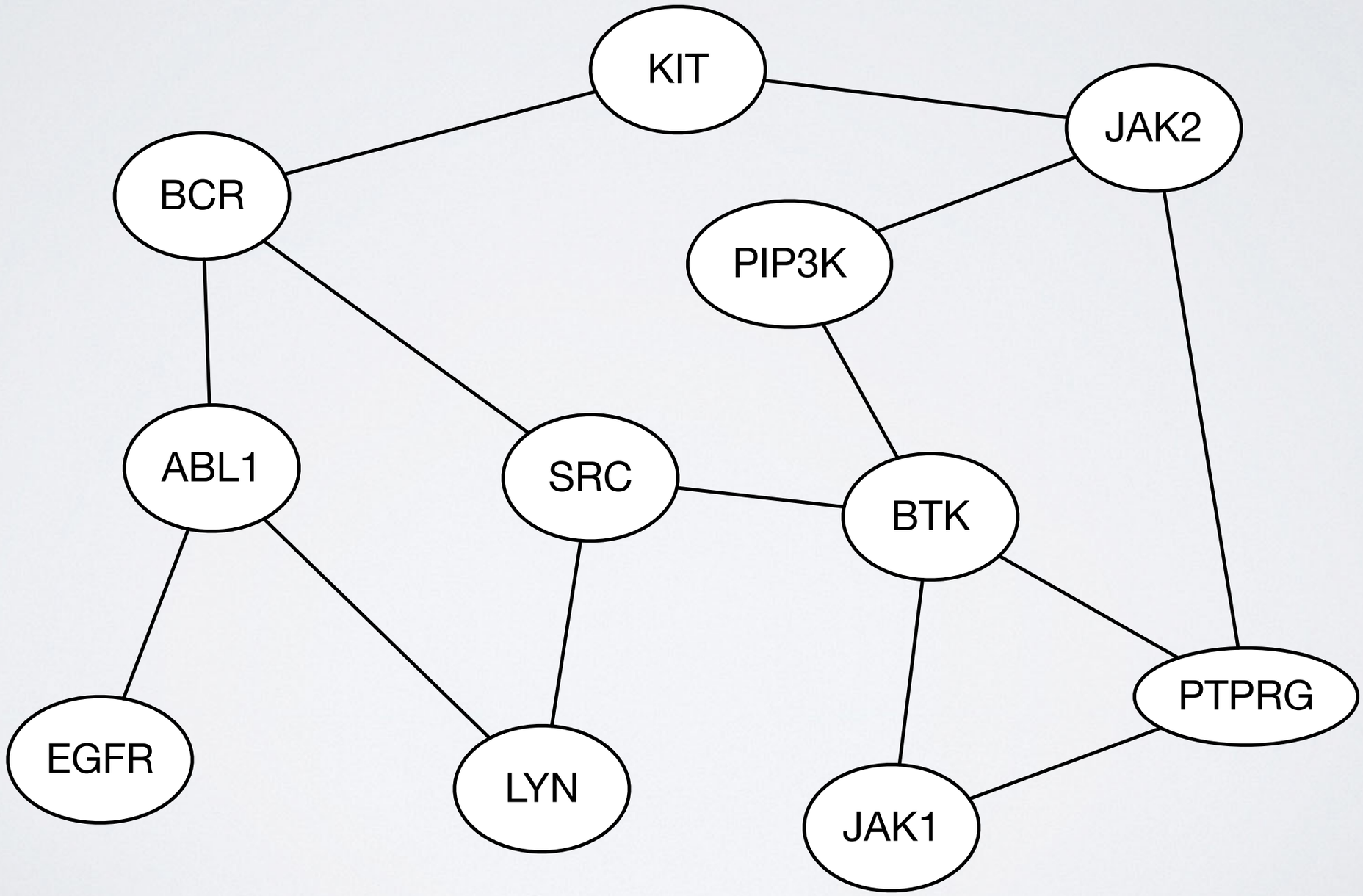


4.Calculating centralities (es. Degree)

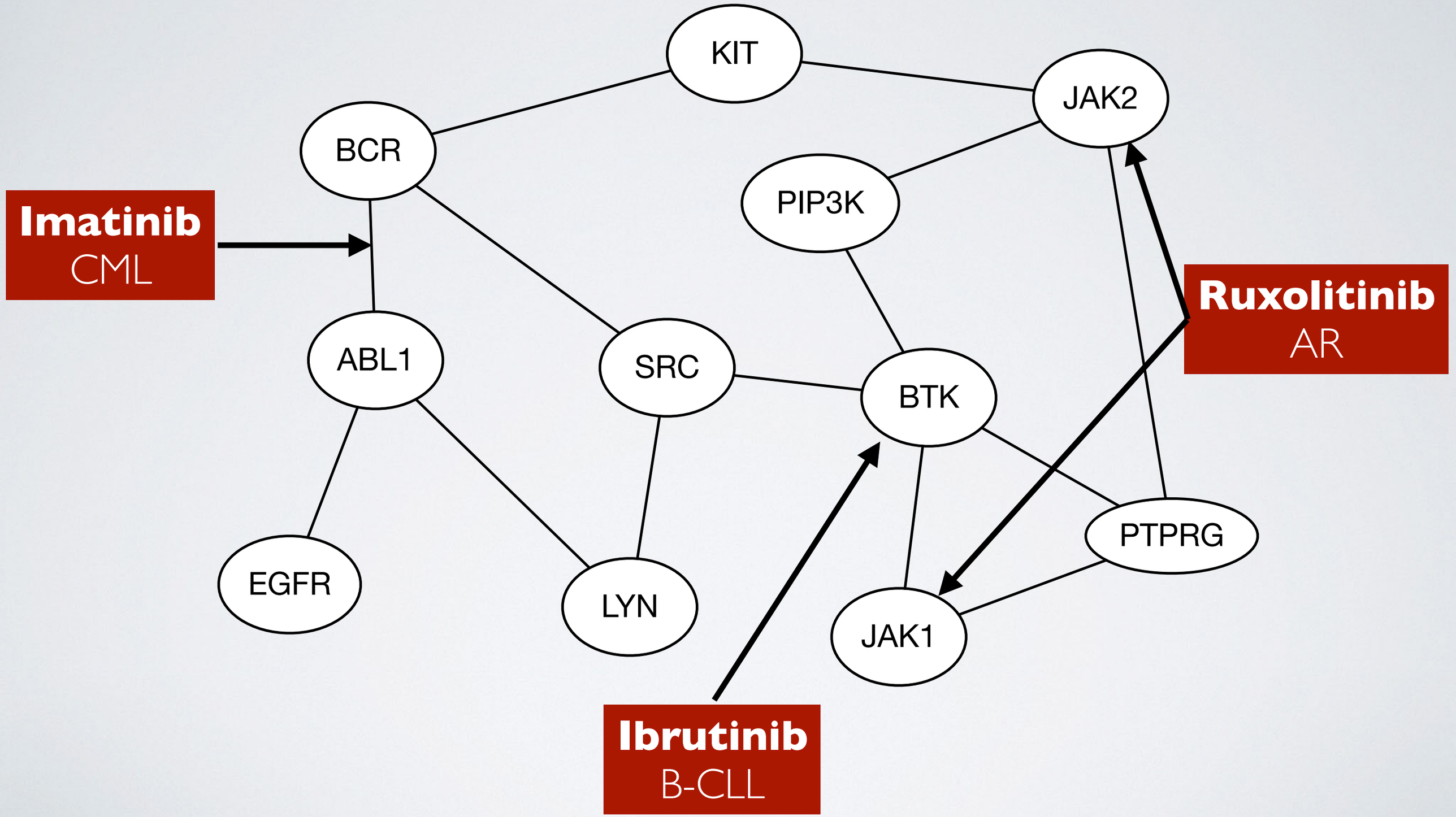


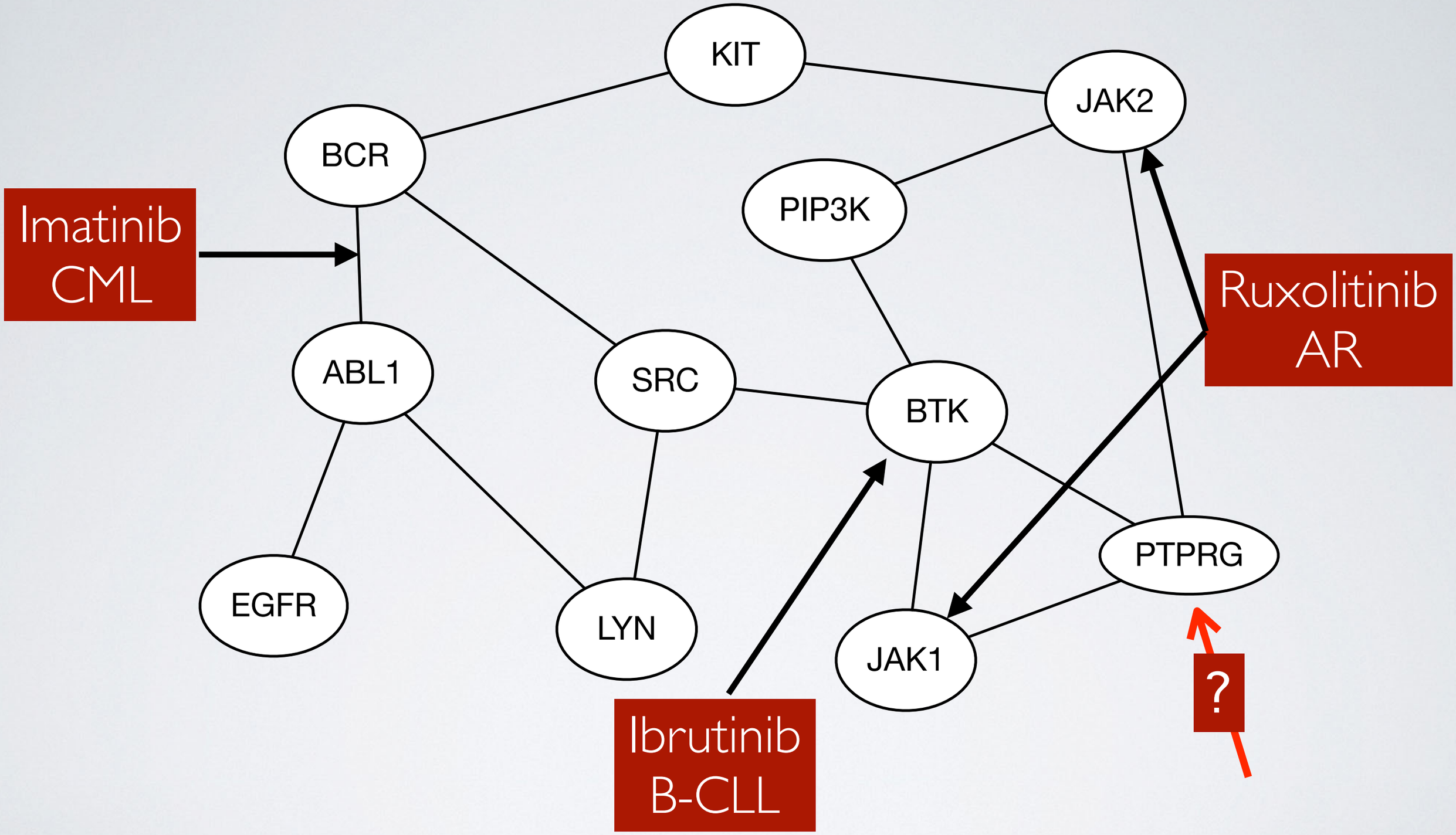
5. Identification of kinases and phosphatases with higher centrality indexes (three interpolated indexes)

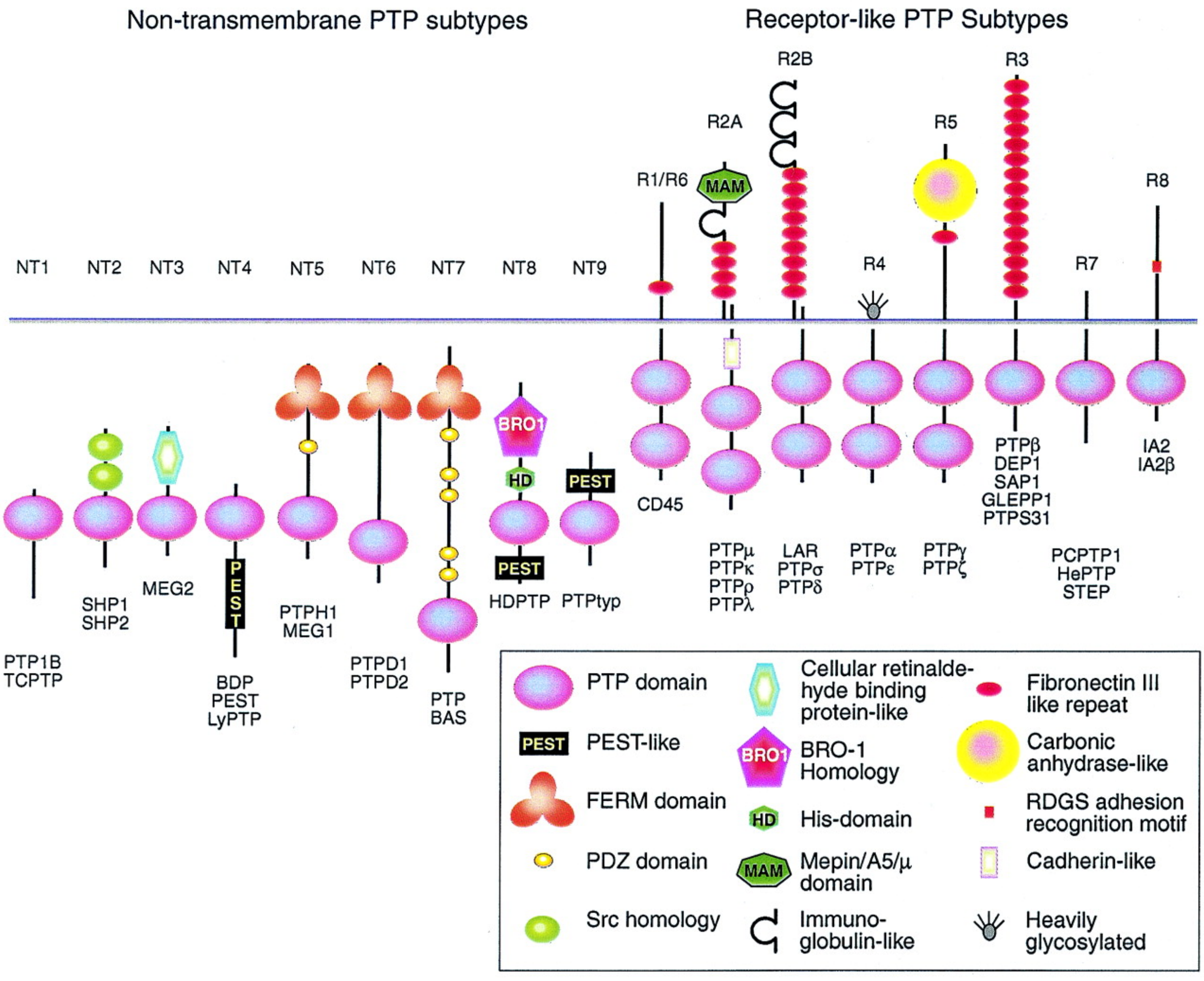
"Distillate" from CML_B-CLL intersection
474 nodes (proteins)
13791 binary interactions

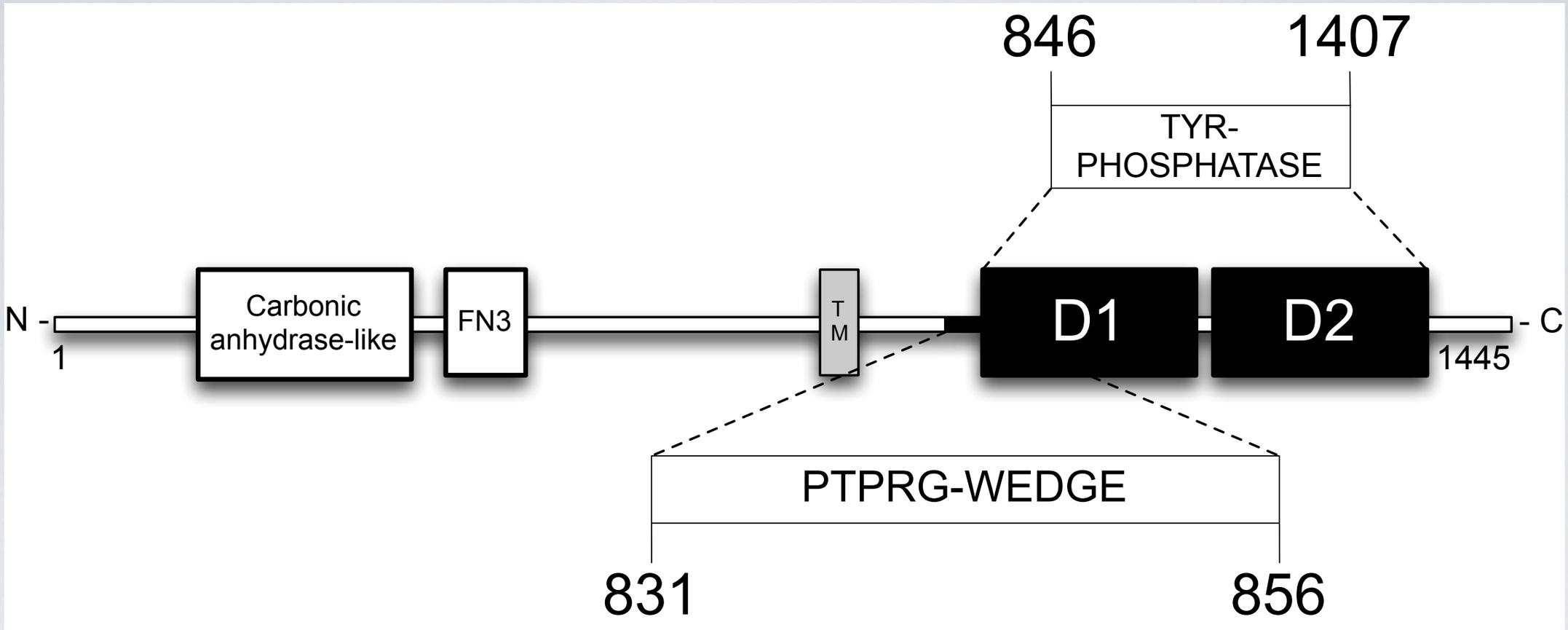


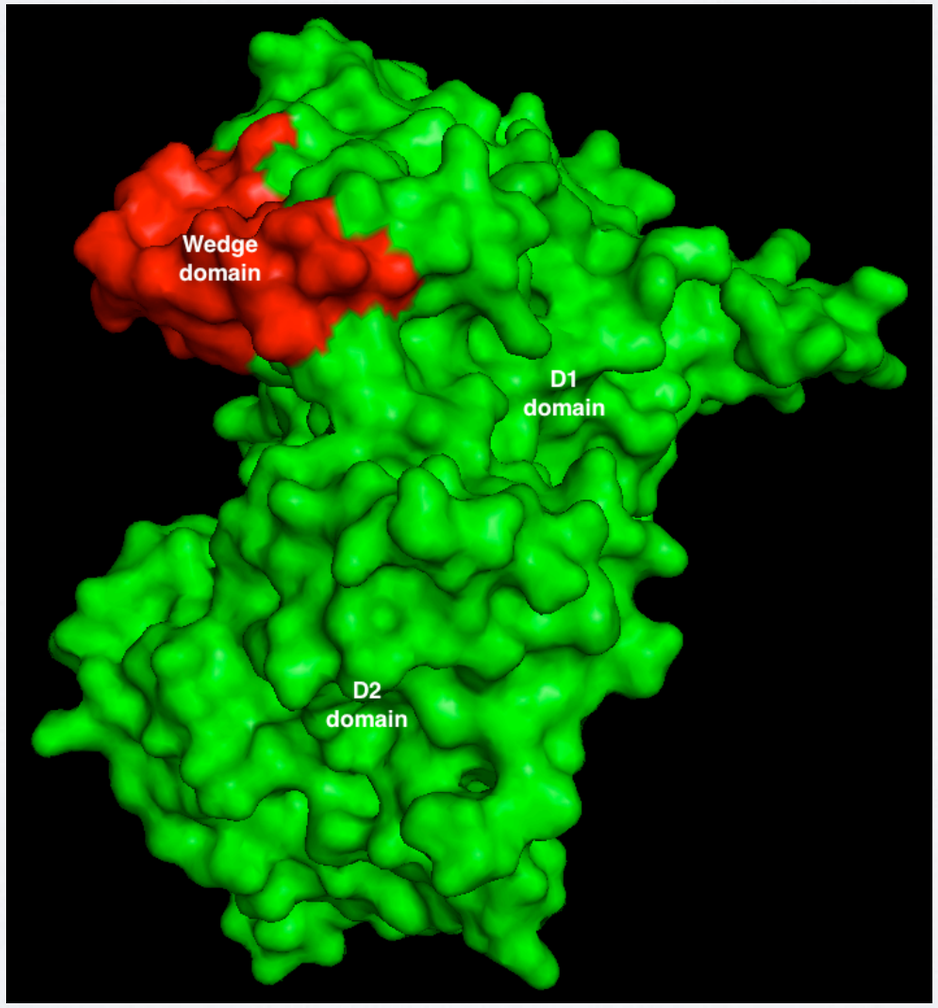
6. Possible targets of novel anti-leukemia drugs for both CML and/or B-CLL




KIT
JAK2
BCR
PIP3K
Imatinib
CML
Ruxolitinib
AR
ABL1
SRC
BTK
EGFR
LYN
PTPRG
JAK1
?
Ibrutinib
B-CLL


Non-transmembrane PTP subtypes
Receptor-like PTP Subtypes
NT1
NT2
NT3
NT4
NT5
NT6
NT7
NT8
NT9
R1/R6
R2A
R2B
R3
R4
R5
R7
R8
MAM
BRO1
HD
PEST
PTP1B
TCPTP
SHP1
SHP2
MEG2
BDP
PEST
LyPTP
PTPH1
MEG1
PTPD1
PTPD2
PTP
BAS
HDPTP
PTPtyp
CD45
PTPβ
DEP1
SAP1
GLEPP1
PTPS31
IA2
IA2β
PTPμ
PTPκ
PTPρ
PTPλ
LAR
PTPσ
PTPδ
PTPα
PTPε
PTPγ
PTPζ
PCPTP1
HePTP
STEP
PTP domain
PEST-like
FERM domain
PDZ domain
Src homology
Cellular retinalde-hyde binding protein-like
BRO-1 Homology
His-domain
Mepin/A5/μ domain
Immuno-globulin-like
Fibronectin III like repeat
Carbonic anhydrase-like
RDGS adhesion recognition motif
Cadherin-like
Heavily glycosylated


846
1407
TYR-
PHOSPHATASE
N -
1
Carbonic
anhydrase-like
FN3
T
M
D1
D2
1445
- C
PTPRG-WEDGE
831
856


Wedge
domain
D1
domain
D2
domain

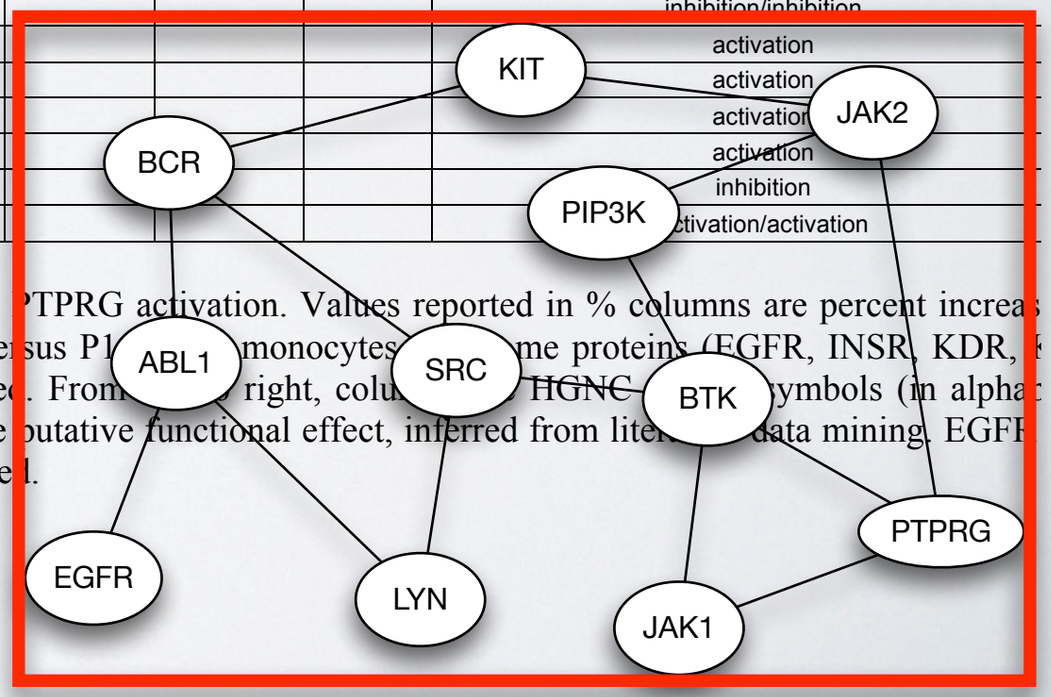
Proteomic analysis of PTPRG targets of tyrosine phosphorylation

Table I. Summary of Kinexus high-throughput phospho-proteomics data

(Mirenda M, Toffali L, Montresor A, Scardoni G, Sorio C and Laudanna C. Protein tyrosine phosphatase, receptor type, gamma (PTPRG) is a JAK phosphatase and negatively regulates leukocyt integrin activation. Journal of Immunology 2015 PubMed PMID: 25624455.)

HGNC protein symbols	p-Site	%	p-Site	%	p-Site	%	p-Site	%	Functional effect of P1-WD
ABL1	Y412	-40							inhibition
BLNK	Y84	19							activation
BMX	Y40	-7							inhibition
BTK	Y223	-58							inhibition
CDK2	Y15	282							inhibition
CTTN	Y470	21							inhibition
DAB1	Y198	-34							inhibition
DOK2	Y142	9							activation
EGFR	Y1068	-13	Y1148	-16	Y1173	68			inhibition/inhibition/activation
ERBB2	Y1248	34							activation
GRIN2B	Y1474	29							activation
INSR	Y999	58	Y1189/Y1190	27					activation/activation
ITGB1	Y783	-11							inhibition
JAK2	Y1007/Y1008	-33							inhibition
KDR	Y1054	-37	Y1054+Y1059	22					inhibition
KIT	Y703	-60	Y730	-47					inhibition
LIMK1	Y507	-10							inhibition
MET	Y1003	-47							inhibition
PDGFRA	Y742	133	Y754	86					activation
PDGFRB	Y716	-36							inhibition
PRKCD	Y313	189							activation
PTK2	Y397	30	Y577	-60	Y576	65	Y861	-11	activation/inhibition/activation/inhibition
PXN	Y118	144	Y31	19					activation/activation
SHC1	Y349+Y350	-10							inhibition
SRC	Y419	-6	Y530	94					inhibition/inhibition
STAT1	Y701	38							activation
STAT2	Y690	61							activation
STAT3	Y705	182							activation
STAT5A	Y694	50							activation
VCL	Y821	-20							inhibition
ZAP70	Y292	-46	Y315+Y319	12					activation/activation

The table shows the 31 identified proteins whose tyrosine phosphorylation is affected by PTPRG activation. Values reported in % columns are percent increas decrease of protein tyrosine phosphorylation upon fMLP triggering, in P1-WD-treated versus P1 monocytes. Some proteins (EGFR, INSR, KDR, PDGFRA, PTK2, PXN, SRC and ZAP70) multiple phosphotyrosine residues are detected. From right, columns HGNC symbols (in alphabetical order), phosphosites (p-Sites), % changes of phosphorylation (induced by P1-WD) and the putative functional effect, inferred from literature data mining. EGFR, PTK2 are highlighted in gray since the functional effect could not be unambiguously inferred.



Conclusion

Comparative network topological analysis coupled to phospho-proteomic analysis shows that:

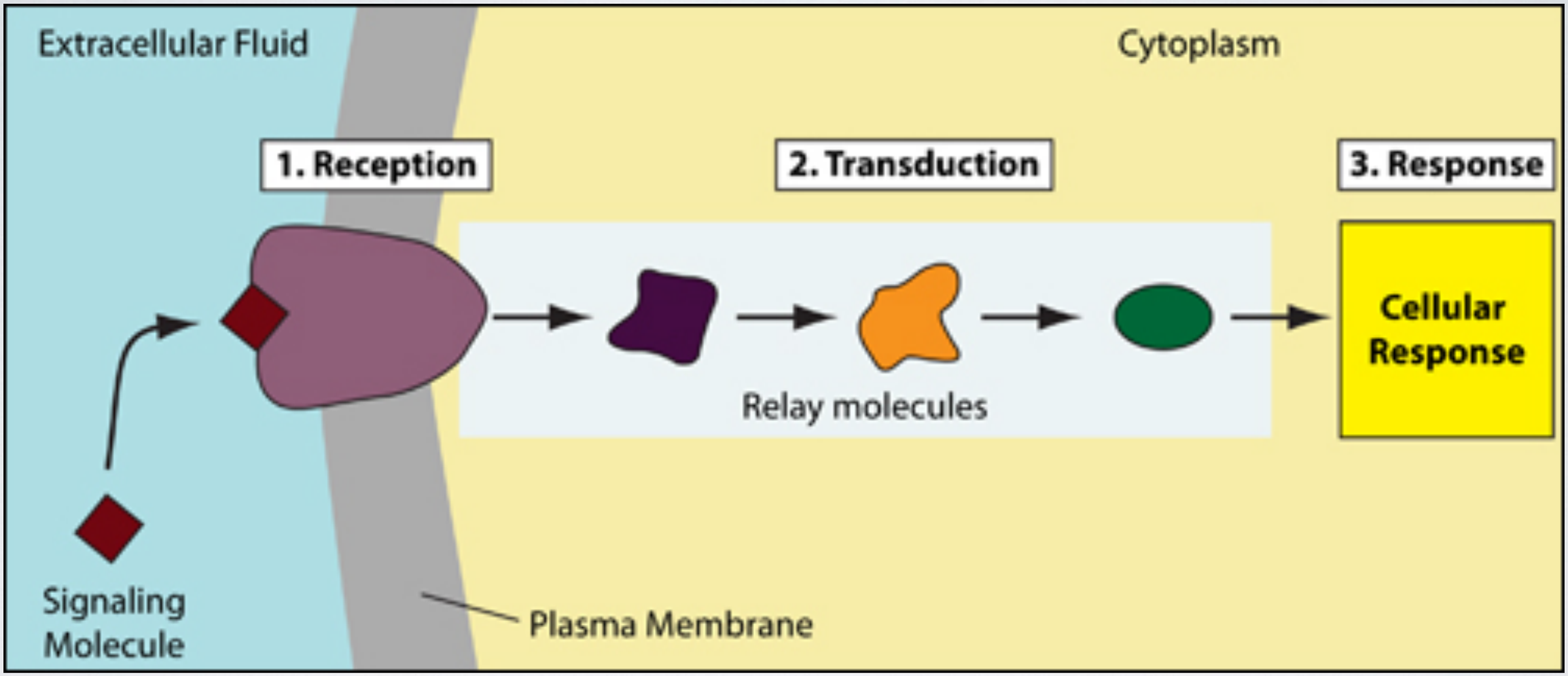
B-CLL and CML may have at least 4 common drive genes of the category of protein kinases and phosphatases

At least 3 known compounds can be **re-purposed**

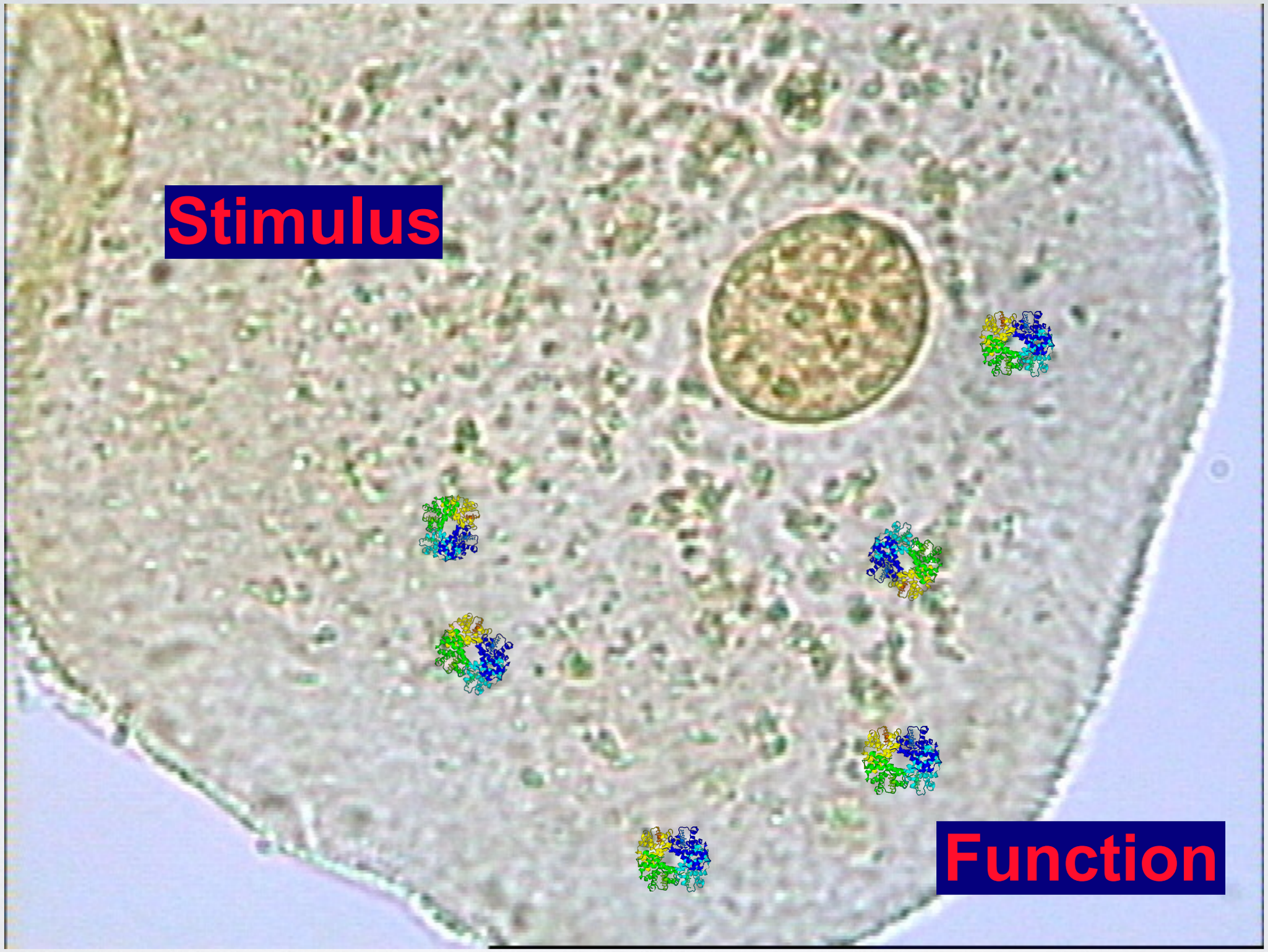
PTPRG may be a novel target for drug development and therapy in leukaemia

But why the topology of a “graph” simply drawn on a 2D plan … should predict a “real” biological function within a living cell?

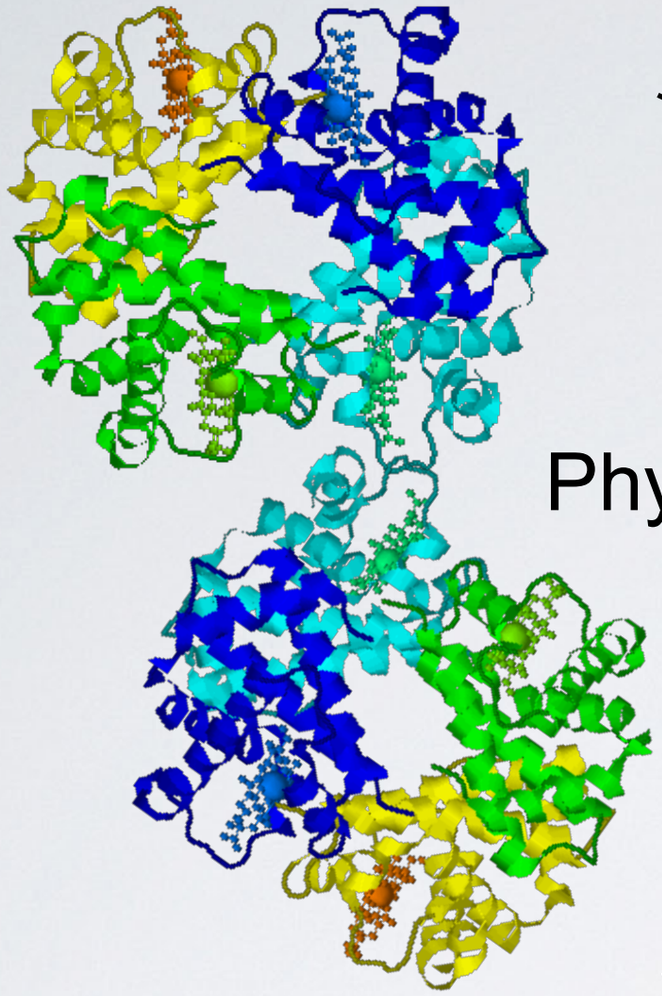
The receptor-function paradigm



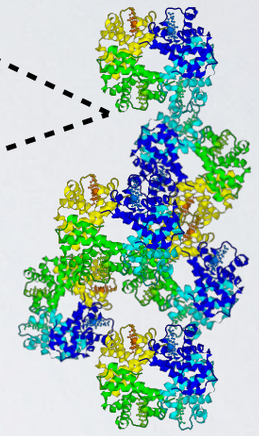
Signal transduction is relies on molecular domains

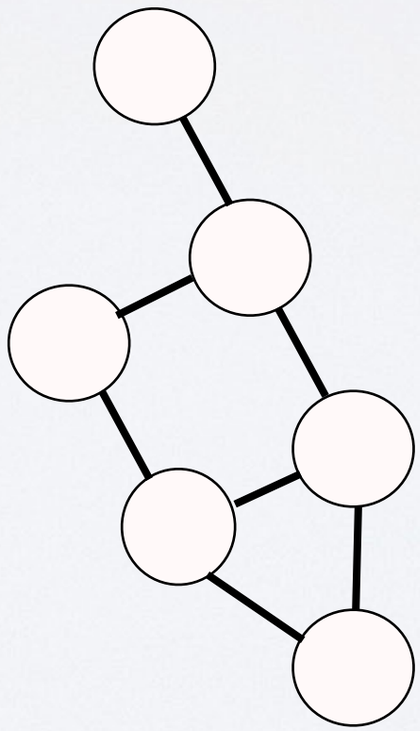

Stimulus
Function

Graphs are STATIC ABSTRACTIONS
of signalling networks

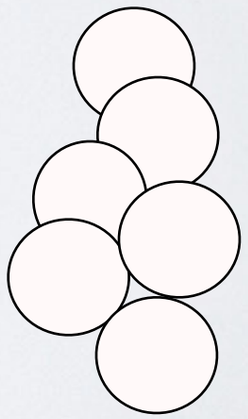


Physical interaction





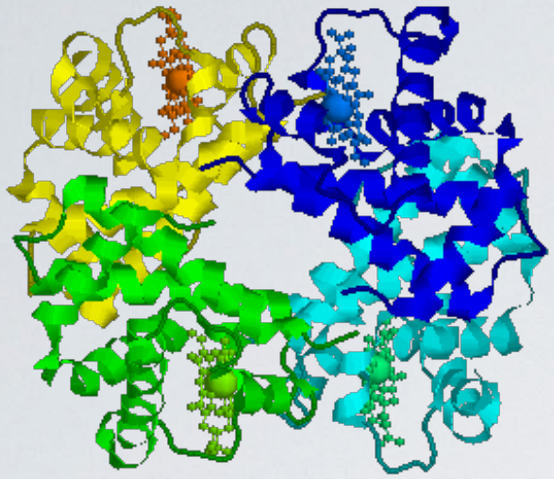
Euler abstraction



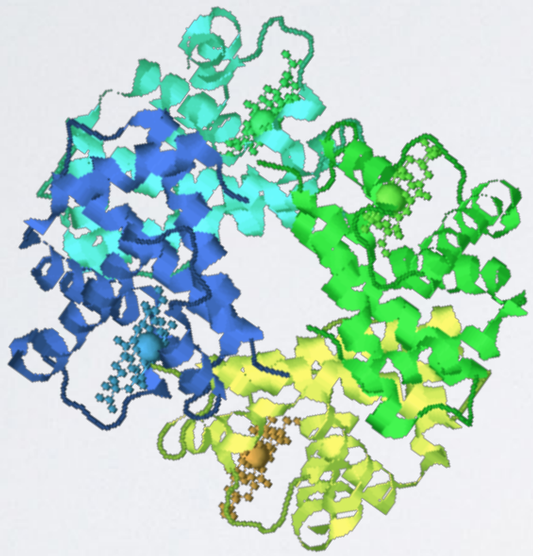
The graph is "*collinear*" with the physical interaction between proteins

Graph topology is "*collinear*" with the topology of the protein complex

a



b



Molecular domains

A

B

A

B

A A A

MTRLPDSF ..... STALYY = GCAATGATA ..... GGTAAGCT

MFFDSTYA ..... PPRPPY = AATGCCTTAG .... AATCCGTA

B B B

The topological structure of a graph is collinear with the topological structure of the protein networks which, in turn, is collinear with the architecture of the protein domains, which, in turn, are collinear with the structure of genes.

THEN

The topological structure of protein networks is written in the sequence of the genes.

The genome is fully collinear with the topology of the networks and determines it!

The topology of protein-protein interaction (PPI) networks does reflect the DNA sequence!

PPI network topology is written by evolution (genetic drift)

Evolution must follow specific rules (mutation on specific gene sites coding for protein domains), to generate gene/protein varieties, still capable of conserving specific network topological properties, facilitating cell adaptation and survival in the environment.

Mutations not respecting previous network architecture may generate *new properties* (evolution)
OR
may lead to *alteration of homeostasis* (cancer)!

Network building
(from experiments)

+

Computation
(from mathematics)

Topological
(contextual)

Is topology alone enough?

Inferred
mechanisms

Extracting information
from the system complexity
(training complexity)

Systems Biology is an integrated, multidisciplinary discipline that is concerned with studying complex biological systems (living).

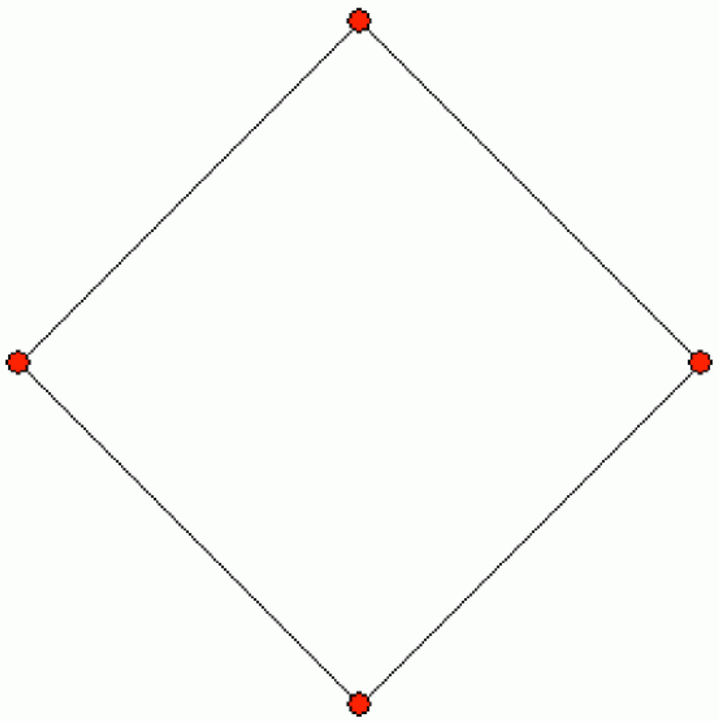
"Theoretical biology, integrative biology, network biology, **multidimensional biology**, network medicine, personalized medicine, precision medicine, etc.”

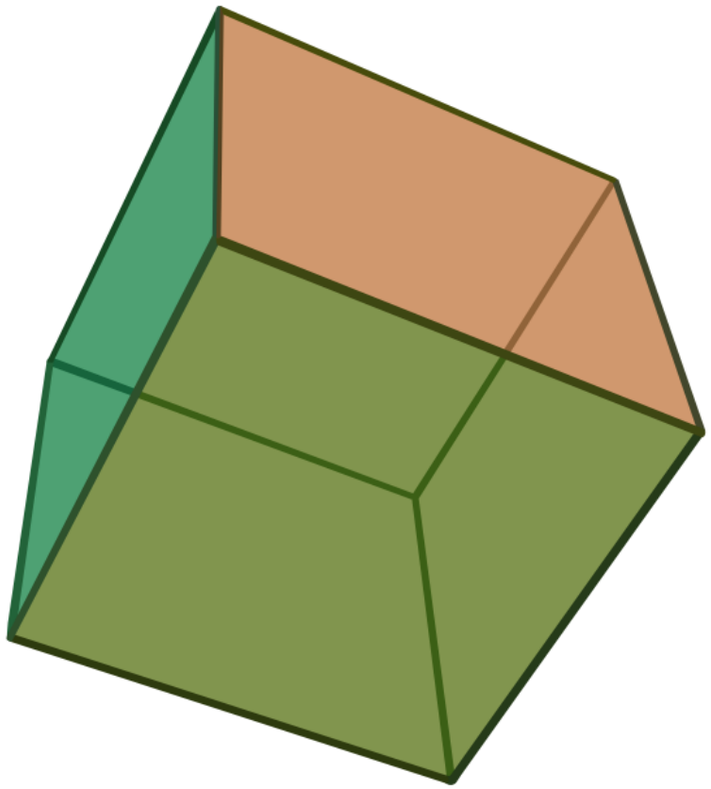
Multi-dimensional biology?

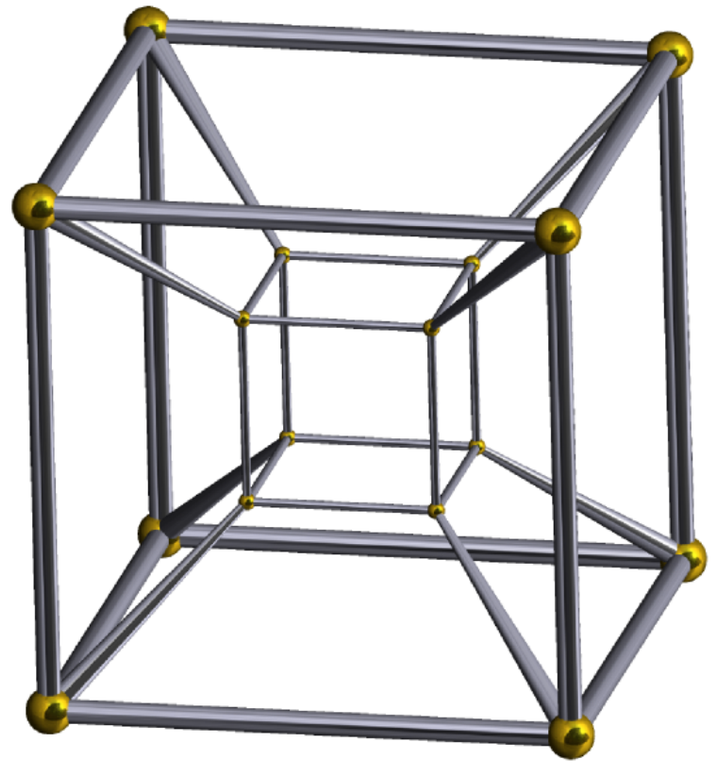
Definition:

A **dimension** corresponds to the number of degrees of freedom available to "*movement*" in a "space" (*informational space*).

In common language, the dimensions of an object are the measurements (numbers) defining its shape and its size, that is the 3D space it occupies.







2

3
+
Time

n

Degree of freedom in space-time

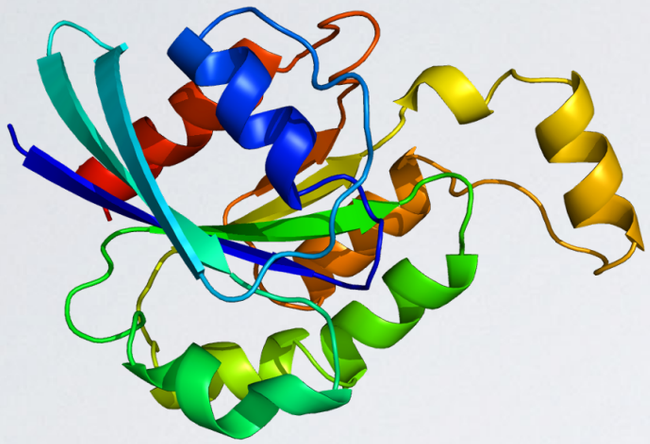
This definition derives from to the common use made in normal geometry, but it may greatly differs with regard to different physical contexts.

Object ⟶ Dimension = **Degree of freedom**

A physical object can have <u>many</u> degree of freedom, NOT necessarily limited to a 4-dimensional space BUT in a space of functions (properties) (what is called **<u>phase-space</u>**)

In dynamical systems theory is called **<u>phase-space</u>** of a system the space whose points represent uniquely all and only the possible functional states of the system. In general the phase space has <u>as many dimensions as degrees of freedom</u> of the system.

Small GTPase RhoA



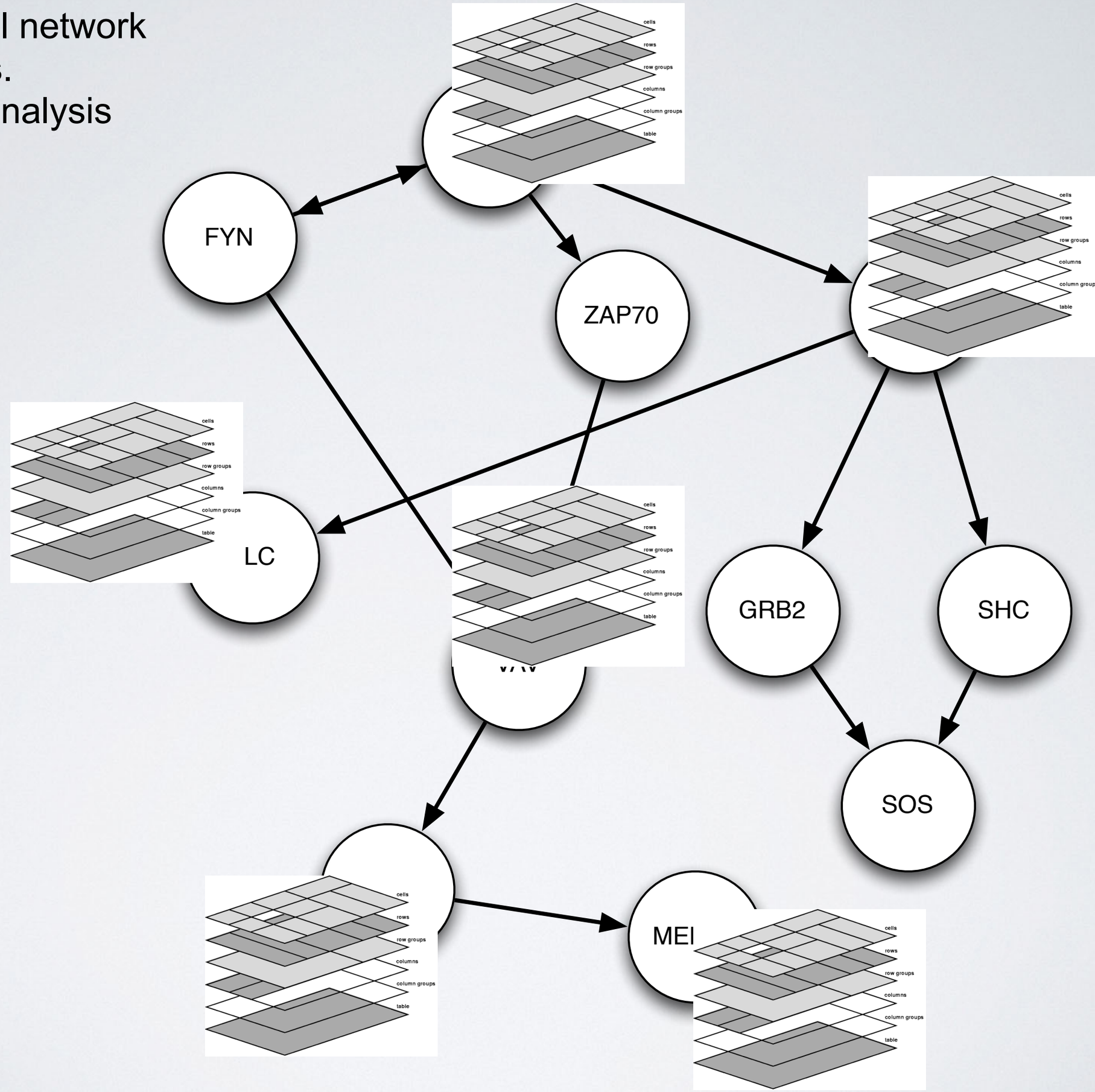
3 space dimensions

- Cytoskeleton assembly
- Adhesion regulation
- Integrin activation
- Gene expression
- Cell proliferation
- ... *n*

n dimensions in space of “functions” (= phase space)

RhoA has 3 space dimensions + *n* functional dimensions

Multidimensional network analysis. Multilayered analysis



LETTERS

A network-based analysis of systemic inflammation in humans

Steve E. Calvano[1]*, Wenzhong Xiao[2]*, Daniel R. Richards[3], Ramon M. Felciano[3], Henry V. Baker[4,5], Raymond J. Cho[3], Richard O. Chen[3], Bernard H. Brownstein[6], J. Perren Cobb[6], S. Kevin Tschoeke[5], Carol Miller-Graziano[7], Lyle L. Moldawer[5], Michael N. Mindrinos[2], Ronald W. Davis[2], Ronald G. Tompkins[8], Stephen F. Lowry[1] & the Inflammation and Host Response to Injury Large Scale Collaborative Research Program†

Oligonucleotide and complementary DNA microarrays are being used to subclassify histologically similar tumours, monitor disease progress, and individualize treatment regimens[1–5]. However, extracting new biological insight from high-throughput genomic studies of human diseases is a challenge, limited by difficulties in recognizing and evaluating relevant biological processes from huge quantities of experimental data. Here we present a structured network knowledge-base approach to analyse genome-wide transcriptional responses in the context of known functional interrelationships among proteins, small molecules and phenotypes. This approach was used to analyse changes in blood leukocyte gene expression patterns in human subjects receiving an inflammatory stimulus (bacterial endotoxin). We explore the known genome-wide interaction network to identify significant functional modules perturbed in response to this stimulus. Our analysis reveals that the human blood leukocyte response to acute systemic inflammation includes the transient dysregulation of leukocyte bioenergetics and modulation of translational machinery. These findings provide insight into the regulation of global leukocyte activities as they relate to innate immune system tolerance and increased susceptibility to infection in humans.

Inflammation is a hallmark of many human diseases[6–8]. We focus on blood leukocytes and other tissues of critically injured patients, in order to better elucidate the mechanisms underlying systemic inflammatory responses[9]. This approach cannot be fully replicated using animal models or human cell lines, and studies of injury in humans can be complicated by antecedent illnesses and concurrent treatment regimes that may alter the recovery process. To our knowledge, no study has evaluated the genome-wide response to systemic inflammation in the context of a fully predictable recovery. Here we combine genome-wide expression analysis with a new bioinformatics method to identify functional networks responsible for the systemic activation and spontaneous resolution of a well-defined inflammatory challenge.

Gene expression in whole blood leukocytes was determined immediately before and at 2, 4, 6, 9 and 24 h after the intravenous administration of bacterial endotoxin to four healthy human subjects. Four additional subjects were studied under identical conditions but without endotoxin administration. The infusion of endotoxin activates innate immune responses and presents with physiological responses of brief duration[10]. Notably, there is an initial proinflammatory phase and a subsequent counterregulatory phase, with resolution of virtually all clinical perturbations within 24 h.

K-means cluster and principal component analyses were first used to visualize the overall response to endotoxin administration. Figure 1a reveals probe sets clustered by *K*-mean analysis, where each bin has a distinct endotoxin-induced temporal pattern. The signal intensity of 5,093 probe sets—representing 3,714 unique genes—out of a total of >44,000 probe sets changed significantly in response to endotoxin, whereas no significant changes were observed in control subjects (estimated false discovery rate <0.1%). Of the 5,093 probe sets identified, over half showed reduced abundance at 2, 4, 6 and 9 h, returning to baseline by 24 h (see bins 0–4). In contrast, a smaller number of probe sets were induced by 2 h (bins 5, 6), and the remaining probe sets showed a delayed response, peaking at 4–9 h but returning to baseline by 24 h (bins 7–9).

Cluster and principal component analyses describe overall changes in apparent gene expression, but provide few insights into the biological processes and signalling networks invoked in propagation and resolution of the inflammatory response. Identifying the perturbed biological networks underlying this complex clinical phenotype requires systematic analysis in the context of known mammalian biology, derived from basic and clinical research.

Using a web-based entry tool developed by Ingenuity Systems Inc., findings presented in peer-reviewed scientific publications were systematically encoded into an ontology by content and modelling experts. Using over 200,000 full-text scientific articles, a knowledge base of more than 9,800 human, 7,900 mouse and 5,000 rat genes was manually curated and supplemented with curated relationships parsed from MEDLINE abstracts. A molecular network of direct physical, transcriptional and enzymatic interactions observed between mammalian orthologues—the observed 'interactome'—was computed from this knowledge base. The resulting network contains molecular relationships involving over 8,000 orthologues with a high degree of connectivity. On average, individual genes have 11.5 interaction partners (median 4.0), of which 7.2 represent direct physical interactions (median 3.0). Every gene interaction in the network is supported by published information. For example, the

[1]Department of Surgery, UMDNJ-Robert Wood Johnson Medical School, New Brunswick, New Jersey 08903, USA. [2]Stanford Genome Technology Center, Palo Alto, California 94304, USA. [3]Ingenuity Systems Inc, Mountain View, California 94043, USA. [4]Departments of Molecular Genetics and Microbiology, and [5]Department of Surgery, University of Florida College of Medicine, Gainesville, Florida 32610, USA. [6]Department of Surgery, Washington University in St Louis, St Louis, Missouri 63110, USA. [7]Department of Surgery, University of Rochester School of Medicine, Rochester, New York 14642, USA. [8]Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA.
*These authors contributed equally to this work.
†Lists of participants and affiliations appear at the end of the paper

Cell, Vol. 118, 217–228, July 23, 2004, Copyright ©2004 by Cell Press

Single Cell Profiling of Potentiated Phospho-Protein Networks in Cancer Cells

Jonathan M. Irish,[1] Randi Hovland,[2,3] Peter O. Krutzik,[1] Omar D. Perez,[1] Øystein Bruserud,[3,4] Bjørn T. Gjertsen,[3,4] and Garry P. Nolan[1,*]
[1]Department of Microbiology & Immunology
Baxter Laboratory of Genetic Pharmacology
Stanford University
Stanford, California 94305
[2]Center for Medical Genetics and Molecular Medicine
Haukeland University Hospital and Proteomic Unit (PROBE)
University of Bergen
Bergen
Norway
[3]Department of Internal Medicine
Haematology Section
Haukeland University Hospital
Bergen
Norway
[4]Institute of Medicine
Haematology Section
University of Bergen
Bergen
Norway

Summary

Altered growth factor responses in phospho-protein-driven signaling networks are crucial to cancer cell survival and pathology. Profiles of cancer cell signaling networks might therefore identify mechanisms by which such cells interpret environmental cues for continued growth. Using multiparameter flow cytometry, we monitored phospho-protein responses to environmental cues in acute myeloid leukemia at the single cell level. By exposing cancer cell signaling networks to potentiating inputs, rather than relying upon the basal levels of protein phosphorylation alone, we could discern unique cancer network profiles that correlated with genetics and disease outcome. Strikingly, individual cancers manifested multiple cell subsets with unique network profiles, reflecting cancer heterogeneity at the level of signaling response. The results revealed a dramatic remodeling of signaling networks in cancer cells. Thus, single cell measurements of phospho-protein responses reveal shifts in signaling potential of a phospho-protein network, allowing for categorizing of cell network phenotypes by multidimensional molecular profiles of signaling.

Introduction

Intracellular signaling and interpretation of environmental cues play central roles in cancer cell initiation and maintenance. Actions that lead to cancer cell progression include mutations to key signaling proteins as well as epigenetic changes to gene expression patterns (Hanahan and Weinberg, 2000). Cancer genesis occurs in a stepwise progression, has underlying stochastic elements, and is reflective of the genetic selection of the cancer in the face of both immune system action and the environmental requirements of cancer cells. The signaling profile of any given cancer cell is therefore the sum of numerous influences: epigenetic, genetic, and microenvironmental. The current molecular understanding of cancer signaling rests largely on extrapolations from studies of cell lines and as such is not adequately representative of the signaling phenotypes of a complex population of cancer cells in the body. In contrast, the heterogeneity of cancer cell responses to therapy can be thought of as mirroring the signaling differences that have arisen during evolution of the cancer cell population in the body. Until now, this cell by cell information on cancer cell populations—required to model signaling network pathologies that relate to cancer cell subsets and disease progression—has not been available for analysis.

Phospho-protein members of signaling cascades, and the kinases and phosphatases that interact with them, are required to initiate and regulate proliferative signals within cells. It might be predicted that genetic changes common in cancers, such as receptor tyrosine kinase mutations and other signaling-related cytogenetic alterations (Spiekermann et al., 2002; Wheatley et al., 1999), would change the potential of pre-existing signaling networks to respond to external stimuli and lead to identifiable patterns of signal transduction associated with gene mutation. For instance, acute myeloid leukemia (AML) is a cancer wherein dysregulated growth and inhibition of apoptosis lead to the accumulation of immature myeloid progenitor cells and oncogenic progression (Lowenberg et al., 1999). Two key parallel signal transduction networks active in cells that are considered progenitors of AML (Reya et al., 2001) are the STAT pathway (Coffer et al., 2000; Smithgall et al., 2000) and the Ras/MAPK pathway (Platanias, 2003). Several reports suggest that STATs, such as Stat3 and Stat5, are constitutively activated in AML (Benekli et al., 2002; Birkenkamp et al., 2001; Turkson and Jove, 2000; Xia et al., 1998). But, a causal link between basal STAT phosphorylation and leukemogenesis in primary patient material has not been demonstrated, despite significant evidence implicating these proteins in oncogenic processes (Benekli et al., 2003; Bowman et al., 2000; Buettner et al., 2002; Calo et al., 2003; Nieborowska-Skorska et al., 1999). Thought to act upstream of these pathways, abnormalities of the Flt3 (*fms-like tyrosine kinase 3*) receptor tyrosine kinase are detected in approximately 30% of AML patients and are well established as a negative prognostic indicator in AML (Gilliland and Griffin, 2002; Kottaridis et al., 2001; Thiede et al., 2002). Expression of mutant, activated Flt3 in cell lines has been observed to activate STAT and Ras/MAPK signaling (Hayakawa et al., 2000; Mizuki et al., 2000). However, basal levels of Stat5 phosphorylation have been reported to

*Correspondence: gnolan@stanford.edu

Example (1) of network multidimensional analysis

(*De Franceschi et al.*)

Virtual cell

=

GLOBAL Network = 14500 nodes (proteins) - 94600 interactions

The experiment: **red blood cells, either normal (WT) or PTPRE -/- (KO)**

↓

Protein phosphorylation patterns (2D-SDS-PAGE)

↓

Two LISTs of protein phosphorylated in WT and KO red blood cells

The LISTs are what are called “bioinformatic probes”

The lists are used to extract, from the GLOBAL network, two sub-networks specifically built upon the two sets of phosphorylated protein experimentally identified.

The phosphorylation level is called “**node attribute**” of the networks

Set-specific sub-networks (WT and KO) are ANNOTATED

WB

4 pH 8

~220
~100
~50
~40
~25
~15

4 pH 8

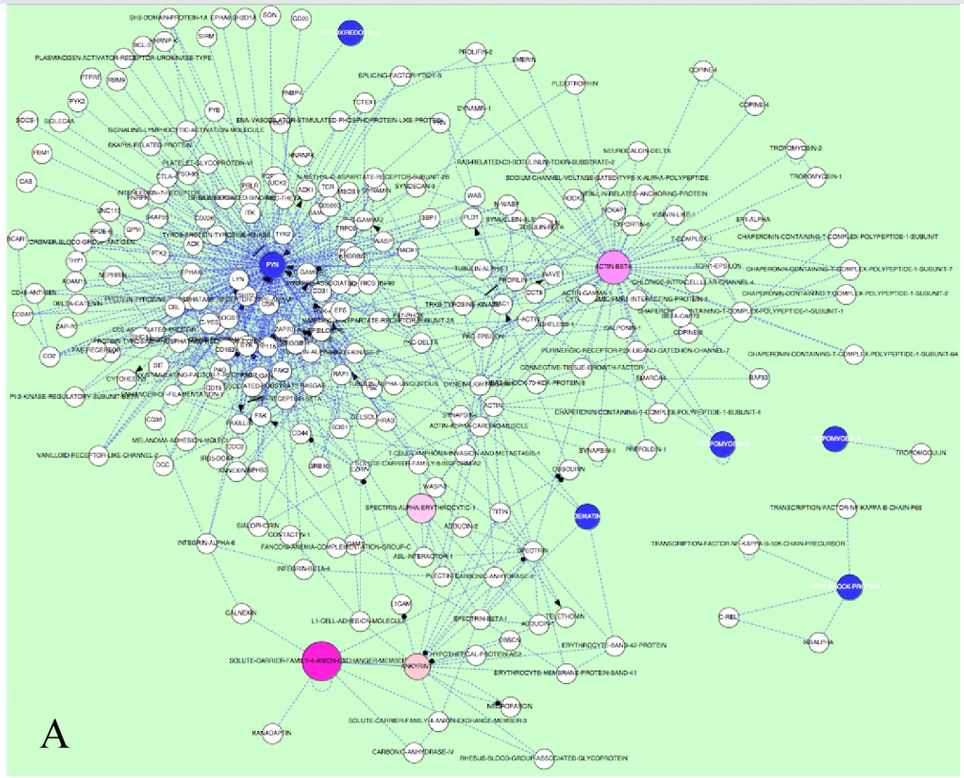
2D

~220
~100
~50
~40
~25
~15

+/+

-/-

WT

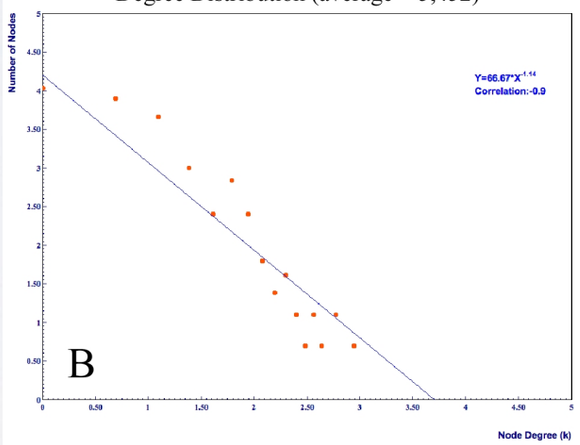


Nodes = 241
Edges = 724

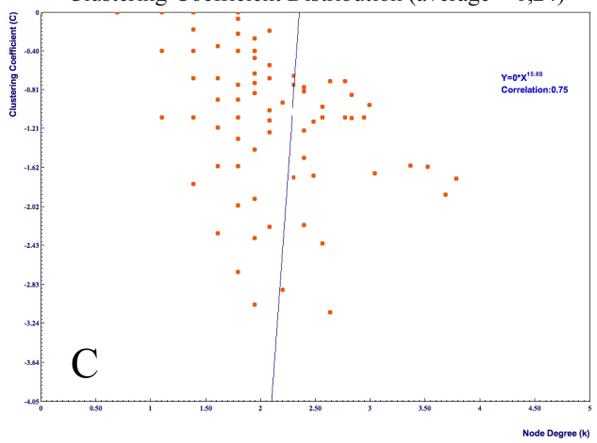
5

120

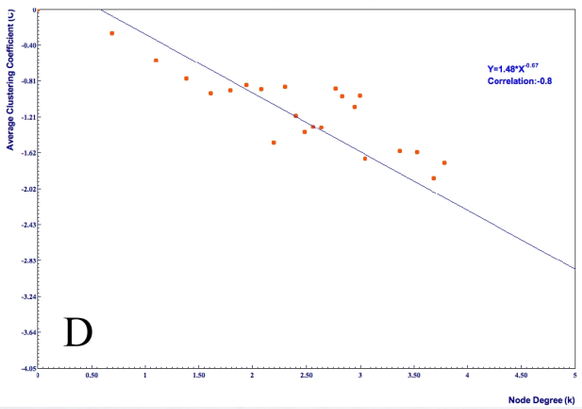
Degree Distribution (average = 5,452)



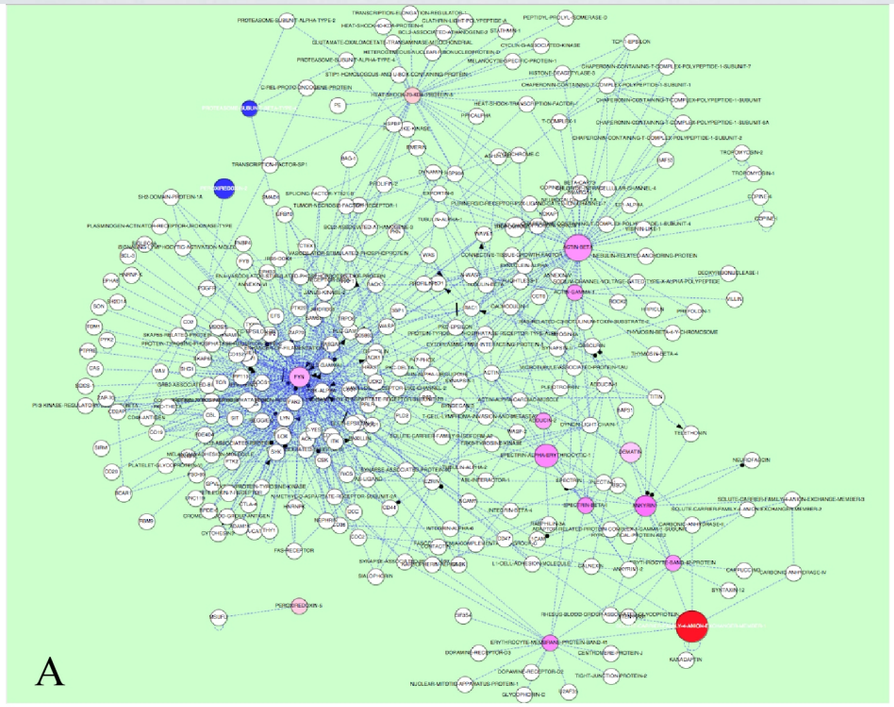
Clustering-Coefficient Distribution (average = 0,24)



Average Clustering Coefficient Distribution over Same Node Degree (average = 0,18)

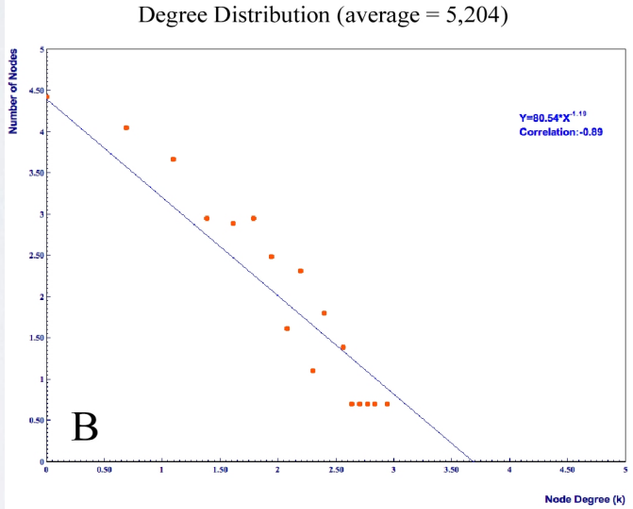


PTPRE -/-

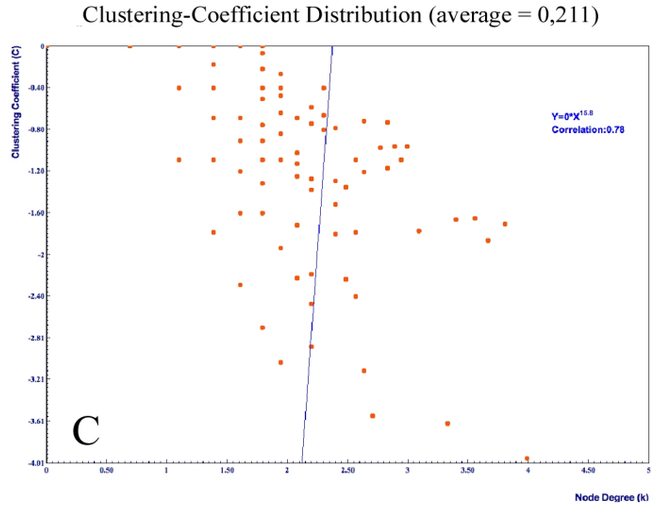


Nodes = 294
Edges = 838

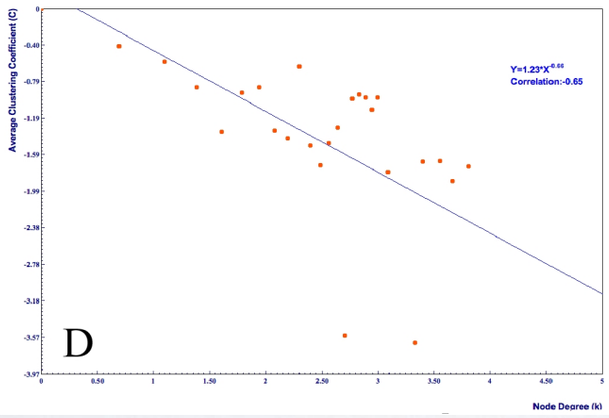
Degree Distribution (average = 5,204)



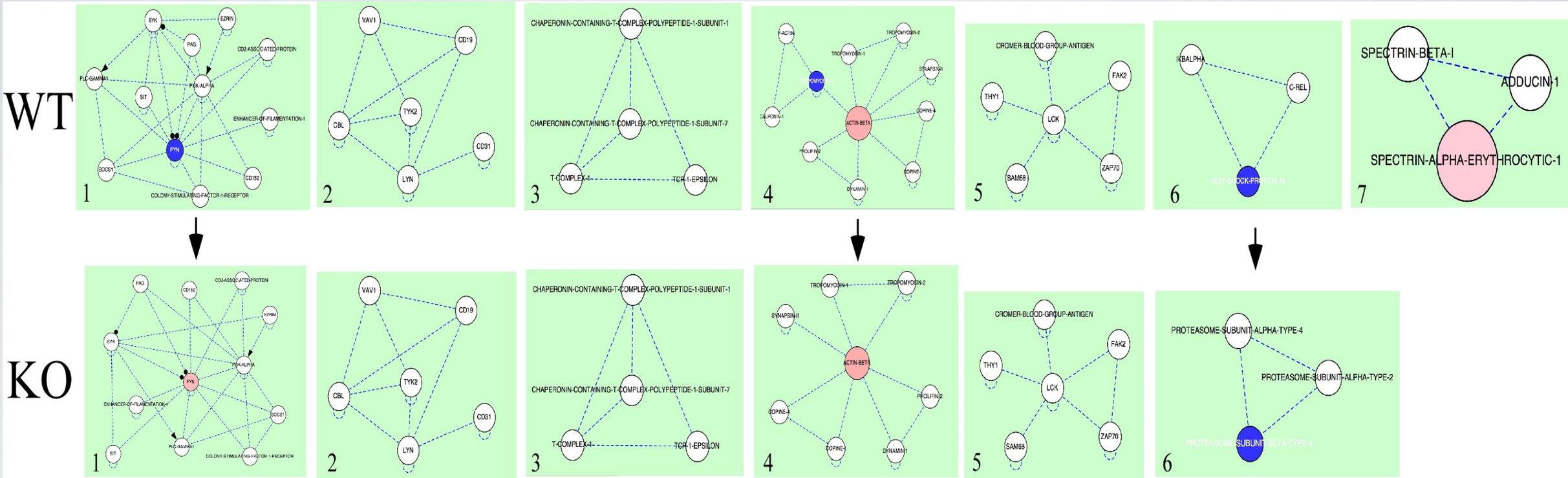
Clustering-Coefficient Distribution (average = 0,211)



Average Clustering Coefficient Distribution over Same Node Degree (average = 0,172)



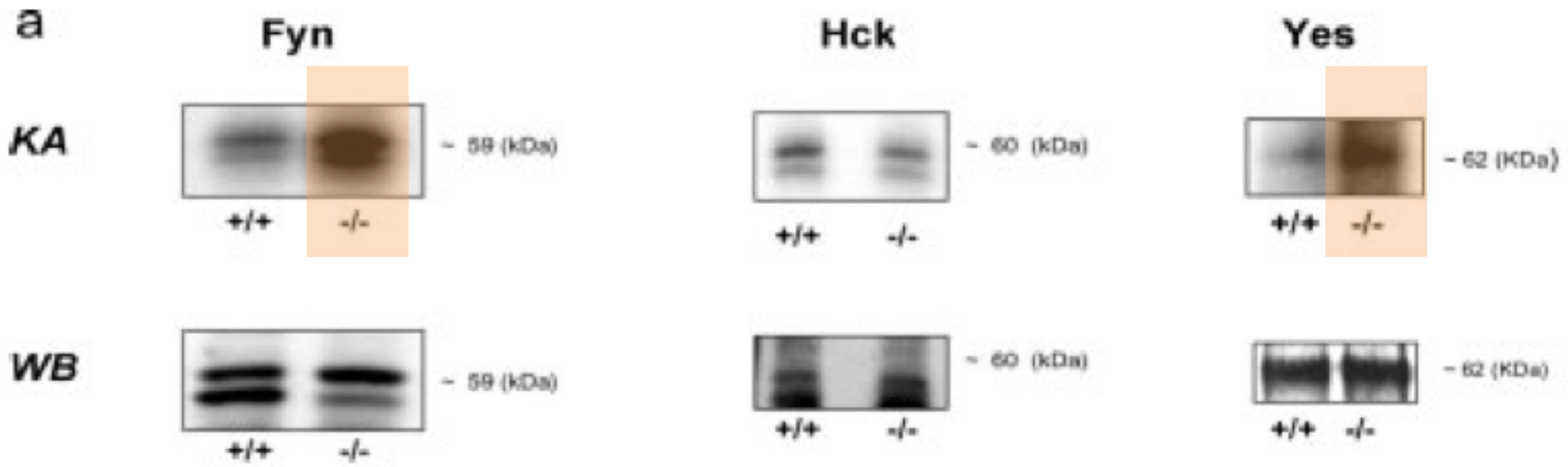
Decomposizione modulare



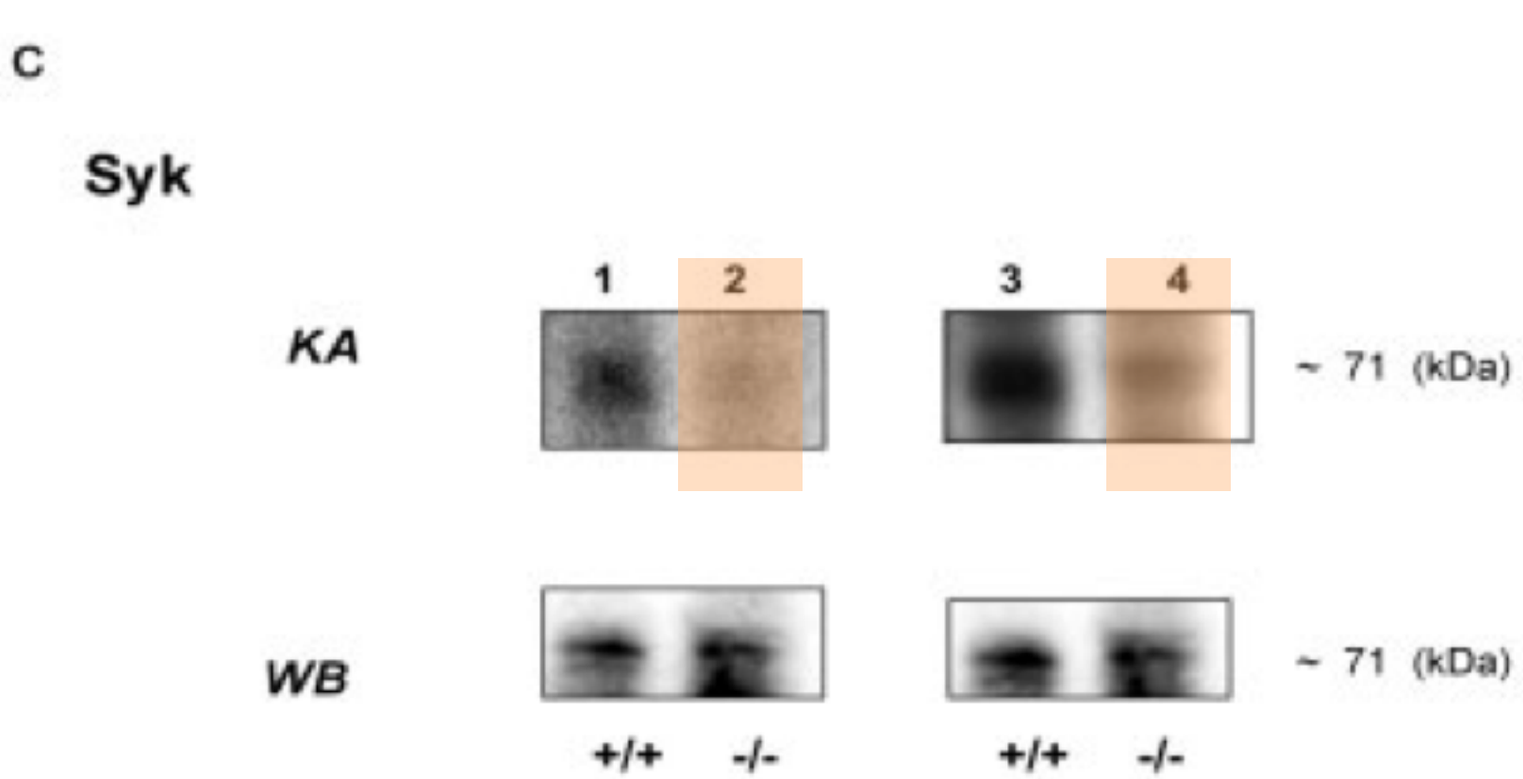
Fyn - Yes - Syk

Most relevant hits suggesting the functional effect of PTPRE deficiency

a



C

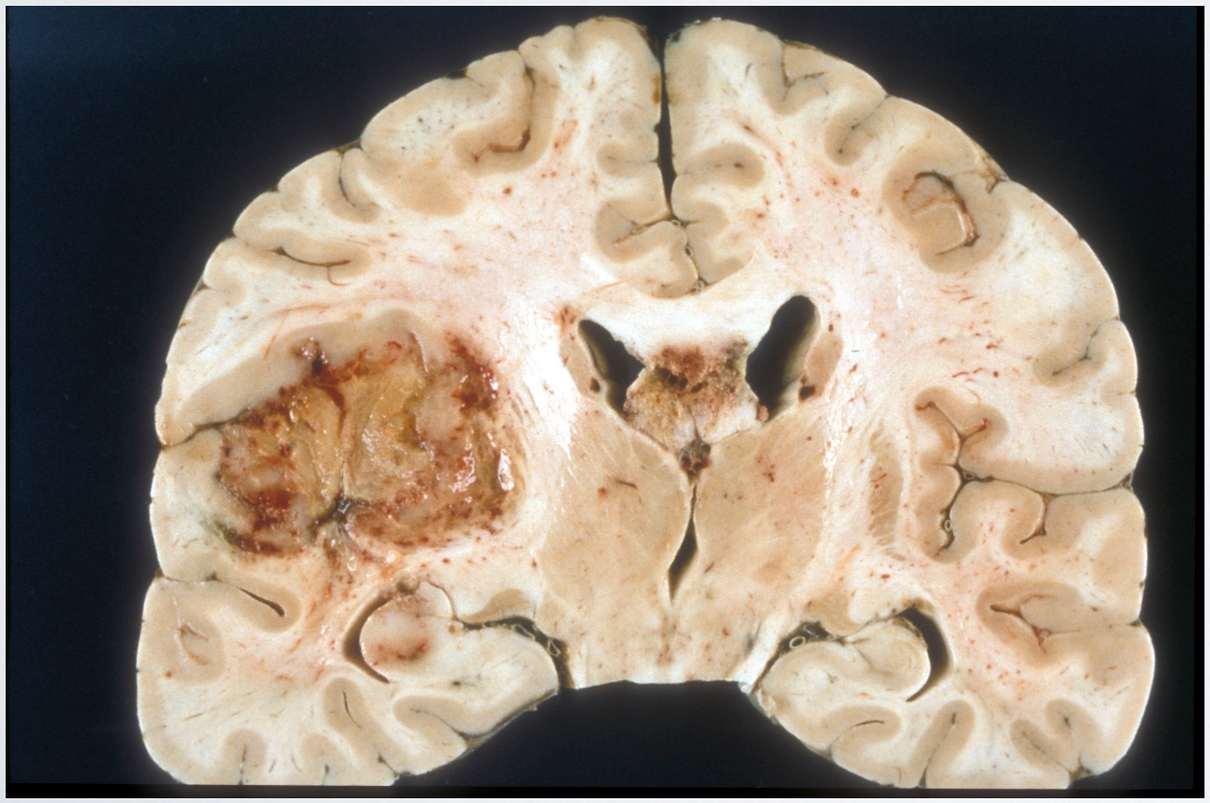


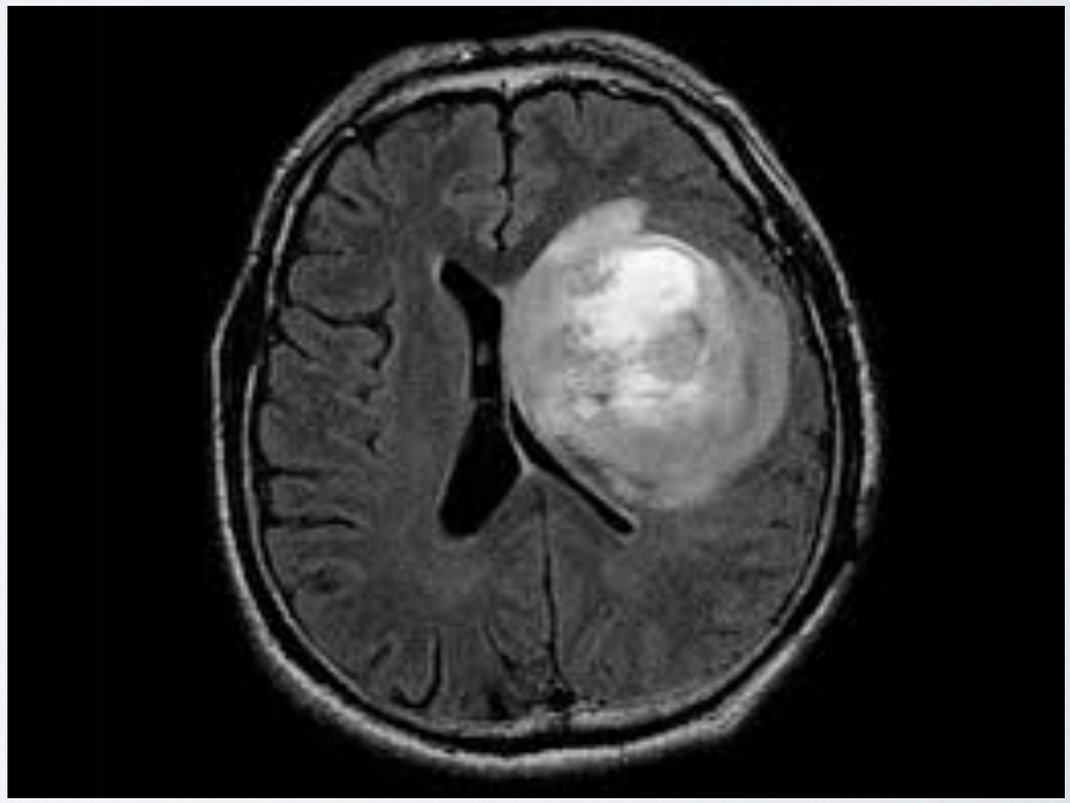
Example (2) of network multidimensional analysis - GBM

Glioblastoma multiforme (**GBM**), also known as **glioblastoma** and **grade IV astrocytoma**, is the most common and aggressive brain cancer

Worsening of symptoms is often rapid. The most common length of survival following diagnosis is 12 to 15 months with less than 3 to 5% of people surviving greater than five years. Without treatment survival is typically 3 months.

About 3 per 100,000 people develop the disease a year.





The experimental data:

Activated PTPRG triggers *in vitro apoptosis* in GBM cells

The questions:

Which is the mechanics of action of activated PTPRG on triggered apoptosis in **Glioblastoma multiforme**?

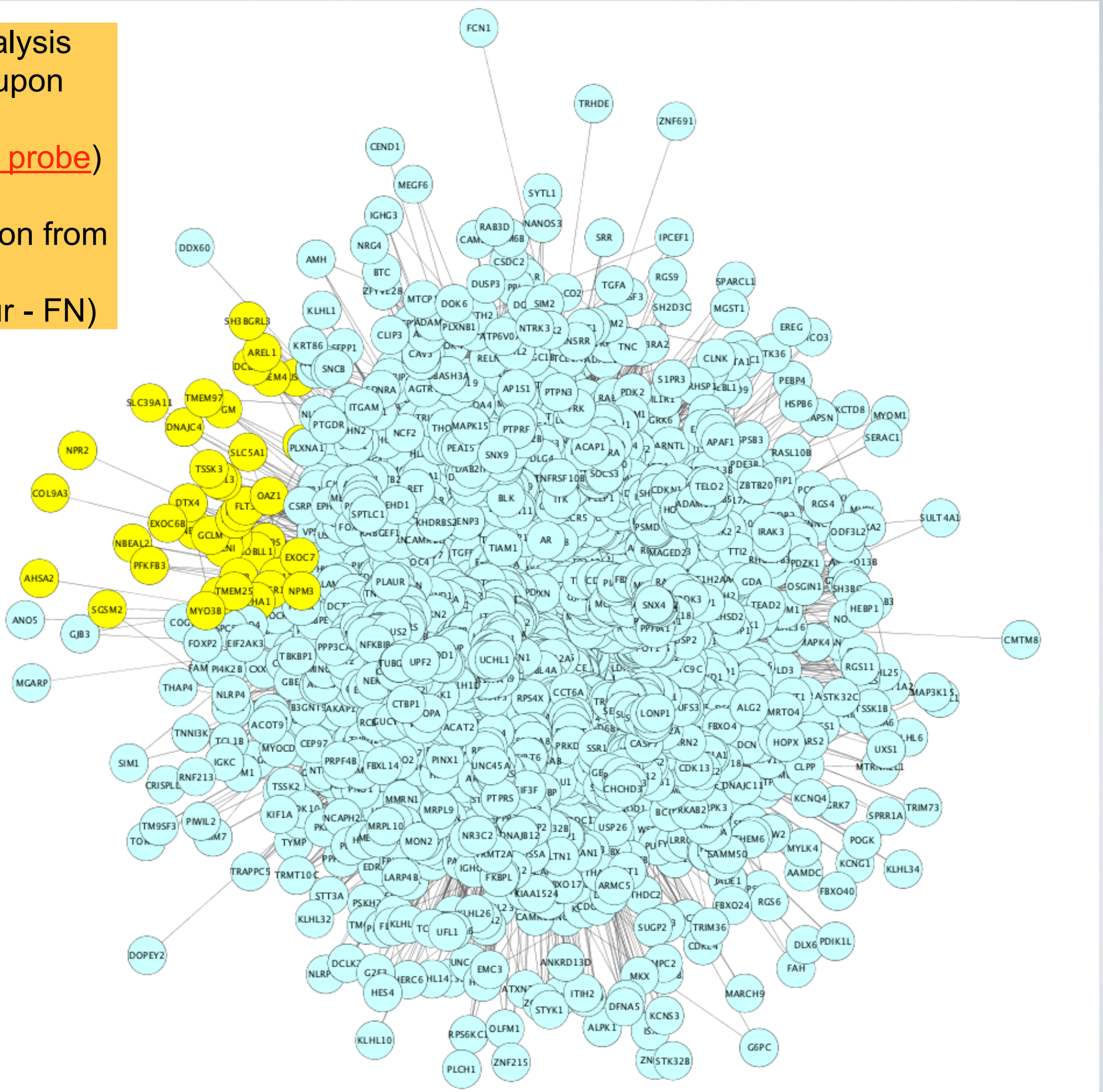
Is PTPRG a tumor-suppressor gene?

Analysis flow-chart

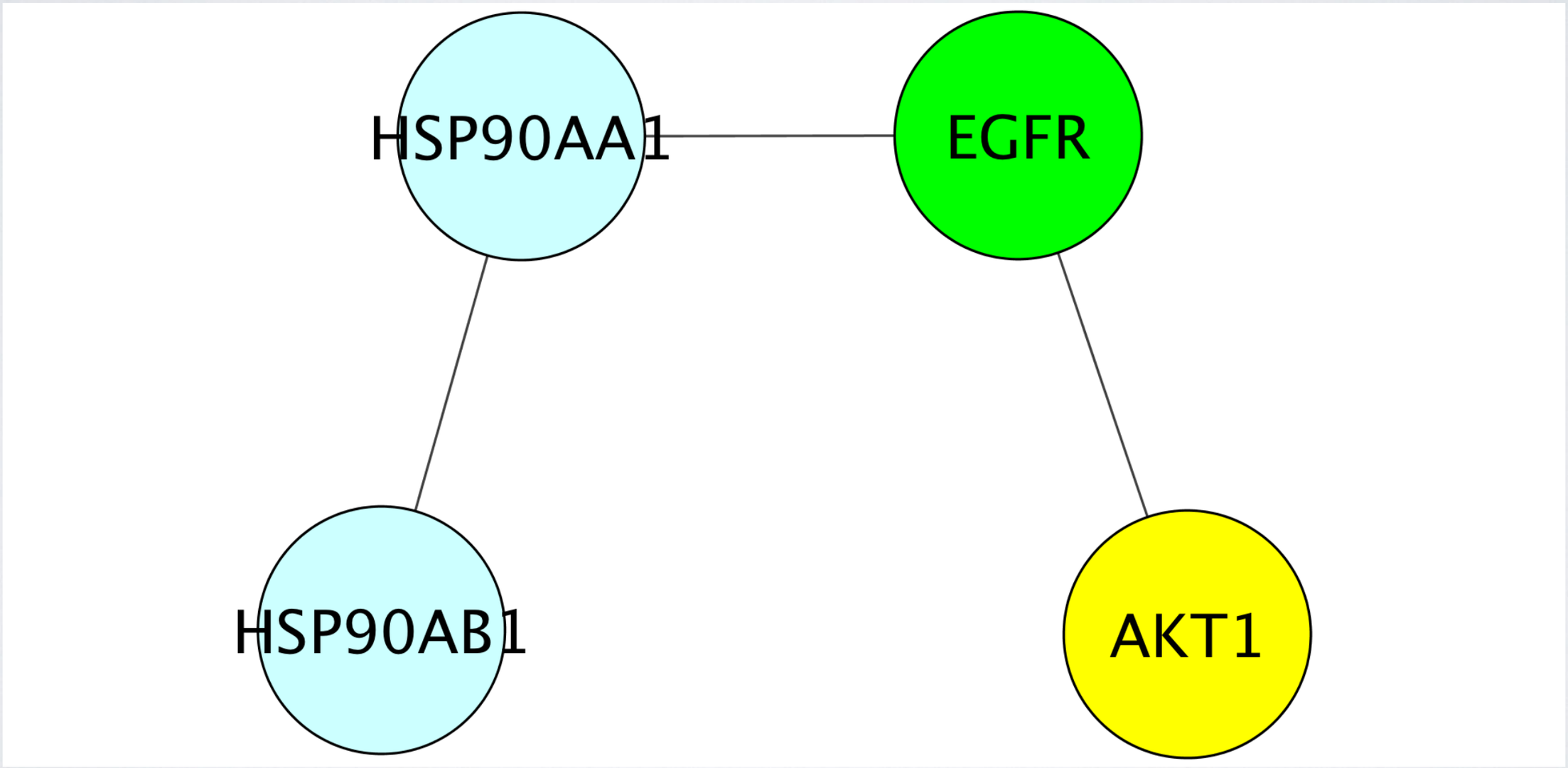
- ☑ Phosphoproteomics analysis (Kinexus) of GBM cell upon PTPRG activation - concept of bioinformatics probe (node dimensions - functional space)
- ☑ GBM network reconstruction from Kinexus data-set (calculating first neighbor - FN)
- ☑ Calculating centralities and identification of most relevant nodes (nodes with betweenness + centroid over network average)
- ☑ Shortest-path (SP) calculation from PTPRG to “extracted” nodes

Phosphoproteomics analysis (Kinexus) of GBM cell upon PTPRG activation (concept of bioinformatics probe)

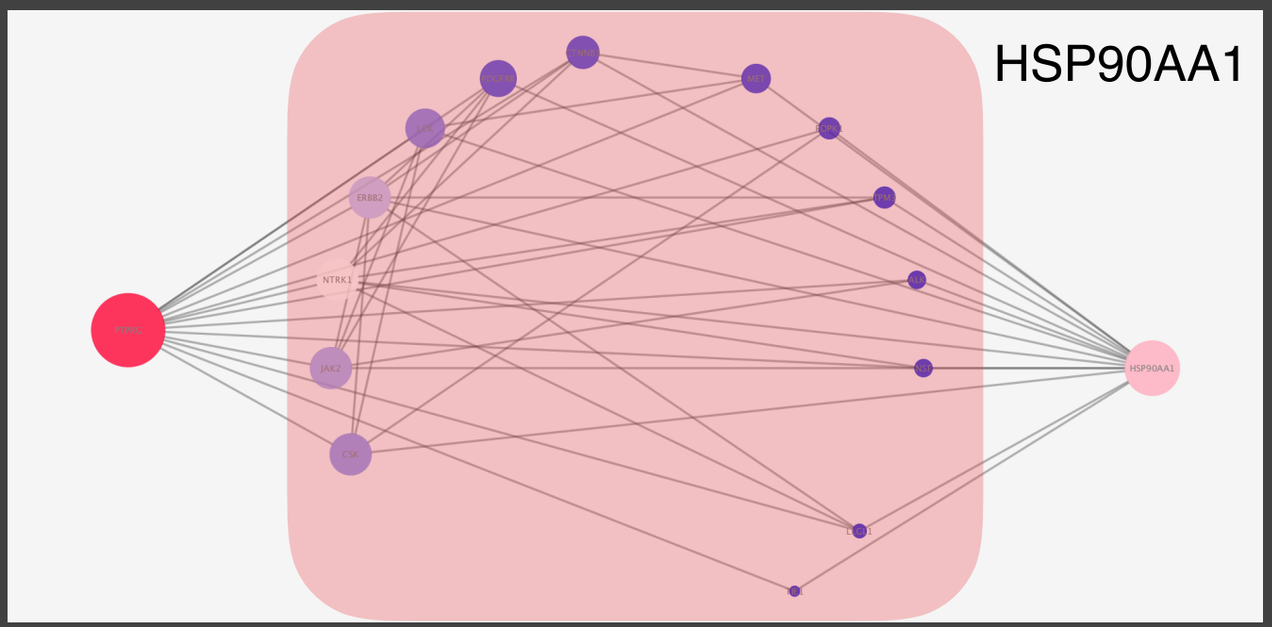
GBM network reconstruction from Kinexus data-set (calculating first neighbour - FN)

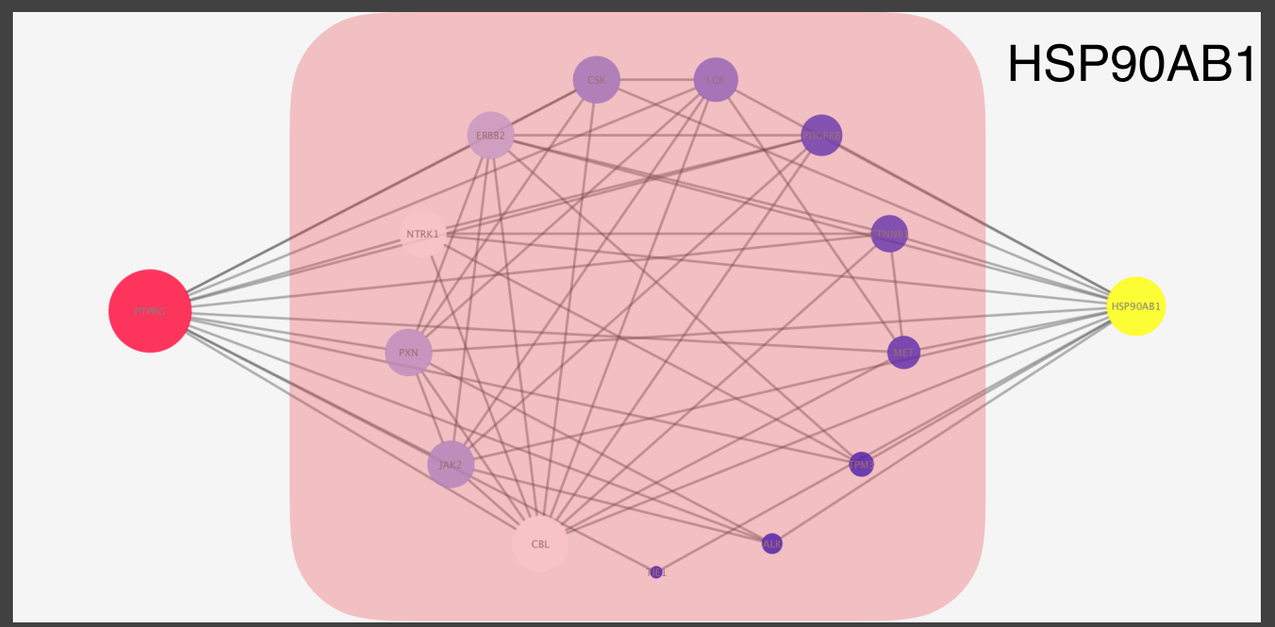


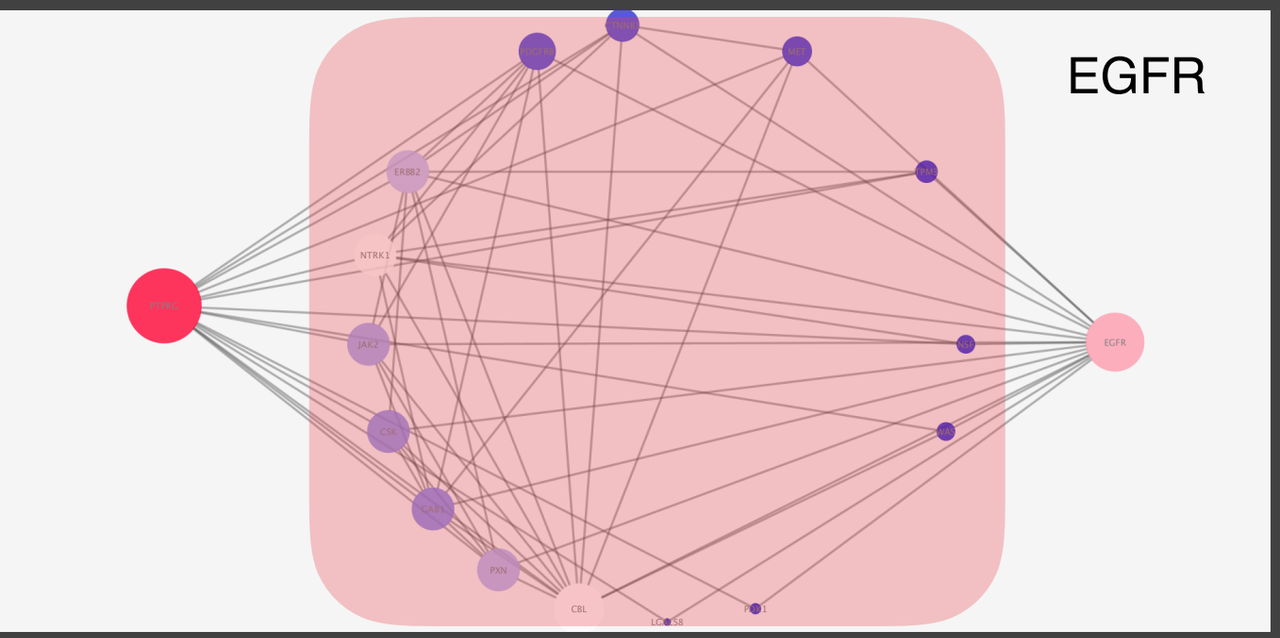
Calculating centralities and identification of most relevant nodes (nodes with betweenness + centroid over network average)

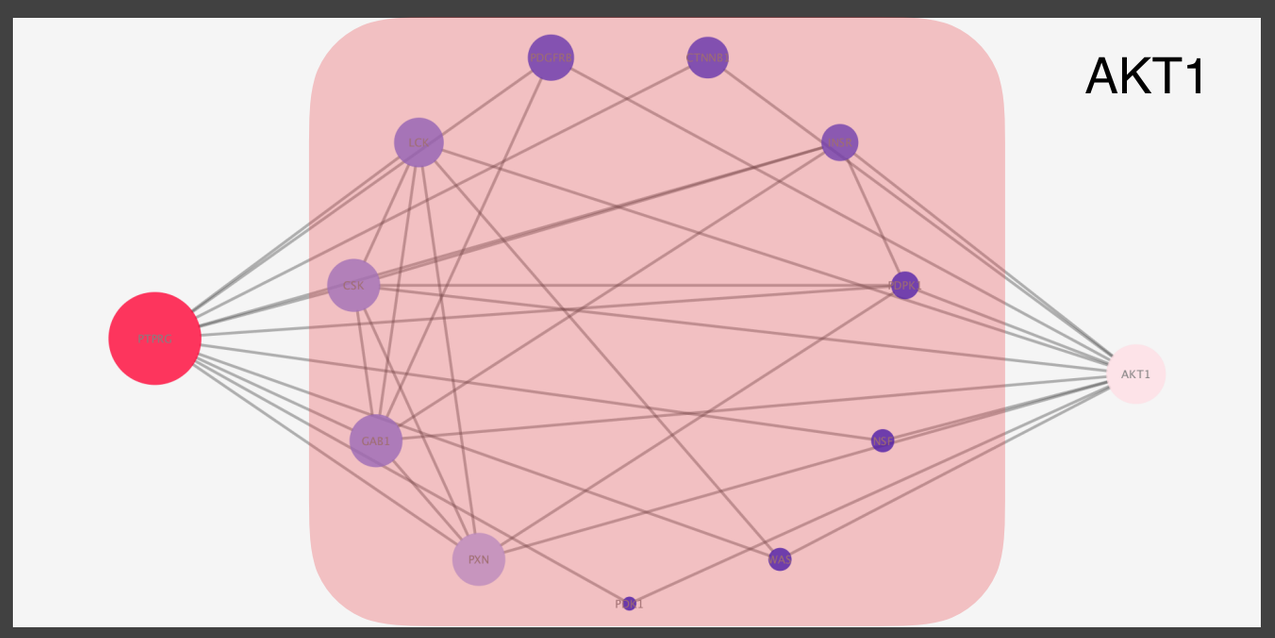


Shortest-path (SP) calculation from PTPRG to “extracted” most relevant nodes



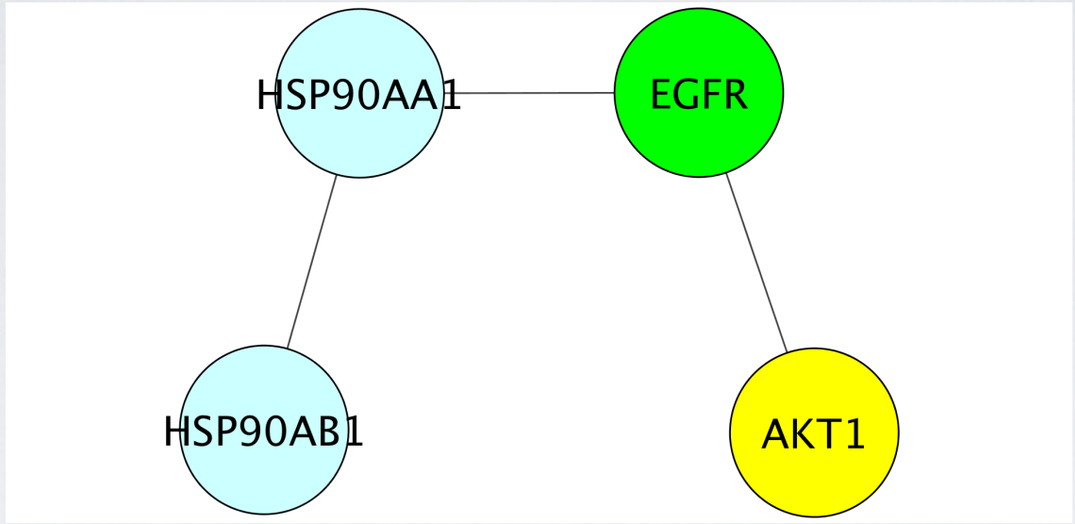




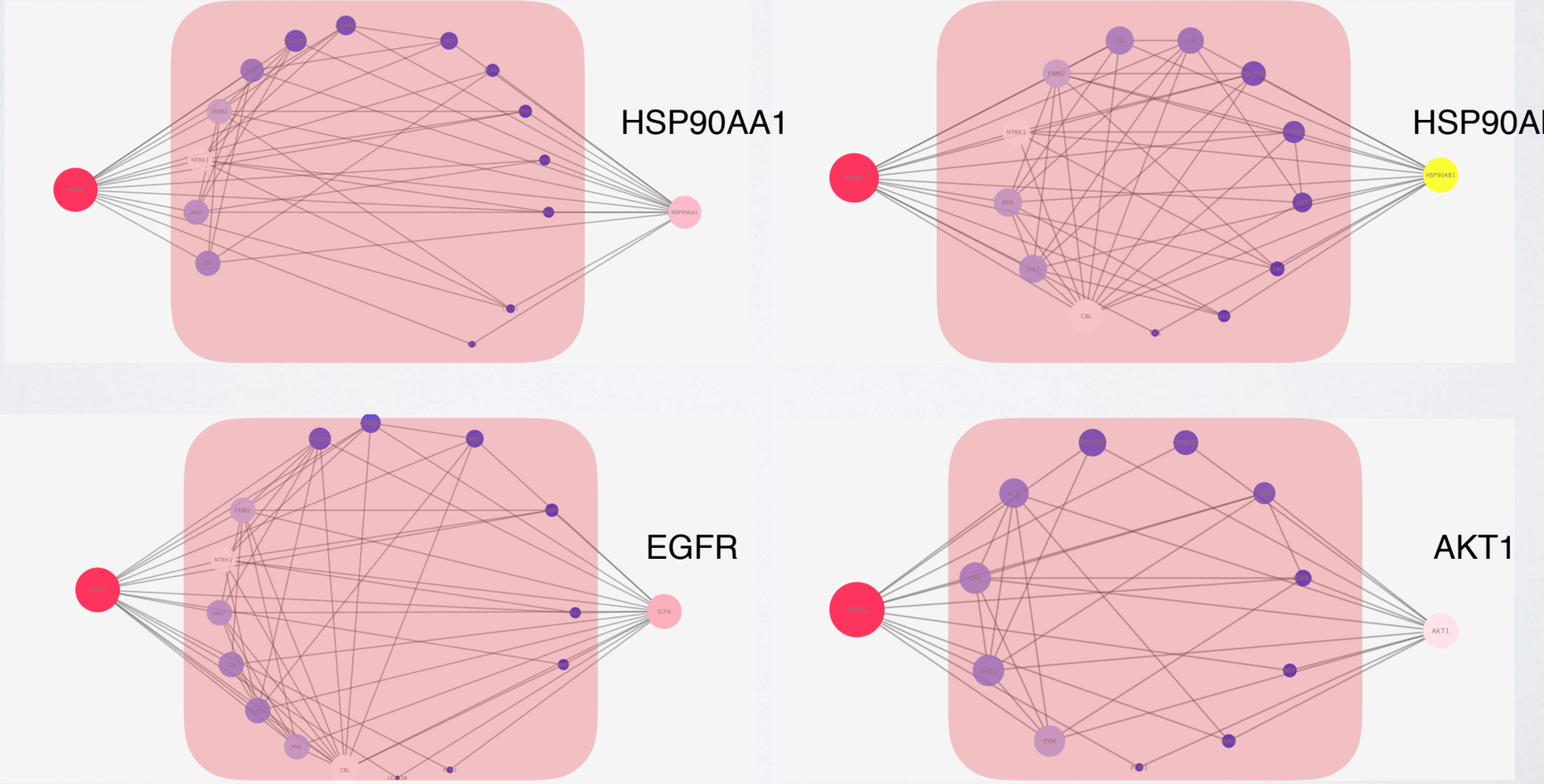


Conclusions

1) Most relevant signaling proteins possibly leading to PTPRG-triggered GBM apoptosis



2) Inferred signaling protein from SP analysis linking PTPRG to most relevant signaling proteins



(8)
Dynamics models: kinetics in biology

Graphs are “STATIC” ABSTRACTIONS of biological networks
BUT
Biology is mainly change!
(Mutation, Selection, Adaptation, Evolution)

To model the **change** in biology we need to take into account:

- **The system structure (topological architecture)**
- **The rules governing the change**
- **The quantitative data characterising the binary relationships**
- **The time factor**

The frequencies of calcium oscillations are optimized for efficient calcium-mediated activation of Ras and the ERK/MAPK cascade

Sabine Kupzig*, Simon A. Walker†, and Peter J. Cullen*‡

*Henry Wellcome Integrated Signalling Laboratories, Department of Biochemistry, School of Medical Sciences, University of Bristol, Bristol BS8 1TD, United Kingdom; and †Signalling Programme, The Babraham Institute, Babraham Research Campus, Cambridge CB2 4AT, United Kingdom

Edited by Solomon H. Snyder, The Johns Hopkins University School of Medicine, Baltimore, MD, and approved March 10, 2005 (received for review December 22, 2004)

Ras proteins are binary switches that, by cycling through inactive GDP- and active GTP-bound conformations, regulate multiple cellular signaling pathways, including those that control growth and differentiation. For some time, it has been known that receptor-mediated increases in the concentration of intracellular free calcium ($[Ca^{2+}]_i$) can modulate Ras activation. Increases in $[Ca^{2+}]_i$ often occur as repetitive Ca^{2+} spikes or oscillations. Induced by electrical or receptor stimuli, these repetitive Ca^{2+} oscillations increase in frequency with the amplitude of receptor stimuli, a phenomenon critical for the induction of selective cellular functions. Here, we show that Ca^{2+} oscillations are optimized for Ca^{2+}-mediated activation of Ras and signaling through the extracellular signal-regulated kinase (ERK)/mitogen-activated protein kinase (MAPK) cascade. We present additional evidence that Ca^{2+} oscillations reduce the effective Ca^{2+} threshold for the activation of Ras and that the oscillatory frequency is optimized for activation of Ras and the ERK/MAPK pathway. Our results describe a hitherto unrecognized link between complex Ca^{2+} signals and the modulation of the Ras/ERK/MAPK signaling cascade.

RASAL | CAPRI | GRF | GDP

For a wide variety of cell surface receptors, activation leads to an increase in the concentration of intracellular free calcium ($[Ca^{2+}]_i$) (1–3). Once induced, the elevation in $[Ca^{2+}]_i$ is responsible for controlling a diverse array of cellular processes, including secretion, contraction, learning, and proliferation (3). Understanding how receptor-mediated increases in $[Ca^{2+}]_i$ are capable of modulating so many physiological processes is one of the major challenges in the study of Ca^{2+} signaling. It appears that such control is achieved through a complex relationship between the amplitude and spatiotemporal patterning of the Ca^{2+} signal and its resultant ability to couple to an extensive molecular repertoire of Ca^{2+}-sensing proteins (3).

Receptor-mediated increases in $[Ca^{2+}]_i$ are often observed as repetitive Ca^{2+} spikes or oscillations that increase their frequency with the amplitude of the receptor stimuli (refs. 4 and 5; reviewed in ref. 3). These frequency-encoded signals appear to be critical for the induction of selective cellular functions (3). For example, the frequency of receptor-mediated Ca^{2+} oscillations determines the efficiency of gene expression driven by the transcription factors NF-AT, OAP, and NF-κB (6–8) and mitochondrial ATP production (9). To decode the information contained within Ca^{2+} oscillations, cells have evolved a number of frequency-modulated decoders. Such proteins include calmodulin (10), protein kinase C (11–15), calpain (16), calmodulin-dependent protein kinase II (17, 18), and the Ras GTPase-activating protein RASAL (19).

Ras proteins are binary molecular switches that regulate multiple signaling pathways, including those controlling growth and differentiation, through an ability to cycle between inactive GDP- and active GTP-bound conformations (20–23). The magnitude and duration of Ras signaling is controlled by two classes of proteins: Guanine nucleotide exchange factors modulate Ras activation by enhancing the exchange of GDP for GTP, and GTPase-activating proteins regulate inactivation by increasing the intrinsic Ras GTPase activity (20–23). Although it has been known for some time that increases in $[Ca^{2+}]_i$ can modulate Ras activation (for example, Ca^{2+} influx through voltage-operated ion channels or release from internal stores can activate Ras in neuronal cells) (24), only recently have molecular entities been described that allow for this coupling (reviewed in ref. 25).

Two families of Ras guanine nucleotide exchange factors (GEFs), RasGRFs (26–29) and RasGRPs (30–36), the latter also being known as CalDAG-GEFs, are modulated by increases in $[Ca^{2+}]_i$. For RasGRFs, this modulation occurs indirectly through association with Ca^{2+}/calmodulin, whereas for RasGRPs, a more direct control is achieved through association of Ca^{2+} with atypical EF hands (25). In addition to stimulating Ras activation, increases in $[Ca^{2+}]_i$ also mediate Ras inactivation through the Ca^{2+}-triggered RasGTPase-activating proteins (RasGAPs) RASAL and CAPRI (19, 37). These proteins are cytosolic, inactive RasGAPs that, upon a receptor-mediated elevation in $[Ca^{2+}]_i$, undergo a rapid, C_2 domain-dependent association with the plasma membrane, an association that leads to an increase in their RasGAP activity (19, 37). Unlike CAPRI, which undergoes a transient association with the plasma membrane and does not sense receptor-mediated Ca^{2+} oscillations, the plasma membrane association of RASAL occurs in an oscillatory manner (19). This oscillatory association occurs in synchrony with underlying receptor-mediated Ca^{2+} oscillations and is frequency-modulated such that, upon increasing the amplitude of receptor stimuli, the frequency of RASAL membrane association is enhanced (19). CAPRI and RASAL therefore constitute molecular entities that can sense the amplitude and frequency, respectively, of complex Ca^{2+} signals, decoding these distinct temporal signals through a modulation of plasma-membrane-associated Ras.

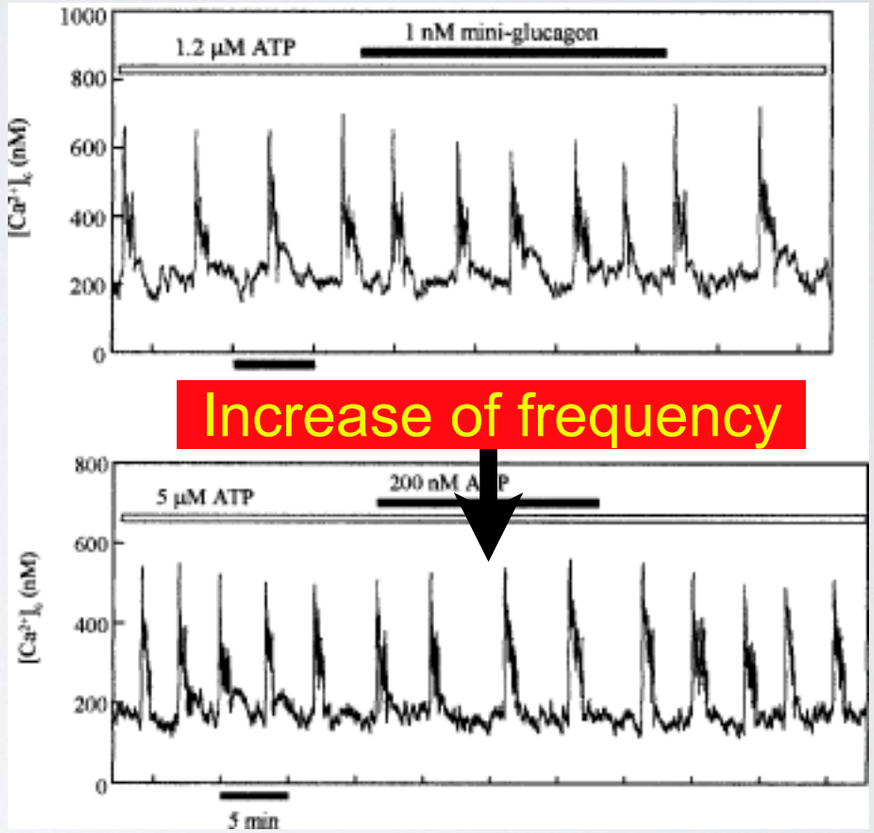
The characterization of such distinct Ca^{2+} sensors, tuned to detect different temporal Ca^{2+} signals, has raised the issue of whether the temporal dynamics of receptor-mediated Ca^{2+} oscillations are optimized for efficient Ca^{2+}-mediated activation of Ras and downstream Ras-dependent signaling (25, 38). Here, we have addressed this issue, presenting data showing that the temporal dynamics of Ca^{2+} signals are indeed optimized for activation of Ras and the downstream extracellular signal-

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviations: $[Ca^{2+}]_i$, concentration of intracellular free calcium; EGF, epithelial growth factor; ERK, extracellular signal-regulated kinase; MAPK, mitogen-activated protein kinase.

‡To whom correspondence should be addressed. E-mail: pete.cullen@bris.ac.uk.



HRas-MAPK pathway

The H-Ras-MAPK pathway is not regulated by the amount (concentration) of absolute Ca_i released (amplitude)

BUT

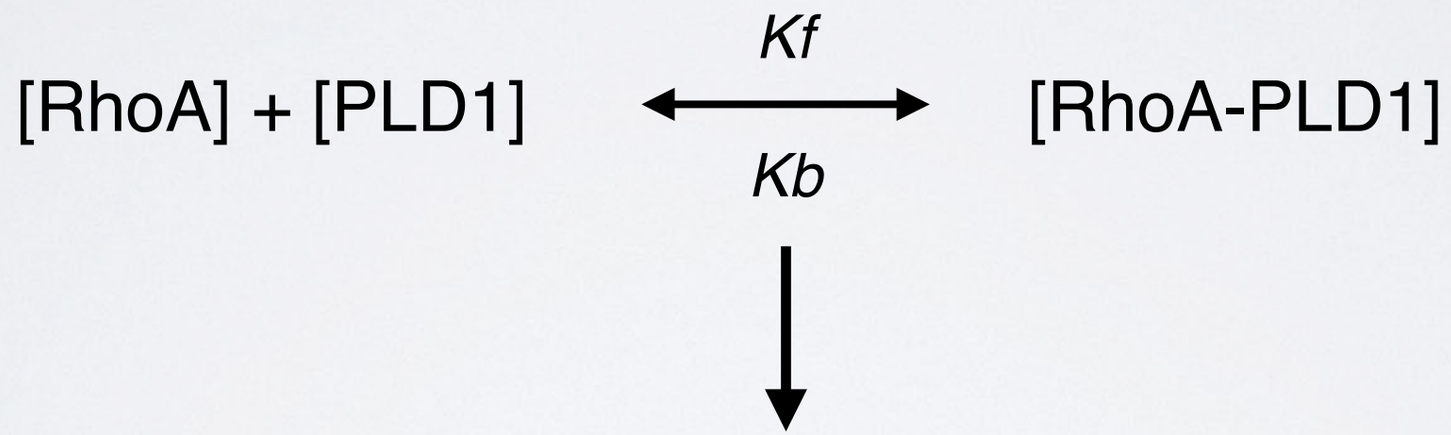
by the <u>frequency of the oscillations</u> (time flow)

An emergent property that directly regulates a biological function

It is not the function of a single molecule that regulates.

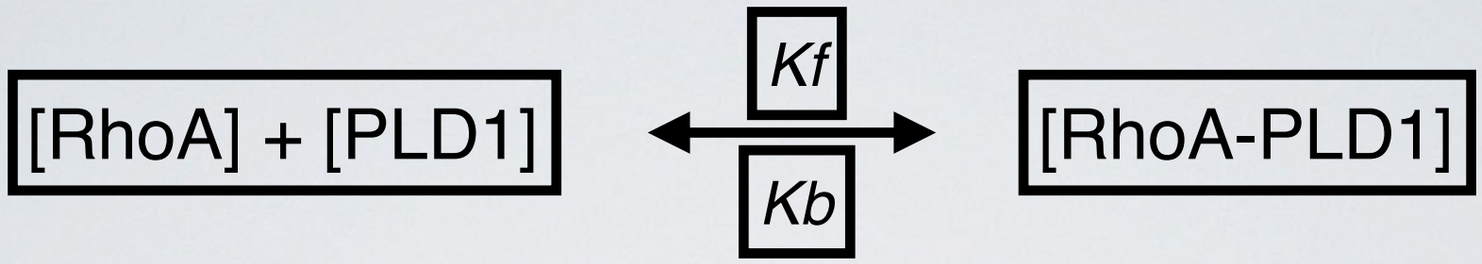
Instead, the ensemble of interacting molecules, appropriately combined, generates circuits generating <u>emergent functions</u>.

Dynamic analysis of molecular networks



Information about:

1. binary physical interaction
2. formation of sufficiently stable complex allowing PLD1 activation
3. reversibility and shut-off of PLD1 activity
4. kinetic constants governing the reaction equilibrium



Set of differential equations, etc.

↓

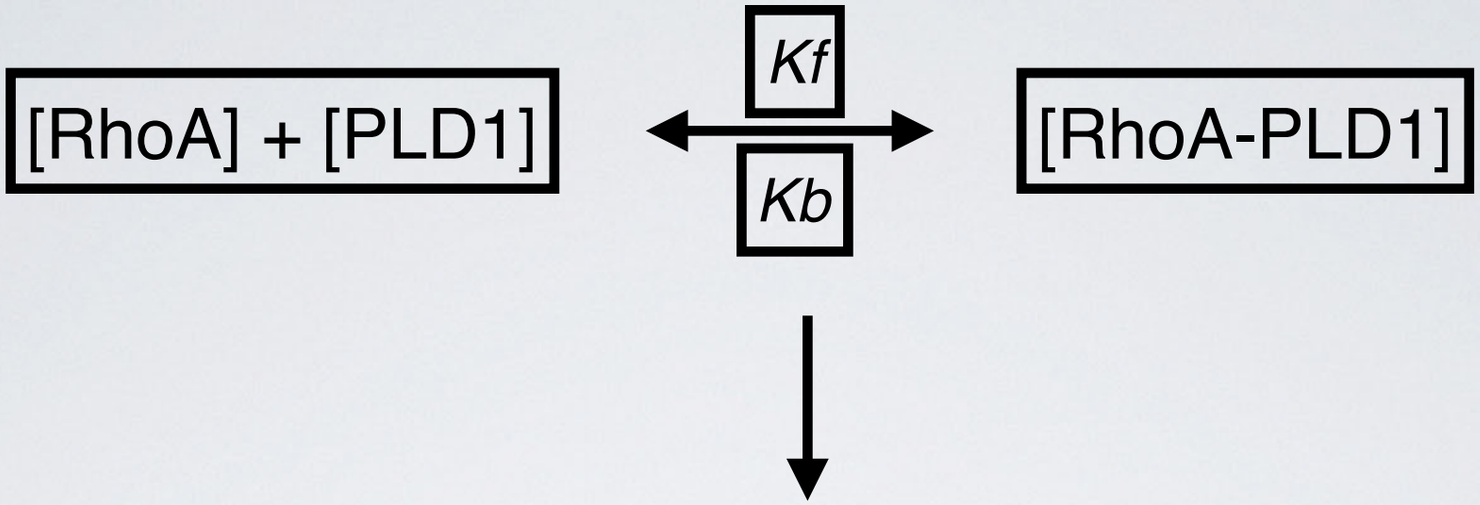
Software implementation of equation sets

↓

Computer simulation of cell molecular dynamics

↓

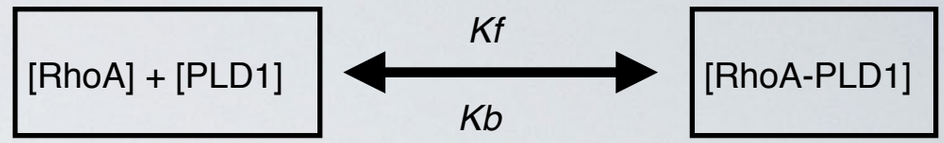
Identification and analysis of “emergent properties”: oscillators, by-stability, feedbacks, etc.



Michaelis-Menten Equation

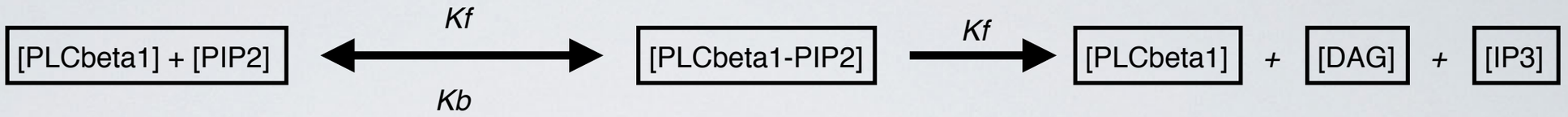
$$V_0 = \frac{V_{max}\,[S]}{K_M + [S]}$$

3 special cases



$$A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} AB \qquad (1)$$

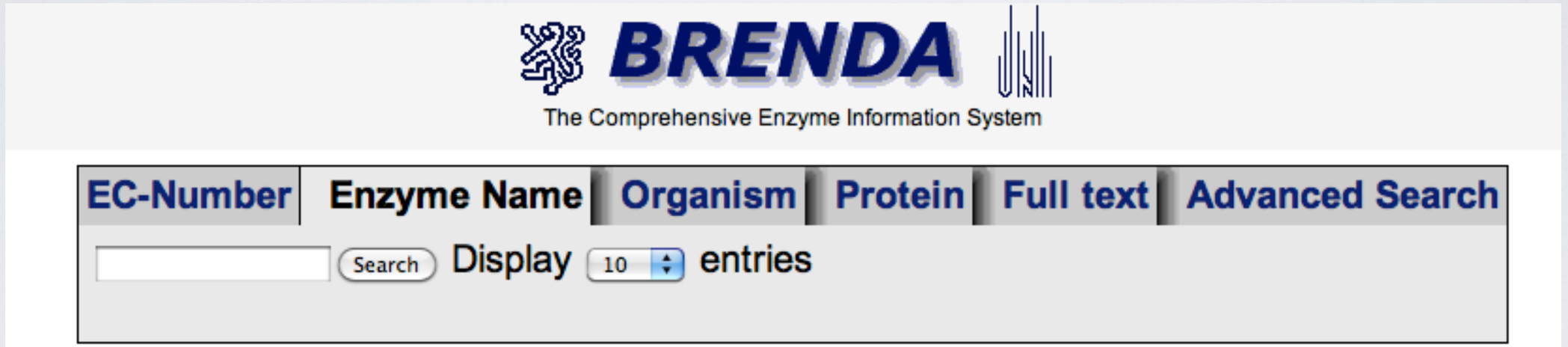
$$A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C + D \qquad (2)$$



$$E + S \underset{k_2}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_3} E + P \qquad (3)$$

To dynamically model molecular networks we need to know the quantitative parameters governing biochemical reactions under physiological conditions!

BRENDA: a data-base of enzymatic kinetics data



http://www.brenda-enzymes.info/index.php4

In chemistry, and more in general in systems theory, **the steady state** of a system or a process (such as a biochemical reaction) occurs when the variables (for instance an enzyme K*m)* defining the process kinetics do not change over time.

In a context of continuous or discrete time (continuous or discrete kinetics) this implies that *variables are always identical* (constant)

with respect to time 0

The problem of the “starting conditions”:

which 'the starting point? The zero state?

It is generally assumed that the <u>system is modelled at steady state</u> (kinetics equilibrium)

CAUTION:
The cell is not alive when I say so (at the time of observation)!
It is just an “obligated” assumption!

Notably: metabolism normally works at steady state; signal transduction not necessarily!

those of the analogous A_2/G reaction, as expected for the higher driving force reaction.

34. J. W. Evenson and M. Karplus, *Science* **262**, 1247 (1993); S. Priyadarshy, S. M. Risser, D. N. Beratan, *J. Bioinorg. Chem.* **3**, 196 (1998); J. J. Regan *et al.*, *Chem. Biol.* **2**, 489 (1995).
35. W. B. Davis, W. A. Svec, M. A. Ratner, M. R. Wasielewski, *Nature* **396**, 60 (1998).
36. Alternatively, the nitrogen-to-carbon substitution in Z compared with the natural base may decrease the strength of stacking interactions for this modified base. However, the use of deazaadenine (which has an oxidation potential quite similar to that of G) as an electron donor in analogous experiments produced an interstrand distance dependence almost identical to that observed with G (S. O. Kelly and J. K. Barton, data not shown). Therefore, it appears that the increased distance dependence for the A_2-Z reaction results from energetic effects.
37. S. Georghiou, T. D Bradrick, A. Philippetis, J. M. Beechem, *Biophys. J.* **70**, 1909 (1996).
38. J. Jortner, M. Bixon, T. Langenbacher, M. E. Michel-Beyerle, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12759 (1998); A. K. Felts, W. T. Pollard, R. A. Freisner, *J. Phys. Chem.* **99**, 2929 (1995); A. Okada, V. Chernyak, S. Mukamel, *idib*. **102**, 1241 (1998).
39. A_ε, A_2, I, and Z phosphoramidites were all obtained from Glen Research. Oligonucleotides were prepared by standard automated techniques (with the exceptions described below) on a 394 ABI synthesizer and purified by reverse phase high-performance liquid chromatography. For Z, the oxidation step was carried out with 10-camphorsulfonyl oxaziridine as previously described (*5*). Base deprotection of ethenoadenine-containing oligonucleotides was carried out for 24 hours at room temperature. Oligonucleotides containing modified bases were characterized by electrospray mass spectroscopy. Samples were prepared as follows: on the basis of the calculated extinction coefficients for DNA sequences [ε_{260} (M^{-1} cm^{-1}): dC = 7.4×10^3; dG = 12.3×10^3; dT = 6.7×10^3; dA = 15.0×10^3; dZ-G = 10.5×10^3; dI = 11.0×10^3 ; d(A_ε) = 4.5×10^3; d(A_2) = 2.5×10^3], appropriate amounts of complementary materials were combined at 1:1 stoichiometry and dissolved in 100 mM sodium phosphate (pH 7) to give a final duplex concentration of 100 μM. The resulting solutions were heated to 90°C and slowly cooled to ambient temperature over 2 to 3 hours to anneal the duplex. The ultraviolet-visible spectra of the duplex samples were carefully measured to ensure that the absorbance at the excitation wavelength was identical for every sample. Thermal denaturation experiments were performed on a HP8452A diode array spectrophotometer with samples at a duplex concentration of 25 μM in 100 mM phosphate (pH 7). Absorbance was monitored every 2°C with 3-min equilibration times. All duplexes used in these experiments exhibited cooperative thermal denaturation profiles with melting temperatures >25°C with 25 μM duplex and therefore were fully hybridized under the conditions of all fluorescence experiments (100 μM, 20°C).
40. Measurements were performed on a TCSPC apparatus previously described (*4*). Excitation was performed at 325 nm, and emission was monitored at 350 nm for A_2 and 400 nm for A_ε. Two data sets were obtained for each sample, one containing >10,000 counts for the determination of decay lifetimes, and another taken over a 120-s time interval to quantitate static quenching. Experiments were otherwise performed under the same conditions as steady-state experiments (100 μM duplex, 100 mM sodium phosphate, pH 7).
41. C. Santhosh and P. C. Mishra, *Spectrochim. Acta* **47A**, 1685 (1991).
42. We thank E. Stemp for assistance with transient absorption experiments, R. Villahermosa for assistance with single-photon counting measurements, and T. Fiebig for discussions and assistance with molecular modeling. In addition, we acknowledge the NIH (grant GM49216 to J.K.B., predoctoral traineeship to S.O.K.) for financial support.

27 August 1998; accepted 7 December 1998

Emergent Properties of Networks of Biological Signaling Pathways

Upinder S. Bhalla and Ravi Iyengar*

Many distinct signaling pathways allow the cell to receive, process, and respond to information. Often, components of different pathways interact, resulting in signaling networks. Biochemical signaling networks were constructed with experimentally obtained constants and analyzed by computational methods to understand their role in complex biological processes. These networks exhibit emergent properties such as integration of signals across multiple time scales, generation of distinct outputs depending on input strength and duration, and self-sustaining feedback loops. Feedback can result in bistable behavior with discrete steady-state activities, well-defined input thresholds for transition between states and prolonged signal output, and signal modulation in response to transient stimuli. These properties of signaling networks raise the possibility that information for "learned behavior" of biological systems may be stored within intracellular biochemical reactions that comprise signaling pathways.

Studies on the cyclic adenosine monophosphate (cAMP) signaling pathway led to the identification of several general mechanisms of signal transfer, such as regulation by protein-protein interactions, protein phosphorylation, regulation of enzymatic activity, production of second messengers, and cell surface signal transduction systems (*1*). These mechanisms of signal transfer have subsequently been shown to occur in many pathways, including Ca^{2+} signaling pathways (*2*), tyrosine kinase pathways (*3*), and other protein kinase cascades, and recently in the intracellular protease cascades in apoptosis (*4*). Initially, signaling pathways were studied in a linear fashion, and it was shown that many important biological effects are obtained through linear information transfer. However, it has become increasingly clear that signaling pathways interact with one another and the final biological response is shaped by interaction between pathways. These interactions result in networks that are quite complex and may have properties that are nonintuitive. A systematic analysis of interactions between signaling pathways could be useful in understanding the properties of these networks. We developed models for simple networks consisting of up to four signaling pathways to determine if the network has properties that the individual pathways do not and if networking results in persistent activation of protein kinases after transient stimulus. Persistent activation of protein kinases is a general mechanism for eliciting biological effects. Cholera toxin continuously elevates cAMP, resulting in persistent activation of protein kinase A (PKA), inhibition of intestinal water reabsorption, and diarrhea, key pathological manifestations of cholera (*5*). Since this original demonstration, persistent activation of protein kinases has been implicated in diverse processes such as neoplastic transformation (*6*) and learning and memory (*7*). Although mutations or altered gene expression can result in persistent activation of protein kinases, we wished to ask the following question: Do connections between preexisting signaling pathways result in persistently activated protein kinases capable of eliciting end-point biological effects?

To develop models of signaling pathways, it is necessary to consider the mechanisms by which signal transfer occurs. In biological systems, signal transmission occurs mostly through two mechanisms: (i) protein-protein interactions and enzymatic reactions such as protein phosphorylation and dephosphorylation (ii) or protein degradation or production of intracellular messengers. In an approach that would include all of these reactions, we used the basic chemical reaction schemes of

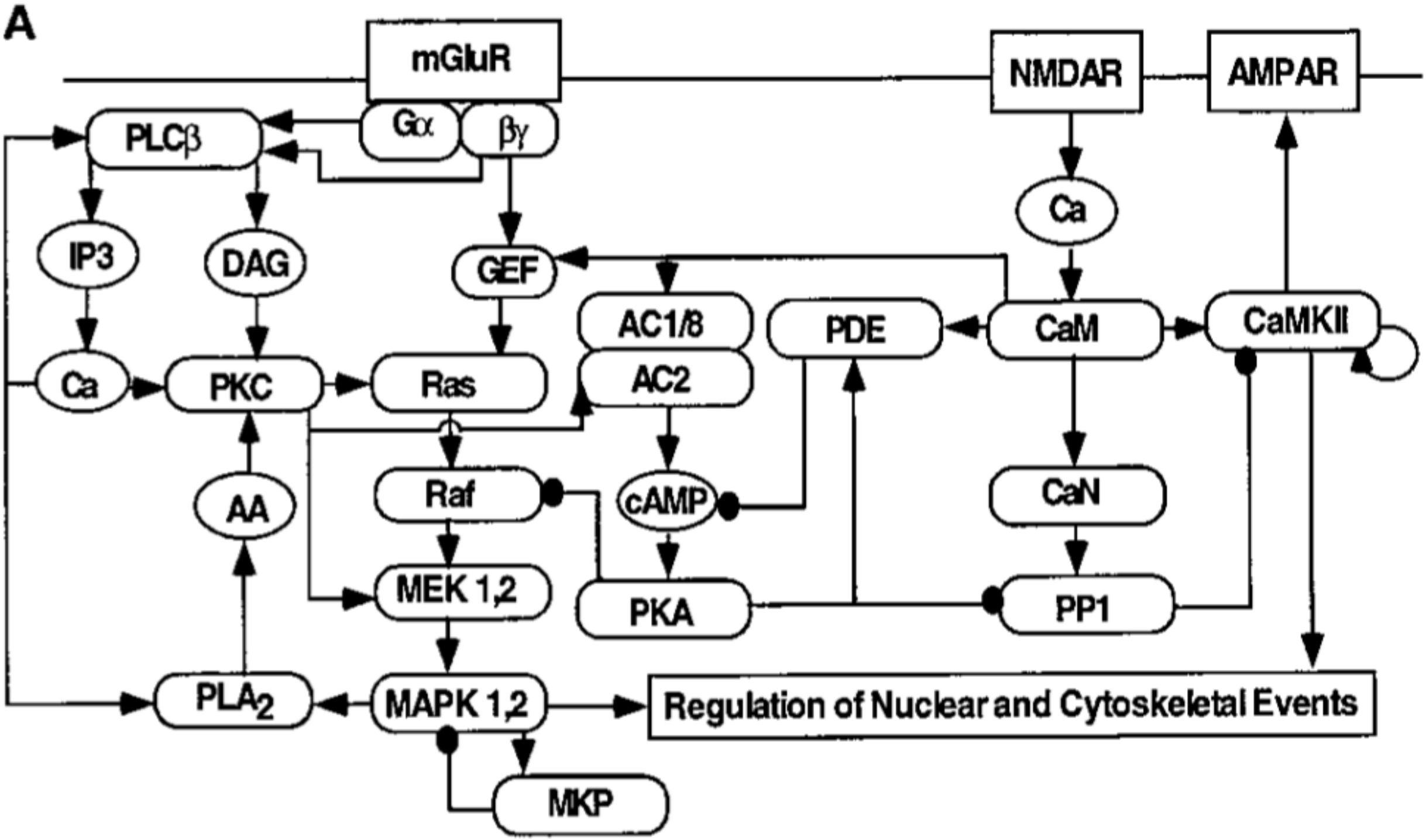
$$A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} AB \qquad (1)$$

$$A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C + D \qquad (2)$$

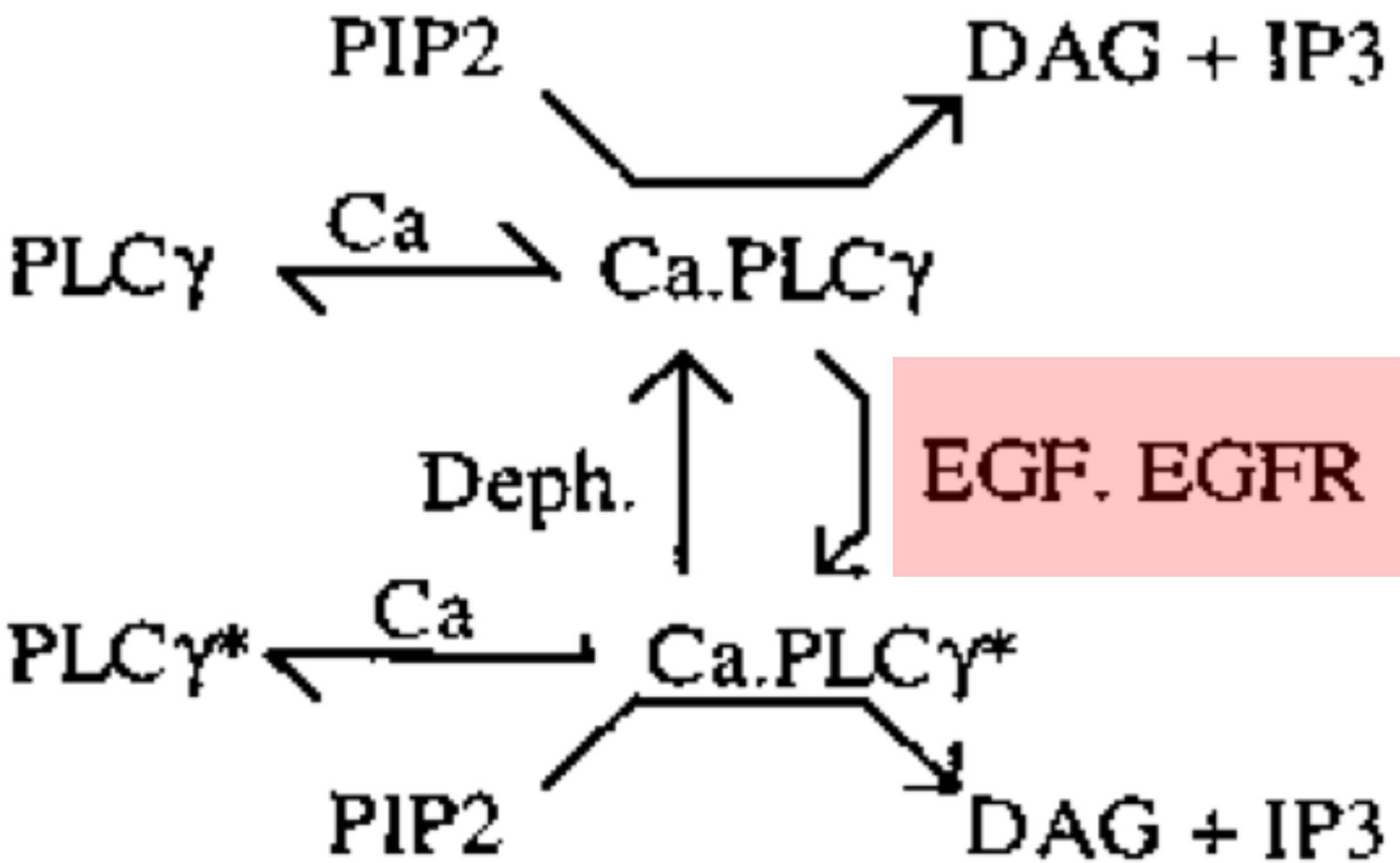
U. S. Bhalla, Department of Pharmacology, Mount Sinai School of Medicine, New York, NY 10029, USA, and National Centre for Biological Sciences, Post Office Box 1234, Bangalore 560012, India. R. Iyengar, Department of Pharmacology, Mount Sinai School of Medicine, New York, NY 10029, USA.

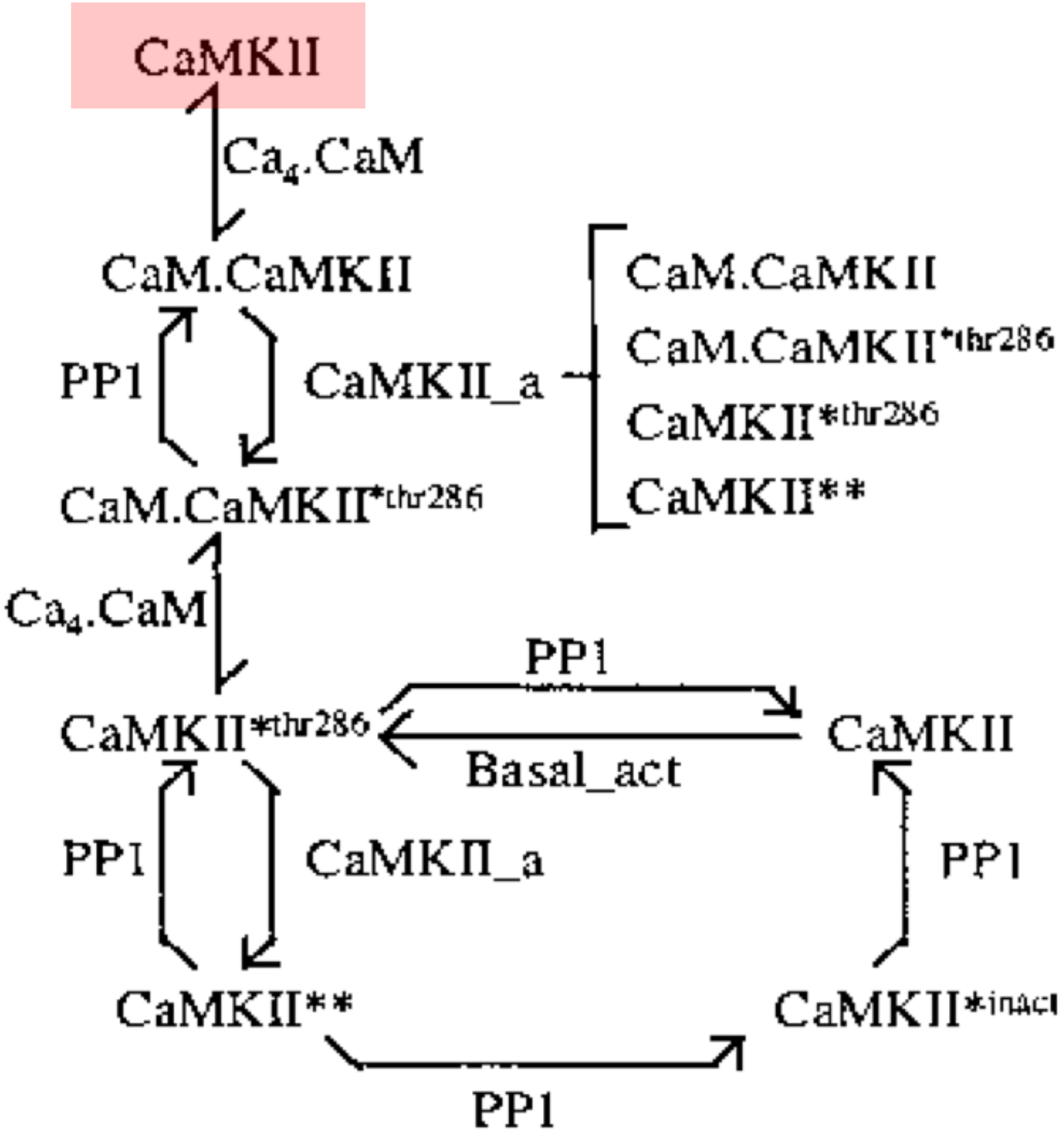
*To whom correspondence should be addressed.

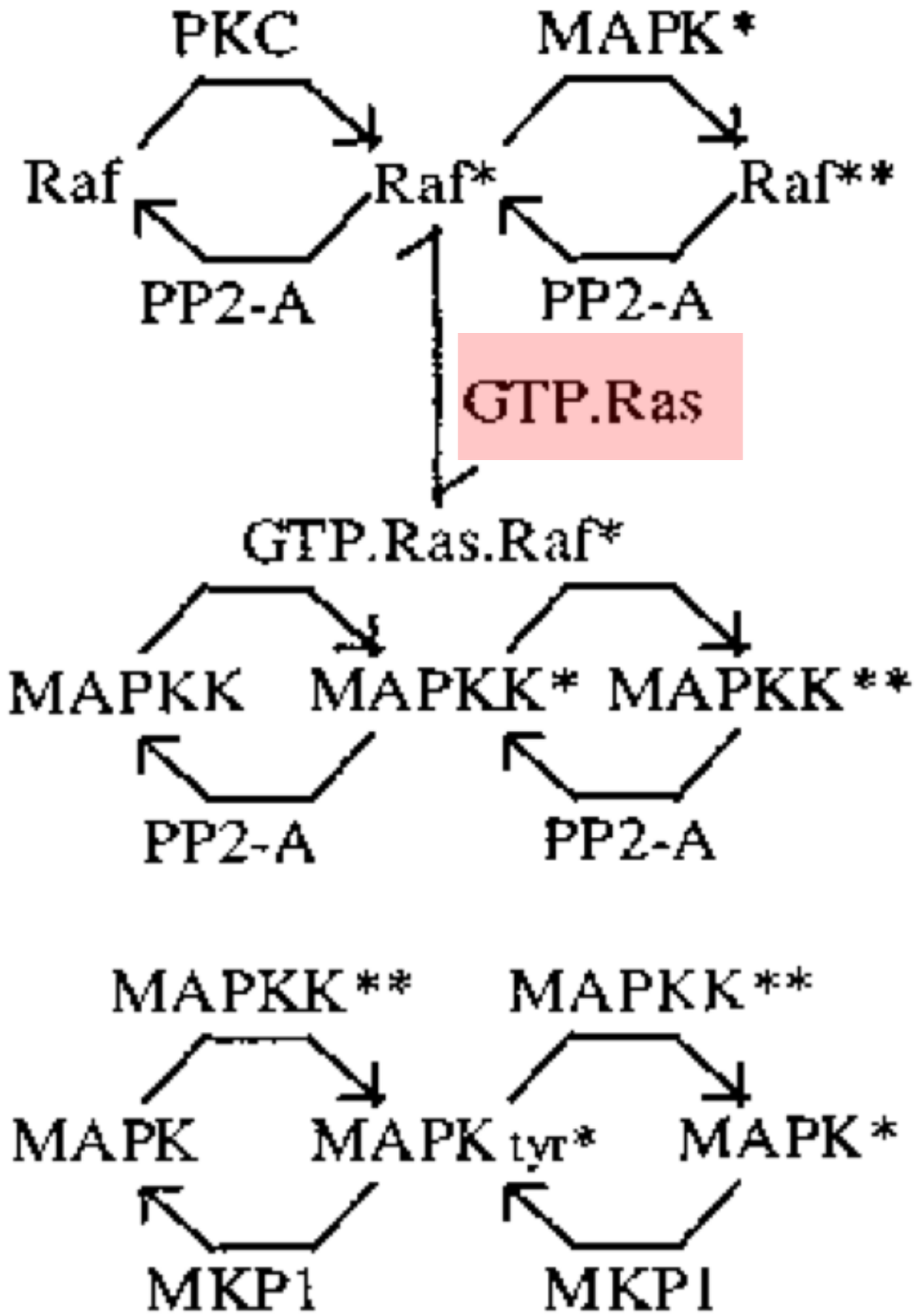
Signal transduction networks



The issue: kinetic modelling at balancing

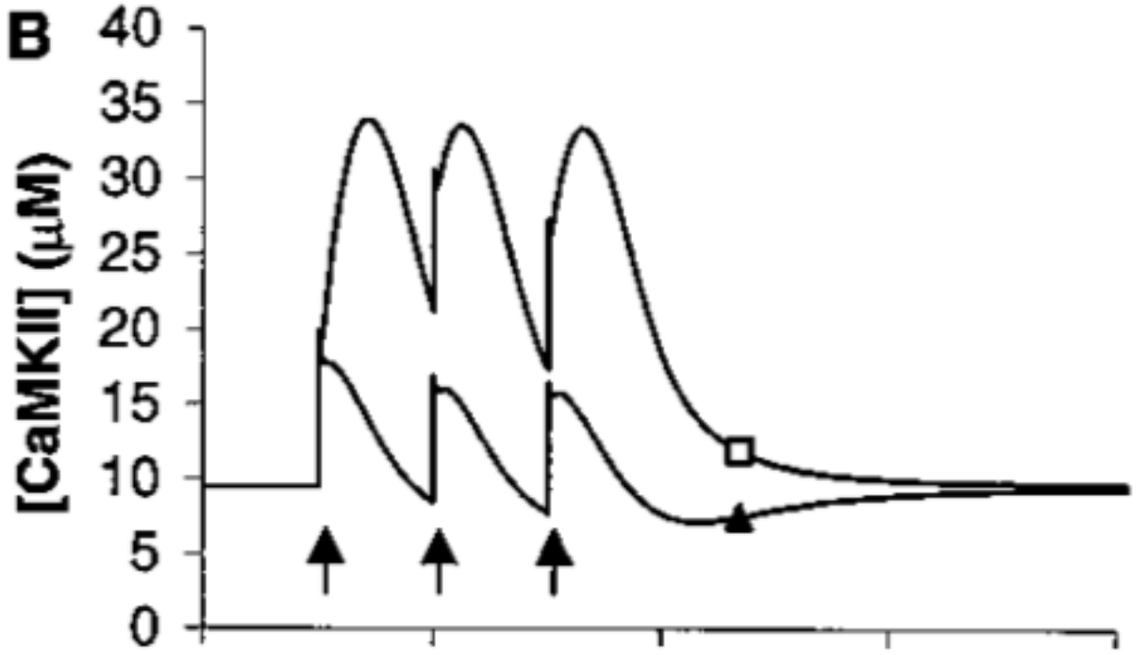

PIP2
DAG + IP3
Ca
PLCγ
Ca.PLCγ
Deph.
EGF. EGFR
Ca
PLCγ*
Ca.PLCγ*
PIP2
DAG + IP3

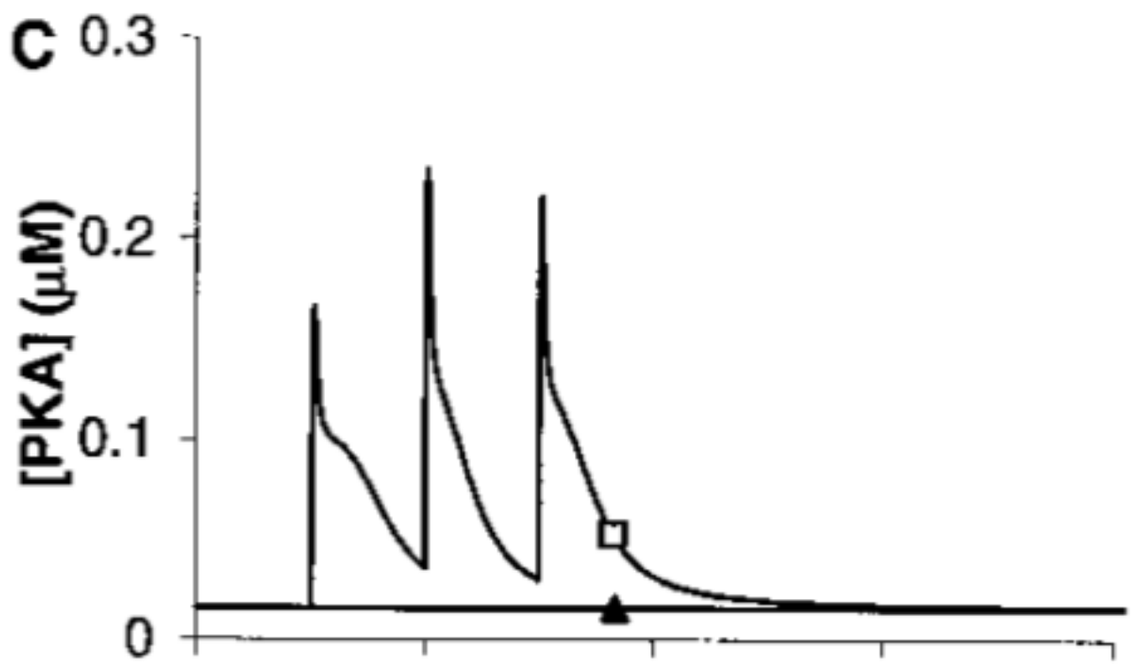

CaMKII
Ca4.CaM
CaM.CaMKII
PP1
CaMKII_a
CaM.CaMKII
CaM.CaMKII*thr286
CaMKII*thr286
CaMKII**
CaM.CaMKII*thr286
Ca4.CaM
PP1
CaMKII*thr286
CaMKII
Basal_act
PP1
CaMKII_a
PP1
CaMKII**
CaMKII*inact
PP1

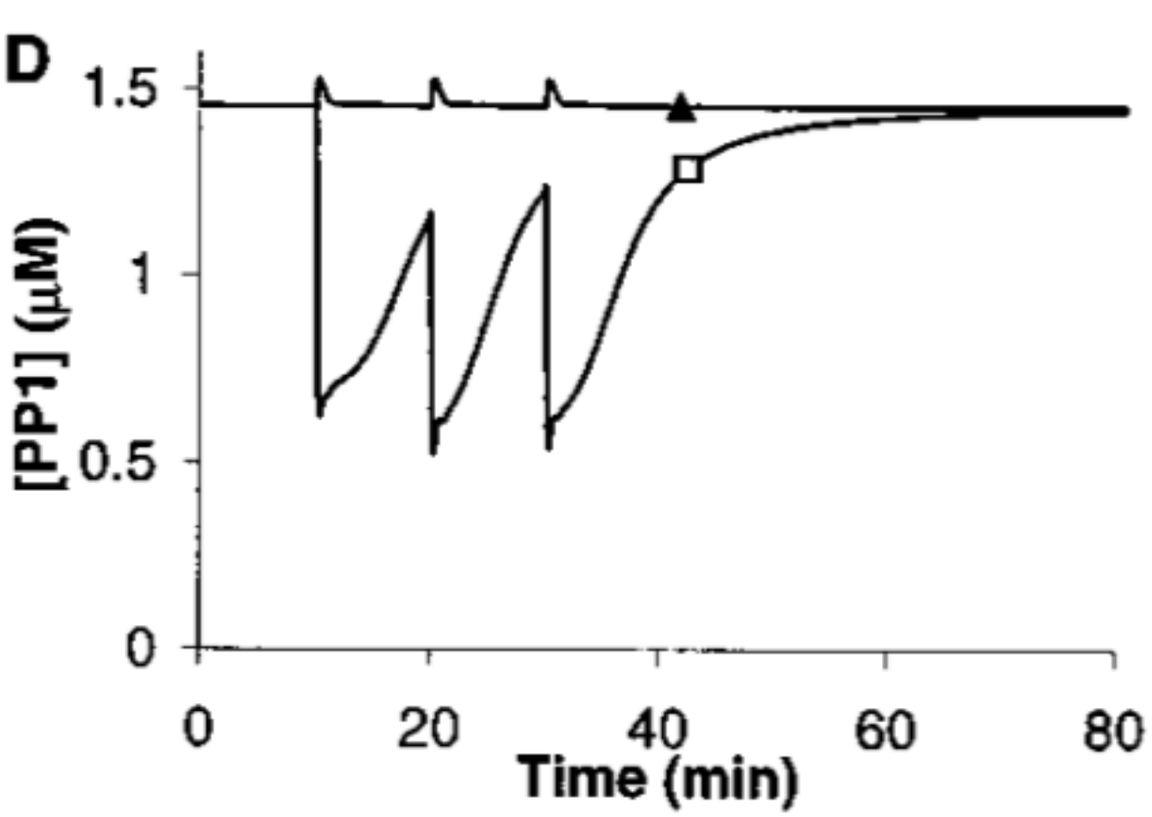

PKC
MAPK*
Raf
Raf*
Raf**
PP2-A
PP2-A
GTP.Ras
GTP.Ras.Raf*
MAPKK
MAPKK*
MAPKK**
PP2-A
PP2-A
MAPKK**
MAPKK**
MAPK
MAPK tyr*
MAPK*
MKP1
MKP1

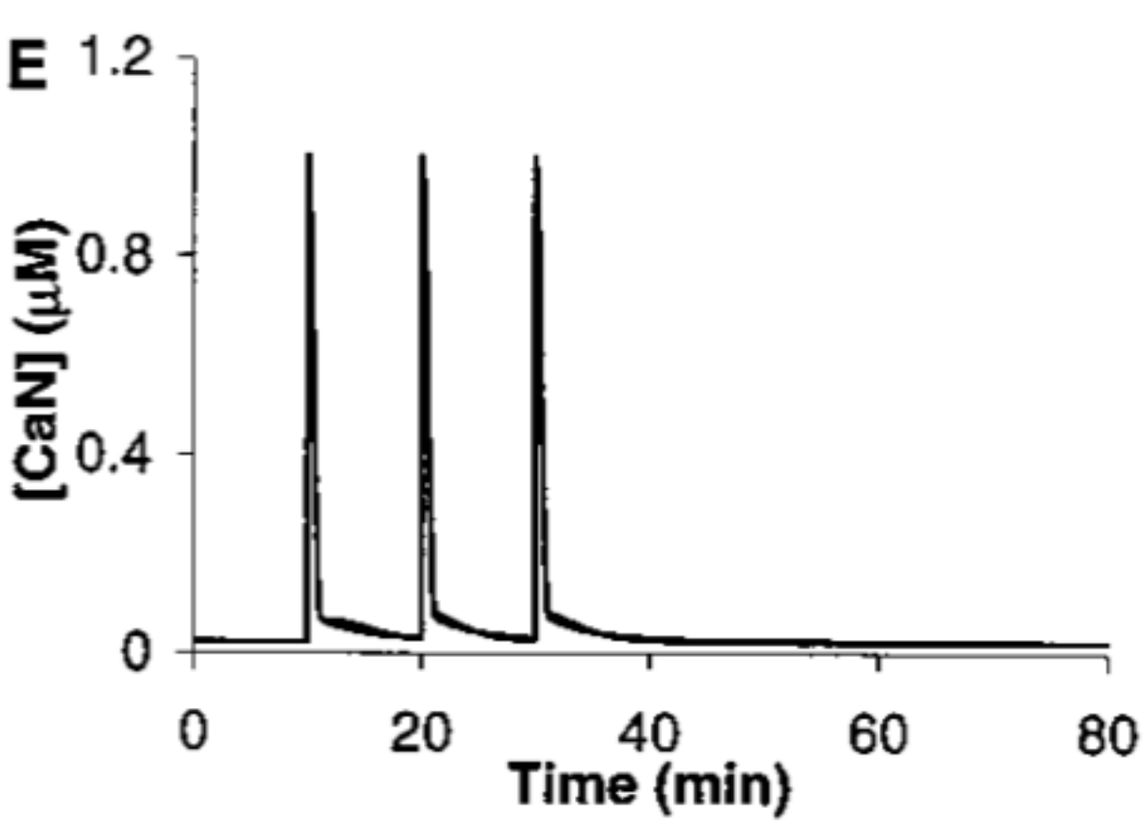
$$d[A]/dt = k_b[\mathrm{C}][\mathrm{D}] - k_f[\mathrm{A}][\mathrm{B}]$$

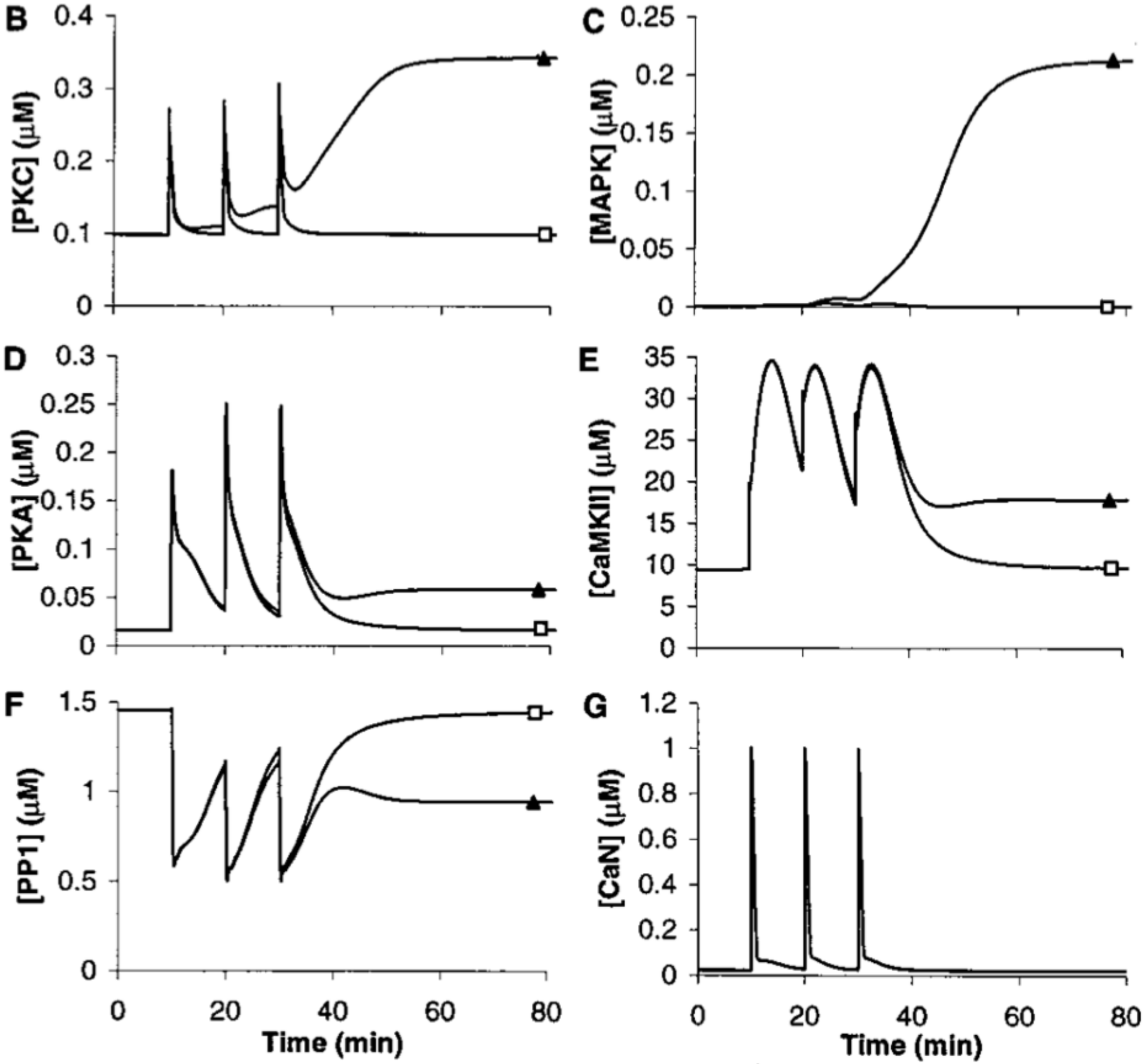
Emergent properties: molecular oscillators










B
[PKC] (μM)
0.4
0.3
0.2
0.1
0
C
[MAPK] (μM)
0.25
0.2
0.15
0.1
0.05
0
D
[PKA] (μM)
0.3
0.25
0.2
0.15
0.1
0.05
0
E
[CaMKII] (μM)
35
30
25
20
15
10
5
0
F
[PP1] (μM)
1.5
1
0.5
0
G
[CaN] (μM)
1.2
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
Time (min)

Emergent Properties of Networks of Biological Signaling Pathways

Upinder S. Bhalla and Ravi Iyengar*

SCIENCE VOL 283 15 JANUARY 1999

Signaling networks can "memorise" (store) information coming from extracellular stimuli by means of self-assembly of dynamic molecular circuits spontaneously generating emergent properties (bistability)!

in bistable behaviour with discrete steady-state activities, well-defined input thresholds for transition between states and prolonged signal output, and signal modulation in response to transient stimuli. These properties of signaling networks raise the possibility that information for "learned behaviour" of biological systems may be stored within intracellular biochemical reactions that comprise signaling pathways.

Emergent properties in cell cycle: Hysteresis

COMMENTARY

Hysteresis meets the cell cycle

Mark J. Solomon*

Department of Molecular Biophysics and Biochemistry, Yale University School of Medicine, C123 SHM, New Haven, CT 06520-8024

What, you may well ask, does hysteresis have to do with cell cycle progression? The last time most of us heard about hysteresis was in the context of the ferromagnetism that underlies tape players and floppy drives. In general, hysteresis means that it takes more of something to push a system from state A to state B than it does to keep the system in state B. Sha *et al.* (1) report in this issue of PNAS that the cell cycle of *Xenopus* egg extracts exhibits hysteresis in that the amount of cyclin needed to induce entry into mitosis is larger than the amount of cyclin needed to hold the extract in mitosis. This effect creates a nice bistable system with a ratchet to prevent slipping back from mitosis to interphase. This work also represents an excellent marriage of theory and experiment from the labs of John Tyson and Jill Sible, respectively, at Virginia Polytechnic Institute in Blacksburg. Some background, both experimental and theoretical, is necessary before discussing the importance of the current findings.

The eukaryotic cell cycle is driven by sequential activation and inactivation of cyclin-dependent protein kinases (CDKs) (2, 3). The CDK for entry into mitosis is Cdc2. Cdc2 activation requires binding to a regulatory protein (cyclin B) and activating phosphorylation (carried out by CDK-activating kinases, or CAKs). Even in the presence of cyclin and activating phosphorylation, Cdc2 can be inactivated by inhibitory phosphorylations (carried out by the Wee1 and Myt1 protein kinases). Inhibitory phosphorylations are removed by the Cdc25 protein phosphatases, which are the immediate triggers for entry into mitosis. CDKs regulating other cell cycle transitions can in addition be inhibited by direct binding of inhibitory proteins. Mitotic cyclins are subject to ubiquitin-mediated degradation at the end of mitosis by the action of the anaphase-promoting complex (APC; an E3, or ubiquitin ligase, of the ubiquitin system). Extremely important to the proper functioning of the cell cycle are checkpoints that ensure that key cell cycle events are not initiated until prior steps are completed. For example, a DNA replication checkpoint prevents Cdc2 activation until DNA replication is complete and the spindle assembly checkpoint prevents cyclin degradation via the APC until all chromosomes are properly aligned on the metaphase plate.

Many important advances in understanding entry into and exit from mitosis have come from biochemical studies of *Xenopus* egg extracts (2). This "simple" system is obtained by crushing frog eggs in the presence of minimal amounts of buffer. These extracts can undergo multiple rapid cell cycles, monitored either by the morphology of added nuclei or by assays of Cdc2 activity. These cell cycles can be driven by the endogenous synthesis and degradation of cyclin, or, if protein synthesis is inhibited, by the addition and subsequent degradation of exogenous cyclin. This system is probably the closest we can come to having the biochemist's ideal: a homogeneous bag of enzymes performing a complex task. The lack of checkpoints (unless one performs special tricks) and nonessential features one associates with more complex cell cycles in tissue culture (such as the need for nuclei, microtubules, subcompartments, DNA, etc.) allows the fundamental cell cycle oscillator to be studied in relative isolation. This system also exhibits conserved features of eukaryotic cell cycles: feedback loops. Thus, the rates of inhibitory phosphorylation of Cdc2 (mediated by Wee1 and Myt1) are decreased and the rates of dephosphorylation of these sites (mediated by Cdc25 proteins) are increased at the transition into mitosis (4). A second critical feedback is that cyclin synthesis, by inducing entry into mitosis and consequent activation of the anaphase-promoting complex, also leads to its own destruction. From these components and this basic wiring diagram one can hand wave one's way through the interphase-to-mitosis transition. But do we really understand this process? That's where mathematical modeling comes in.

This effect creates a bistable system with a ratchet to prevent slipping back to interphase.

There are many reasons to model a complex process quantitatively. It can be reassuring to plug the complete set of enzyme and substrate concentrations and their respective kinetic properties into a set of equations and have one's favorite process pop out. But that doesn't happen until you already know everything. More useful is modeling when significant ignorance remains so that the act of modeling points to important gaps to be filled or suggests interesting behavioral aspects of the system that can be tested. That's the goal John Tyson and collaborators have tackled for over a decade. Most of this work was done in collaboration with Bela Novak (Budapest University of Technology and Economics) who brought his experimental experience studying the *Schizosaccharomyces pombe* cell cycle. Their first joint effort was a model of the basic cell cycle in *Xenopus* egg extracts (5). The underlying logical approach was remarkably straightforward: Compile a large set of conceptually simple differential equations describing the rate of change of given components (say cyclin concentration or Cdc2 with inhibitory phosphorylations) and let a computer sort out the resulting complex web of interrelationships. What resulted was a good description of the state of knowledge at the time. Many models could do that, however, because the degrees of freedom are unbounded. However, what also came out of the simulations were some nonintuitive predictions not made by other published models. The key one tested in the Sha *et al*. article (1) was that the system would display hysteresis, thereby explaining its bistability: two stable states, interphase and mitosis.

The approach to testing this prediction is at once elegant and technically tricky. What is needed are accurate measurements of how much cyclin it takes to push an egg extract into mitosis and how much it takes to hold it there. The first measurement is easy: Just determine how much of a nondegradable cyclin needs to be added to an extract in interphase lacking any endogenous cyclin to induce entry into mitosis. The second measurement is more complicated because the determination of how much cyclin is needed to maintain mitosis requires that some of the cyclin that was necessary to get you into mitosis in the first place be removed. The solution was, essentially, to use a mixture of de-

See companion article on page 975.

*E-mail: Mark.Solomon@Yale.edu.

SEE COMMENTARY

CELL BIOLOGY

Hysteresis drives cell-cycle transitions in *Xenopus laevis* egg extracts

Wei Sha*, Jonathan Moore†, Katherine Chen*, Antonio D. Lassaletta*, Chung-Seon Yi*, John J. Tyson*, and Jill C. Sible*‡

*Department of Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0406; and †Cancer Research UK London Research Institute, Clare Hall Labs, South Mimms, Herts, EN6 3LD, United Kingdom

Edited by Thomas D. Pollard, Yale University, New Haven, CT, and approved November 21, 2002 (received for review September 3, 2002)

Cells progressing through the cell cycle must commit irreversibly to mitosis without slipping back to interphase before properly segregating their chromosomes. A mathematical model of cell-cycle progression in cell-free egg extracts from frog predicts that irreversible transitions into and out of mitosis are driven by hysteresis in the molecular control system. Hysteresis refers to toggle-like switching behavior in a dynamical system. In the mathematical model, the toggle switch is created by positive feedback in the phosphorylation reactions controlling the activity of Cdc2, a protein kinase bound to its regulatory subunit, cyclin B. To determine whether hysteresis underlies entry into and exit from mitosis in cell-free egg extracts, we tested three predictions of the Novak–Tyson model. (*i*) The minimal concentration of cyclin B necessary to drive an interphase extract into mitosis is distinctly higher than the minimal concentration necessary to hold a mitotic extract in mitosis, evidence for hysteresis. (*ii*) Unreplicated DNA elevates the cyclin threshold for Cdc2 activation, indication that checkpoints operate by enlarging the hysteresis loop. (*iii*) A dramatic "slowing down" in the rate of Cdc2 activation is detected at concentrations of cyclin B marginally above the activation threshold. All three predictions were validated. These observations confirm hysteresis as the driving force for cell-cycle transitions into and out of mitosis.

The biochemical oscillations that characterize early cell cycles of South African clawed frog, *Xenopus laevis*, can be reconstituted in cell-free egg extracts (1, 2). In this system, newly synthesized cyclin B associates with the cyclin-dependent kinase (Cdk) Cdc2 (3, 4). (Cyclin B/Cdc2 dimers are referred to as M-phase promoting factor.) Cyclin B/Cdc2 is rapidly inhibited by phosphorylation of Cdc2 on tyrosine 15 by two kinases, Wee1 and Myt1 (5, 6). Cdc2 remains inactive until this phosphate group is removed by the phosphatase, Cdc25 (7, 8). In turn, active Cdc2 phosphorylates and inhibits Wee1 (9) and phosphorylates and activates Cdc25 (10, 11). These positive feedback loops are responsible for the abrupt activation of cyclin B/Cdc2 at the G_2/M transition. Also important to this control system is a negative feedback loop in which active Cdc2 indirectly activates Fizzy, a protein that targets cyclin B for degradation via the ubiquitin–proteasome pathway (12, 13). Entry into mitosis is triggered by synthesis of cyclin B (1, 4), and exit from mitosis is triggered by degradation of cyclin B (14–16).

The cell cycle of frog egg extracts was selected as the first case for building a comprehensive mathematical model of the cell-cycle engine (17) because egg extracts contain the simplest functional control system for activation of the Cdks that drive cell-cycle transitions. The Novak–Tyson equations model this network of interlocking positive and negative feedback loops. In the model, the positive feedback loops create alternative states of low and high Cdc2 activity (interphase and M phase, respectively), and the negative feedback loop drives the control system back and forth between these states (Fig. 1*a*). During interphase, Cdc2 activity is low (because Cdc2 is phosphorylated), the rate of cyclin synthesis exceeds the rate of cyclin degradation, and cyclin accumulates in the extract. When total cyclin concentration exceeds an activation threshold (Fig. 1*a*), Cdc2 is abruptly activated by removal of the inhibitory phosphate groups. Because Cdc2 activates cyclin proteolysis, the rate of cyclin degradation in M phase exceeds its rate of synthesis, and cyclin concentration falls. However, according to the model, the extract stays in the "activated state" (unphosphorylated Cdc2 and rapid cyclin degradation) until Cdc2 activity falls below an inactivation threshold (Fig. 1*a*), when Cdc2 is abruptly inactivated by tyrosine phosphorylation. This cycle of events is called a hysteresis loop. Hysteresis underlies behaviors like ferromagnetism and DNA melting/reannealing. In both cases, the value of a control parameter (magnetic field, temperature) that induces a transition from one state to another is quite different from the value needed to induce the reverse transition.

Hysteretic transitions are discontinuous. Once the system has been switched on by moving the control parameter across the activation threshold, it cannot be switched off by bringing the control parameter back across the activation threshold in the opposite direction. Nonhysteretic switches behave differently, switching on and off at the same value. A reversible Cdc2 switch would look like Fig. 1*b*.

Although several authors have suggested that progress through the cell cycle is governed by a hysteresis loop like Fig. 1*a* (17, 20–23), there is another theoretically plausible explanation for switch-like behavior at mitosis. Periodic cyclin degradation could be driven by a time-delayed negative feedback loop involving Cdc2 activation of Fizzy, without participation from Wee1 and Cdc25. Such a model was proposed by Goldbeter (19) and is consistent with a nonhysteretic switch (Fig. 1*b*). The distinction between these two pictures had not been investigated experimentally until now.

Cyclin thresholds for entry into or exit from mitosis have been measured experimentally. Solomon *et al.* (4) demonstrated that there is a cyclin threshold for Cdc2 activation at mitosis 1 in frog egg extracts. Subsequently, Holloway *et al.* (15) and Stemmann *et al.* (16) demonstrated a cyclin threshold for exit from mitosis. All of these experiments are consistent with either Fig. 1 *a* or *b*. In this study, we measure the thresholds for Cdc2 activation and inactivation going into and out of the same mitosis to distinguish between the mechanisms proposed in Fig. 1*a* (hysteretic) and Fig. 1*b* (nonhysteretic).

Another distinction between Fig. 1 *a* and *b* is that, in the case of hysteresis, the underlying dynamical system is bistable. That is, for certain fixed values of the control parameter, the governing dynamical equations admit two different stable steady-state solutions separated by an unstable steady state (Fig. 1*a*). (Stable and unstable steady states are illustrated by a ball rolling on an undulating landscape. At the bottom of any pit ∪, the ball is in a stable steady state, whereas, if balanced at the top of a hill ∩,

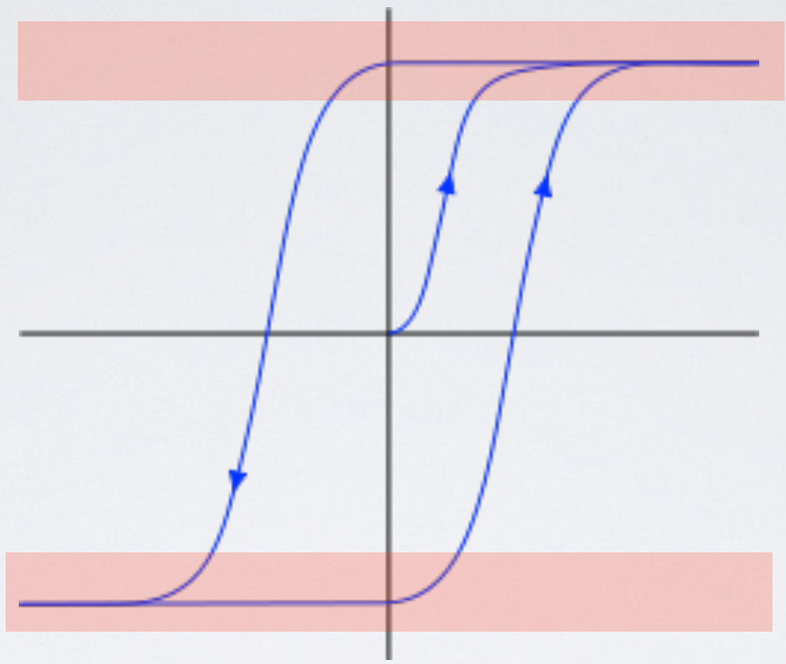
This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: Cdk, cyclin-dependent kinase; Δcyclin B, recombinant, nondegradable human cyclin B; CSF, cytostatic factor; CHX, cycloheximide; APH, aphidicolin.

See commentary on page 771.

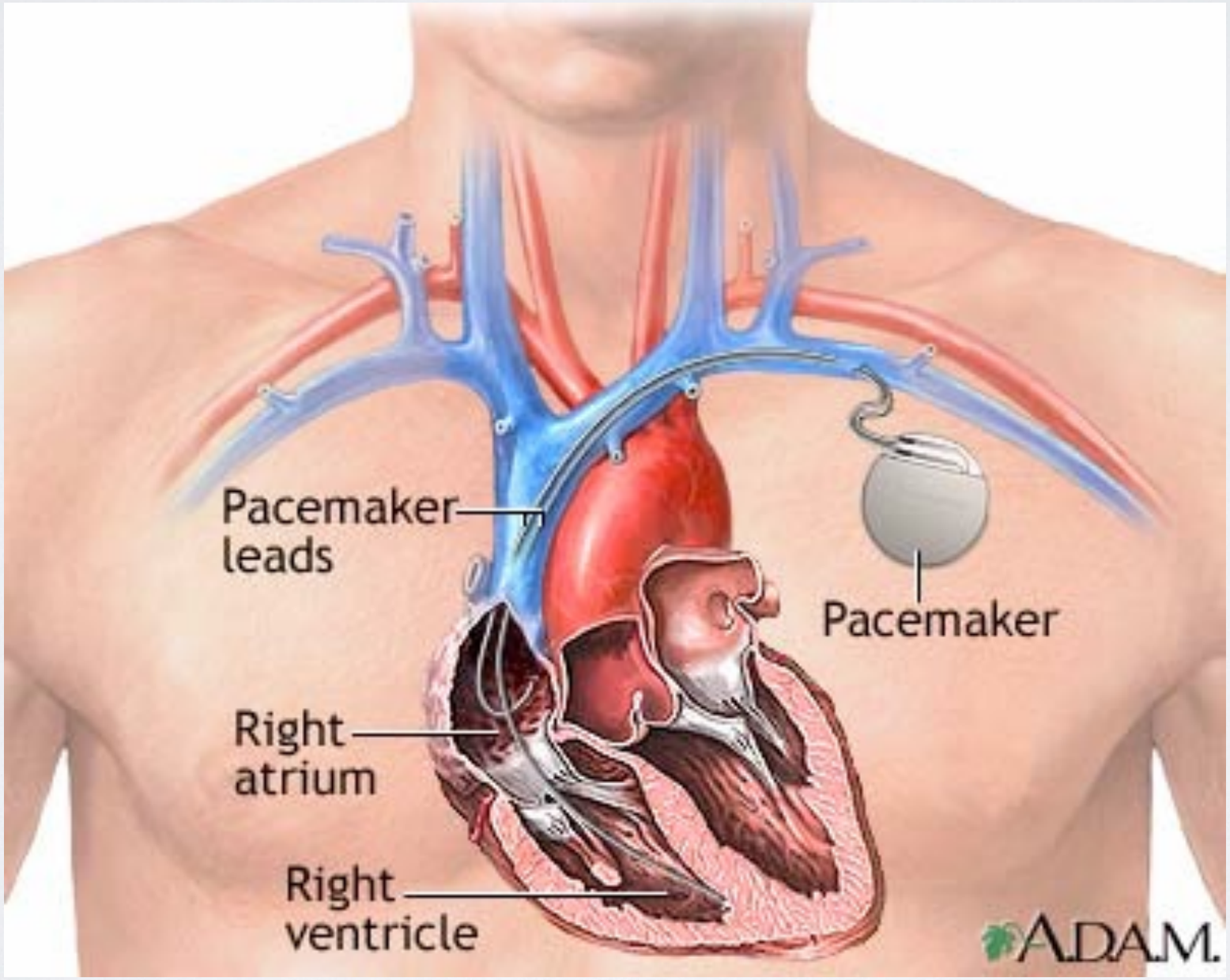
‡To whom correspondence should be addressed. E-mail: siblej@vt.edu.

Hysteresis **is a complex system property characterised by a delayed reaction to an applied stress in dependence on the previous state.**

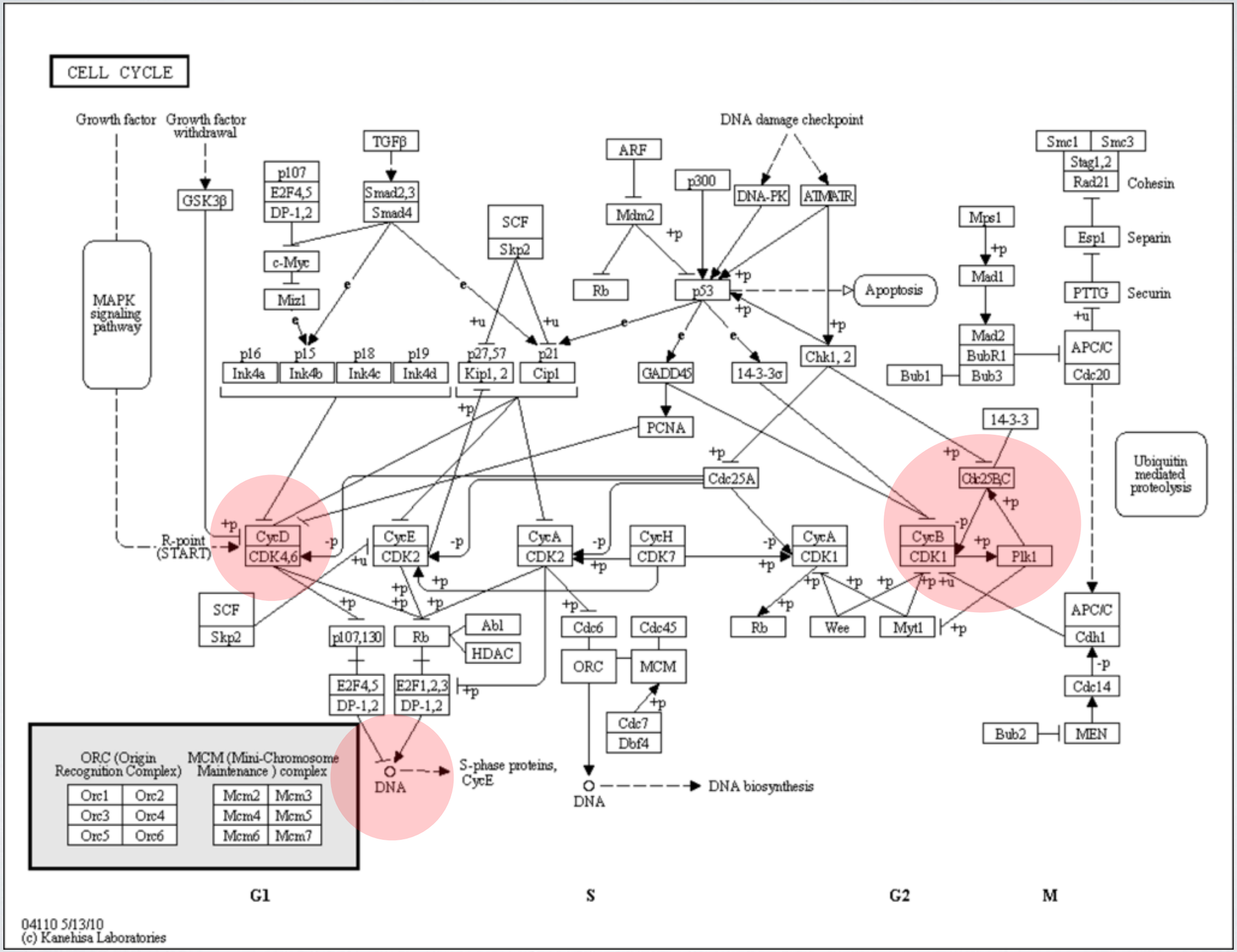


System hysteresis is represented **on a graph as a function of the stimulus**, with a characteristic closed curve. In presence of hysteresis, it is obtained a doubling of the curve: if traversed from left to right there is a path, if traversed in the reverse direction an opposite path is obtained. Often two horizontal sections are obtained: one upper and one lower, which represent the saturation limits. The amplitude of the closed curve is the entity of hysteresis. In a system without hysteresis curves just generate a single line (no-loop kinetics).

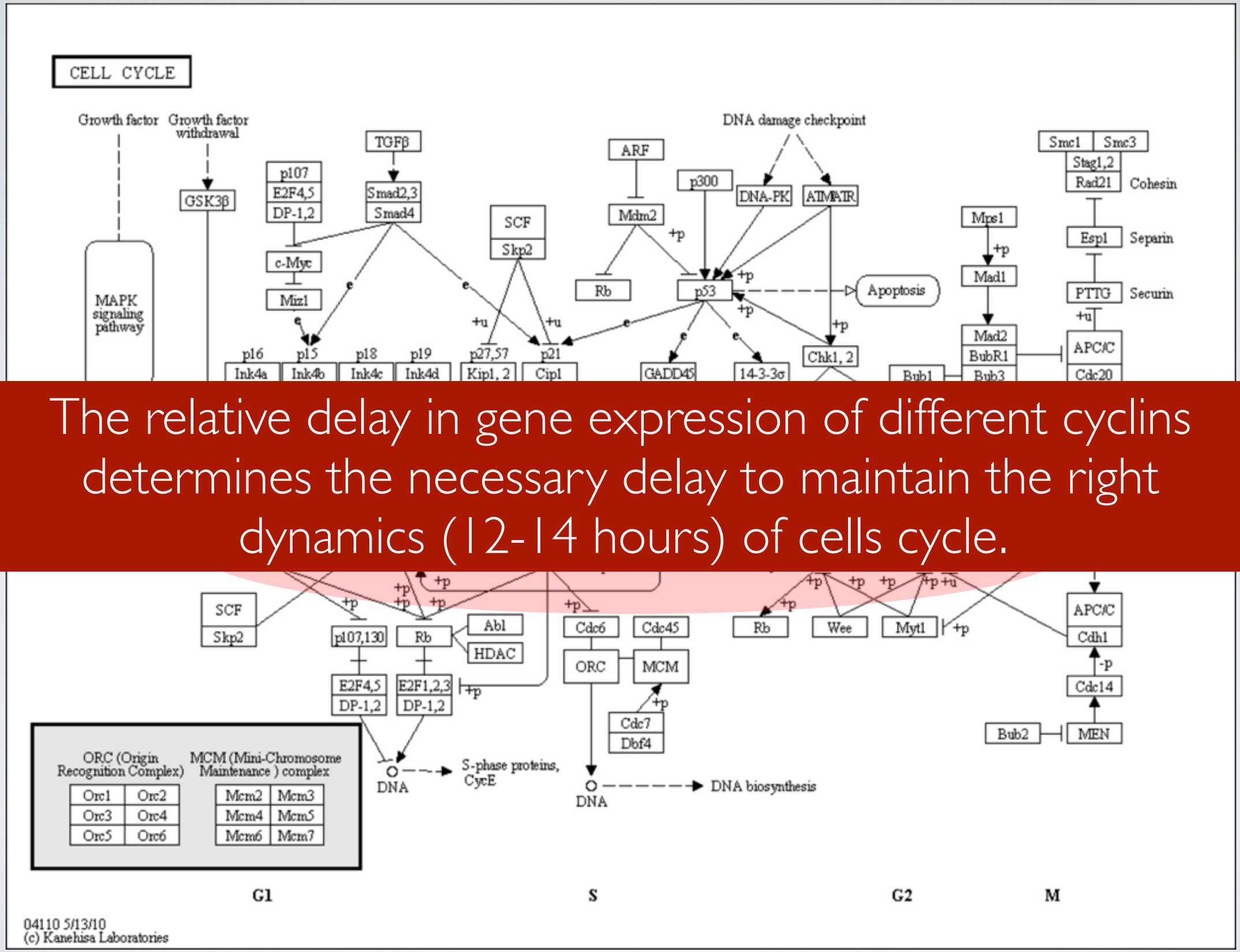
In medicine hysteresis is the programmed delay in heart pacemakers to avoid interference with the proper activity of the sinoatrial node.



Cell cycle


CELL CYCLE
Growth factor
Growth factor withdrawal
MAPK signaling pathway
GSK3β
TGFβ
p107
E2F4,5
DP-1,2
Smad2,3
Smad4
c-Myc
Miz1
SCF
Skp2
ARF
Mdm2
p300
DNA damage checkpoint
DNA-PK
ATM/ATR
Rb
p53
Apoptosis
Chk1, 2
GADD45
14-3-3σ
p16
Ink4a
p15
Ink4b
p18
Ink4c
p19
Ink4d
p27,57
Kip1, 2
p21
Cip1
PCNA
Cdc25A
R-point
(START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Cdc25B,C
14-3-3
Plk1
Mps1
Mad1
Mad2
BubR1
Bub3
Bub1
Smc1
Smc3
Stag1,2
Rad21
Cohesin
Esp1
Separin
PTTG
Securin
APC/C
Cdc20
Ubiquitin mediated proteolysis
SCF
Skp2
p107,130
Rb
Abl
HDAC
E2F4,5
DP-1,2
E2F1,2,3
DP-1,2
Cdc6
Cdc45
ORC
MCM
Cdc7
Dbf4
Rb
Wee
Myt1
APC/C
Cdh1
Cdc14
Bub2
MEN
ORC (Origin Recognition Complex)
Orc1
Orc2
Orc3
Orc4
Orc5
Orc6
MCM (Mini-Chromosome Maintenance) complex
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
DNA
S-phase proteins, CycE
DNA
DNA biosynthesis
G1
S
G2
M
04110 5/13/10
(c) Kanehisa Laboratories

Cell cycle is controlled by hysteresis at the level of ciclins



The relative delay in gene expression of different cyclins determines the necessary delay to maintain the right dynamics (12-14 hours) of cells cycle.

glume, which can secondarily alter the fate of the SMs when de-repressed.

None of these models adequately address the fact that the *bd1* SM has different fates in the tassel and ear. It is unlikely that *bd1b* partially compensates for the loss of BD1 in the tassel, as proposed for the *zag1/zmm2* duplication in maize (*14*), because we have been unable to detect *bd1b* transcript in any tissues (*15*). It is possible that a different tassel-specific factor may function redundantly with *bd1*. Given the intense selective pressure on the maize ear, it is not surprising that the ear and tassel are genetically distinct.

The expression pattern and mutant phenotype of *bd1* show similarities to the *FIMBRIATA/UFO* genes of *Antirrhinum* and *Arabidopsis,* respectively (*16*, *17*). Both genes are expressed in a ring at the base of the floral meristem adjacent to the sepals, and the *Antirrhinum* mutant shows a partial loss of lateral determinacy within the meristem. In the case of *UFO*, the basal floral meristems may be replaced with coflorescence branches (*18*). In *Arabidopsis*, the *UFO* and *LEAFY* genes have been proposed to be coregulators of floral meristem identity (*19*). Therefore, BD1 may interact with other SM identity factors to impose determinate meristem fates. As in wild type, the maize *LEAFY* ortholog is expressed in the SPMs and SMs of *bd1* mutants (*5*). However, the genetic interaction between *bd1* and *leafy* is unknown and awaits identification of *leafy* mutants in maize.

To date, *bd1* is the only maize mutant that specifically displays altered SM identity. Several maize mutants that affect SM determinacy have been described, such as *Tassel-seed6* (*20*) and *indeterminate spikelet1* (*21*). Both these mutants display SMs that initiate more than two florets per spikelet, and interestingly, both show normal patterns of *bd1* expression in the SM (fig. S1). The latter result indicates that SM identity is acquired before SM determinacy. Recently, it has been shown that SM identity and determinacy are interdependent, as two genes that control SM determinacy, *indeterminate spikelet1* and *indeterminate floral apex1,* also show SM identity defects as a double mutant (*22*).

The grass spikelet is conventionally interpreted as a strongly contracted branch system—literally, a little spike (*23*). If this interpretation is correct, then genes should exist that, when mutated, cause the spikelet to revert to a branchlike structure. We have identified a gene that regulates spikelet versus branch meristem fates within the inflorescence of maize, and whose sequence and expression are conserved in other grasses such as rice and sorghum. Our data suggest that the expression of *bd1* is fundamental to grass spikelet formation and may have played a role in the origin of this evolutionary novelty.

References and Notes

1. P. McSteen, D. Laudencia-Chingcuanco, J. Colasanti, *Trends Plant Sci.* **5**, 61 (2000).
2. P. C. Cheng, R. I. Greyson, D. B. Walden, *Am. J. Bot.* **70**, 450 (1983).
3. J. H. Kempton, *J. Hered.* **25**, 29 (1934).
4. L. Colombo *et al.*, *Plant J.* **16**, 353 (1998).
5. Materials and methods are available as supporting material at *Science* Online.
6. T. Helentjaris, D. Weber, S. Wright, *Genetics* **118**, 353 (1988).
7. S. Mathews, R. J. Mason-Gamer, R. E. Spangler, E. A. Kellogg, *Int. J. Plant Sci.* **163**, 441 (2002).
8. S. Y. Fujimoto, M. Ohta, A. Usui, H. Shinshi, M. Ohme-Takagi, *Plant Cell* **12**, 393 (2000).
9. J. L. Riechmann, E. M. Meyerowitz, *Biol. Chem.* **379**, 633 (1998).
10. E. van der Graaff, A. D. Dulk-Ras, P. J. Hooykaas, B. Keller, *Development* **127**, 4971 (2000).
11. K. Wilson, D. Long, J. Swinburne, G. A. Coupland, *Plant Cell* **4**, 659 (1996).
12. H. Banno, Y. Ikeda, Q.-W. Niu, N.-H. Chua, *Plant Cell* **13**, 2609 (2001).
13. A. Barkan, R. Martienssen, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 3502 (1991).
14. M. Mena *et al.*, *Science* **274**, 1537 (1996).
15. G. Chuck, unpublished observations.
16. R. Simon, R. Carpenter, S. Doyle, E. Coen, *Cell* **78**, 99 (1994).
17. G. C. Ingram *et al.*, *Plant Cell* **7**, 1501 (1995).
18. J. Z. Levin, E. M. Meyerowitz, *Plant Cell* **7**, 529 (1995).
19. I. Lee, D. S. Wolfe, O. Nilsson, D. Weigel, *Curr. Biol.* **7**, 95 (1997).
20. E. E. Irish, *Am. J. Bot.* **84**, 1502 (1997).
21. G. Chuck, R. Meeley, S. Hake, *Genes Dev.* **12**, 1145 (1998).
22. D. Laudencia-Chingcuanco, *Development* **129**, 2629 (2002).
23. W. D. Clayton, in *Reproductive Versatility in the Grasses*, G. P. Chapman, Ed. (Cambridge Univ. Press, Cambridge, 1990), pp. 32–51.

24. M. Mena, M. A. Mandel, D. R. Lerner, M. F. Yanofsky, R. J. Schmidt, *Plant J.* **8**, 845 (1995).
25. We thank I. Murray, C. Whipple, and M. Ritter for technical assistance, and members of the Hake laboratory and J. Fletcher for comments on the manuscript. This work was supported by grants from the National Science Foundation to R.S. and S.H.

Supporting Online Material

www.sciencemag.org/cgi/content/full/298/5596/1238/DC1
Materials and Methods
Fig. S1

2 August 2002; accepted 17 September 2002

The IκB–NF-κB Signaling Module: Temporal Control and Selective Gene Activation

Alexander Hoffmann,[1]* Andre Levchenko,[2]* Martin L. Scott,[3]† David Baltimore[1]‡

Nuclear localization of the transcriptional activator NF-κB (nuclear factor κB) is controlled in mammalian cells by three isoforms of NF-κB inhibitor protein: IκBα, -β, and -ε. Based on simplifying reductions of the IκB–NF-κB signaling module in knockout cell lines, we present a computational model that describes the temporal control of NF-κB activation by the coordinated degradation and synthesis of IκB proteins. The model demonstrates that IκBα is responsible for strong negative feedback that allows for a fast turn-off of the NF-κB response, whereas IκBβ and -ε function to reduce the system's oscillatory potential and stabilize NF-κB responses during longer stimulations. Bimodal signal-processing characteristics with respect to stimulus duration are revealed by the model and are shown to generate specificity in gene expression.

The transcription factor NF-κB regulates numerous genes that play important roles in inter- and intracellular signaling, cellular stress responses, cell growth, survival, and apoptosis (*1*–*3*). As such, the specificity and temporal control of gene expression are of crucial physiological interest. Furthermore, the realization of the potential of NF-κB as a drug target for chronic inflammatory diseases or within chemotherapy regimens (*4*, *5*) is dependent on understanding the specificity mechanisms that govern NF-κB–responsive gene expression.

Five related mammalian gene products participate in NF-κB functions (RelA/p65, cRel, RelB, p50, p52), but the predominant species in many cell types is a p65:p50 heterodimer. Its activity is largely controlled by three IκB isoforms (IκBα, -β, and -ε) that bind to NF-κB, preventing its association with DNA and causing its localization to the cytoplasm. Signals from various stimuli are transduced to the IκB kinase (IKK) complex, which phosphorylates each IκB isoform, leading to its ubiquitination and proteolysis (*6*). IκB degradation allows NF-κB to translocate to the nucleus and bind DNA (Fig. 1A). The specific role of each IκB protein in regulating NF-κB is not understood. Mice

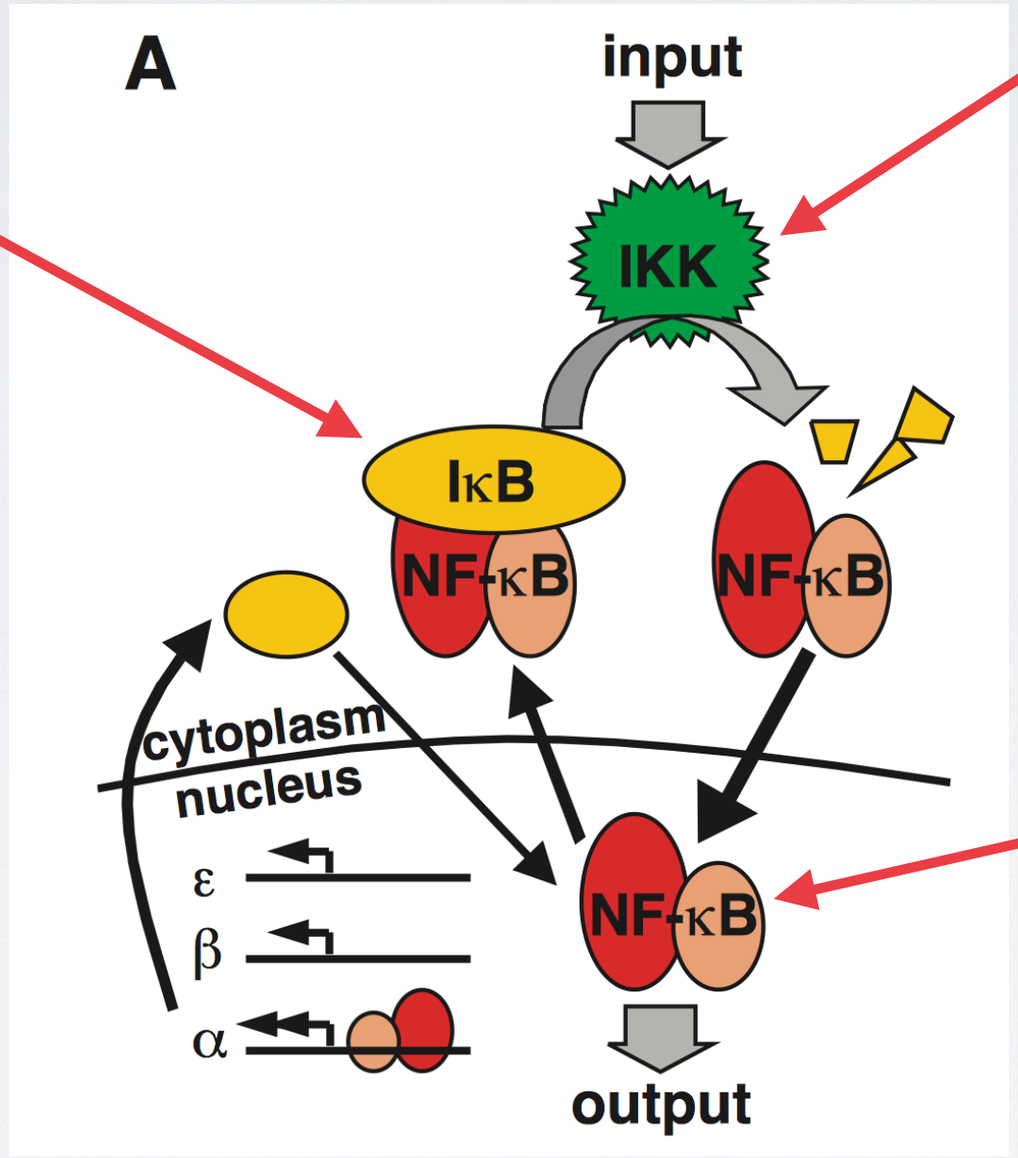
[1]Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA. [2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*These authors contributed equally to this work.
†Present address: Biogen, Incorporated, Cambridge, MA 02142, USA.
‡To whom correspondence should be addressed. E-mail: baltimo@caltech.edu

The IkB–NF-kB signaling module.

1-

NF-kB is held <u>inactive</u> in the cytoplasm by three IkB isoforms.



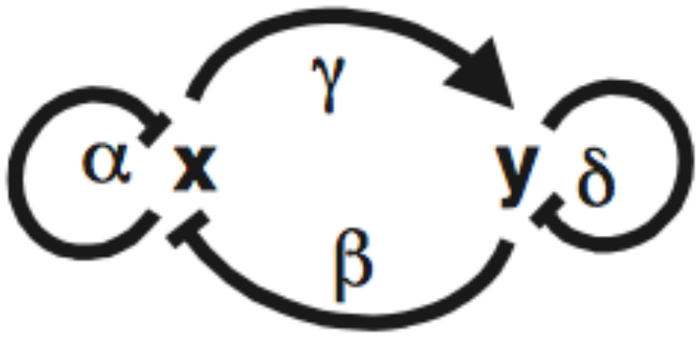
2-

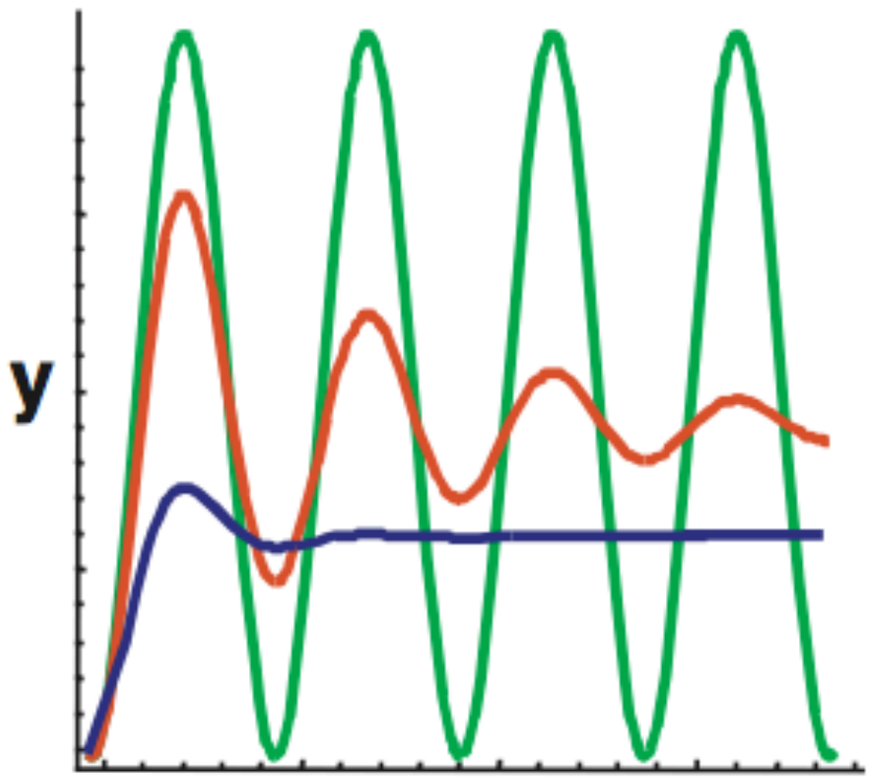
<u>Cell stimulation</u> activates the IKK complex, leading to phosphorylation and degradation of IkB proteins.

3-

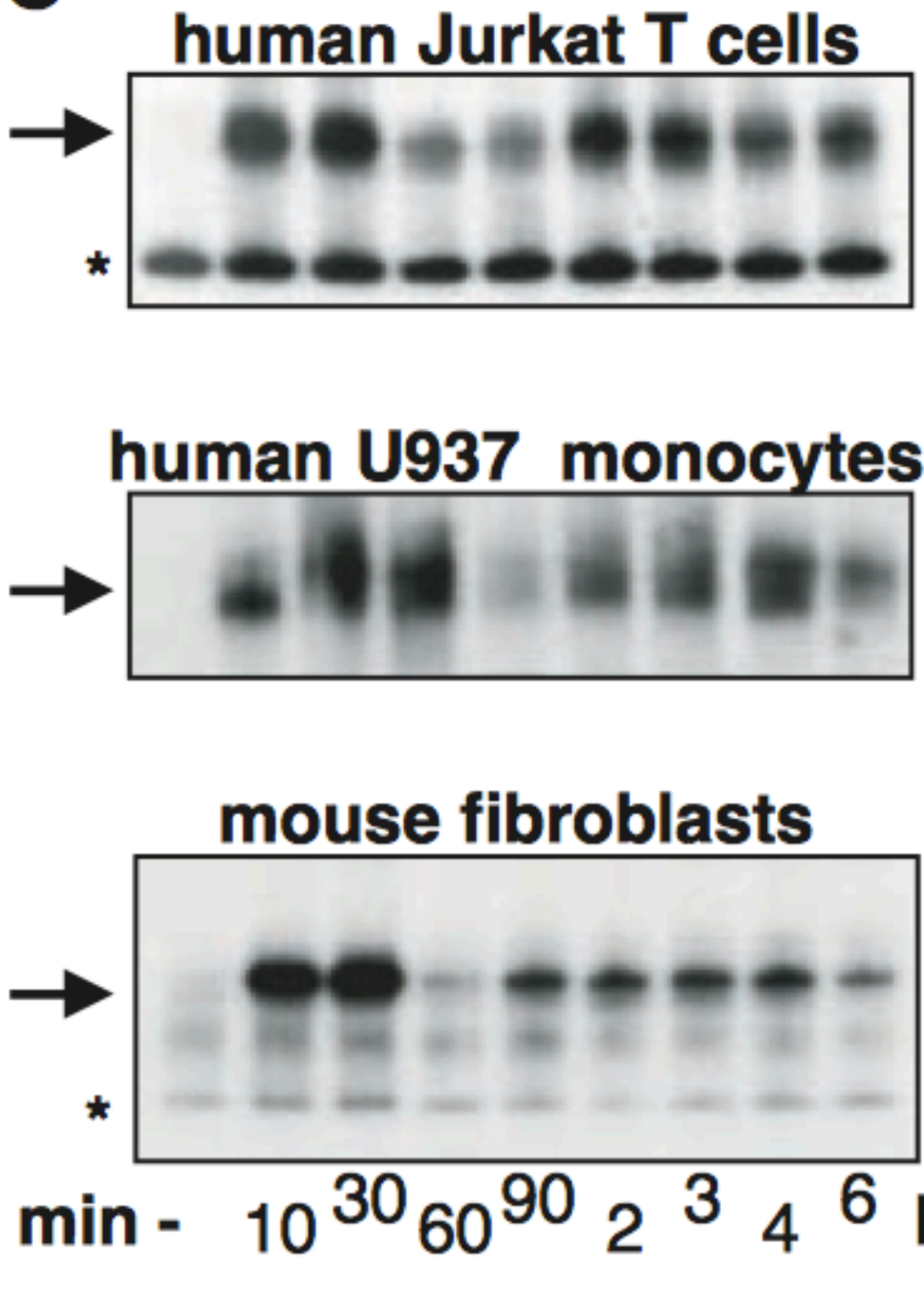
Free NF-kB translocates to the nucleus, activating genes, including IkBs. IkBs <u>are synthesised at a steady rate</u>, allowing for complex temporal control of NF-kB activation involving a negative feedback.

B


γ
α
x
y
δ
β


y
time

C


human Jurkat T cells
*
human U937 monocytes
mouse fibroblasts
*
min - 10 30 60 90 2 3 4 6 hr

The IkB–NF-kB Signaling Module: Temporal Control and Selective Gene Activation

Alexander Hoffmann, 1 * Andre Levchenko, 2 * Martin L. Scott, 3 † David Baltimore 1 ‡

Nuclear localization of the transcriptional activator NF-kB (nuclear factor kB) is controlled in mammalian cells by three isoforms of NF-kB inhibitor protein: IkB, α-, β- and -ε. Based on simplifying reductions of the IkB–NF-kB signaling module in knockout cell lines, **we present a computational model that describes the temporal control of NF-kB activation by the coordinated degradation and synthesis of IkB proteins**. The model demonstrates that IkB α is responsible for ***strong negative feedback*** that allows for a fast turn-off of the NF-kB response, whereas IkB α- and -ε function to ***reduce the system's oscillatory potential and stabilize NF-kB responses during longer stimulations***. **Bimodal signal-processing characteristics with respect to stimulus duration are revealed by the model and are shown to generate specificity in gene expression.**

Overall conclusion

To generate dynamic models of biological networks it is mandatory to be able to:

1- reconstruct the network architecture

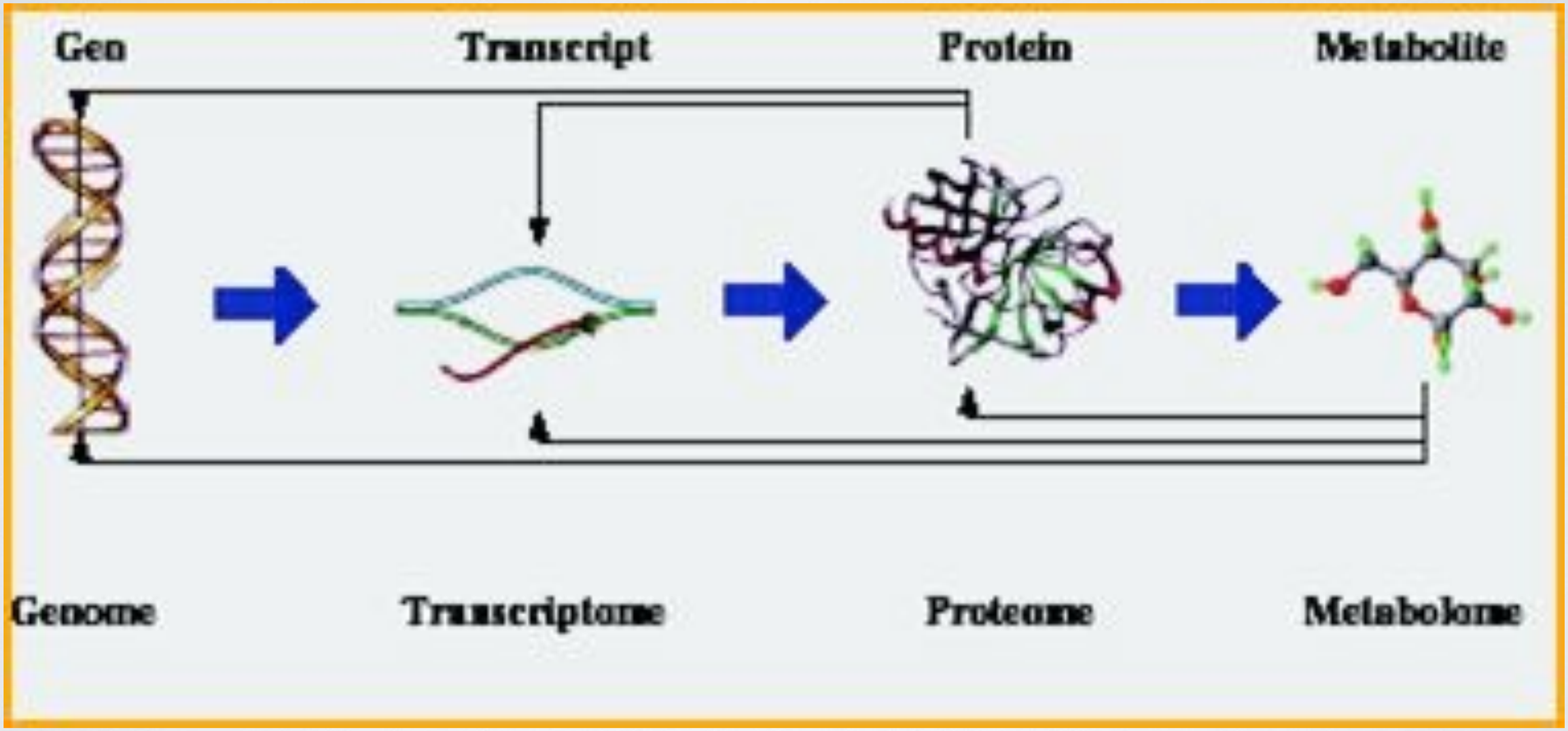
2- to know all kinetics parameters, including kinetic constants, reactant concentration, intracellular localisation-diffusion and time-frame state evolution

3- to decide whether to apply a continuous (differential equation) versus discrete model

4- a good …. software allowing time-evolution computation

5- possibly an expected phenomenological output (e.g. gene expression) to correlate the network emergent behaviour to cell events

(9)
Technologies in Systems Biology: the high-performance technologies (high throughput technologies)



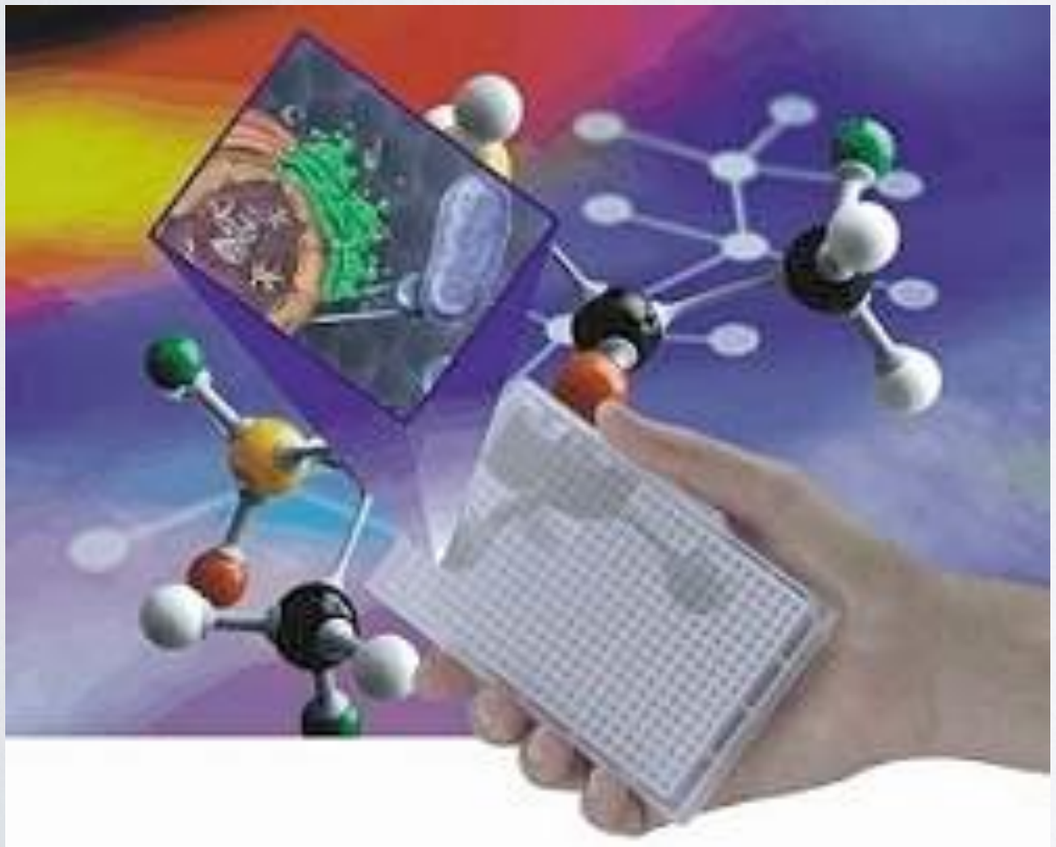
Thousands to millions of context-specific variations!

To cope with the enormous space of molecular variability occurring in the organisms we need powerful, yet cost efficient, innovative technologies

The experimental high-performance technologies (high-throughput screening (HTS)) are a set of advanced technologies that allow running thousands or millions of experimental quantifications simultaneously.

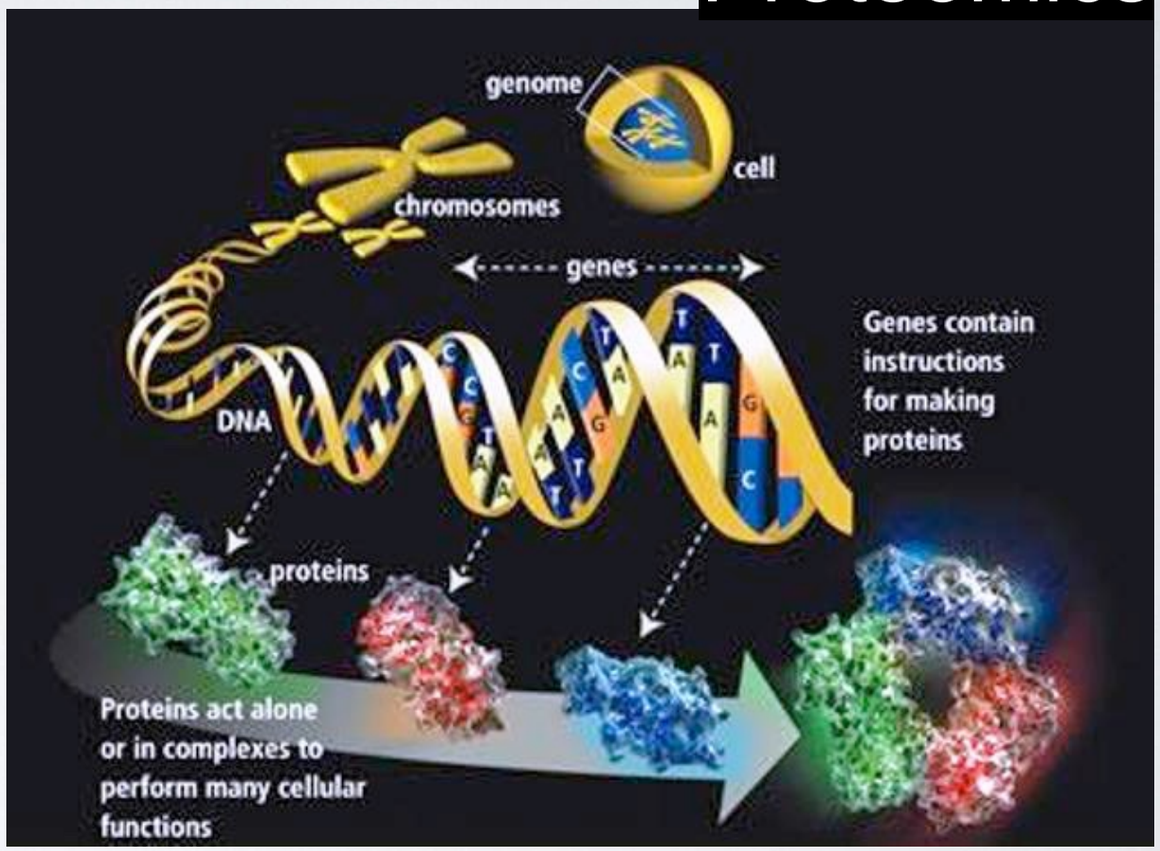
"Omics" technologies

Genomics



Metabolomics

Proteomics



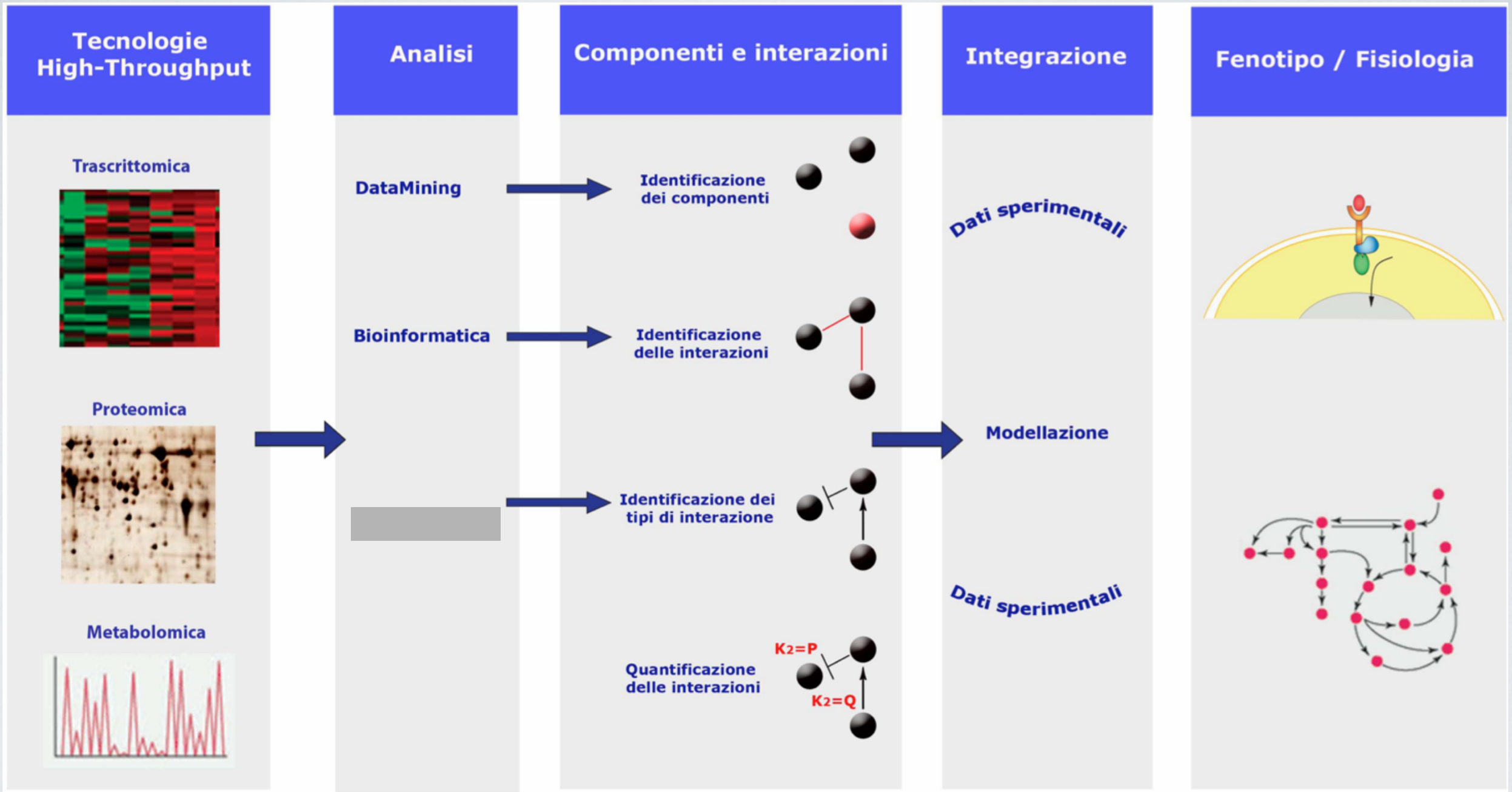
HTSs are based on the use of robotic technologies, highly miniaturised and automated, producing enormous amounts of experimental data regarding the molecular state of a biological sample, at the genomic, proteomic and / or metabolomic level.

HTS-based analysis must, therefore, always be associated with **powerful bioinformatics and computational analysis tools** in order to reconstruct and analyse the molecular network and mechanisms that control biological phenomena.

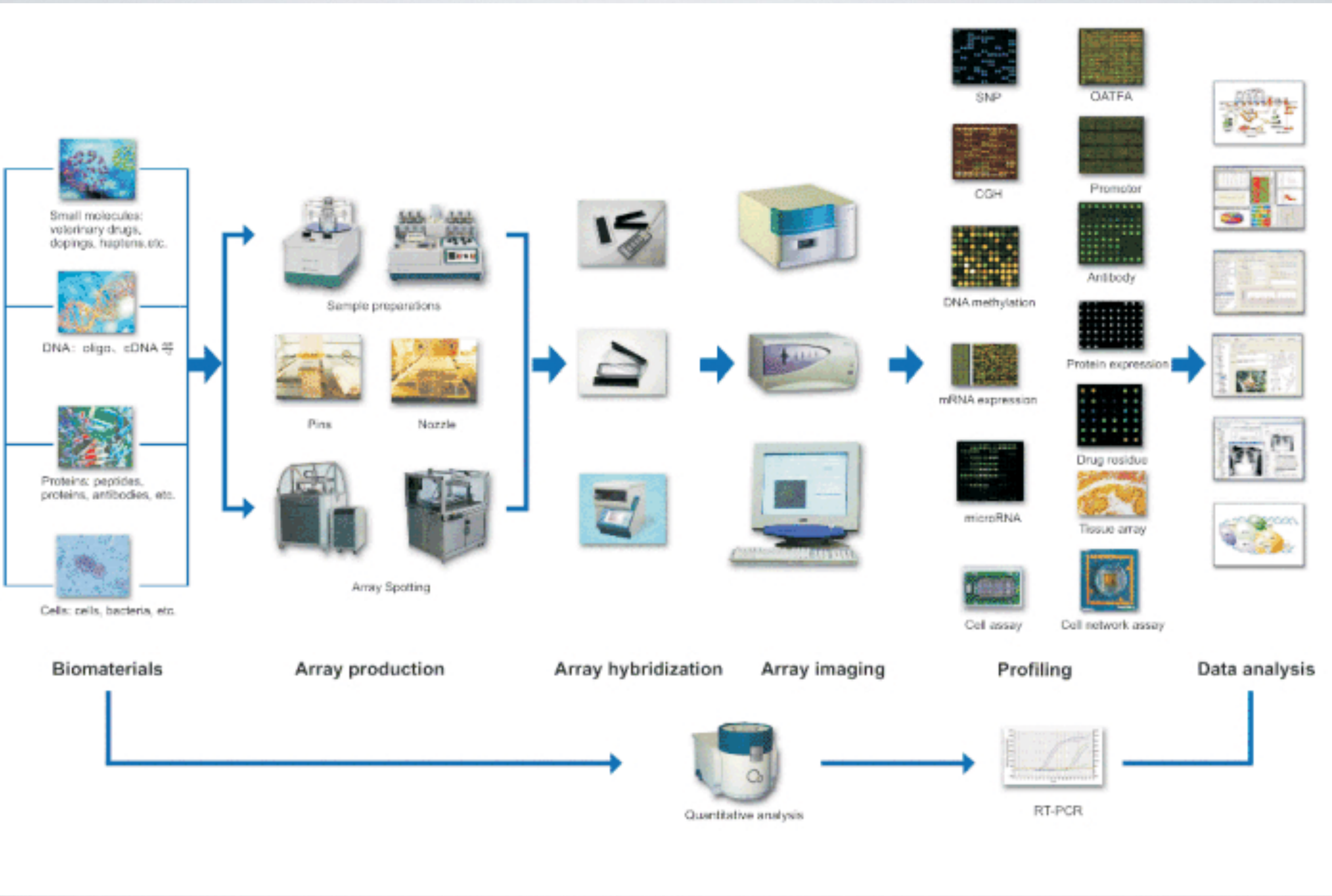
Notably:

The HTS are propadeutic to systems biology.

“Omics” technologies flow-chart

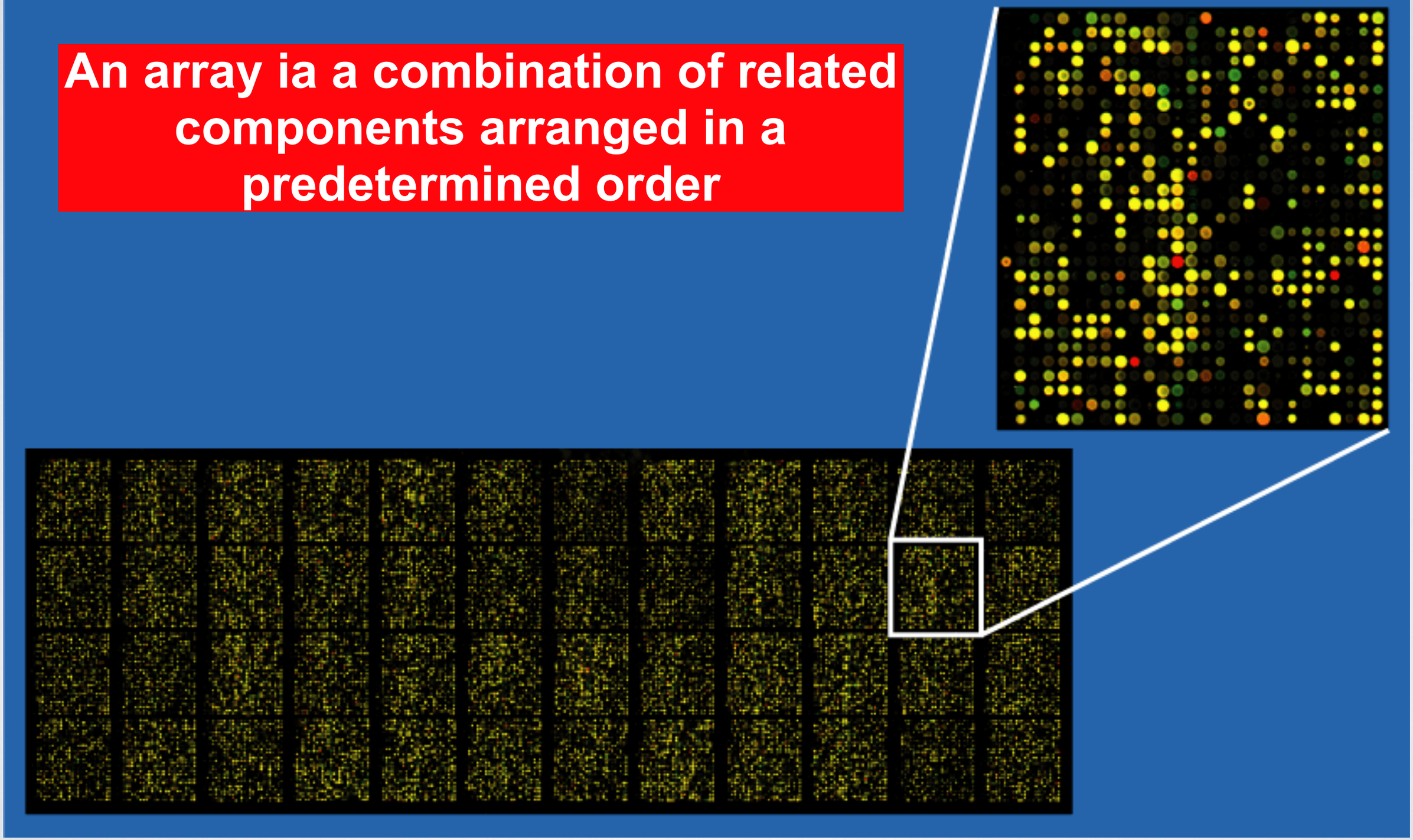


Gene microarray analysis flow-chart



Microarrays (DNA, proteins)

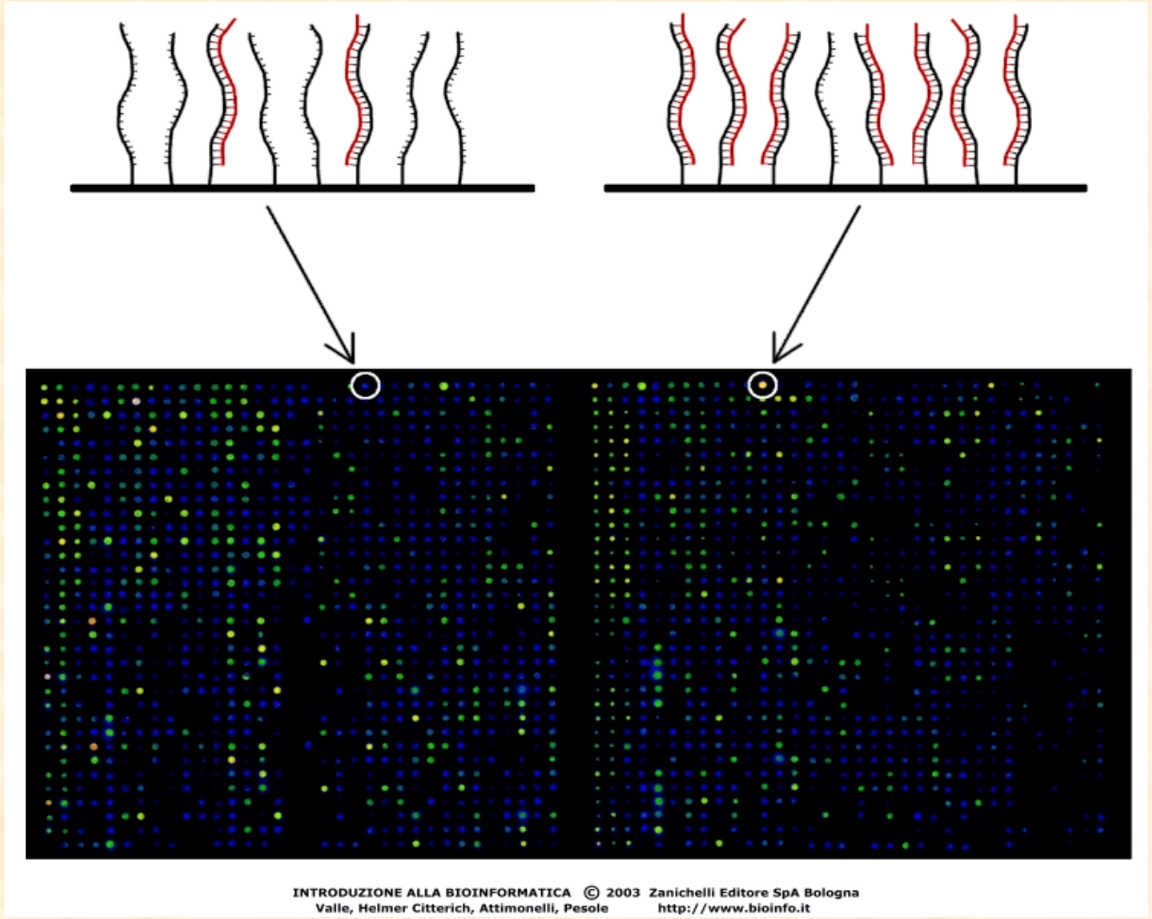
An array ia a combination of related components arranged in a predetermined order



DNA microarrays

TRANSCRIPTOMICS

Generation of mRNA expression profiles of a specific cell to generate context-specific gene expression data



•ARRAYEXPRESS e MIANE: Un database ed uno standard per gli esperimenti di microarray http://www.ebi.ac.uk/Databases/microarray.html

DNA microarray

cDNA microarray

- probes generated before glass binding/adsorption)
- probe length: 200-400 mer
- probes spotted upon generation

Oligonucleotide microarray

- probes generated directly on glass -> *in situ-synthesis*
- short probes: 20-40 mer (Affimetrix GeneChip)
- long probes: 60 mer /Agilent)

cDNA microarray

Treated cells

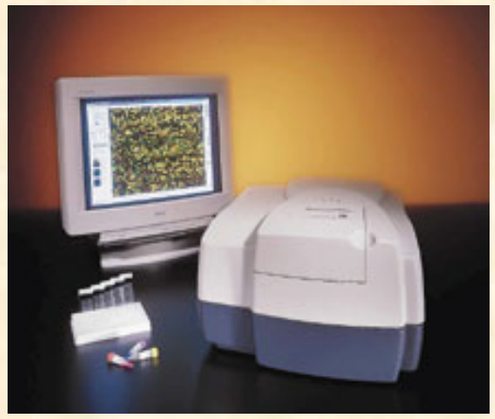
Un-treated cells

RNA

Cy5

cDNA

Cy3



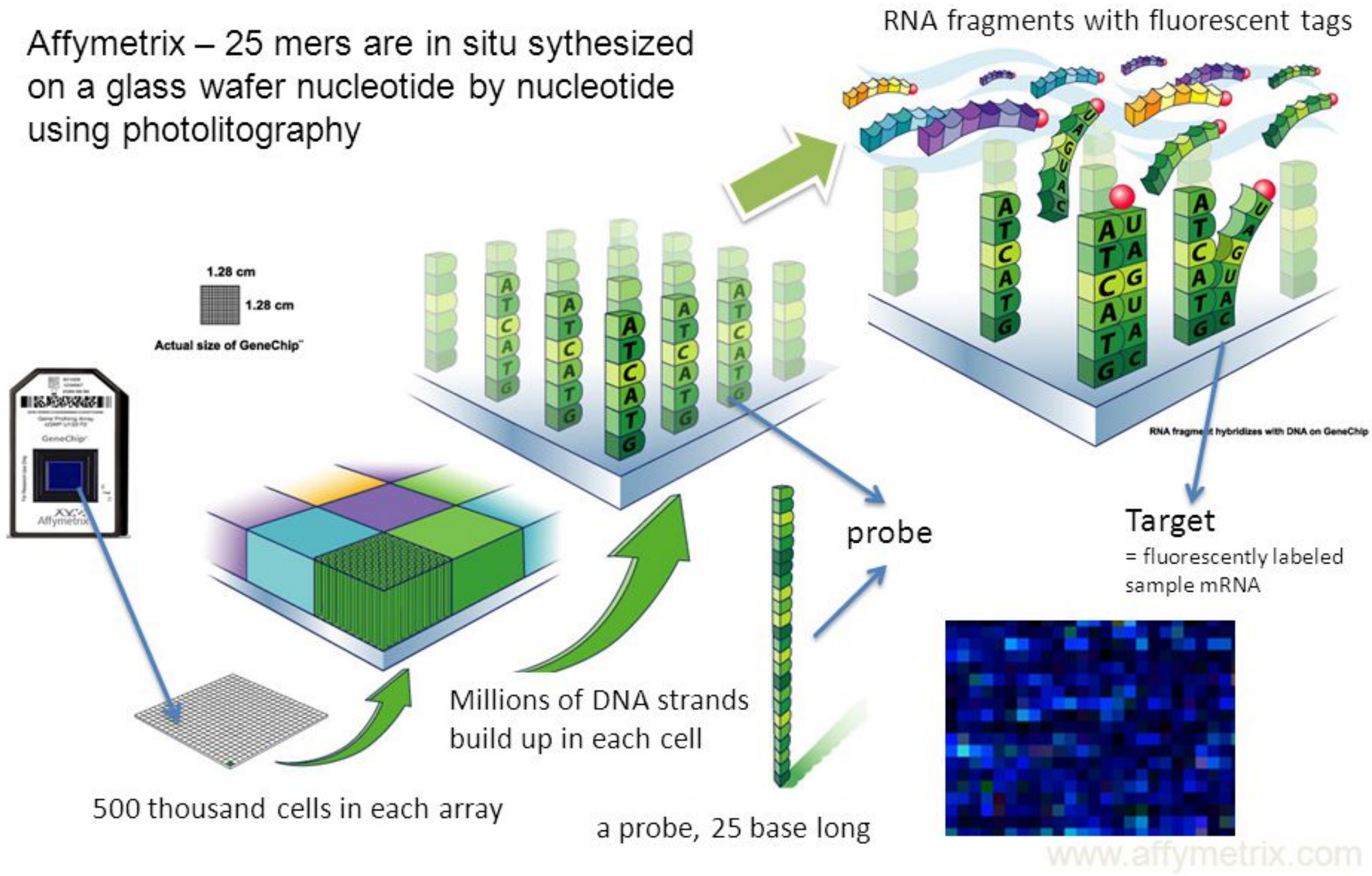
1- Total RNA extraction
2- mRNA isolation and reverse transcription to cDNA
3- Fluorescente labelling
4- Ibridization
5- Glass scanning

Oligonucleotide microarray

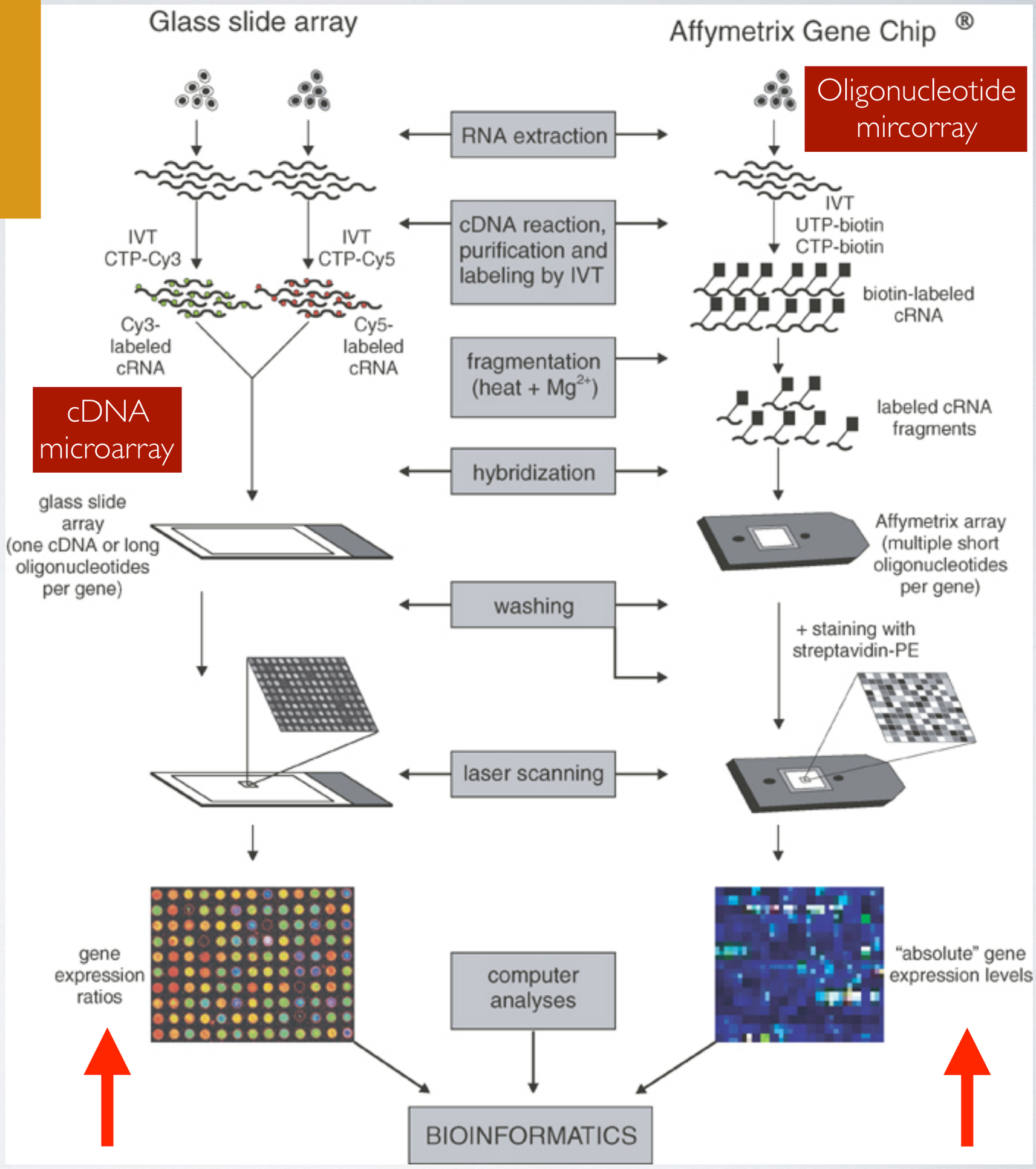
Affymetrix Microarrays

photolithographic synthesis of oligonucleotide on microarrays

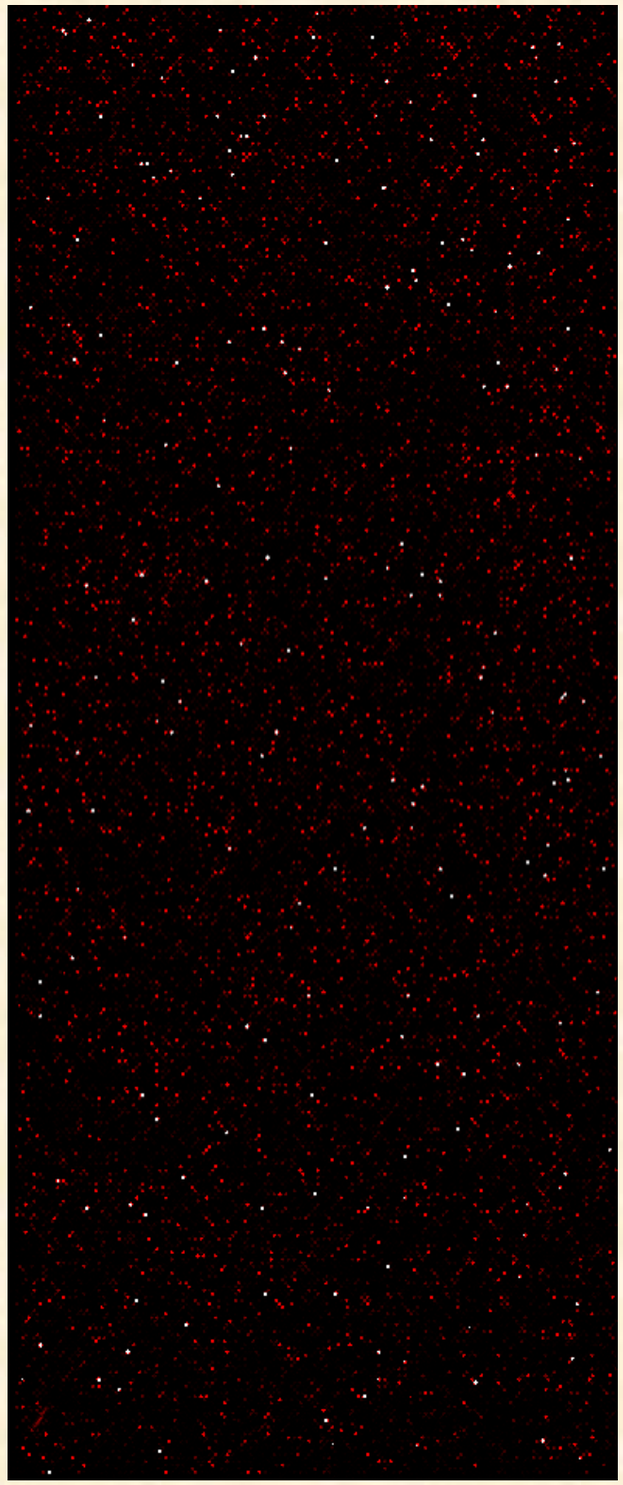
Affymetrix – 25 mers are in situ sythesized on a glass wafer nucleotide by nucleotide using photolitography

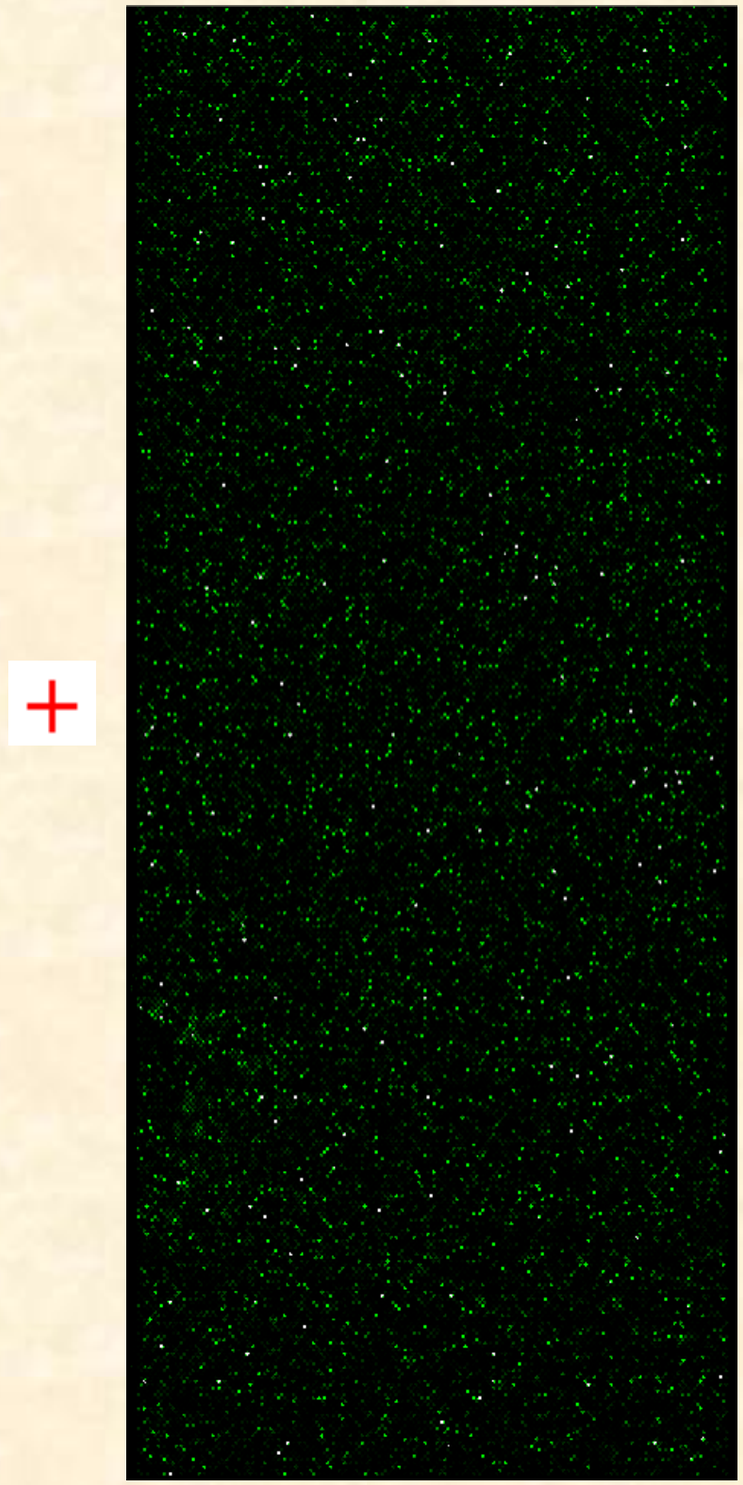


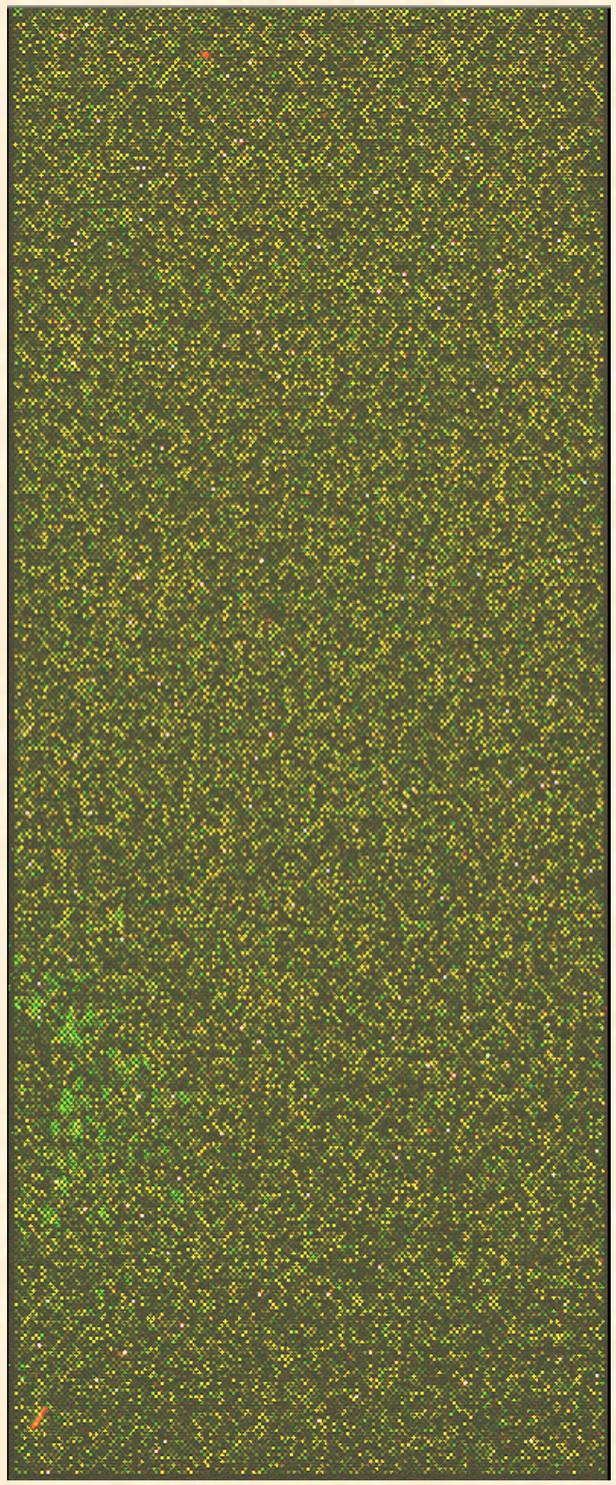
Generation of mRNA expression profiles: two different approaches

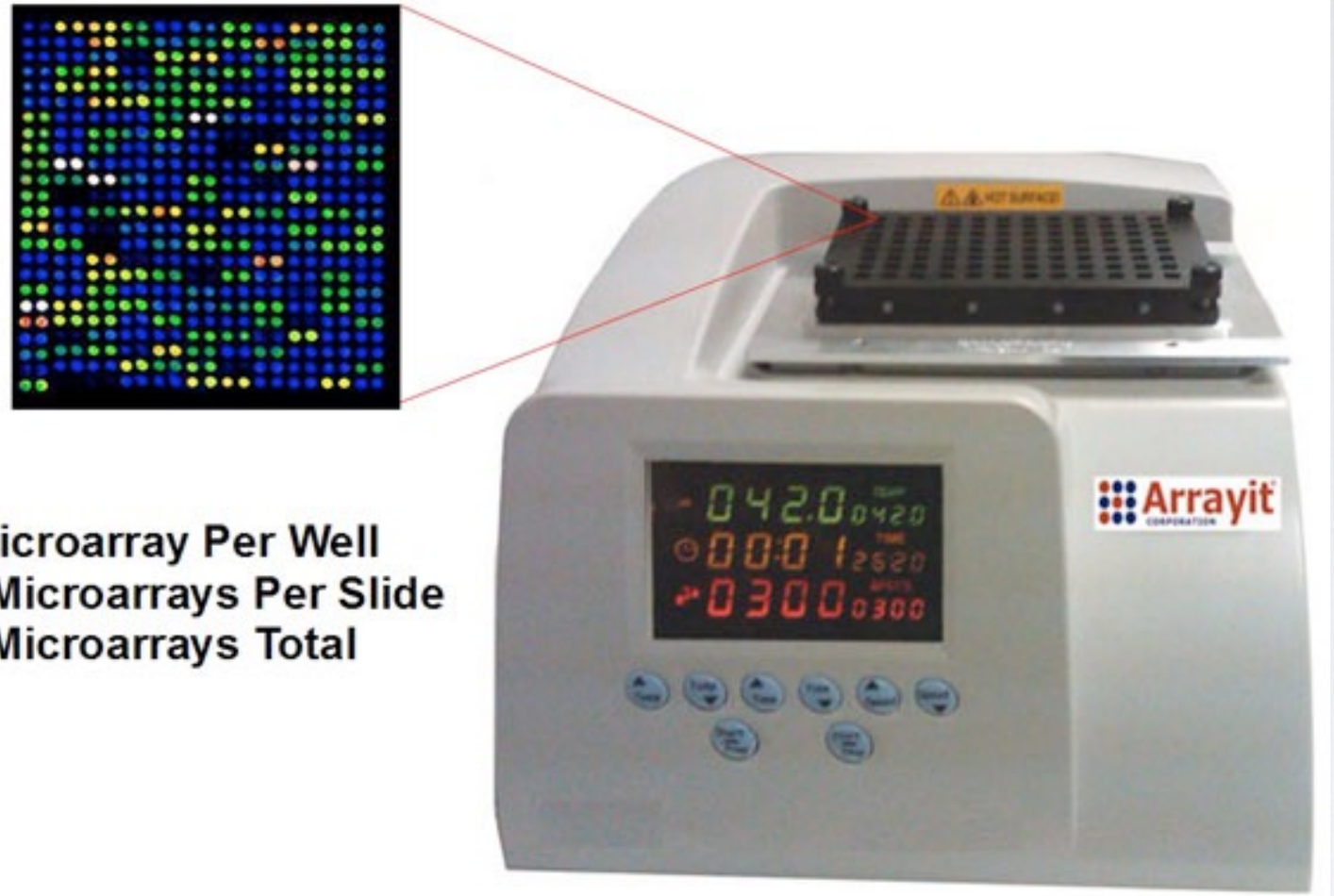


Glass laser scanning with two excitation wavelengths (635 and 532 nm) 16 bit image codification (65536 color levels)



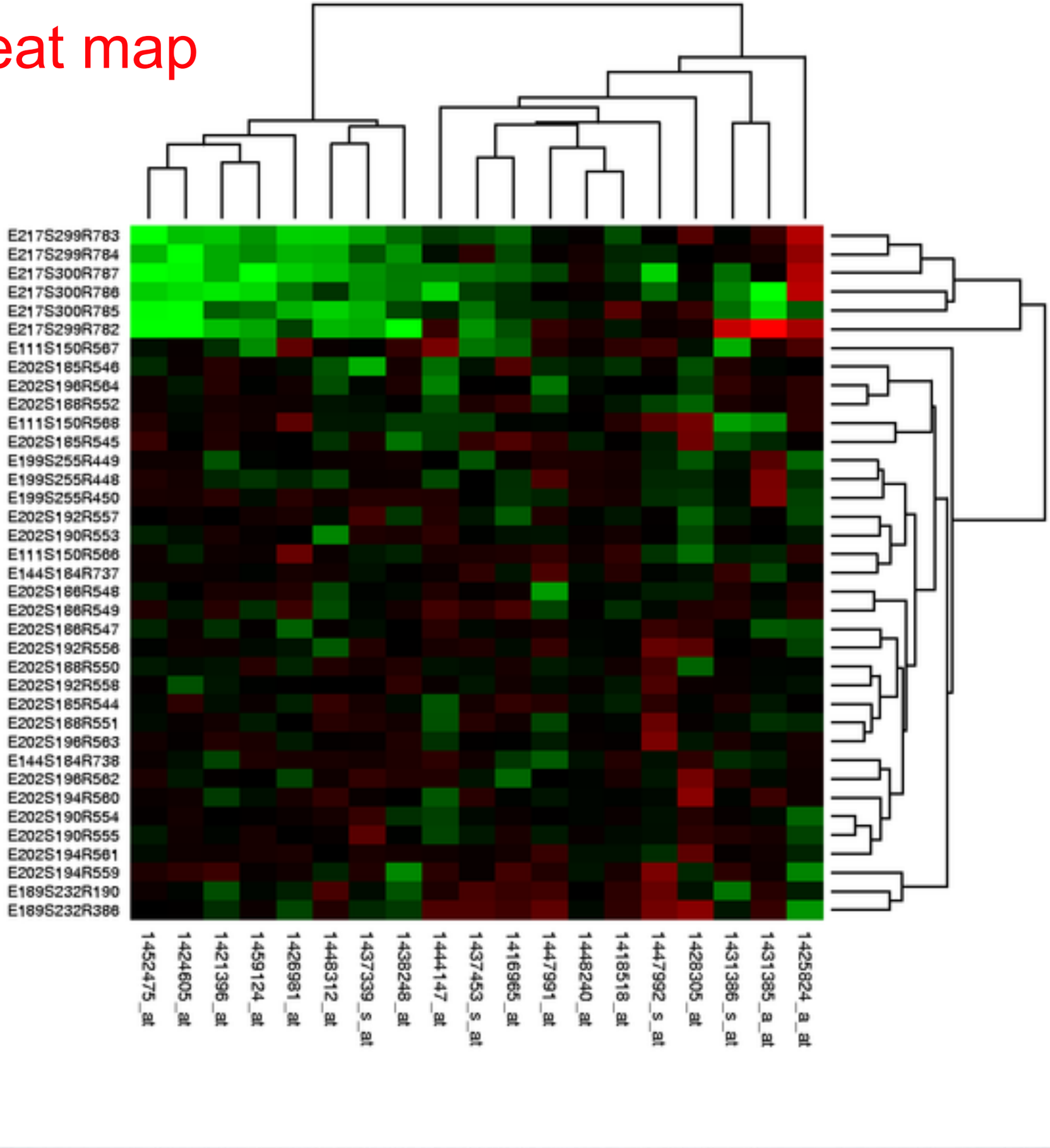


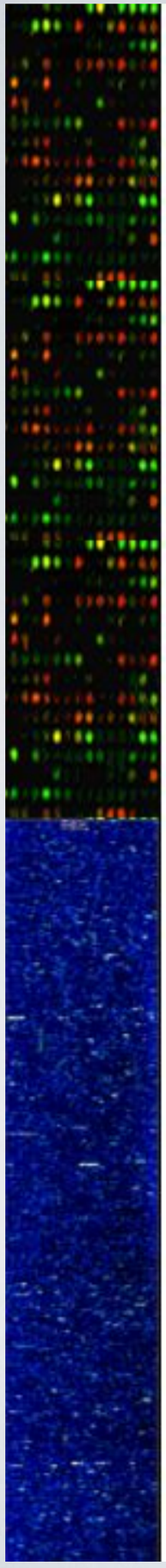




1 Microarray Per Well
24 Microarrays Per Slide
96 Microarrays Total

Heat map


E217S299R783
E217S299R784
E217S300R787
E217S300R786
E217S300R785
E217S299R782
E111S150R567
E202S185R546
E202S196R564
E202S188R552
E111S150R568
E202S185R545
E199S255R449
E199S255R448
E199S255R450
E202S192R557
E202S190R553
E111S150R566
E144S184R737
E202S186R548
E202S186R549
E202S186R547
E202S192R556
E202S188R550
E202S192R558
E202S185R544
E202S188R551
E202S196R563
E144S184R738
E202S196R562
E202S194R560
E202S190R554
E202S190R555
E202S194R561
E202S194R559
E189S232R190
E189S232R388
1425824_a_at
1431385_a_at
1431386_s_at
1428305_at
1447992_s_at
1418518_at
1448240_at
1447991_at
1416965_at
1437453_s_at
1444147_at
1438248_at
1437339_s_at
1448312_at
1426981_at
1459124_at
1421396_at
1424605_at
1452475_at



Practical Applications of Microarrays

Gene Target Discovery
By allowing scientists to compare diseased cells with normal cells, arrays can be used to discover sets of genes that play key roles in diseases. Genes that are either overexpressed or underexpressed in the diseased cells often present excellent targets for therapeutic drugs.

Pharmacology and Toxicology
Arrays can provide a highly sensitive indicator of a drug's activity (pharmacology) and toxicity (toxicology) in cell culture or test animals. This information can then be used to screen or optimize drug candidates prior to launching costly clinical trials.

Diagnostics
Array technology can be used to diagnose clinical conditions by detecting gene expression patterns associated with disease states in either biopsy samples or peripheral blood cells.

PROTEOMICS

•collezioni di sequenze di proteine di un organismo (proteoma) e loro analisi

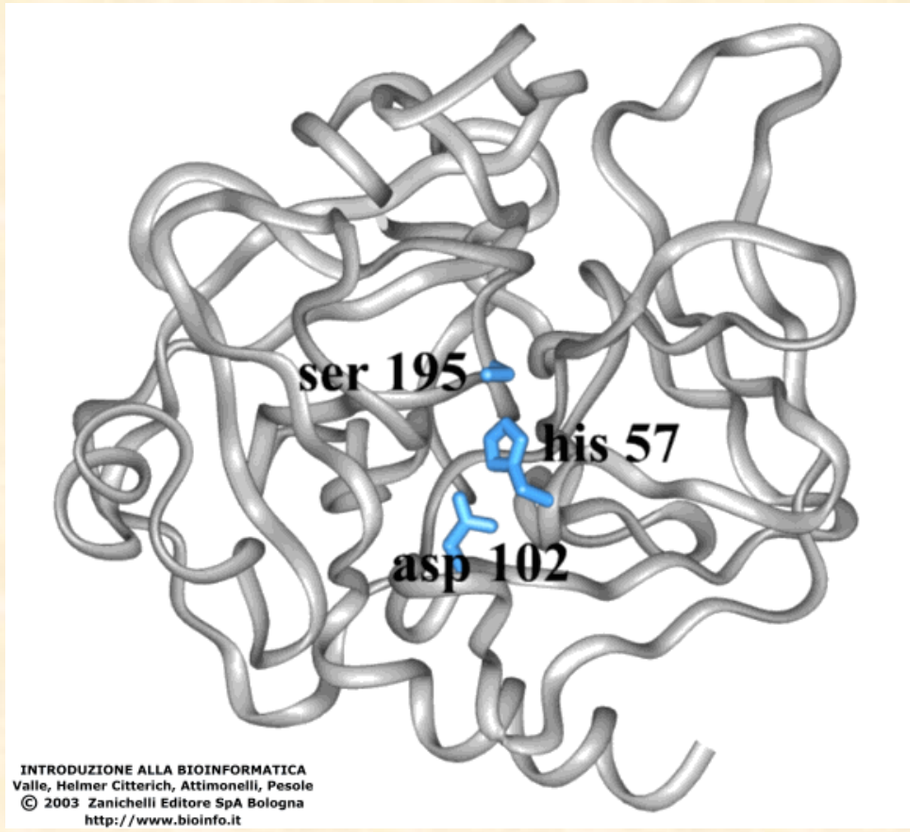
•determinazione della struttura 3D delle proteine (cristallografia e raggi X, NMR)

•predizione della struttura di proteine di cui sia nota solo la sequenza

PDB: database di strutture di proteine

http://www.rcsb.org/pdb/

RCSB PDB
PROTEIN DATA BANK



Proteomics

Proteomics is the large-scale study of proteins.

The term *proteomics* was coined in 1997 in analogy with genomics, the study of the genome.

The word *proteome* is a fusion term of *prote*in and gen*ome*, and was coined by Marc Wilkins in 1994 while working on the concept as a PhD student.

The proteome *is the entire set of proteins*, produced or modified by an organism or system. This ***varies*** with time and distinct requirements, or stresses, that a cell or organism undergoes.

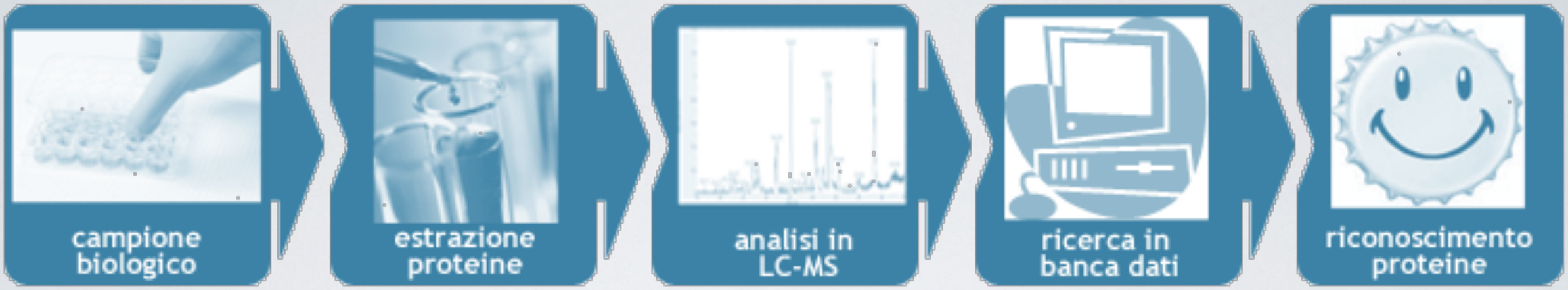
Proteomics is an interdisciplinary domain that has benefited greatly from the genetic information of the Human Genome Project.

It is mainly focused on the exploration of cell proteomes from the overall level of intracellular protein composition, structure, and its own unique activity patterns and functions.

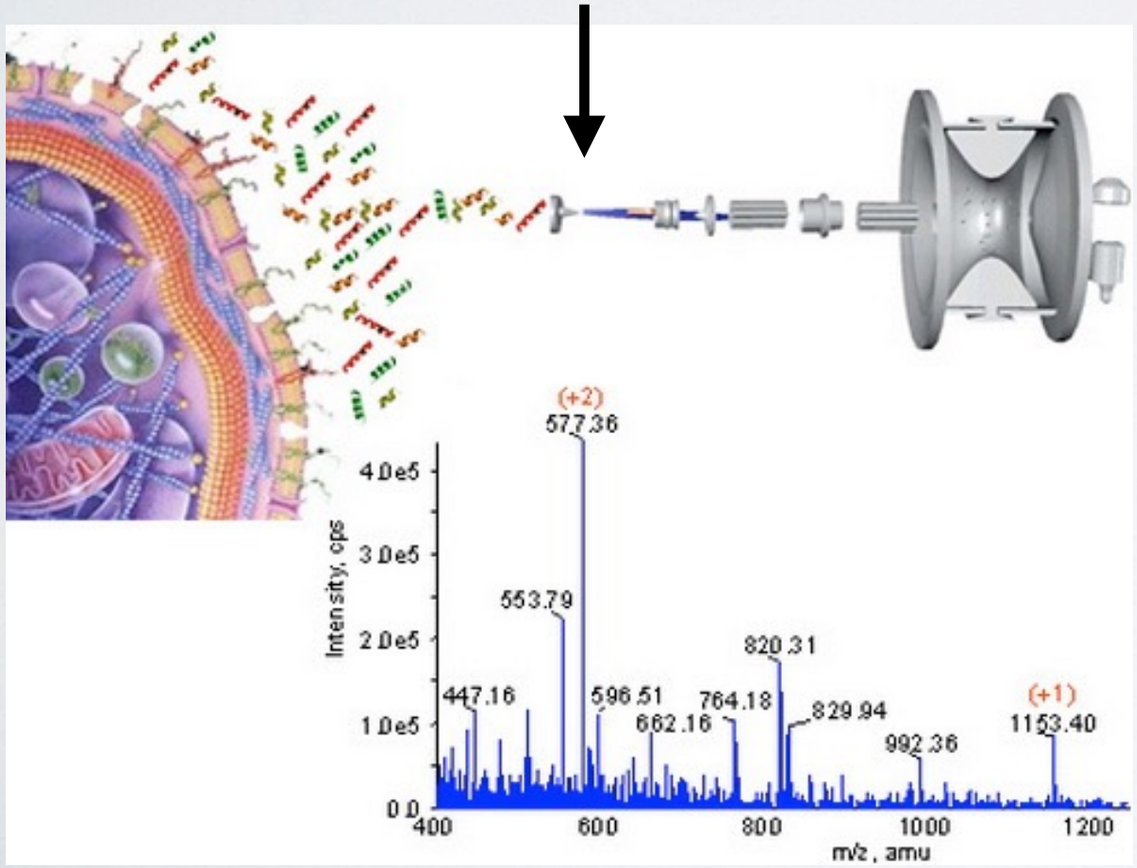
It is an important component of **functional genomics.**

While *proteomics* generally refers to the large-scale experimental analysis of proteins, the term it is often more specifically used for protein purification and mass spectrometry (Mass-Spec)

Proteomic analysis flow-chart

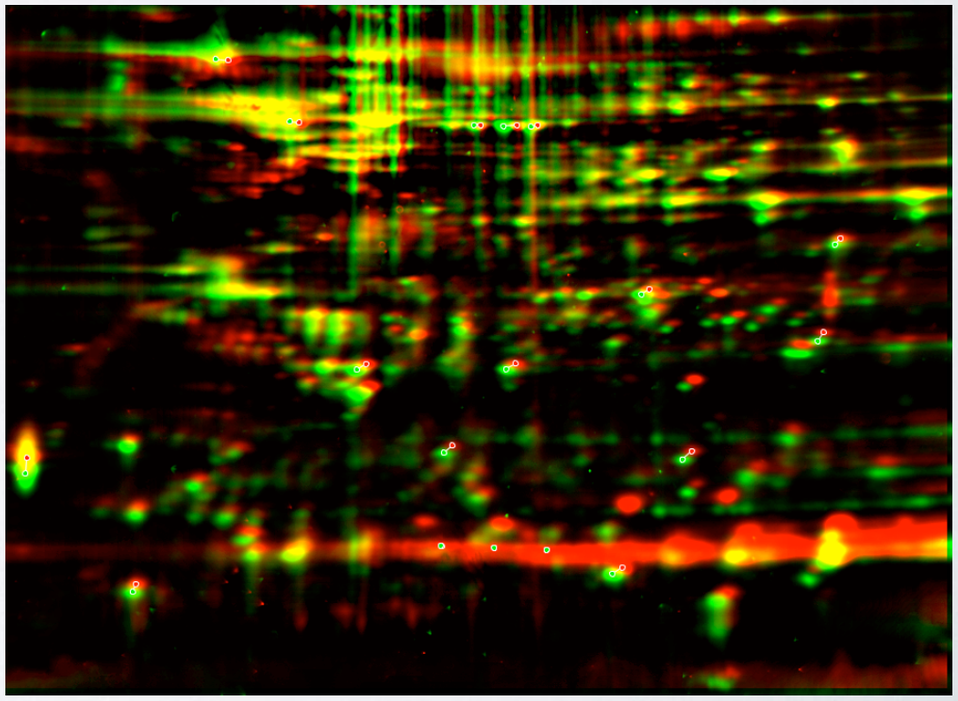


Peptide detection



2D - PAGE

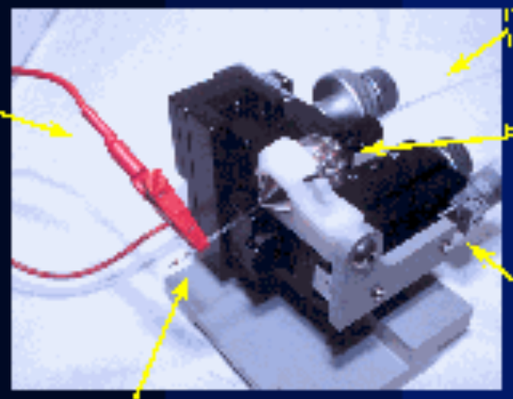
(iso-electric-focusing)



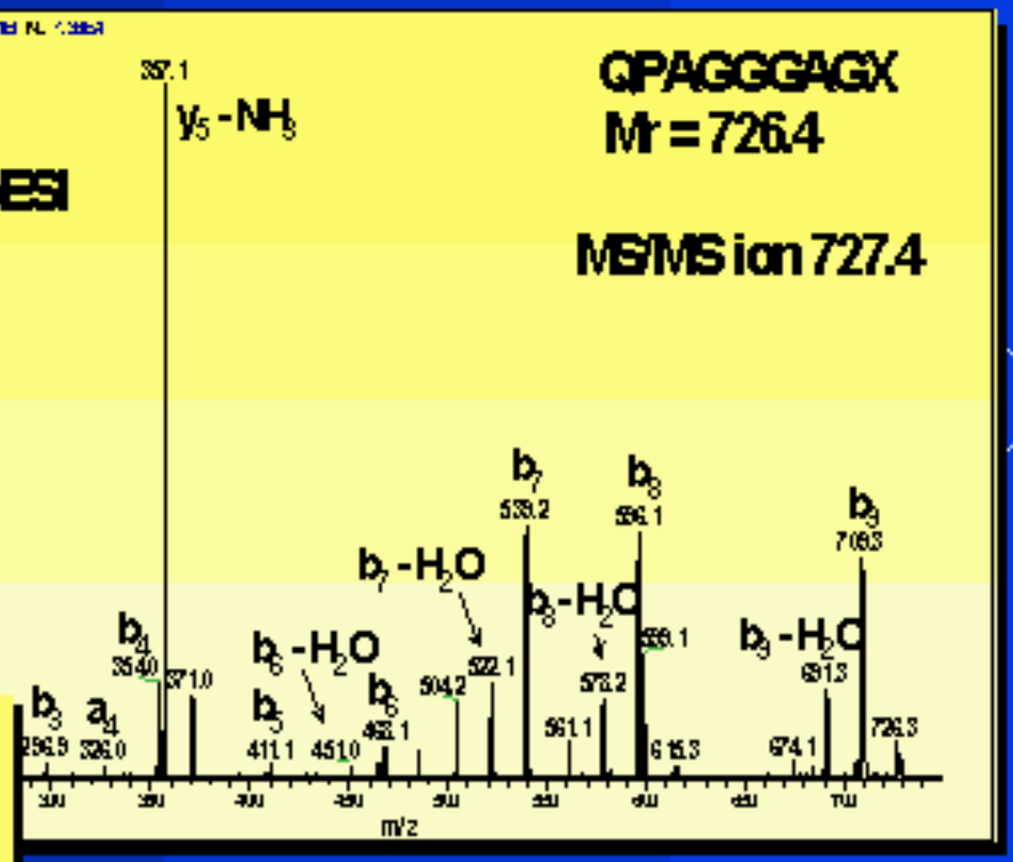
Mass-Spec

PROTEIN EMEB IIBB-IDIBAPS ANALYSIS

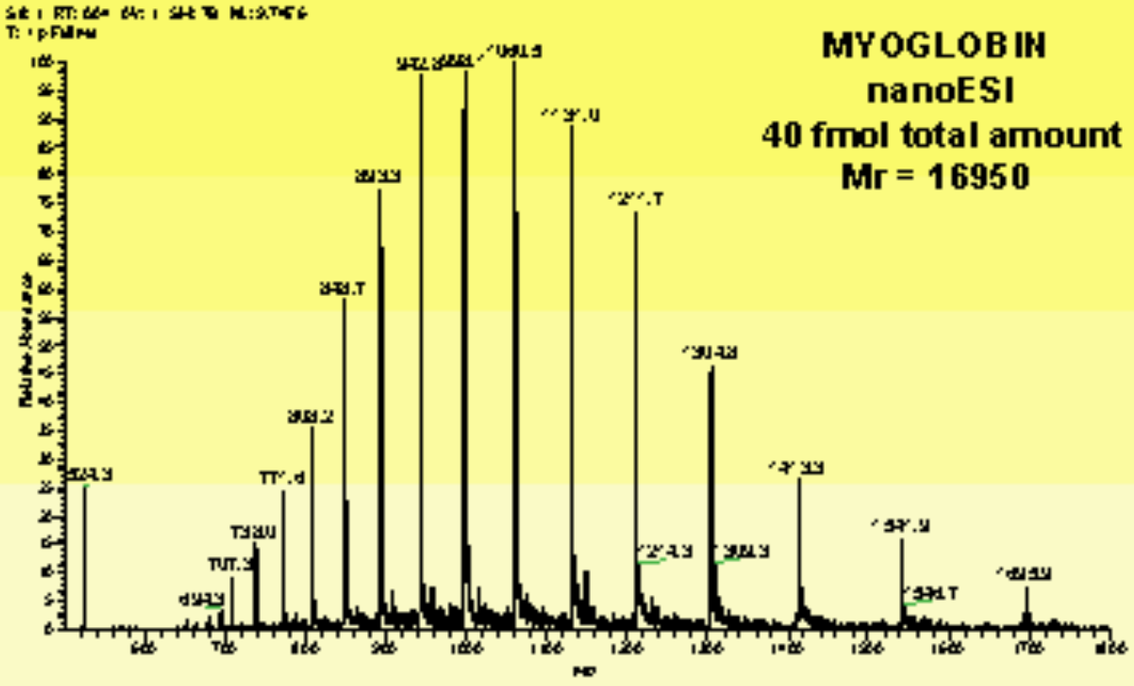
nanoESI SOURCE



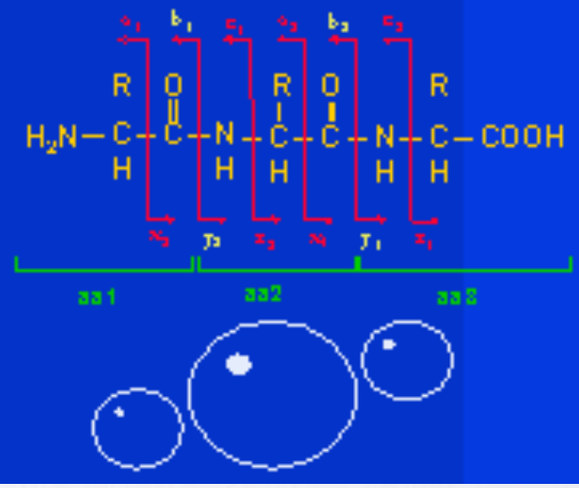
nanoESI



MYOGLOBIN
nanoESI
40 fmol total amount
Mr = 16950

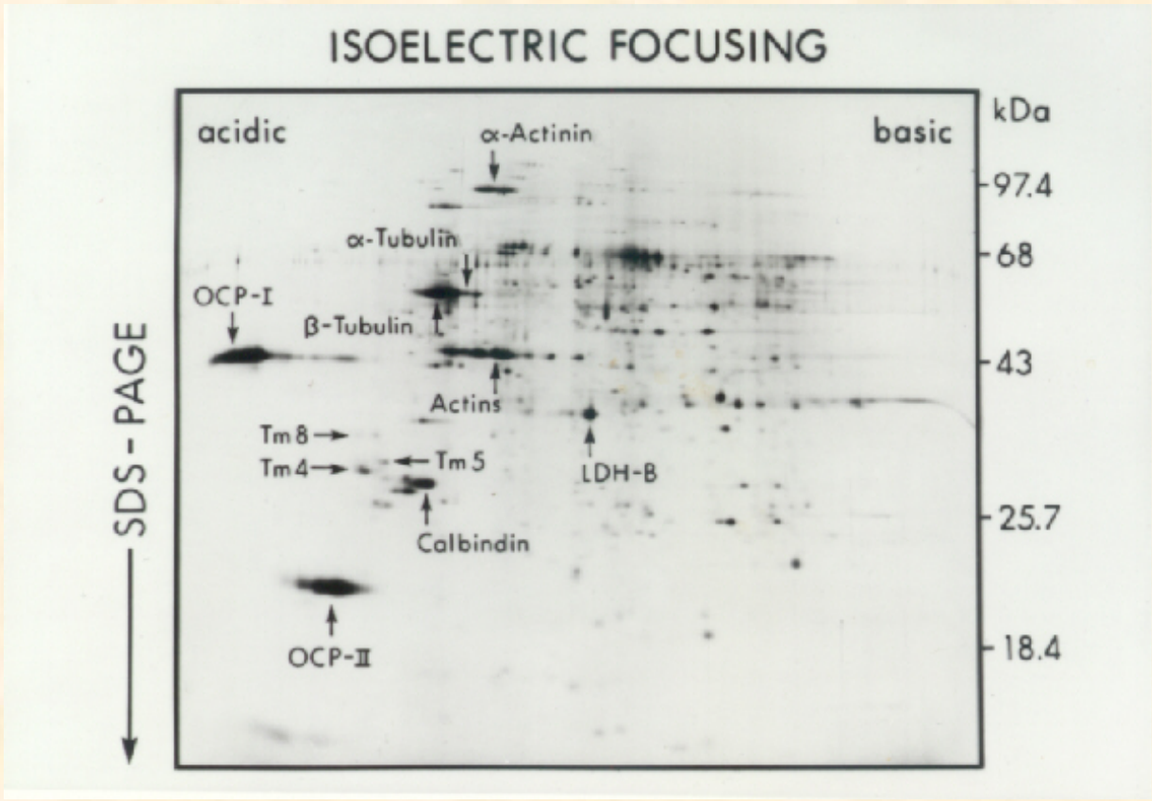


MS/MS PEPTIDE FRAGMENTATION



•separazione di proteine in base a caratteristiche chimico-fisiche (massa, PH)

http://www.lecb.ncifcrf.gov/flicker/



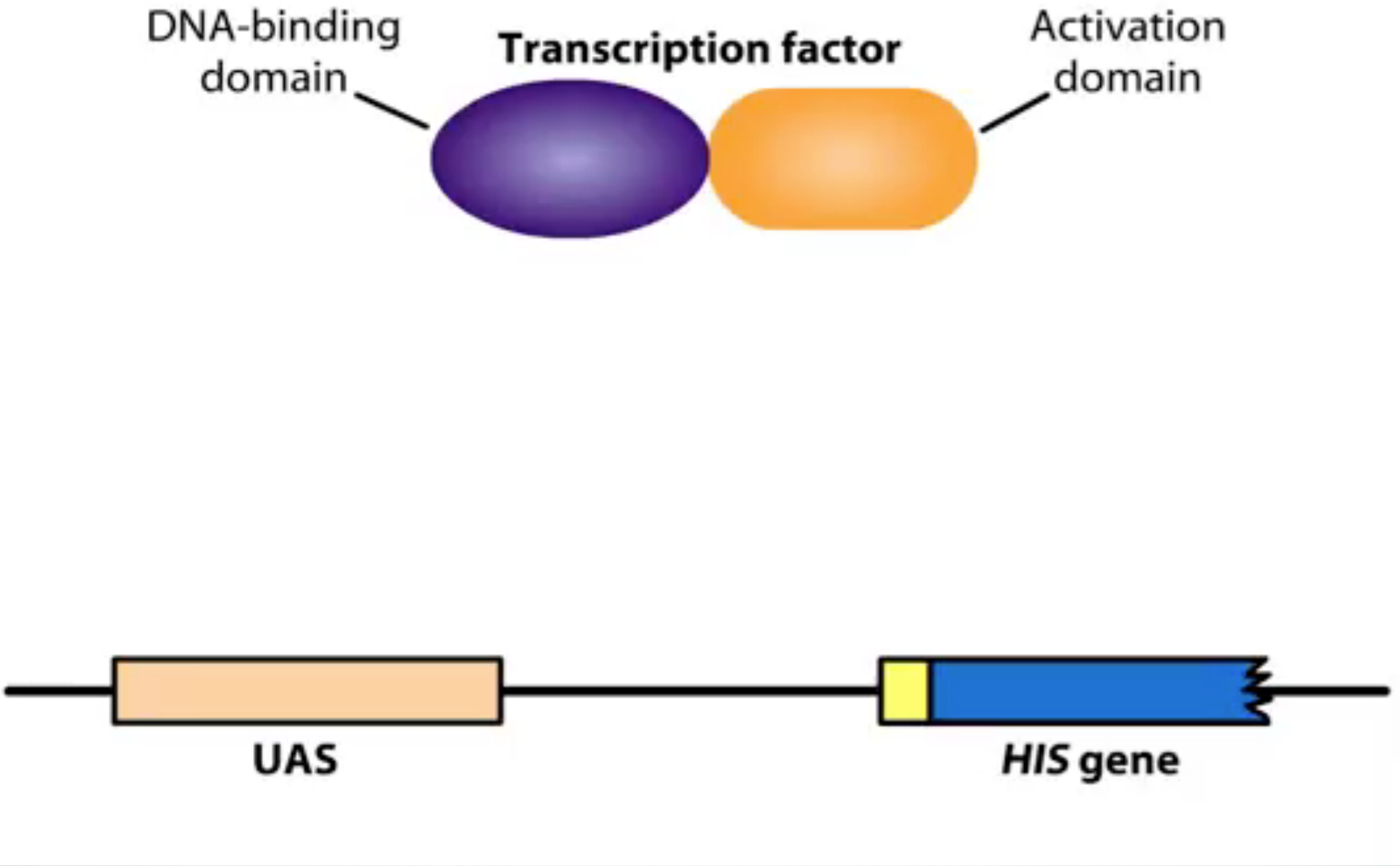
•interazioni proteina-proteina, proteina-acidi nucleici, proteina-metaboliti

Detection of Protein-Protein Interaction by

Yeast two-hybrid system

The overall idea is to be able to identify PPIs in a permissive, enzymatically defective, eukaryotic cell, such as YEAST, and capable of surviving and duplicating in selective medium only if a transcription initiation complex for a specific enzyme is reconstituted upon interaction between a bait protein domain and fish protein domain. Yeast cells are transfected with the **bait** protein of interest, for which interactors have to be discovered, and with a library of many **fish** proteins, all potential interactors for the bait. When the PPI complex is generated, the transcription initiation complex trigger the transcription of a gene leading to expression of enzyme allowing survival in media lacking of histidine.

Thus, only when the PPI interaction is established, cells can survive and generate clones. Clones are, then, plasmid sequenced to identify the fish. The identified fish is a ligand for the bait of interest.


DNA-binding domain
Transcription factor
Activation domain
UAS
HIS gene

Dimension of human protein interactome (PPI)

SEE COMMENTARY

EVOLUTION

APPLIED MATHEMATICS

Estimating the size of the human interactome

Michael P. H. Stumpf[†‡§], Thomas Thorne[†], Eric de Silva[†], Ronald Stewart[†], Hyeong Jun An[¶], Michael Lappe[¶], and Carsten Wiuf[§‖]

[†]Division of Molecular Biosciences, Imperial College London, Wolfson Building, London SW7 2AZ, United Kingdom; [‡]Institute of Mathematical Sciences, Imperial College London, London SW7 2AZ, United Kingdom; [‖]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; and [¶]Bioinformatics Research Center, University of Aarhus, 8000 Aarhus C, Denmark

Edited by Burton H. Singer, Princeton University, Princeton, NJ, and approved February 19, 2008 (received for review August 27, 2007)

After the completion of the human and other genome projects it emerged that the number of genes in organisms as diverse as fruit flies, nematodes, and humans does not reflect our perception of their relative complexity. Here, we provide reliable evidence that the size of protein interaction networks in different organisms appears to correlate much better with their apparent biological complexity. We develop a stable and powerful, yet simple, statistical procedure to estimate the size of the whole network from subnet data. This approach is then applied to a range of eukaryotic organisms for which extensive protein interaction data have been collected and we estimate the number of interactions in humans to be ≈650,000. We find that the human interaction network is one order of magnitude bigger than the *Drosophila melanogaster* interactome and ≈3 times bigger than in *Caenorhabditis elegans*.

evolutionary systems biology | network inference | network sampling theory | network evolution

One of the perhaps most surprising results of the genome-sequencing projects was that the number of genes is much lower than had been expected and is, in fact, surprisingly similar for very different organisms (1, 2). For example, the nematode *Caenorhabditis elegans* appears to have a similar number of genes as humans, whereas rice and maize appear to have even more genes than humans. It was then quickly suggested that the biological complexity of organisms is not reflected merely by the number of genes but by the number of physiologically relevant interactions (1, 3). In addition to alternative splice variants (4), posttranslational processes (5), and other (e.g., genetic) factors influencing gene expression (6, 7), the structure of interactome is one of the crucial factors underlying the complexity of biological organisms. Here, we focus on the wealth of available protein interaction data and demonstrate that it is possible to arrive at a reliable statistical estimate for the size of these interaction networks. This approach is then used to assess the complexity of protein interaction networks in different organisms from present incomplete and noisy protein interaction datasets.

There are now fairly extensive protein interaction network (PIN) datasets in a number of species, including humans (8, 9). These have been generated by a variety of experimental techniques (as well as some *in silico* inferences). Although these techniques and the resulting data are (*i*) notoriously prone to false positives and negatives (10, 11), and (*ii*) result in highly idealized and averaged network structures (12), such interaction datasets are increasingly turning into useful tools for the analysis of the functional (e.g., ref. 13) and evolutionary properties (14) of biological systems. In particular, in *Saccharomyces cerevisiae* we are beginning to have a fairly complete description of the protein interaction network that is accessible with current experimental technologies; the recent high-quality literature-curated dataset of Reguly *et al.* (15) provides us with a dataset that should be almost completely free from false positives. For most other organisms, however, interaction data are still far from complete and it has recently been shown that subnetworks, in general, have qualitatively different properties from the true network (16–18). Although the importance of network-sampling properties had only been realized relatively recently, this aspect of most systems biology data are increasingly being recognized (11, 19) as important.

There are, however, some properties of the true network that can be inferred even from subnet data, and here we show that the total network size is one property for which this is the case. Present protein-interaction datasets enable us to estimate the size of the interactomes in different species by using graph theoretical invariants. This is particularly interesting for species where more than one experimental dataset is available. Below we first describe a robust and very general estimator of network size from partial network data that overcomes this problem. We then apply it to available PIN data in a range of eukaryotic organisms. In supporting information (SI) *Text* we demonstrate the power of this approach by using extensive simulation studies.

Estimating Interactome Size

Here, we develop an approach for estimating the size of a network from incomplete data. We will show below (and by using extensive simulations in *SI Text*) that for a given species estimates from different independent datasets—generated by different methods such as yeast-two-hybrid and TAP tagging—yield estimates for the interactome size that are in excellent agreement.

We are concerned with a true network, $\mathcal{N}$, which has $N_{\mathcal{N}}$ nodes and $M_{\mathcal{N}}$ edges. The sets of nodes and edges are given by $\mathcal{V}_{\mathcal{N}}$ and $\mathcal{E}_{\mathcal{N}}$, respectively; these define the graph representation of the true network:

$$G_{\mathcal{N}} = (\mathcal{V}_{\mathcal{N}}, \mathcal{E}_{\mathcal{N}}). \quad [1]$$

We pick a subset of nodes $\mathcal{V}_{\mathcal{S}} \subseteq \mathcal{V}_{\mathcal{N}}$ and study properties of the subgraph $G_{\mathcal{S}}$ induced by the nodes in $\mathcal{V}_{\mathcal{S}}$

$$G_{\mathcal{S}} = (\mathcal{V}_{\mathcal{S}}, \mathcal{E}_{\mathcal{S}}), \quad [2]$$

where the set of edges observed in the $\mathcal{S}$ is a subset of the total set of edges, $\mathcal{E}_{\mathcal{S}} \subseteq \mathcal{E}_{\mathcal{N}}$. Our aim is to predict the number of interactions in the true network $G_{\mathcal{N}}$ based on the available data in the subnet, $G_{\mathcal{S}}$.

We assume that the network, $G_{\mathcal{N}}$, is generated according to some (unknown) model characterized by a parameter (vector) θ, and subsequently the observed network, $G_{\mathcal{S}}$ is sampled from it. Then

$$P_{\theta,p}(G_{\mathcal{S}}) = \sum_{G_{\mathcal{N}} \supseteq G_{\mathcal{S}}} P_p(G_{\mathcal{S}}|G_{\mathcal{N}}) P_\theta(G_{\mathcal{N}}), \quad [3]$$

Author contributions: M.P.H.S., M.L., and C.W. designed research; M.P.H.S., T.T., E.d.S., M.L., and C.W. performed research; M.P.H.S., T.T., E.d.S., R.S., H.J.A., and C.W. analyzed data; and M.P.H.S., M.L., and C.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

See Commentary on page 6795.

[§]To whom correspondence may be addressed. E-mail: m.stumpf@imperial.ac.uk or wiuf@birc.au.dk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708078105/DCSupplemental.

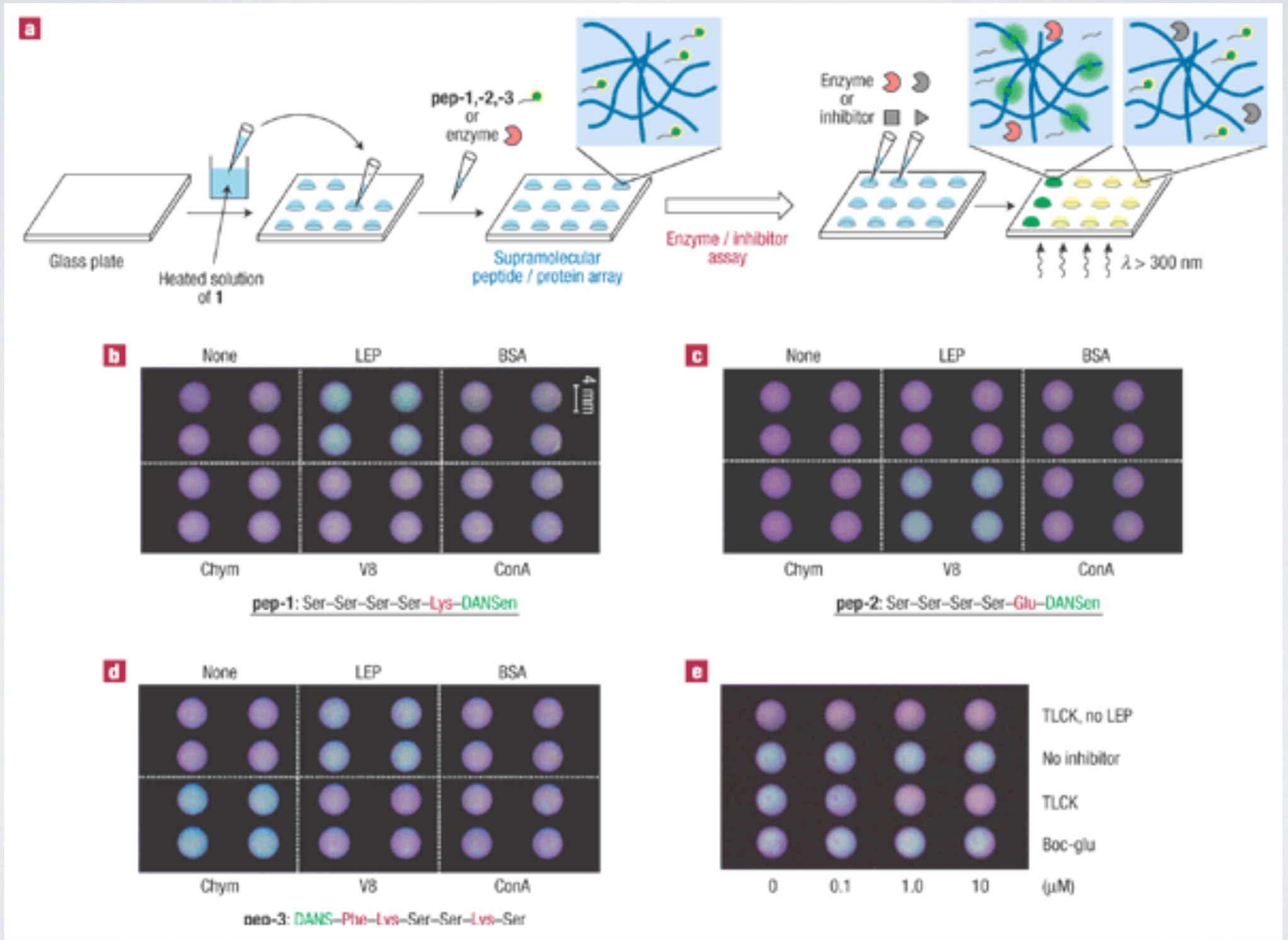
~ 24000 proteine
(da ORFs)

~ 650000
interazioni binarie

PPI DATA BASES

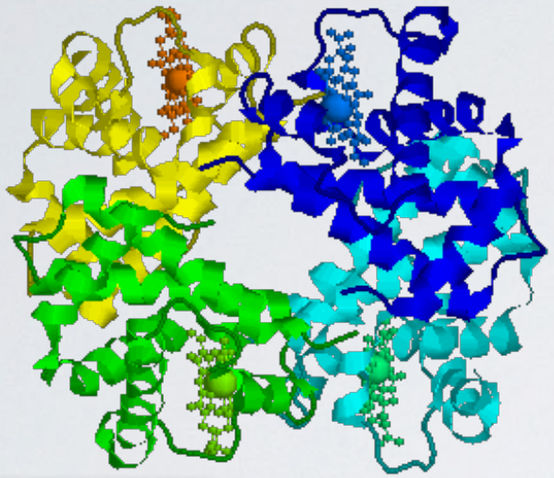
InBio Map,
BioGRID,
Pathway commons_hs,
MiMI,
BCI,
DIP,
HPRD,
HumanNet,
IntAct,
MINT,
PiNA,
HiNT,
HuRI, P
athPPI,
UniHI,
ConsensusPathDB,
OmniPath,
PSICQUIC,
SignaLink 2.0,
STRING,
STITCH,
HIPPIE,
HAPPI,
MENTHA,
SIGNOR

Protein arrays

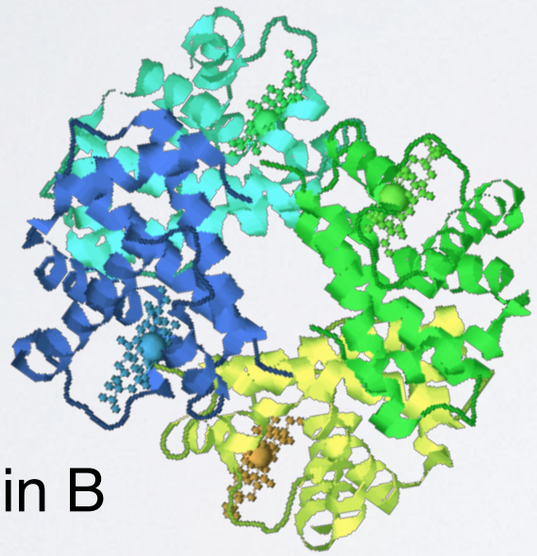


Protein domains array and drug screening

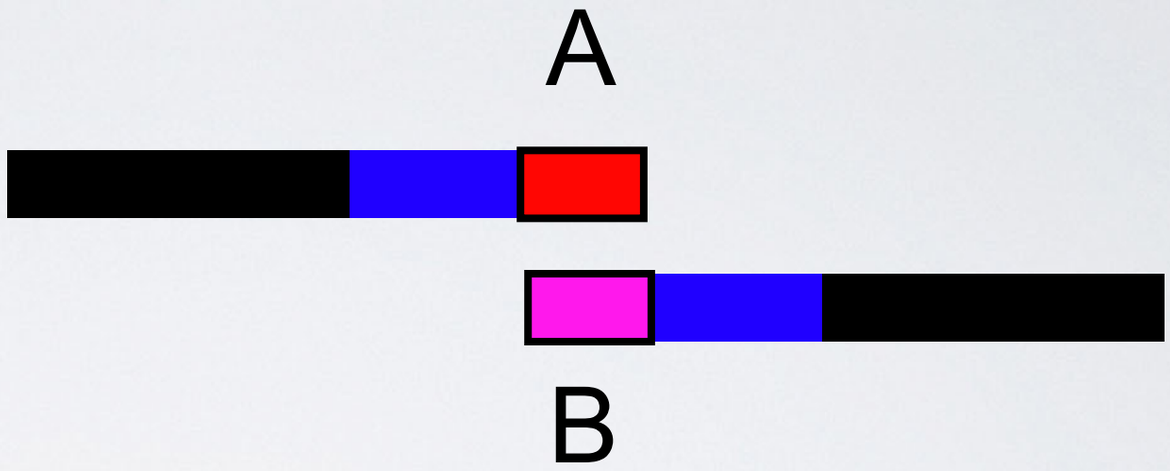
Protein A

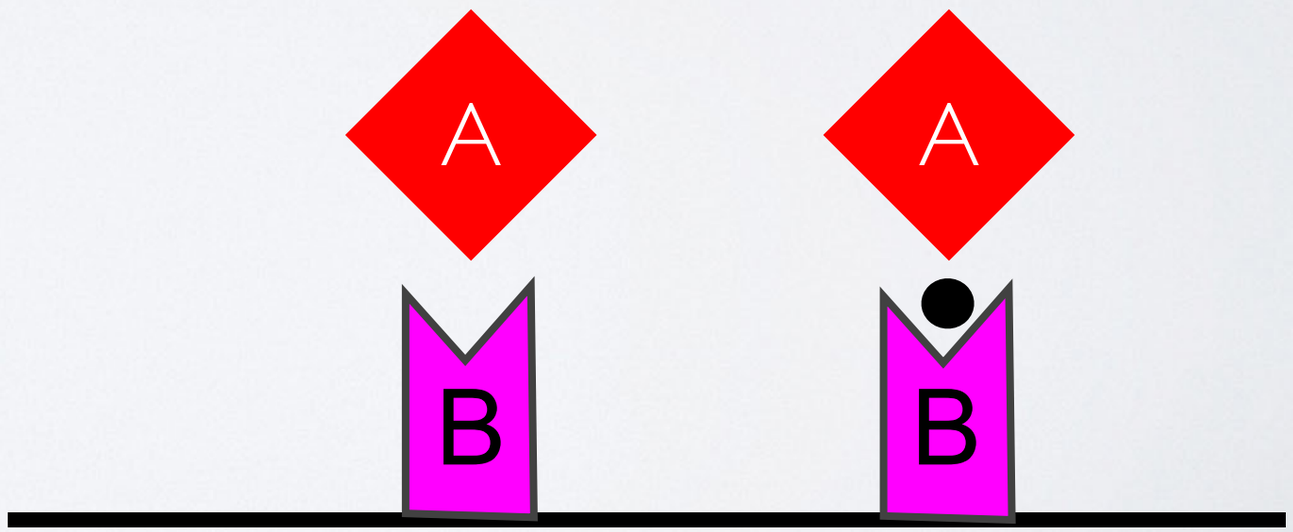


Domain-mediated binary interaction



Protein B

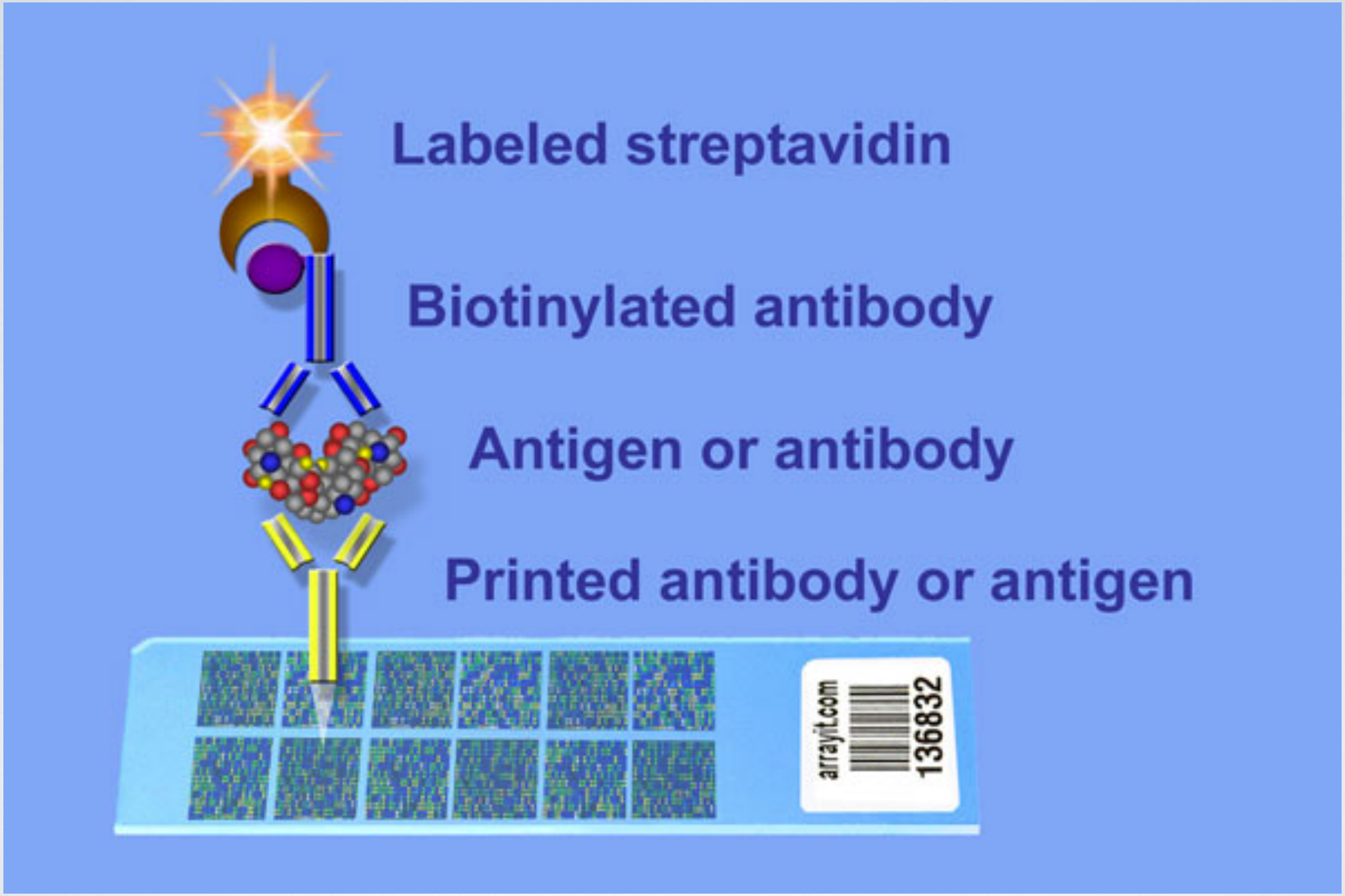




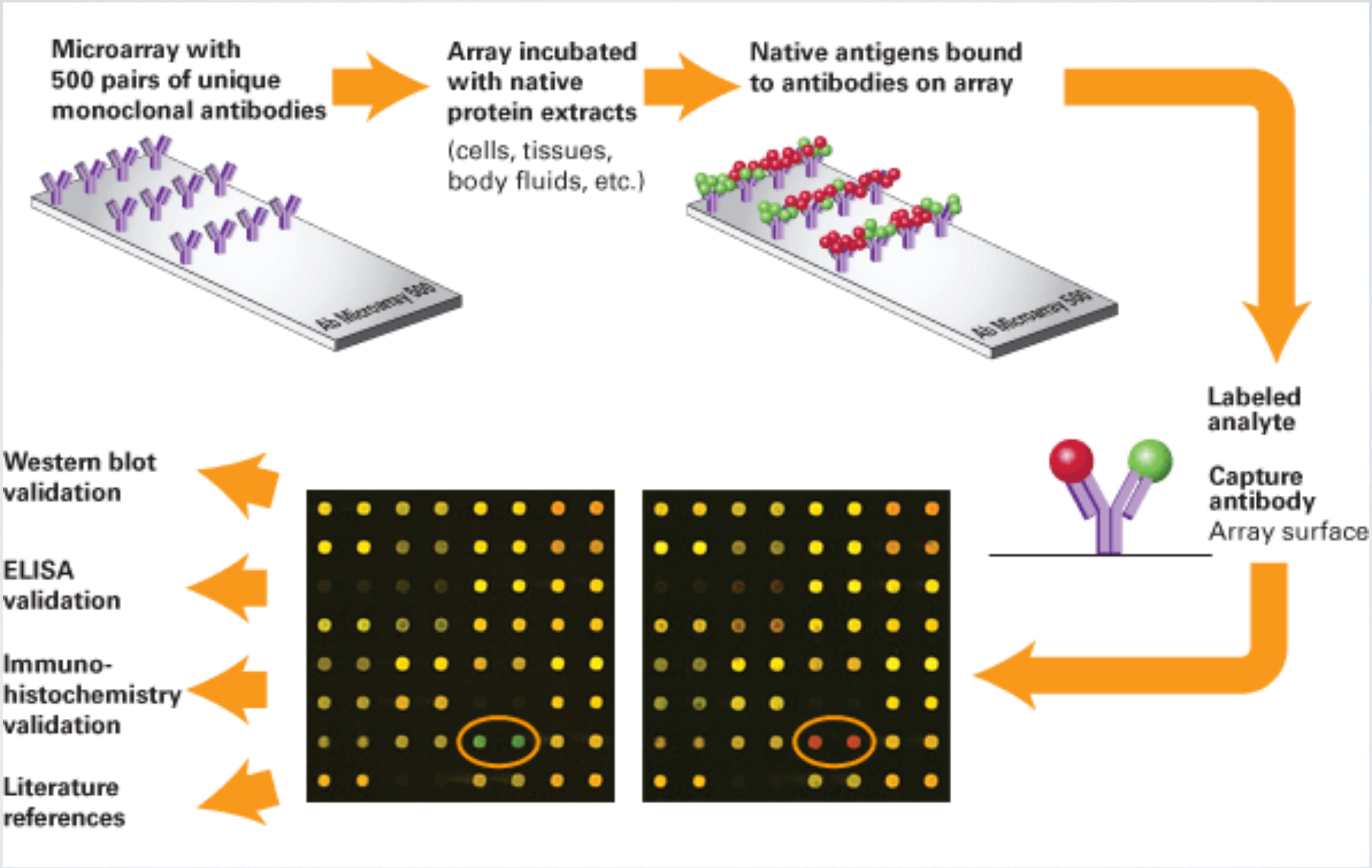
Protein (domain) array

Compound screening by means of inhibition of domain-domain interaction

Antibody array



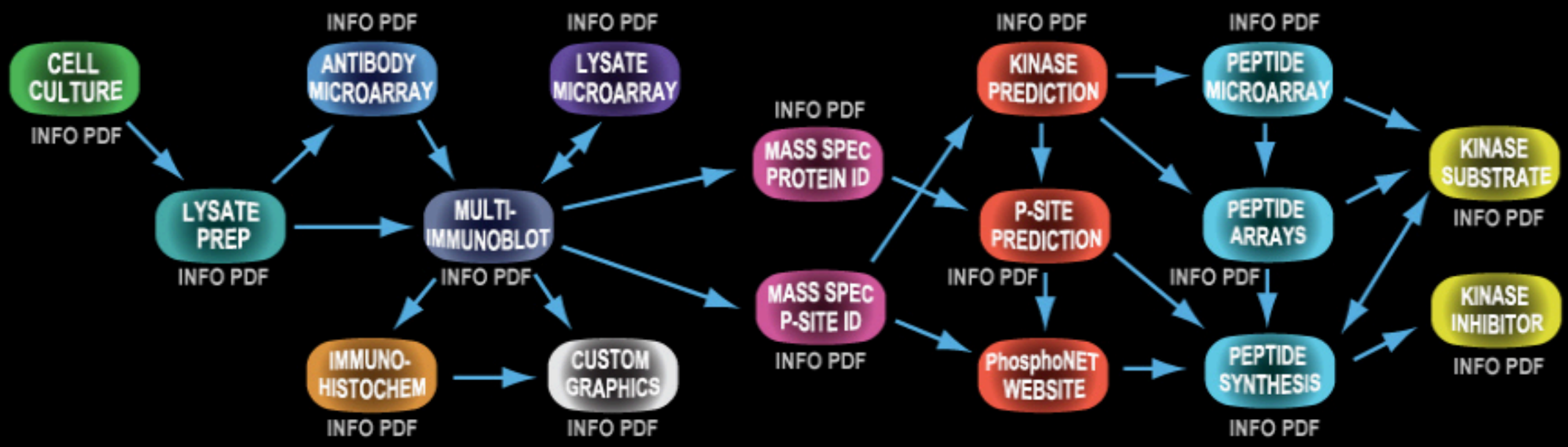
Antibody array



This page requires Adobe Flash Player to display correctly.

This ticker requires Adobe Flash Player to display correctly.

Kinexus Services Map



Click on button to open a webpage with more information

Click INFO PDF to download a Customer Info Package

Metabolomics

Metabolomics is the systematic study of the unique chemical fingerprints left behind by specific cellular processes - specifically, the study of their metabolic small molecules profiles, as derived from metabolic pathways, either processing energy as well as mass.

The metabolome is the set of all metabolites in a biological organism **principal =** ~ 8000**; total** (including minor intermediates) = ~ 10^6.

Note that, while gene expression and proteomics data do not fully explain what happens in a cell, the metabolic profile may provide a snapshot of the ongoing physiology of the cell.

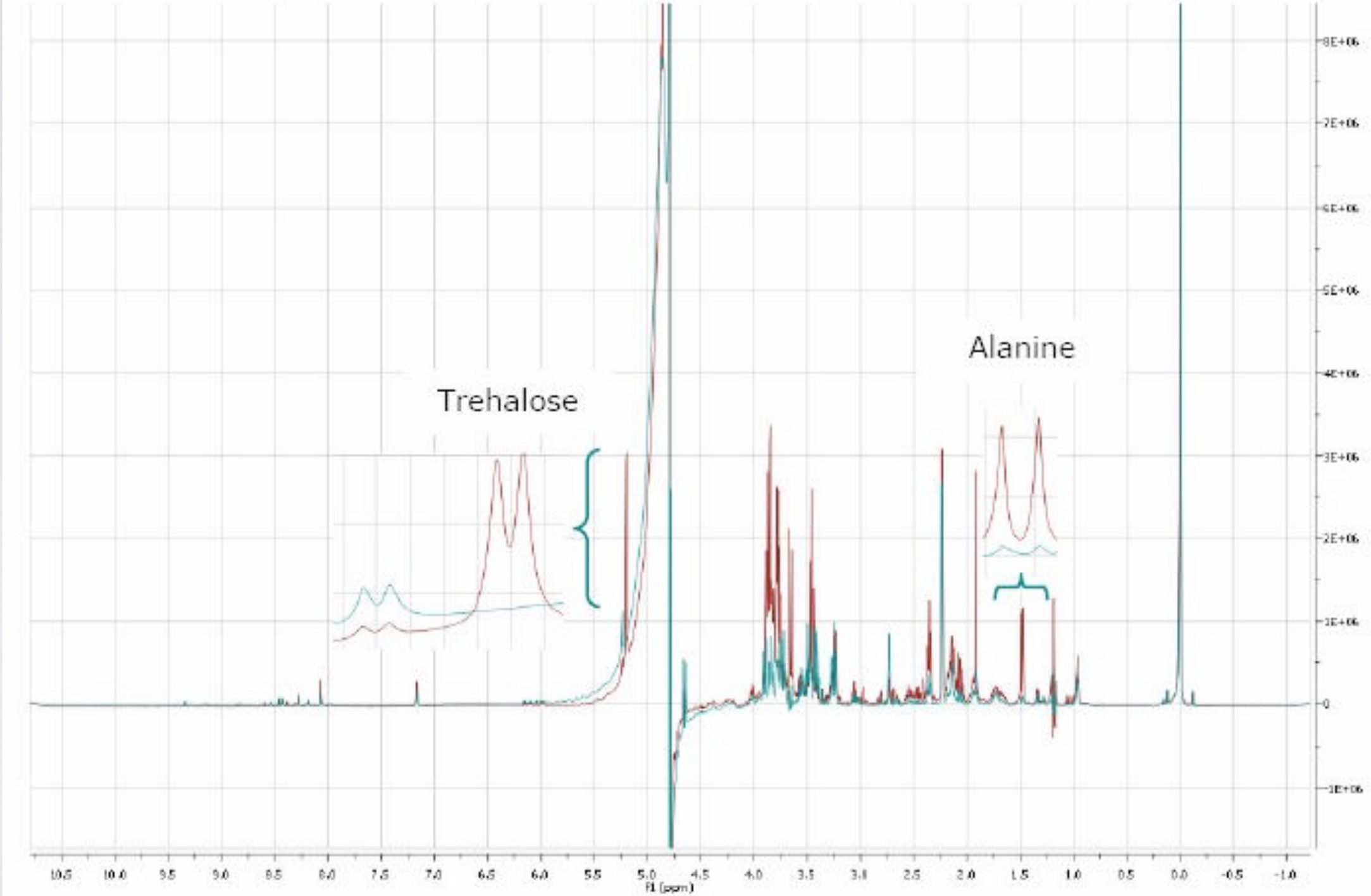
Notably, metabolic and signal transduction pathways are now considered strongly connected.

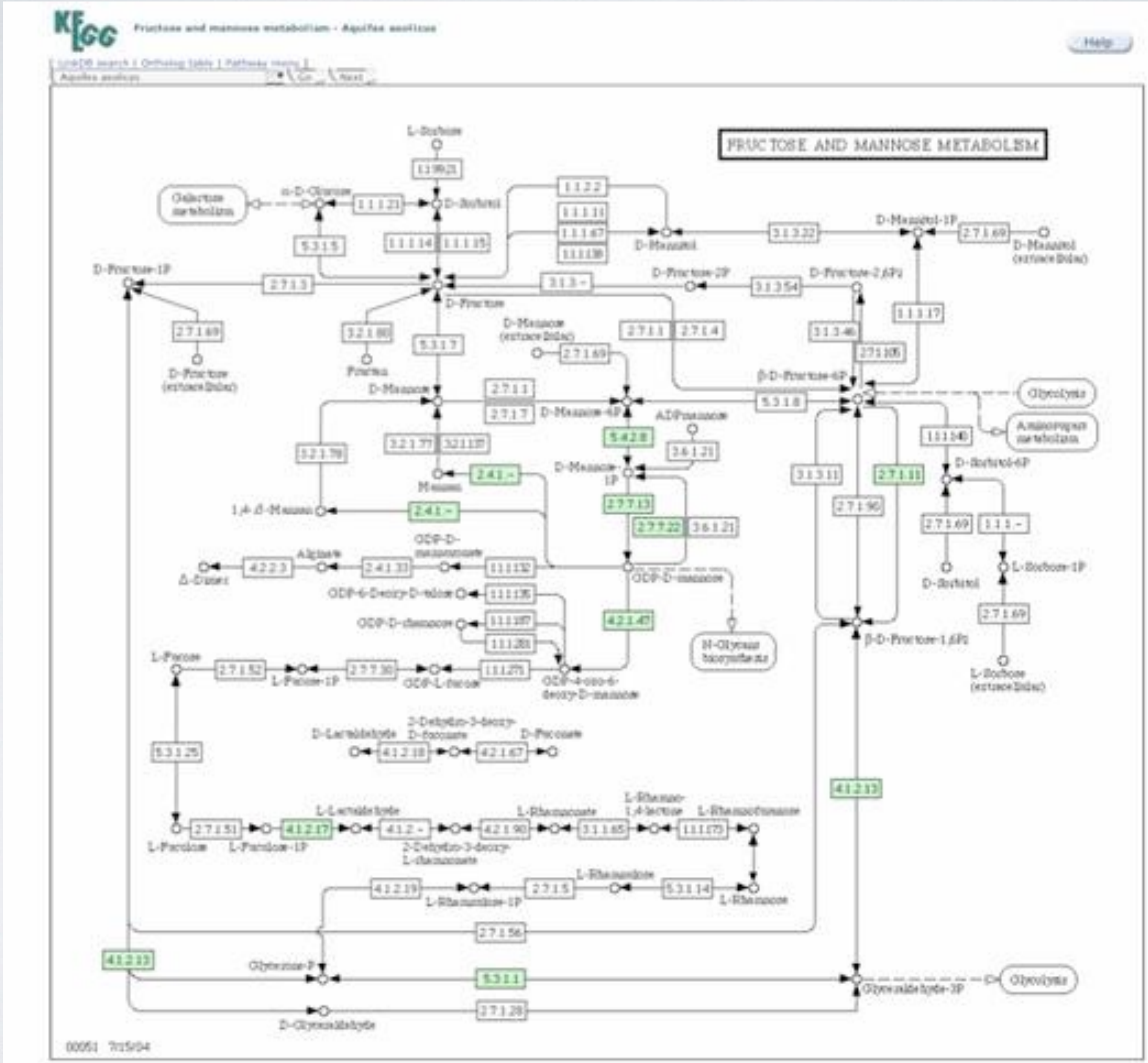
One of the challenges of systems biology is to integrate genomics, proteomics and metabolomics to have a complete overview of living organisms.

Metabolomics analysis

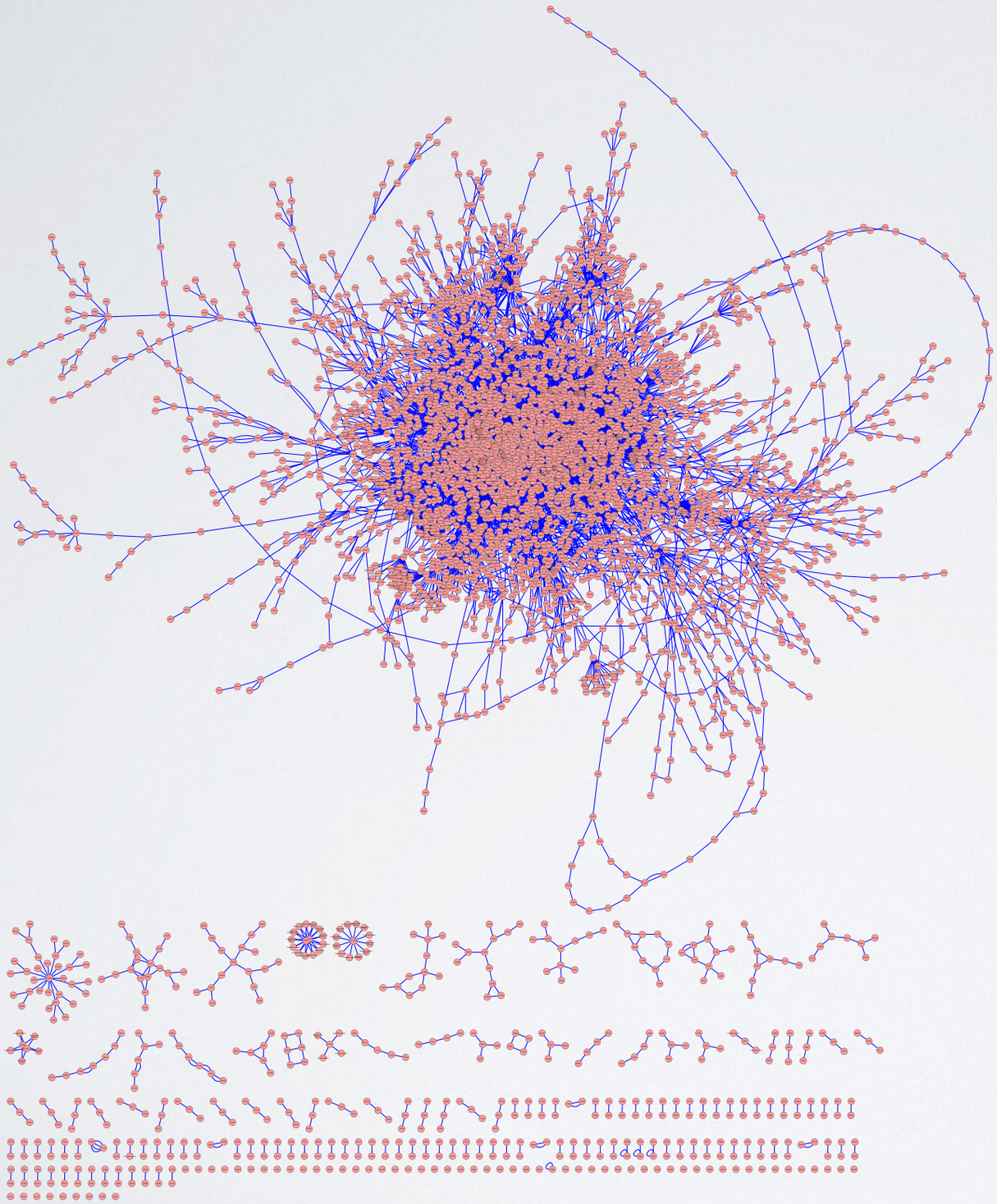
Capillary electrophoresis (CE)
Gas-chromatography (GC)
High Performance liquid chromatography (HPLC)
Mass spectroscopy (Mass Spec)

NMR and metabolomics


Trehalose
Alanine
8E+06
7E+06
6E+06
5E+06
4E+06
3E+06
2E+06
1E+06
0
-1E+06
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
f1 (ppm)


Fructose and mannose metabolism - Aquifex aeolicus
Help
FRUCTOSE AND MANNOSE METABOLISM
L-Sorbose
D-Sorbitol
Galactose metabolism
D-Fructose-1P
D-Fructose
D-Mannitol
D-Mannitol-1P
D-Fructose-2P
D-Fructose-2,6P2
D-Fructose (extracellular)
D-Mannose (extracellular)
D-Mannitol (extracellular)
D-Mannose
D-Mannose-6P
β-D-Fructose-6P
Glycolysis
Aminosugars metabolism
D-Sorbitol-6P
D-Mannose-1P
Alginate
GDP-D-mannose
L-Sorbose-1P
β-D-Fructose-1,6P2
N-Glycans biosynthesis
L-Fucose
L-Fucose-1P
GDP-L-fucose
L-Sorbose (extracellular)
D-Lactaldehyde
D-Fuconate
L-Rhamnonate
L-Rhamnose
L-Rhamnulose
L-Rhamnulose-1P
L-Fuculose
L-Fuculose-1P
L-Lactaldehyde
Glycerone-P
Glyceraldehyde-3P
D-Glyceraldehyde
00051 7/15/04

Size of human metabolic network



~ 8000 main metabolites
x
~ 22000 interactions

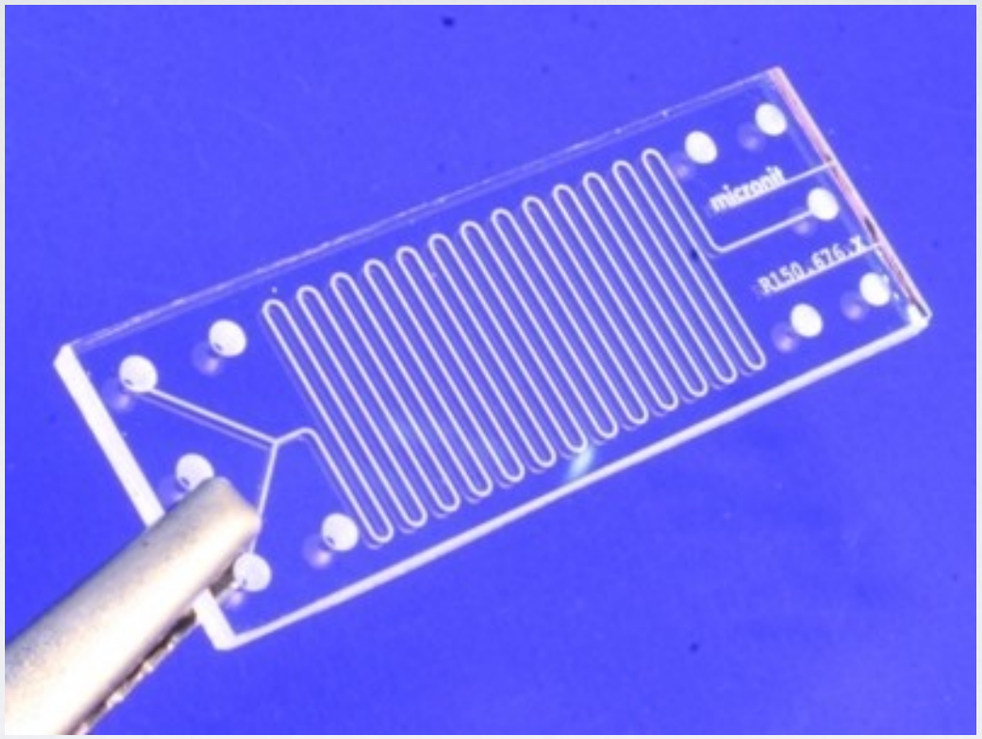
BUT ~ 10^6 total metabolites
x
Interazioni ?

METABOLON®

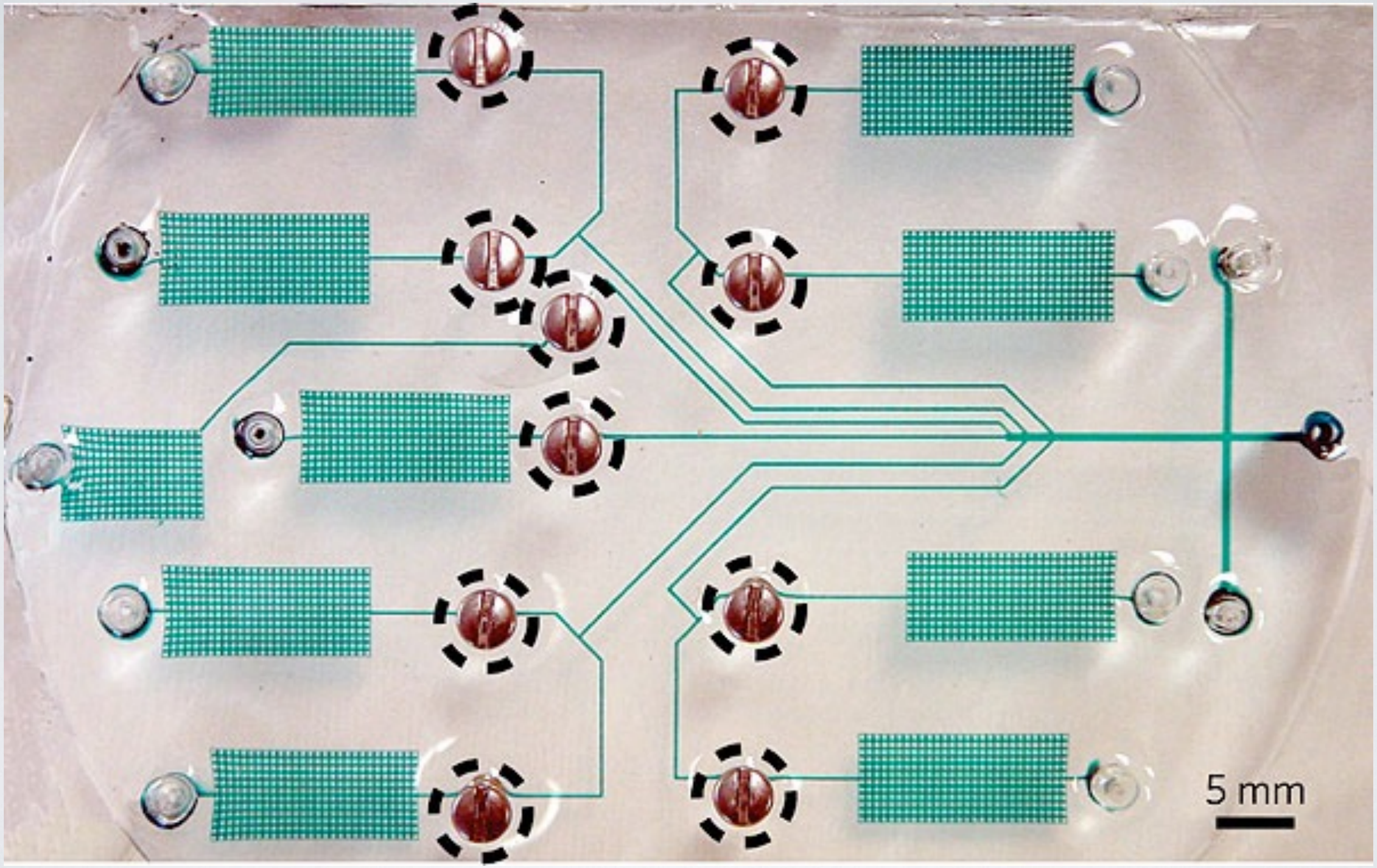
search

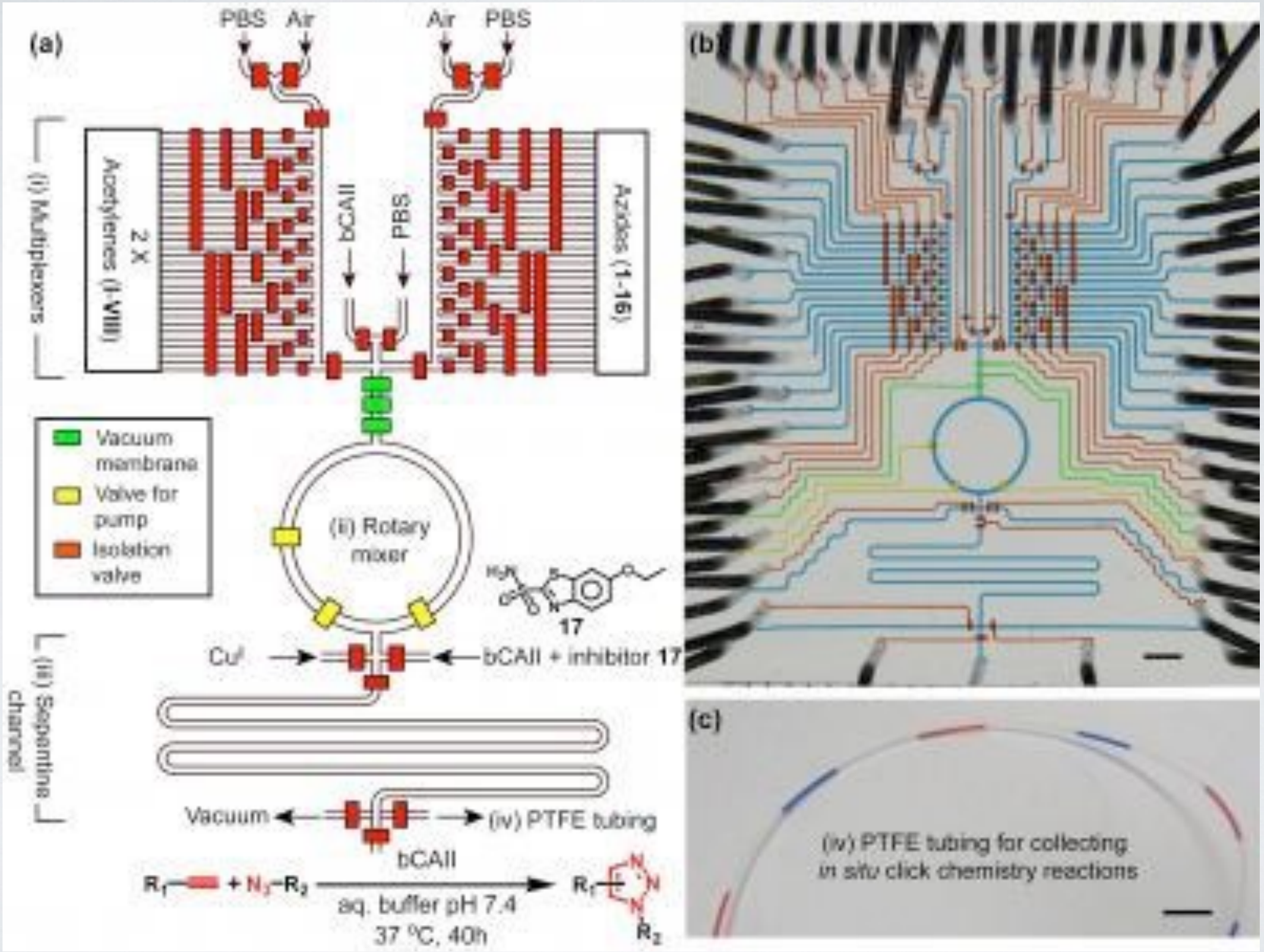
Precision Medicine | Clinical Diagnostics | Metabolomics Services | Technology | Resources | About Us

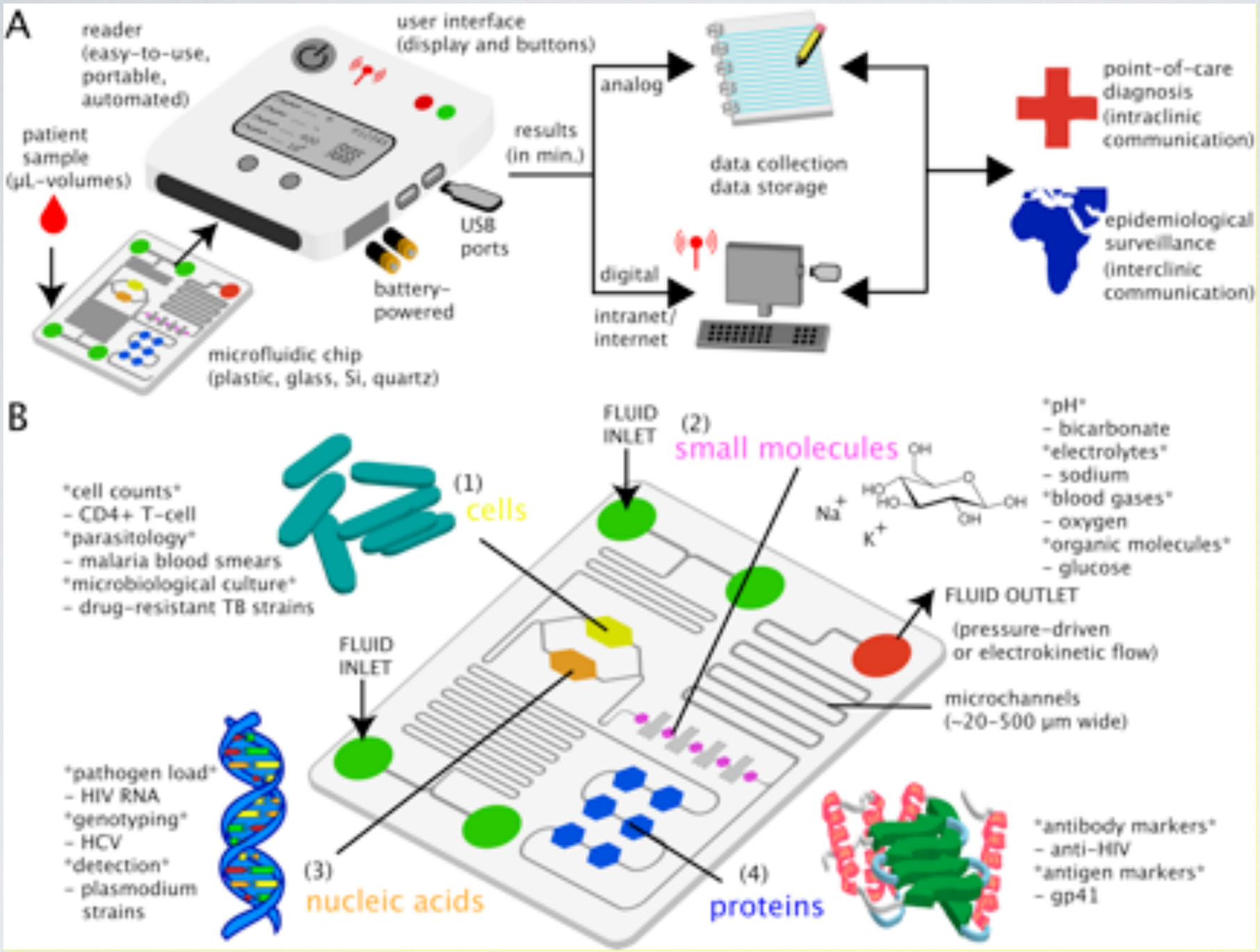
Microfluidics: “lab-on-chip”



A lab-on-chip is not simply a network of microchannels.
It also includes other functions depending on the application such as **pumps**, **valves**, **sensors**, **electronics**, etc.
Therefore, it can be considered as a complex microsystem including mechanical, electronic, fluid functions, etc.


5 mm


(a)
PBS
Air
Air
PBS
(i) Multiplexers
2 X
Acetylenes (I-VIII)
bCAII
PBS
Azides (1-16)
Vacuum membrane
Valve for pump
Isolation valve
(ii) Rotary mixer
17
Cu
bCAII + inhibitor 17
(iii) Serpentine channel
Vacuum
(iv) PTFE tubing
bCAII
aq. buffer pH 7.4
37 °C, 40h
(b)
(c)
(iv) PTFE tubing for collecting
in situ click chemistry reactions


A
reader
(easy-to-use,
portable,
automated)
user interface
(display and buttons)
patient
sample
(µL-volumes)
USB
ports
battery-
powered
microfluidic chip
(plastic, glass, Si, quartz)
results
(in min.)
analog
digital
intranet/
internet
data collection
data storage
point-of-care
diagnosis
(intraclinic
communication)
epidemiological
surveillance
(interclinic
communication)
B
cell counts
- CD4+ T-cell
parasitology
- malaria blood smears
microbiological culture
- drug-resistant TB strains
(1)
cells
FLUID
INLET
(2)
small molecules
pH
- bicarbonate
electrolytes
- sodium
blood gases
- oxygen
organic molecules
- glucose
FLUID OUTLET
(pressure-driven
or electrokinetic flow)
microchannels
(~20-500 µm wide)
FLUID
INLET
pathogen load
- HIV RNA
genotyping
- HCV
detection
- plasmodium
strains
(3)
nucleic acids
(4)
proteins
antibody markers
- anti-HIV
antigen markers
- gp41

BIOMICROFLUIDICS 7, 011302 (2013)

Chip in a lab: Microfluidics for next generation life science research

Aaron M. Streets[1,2] and Yanyi Huang[1,2,3,a)]
[1]*Biodynamic Optical Imaging Center (BIOPIC), Peking University, Beijing 100871, China*
[2]*College of Engineering, Peking University, Beijing 100871, China*
[3]*College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China*

(Received 3 December 2012; accepted 14 January 2013; published online 31 January 2013)

Microfluidic circuits are characterized by fluidic channels and chambers with a linear dimension on the order of tens to hundreds of micrometers. Components of this size enable *lab-on-a-chip* technology that has much promise, for example, in the development of point-of-care diagnostics. Micro-scale fluidic circuits also yield practical, physical, and technological advantages for studying biological systems, enhancing the ability of researchers to make more precise quantitative measurements. Microfluidic technology has thus become a powerful tool in the life science research laboratory over the past decade. Here we focus on *chip-in-a-lab* applications of microfluidics and survey some examples of how small fluidic components have provided researchers with new tools for life science research. [http://dx.doi.org/10.1063/1.4789751]

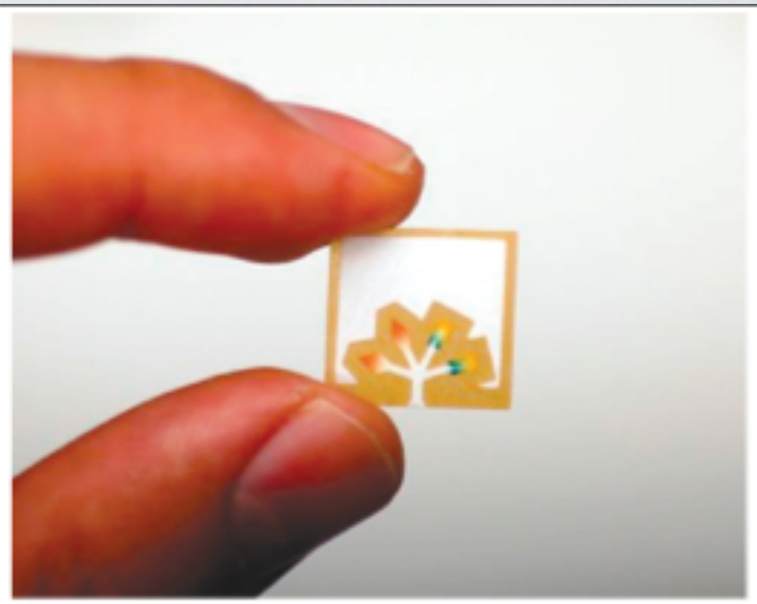
INTRODUCTION

Integrated microfluidic devices have been used in research labs for over twenty years[1] and within the last decade their use in life science research has increased dramatically in large part because of the invention of soft lithography[2] and microfluidic large scale integration.[3] These technological advances have increased the throughput of device production, improved rapid prototyping efforts, and have enabled researchers to enhance the complexity and sophistication of experiments that can be performed on a microfluidic chip. As a result, microfluidic technology is beginning to realize its immense potential for research in life science and medicine.

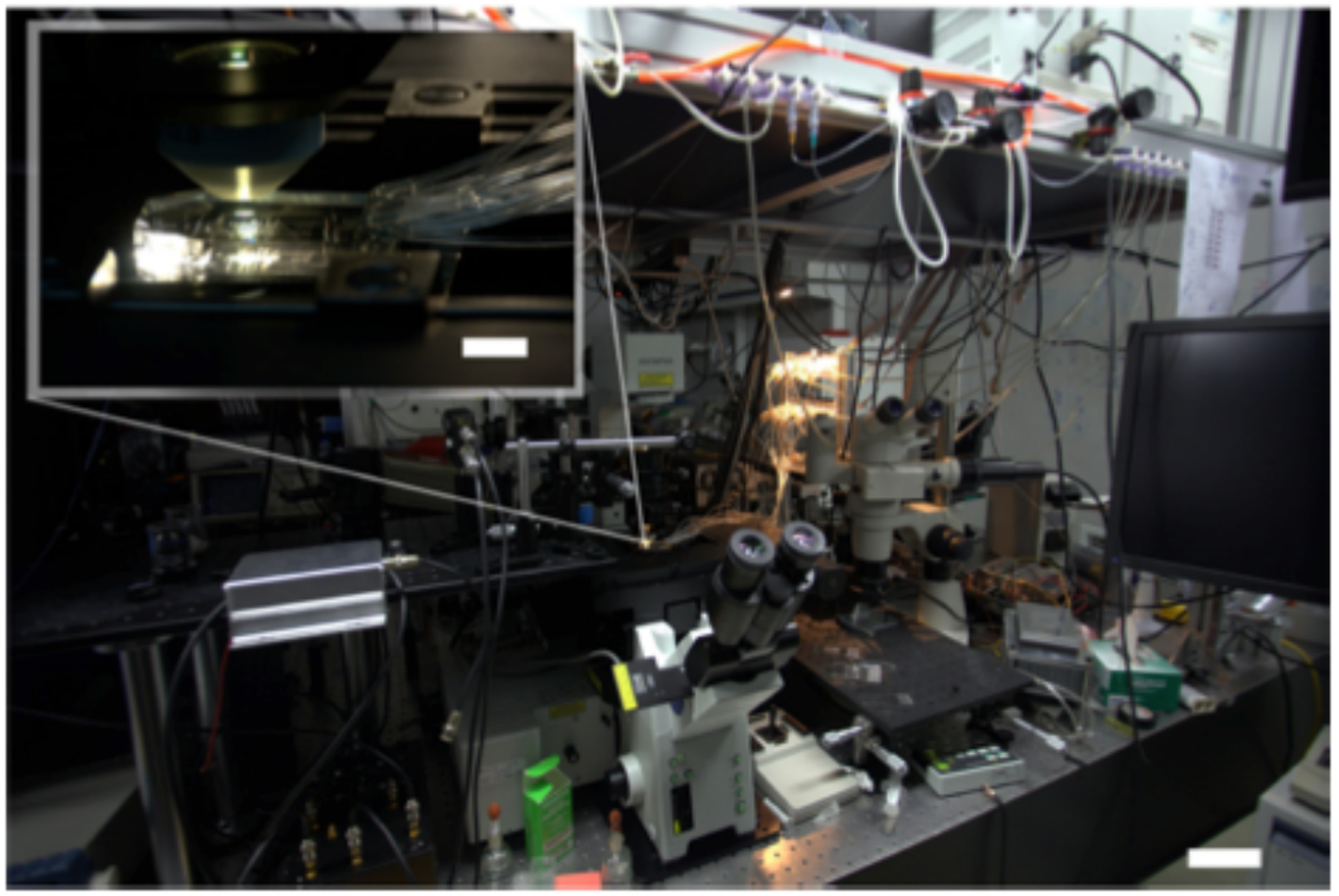
The realm of microfluidic technology is home to a wide variety of embodiments and applications, but the motivation for its use in medicine and life science research can be reduced to two essential themes. One reason that microfluidic solutions are desired is because the size of the device itself is small. The other motivation to use microfluidic technology comes from the advantages gained from the small size of individual fluidic components. The small device size makes microfluidic technology the ideal platform for portable, point-of-care diagnostic devices. In addition to being small, microfluidic devices can be easy to use, cheap to fabricate and operate, require very little sample, and they can be easily disposed of. For these reasons and more the handheld diagnostic device has been recognized as a potential killer application of microfluidics and is the motivation for much of the research effort in microfluidic technology development.

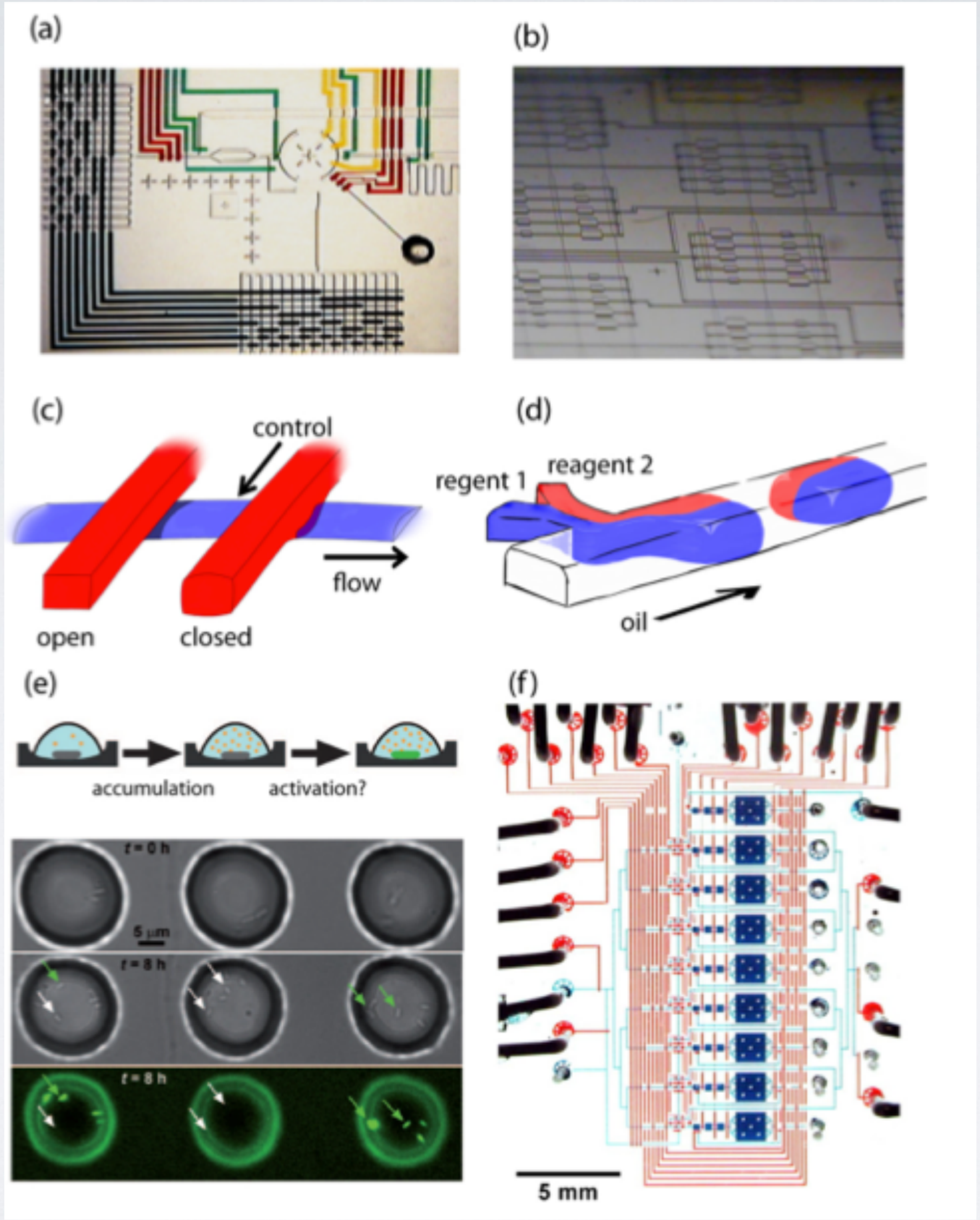
The commercially available glucose meter, for example, is often regarded as the archetype for a handheld diagnostic device[4] (Figure 1(a)). It is cheap ($10–$20 US), easy to operate, provides a clear digital readout of blood glucose level, and uses disposable paper strips for sample delivery. By combining a miniature bio-sensor with a simple and passive microfluidic delivery system the glucose meter requires only a small drop of blood from a finger prick of the operator. The use of paper in microfluidic device fabrication is emerging as a popular strategy

a)Author to whom correspondence should be addressed. Electronic mail: yanyi@pku.edu.cn.



(c)




(a)
(b)
(c)
control
flow
open
closed
(d)
regent 1
reagent 2
oil
(e)
accumulation
activation?
t = 0 h
5 μm
t = 8 h
t = 8 h
(f)
5 mm

analytical chemistry

Article

pubs.acs.org/ac

Microfluidics-Based Assessment of Cell Deformability

Andrea Adamo,[†] Armon Sharei,[†] Luigi Adamo,[‡] ByungKun Lee,[§] Shirley Mao,[†] and Klavs F. Jensen*[,†]

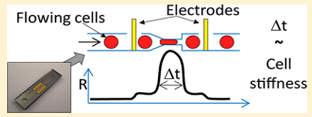
[†]Department of Chemical Engineering, MIT, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

[‡]Department of Medicine, Washington University School of Medicine St Louis, 660 S. Euclid Avenue, St. Louis, Missouri 63110, United States

[§]Department of Electrical Engineering and Computer Science, MIT, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

Supporting Information

ABSTRACT: Mechanical properties of cells have been shown to have a significant role in disease, as in many instances cell stiffness changes when a cell is no longer healthy. We present a high-throughput microfluidics-based approach that exploits the connection between travel time of a cell through a narrow passage and cell stiffness. The system resolves both cell travel time and relative cell diameter while retaining information on the cell level. We show that stiffer cells have longer transit times than less stiff ones and that cell size significantly influences travel times. Experiments with untreated HeLa cells and cells made compliant with latrunculin A and cytochalasin B further demonstrate that travel time is influenced by cell stiffness, with the compliant cells having faster transit time.



Mechanical properties of living cells have been shown to play a role in a number of physiological and pathological processes.[1] Specifically, cells often change their stiffness in disease states including infectious diseases and cancer. For instance, red blood cells of patients with malaria have increased stiffness,[2] human epithelial pancreatic cancer cells have reduced elastic stiffness,[3] and cancerous human bladder epithelial cells are 1 order of magnitude more deformable than healthy ones.[4] A comprehensive review of the literature on this topic with specific attention to cancer has been recently published.[5]

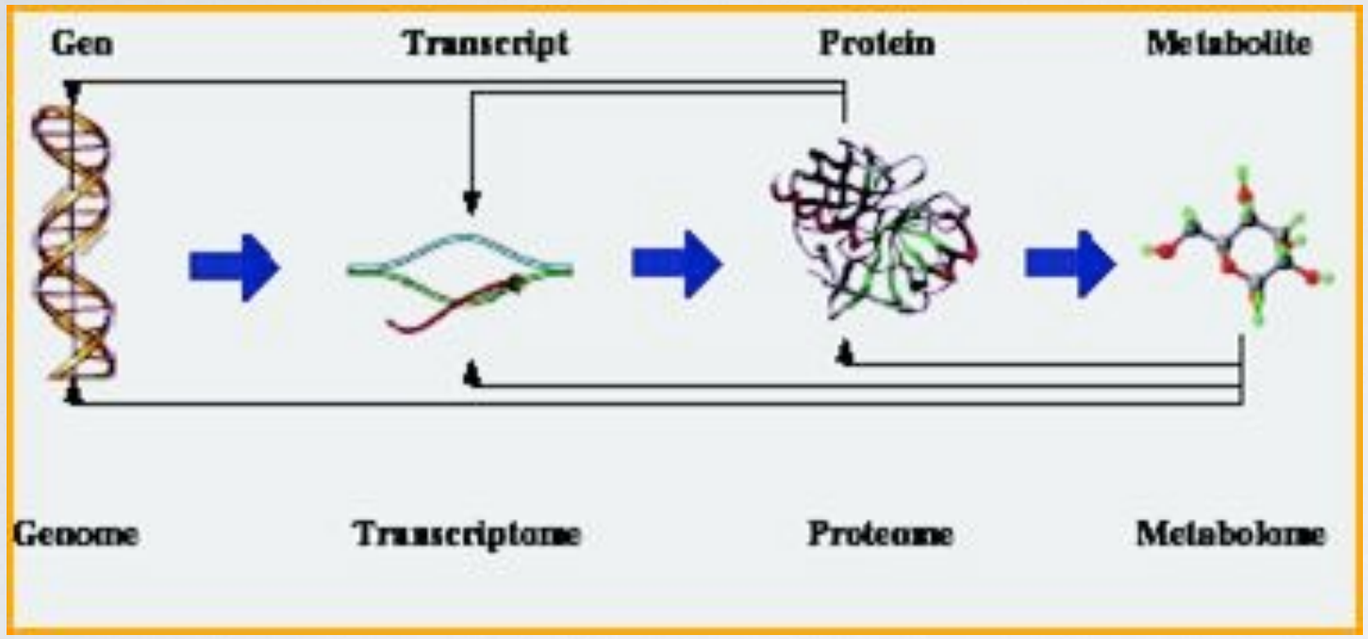
Mechanical characteristics of living cells can be probed with techniques that have either very low throughput or very low accuracy.[6] Among the most accurate methods to measure cell mechanical parameters is atomic force microscopy (AFM) in which the tip of an atomic force microscope is used to indent the cell membrane and measure its mechanical response. The technique is very accurate and provides quantitative data. However, AFM measurements have extremely low throughput and require very skilled operators.[7] Another popular and elegant method to measure cell mechanical properties is optical tweezers,[8] a system that allows applying tensile loads to single cells. This system shares the same advantages and disadvantages of AFM. Similar to optical tweezers but without the need for "handles" is the optical stretcher method.[9,10] A third quantitative technique for evaluation of cellular mechanical properties is micropipet aspiration.[11] In this technique, a cell is aspirated into a micropipet using a negative pressure. Displacements of the cell membrane are recorded to infer cell mechanical properties. This approach provides fairly accurate results with higher throughput than the two previously mentioned techniques, but it is still very slow, with throughput in the order of a few cells/hour at most. A few approaches for the quick nonquantitative evaluation of cellular mechanical properties have been developed. Among them, filtration of cell suspensions through micropores has been used to evaluate the overall ability of a cell population to deform enough to go through the filter.[12] In more sophisticated systems, flowing of cells at a specified pressure through a microfluidic device with defined channels has been optimized to filter out and identify subpopulation of stiffer cells.[13] Alternatively, image analysis of transit times through artificial microcapillaries has yielded information about cell deformability.[14] These approaches provide increased throughput but generally do not take into account the cell diameter. Recently, microfluidic systems in which cells flow through microfabricated channels with known compliance for the purpose of assessing cell deformability have been implemented with optical analysis tools to enable further characterization of the cells of interest.[15] These systems have particularly been optimized for red blood cells that have very similar cell diameters. In addition, a deformability-based cell classification device that uses inertial microfluidics has been proposed;[16] this system offers an interesting and promising approach but currently the separation confounds the roles of cell deformability and diameter.

Herein we describe a microfluidic chip for semiquantitative high-throughput (up to 800 cells/min) probing of cellular mechanical properties of a cell population. This system is based on the observation that, for a given diameter, the travel time of an object through a funnel with diameter smaller than the characteristic diameter of the object is influenced by the stiffness of the object of interest. Applying this to cells, we design a system in which cells are flowed through microfluidic channels with an appropriately designed funnel-shaped narrowing. Because cells behave as dielectrics, at low frequencies we generate an electric field across the microchannel constriction and investigate the

Received: January 26, 2012
Accepted: June 20, 2012
Published: June 20, 2012

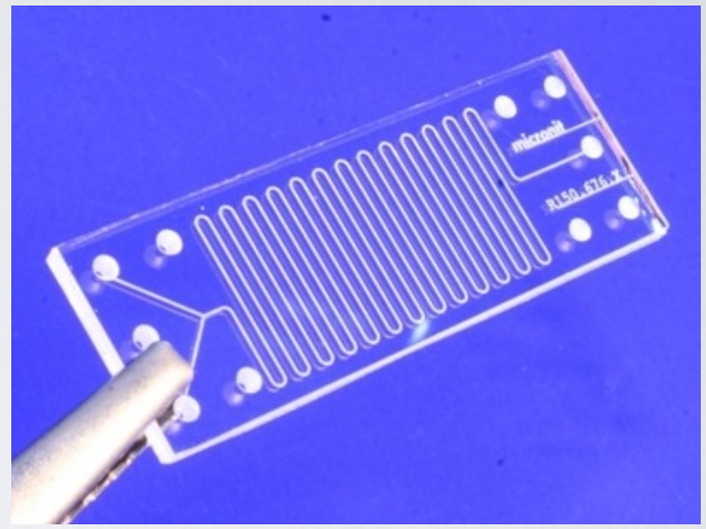
   dx.doi.org/10.1021/ac300264v | *Anal. Chem.* 2012, 84, 6438−6443

“Omics” technologies



Lab-on-chip

+



Systemic functional reconstruction of cells

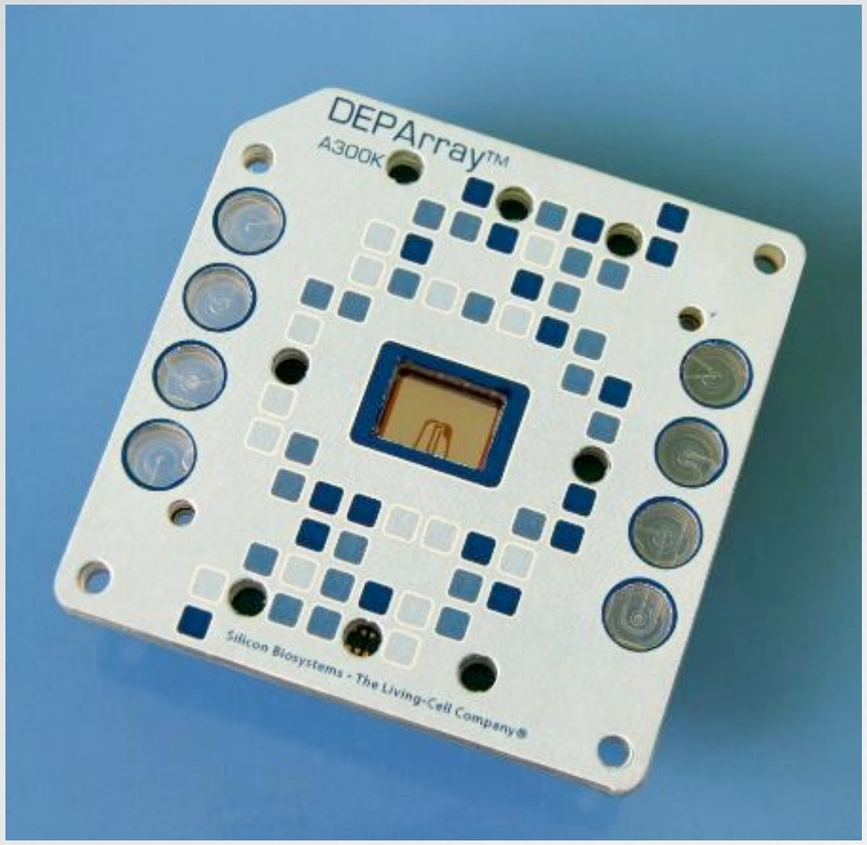
↓

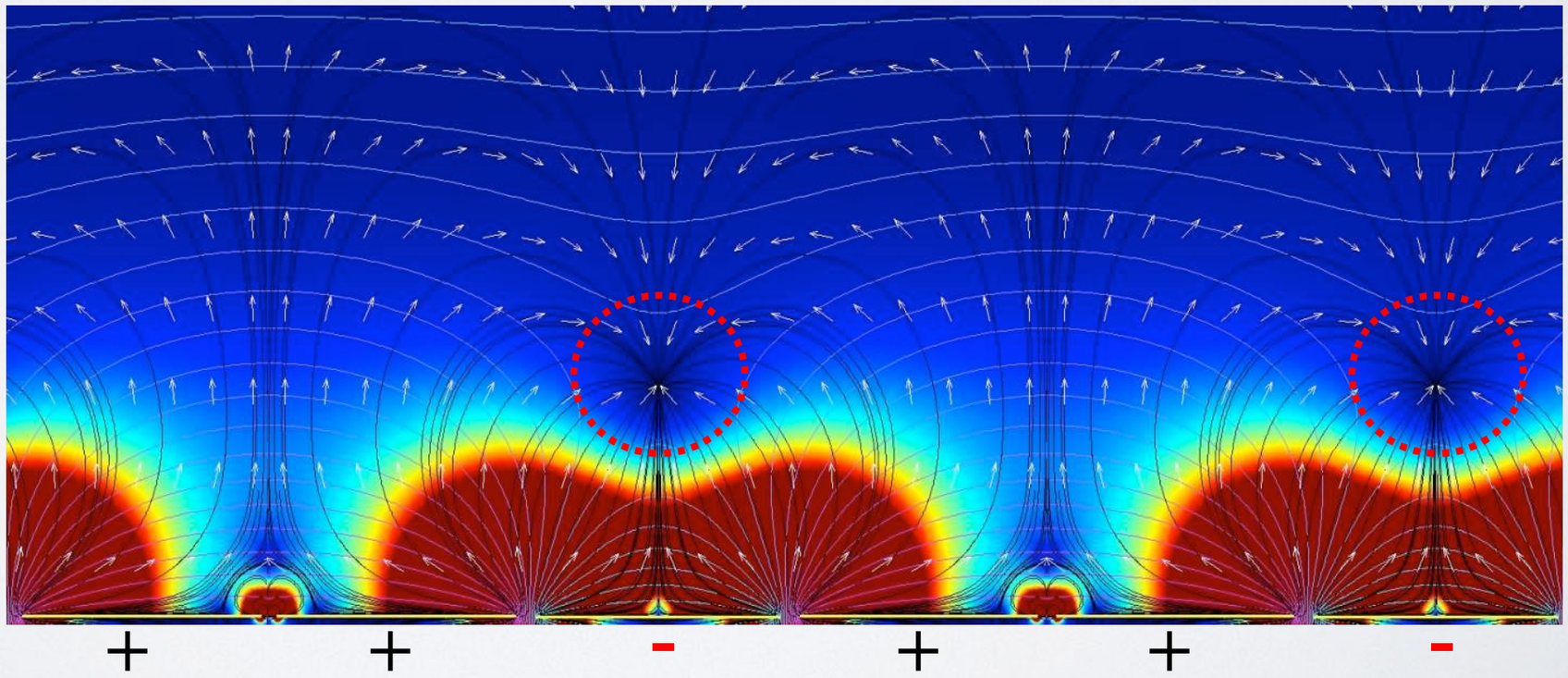
Abstraction (network) → Simulation → Prediction

ImageStream

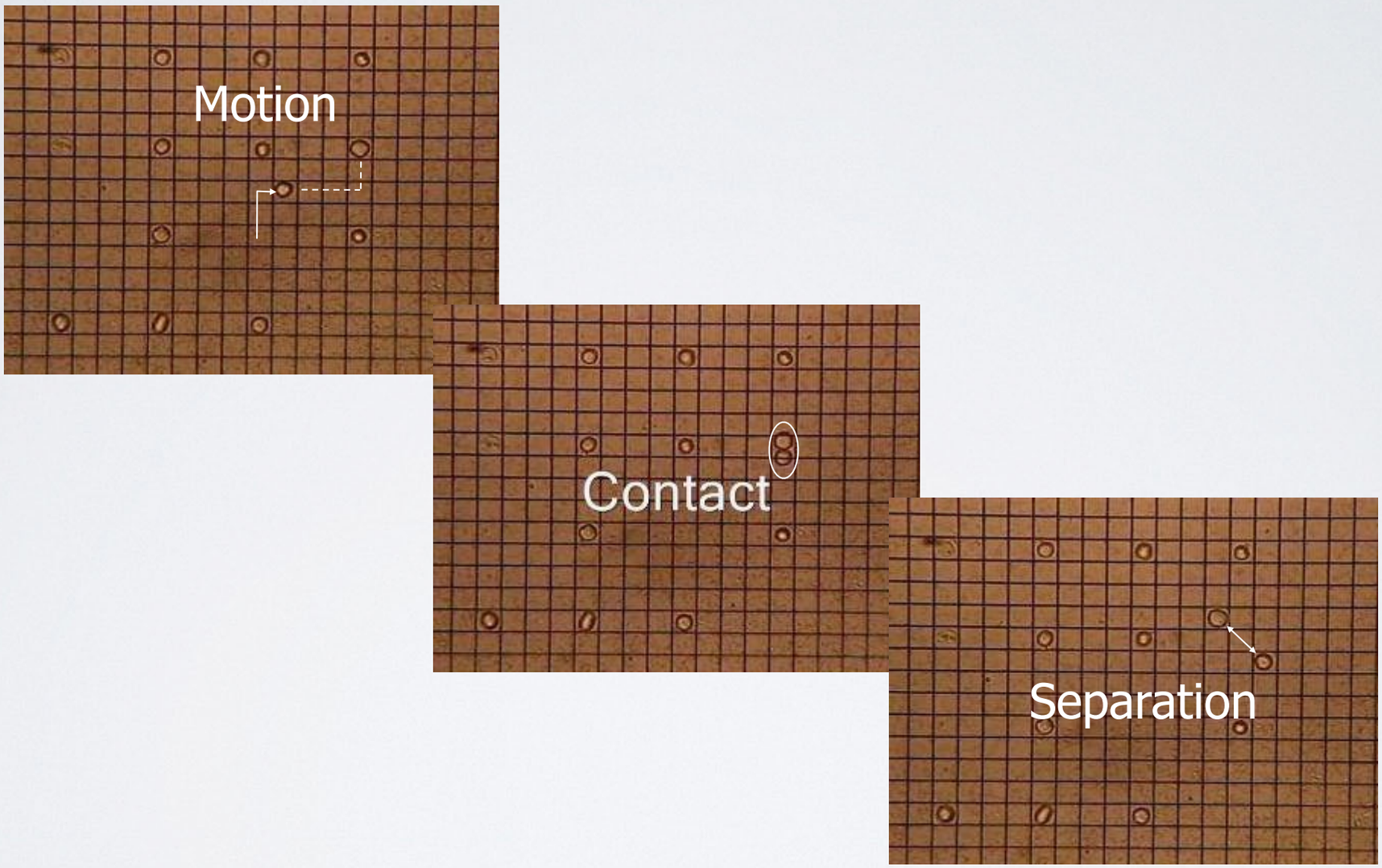
Flux cytometry + confocal microcopy
(1000 cells/sec)

DEPArray


DEPArray™
A300K
Silicon Biosystems - The Living-Cell Company®

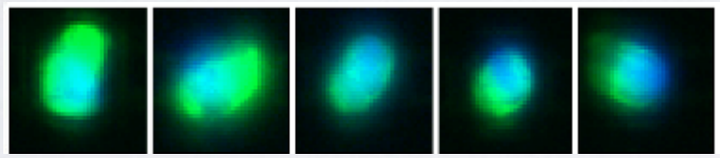

+
+
-
+
+
-

DEPArray™ Enables Individually Organised and Timed Cell-Cell Interactions

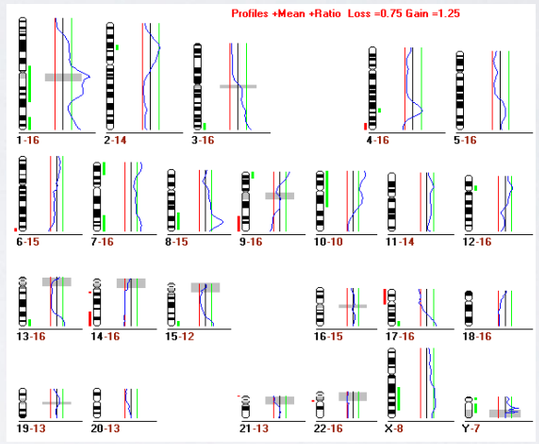


DEPArray™ Enables the Isolation of Circulating Tumor Cells and their Molecular Analysis

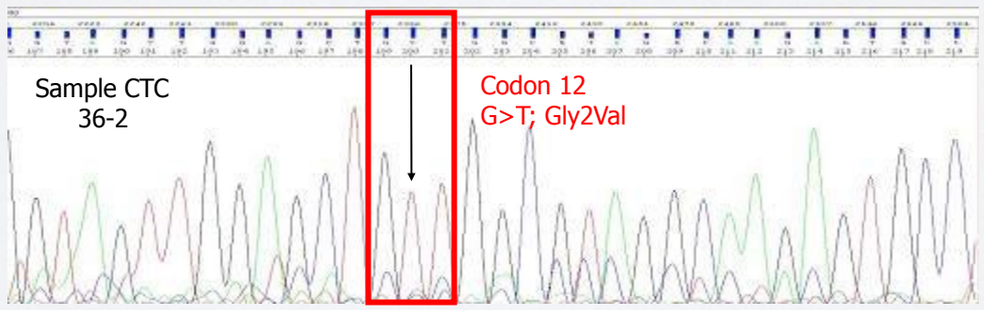
- Genome-wide detection of chromosomal unbalances



CTCs (CK-green/DAPI-blue) isolated with DEPArray™ from a metastatic breast cancer patient



- Detection of individual cell point mutations (e.g. KRAS)



CTCs isolated with DEPArray™ from a metastatic colorectal cancer patient

(10)
Methods in Systems Biology: bioinformatics

The term **Bioinformatics**, (Hwa Lim, late ’80s) generally indicates the application of computer technologies to the study of life sciences.

More precisely it does refer to:

"The study of information content and informational flow in the systems and functional processes related to biology".

<u>A more common connotation</u>

Bioinformatics is the field of science in which biology and information technology merge into a single discipline to facilitate:

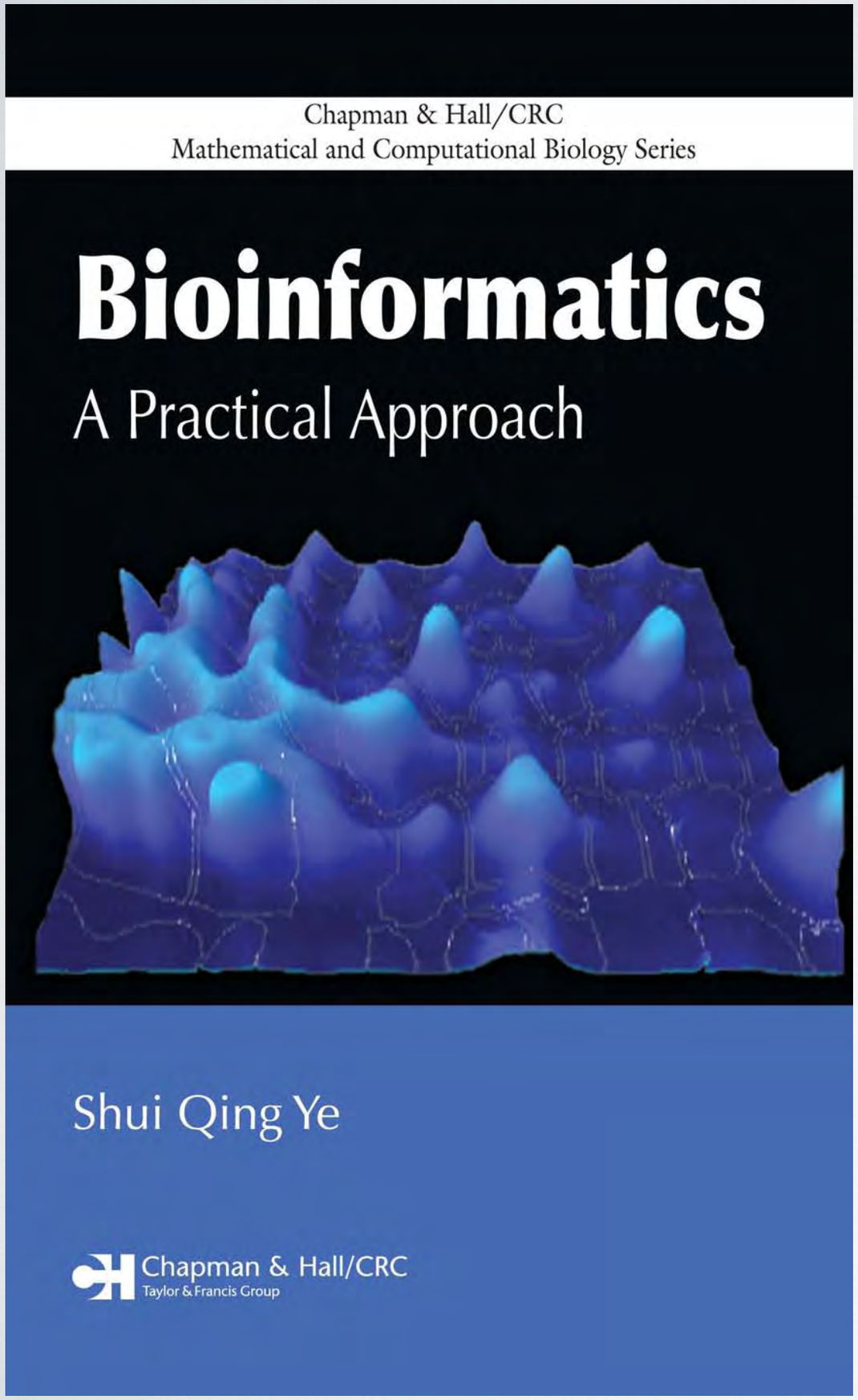
- new biological discoveries from experimental data analysis
- building new computational paradigms from the analysis of living systems.

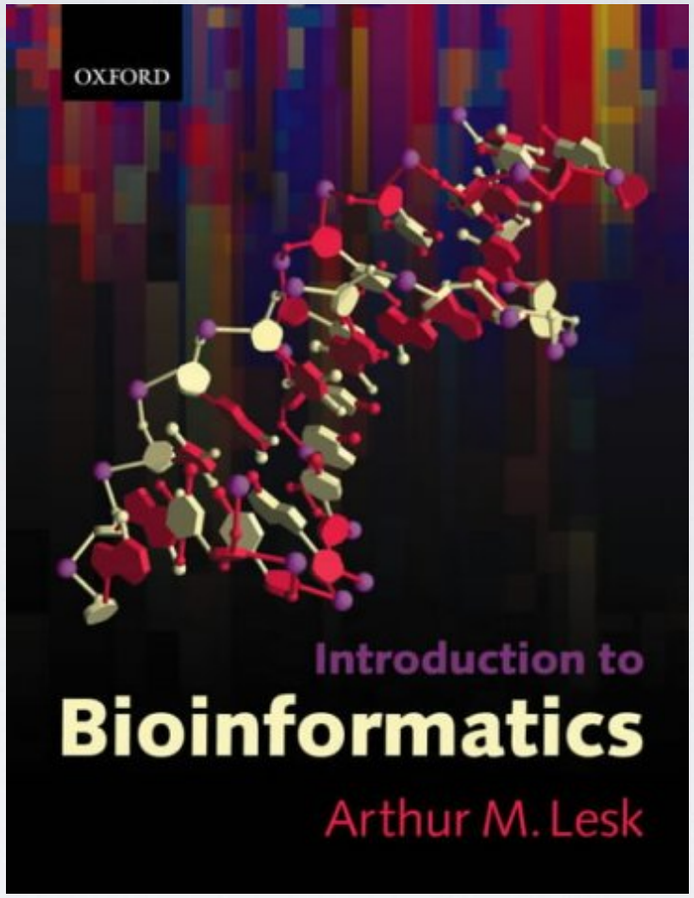
Few definitions related to bioinformatics

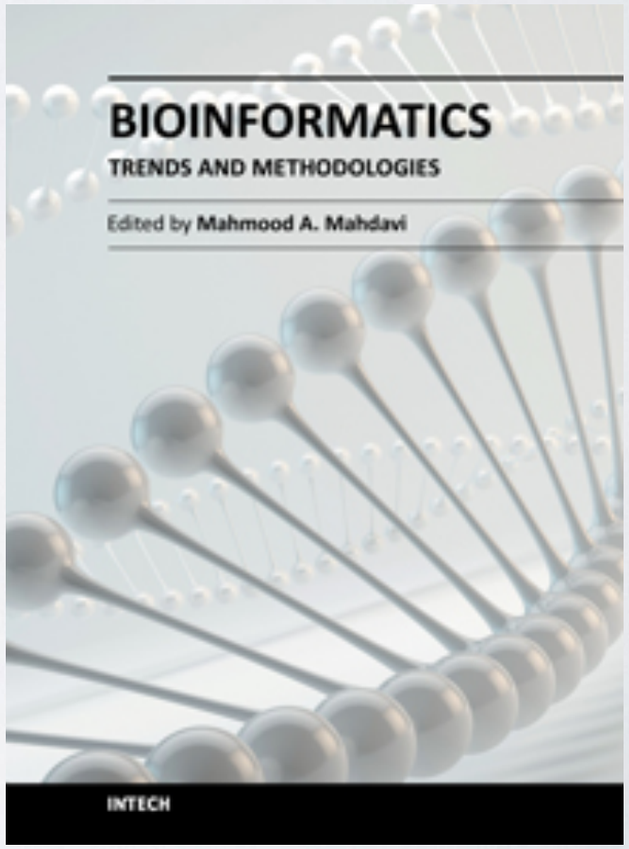
Computational Biology focuses on the algorithmic aspects of biological problems and the efficiency of their solutions.

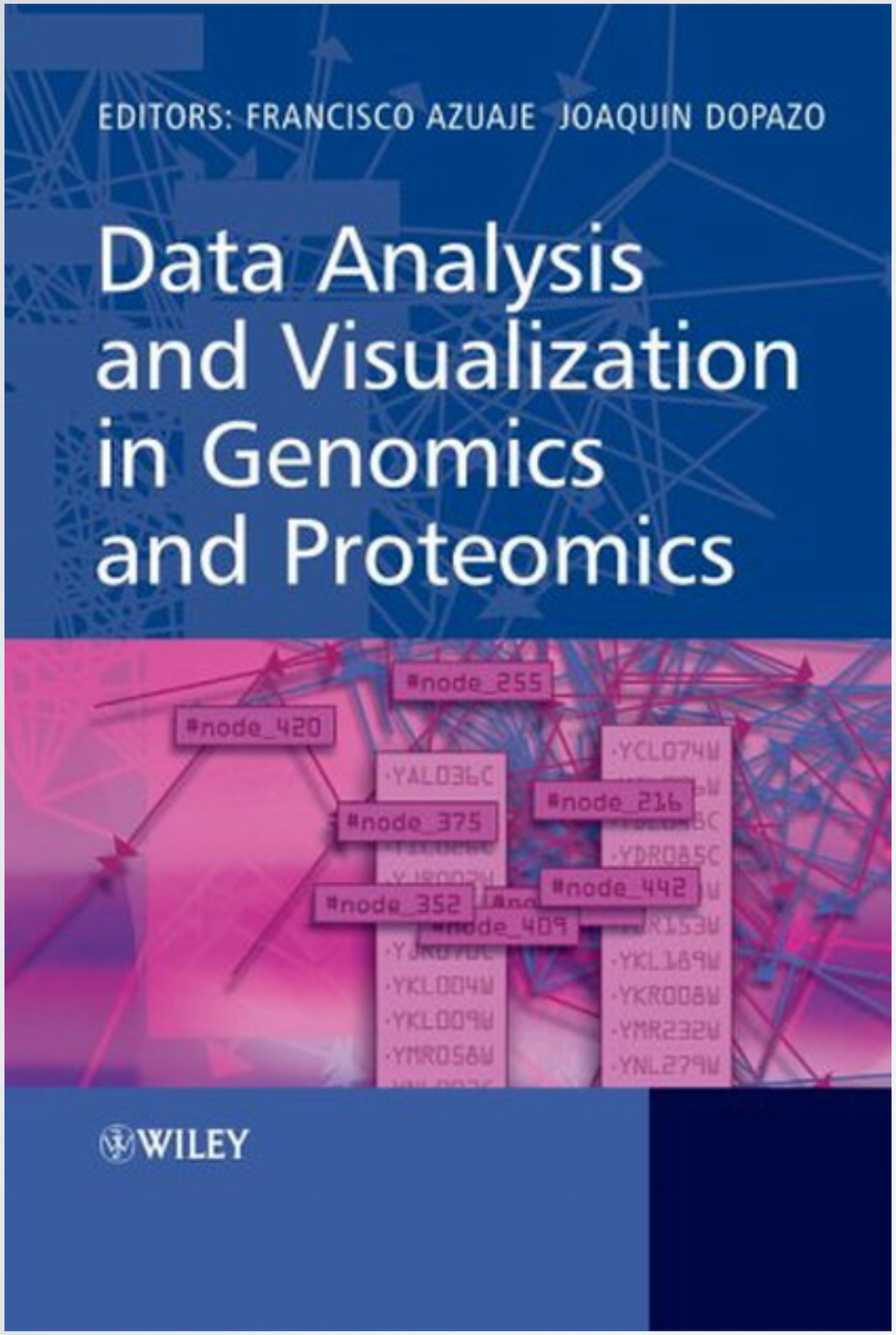
DNA computing is a form of computation that uses DNA and molecular biology, instead of traditional computational technologies (silicon-based).

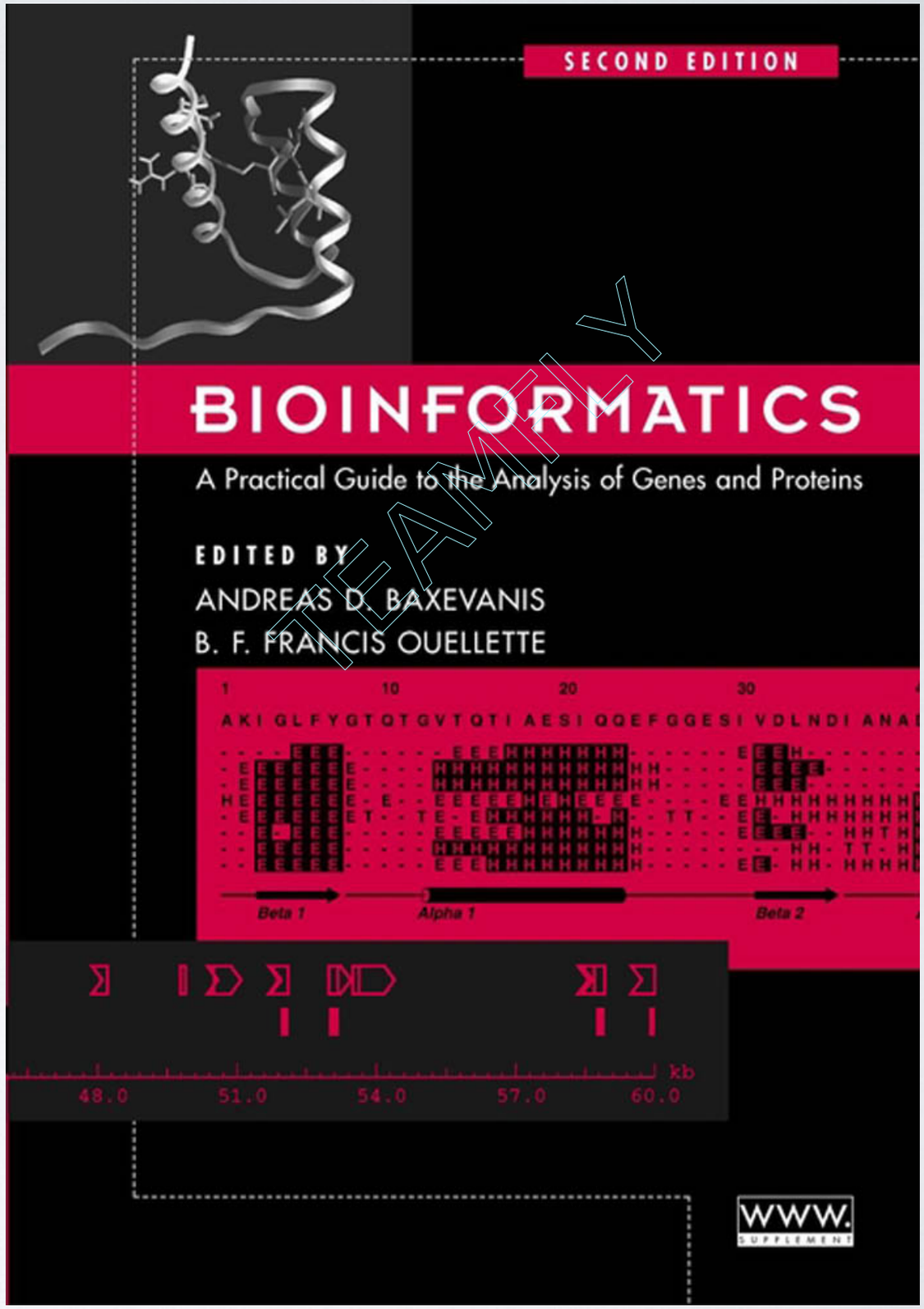
Systems Biology the recently introduced approach based on the theory of complex systems to study biological phenomena from the view point of complexity and emergent properties.


Chapman & Hall/CRC
Mathematical and Computational Biology Series
Bioinformatics
A Practical Approach
Shui Qing Ye
Chapman & Hall/CRC
Taylor & Francis Group


OXFORD
Introduction to
Bioinformatics
Arthur M. Lesk


BIOINFORMATICS
TRENDS AND METHODOLOGIES
Edited by Mahmood A. Mahdavi
INTECH


EDITORS: FRANCISCO AZUAJE JOAQUIN DOPAZO
Data Analysis and Visualization in Genomics and Proteomics
#node_255
#node_420
#node_375
#node_216
#node_352
#node_442
#node_409
WILEY


SECOND EDITION
BIOINFORMATICS
A Practical Guide to the Analysis of Genes and Proteins
EDITED BY
ANDREAS D. BAXEVANIS
B. F. FRANCIS OUELLETTE
Beta 1
Alpha 1
Beta 2
kb
48.0
51.0
54.0
57.0
60.0
WWW.
SUPPLEMENT

Bioinformatics milestones

1962	Pauling's theory of molecular evolution
1965	Margaret Dayhoff's Atlas of Protein Sequences
1970	Needleman-Wunsch algorithm
1977	DNA sequencing and software to analyze it (Staden)
1981	Smith-Waterman algorithm developed
1981	The concept of a sequence motif (Doolittle)
1982	GenBank Release 3 made public
1982	Phage lambda genome sequenced
1983	Sequence database searching algorithm (Wilbur-Lipman)
1985	FASTP/FASTN: fast sequence similarity searching
1988	National Center for Biotechnology Information (NCBI) created at NIH/NLM
1988	EMBnet network for database distribution
1990	BLAST: fast sequence similarity searching
1991	EST: expressed sequence tag sequencing
1993	Sanger Centre, Hinxton, UK
1994	EMBL European Bioinformatics Institute, Hinxton, UK
1995	First bacterial genomes completely sequenced
1996	Yeast genome completely sequenced
1997	PSI-BLAST
1998	Worm (multicellular) genome completely sequenced
1999	Fly genome completely sequenced

A technology driven revolution

Over the past 25 years we have witnessed a real **explosion** in the accumulation of massive <u>biological experimental data</u>

- Started with increasingly fast biotechnology development leading to massive DNA sequences of cells and organisms
- Followed by widespread modern advent of “omcis” technologies

» Need for computers to store **big data**

» Need for programs for management and big-data analysis ("deciphering")

THE TWO “SOULS” OF BIOINFORMATICS

DATA MANAGEMENT → DATABASE (***BIG DATA***)

DATA ANALYSIS → COMPUTATIONAL BIOLOGY

DATABASEs vs. COMPUTATIONAL BIOLOGY

DATABASE:

careful storage, organisation, indexing and maintaining of experimental biological information

COMPUTATIONAL BIOLOGY:

application of algorithms to perform:

- **similarity search between DNA sequences (search of functional homology)**
- **search for genes (ORF) in DNA sequences (decryption)**
- **search of functional motifs in DNA (eg. binding sites for transcription factors)**
- **comparison between global genomes (inter-species)**
- **multiple sequence alignment and phylogenetic analysis**
- **analysis of protein 3D structural data (prediction of protein structure)**
- **analysis of the results of microarray experiments**
- **Protein-Protein interatomic data analysis**

etc.

(11)
Methods in Systems Biology: data-bases

In informatics, the term database indicates a structured electronic archive that allows access and management of data (entering, searching, deleting and updating) by special software applications dedicated to them.

The database is "populated" with a set of information (data) that are divided by topics in a logical order (tables) and then these topics are organised into categories (fields).

Data-base structure

Hierarchical (represented by a tree - ’60s)

Reticular (represented by a graph - ’60s)

Relational (currently the most widespread, represented by tables and relationships between them - ’70s)

Object (extension to the data bases of the paradigm "object oriented**", typical of object-oriented programming - ’80s)**

Semantic (represented by a relational graph - beginning 2000).

bioinformatics.ca
links directory

Bioinformatics Links Directory

Search Bioinformatics Links Directory | Search all | Search Directory

Bioinformatics Links Directory

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

Hide Resources (176) | Hide Databases (621) | Hide Tools (1548)

Computer Related (85)
This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

DNA (604)
This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

Education (75)
Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current news headlines, literature sources, educational material and links to bioinformatics courses and workshops.

Expression (396)
Links to tools for predicting the expression, alternative splicing, and regulation of a gene sequence are found here. This section also contains links to databases, methods, and analysis tools for protein expression, SAGE, EST, and microarray data.

Human Genome (240)
This section contains links to draft annotations of the human genome in addition to resources for sequence polymorphisms and genomics. Also included are links related to ethical discussions surrounding the study of the human genome.

Literature (87)
Links to resources related to published literature, including tools to search for articles and through literature abstracts. Additional text mining resources, open access resources, and literature goldmines are also listed.

Model Organisms (378)
Included in this category are links to resources for various model organisms ranging from mammals to microbes. These include databases and tools for genome scale analyses.

Other Molecules (117)
Bioinformatics tools related to molecules other than DNA, RNA, and protein. This category will include resources for the bioinformatics of small molecules as well as for other biopolymers including carbohydrates and metabolites.

Protein (1007)
This category contains links to useful resources for protein sequence and structure analyses. Resources for phylogenetic analyses, prediction of protein features, and analyses of interactions are also found here.

RNA (203)
Resources include links to sequence retrieval programs, structure prediction and visualization tools, motif search programs, and information on various functional RNAs.

Sequence Comparison (271)
Tools and resources for the comparison of sequences (nucleic acid or protein) including sequence similarity searching, alignment tools, classification and general comparative genomics resources.

Change Text Size: A A A

Username: *

Password: *

CAPTCHA

This question is for testing whether you are a human visitor and to prevent automated spam submissions.

Math question: *
1 + 10 =

Solve this simple math problem and enter the result. E.g. for 1+3, enter 4.

Log in

Create new account
Request new password

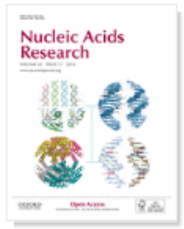
- Main Page
- Sets
- Citations
- Acknowledgements
- News
- Suggest a URL
- NAR Collaboration
- RSS Feeds
- Help / Tutorial
- Support

FEATURED LINK

GENSCAN
http://genes.mit.edu/GENS...
Identification of complete gene structures in genomic DNA.

- 176 Resources
- 621 Databases
- 1548 Tools
- 5 Sets

Follow @biolinksdir



bioinformatics.ca

Return to BioInformatics.ca

Bioinformatics Links Directory

http://www.bioinformatics.ca/links_directory

3463 web-based data bases

(03-04-2017)

Database Commons a catalog of biological databases

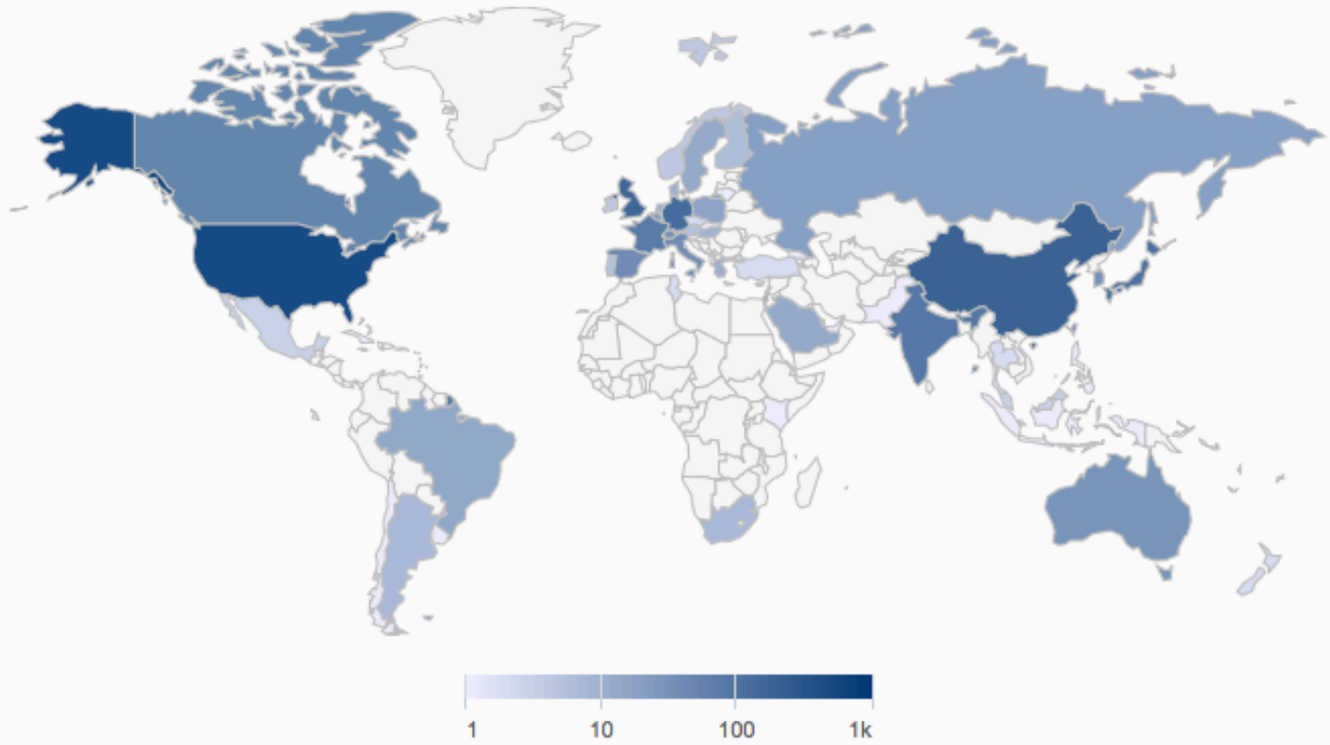
Statistics
- Databases: 1713
- Data types: 14
- Organisms: 400
- Publications: 2609
- Institutions: 746
- Countries/regions: 54

New entries
- HGMD [2016-08-12]
- OGDD [2016-06-13]
- KGCAK [2016-05-06]
- Database Commons [2016-05-06]
- HRGD [2016-05-06]

Recent updates
- AnimalTFDB [2016-09-26]
- TFClass [2016-09-26]
- OnTheFly [2016-09-26]
- ZiFDB [2016-09-26]
- MAPPER2 [2016-09-26]

Worldwide biological databases

1713 databases distributed in 54 countries/regions



Note: All data is derived from papers published from 2005 primarily in *Nucleic Acids Research*, *Database*, etc.

Pathguide» the pathway resource list

Nucleic Acids Research

NAR Database Summary Paper Alphabetic List

1 2 3 4 5 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

OMIC TOOLS

Search engine for biological data analysis

Search among 13,659 omic tools

JOIN OUR COMMUNITY | SIGN IN

Browse by **OMIC TECHNOLOGIES**

High-throughput sequencing
WGS analysis, WES analysis, De novo sequencing analysis, RNA-seq analysis, ChIP-seq analysis, BS-seq

Microarray
aCGH and SNP microarray, Gene expression microarray, DNA methylation microarray

Mass spectrometry
MS-based untargeted proteomics, MS-based targeted proteomics, MS-based untargeted metabolomics

NMR spectroscopy
NMR-based proteomics, NMR-based metabolomics

PCR
qPCR, dPCR, Single-cell qPCR

Flow cytometry & mass cytometry
Flow cytometry, Mass cytometry

Bioimaging
Small-angle scattering, Super-resolution imaging, Mass spectrometry imaging

Other omic technologies
Sanger sequencing, DNA fingerprinting, nCounter System

Browse by **OMIC INTERPRETATION**

Genomics
Genome annotation, Comparative genomics, Phylogenetics & phylogenomics, Genome editing

Epigenomics
Epigenomic analysis software tools and databases

Transcriptomics
Transcriptomic analysis software tools and databases

Proteomics
Protein sequence analysis, Protein comparison analysis, PTM analysis

Metabolomics
Metabolomic analysis software tools and databases

Fluxomics
Fluxomic analysis software tools and databases

Pathway analysis
Biological pathway analysis software tools and databases

Genotype-phenotype analysis
GWAS analysis, Linkage analysis, QTL mapping, eQTL mapping, Phenomics

Browse by **OMIC TOPICS**

Drug discovery

Synthetic biology

Metabolic engineering

Biomolecular structure

Health/Medical research data

Immunology

Text mining

Bioinformatics infrastructure

BIOINFORMATICS TOOLS

The “ID issue”

AKT1

alias

AKT; PKB; RAC; PRKBA; MGC99656; PKB-ALPHA; RAC-ALPHA; AKT1

SYK

alias

FLJ25043; FLJ37489; DKFZp313N1010

HUGO Gene Nomenclature Committee

http://www.genenames.org

ID converter web tools (8)

http://biit.cs.ut.ee/gprofiler/gconvert.cgi

http://biodb.jp

National Center for
Biotechnology Information

Entrez, The Life Sciences Search Engine

Search across databases | GO | Clear | Help

Welcome to the Entrez cross-database search page

- **PubMed:** biomedical literature citations and abstracts
- **PubMed Central:** free, full text journal articles
- **Site Search:** NCBI web and FTP sites
- **Books:** online books
- **Images:** images from full text resources at NCBI
- **OMIM:** online Mendelian Inheritance in Man

- **Nucleotide:** Core subset of nucleotide sequence records
- **EST:** Expressed Sequence Tag records
- **GSS:** Genome Survey Sequence records
- **Protein:** sequence database
- **Genome:** whole genome sequences
- **Structure:** three-dimensional macromolecular structures
- **Taxonomy:** organisms in GenBank
- **SNP:** single nucleotide polymorphism
- **dbVar:** Genomic structural variation
- **Gene:** gene-centered information
- **SRA:** Sequence Read Archive
- **BioSystems:** Pathways and systems of interacting molecules
- **HomoloGene:** eukaryotic homology groups
- **GENSAT:** gene expression atlas of mouse central nervous system
- **Probe:** sequence-specific reagents
- **Genome Project:** genome project information
- **dbGaP:** genotype and phenotype
- **UniGene:** gene-oriented clusters of transcript sequences
- **CDD:** conserved protein domain database
- **UniSTS:** markers and mapping data
- **PopSet:** population study data sets
- **GEO Profiles:** expression and molecular abundance profiles
- **GEO DataSets:** experimental sets of GEO data
- **Epigenomics:** Epigenetic maps and data sets
- **Cancer Chromosomes:** cytogenetic databases
- **PubChem BioAssay:** bioactivity screens of chemical substances
- **PubChem Compound:** unique small molecule chemical structures
- **PubChem Substance:** deposited chemical substance records
- **Protein Clusters:** a collection of related protein sequences
- **Peptidome:** MS/MS proteomic experiments
- **OMIA:** online Mendelian Inheritance in Animals
- **BioSample:** biological material descriptions

- **NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections
- **MeSH:** detailed information about NLM's controlled vocabulary

http://www.ncbi.nlm.nih.gov/gquery

EMBL-EBI European Bioinformatics Institute

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

Explore the EBI:

FIND

Examples: ROA1_HUMAN, tpi1, Sulston...

Help & Suggestions | Support & Feedback

Data Resources &Tools

- **ENA**
- **UniProt**
- **ArrayExpress**
- **Ensembl**
- **InterPro**
- **PDBe**
- Genomes
- Nucleotide Sequences
- Protein Sequences
- Macromolecular Structures
- Small Molecules
- Gene Expression
- Molecular Interactions
- Reactions & Pathways
- Protein Families
- Enzymes
- Literature
- Taxonomy
- Ontologies
- Patent Resources
- Sequence Similarity & Analysis
- Pattern & Motif Searches
- Structure Analysis
- Text Mining
- Downloads
- Web Services

European Bioinformatics Institute

Latest News

EMBL-EBI will continue to support the Sequence Read Archive
Feb 16, 2011
EMBL-EBI will continue to support the Sequence Read Archive. See here for further details.

UK funds earmarked for ELIXIR, a major bioscience data infrastructure project
Feb 09, 2011
UK funding has been earmarked for ELIXIR , an infrastructure for biological information. The project, if approved, will allow the construction of ELIXIR's central hub at EMBL-EBI, ensuring the maintenance and expansion of essential biological data resources. Find out more (PDF).

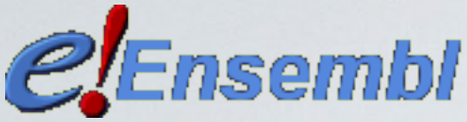
Research Highlights

Does in vitro potency really correlate with the therapeutic dose?
Mar 03, 2011
Members of the EBI's ChEMBL group have been probing the links between in vitro potency, ADMET and physicochemical parameters. Published this week in Nature Reviews Drug Discovery, their review challenges the widely held assumption that compounds with higher in vitro potency at their target have greater potential to translate into successful, low-dose therapeutics... more

Events

- Next Generation Sequencing Workshop 2011
4th-6th April 2011 more
Registration closes 21st March 2011
- Joint EBI-Wellcome Trust Summer school in bioinformatics
220-24 June 2011 more
Registration closes 28th March 2011.
- Joint EBI-Wellcome Trust Proteomics Bioinformatics Workshop
15-19 July 2011 more more
Registration closes 8th April 2011
- Programmatic access to biological databases (Java)
9-13 May 2011 more
Registration closes 9th April 2011

http://www.ebi.ac.uk



http://www.ensembl.org/

http://www.genome.jp/

http://www.pathwaycommons.org/pc/

http://harvester.fzk.de/harvester/

http://www.systemsbiology.org

http://www.systems-biology.org

http://www.pantherdb.org/

http://david.abcc.ncifcrf.gov/

Welcome to Open Targets

For biomedical researchers who need to identify a biological target for a new therapy, Open Targets is a public-private initiative to generate evidence on the validity of therapeutic targets based on genome-scale experiments and analysis. Open Targets is working to create an R&D framework that applies to a wide range of human diseases, and is committed to sharing its data openly with the scientific community.

About Open Targets

Biogen EMBL-EBI gsk wellcome trust sanger institute

Browse Pathways

Analyze Data

Reactome FIViz app

User Guide

Data Download

Contact Us

About Reactome

Reactome is a free, open-source, curated and peer reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education. The current version (v58) of Reactome was released on October 5, 2016.

The development of Reactome is supported by grants from the US National Institutes of Health (P41 HG003751 and 1U54GM114833-01), Ontario Research Fund, and the European Molecular Biology Laboratory.

Tweets

Current Version: Reactome V58

reactome Retweeted

ChEBI
@chebit

Release 145 is live with 50089 fully annotated entities. See our entity of the month: GNF6702.
bit.ly/1kLIxDP

01 Nov

reactome Retweeted

Antonio Fabregat
@AntonioFabregat

Happy to meet my colleague @marcelakarey from @GrameneDatabase presenting the plant @reactome poster. Come to 107 if you're at #BioData16

Embed View on Twitter

Metabolomic data-bases

http://www.hmdb.ca/

SMPDB
Small Molecule Pathway Database

http://www.smpdb.ca/

MBROLE
Metabolites Biological Role

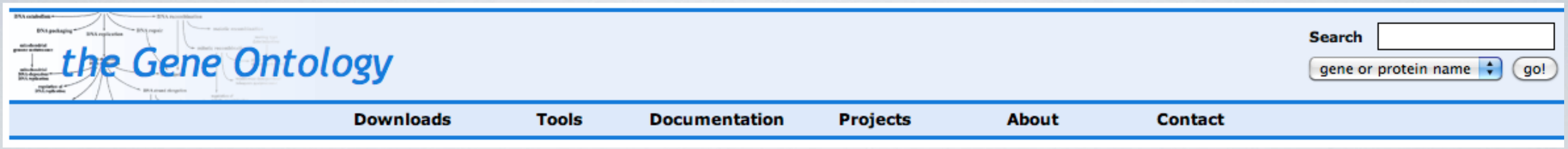
http://csbg.cnb.csic.es/mbrole/index.jsp

http://metpa.metabolomics.ca/MetPA/faces/Home.jsp

http://scopes.biologie.hu-berlin.de/

Metabolomic analysis with Cytoscape + Metscape

Meta-inflammoma and cancer



http://www.geneontology.org/

The Gene Ontology (GO) is a highly **STRUCTURED VOCABULARY,** continuously growing and updated, which enables *dynamic* assignment of functional significance to all proteins coded and expressed by eukaryotic cells.

This vocabulary is defined as CONTROLLED and DYNAMIC since it is written and regularly maintained by a group of specialists in charge: the curators. The ultimate goal is to create a vocabulary that can be applied to all eukaryotes as information about genes, proteins and their functional role in the cell is accumulated.

The **Gene Ontology consortium** provides controlled vocabularies of defined terms representing gene product properties.

These cover three domains:

- **Cellular Component:** the parts of a cell or its extracellular environment;
- **Molecular Function:** the elemental activities of a gene product at the molecular level, such as binding or catalysis;
- **Biological Process:** sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

The three Ontologies:

(1)
Cellular Component

These terms describe a component of a cell that is part of a larger object, such as an sub-cellular structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

Overall, it does refer to intracellular compartment localisation.

(2)
Molecular function

Molecular function terms describes activities that occur at the molecular level, such as "catalytic activity" or "binding activity".

GO molecular function terms represent activities rather than the entities (such as, molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place.

Molecular functions generally correspond to biochemical activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products.

Examples of broad functional terms are "catalytic activity" and "transporter activity";

Examples of narrower functional terms are "adenylate cyclase activity" or "Toll receptor binding".

It is easy to confuse a gene product name with its molecular function; for that reason GO molecular functions are often appended with the word "activity".

(3)
Biological Process

A biological process term describes a *series of events accomplished by one or more organized assemblies of molecular functions.*

Examples of broad biological process terms are "cellular physiological process" or "signal transduction".

Examples of more specific terms are "pyrimidine metabolic process" or "alpha-glucoside transport".

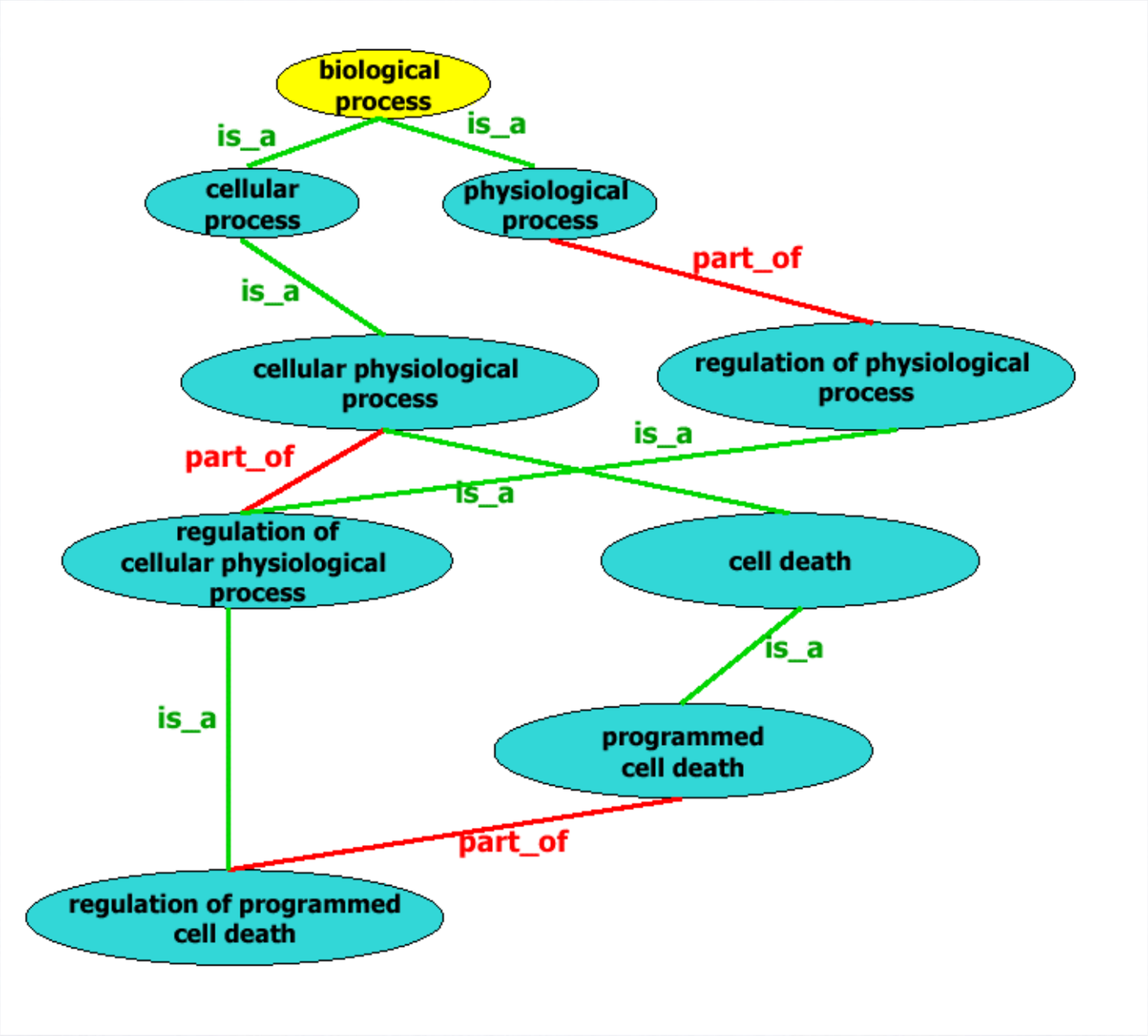
The general rule to assist in distinguishing between a biological process and a molecular function is that a process must have more than one distinct “steps”.

NOTABLY: a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

DAG (Directed Acyclic Graph)

The GO ontology is structured as a **directed acyclic graph** where each GO term has defined relationships to one or more other terms in the same domain, and sometimes to other domains.

The GO vocabulary is designed to be species-agnostic, and includes terms applicable to prokaryotes and eukaryotes, and single and multicellular organisms.



In an example of GO annotation, the gene product "**Cytochrome c**" can be described:

by the **Molecular Function** term "oxidoreductase activity",

by the **Biological Process** terms "oxidative phosphorylation" and "induction of cell death", and

by the **Cellular Component** terms "mitochondrial matrix" and "mitochondrial inner membrane".

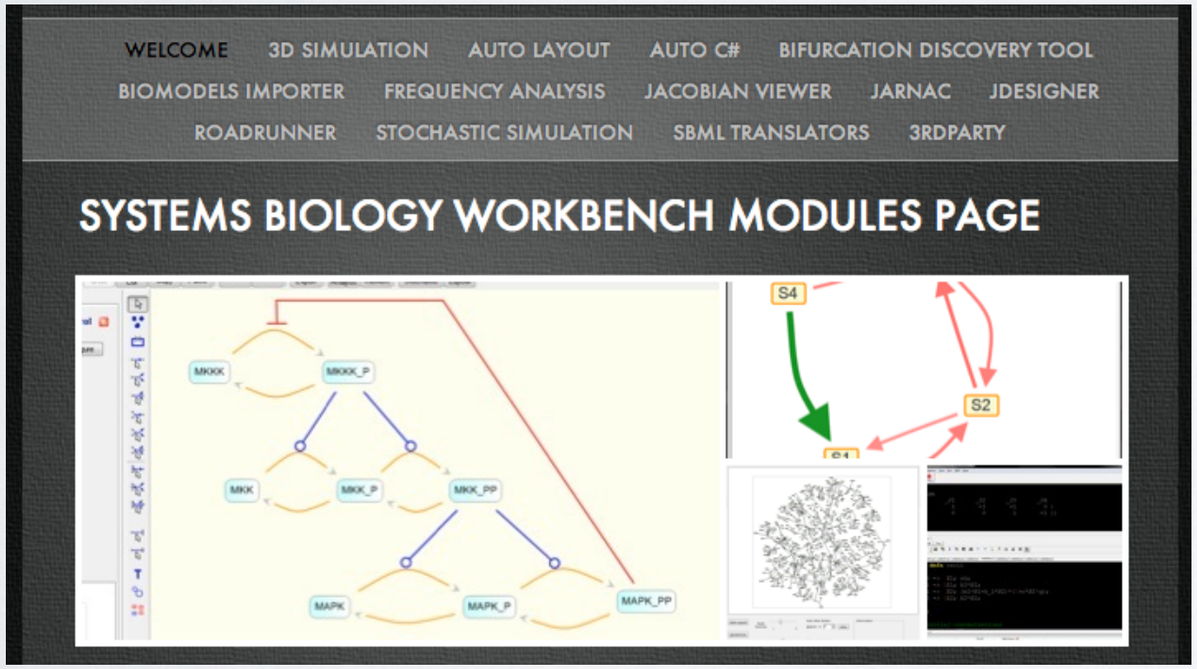
In a context of bioinformatics analysis, these annotations can be considered as

"informational layers" or "dimensions"

(12)
Methods in Systems Biology: software for Systems Biology

There are hundreds of softwares dedicated to various applications in bioinformatics and computational biology! (Java)

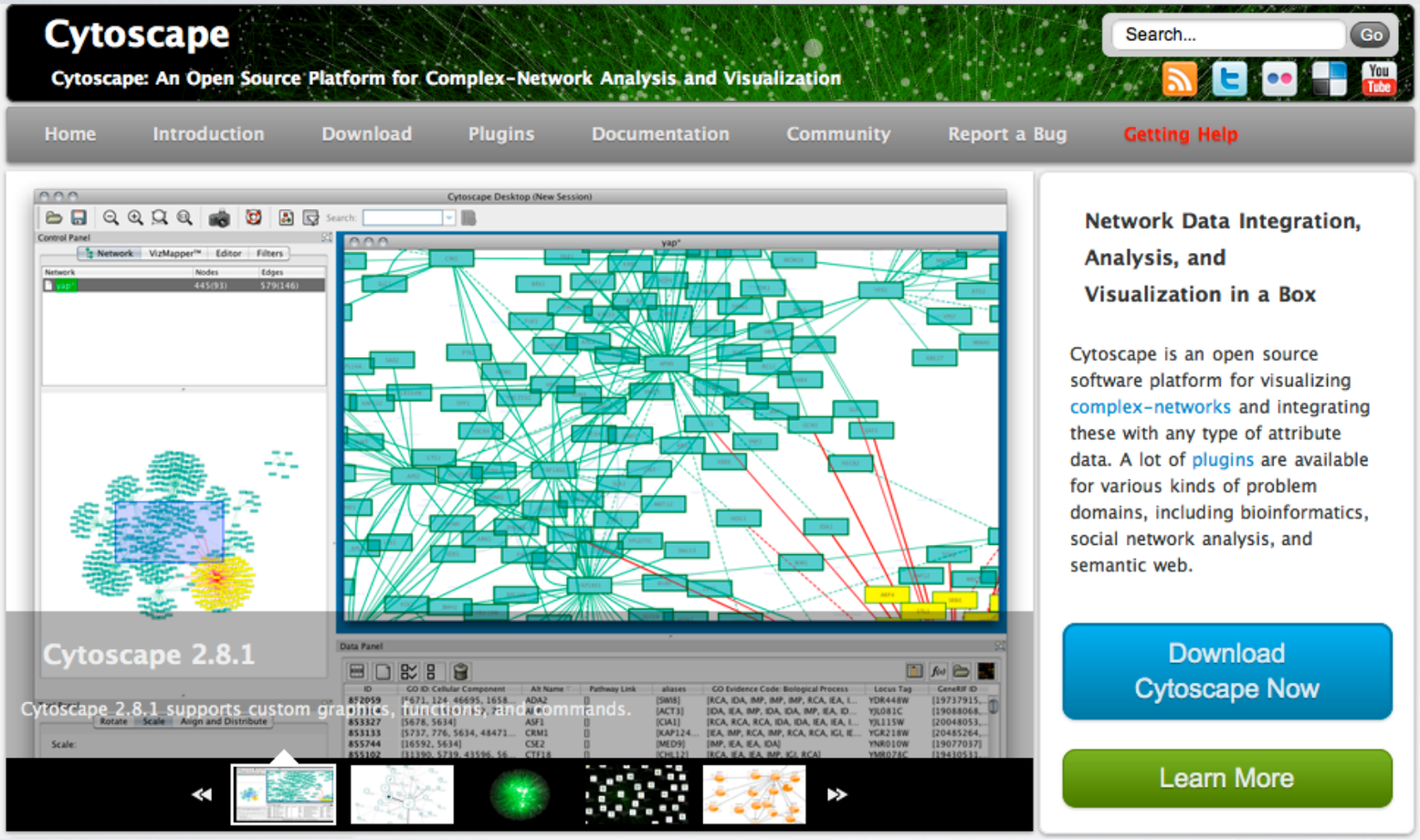
http://systems-biology.org/software/

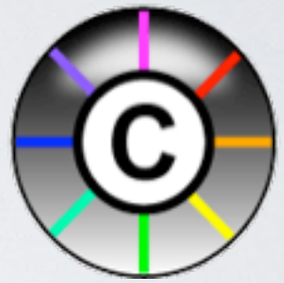


http://jdesigner.sourceforge.net/Site/Welcome.html

SBML.org The Systems Biology Markup Language

http://sbml.org

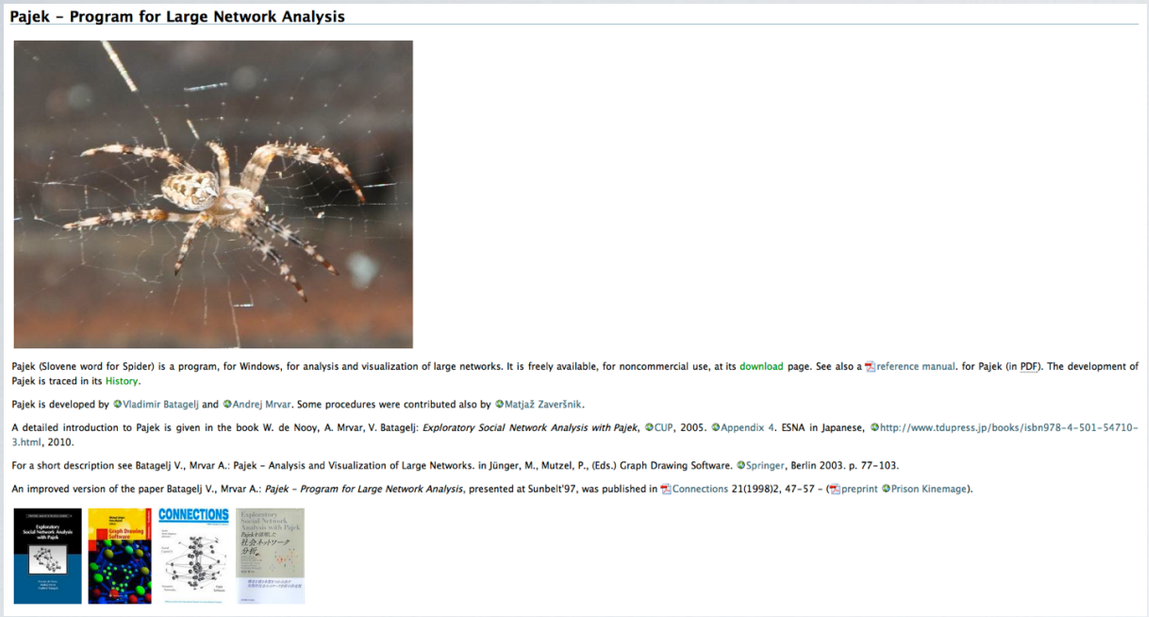




http://
www.cytoscape.org

BIOLOGICAL NETWORKS
Integrated Research Environment

http://biologicalnetworks.net/



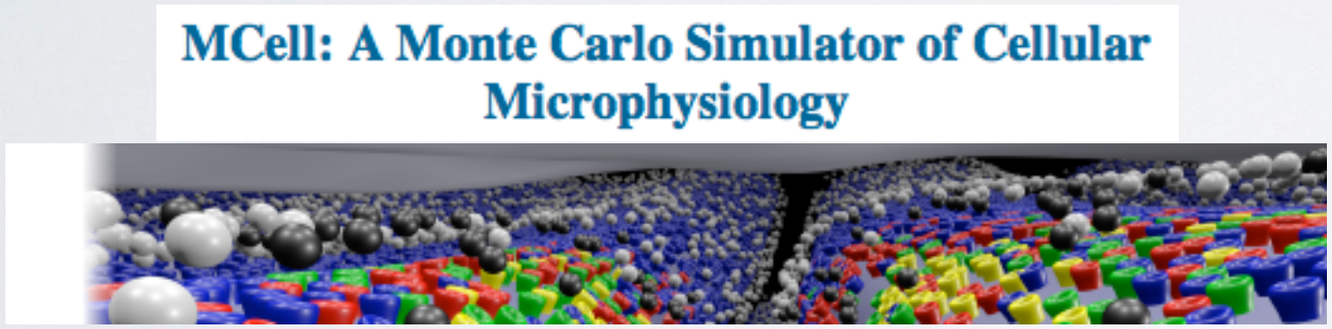
http://pajek.imfm.si/doku.php?id=pajek

http://rsat.ulb.ac.be/rsat/index_neat.html

http://mavisto.ipk-gatersleben.de/

http://visant.bu.edu/

http://www.nrcam.uchc.edu/



http://www.mcell.cnl.salk.edu/

http://www.cellml.org/

http://mplab.sci.univr.it/index.html

http://www.biotapestry.org/

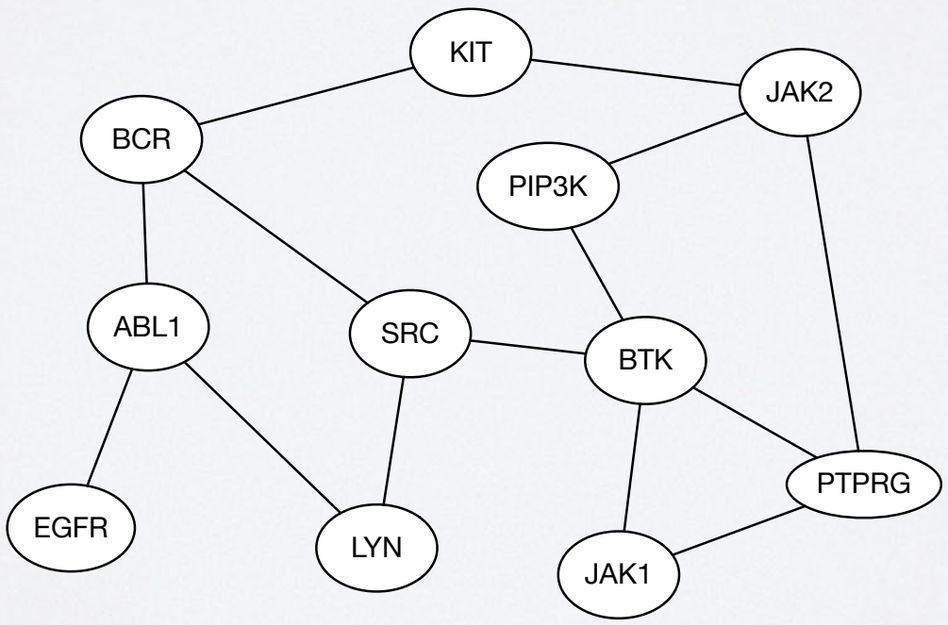
http://www.tm4.org/

(12)
Application contexts for Systems Biology: biological networks - transcriptomics, proteomics, metabolomics, diseasomics, etc.

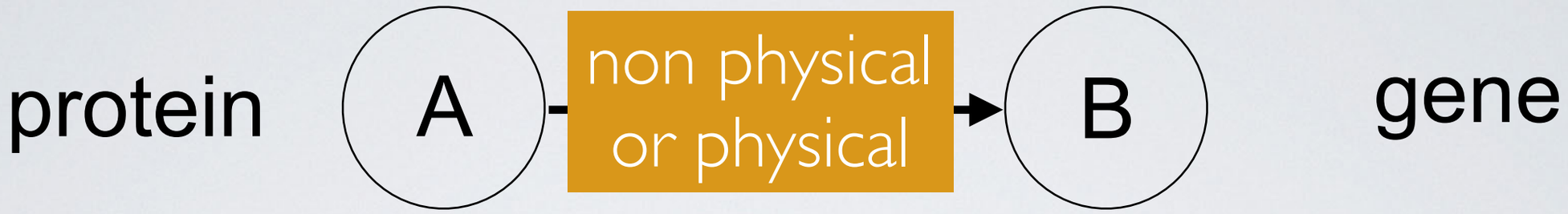
In biology, all systems functions are conceivable as networks and is possible to abstract them as 2D graphs

Fundamental unit

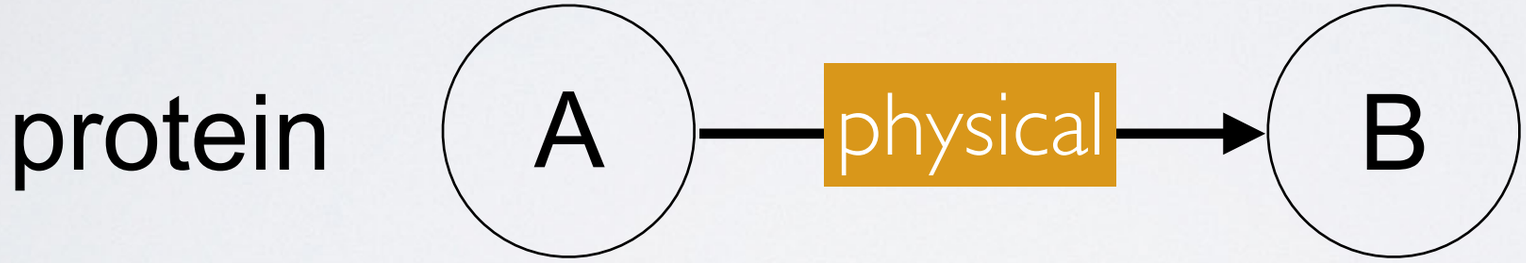
Graph



Transcriptomic Network = protein regulates gene expression (mRNA)

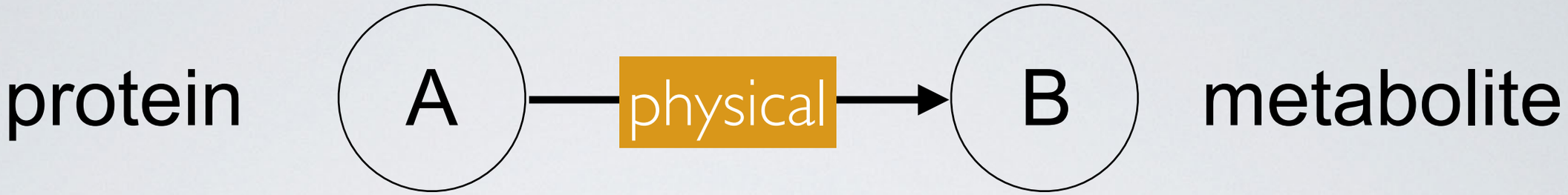


Interactomic Network = protein binds physically to other proteins (PPI)

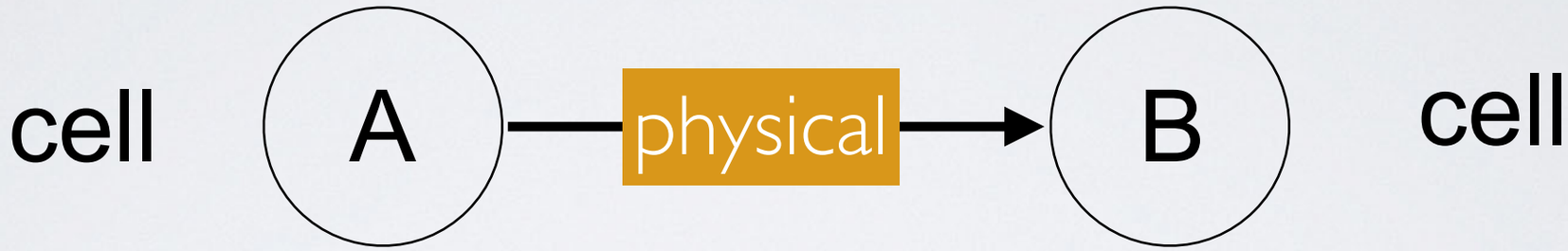


Metabolic Network (mass transfer) = metabolite transform to another metabolite

Metabolic Network (regulators) = a protein catalyzes the transformation of a metabolite to a different one

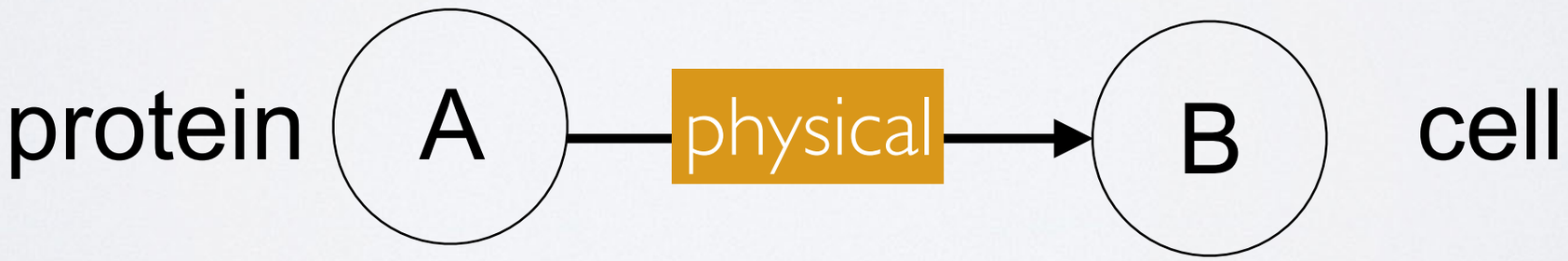


Cell Network = cell regulates other cells



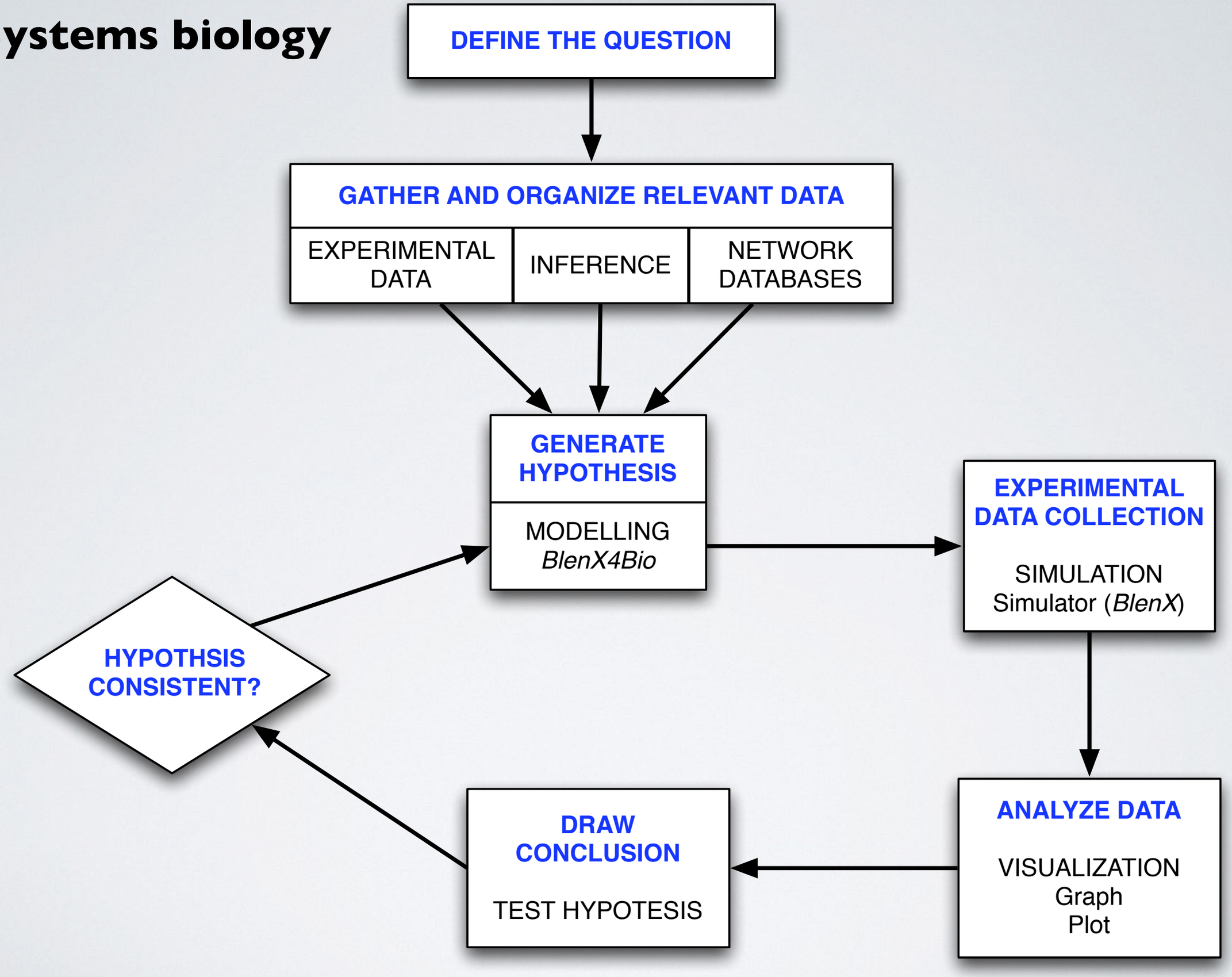
Es. neurons_neurons; APC_T-cells

Protein_Cell Network = protein regulates cells

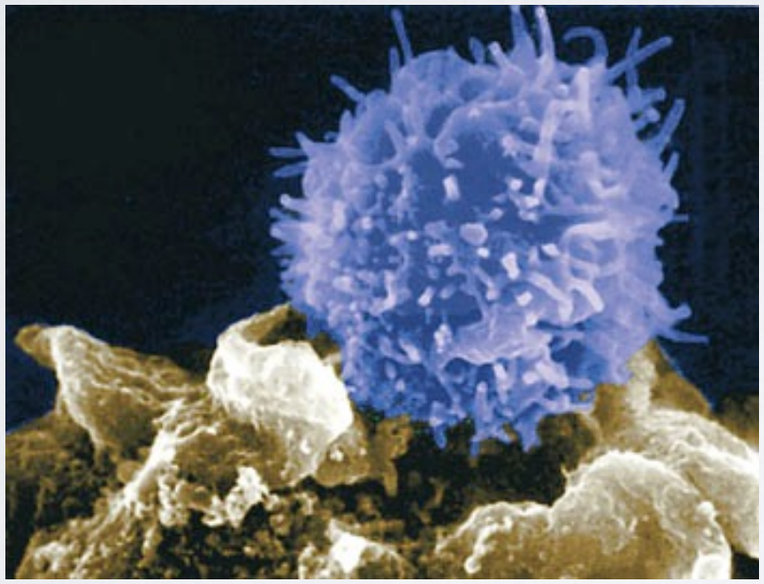


Es. Cytokine_leukocyte; hormone_cell

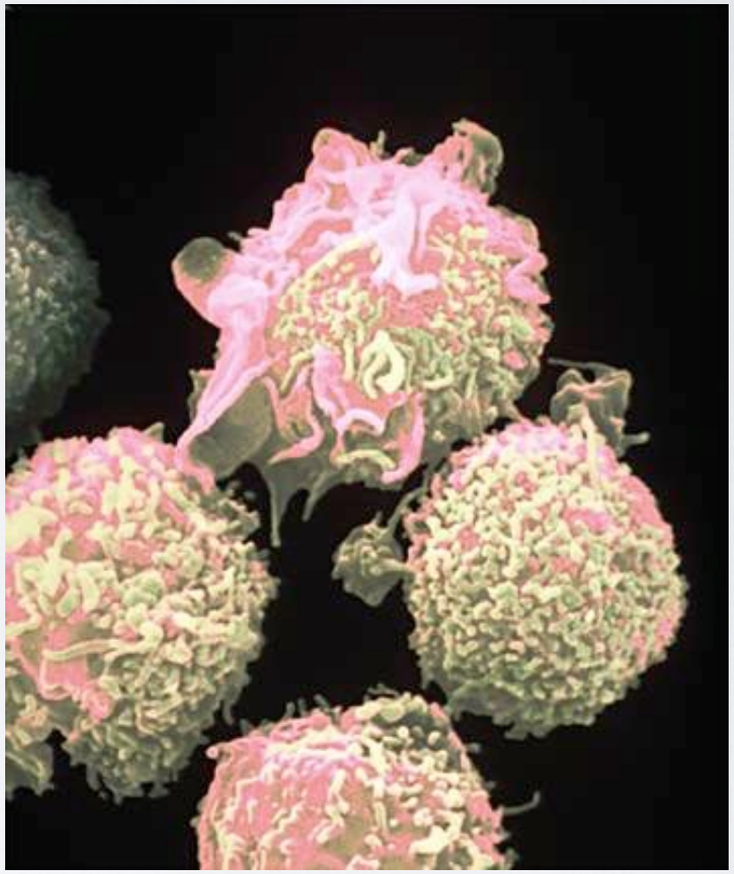
Standard flow-chart in Systems biology



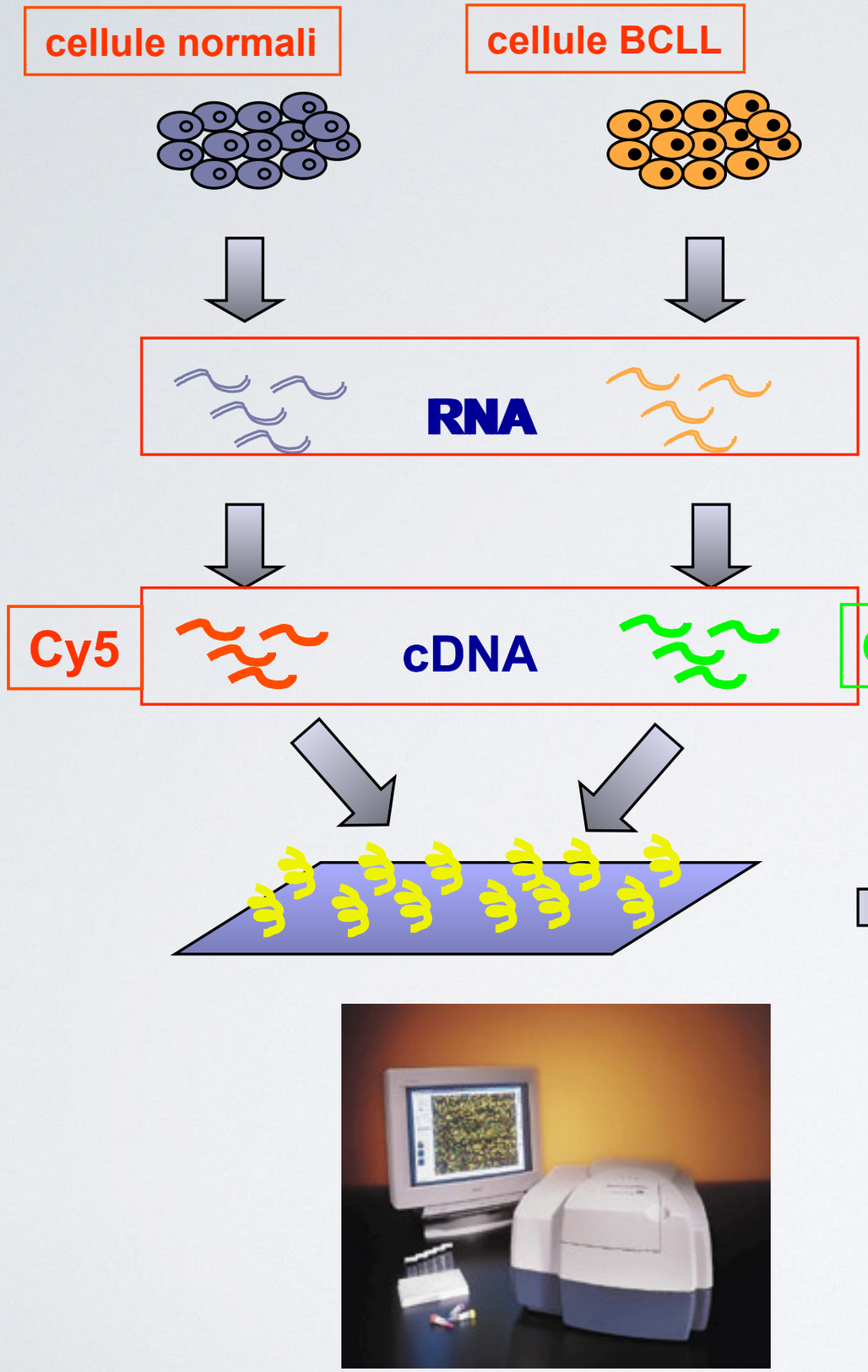
Example:
Differential analysis of gene expression



B-lymphocyte



CLL B-lymphocyte



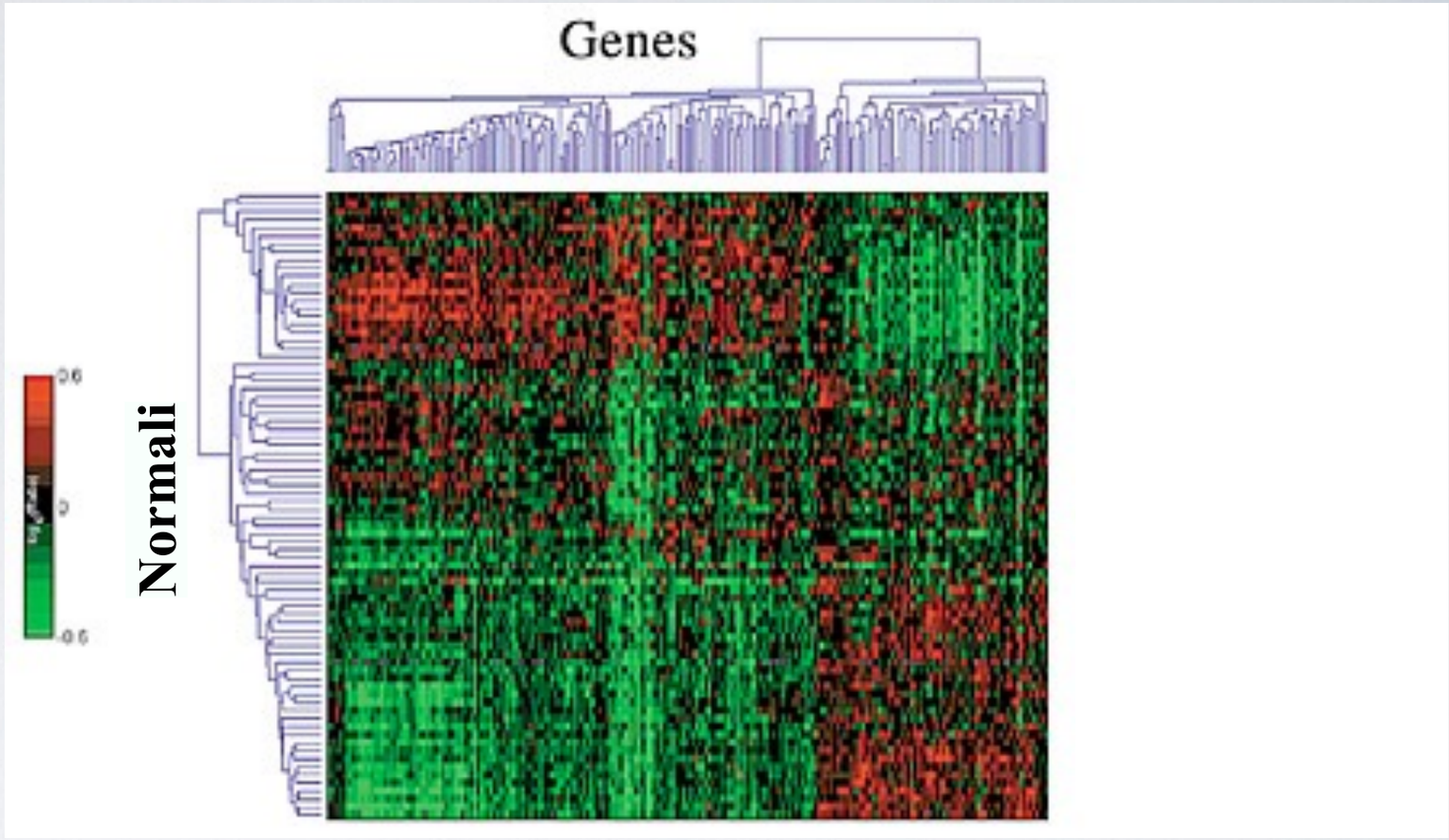
1) Estrazione dell'RNA totale dai campioni

2) Isolamento dell' mRNA, retrotrascrizione in cDNA e marcatura con fluorofori

3) Ibridizzazione

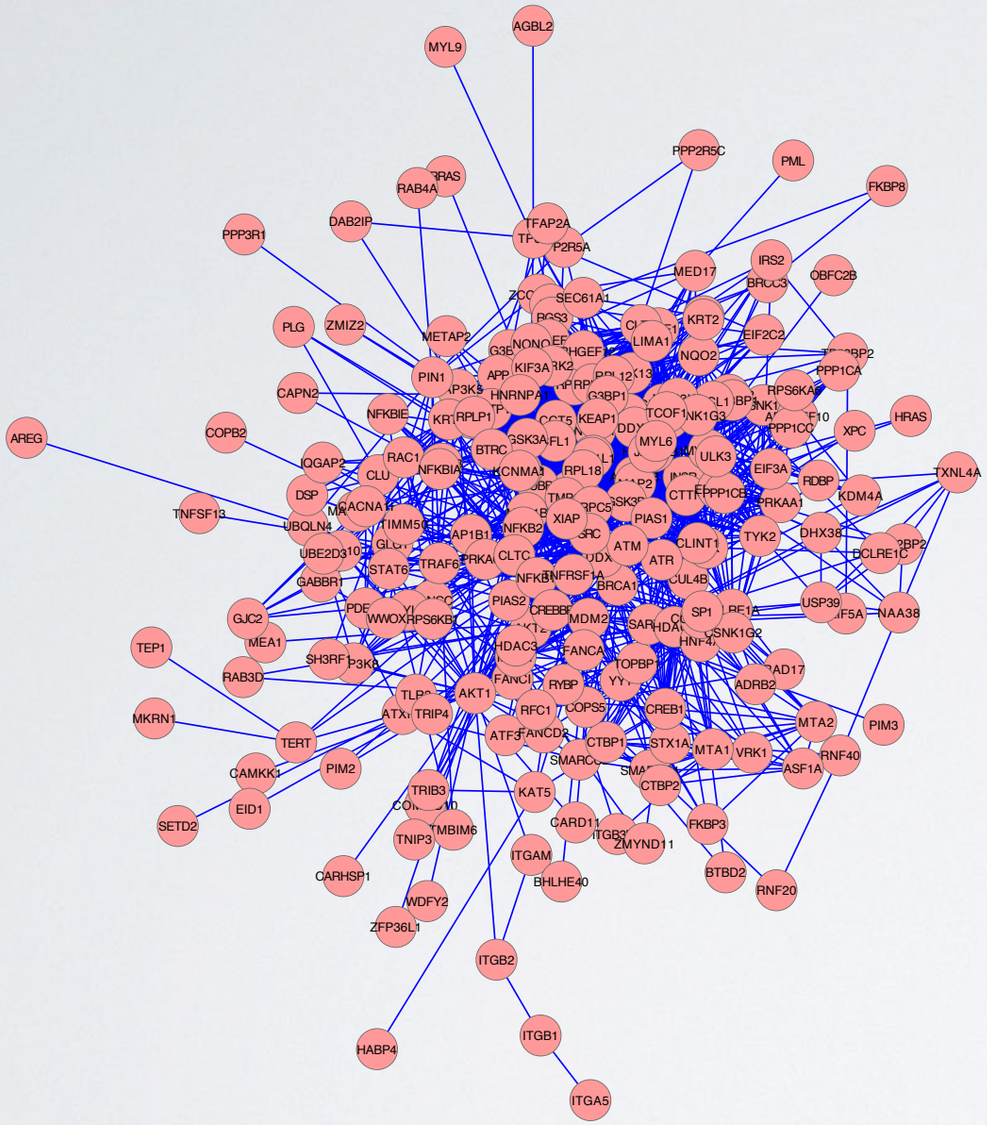
4) Scansione del vertino

Statistical clusters of altered genes

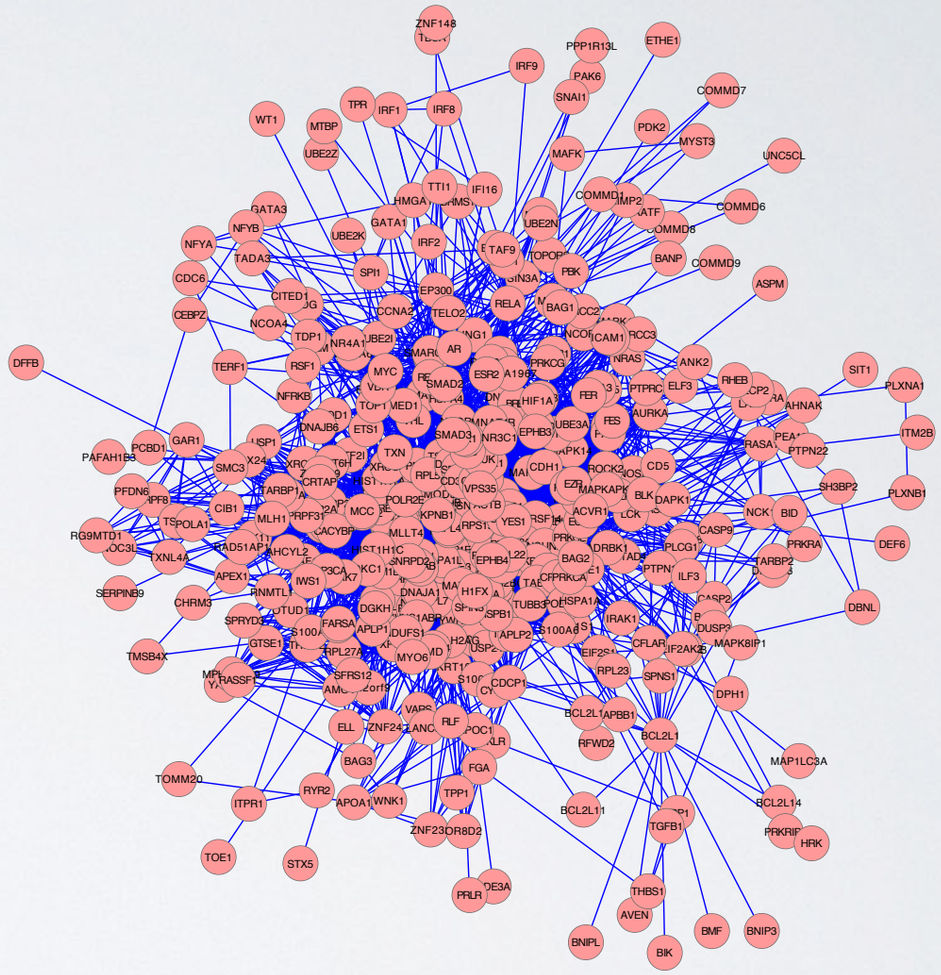


Gene network reconstruction

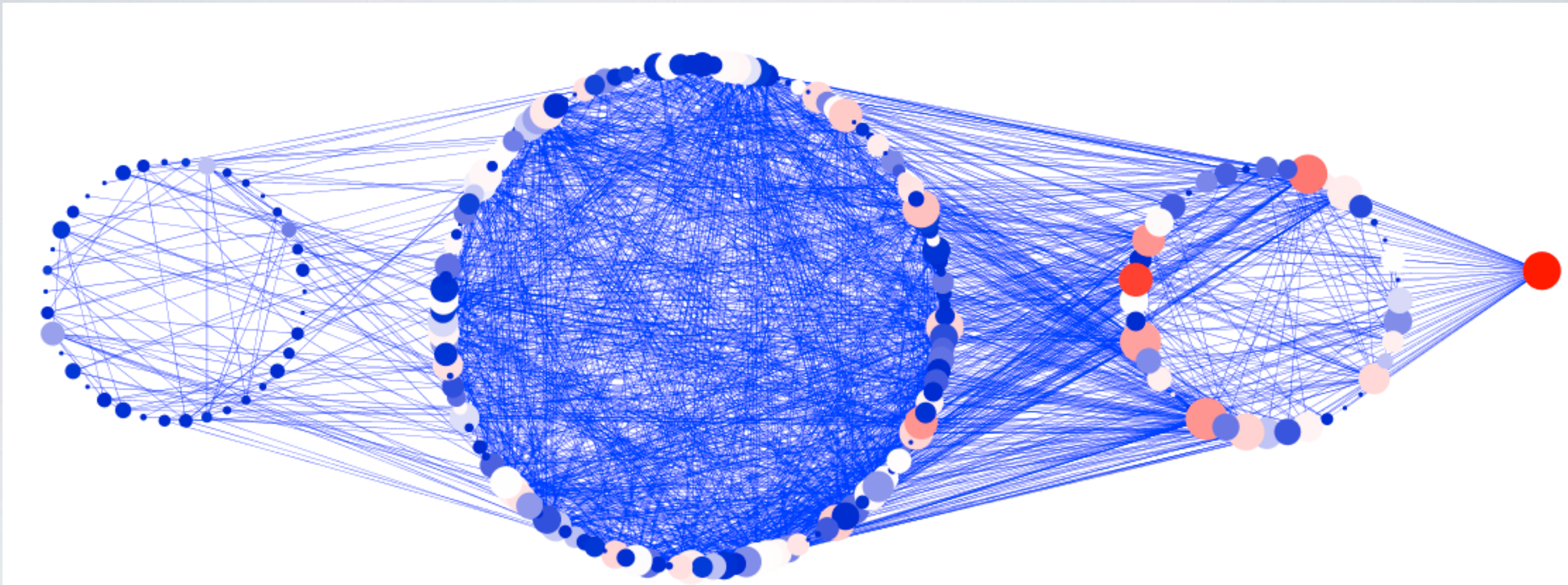
Normal gene network



B-CLL altered gene network

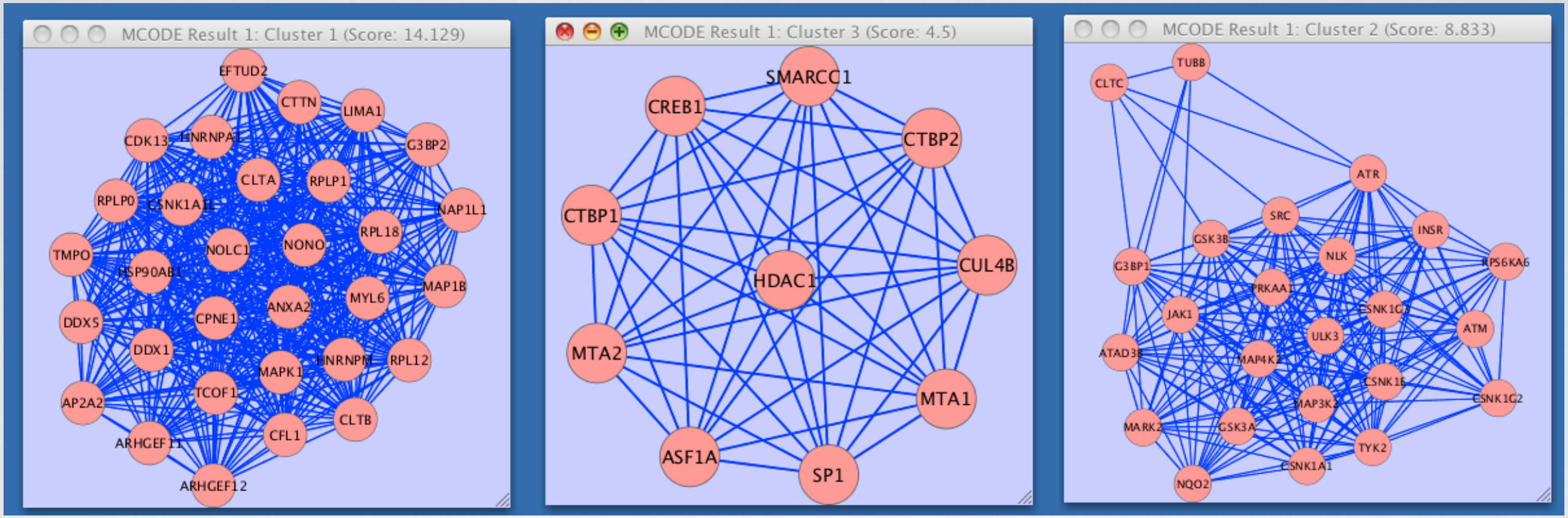


Identification of groups of altered genes
Differential analysis, intersection, etc.

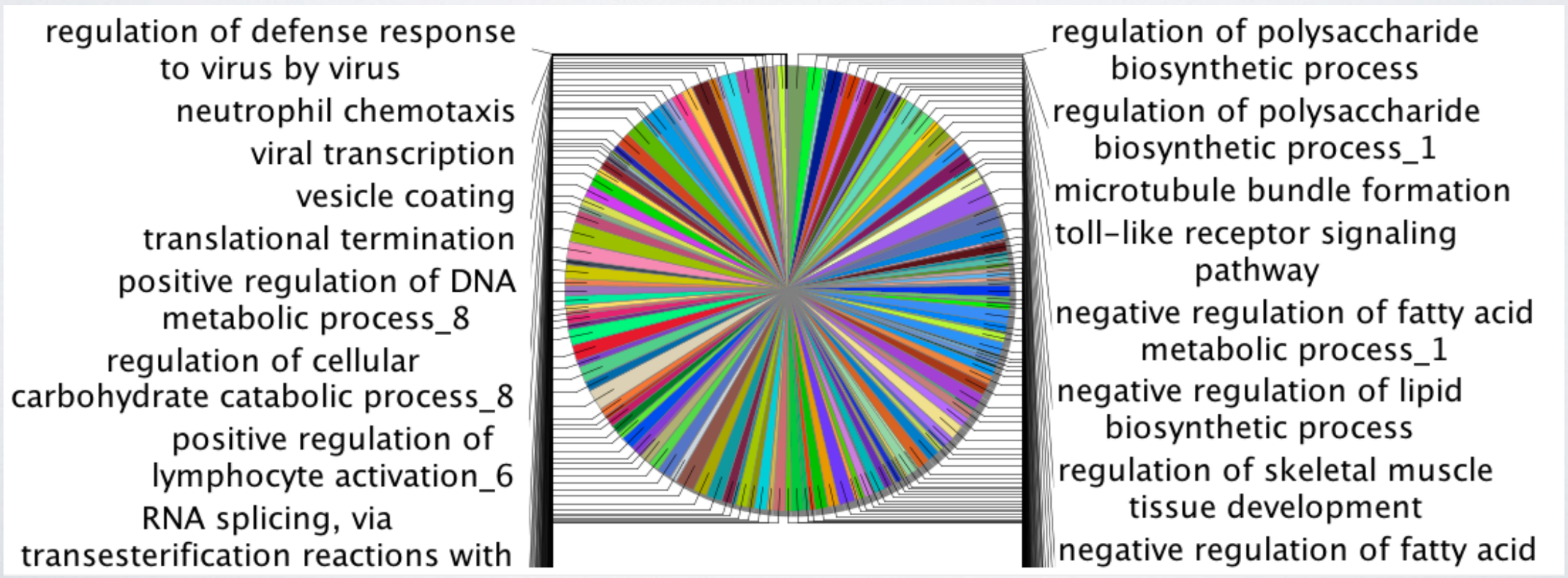


Identification of groups of altered genes
Differential analysis, intersection, modularisation, etc.

Modules of altered genes



Modules GO categories



Molecular Systems Biology 3; Article number 140; doi:10.1038/msb4100180
Citation: *Molecular Systems Biology* 3:140

www.molecularsystemsbiology.com

molecular systems biology

REPORT

Network-based classification of breast cancer metastasis

Han-Yu Chuang[1,5], Eunjung Lee[2,3,5], Yu-Tsueng Liu[4], Doheon Lee[3] and Trey Ideker[1,2,4,*]

[1] Bioinformatics Program, University of California San Diego, La Jolla, CA, USA, [2] Department of Bioengineering, University of California San Diego, La Jolla, CA, USA, [3] Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea and [4] Cancer Genetics Program, Moores Cancer Center, University of California San Diego, La Jolla, CA, USA
[5] These authors contributed equally to this work
* Corresponding author. Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA. Tel.: +1 858 822 4558; Fax: +1 858 534 5722; E-mail: trey@bioeng.ucsd.edu

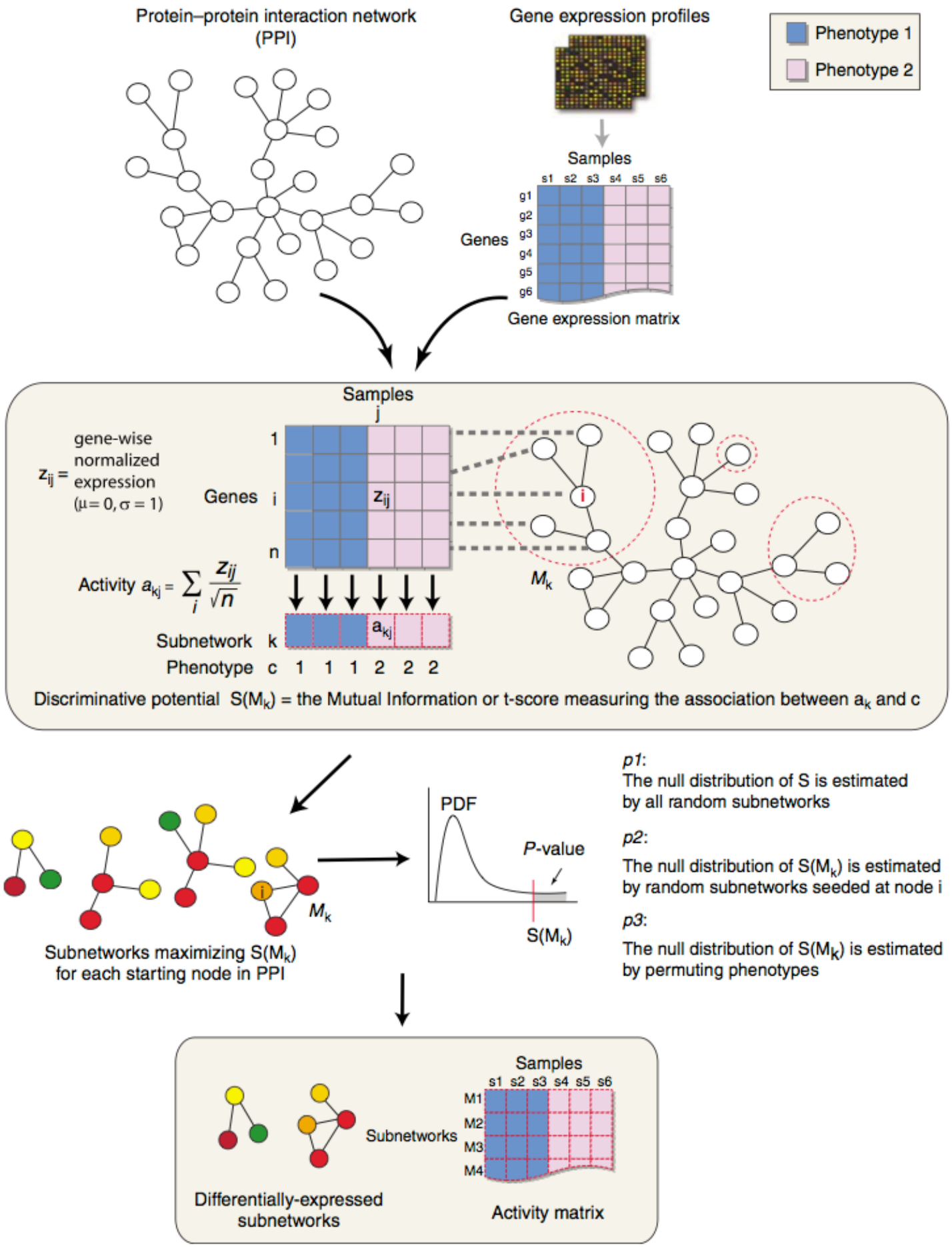
Received 11.6.07; accepted 20.8.07

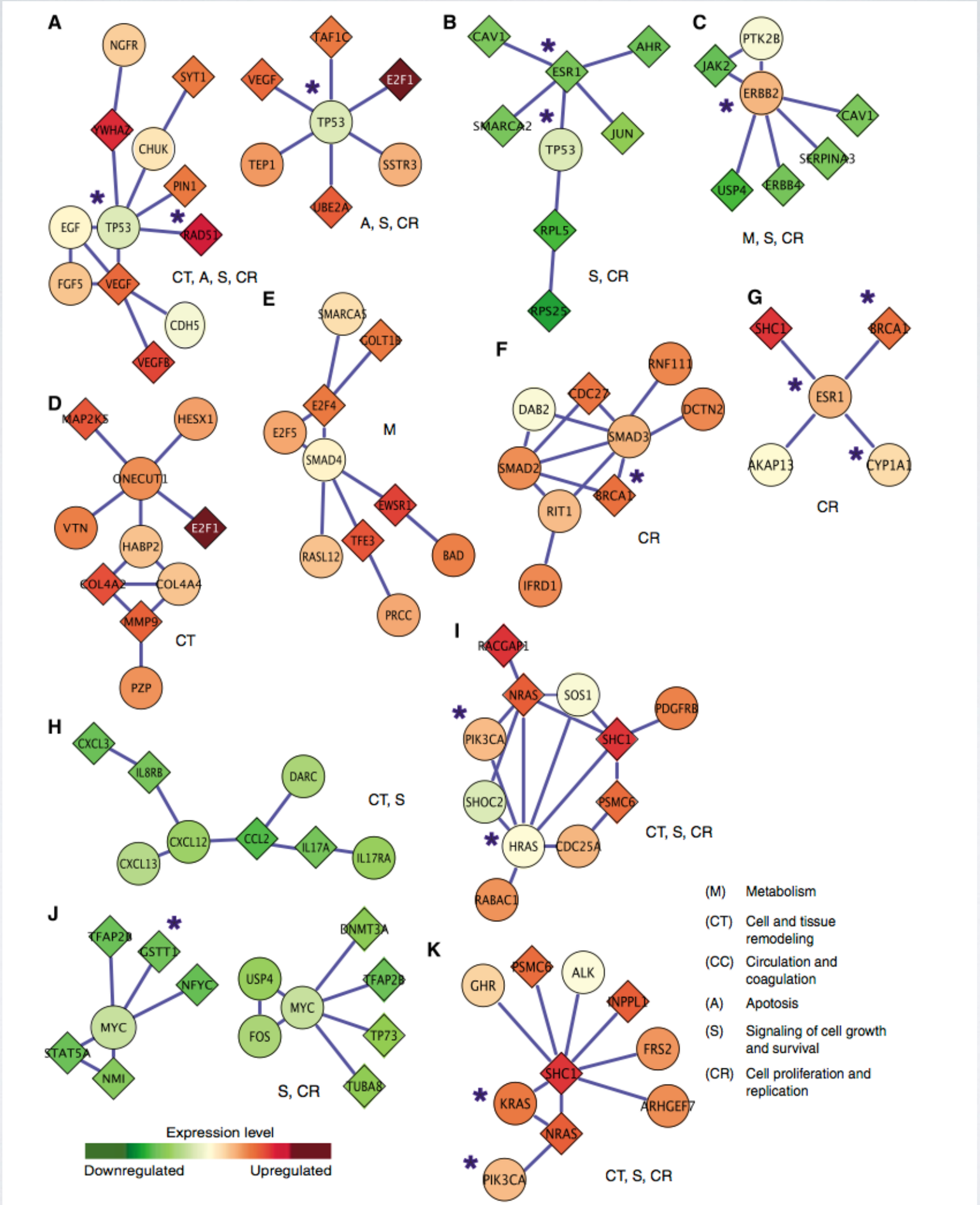
Mapping the pathways that give rise to metastasis is one of the key challenges of breast cancer research. Recently, several large-scale studies have shed light on this problem through analysis of gene expression profiles to identify markers correlated with metastasis. Here, we apply a protein-network-based approach that identifies markers not as individual genes but as subnetworks extracted from protein interaction databases. The resulting subnetworks provide novel hypotheses for pathways involved in tumor progression. Although genes with known breast cancer mutations are typically not detected through analysis of differential expression, they play a central role in the protein network by interconnecting many differentially expressed genes. We find that the subnetwork markers are more reproducible than individual marker genes selected without network information, and that they achieve higher accuracy in the classification of metastatic versus non-metastatic tumors.
Molecular Systems Biology 16 October 2007; doi:10.1038/msb4100180
Subject Categories: molecular biology of disease; metabolic and regulatory networks
Keywords: breast cancer metastasis; classification; protein networks; pathways; microarrays

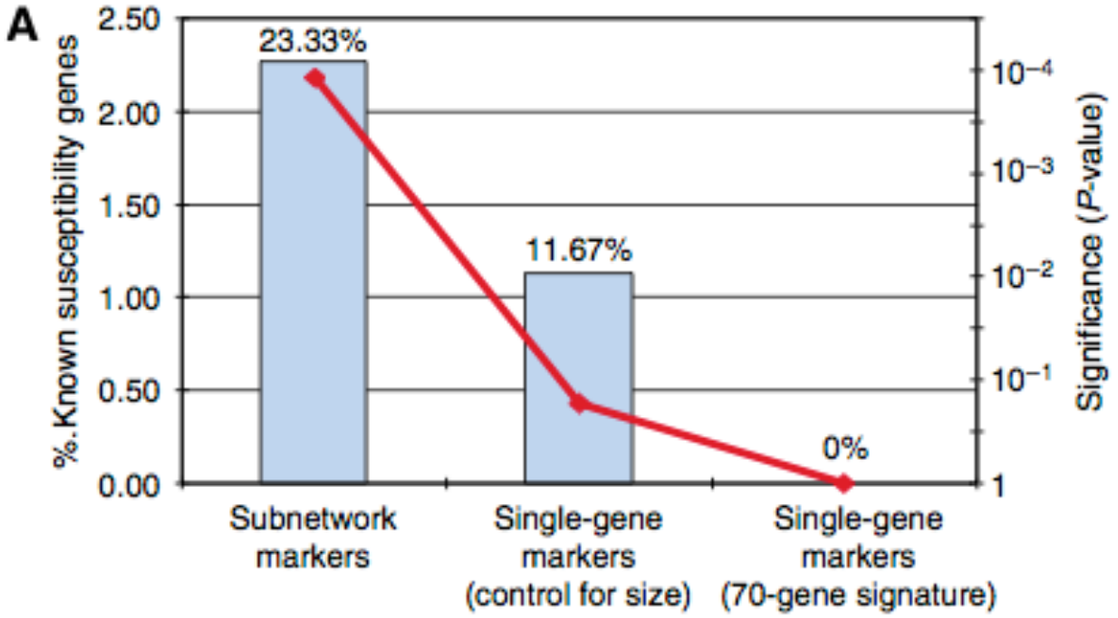
Introduction

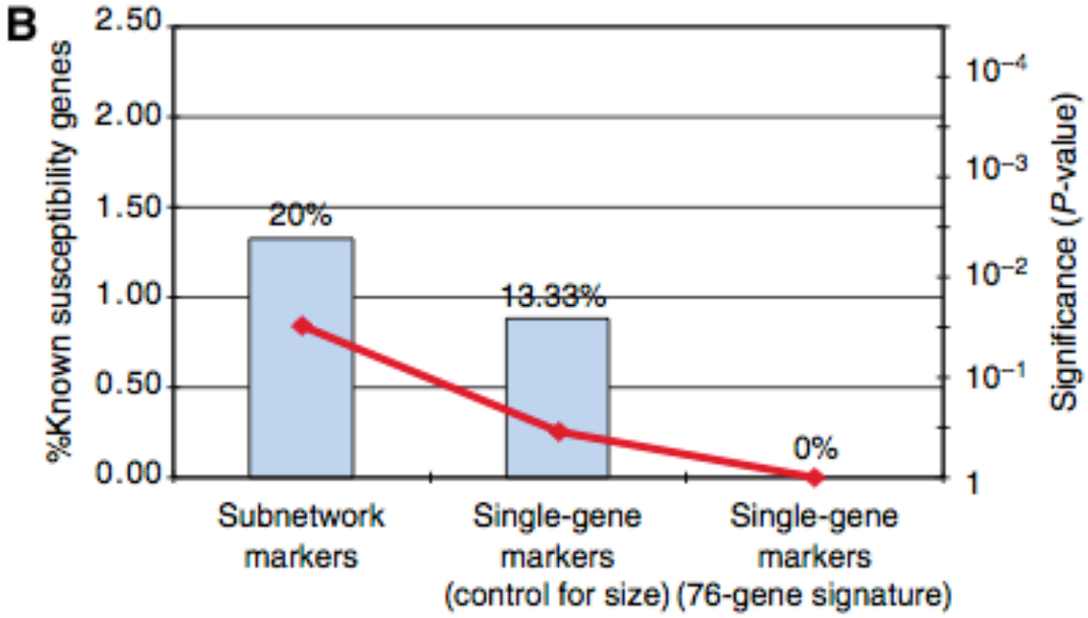
Distant metastases are the main cause of death among breast cancer patients (Weigelt *et al*, 2005). Clinical and pathological risk factors, such as patient age, tumor size, and steroid receptor status, are commonly used to assess the likelihood of metastasis development. When metastasis is likely, aggressive adjuvant therapy can be prescribed which has led to significant decreases in breast cancer mortality rates (Weigelt *et al*, 2005). However, for the majority of patients with intermediate-risk breast cancer, the traditional factors are not strongly predictive (Wang *et al*, 2005). Accordingly, approximately 70–80% of lymph node-negative patients may undergo adjuvant chemotherapy when it is in fact unnecessary (van 't Veer *et al*, 2002). Moreover, it is believed that many of the current risk factors are likely to be secondary manifestations rather than primary mechanisms of disease. An ongoing challenge is to identify new prognostic markers that are more directly related to disease and that can more accurately predict the risk of metastasis in individual patients.

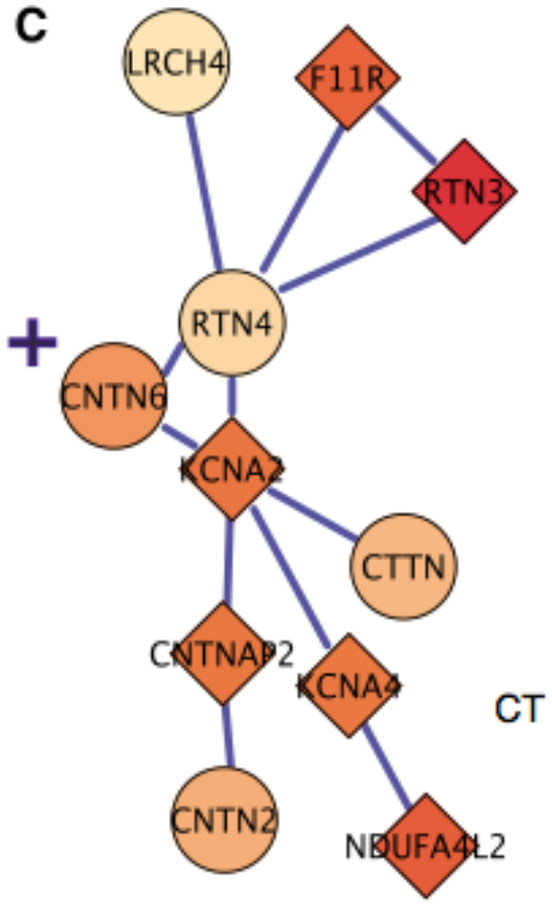
In the recent years, an increasing number of disease markers have been identified through analysis of genome-wide expression profiles (Golub *et al*, 1999; Alizadeh *et al*, 2000; Ben-Dor *et al*, 2000; Ramaswamy *et al*, 2003). Marker sets are selected by scoring each individual gene for how well its expression pattern can discriminate between different classes of disease. In breast cancer, two large-scale expression studies by van 't Veer *et al* (2002) and Wang *et al* (2005) each identified a set of ~70 gene markers that were 60–70% accurate for prediction of metastasis, rivaling the performance of clinical criteria. Strangely, however, these marker sets shared only three genes in common, with the first set of markers predicting metastasis less successfully when scoring patients from the second study, and vice versa (Ein-Dor *et al*, 2006). One possible explanation for the different marker sets is that changes in expression of the relatively few genes governing metastatic potential may be subtle compared to those of the downstream effectors, which may vary considerably from patient to patient (Symmans *et al*, 1995; Ein-Dor *et al*, 2005; Tomlins *et al*, 2005).

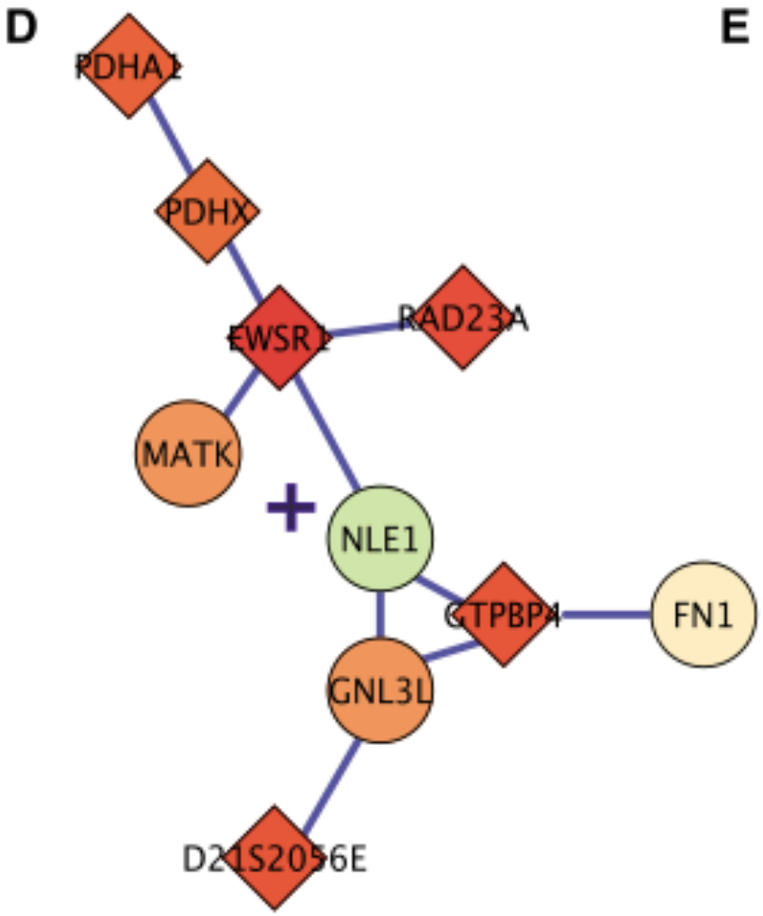

Protein–protein interaction network (PPI)
Gene expression profiles
Phenotype 1
Phenotype 2
Samples
s1 s2 s3 s4 s5 s6
Genes
g1
g2
g3
g4
g5
g6
Gene expression matrix
Samples
j
gene-wise normalized expression
($\mu = 0, \sigma = 1$)
z_{ij} =
Genes i
1
n
z_{ij}
Activity $a_{kj} = \sum_i \frac{z_{ij}}{\sqrt{n}}$
Subnetwork k
a_{kj}
Phenotype c 1 1 1 2 2 2
M_k
Discriminative potential S(M_k) = the Mutual Information or t-score measuring the association between a_k and c
Subnetworks maximizing S(M_k) for each starting node in PPI
PDF
P-value
S(M_k)
p1:
The null distribution of S is estimated by all random subnetworks
p2:
The null distribution of S(M_k) is estimated by random subnetworks seeded at node i
p3:
The null distribution of S(M_k) is estimated by permuting phenotypes
Samples
s1 s2 s3 s4 s5 s6
M1
M2
M3
M4
Subnetworks
Differentially-expressed subnetworks
Activity matrix

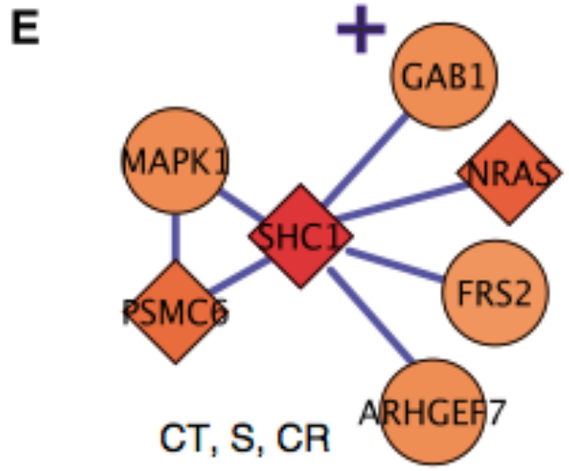

A
NGFR
SYT1
YWHAZ
CHUK
PIN1
EGF
TP53
RAD51
FGF5
VEGF
CDH5
VEGFB
CT, A, S, CR
TAF1C
VEGF
E2F1
TP53
TEP1
SSTR3
UBE2A
A, S, CR
B
CAV1
AHR
ESR1
SMARCA2
TP53
JUN
RPL5
RPS25
S, CR
C
PTK2B
JAK2
ERBB2
CAV1
SERPINA3
USP4
ERBB4
M, S, CR
D
MAP2K5
HESX1
ONECUT1
VTN
E2F1
HABP2
COL4A2
COL4A4
MMP9
PZP
CT
E
SMARCA5
COLT1B
E2F4
E2F5
M
SMAD4
EWSR1
RASL12
TFE3
BAD
PRCC
F
RNF111
CDC27
DAB2
DCTN2
SMAD3
SMAD2
BRCA1
RIT1
IFRD1
CR
G
SHC1
BRCA1
ESR1
AKAP13
CYP1A1
CR
H
CXCL3
IL8RB
DARC
CXCL12
CCL2
IL17A
IL17RA
CXCL13
CT, S
I
RACGAP1
NRAS
SOS1
PDGFRB
PIK3CA
SHC1
SHOC2
PSMC6
HRAS
CDC25A
RABAC1
CT, S, CR
J
TFAP2B
GSTT1
NFYC
MYC
STAT5A
NMI
USP4
DNMT3A
MYC
TFAP2B
FOS
TP73
TUBA8
S, CR
K
GHR
PSMC6
ALK
NPPL1
FRS2
SHC1
KRAS
ARHGEF7
NRAS
PIK3CA
CT, S, CR
Expression level
Downregulated
Upregulated
(M) Metabolism
(CT) Cell and tissue remodeling
(CC) Circulation and coagulation
(A) Apotosis
(S) Signaling of cell growth and survival
(CR) Cell proliferation and replication

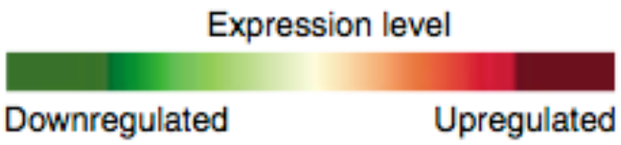












(M) Metabolism

(CT) Cell and tissue remodeling

(CC) Circulation and coagulation

(A) Apotosis

(S) Signaling of cell growth and survival

(CR) Cell proliferation and replication

Molecular Systems Biology 3; Article number 152; doi:10.1038/msb4100200
Citation: *Molecular Systems Biology* 3:152

www.molecularsystemsbiology.com

molecular systems biology

A map of human cancer signaling

Qinghua Cui[1], Yun Ma[2], Maria Jaramillo[3], Hamza Bari[1], Arif Awan[1], Song Yang[4], Simo Zhang[2], Lixue Liu[2], Meng Lu[2], Maureen O'Connor-McCourt[3], Enrico O Purisima[1,5] and Edwin Wang[1,5,*]

[1] Computational Chemistry and Biology Group, Biotechnology Research Institute, National Research Council Canada, Montreal, QC, Canada, [2] Department of Biology, Tianjin Normal University, Tianjin, China, [3] Receptor, Signaling and Proteomics Group, Biotechnology Research Institute, National Research Council Canada, Montreal, QC, Canada, [4] School of Chemical Engineering, Tianjin University, Tianjin, China and [5] Center for Bioinformatics, McGill University, Montreal, QC, Canada
* Corresponding author. Computational Chemistry and Biology Group, Biotechnology Research Institute, National Research Council Canada, 6100 Royalmount, Montreal, QC, Canada H4P 2R2. Tel.: +1 514 496 0914; Fax: +1 514 496 0943; E-mail: edwin.wang@cnrc-nrc.gc.ca

Received 8.6.07; accepted 6.11.07

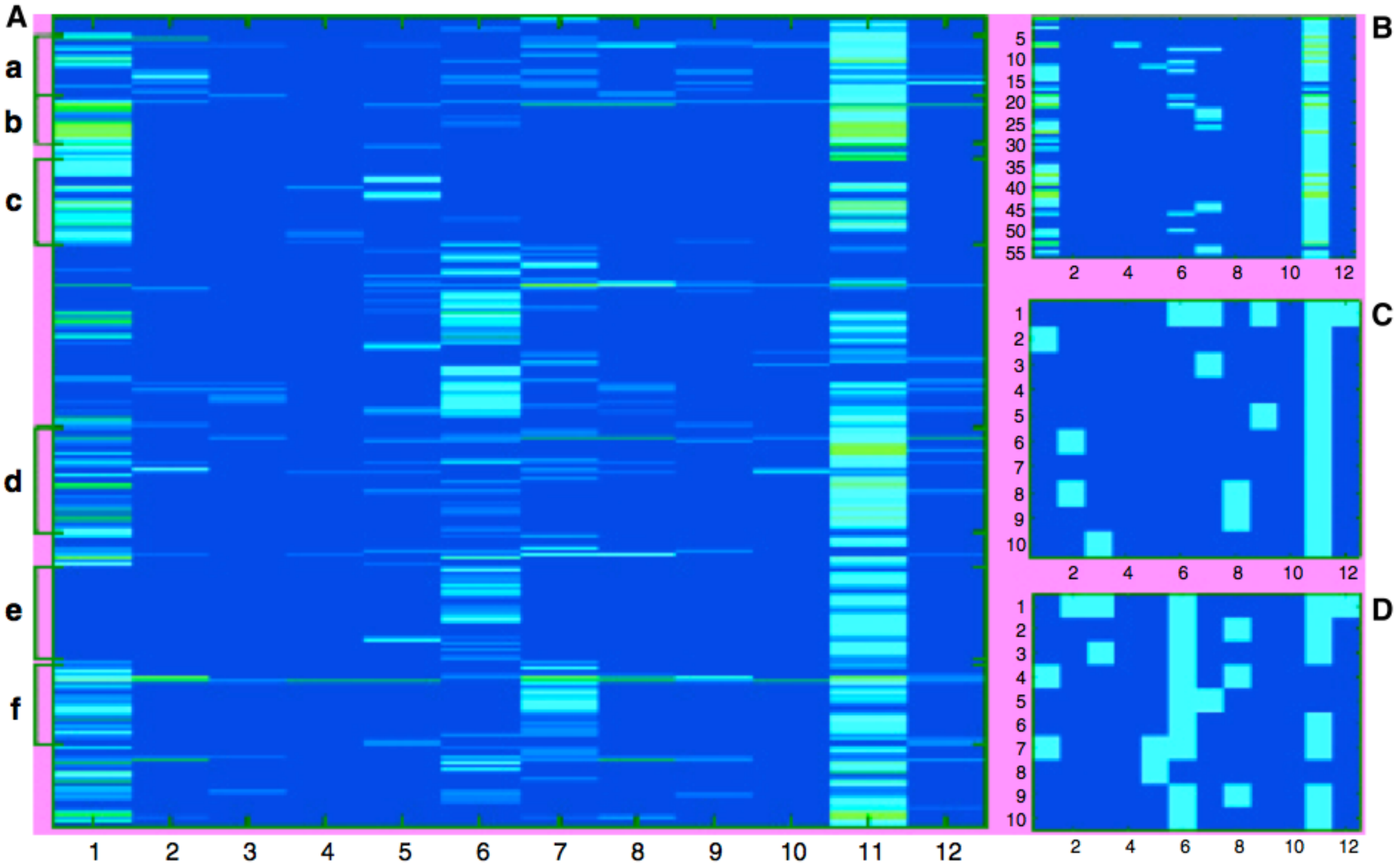
We conducted a comprehensive analysis of a manually curated human signaling network containing 1634 nodes and 5089 signaling regulatory relations by integrating cancer-associated genetically and epigenetically altered genes. We find that cancer mutated genes are enriched in positive signaling regulatory loops, whereas the cancer-associated methylated genes are enriched in negative signaling regulatory loops. We further characterized an overall picture of the cancer-signaling architectural and functional organization. From the network, we extracted an oncogene-signaling map, which contains 326 nodes, 892 links and the interconnections of mutated and methylated genes. The map can be decomposed into 12 topological regions or oncogene-signaling blocks, including a few 'oncogene-signaling-dependent blocks' in which frequently used oncogene-signaling events are enriched. One such block, in which the genes are highly mutated and methylated, appears in most tumors and thus plays a central role in cancer signaling. Functional collaborations between two oncogene-signaling-dependent blocks occur in most tumors, although breast and lung tumors exhibit more complex collaborative patterns between multiple blocks than other cancer types. Benchmarking two data sets derived from systematic screening of mutations in tumors further reinforced our findings that, although the mutations are tremendously diverse and complex at the gene level, clear patterns of oncogene-signaling collaborations emerge recurrently at the network level. Finally, the mutated genes in the network could be used to discover novel cancer-associated genes and biomarkers.

Molecular Systems Biology 18 December 2007; doi:10.1038/msb4100200
Subject Categories: metabolic and regulatory networks; molecular biology of disease
Keywords: cancer signaling; cancer-signaling map; DNA methylation; gene mutation; oncogene-signaling dependence; signaling network; tumorigenesis

Introduction

Cells use sophisticated communication between proteins in order to initiate and maintain basic cellular functions such as growth, survival, proliferation and development. Traditionally, cell signaling is described via linear diagrams and signaling pathways. As many more 'cross-talks' between signaling pathways have been identified (Natarajan *et al*, 2006), a network view of cell signaling emerged: the signaling proteins rarely operate in isolation through linear pathways, but rather through a large and complex network. As cell signaling is crucial to affect cell responses such as growth and survival, alterations of cellular signaling events, such as those that arise by mutations, can result in tumor development. Indeed, cancer is largely a genetic disease that is caused by acquiring genomic alterations in somatic cells. Alterations to the genes that encode key signaling proteins, such as RAS and PI3K, are commonly observed in many types of cancers. During tumor progression, it is proposed that a malignant tumor arises from a single cell, which undergoes a series of evolutionary processes of genetic or epigenetic changes and selections so that a cell within the population can acquire additional selective advantages for cellular growth or survival, resulting in progressive clonal expansion (Nowell, 1976).

Genetic mutations of the signaling proteins might over-activate key cell-signaling properties such as cell proliferation or survival and then give rise to the cell with selective advantages for uncontrolled cellular growth and promoting tumor progression. In addition, mutations may also inhibit the function of tumor-suppressor proteins, resulting in a relief from normal constraints on growth. Furthermore, epigenetic alterations by promoter methylation, resulting in transcriptional


A
a
b
c
d
e
f
1
2
3
4
5
6
7
8
9
10
11
12
B
5
10
15
20
25
30
35
40
45
50
55
2
4
6
8
10
12
C
1
2
3
4
5
6
7
8
9
10
2
4
6
8
10
12
D
1
2
3
4
5
6
7
8
9
10
2
4
6
8
10
12

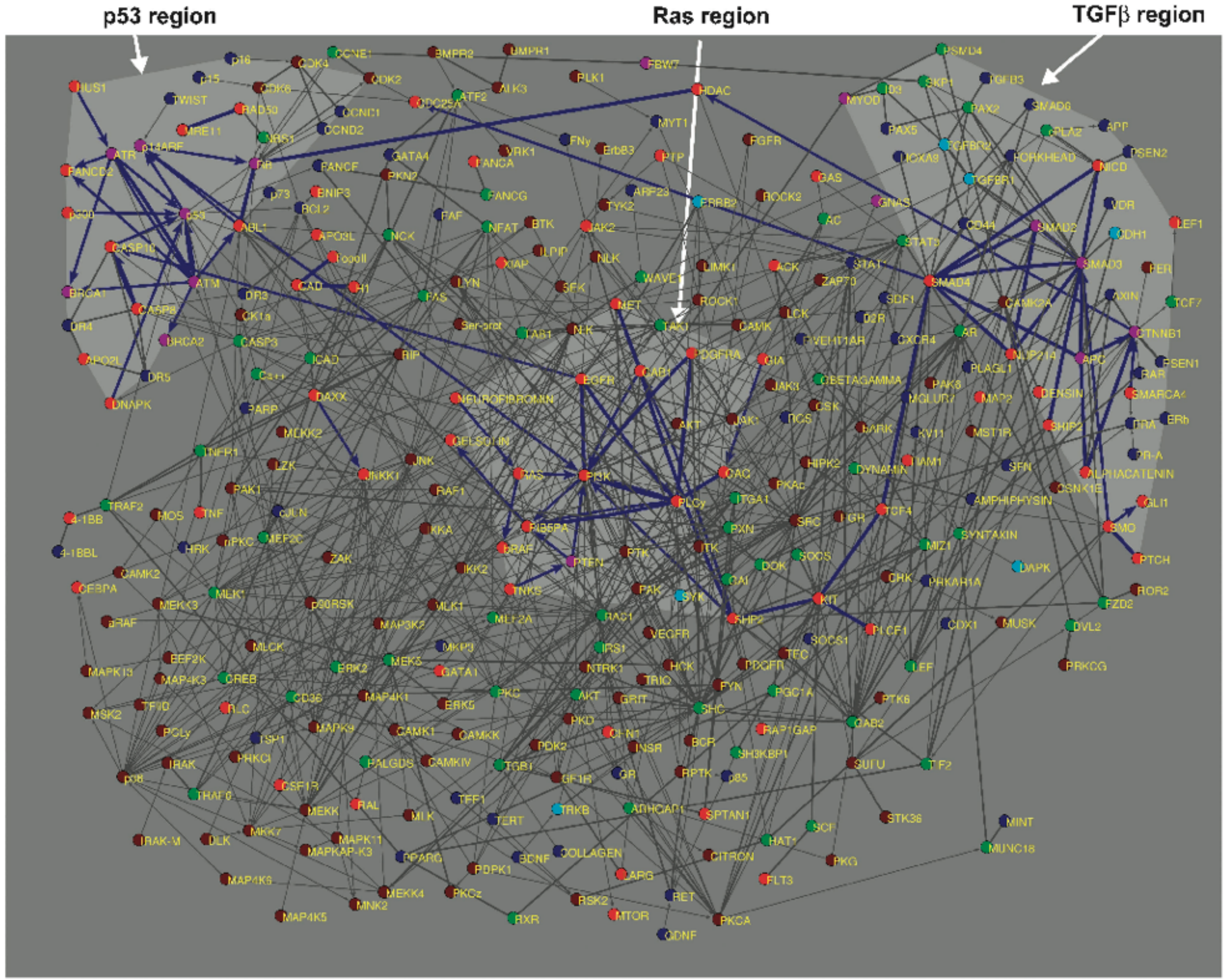


Figure 3 Human oncogene-signaling map. The human cancer-signaling map was extracted from the human signaling network, which was mapped with cancer mutated and methylated genes. The map shows three 'oncogenic-dependent regions' (background in light gray), in which genes of the two regions are also heavily methylated. Nodes represent genes, whereas the links with and without arrows represent signal and physical relations, respectively. Nodes in red, purple, brown, cyan, blue and green represent the genes that are highly mutated but not methylated, both highly mutated and methylated, poorly mutated but not methylated, both poorly mutated and methylated, methylated but not mutated, and neither mutated nor methylated, respectively.

Briefings in Bioinformatics, 2017, 1–15

doi: 10.1093/bib/bbx144
Paper

Precision medicine needs pioneering clinical bioinformaticians

Gonzalo Gómez-López, Joaquín Dopazo, Juan C. Cigudosa, Alfonso Valencia and Fátima Al-Shahrour

Corresponding author. Fátima Al-Shahrour, Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro, 3. ES-28029 Madrid, Spain. Tel.: +34 91 732 80 00; E-mail: falshahrour@cnio.es

Abstract

Success in precision medicine depends on accessing high-quality genetic and molecular data from large, well-annotated patient cohorts that couple biological samples to comprehensive clinical data, which in conjunction can lead to effective therapies. From such a scenario emerges the need for a new professional profile, an expert bioinformatician with training in clinical areas who can make sense of multi-omics data to improve therapeutic interventions in patients, and the design of optimized basket trials. In this review, we first describe the main policies and international initiatives that focus on precision medicine. Secondly, we review the currently ongoing clinical trials in precision medicine, introducing the concept of 'precision bioinformatics', and we describe current pioneering bioinformatics efforts aimed at implementing tools and computational infrastructures for precision medicine in health institutions around the world. Thirdly, we discuss the challenges related to the clinical training of bioinformaticians, and the urgent need for computational specialists capable of assimilating medical terminologies and protocols to address real clinical questions. We also propose some skills required to carry out common tasks in clinical bioinformatics and some tips for emergent groups. Finally, we explore the future perspectives and the challenges faced by precision medicine bioinformatics.

Key words: precision medicine; computing infrastructures; clinical bioinformatics; training; clinical bioinformatician; genomic report

Precision medicine in the real world: the dress rehearsals

The paradigm of precision medicine is defined by combining the use of population-based molecular profiling, clinical data, epidemiological information and other types of data to make clinical decisions that are tailored to individual patients [1]. The potential advantages of this approach, both for patients and doctors, include more accurate diagnosis and treatments, safer drug prescription, better disease prevention and consequently, a reduction in healthcare costs. The integration of genomics into routine clinical practice requires systems and workforces that are equipped and prepared to handle the scale and complexity of genomic data. As such, bioinformatics plays an essential role in

Gonzalo Gómez-López is a senior computational biologist at the Spanish National Cancer Research Centre, and assistant professor of translational genomics and bioinformatics at the School of Medicine of the Autónoma University (Madrid). He is interested in cancer genome interpretation, transcriptomics and immunotherapy.
Joaquín Dopazo is the director of the Clinical Bioinformatics Area of the Fundación Progreso y Salud (Seville). His research focuses on translational genomic data analysis for precision medicine, with an emphasis on systems medicine applications.
Juan Cruz Cigudosa is a clinical geneticist with experience in genomics for research and clinical applications. He is the current president of the Spanish Human Genetics Association and scientific director of NIMGenetics.
Alfonso Valencia is the director of the Spanish National Bioinformatics Institute (INB) and the Life Sciences department at the Barcelona Supercomputing Centre. His main research goals are to analyse the function and interactions of cancer-related proteins, and to develop novel computational methods to examine, represent and interpret cancer genome information.
Fátima Al-Shahrour is head of the Bioinformatics Unit at Spanish National Cancer Research Centre. She is applying computational methods to precision medicine to interpret cancer genomes, for drug repositioning and for the prediction of anticancer therapies.
Submitted: 17 July 2017; **Received (in revised form):** 14 September 2017

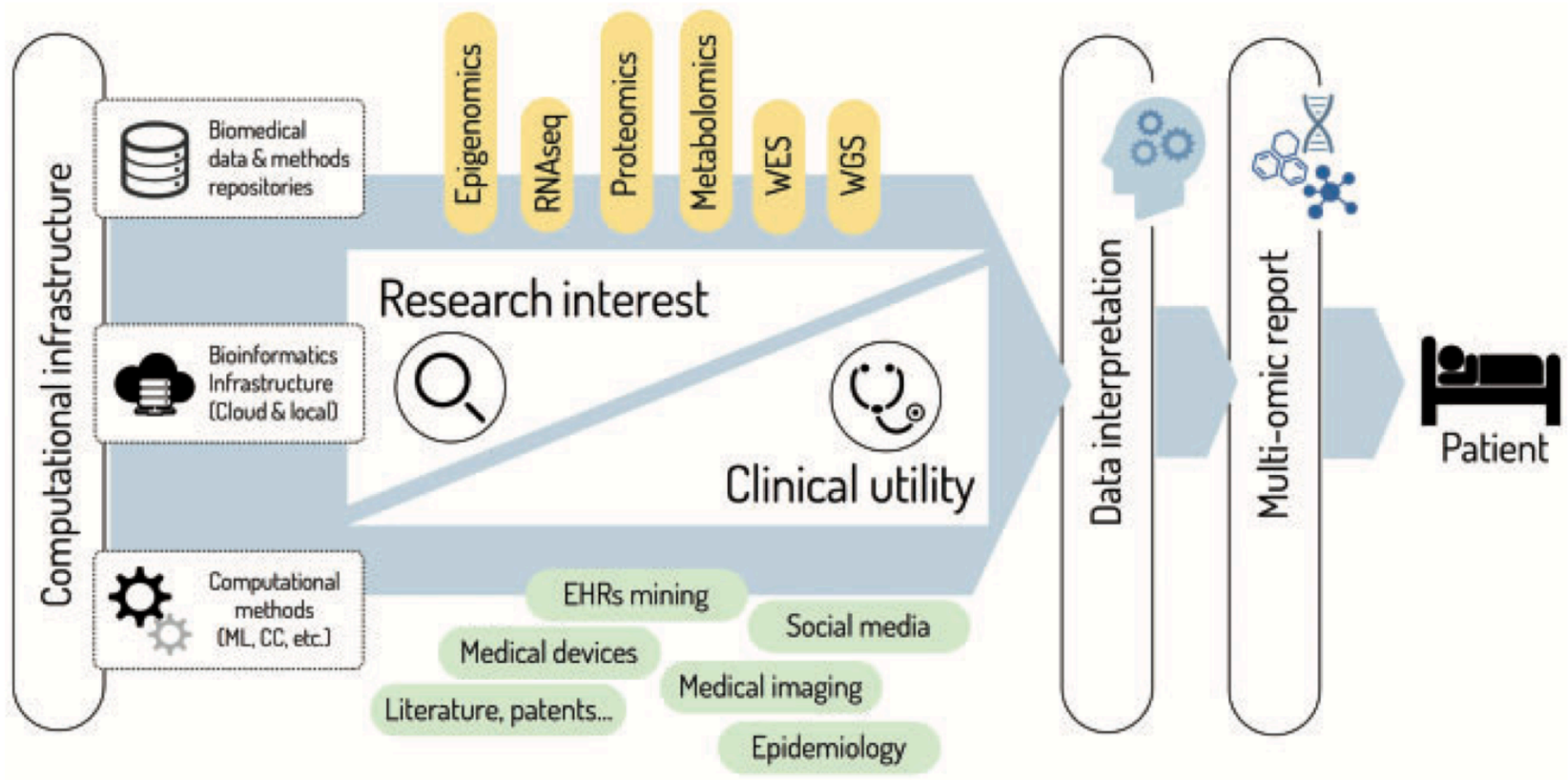


Figure 1. Precision medicine workflow: from data to patient care. Precision medicine requires computational infrastructures to efficiently store and process data on patient genotypes and phenotypes. The biological and clinical interpretation of such data is converted into an integral multi-omics report that will support clinical decision-making: ML, Machine Learning; CC, Cognitive Computing; EHRs, Electronic Health Records.

Box 1. Fundamental technical skills for clinical bioinformaticians

1. Informatics
○ Experience in UNIX command line.
○ A basic programming language (i.e. Python). R as a useful language for handling statistics.
○ Knowledge of big data environments.
2. Life sciences
○ Understand the different types of biological data and databases.
○ Comprehend HTP data analysis methods.
○ Multi-omics data integration and interpretation.
3. Clinical scenario
○ Be familiar with EHRs, clinical terminology and medical procedures and protocols.
○ Get to know medical genomics: diagnosis, predictive and prognosis biomarkers.
○ Understand clinical trial design and monitoring.

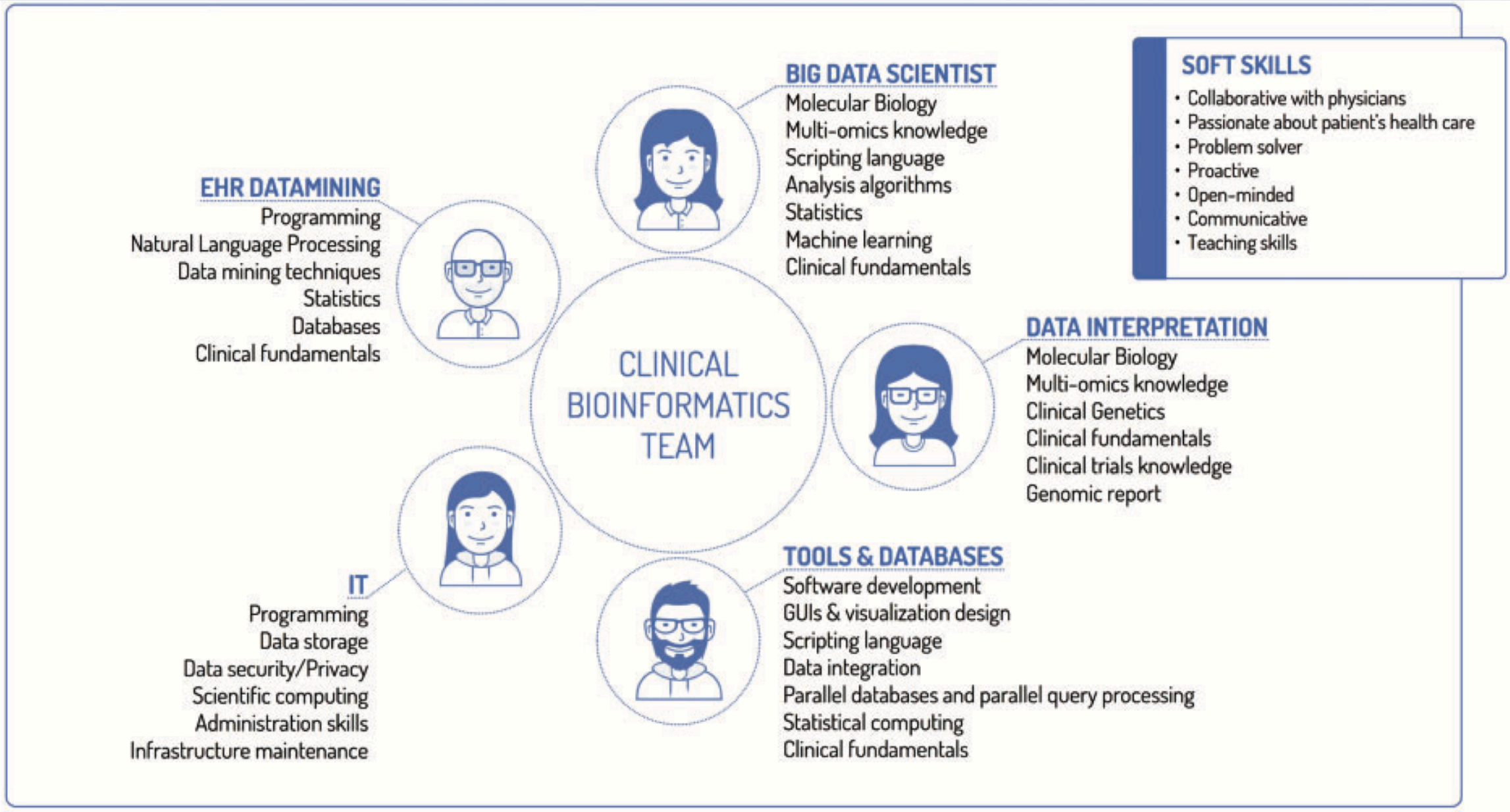


Figure 2. Clinical bioinformatics laboratory profile. Clinical bioinformatics teams require multidisciplinary experts to perform regular tasks. Computational experts will be in charge of servers, databases, development and pipeline optimization. More biologically focused profiles will be responsible for genomic analysis and interpretation. In addition, all team members will have been trained to a varying degree in basic clinical sciences. There will be a continued knowledge exchange among the team members to achieve clinical goals and improve patient healthcare.

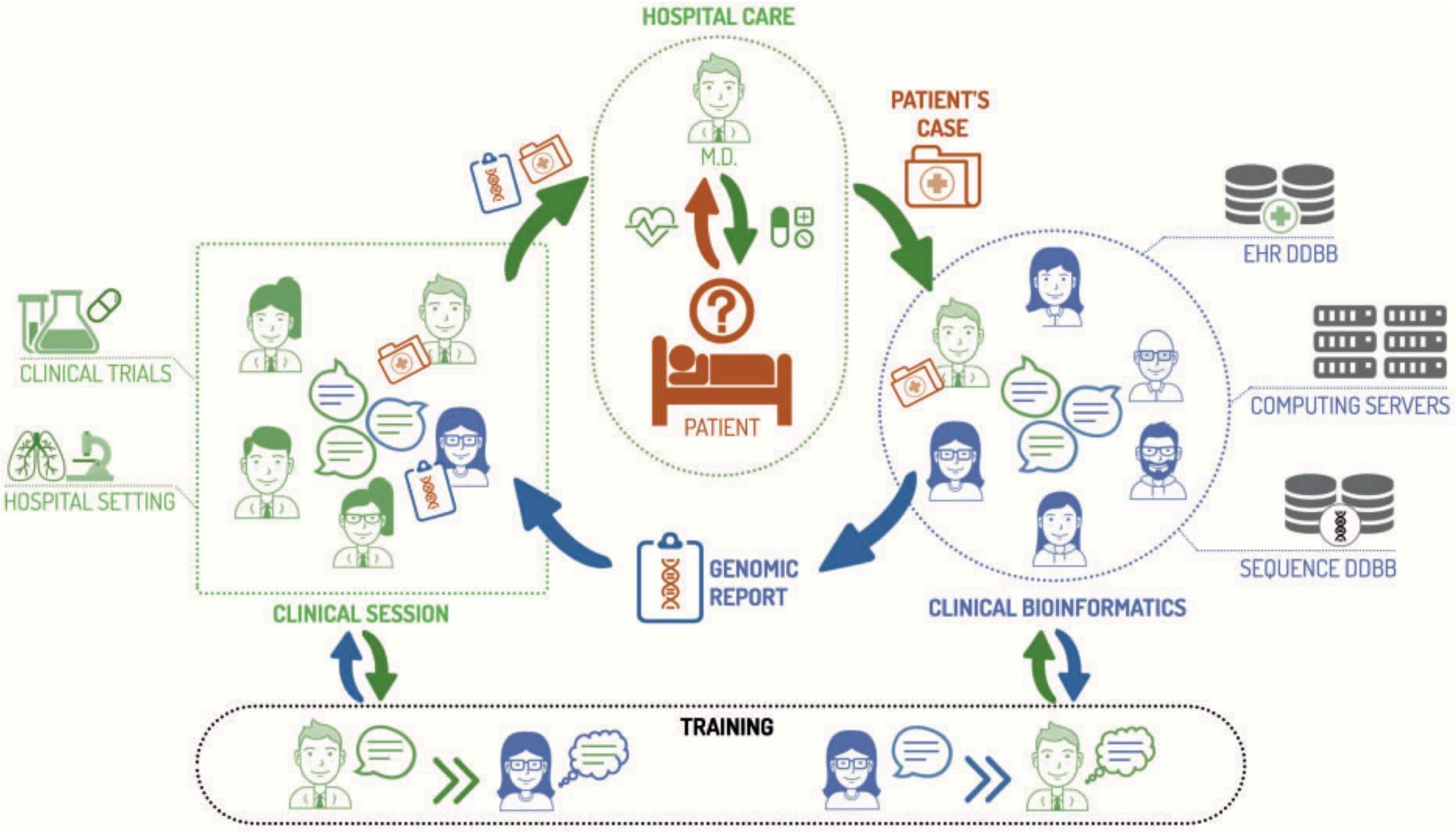


Figure 3. Precision medicine workflow in hospitals. The patient's standard of care in a precision medicine scenario requires specialist clinical bioinformaticians to participate in multidisciplinary clinical sessions with physicians and clinical geneticists, making knowledge exchange easier and facilitating data-driven diagnosis and clinical decision-making. Conversely, physicians and clinical geneticists should be incorporated into clinical bioinformatics discussions to guide and adapt genomic reports to the medical reality. Clinical bioinformaticians, in hospitals should not be simple data extraction technicians but rather, specialized partners of physicians in accordance with a patient-centred model. Mutual training for clinicians and bioinformaticians should provide reciprocal and bidirectional information exchange.

Key Points

- The precision medicine initiatives emerging around the world are facing many challenges.
- Current public and private investments aim to establish the computational infrastructures required to support precision medicine initiatives.
- Precision medicine and clinical bioinformatics can only work efficiently if electronic health records and patient genotypes are accessible to in-house bioinformaticians.
- The successful implementation of precision medicine in health institutions requires bioinformaticians with a basic clinical training, as yet an unfulfilled need.
- The clinical bioinformatician is a novel and specialized profile demanded increasingly by healthcare centres.

Signal transduction mechanisms: What are they?

A signal transduction system is an intracellular biochemical mechanism which processes **environmental information** and translates it into a molecular language that the cell is able to understand.

Signalosome (interactome):

~ 5000 proteins (?)

~ 300,000 interactions (and counting ...)

Signal transduction systems are the most critical biochemical mechanism in all organisms since the determine how a living will adapt to the environments, evolve, duplicate and mutate.

They are also called the “black box” of 21th century medicine.

The functional sequence of signal transduction

Agonista (1° segnale)

Recettore

Small GTP binding proteins

OFF ↔ ON

Tyrosin chinasi

G-protein trimerica

Transduttore

Interruttori (switchers)

Amplificatore

PLC, AC, PI(3)K
PLD, PLA2

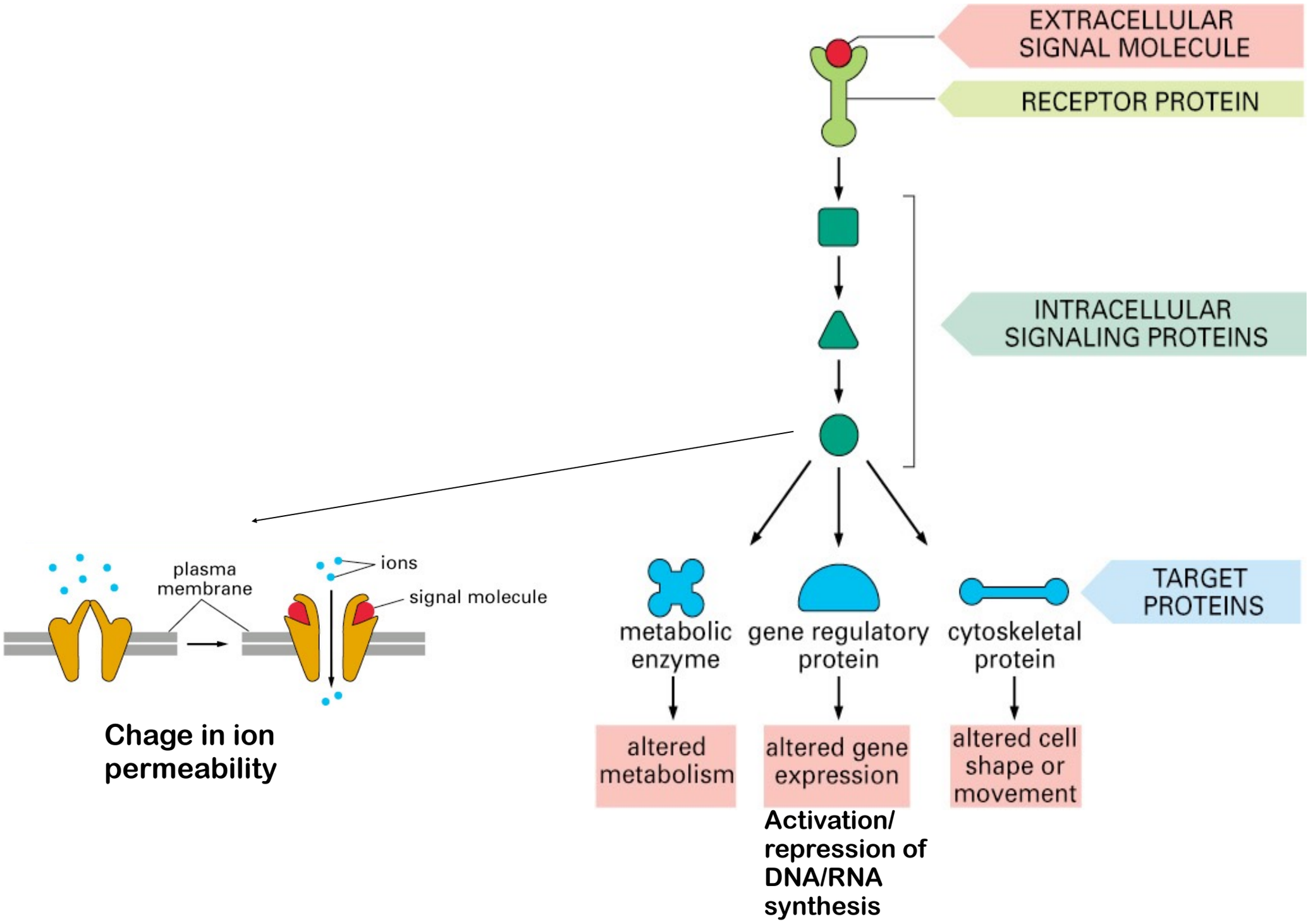
IP3
PIP 1-3-4-5
PA
AA
DAG
cAMP
Ca^{++}

Secondi messaggeri

PKA
PKC
PDK1
PKB/AKT
ROCK
PAK

Effettori

Basic Components and Responses of Cellular Signaling

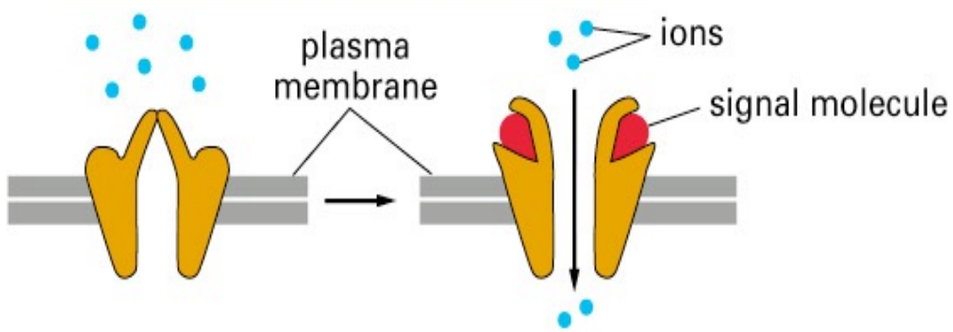


Steps of a signaling pathway

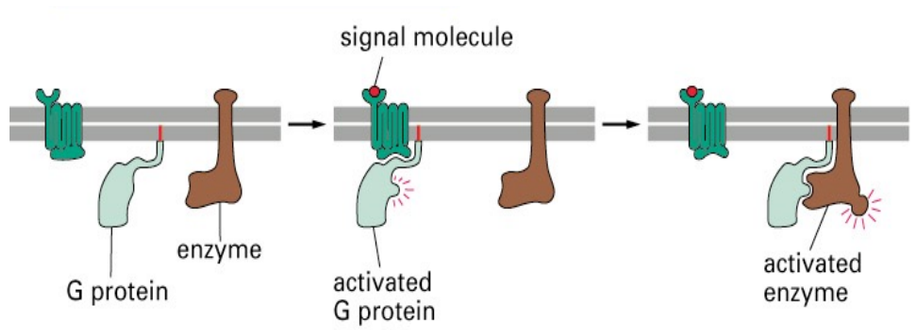
- Recognition of stimulus by cell surface receptor
- Transfer of signal across plasma membrane
- Transmission of the signal to specific targets inside the cells
- Cessation of the responses

Types of Cell-Surface-Receptors

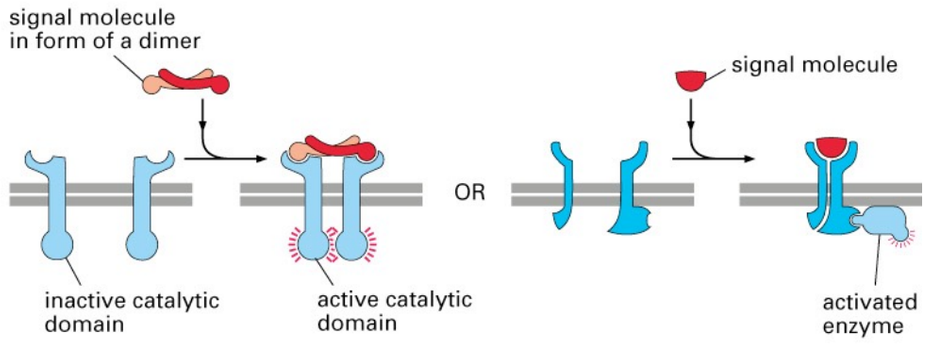
1. Ion-channel-linked receptors



2. G-protein-coupled receptors



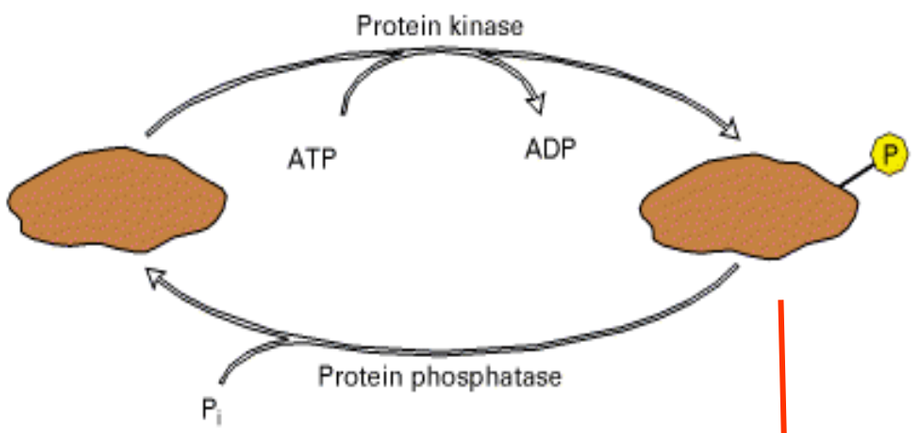
3. Enzyme-linked receptors

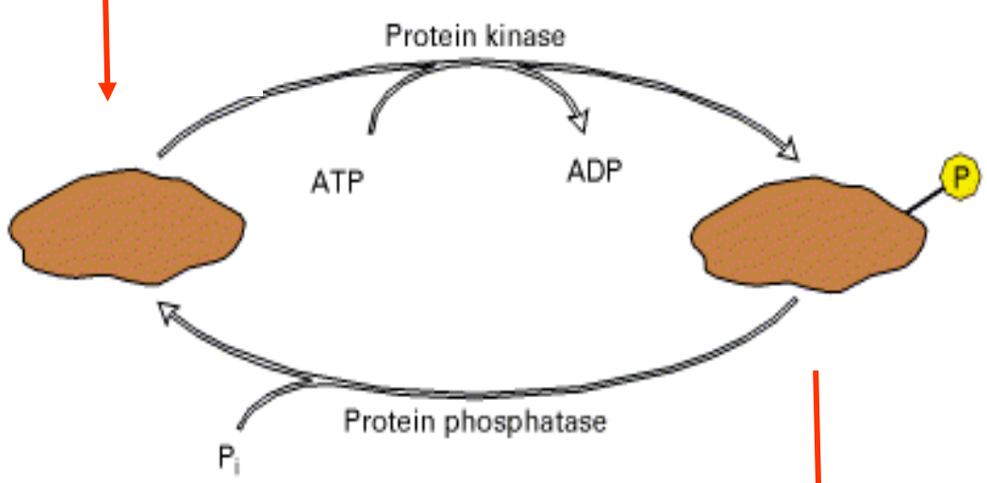


Types of Signaling Proteins

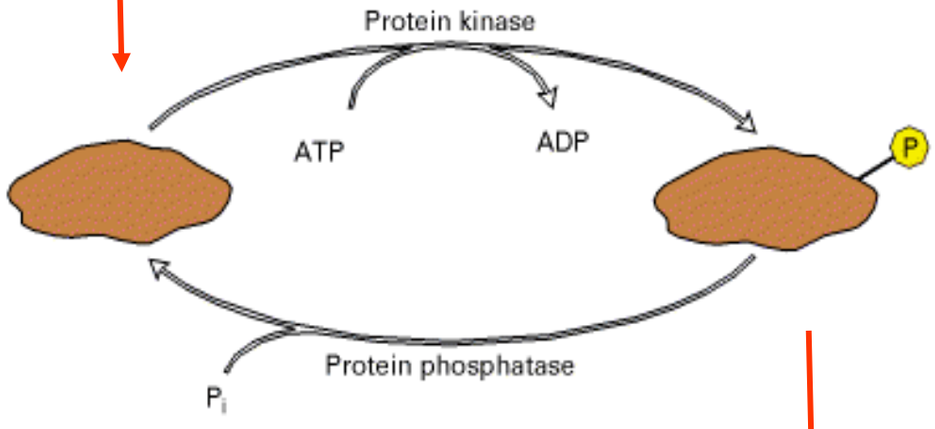
1. Proteins Kinases / Phosphatases. Proteins involved in phosphorylation reactions
2. GTP-binding proteins (switchers)
3. Phospholipases
4. Adaptor and scaffold proteins

Protein Kinases & Phosphatases



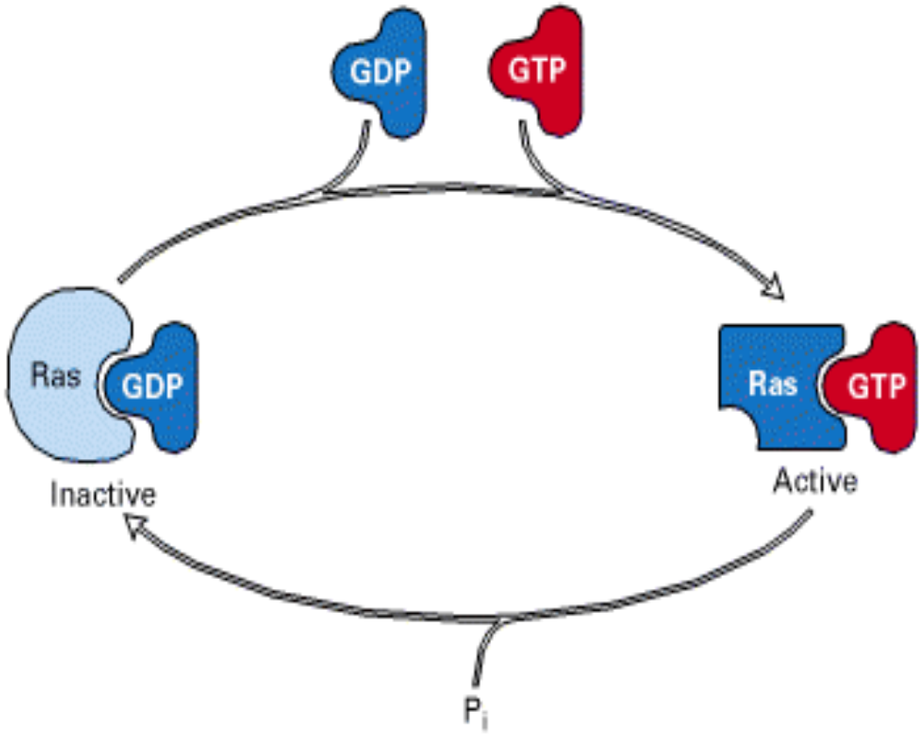


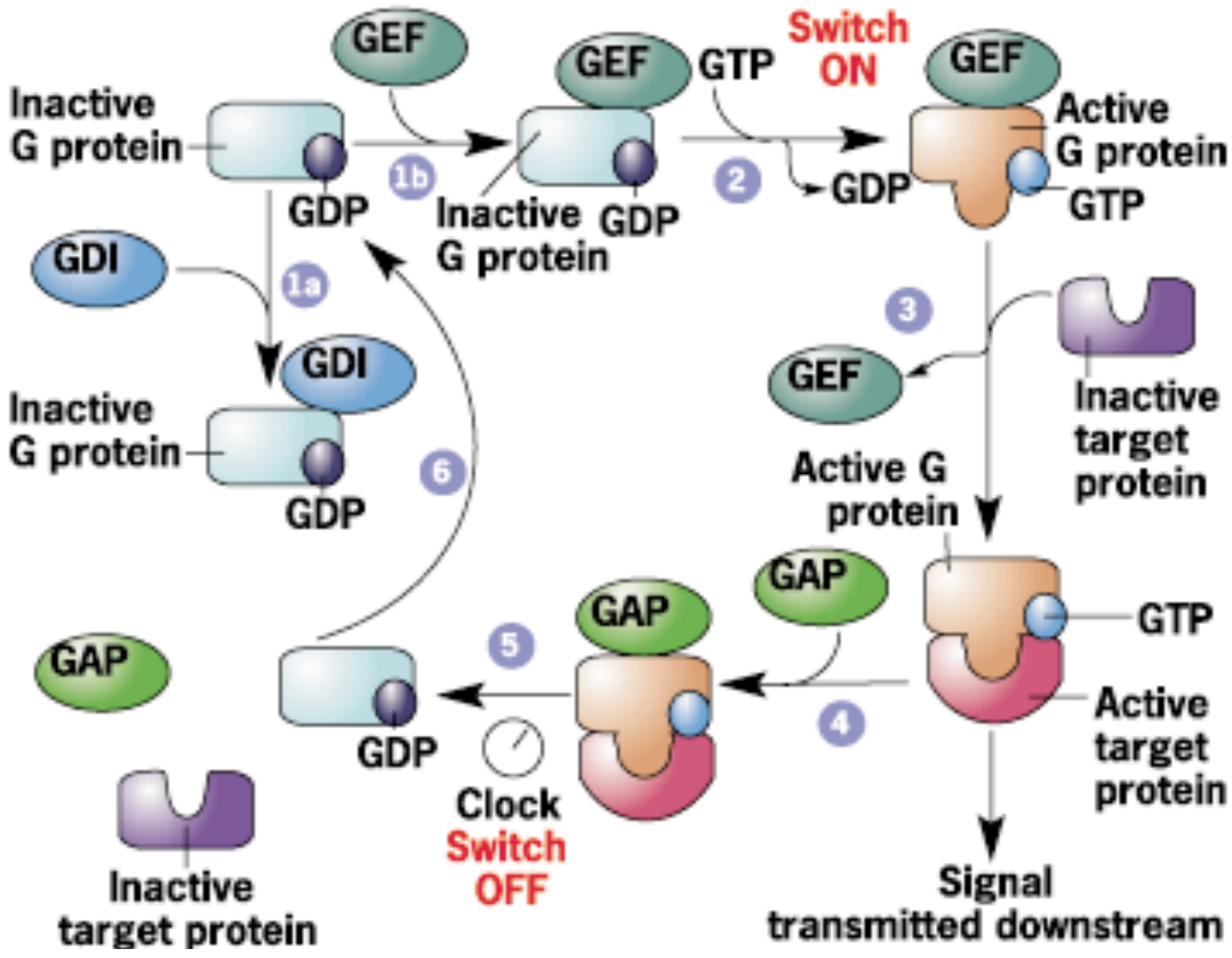
Kinase cascade



Final Target

G-Proteins





Accessory proteins
1. GTPase-activating proteins (GAPs)
2. Guanine nucleotide-exchange factors (GEFs)
3. Guanine nucleotide-dissociation inhibitors (GDIs)

GTPases switchers

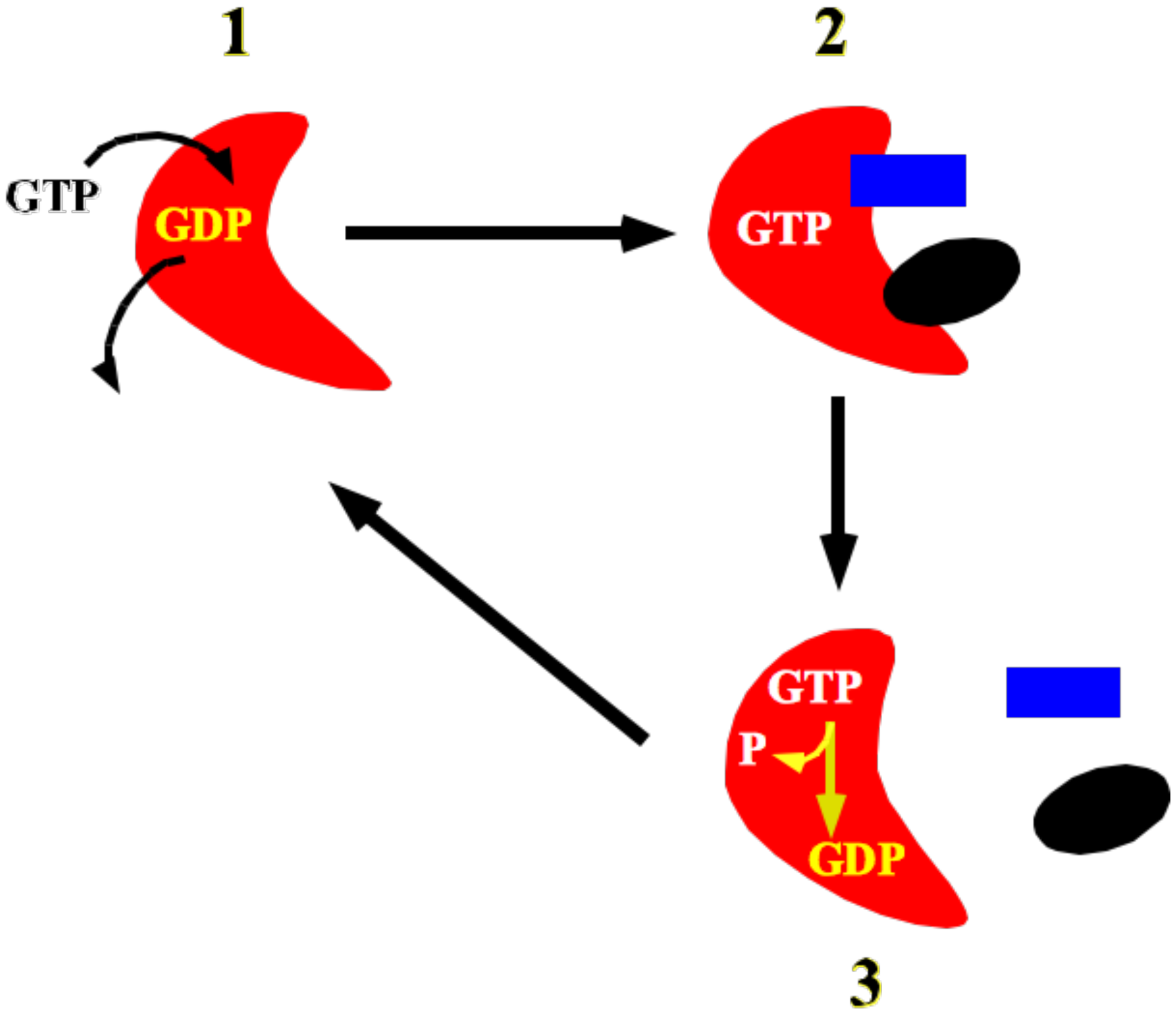
1- Alternano rapidamente lo stato *attivo* (legame al GTP) e *inattivo* (legame al GDP) (ciclo ON-OFF)

2- Hanno attività GTPasica intrinseca (spegnimento intramolecolare)

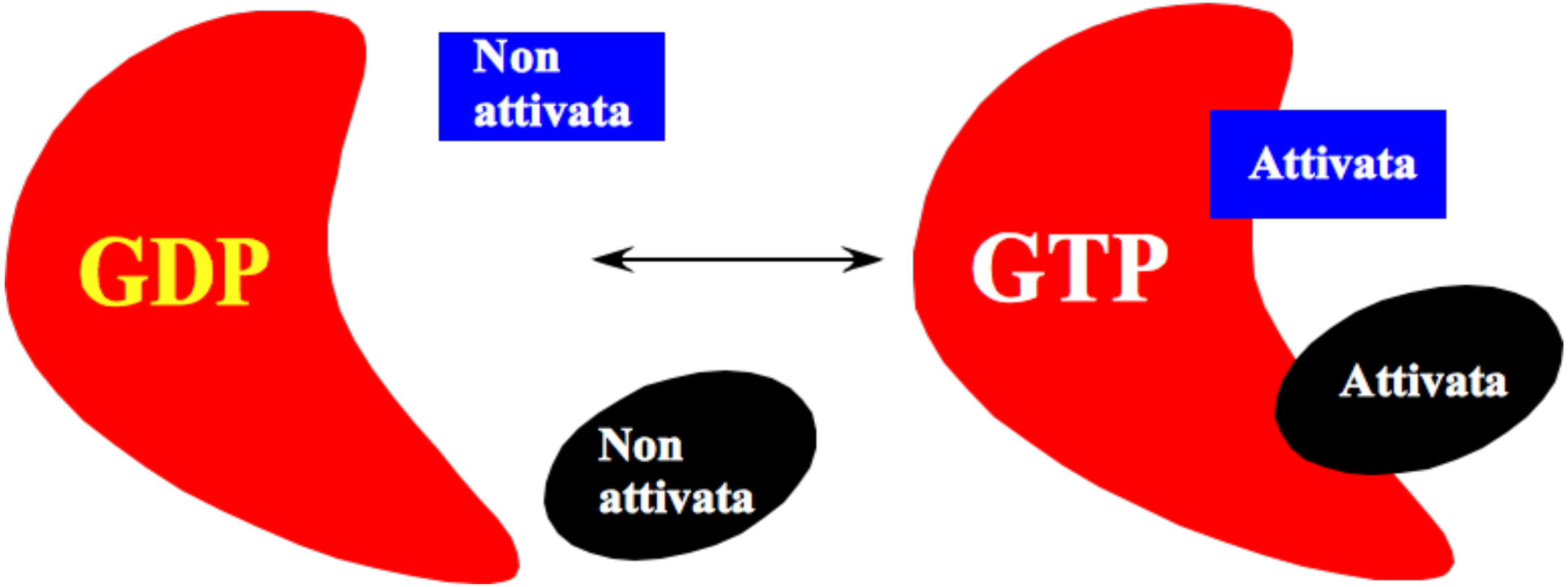
3- Non sono amplificatori diretti

4- Diversificano le vie di transduzione del segnale

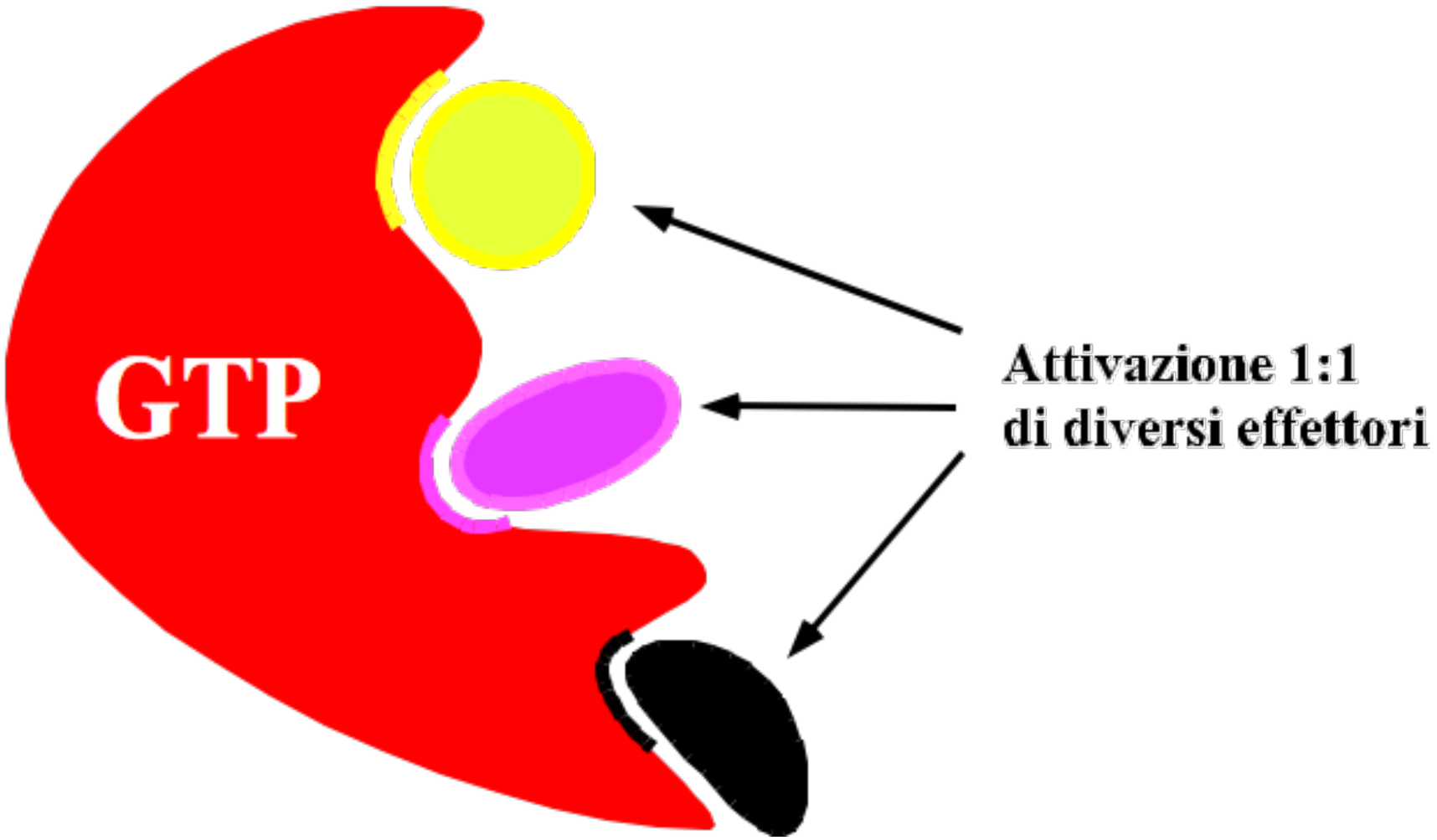
5- Sono modulari


1
2
GTP
GDP
GTP
GTP
P
GDP
3

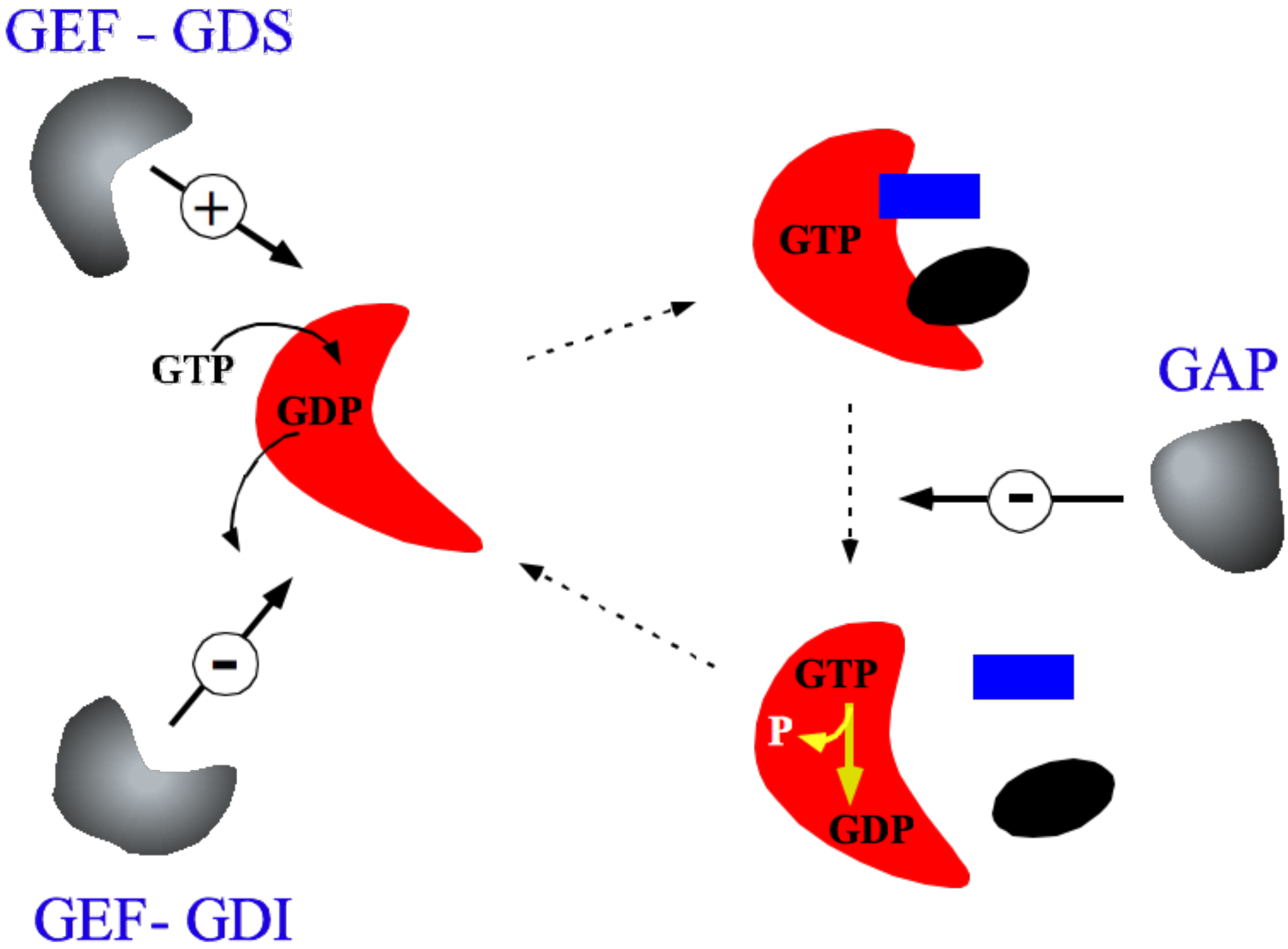
Rapporto di attivazione 1:1
Il sistema non amplifica



Struttura con multipli domini effettori
Il sistema diversifica



GTPases regulators



Trimeric GTP-binding proteins

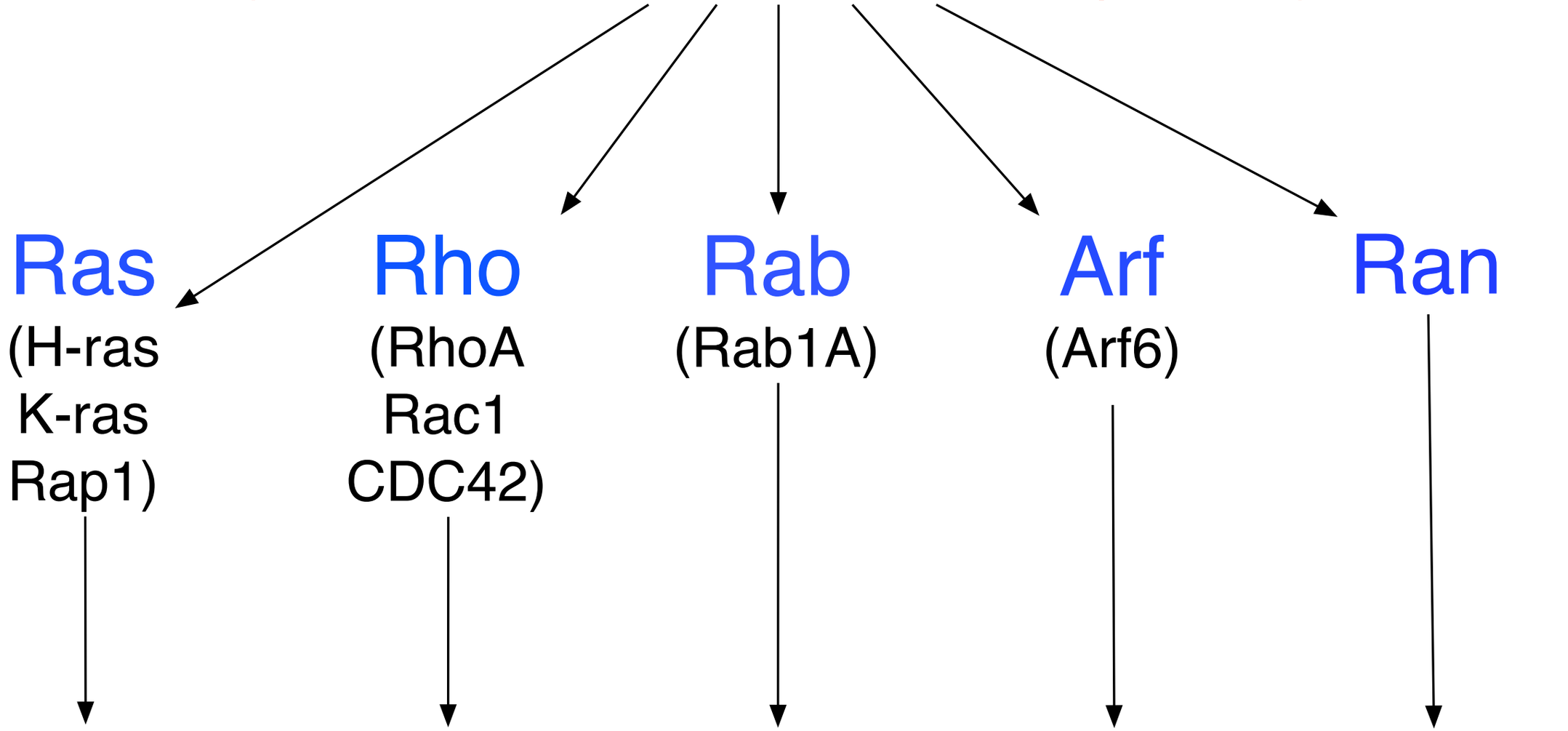
TABLE 15.2 G-protein families and their functions

G_α class	Initiating signal	Downstream signal
$G_{\alpha s}$	β-Adrenergic amines, glucagon, parathyroid hormone, many others	Stimulates adenylate cyclase
$G_{\alpha i}$	Acetylcholine, α-adrenergic amines, many neurotransmitters	Inhibits adenylate cyclase
$G_{\alpha t}$	Photons	Stimulates cGMP phosphodiesterase
$G_{\alpha q}$	Acetylcholine, α-adrenergic amines, many neurotransmitters	Increases IP_3 and intracellular calcium
$G_{\alpha 13}$	Thrombin, other agonists	Stimulates Na^+ and H^+ exchange

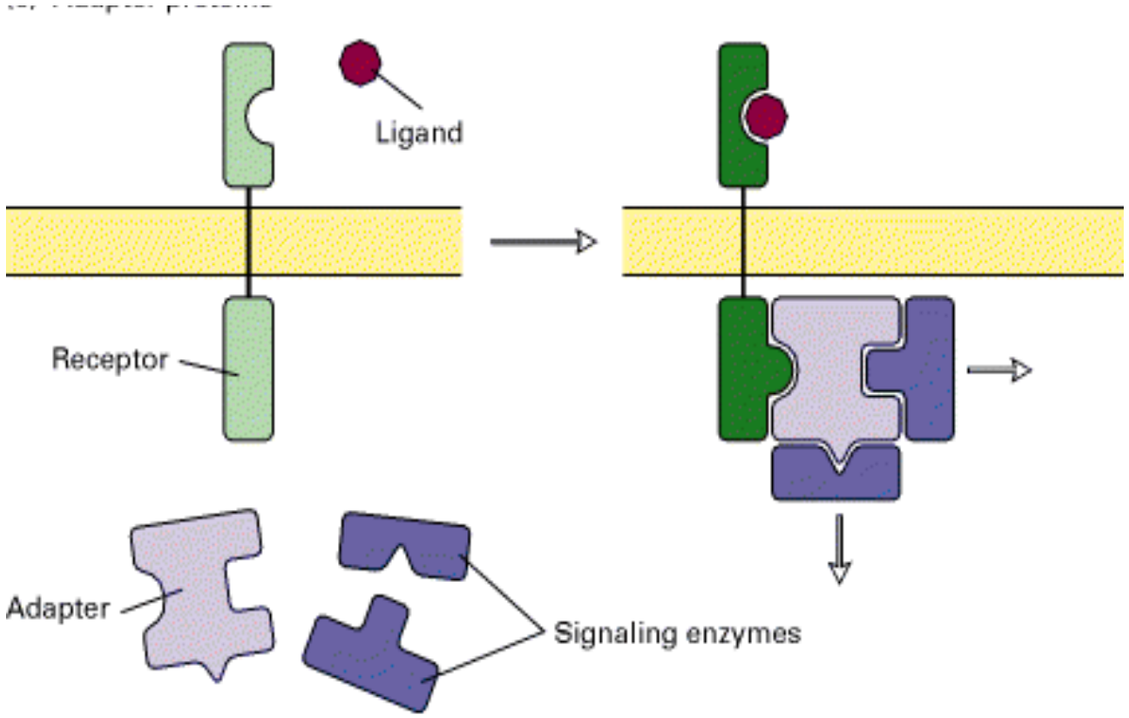
Source: Z. Farfel, H. R. Bourne, and T. Iiri. *N. Engl. J. Med.* 340(1999):1012.

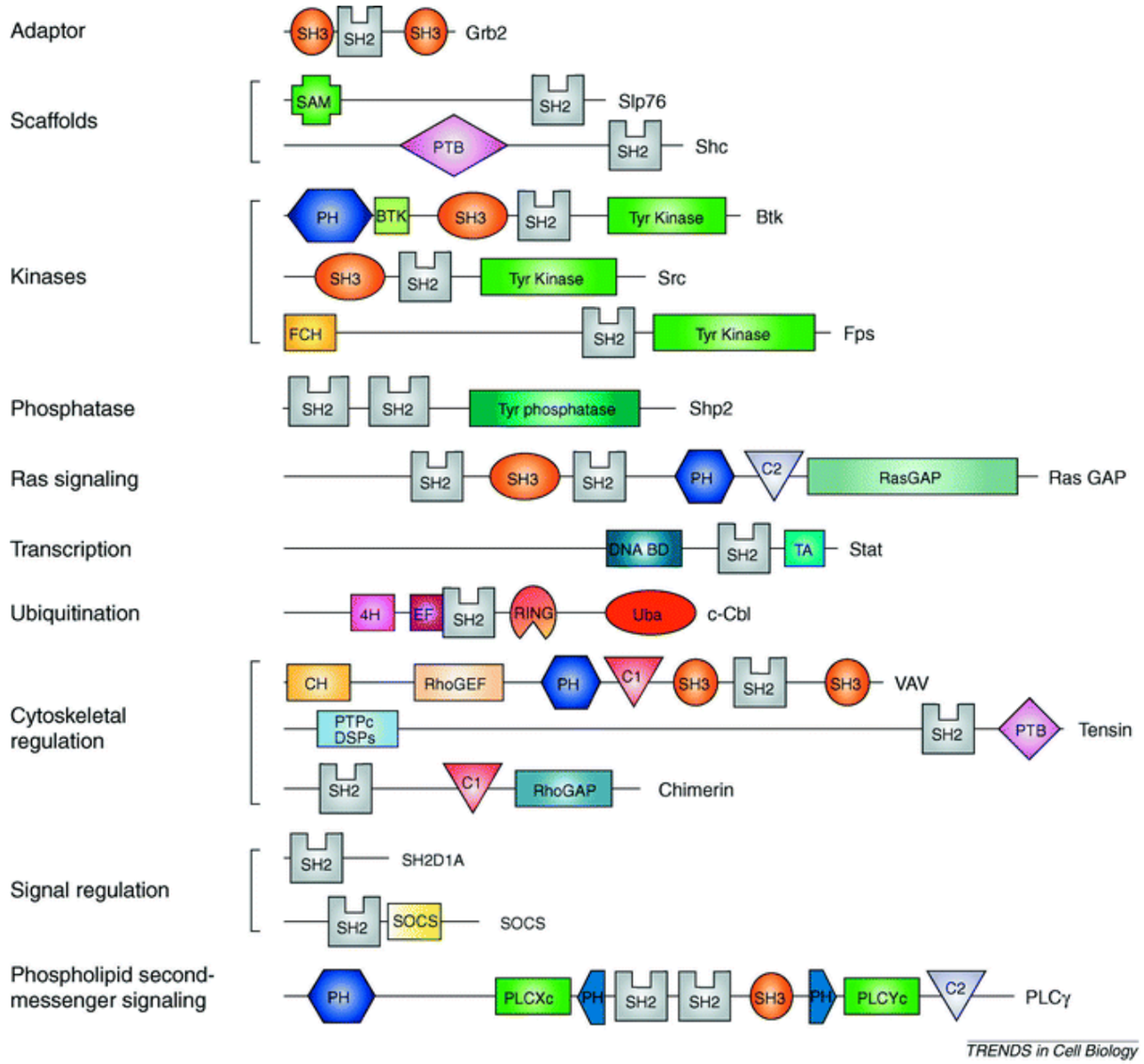
Le small GTP-binding proteins

(monomeri, mw. 21 kDa, circa 100 proteine)

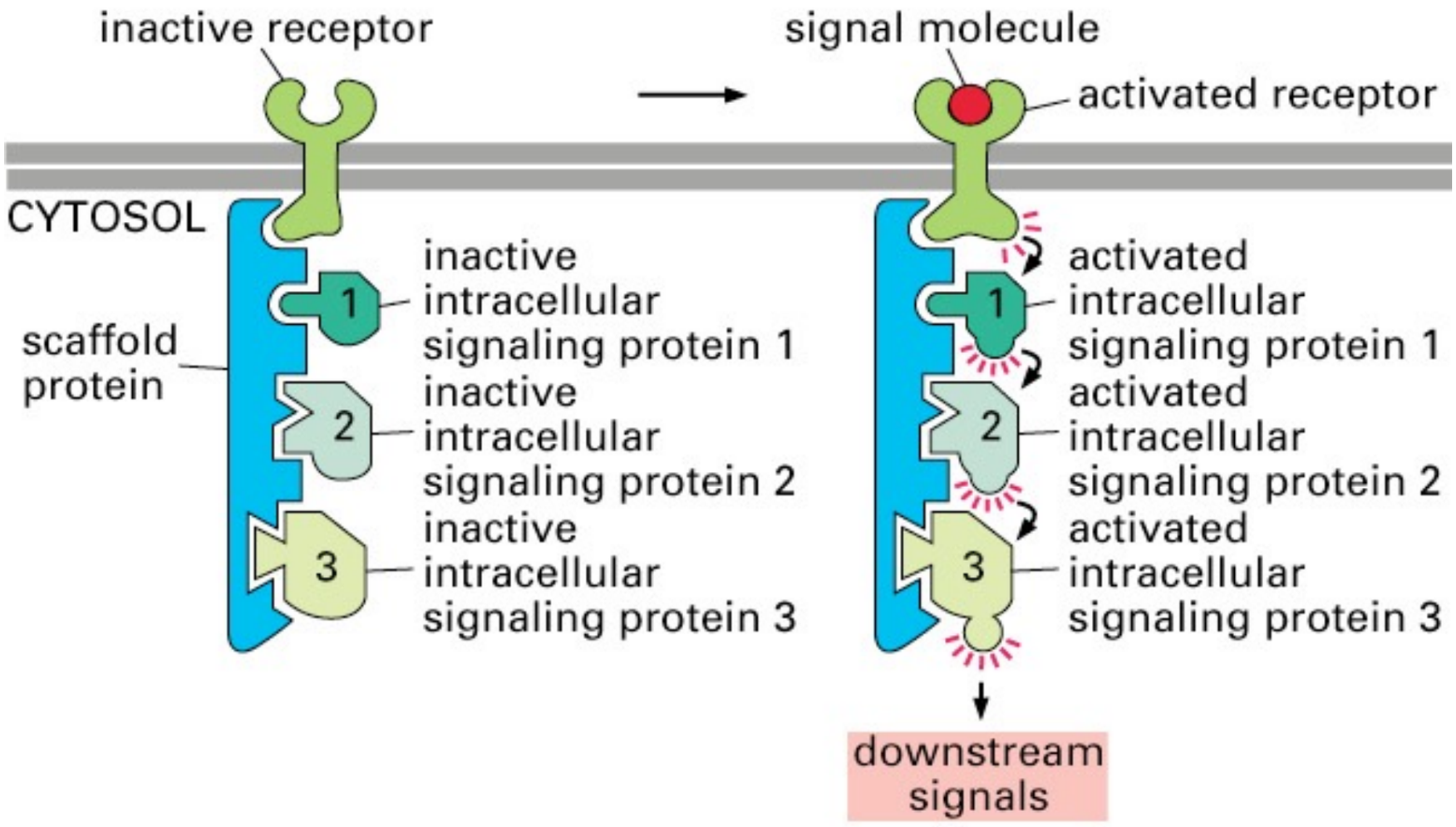


Adaptor Protein

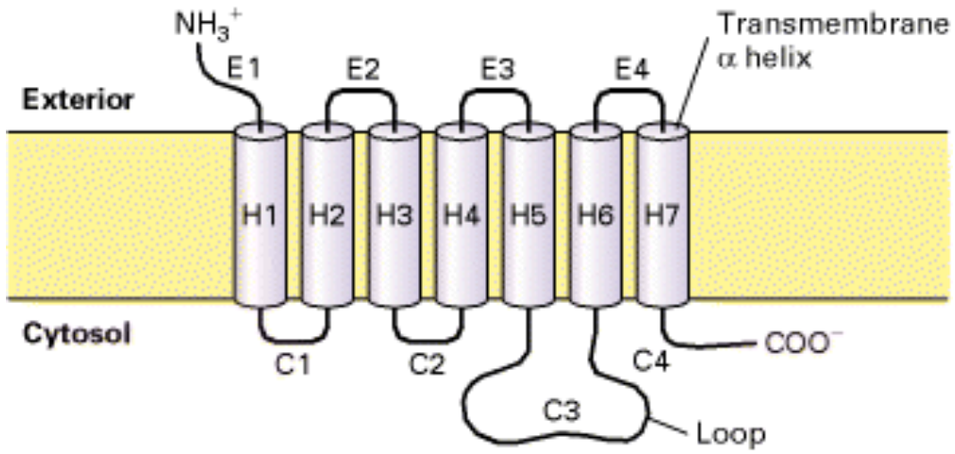


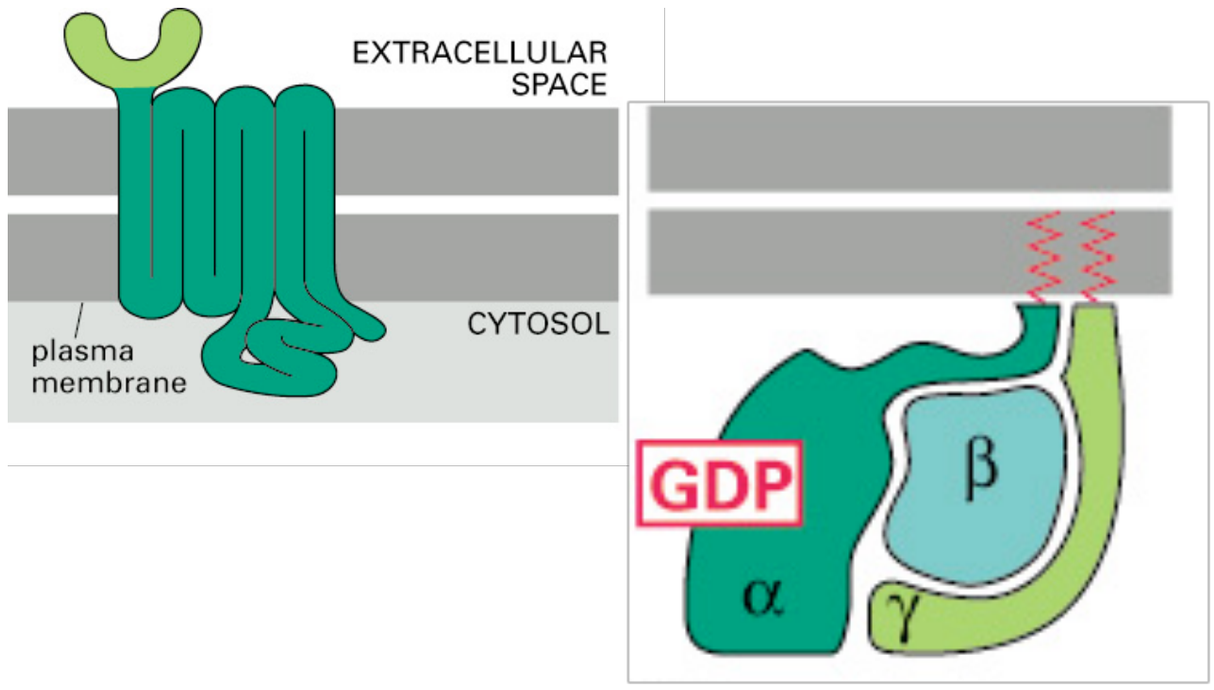


Scaffold Proteins



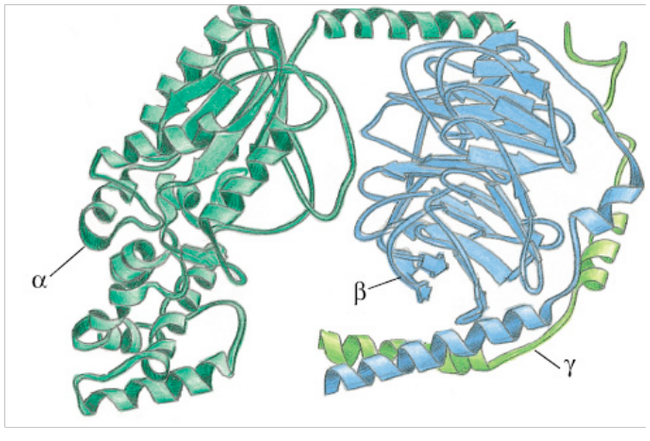
G protein-coupled receptor





Seven membrane spanning α helices

G protein binds to guanine nucleotides, either GDP or GTP. It consists of three different polypeptide subunits, called α, β, and γ.



I. Activation of the G protein by the receptor

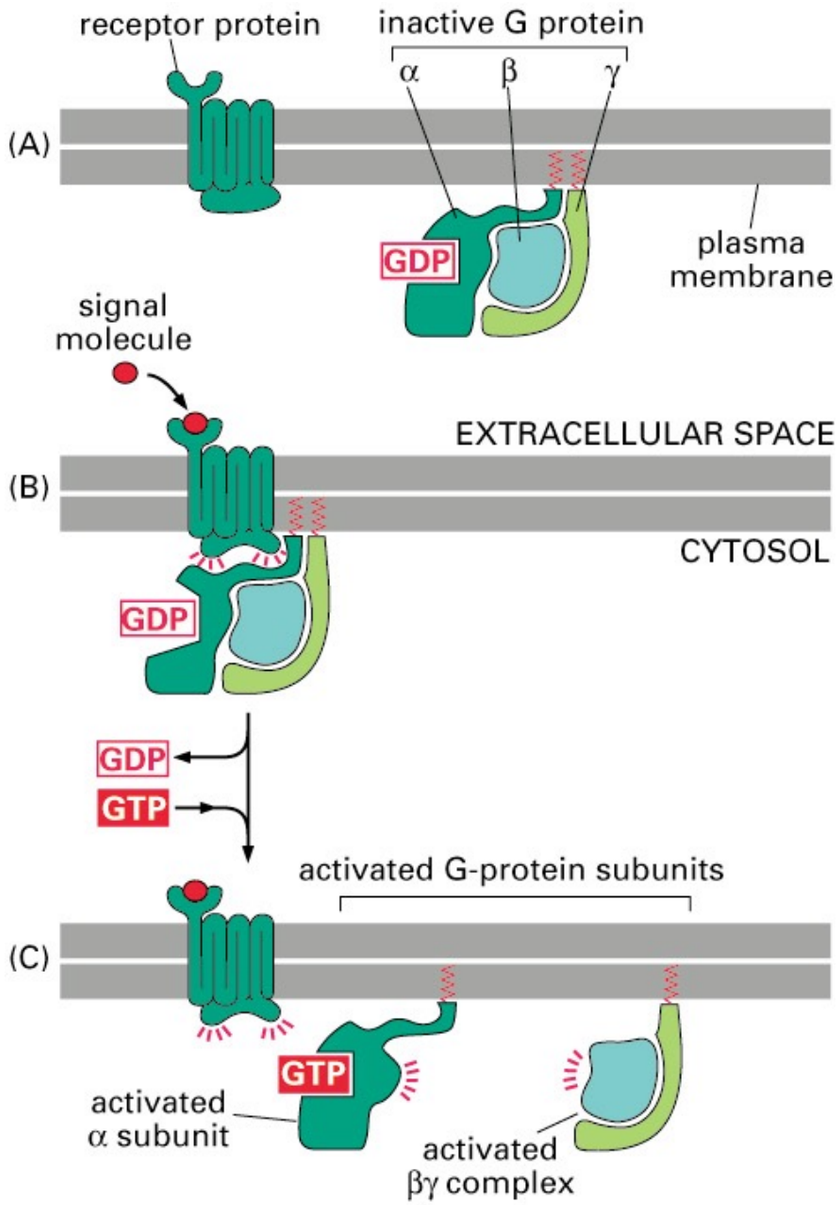


Figure 15–28. Molecular Biology of the Cell, 4th Edition.

II. Relay of the signal from G protein to effector

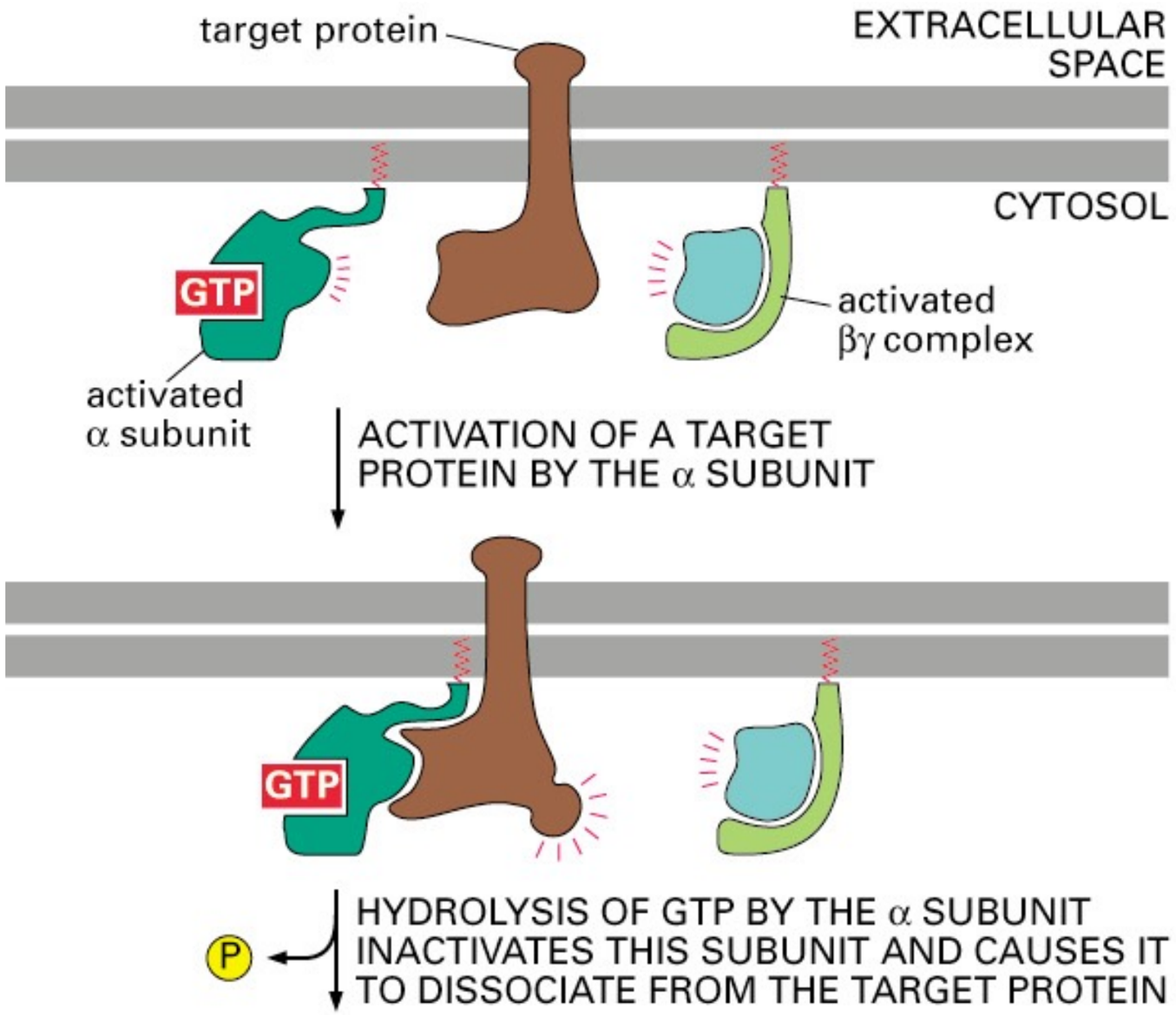


Figure 15–29 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

III. Termination of the response

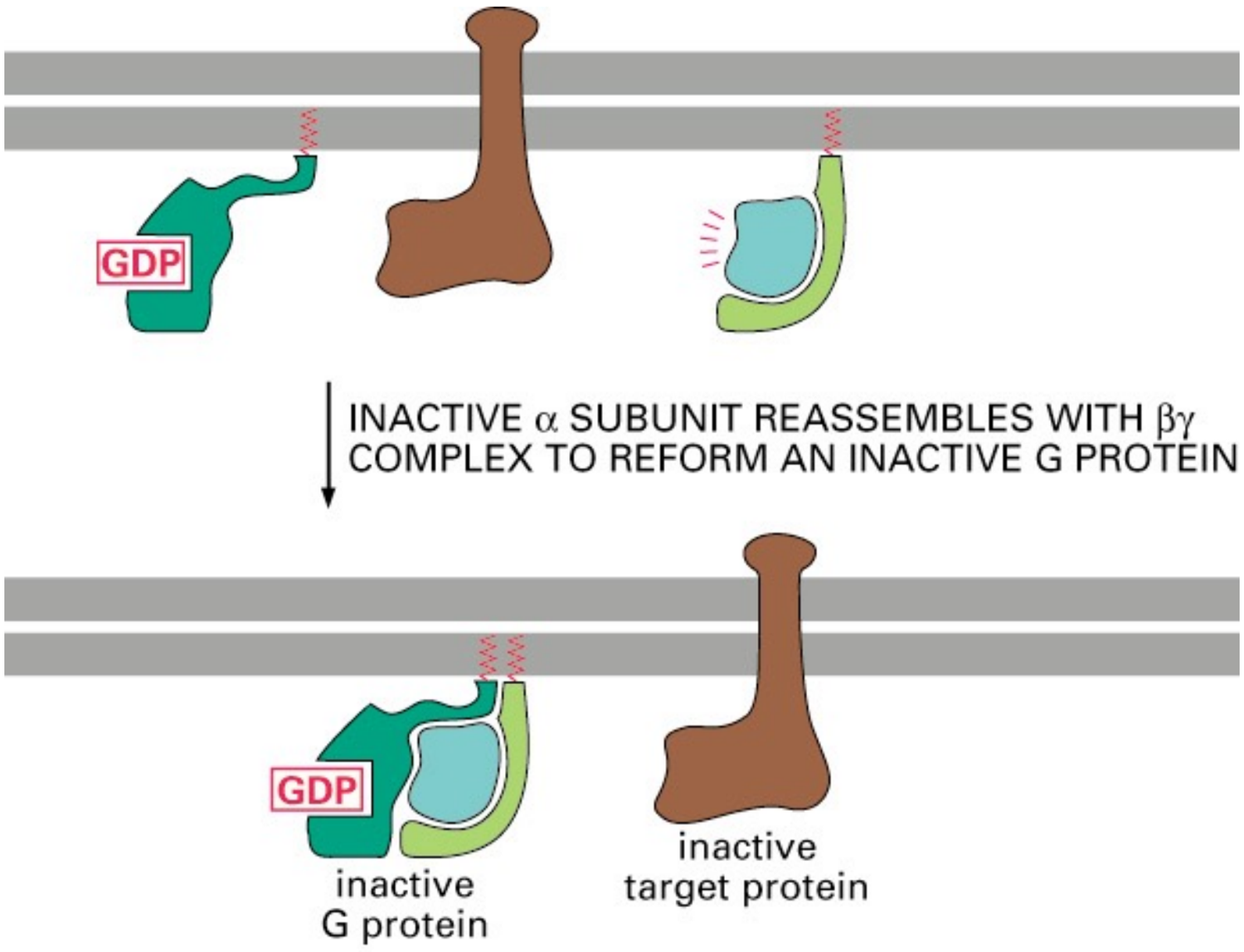


Figure 15–29 part 2 of 2. Molecular Biology of the Cell, 4th Edition.

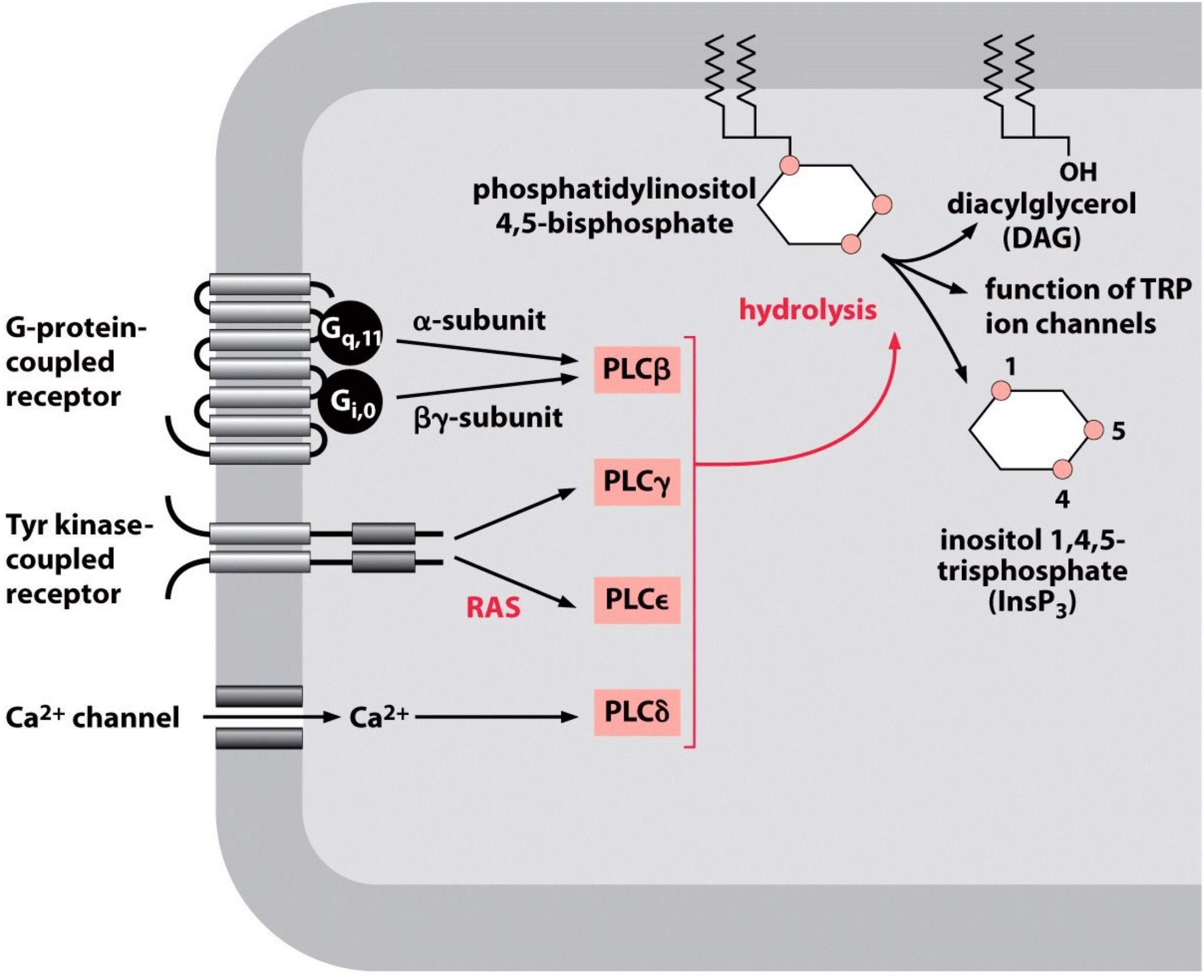


Figure 4.20 Cellular Signal Processing (© Garland Science 2009)

Phospholipases (amplificators)

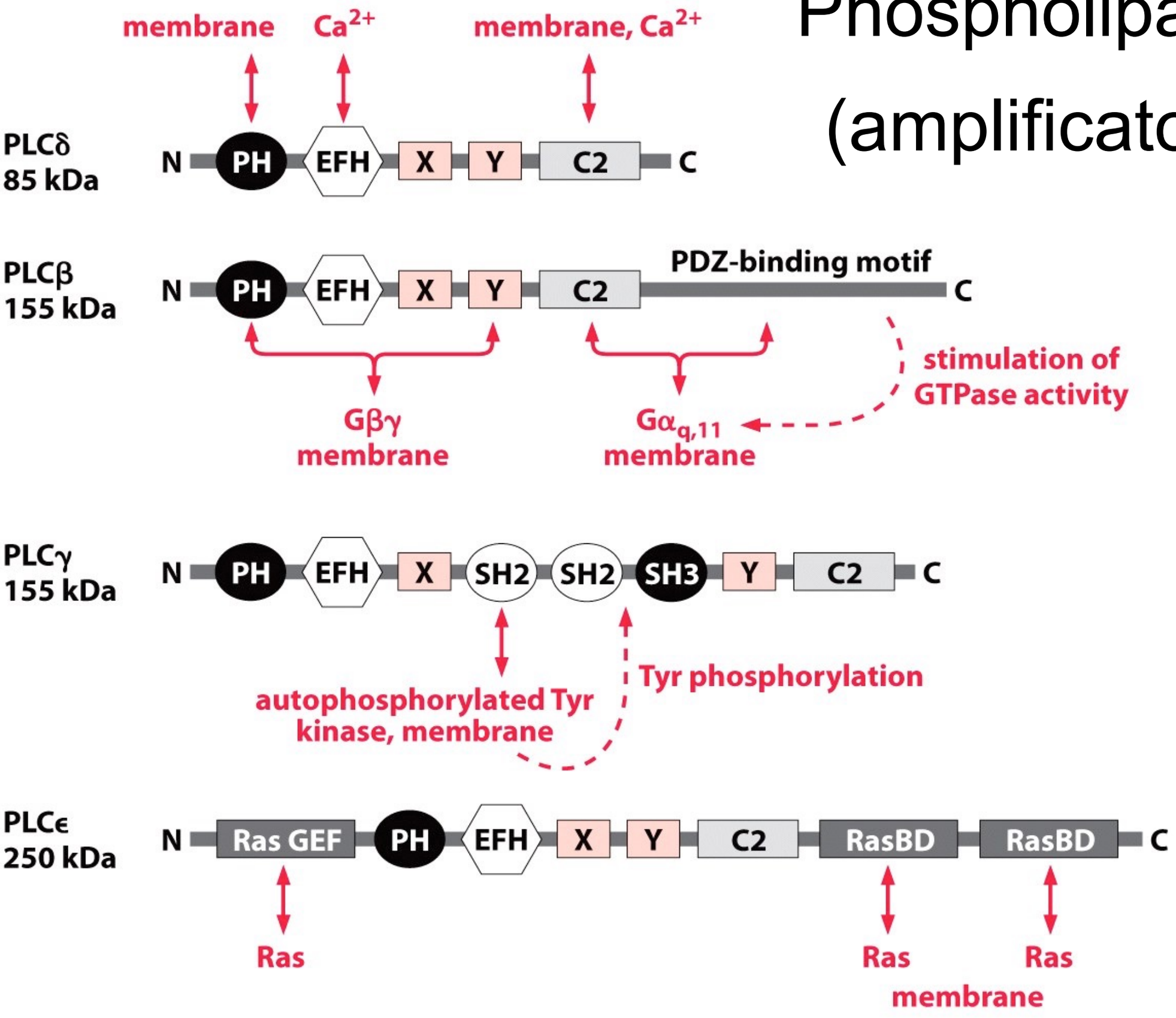


Figure 4.19 Cellular Signal Processing (© Garland Science 2009)

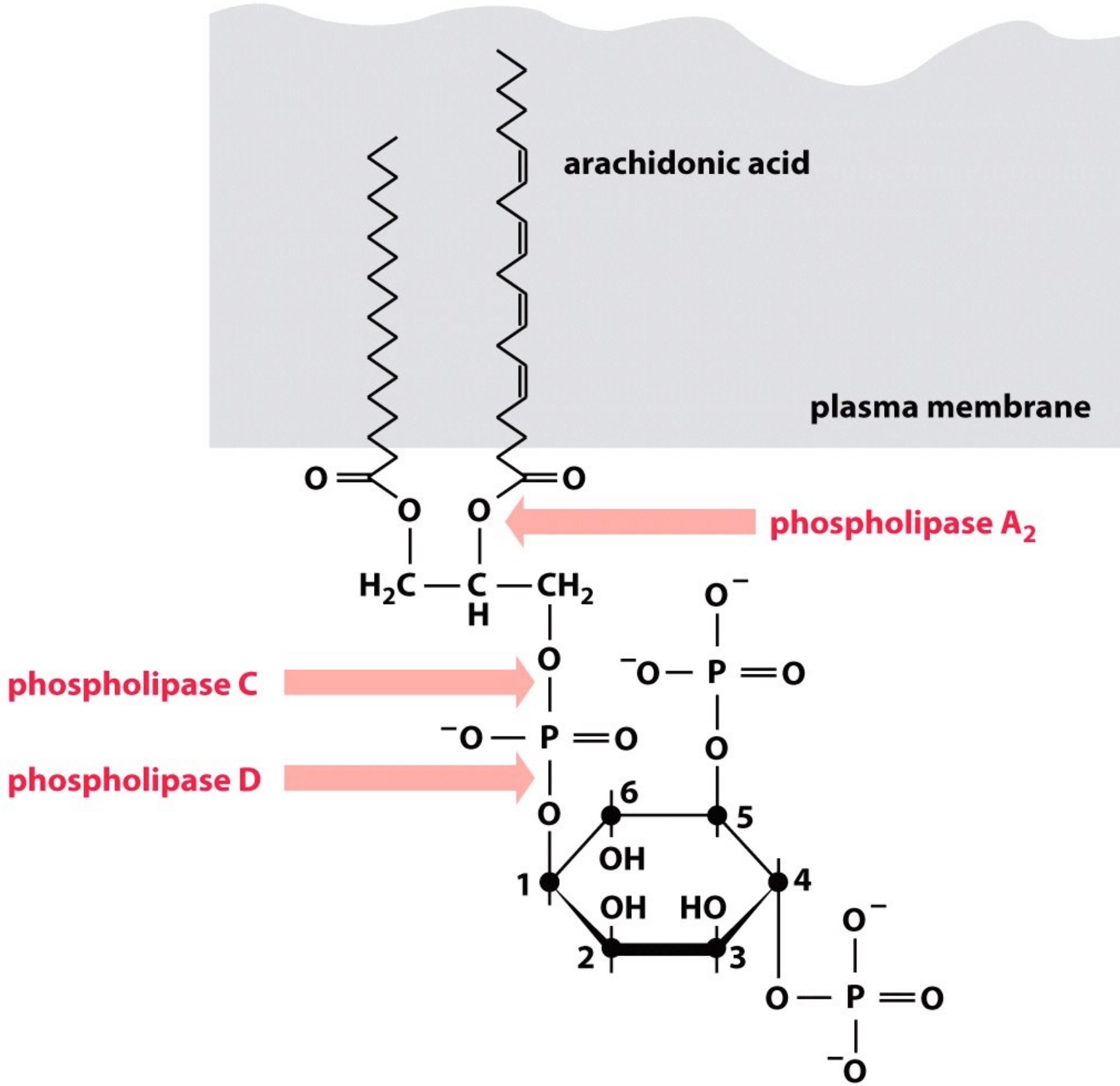
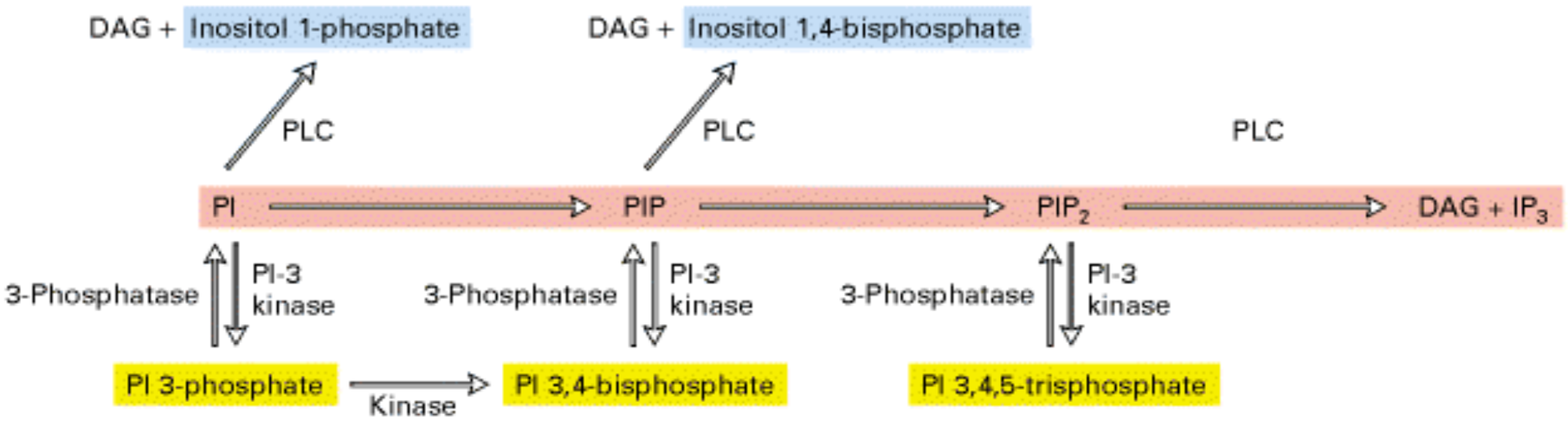
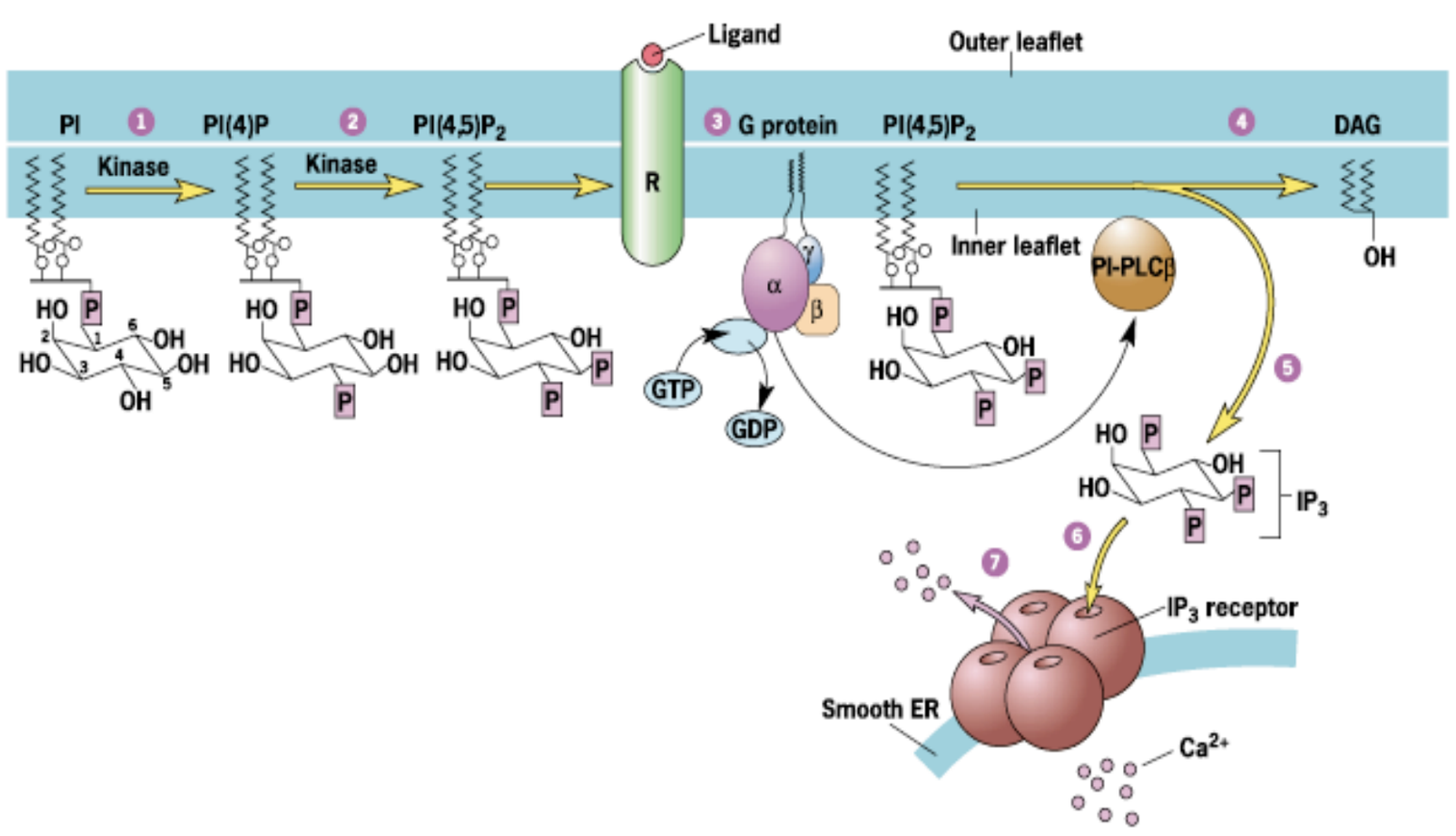


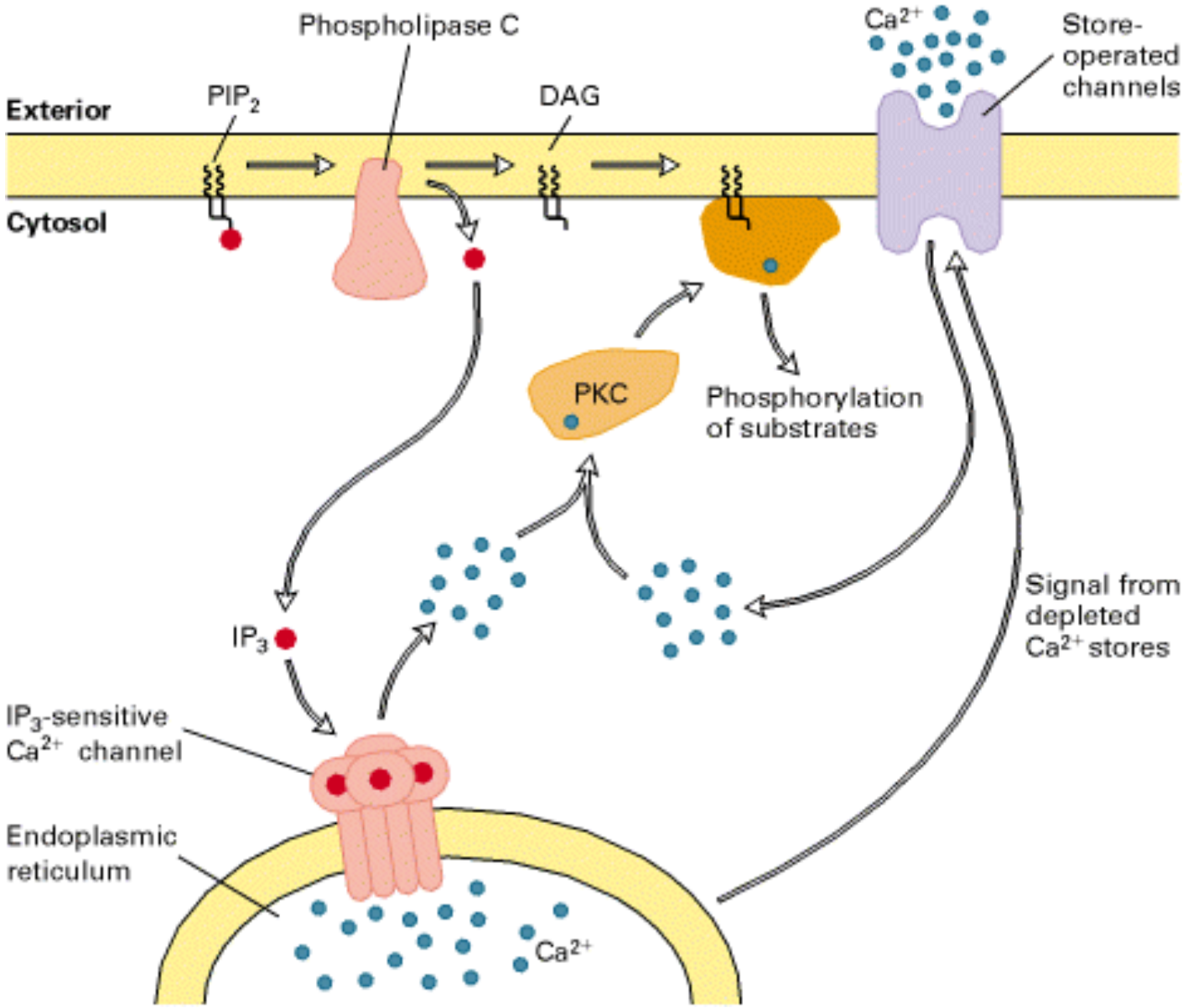
Figure 4.18 Cellular Signal Processing (© Garland Science 2009)

The generation of phosphatidyl inositol-derived second messengers



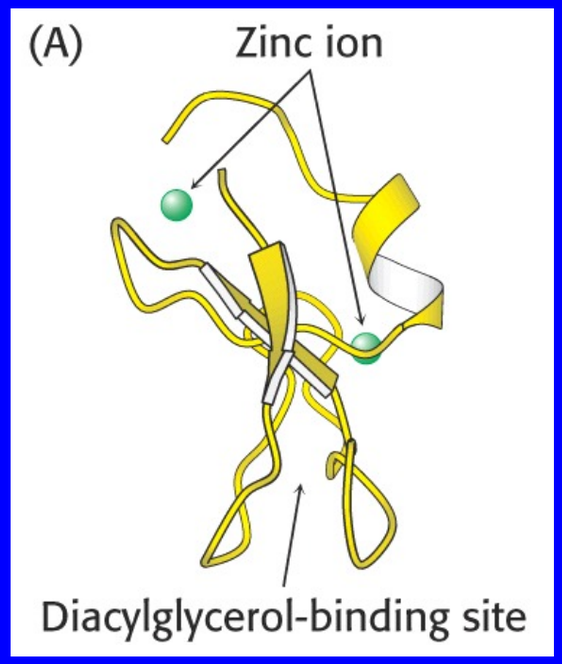


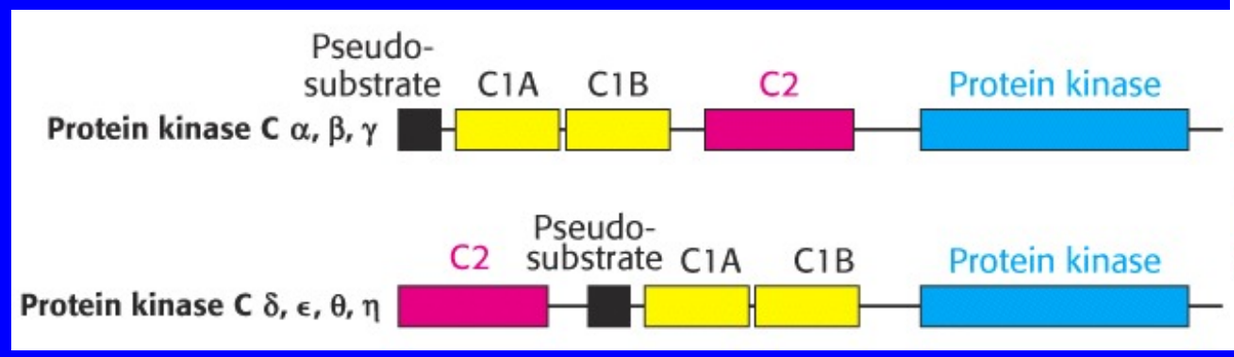
Elevation of Ca^{2+} via the inositol lipid signaling pathway

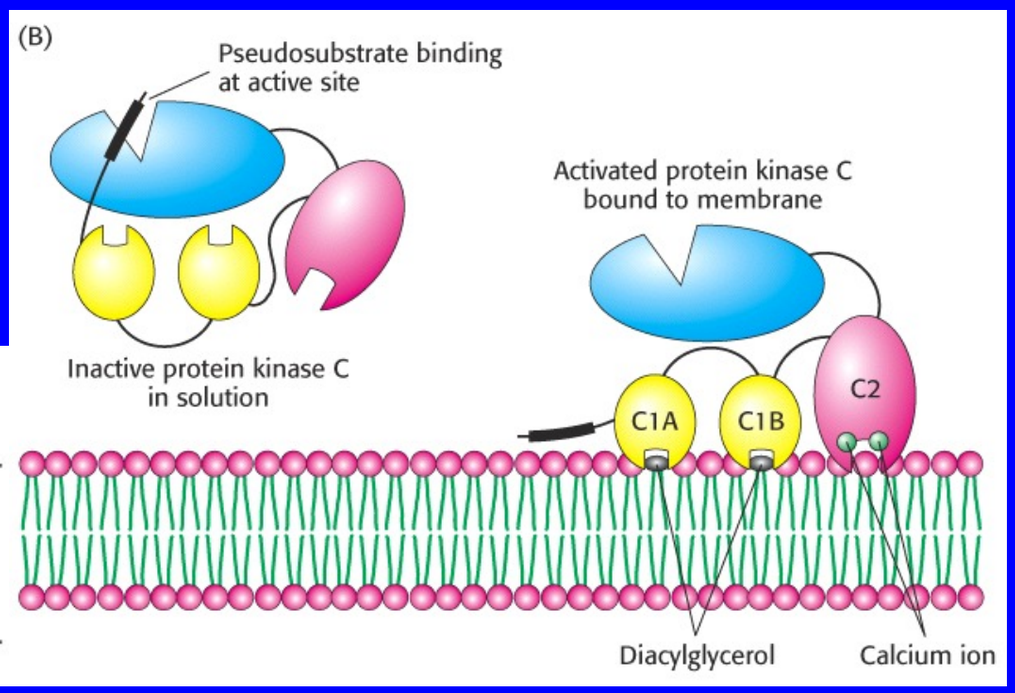


Protein Kinase C

- calcium binds C2 domain of protein kinase C - causes association with membrane
- DAG binds C1 domains of protein kinase C, removing pseudosubstrate from active site
- PKC phosphorylates specific proteins to cause a cellular response
- DAG is also a precursor of arachidonic acid and prostaglandins







PKC family of ser-thr kinase

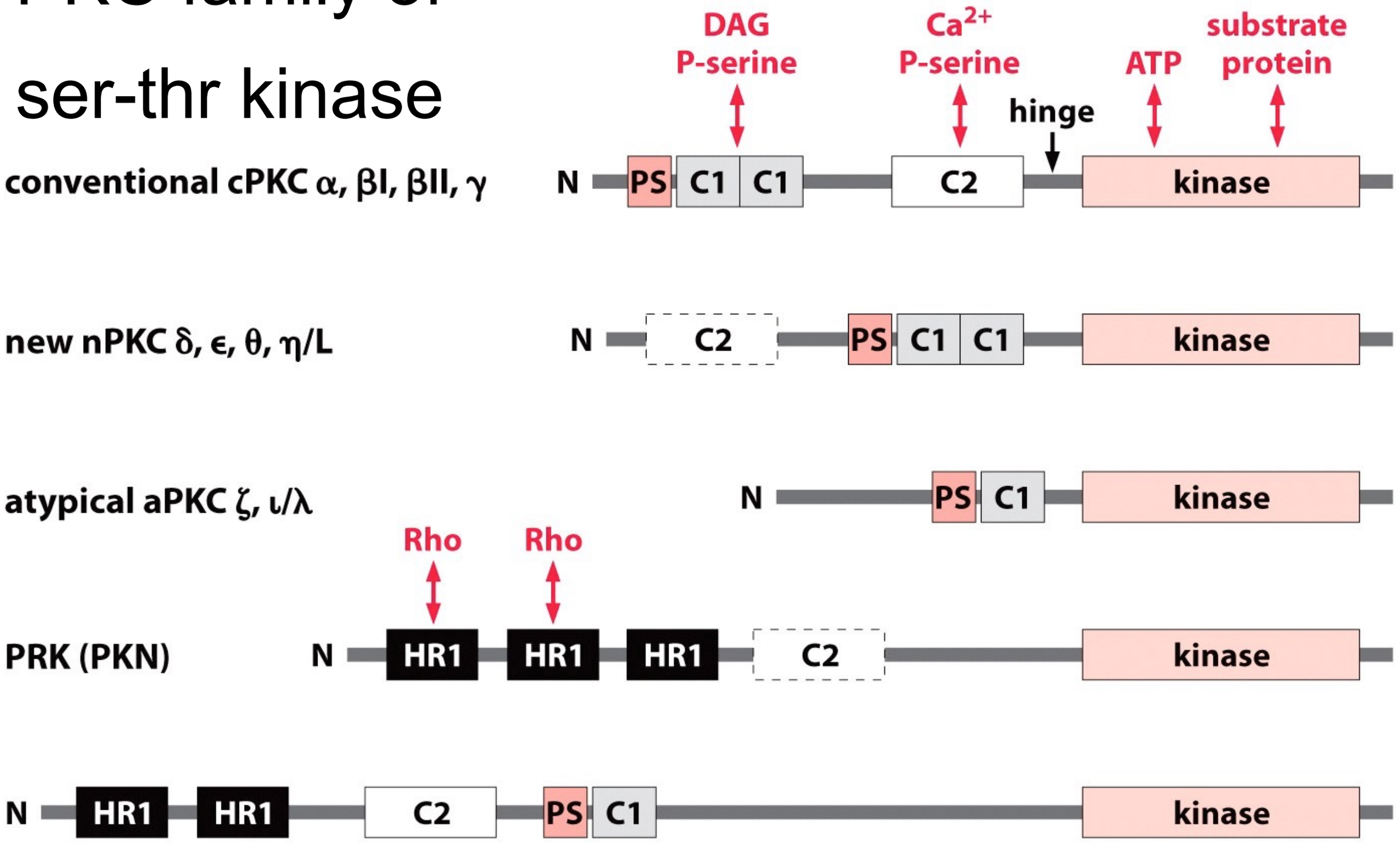


Figure 4.38 Cellular Signal Processing (© Garland Science 2009)

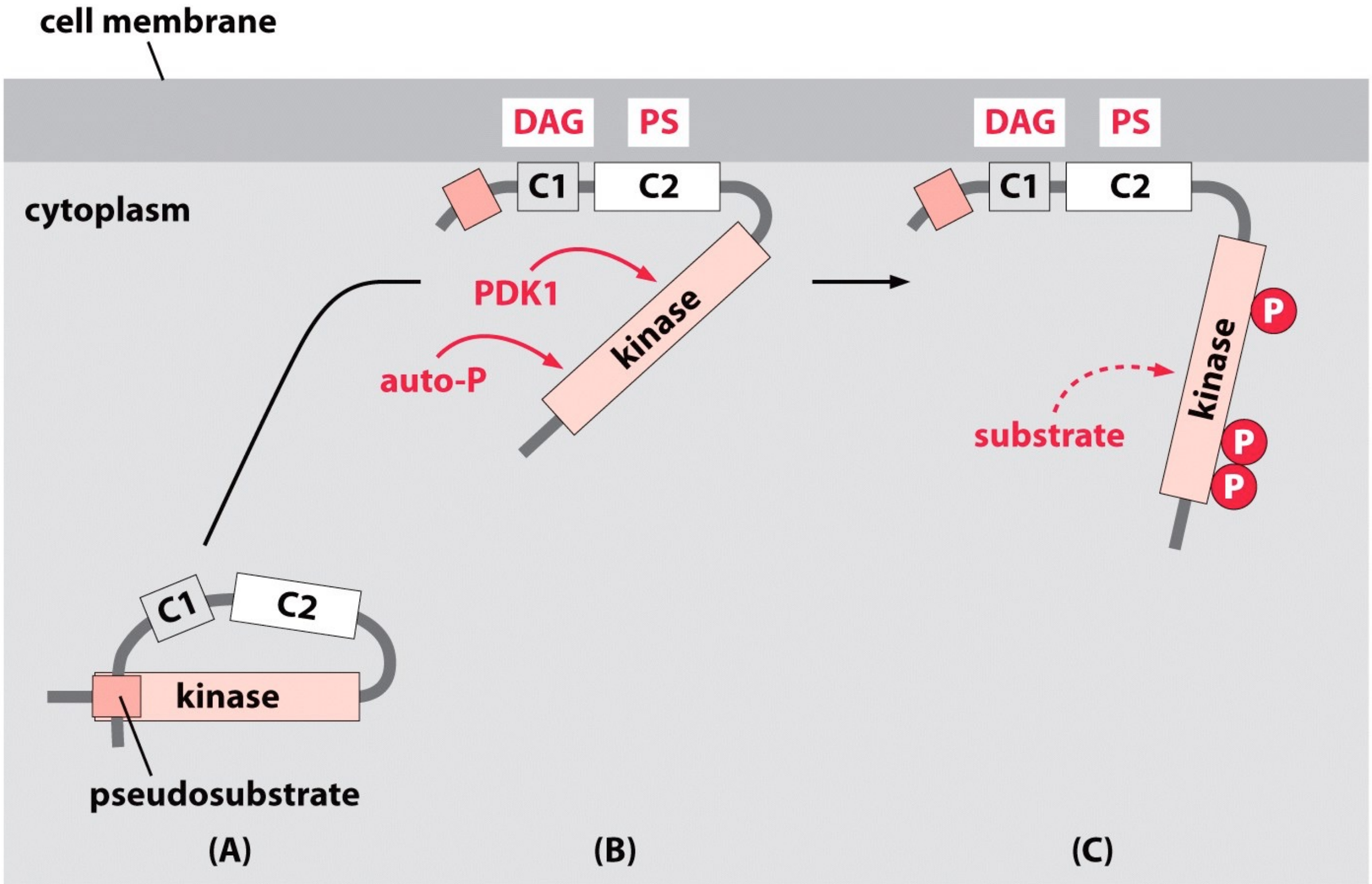


Figure 4.39 Cellular Signal Processing (© Garland Science 2009)

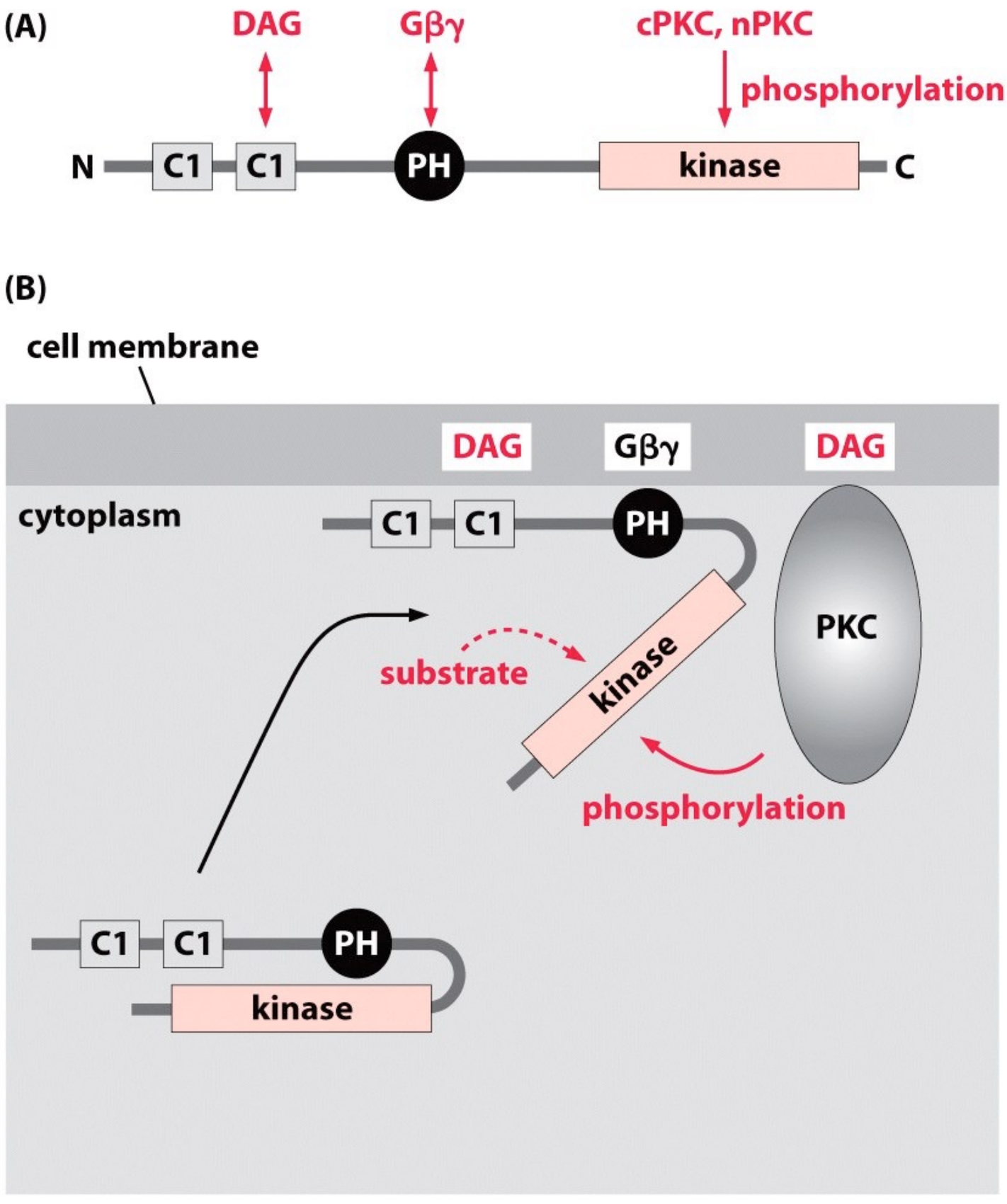


Figure 4.41 Cellular Signal Processing (© Garland Science 2009)

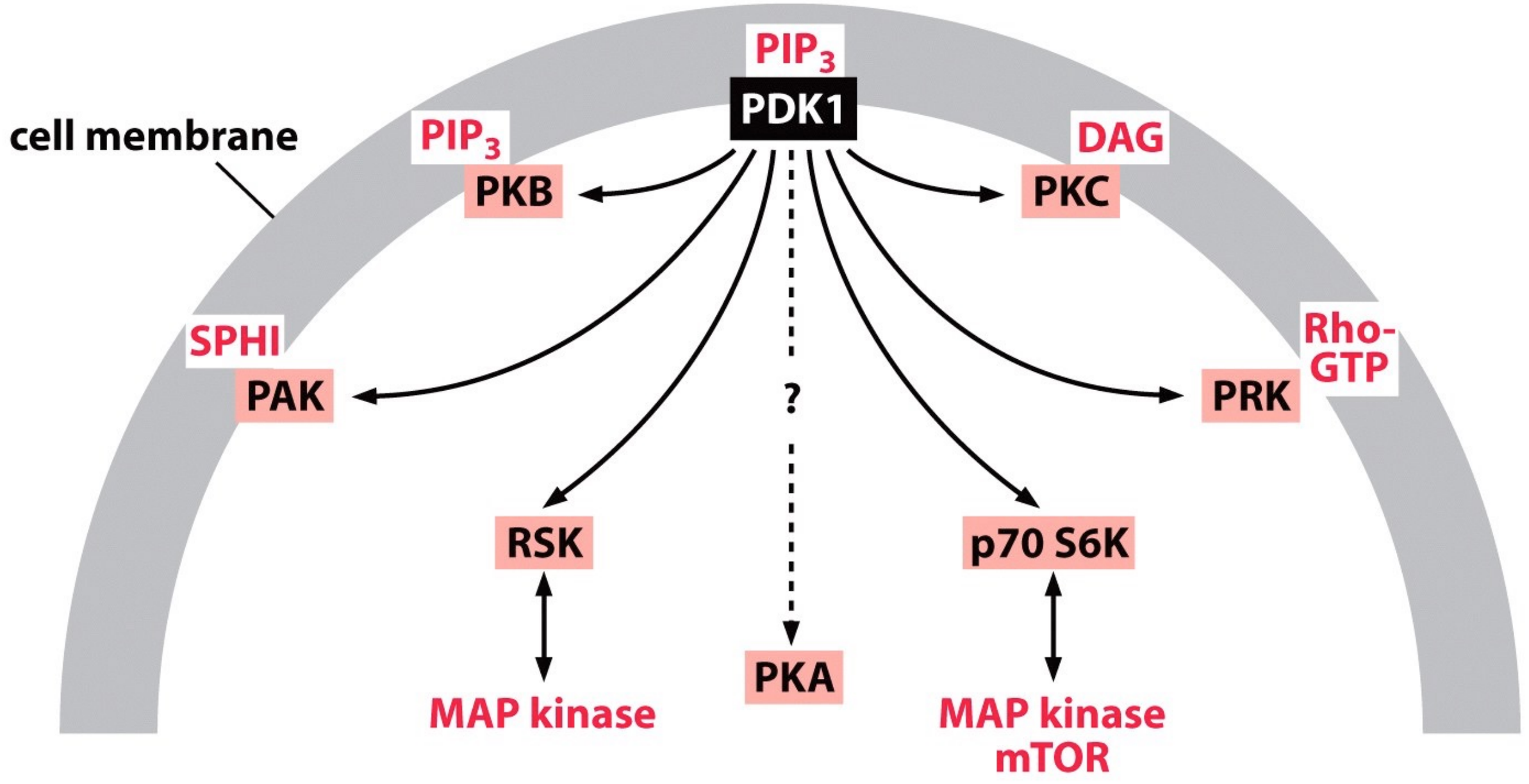
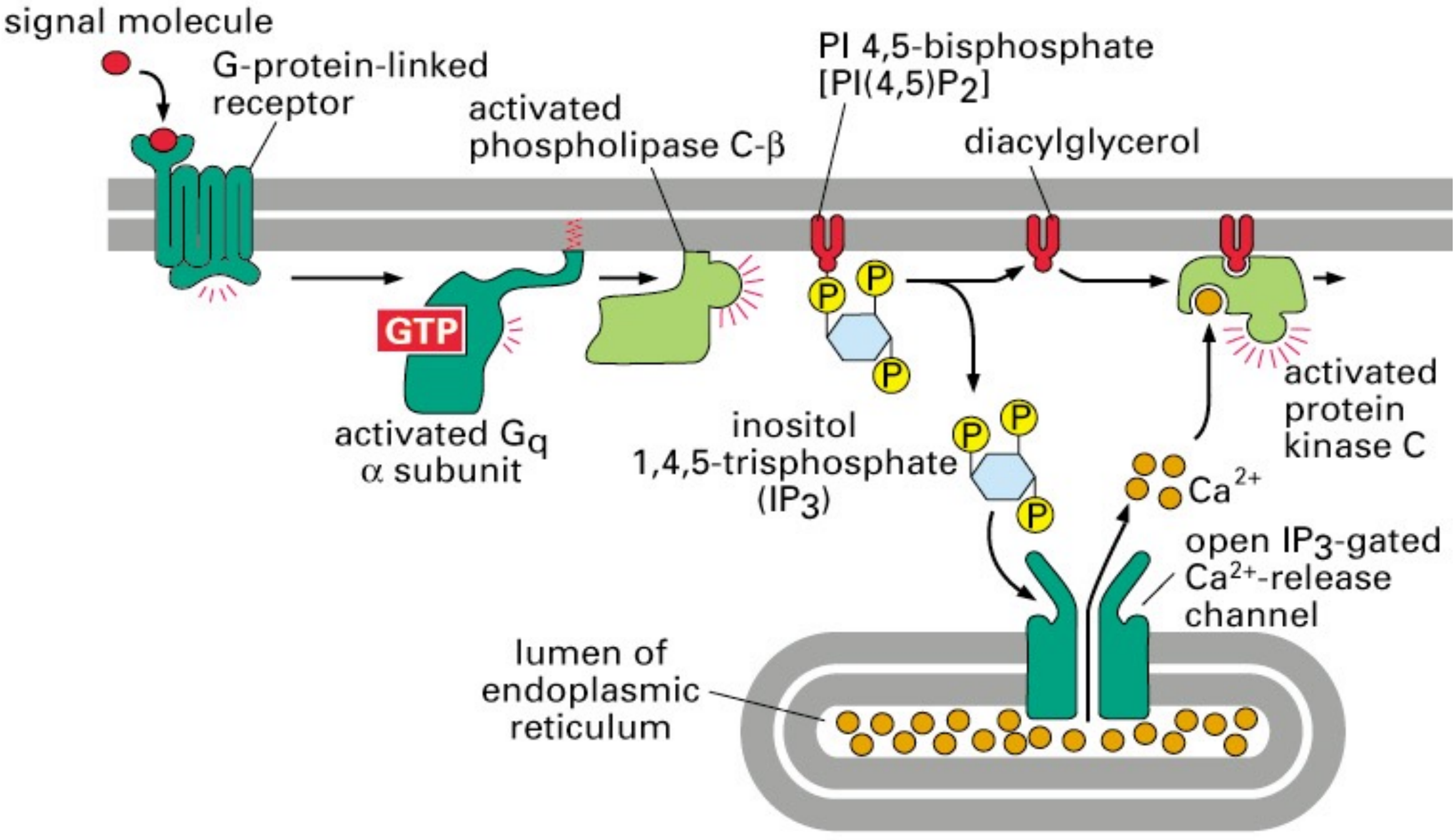
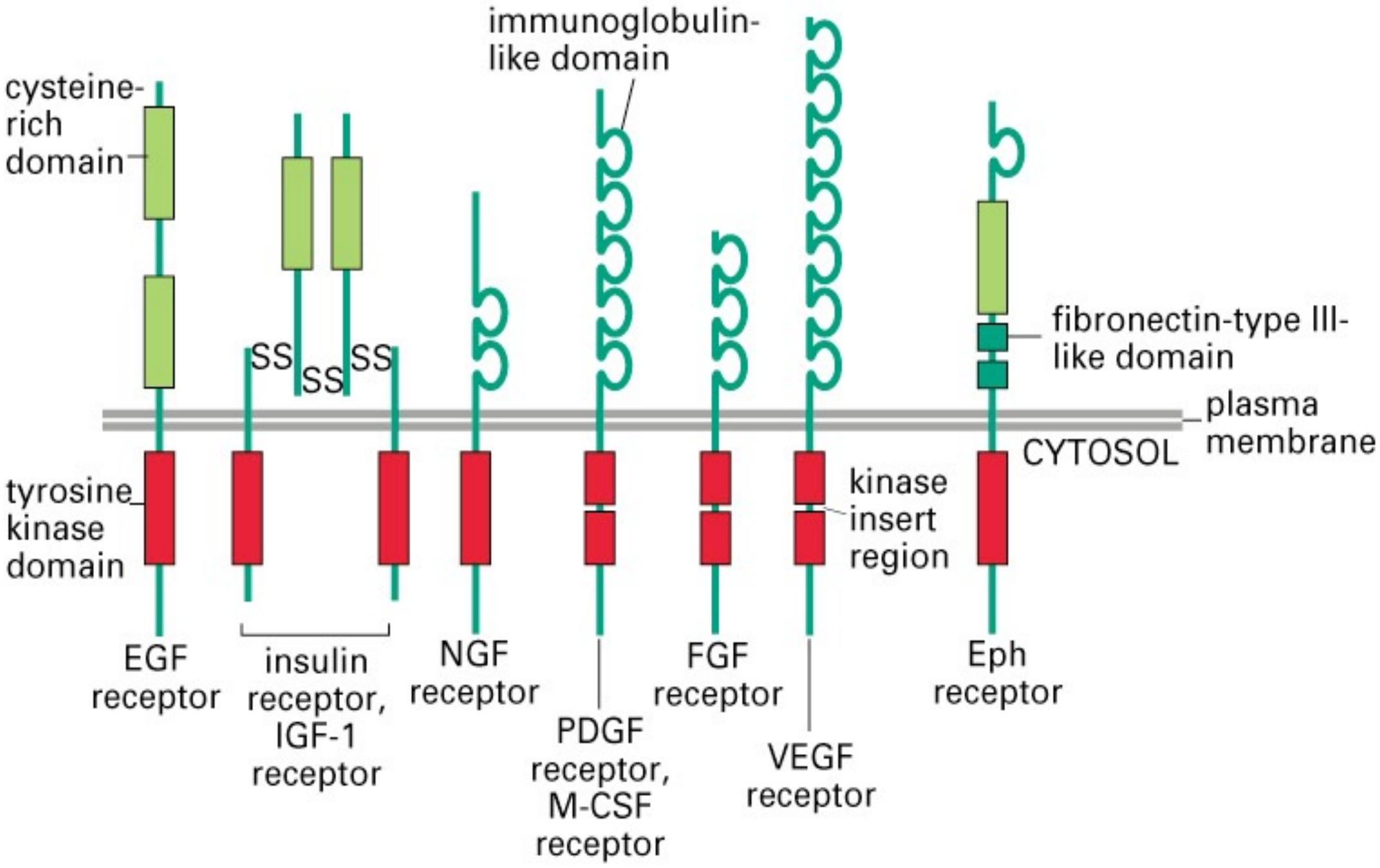


Figure 4.40 Cellular Signal Processing (© Garland Science 2009)

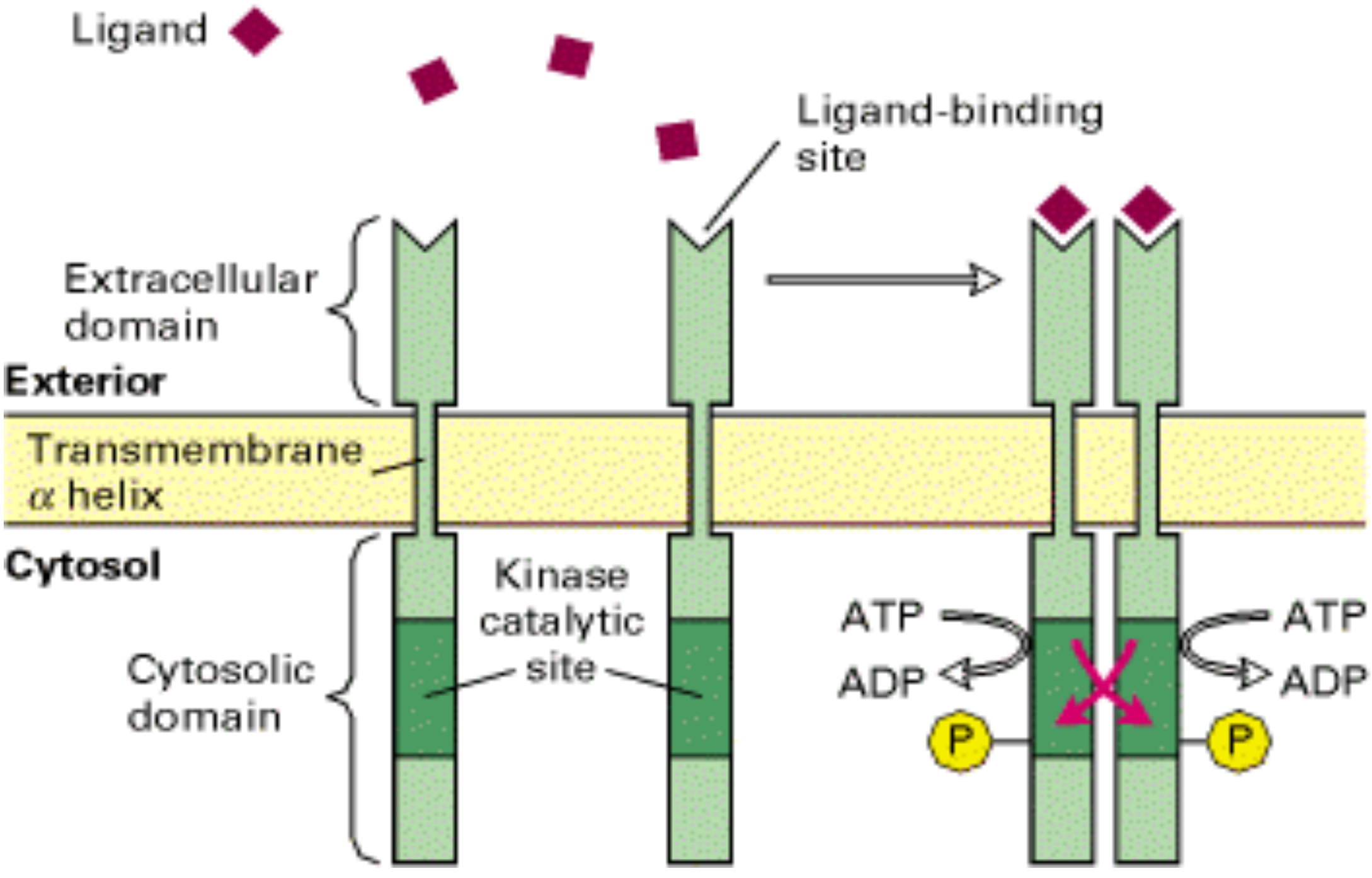
Protein Kinase C (PKC) is activated by inositol phospholipid pathway



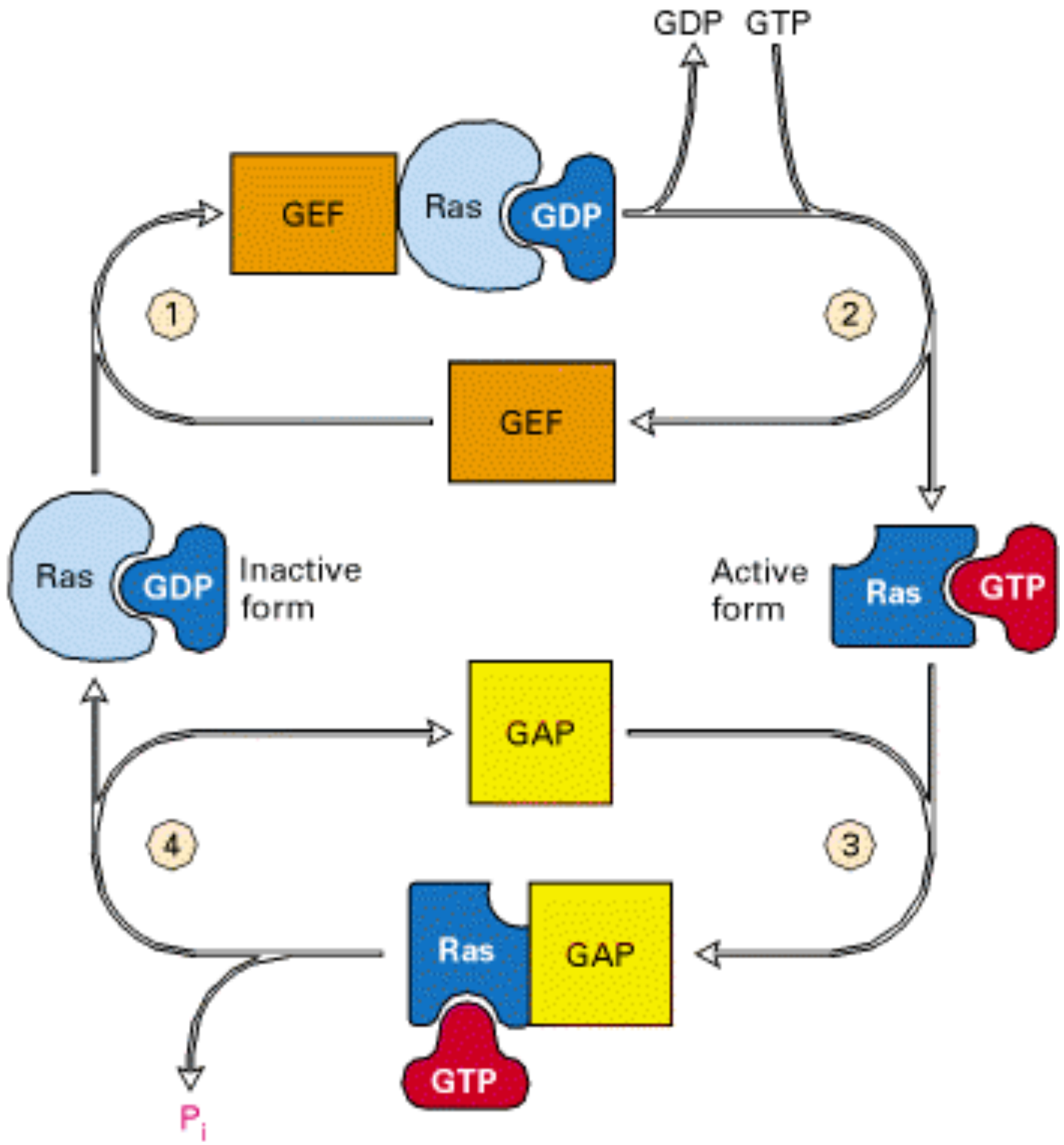
Receptor Tyrosine Kinases (RTKs)



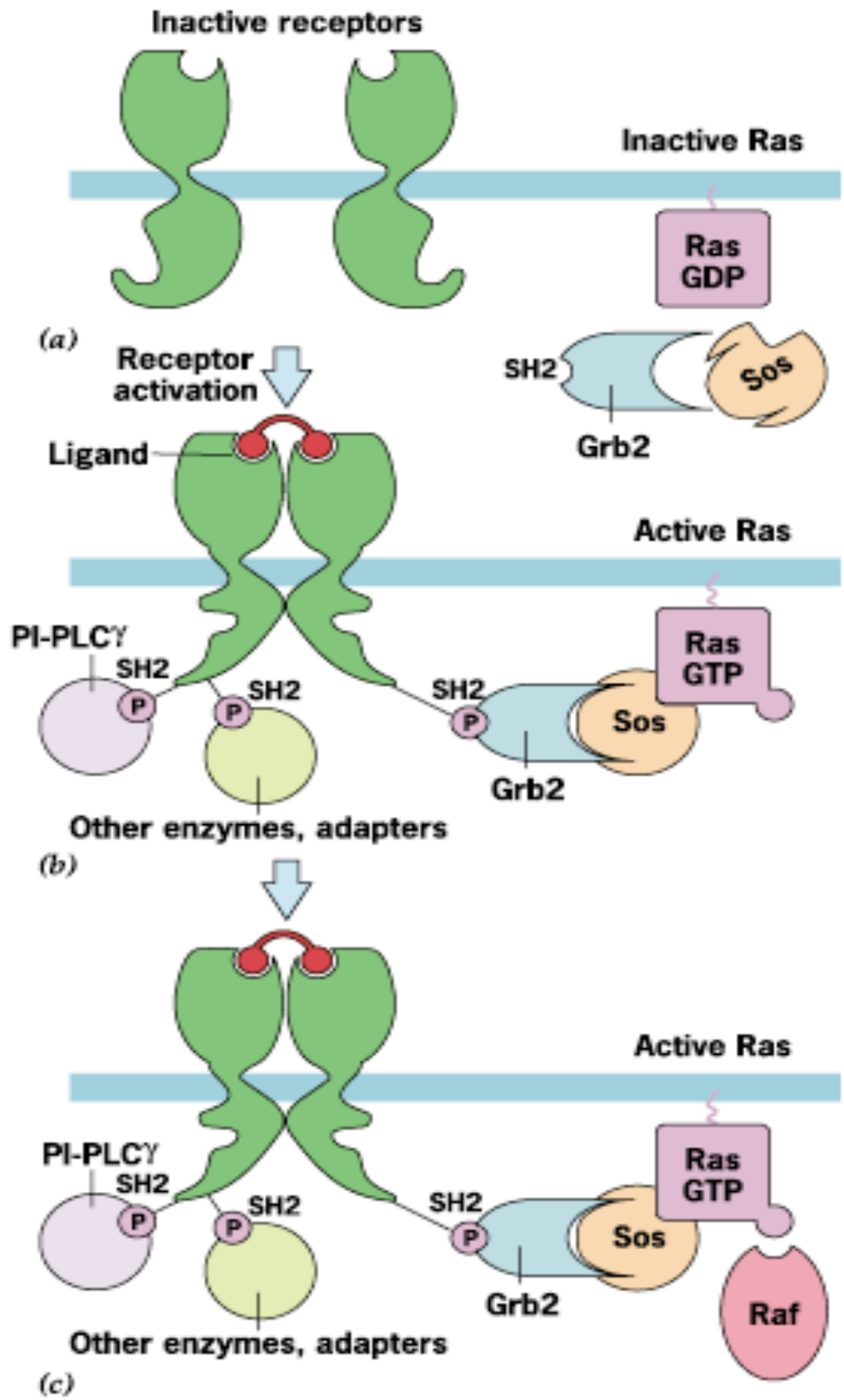
Activation of RTKs



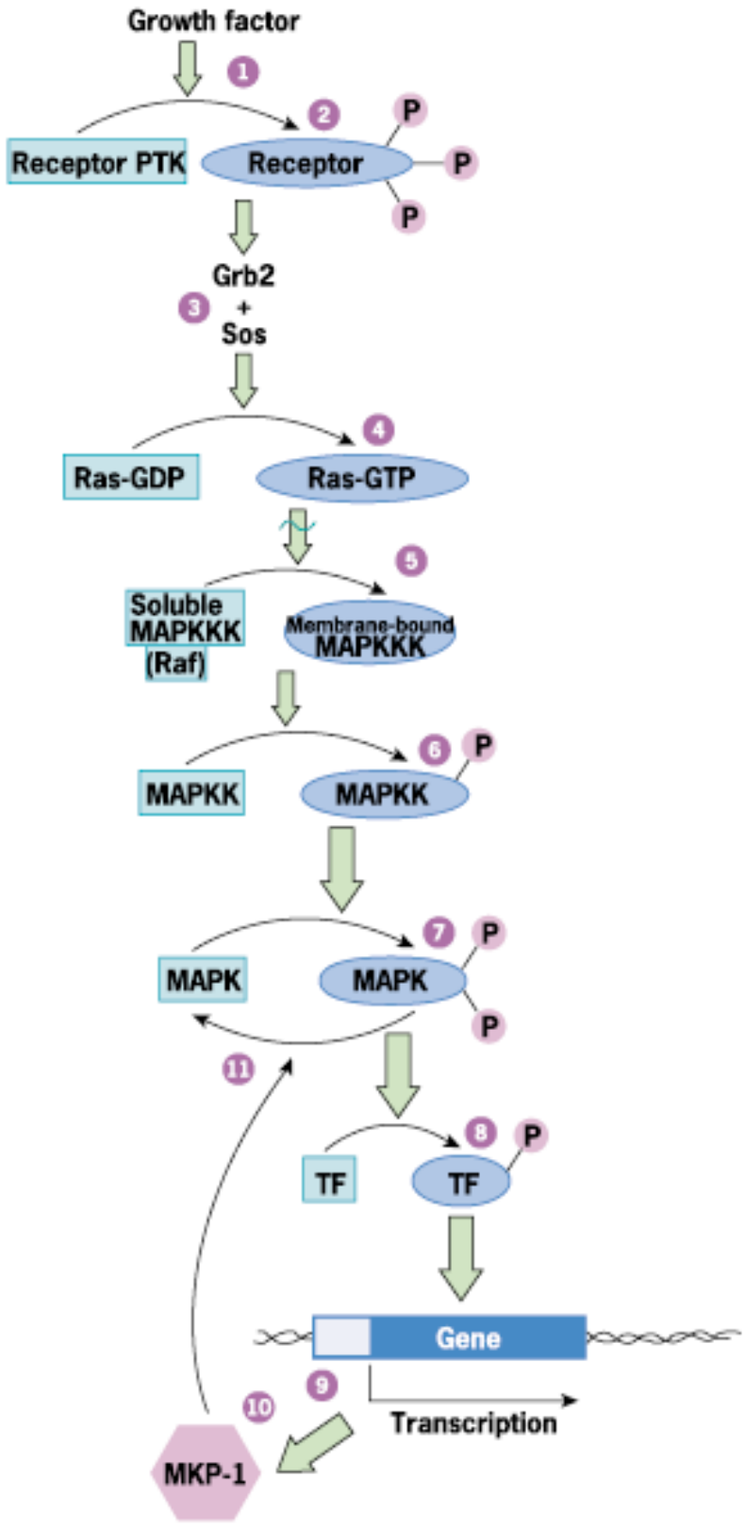
Ras function downstream of RTKs



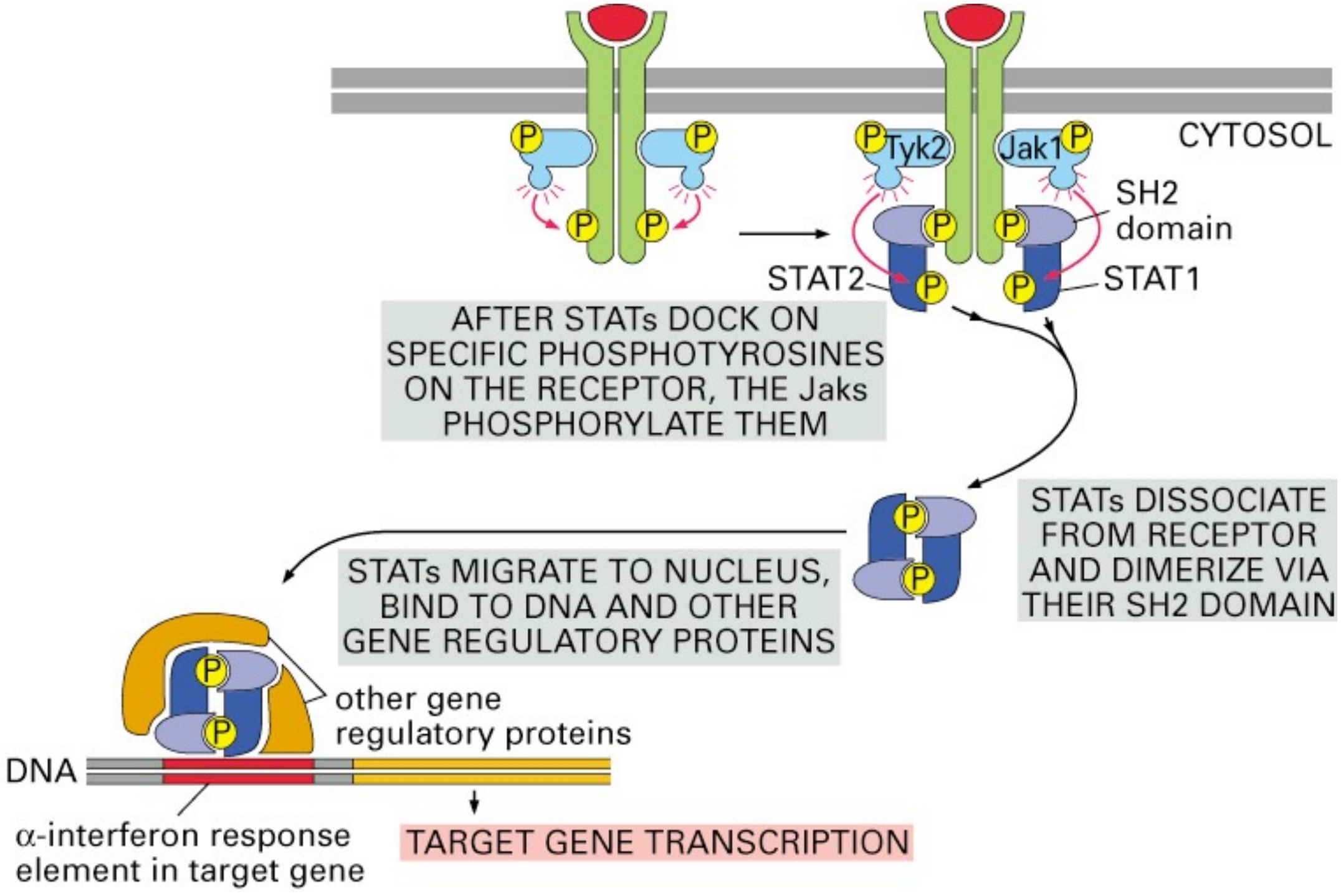
Activation of Ras by RTKs



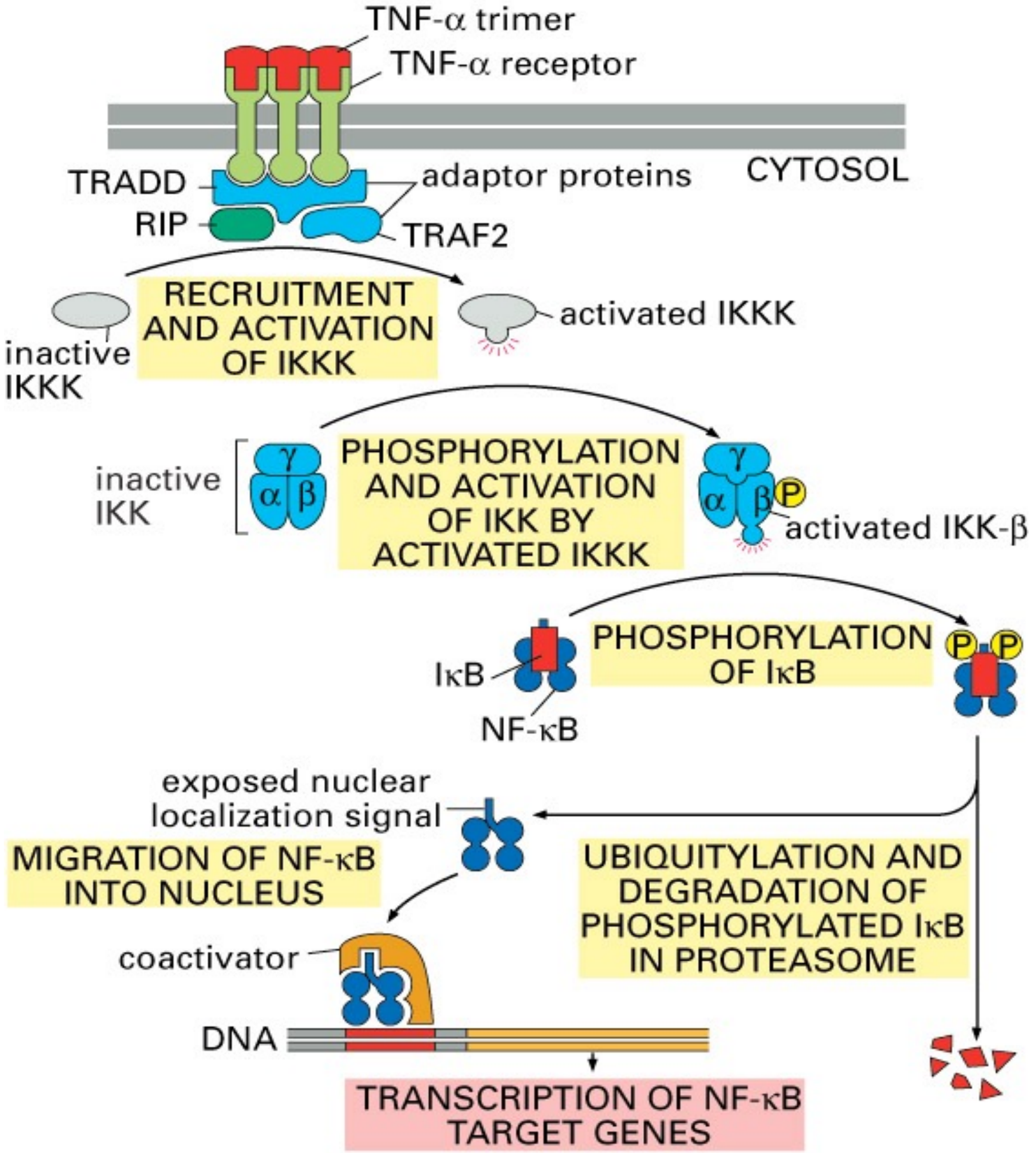
Ras activates MAP Kinase Cascade



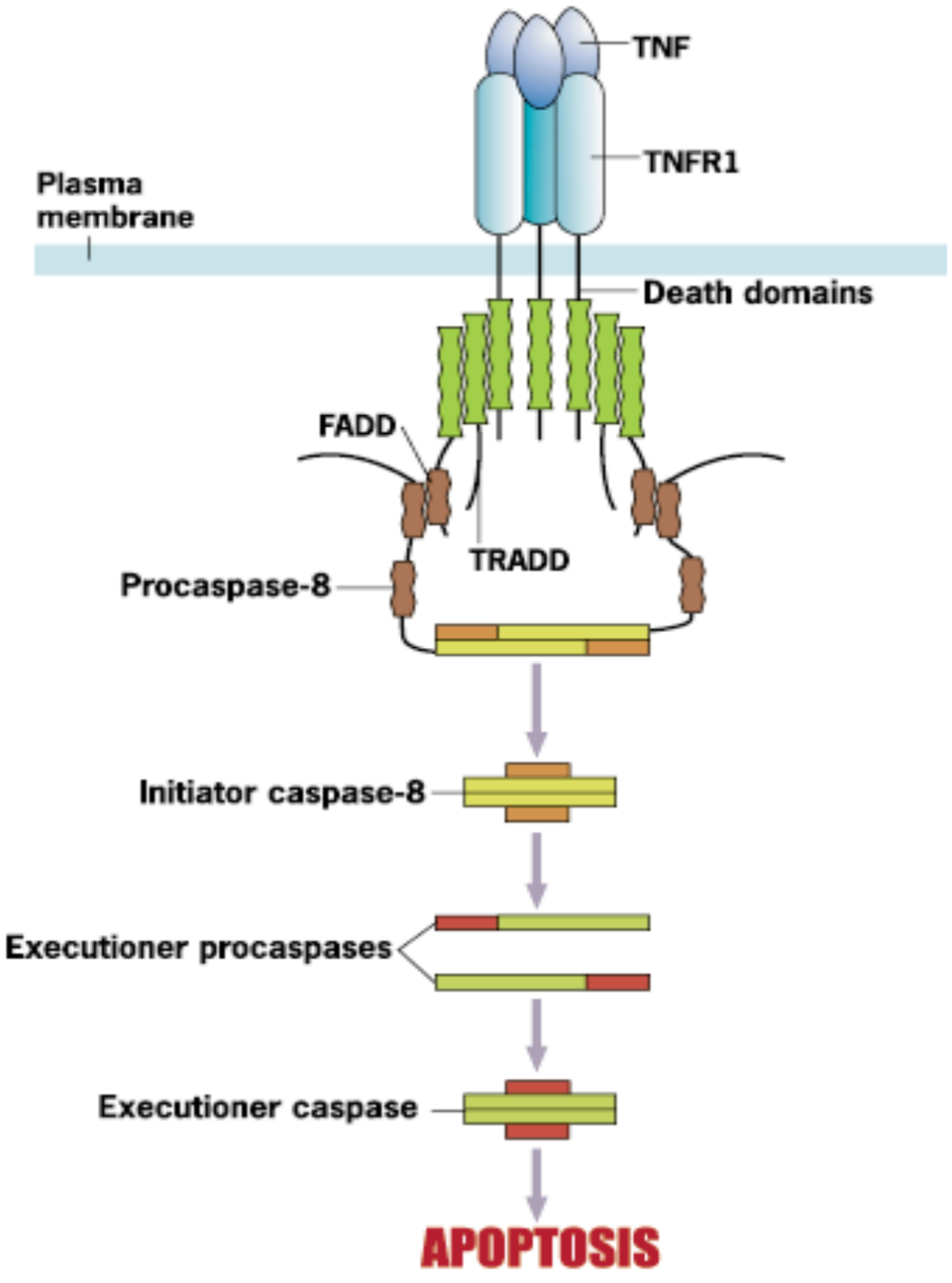
Activation of Jak-STAT pathway by Cytokine Receptors

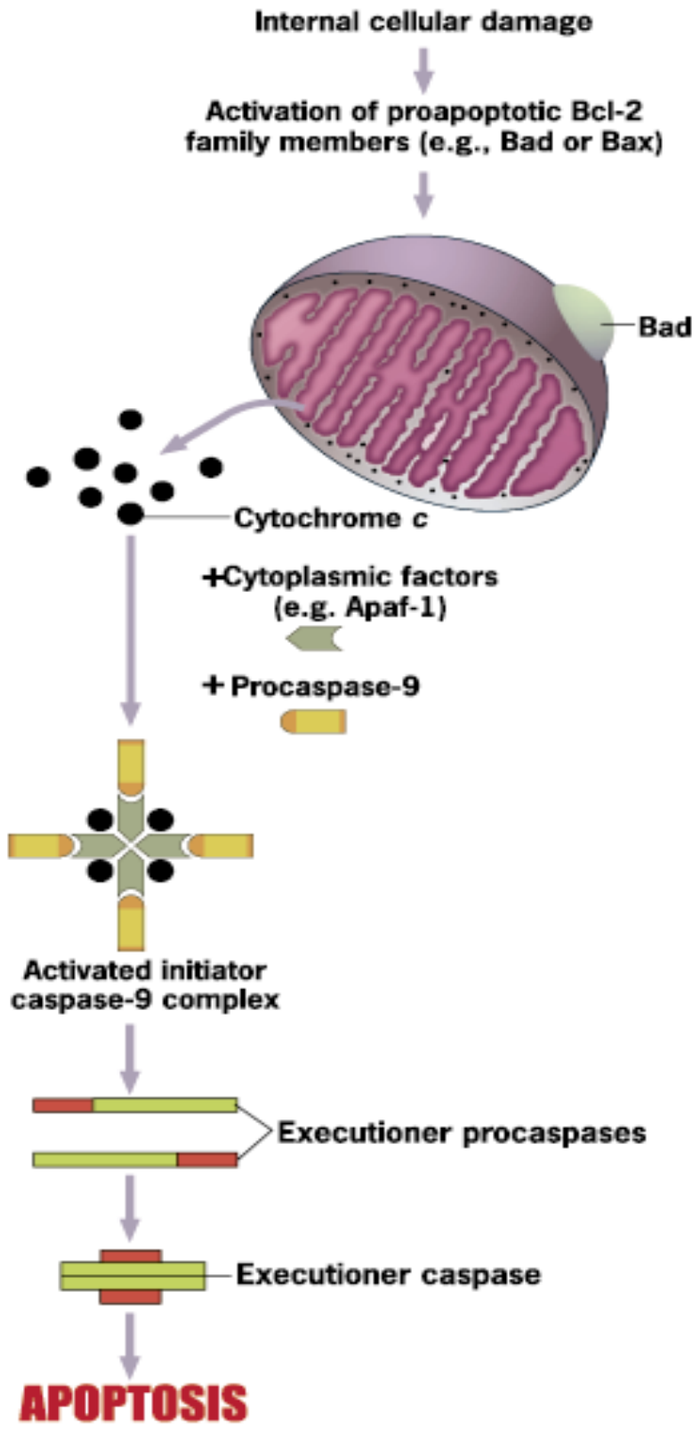


TNF-α signaling Pathway



Apoptotic Pathway





Signaling from contacts between cell surface and the substratum

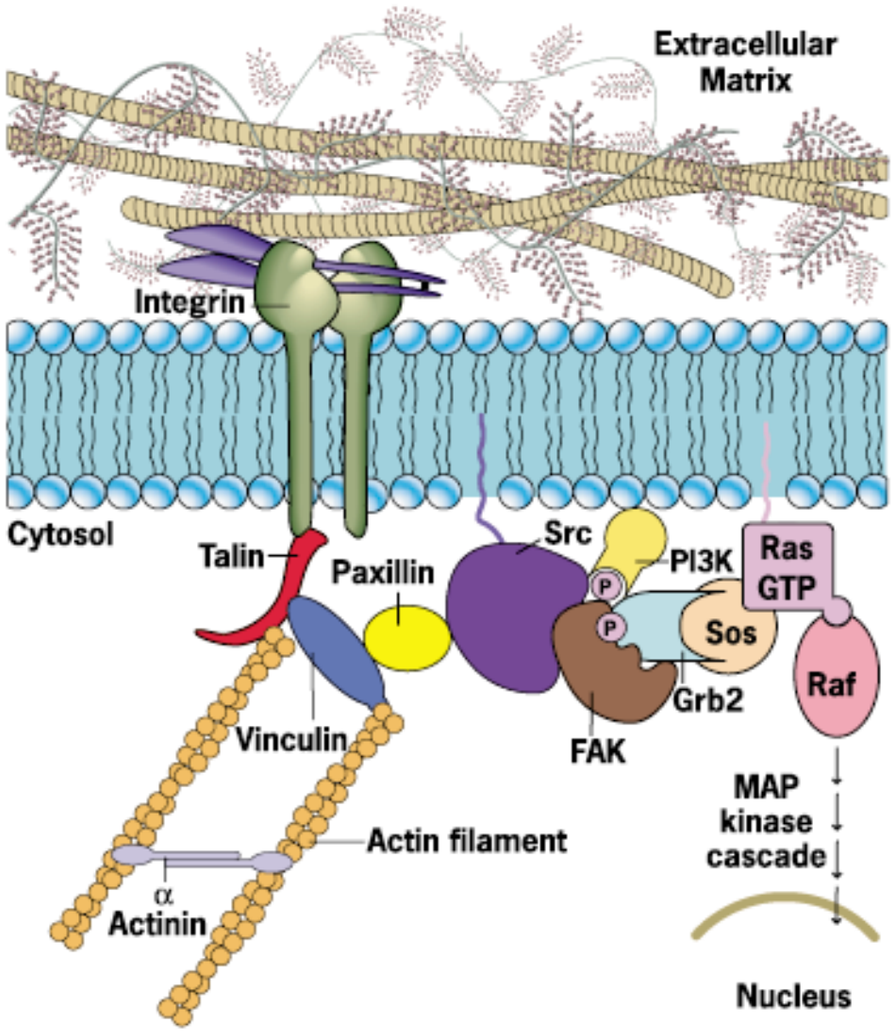


Figure 15.28 Schematic model of the protein-protein interactions of a focal adhesion complex.

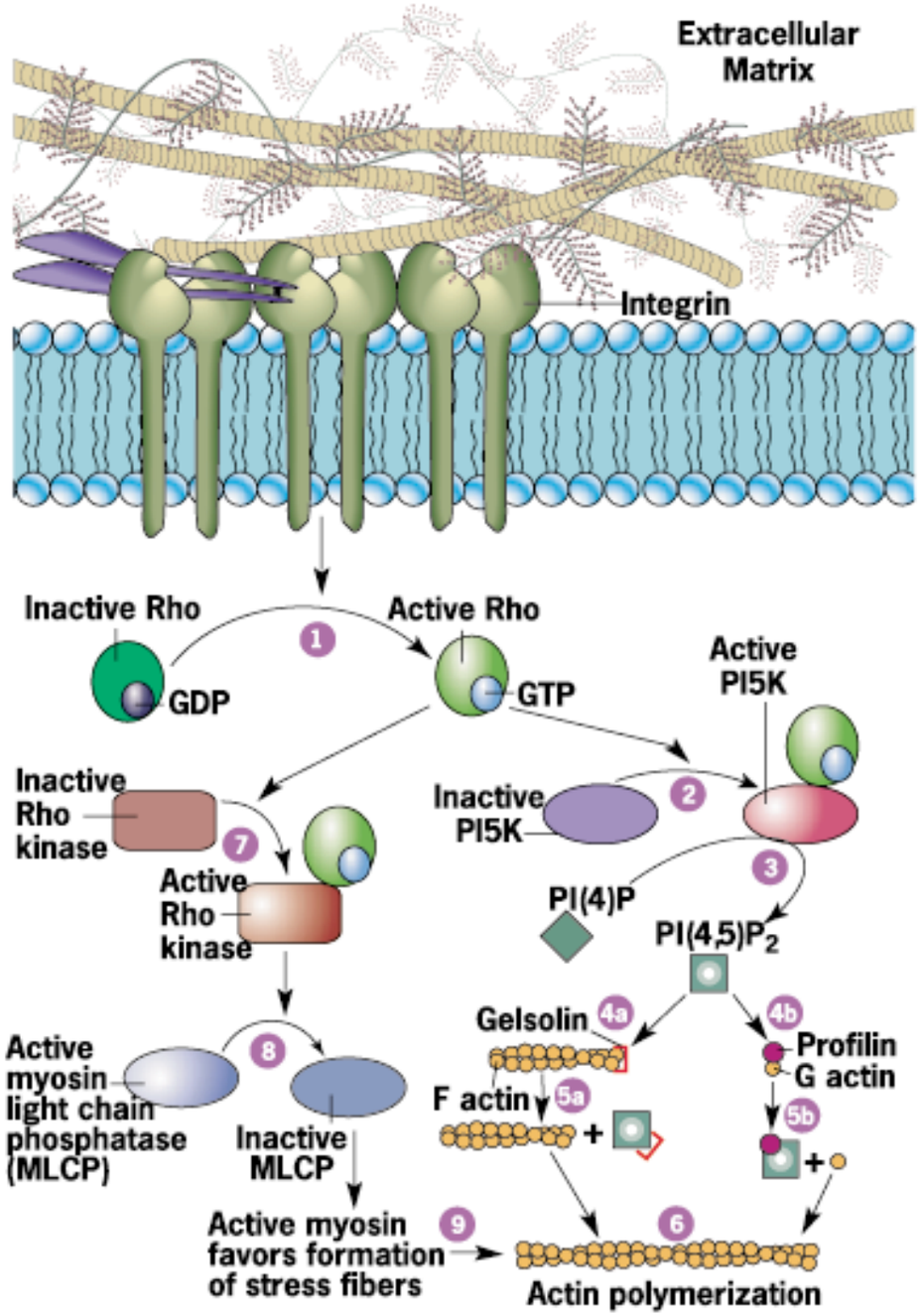
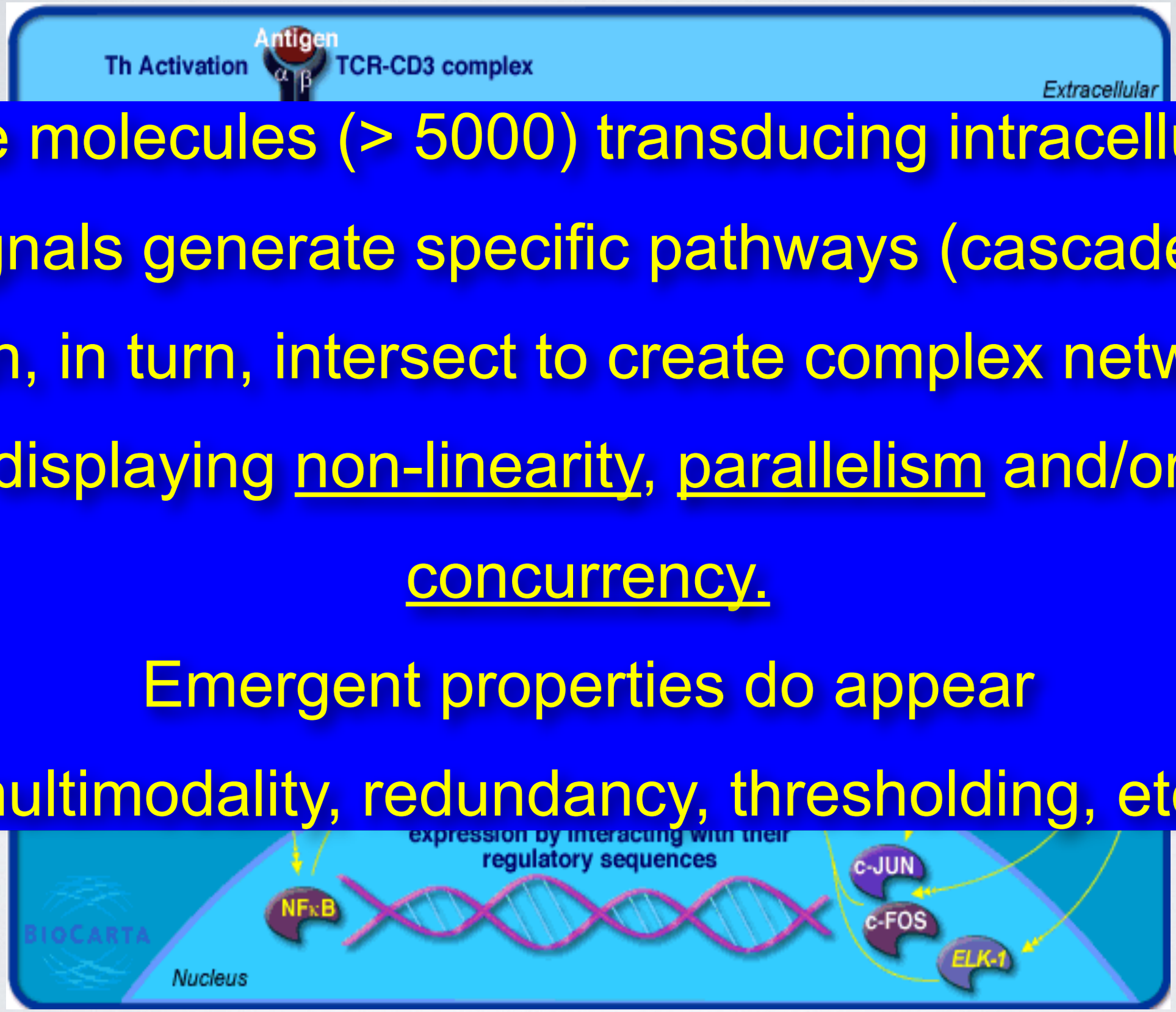


Figure 15.30 Signal pathways that lead to the assembly of the stress fibers of a focal adhesion.

TcR: an example of signalling network

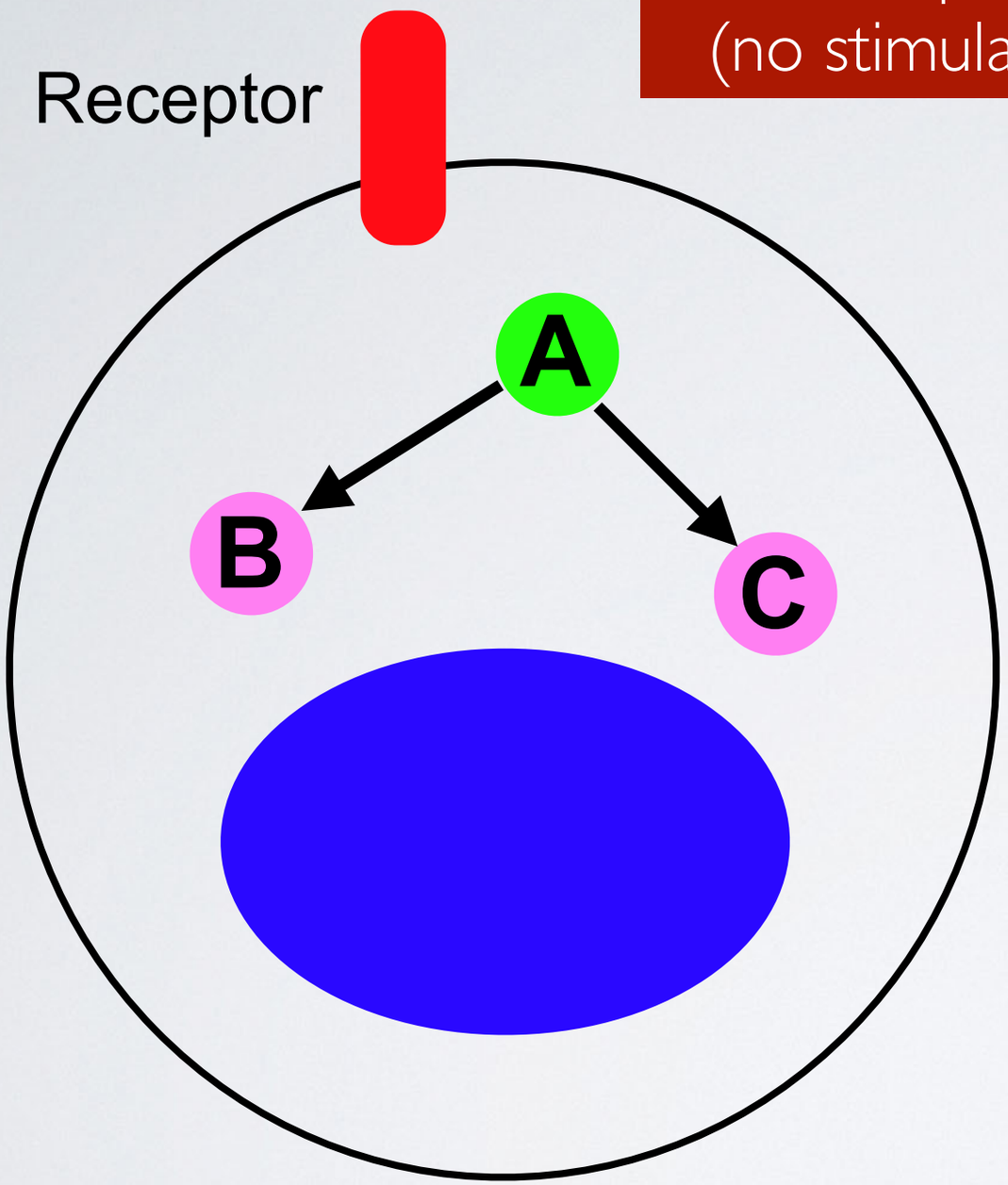


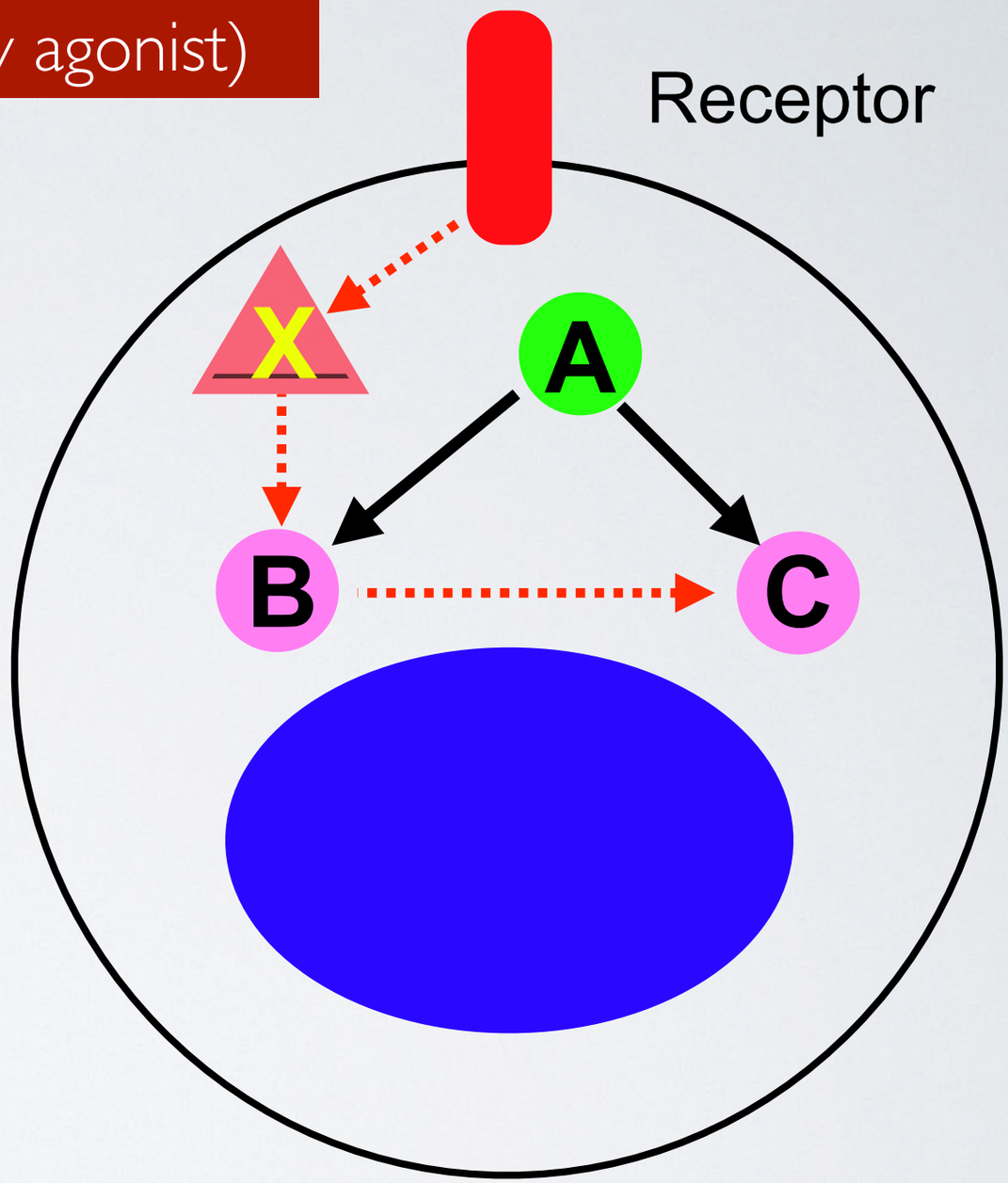
The molecules (> 5000) transducing intracellular signals generate specific pathways (cascades) which, in turn, intersect to create complex networks displaying non-linearity, parallelism and/or concurrency.

Emergent properties do appear (multimodality, redundancy, thresholding, etc.)

The concept of “potentiated” signalling

NON potentiated cells
(no stimulation by agonist)

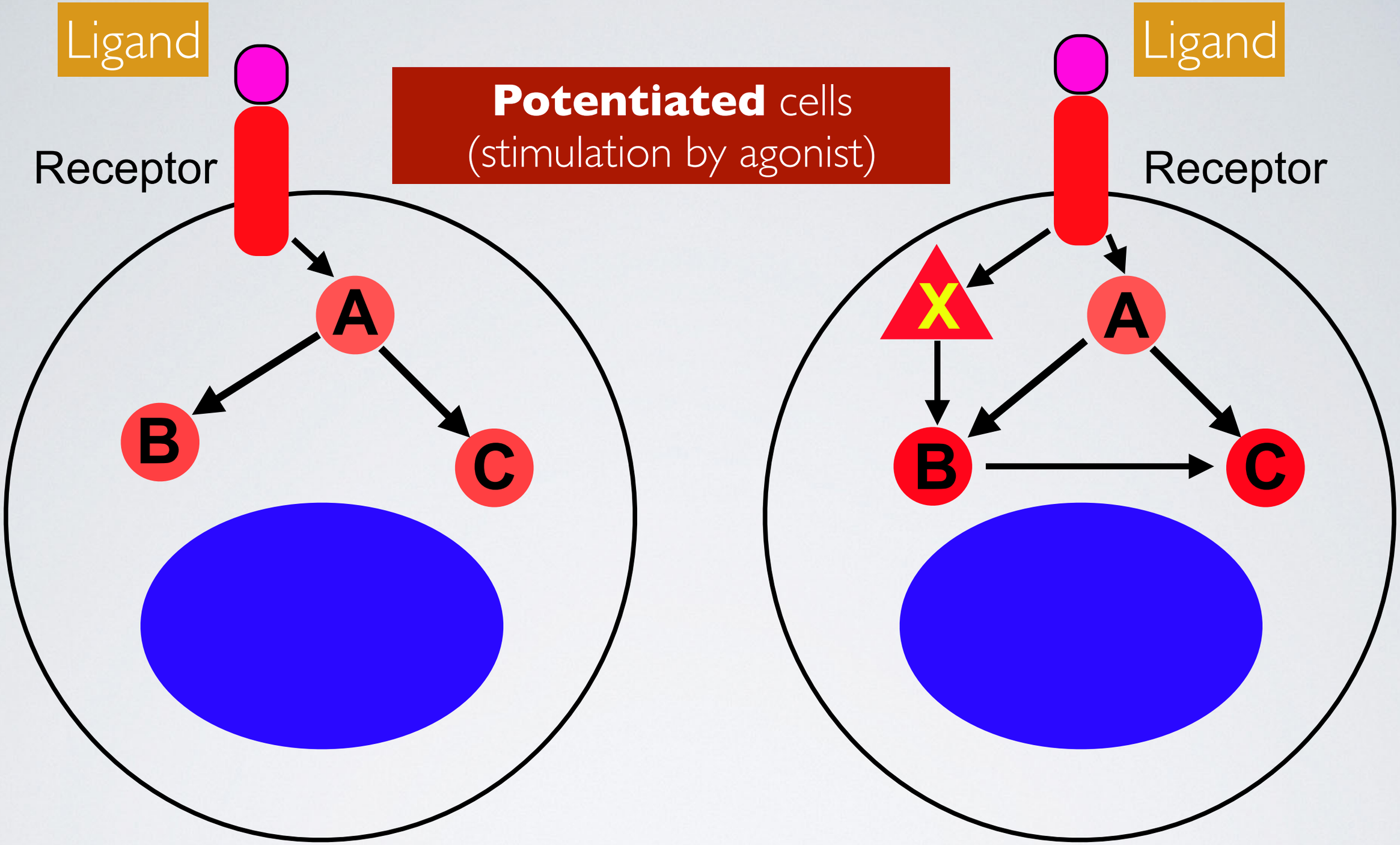




Normal cell **=** **Cancer cell**

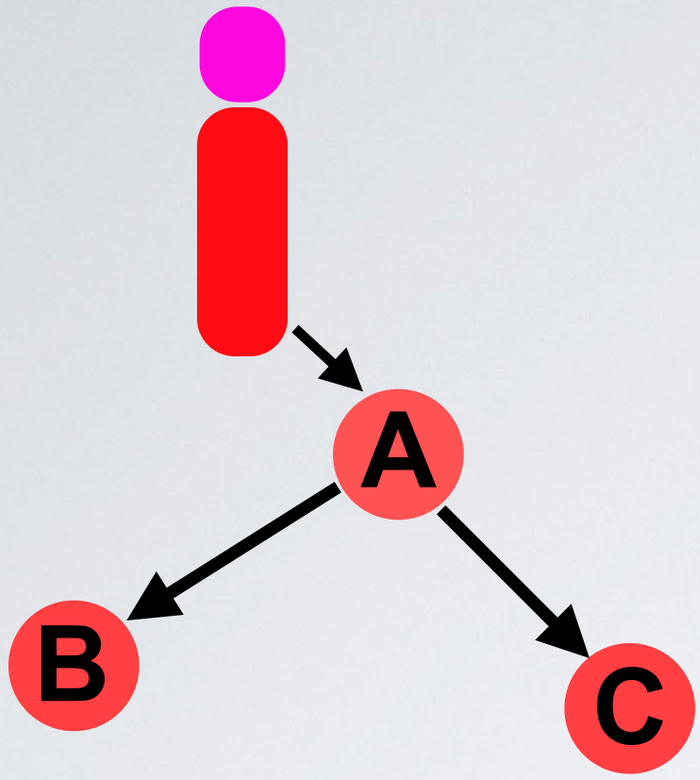
<u>Basal</u> (resting) phosphorylation (housekeeping)

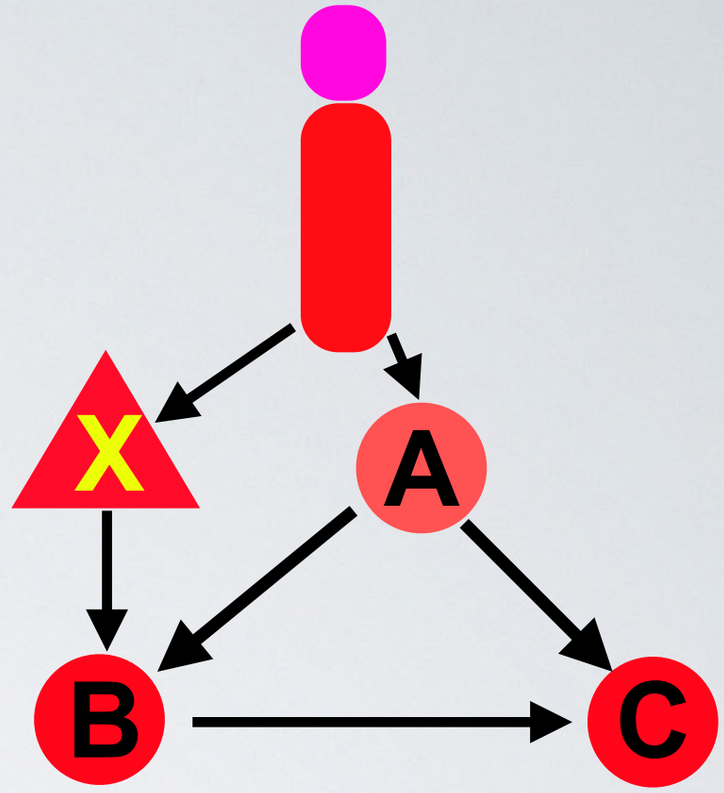
Potentiated cells
(stimulation by agonist)



Normal cell **Cancer cell**

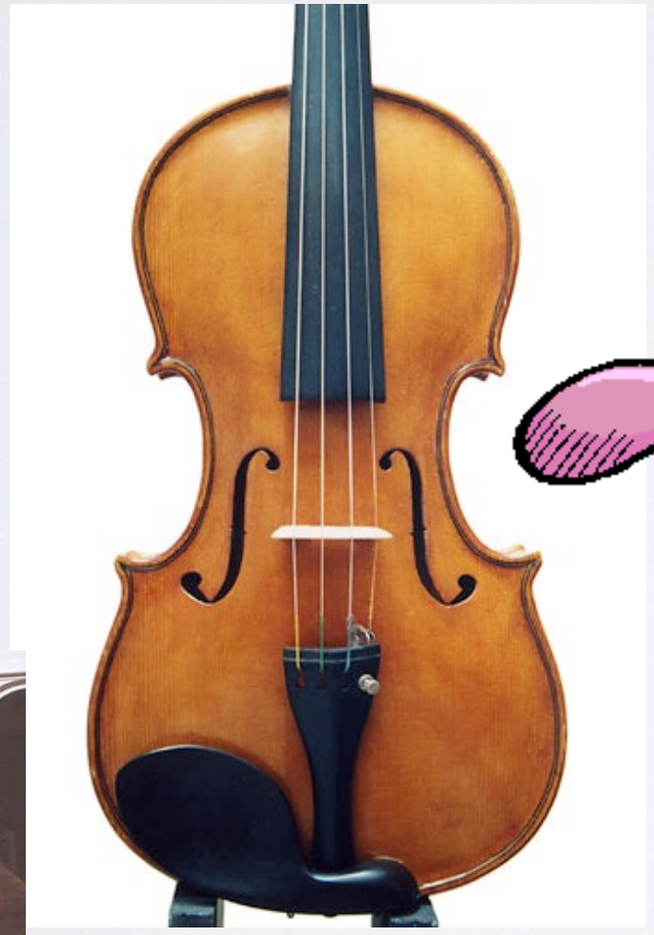
<u>Potentiated</u> phosphorylation
(quantitatively and qualitatively different)





	normale	neoplastica			
A	10	10			
B	20	20	***No***		
C	35	35	*stimolo*		
X		0			
A	20	20			
B	50	100	***SI***		
C	60	140	*stimolo*		
X		30			

The basic concept consists in the fact that in the **potentiated** cells (stimulated) any differences between different cells, either from different subjects or between normal versus pathological cells, are greatly **“enhanced"!**



Normal

Stradivari

Example (4):
Analysis of transduction network alteration in neoplastic cells

Cell, Vol. 118, 217–228, July 23, 2004, Copyright ©2004 by Cell Press

Single Cell Profiling of Potentiated Phospho-Protein Networks in Cancer Cells

Jonathan M. Irish,[1] Randi Hovland,[2,3] Peter O. Krutzik,[1] Omar D. Perez,[1] Øystein Bruserud,[3,4] Bjørn T. Gjertsen,[3,4] and Garry P. Nolan[1,*]

[1]Department of Microbiology & Immunology
Baxter Laboratory of Genetic Pharmacology
Stanford University
Stanford, California 94305

[2]Center for Medical Genetics and Molecular Medicine
Haukeland University Hospital and Proteomic Unit (PROBE)
University of Bergen
Bergen
Norway

[3]Department of Internal Medicine
Haematology Section
Haukeland University Hospital
Bergen
Norway

[4]Institute of Medicine
Haematology Section
University of Bergen
Bergen
Norway

Summary

Altered growth factor responses in phospho-protein-driven signaling networks are crucial to cancer cell survival and pathology. Profiles of cancer cell signaling networks might therefore identify mechanisms by which such cells interpret environmental cues for continued growth. Using multiparameter flow cytometry, we monitored phospho-protein responses to environmental cues in acute myeloid leukemia at the single cell level. By exposing cancer cell signaling networks to potentiating inputs, rather than relying upon the basal levels of protein phosphorylation alone, we could discern unique cancer network profiles that correlated with genetics and disease outcome. Strikingly, individual cancers manifested multiple cell subsets with unique network profiles, reflecting cancer heterogeneity at the level of signaling response. The results revealed a dramatic remodeling of signaling networks in cancer cells. Thus, single cell measurements of phospho-protein responses reveal shifts in signaling potential of a phospho-protein network, allowing for categorizing of cell network phenotypes by multidimensional molecular profiles of signaling.

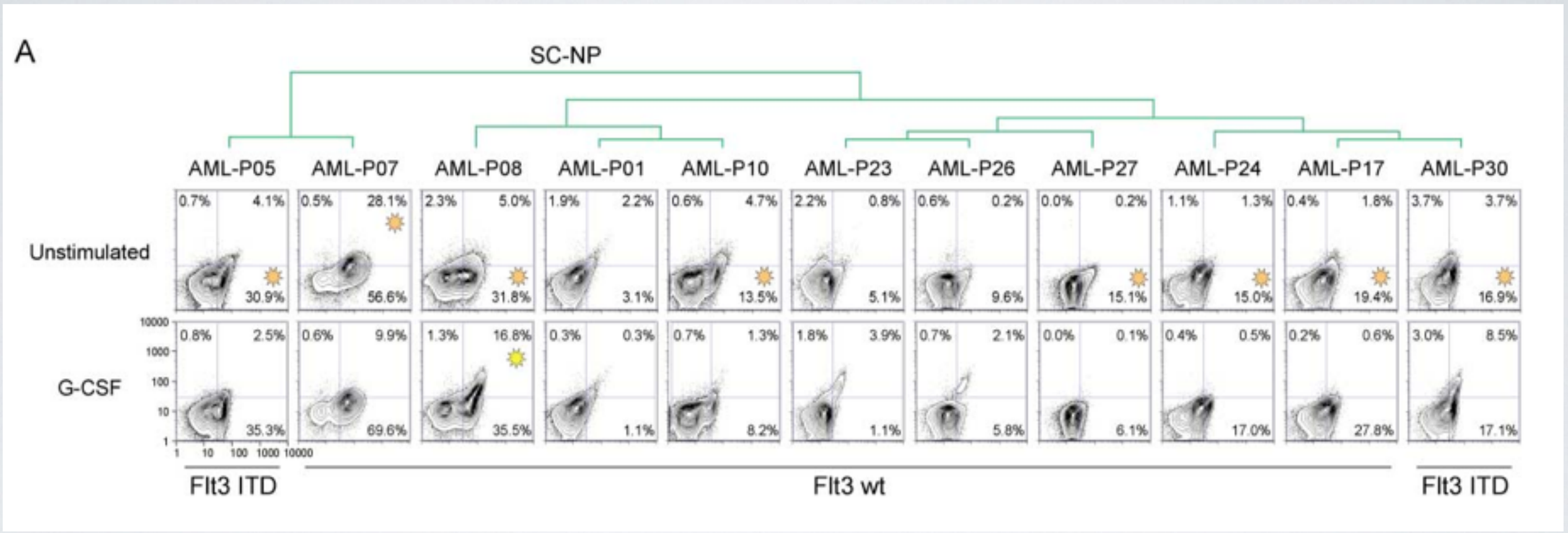
Introduction

Intracellular signaling and interpretation of environmental cues play central roles in cancer cell initiation and maintenance. Actions that lead to cancer cell progression include mutations to key signaling proteins as well as epigenetic changes to gene expression patterns (Hanahan and Weinberg, 2000). Cancer genesis occurs in a stepwise progression, has underlying stochastic elements, and is reflective of the genetic selection of the cancer in the face of both immune system action and the environmental requirements of cancer cells. The signaling profile of any given cancer cell is therefore the sum of numerous influences: epigenetic, genetic, and microenvironmental. The current molecular understanding of cancer signaling rests largely on extrapolations from studies of cell lines and as such is not adequately representative of the signaling phenotypes of a complex population of cancer cells in the body. In contrast, the heterogeneity of cancer cell responses to therapy can be thought of as mirroring the signaling differences that have arisen during evolution of the cancer cell population in the body. Until now, this cell by cell information on cancer cell populations—required to model signaling network pathologies that relate to cancer cell subsets and disease progression—has not been available for analysis.

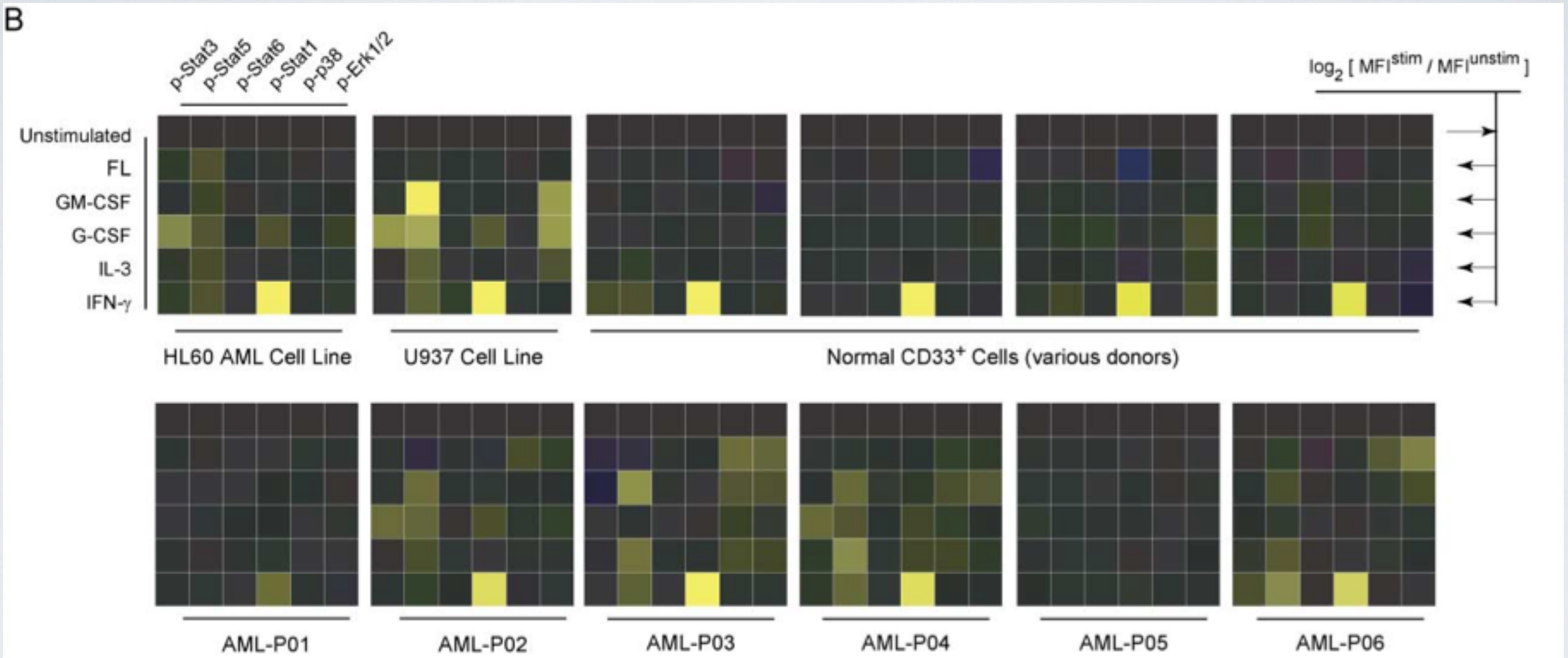
Phospho-protein members of signaling cascades, and the kinases and phosphatases that interact with them, are required to initiate and regulate proliferative signals within cells. It might be predicted that genetic changes common in cancers, such as receptor tyrosine kinase mutations and other signaling-related cytogenetic alterations (Spiekermann et al., 2002; Wheatley et al., 1999), would change the potential of pre-existing signaling networks to respond to external stimuli and lead to identifiable patterns of signal transduction associated with gene mutation. For instance, acute myeloid leukemia (AML) is a cancer wherein dysregulated growth and inhibition of apoptosis lead to the accumulation of immature myeloid progenitor cells and oncogenic progression (Lowenberg et al., 1999). Two key parallel signal transduction networks active in cells that are considered progenitors of AML (Reya et al., 2001) are the STAT pathway (Coffer et al., 2000; Smithgall et al., 2000) and the Ras/MAPK pathway (Platanias, 2003). Several reports suggest that STATs, such as Stat3 and Stat5, are constitutively activated in AML (Benekli et al., 2002; Birkenkamp et al., 2001; Turkson and Jove, 2000; Xia et al., 1998). But, a causal link between basal STAT phosphorylation and leukemogenesis in primary patient material has not been demonstrated, despite significant evidence implicating these proteins in oncogenic processes (Benekli et al., 2003; Bowman et al., 2000; Buettner et al., 2002; Calo et al., 2003; Nieborowska-Skorska et al., 1999). Thought to act upstream of these pathways, abnormalities of the Flt3 (*f*ms-*l*ike *t*yrosine kinase *3*) receptor tyrosine kinase are detected in approximately 30% of AML patients and are well established as a negative prognostic indicator in AML (Gilliland and Griffin, 2002; Kottaridis et al., 2001; Thiede et al., 2002). Expression of mutant, activated Flt3 in cell lines has been observed to activate STAT and Ras/MAPK signaling (Hayakawa et al., 2000; Mizuki et al., 2000). However, basal levels of Stat5 phosphorylation have been reported to

*Correspondence: gnolan@stanford.edu

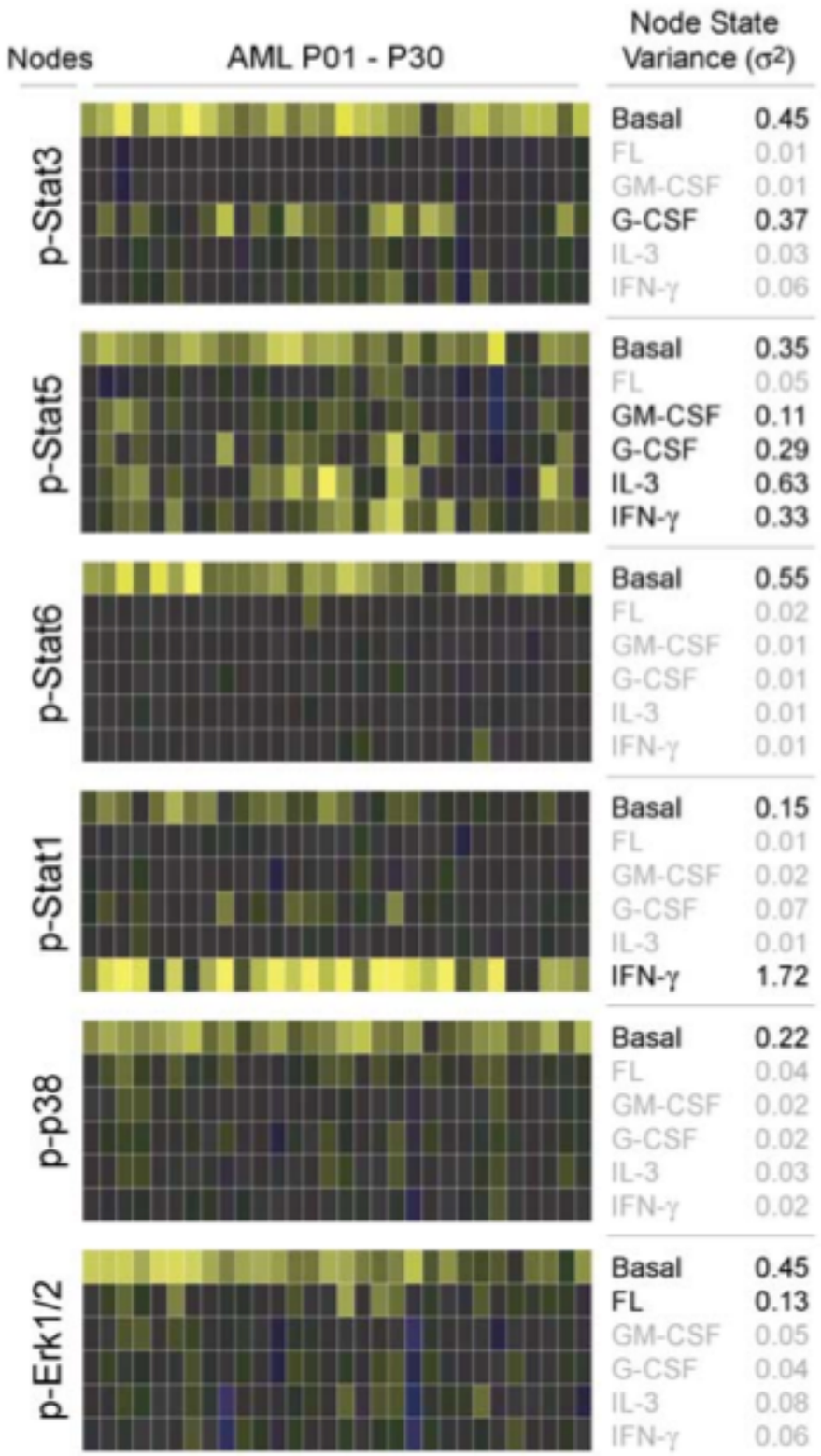
Single cell cyto-fluorimetric profiling of "enhanced" signal transduction in individual patients



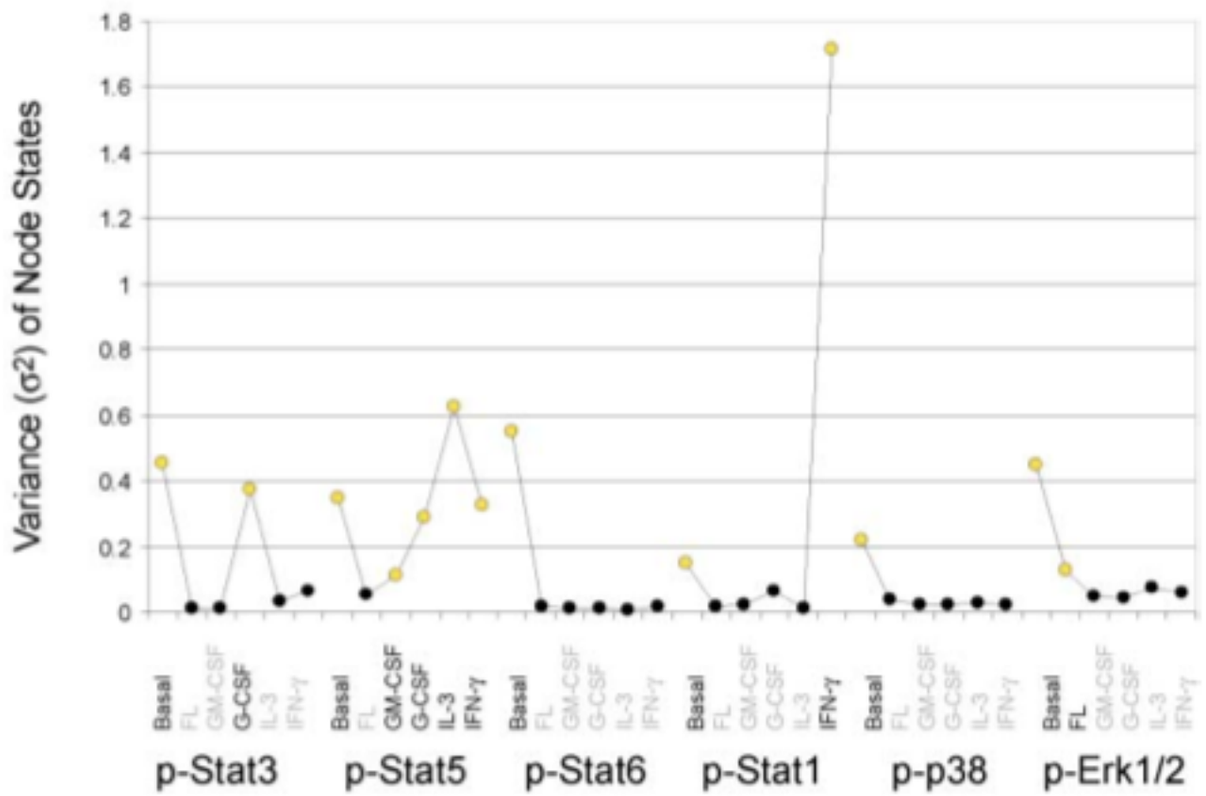
B


p-Stat3
p-Stat5
p-Stat6
p-Stat1
p-p38
p-Erk1/2
log2 [MFIstim / MFIunstim]
Unstimulated
FL
GM-CSF
G-CSF
IL-3
IFN-γ
HL60 AML Cell Line
U937 Cell Line
Normal CD33+ Cells (various donors)
AML-P01
AML-P02
AML-P03
AML-P04
AML-P05
AML-P06

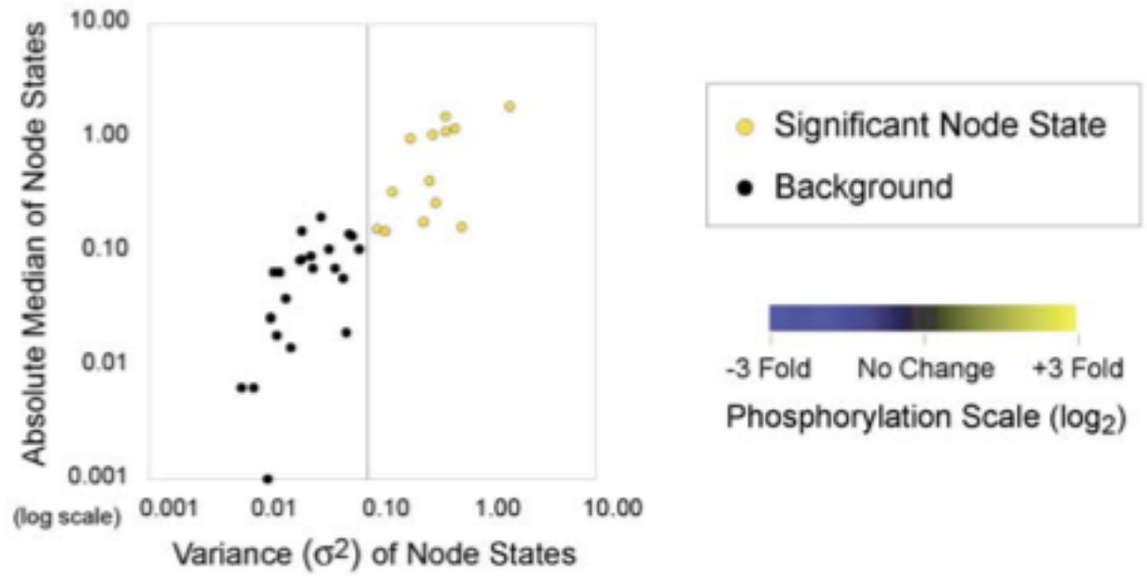
A

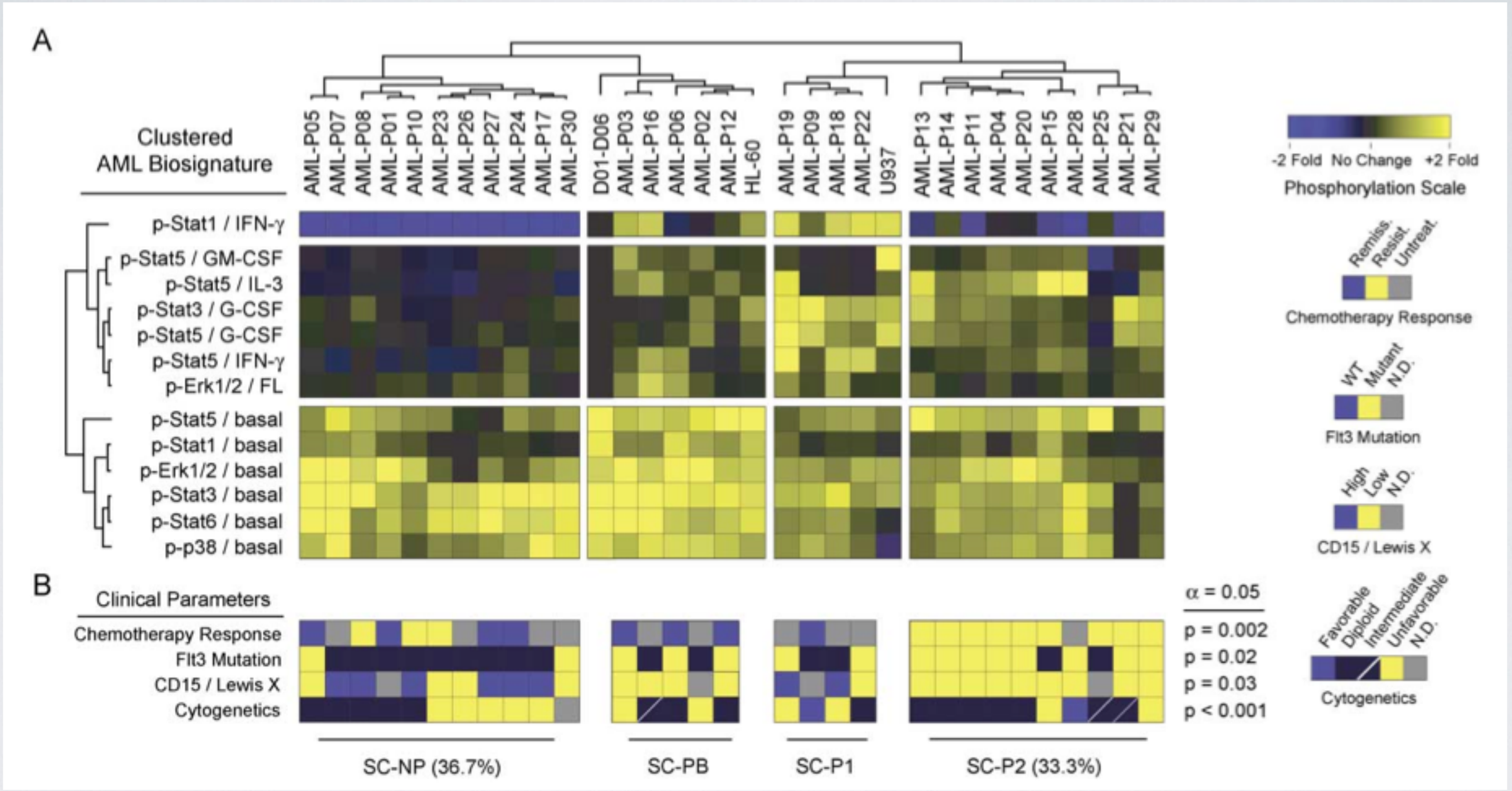

Nodes
AML P01 - P30
Node State Variance (σ²)
p-Stat3
Basal 0.45
FL 0.01
GM-CSF 0.01
G-CSF 0.37
IL-3 0.03
IFN-γ 0.06
p-Stat5
Basal 0.35
FL 0.05
GM-CSF 0.11
G-CSF 0.29
IL-3 0.63
IFN-γ 0.33
p-Stat6
Basal 0.55
FL 0.02
GM-CSF 0.01
G-CSF 0.01
IL-3 0.01
IFN-γ 0.01
p-Stat1
Basal 0.15
FL 0.01
GM-CSF 0.02
G-CSF 0.07
IL-3 0.01
IFN-γ 1.72
p-p38
Basal 0.22
FL 0.04
GM-CSF 0.02
G-CSF 0.02
IL-3 0.03
IFN-γ 0.02
p-Erk1/2
Basal 0.45
FL 0.13
GM-CSF 0.05
G-CSF 0.04
IL-3 0.08
IFN-γ 0.06

B

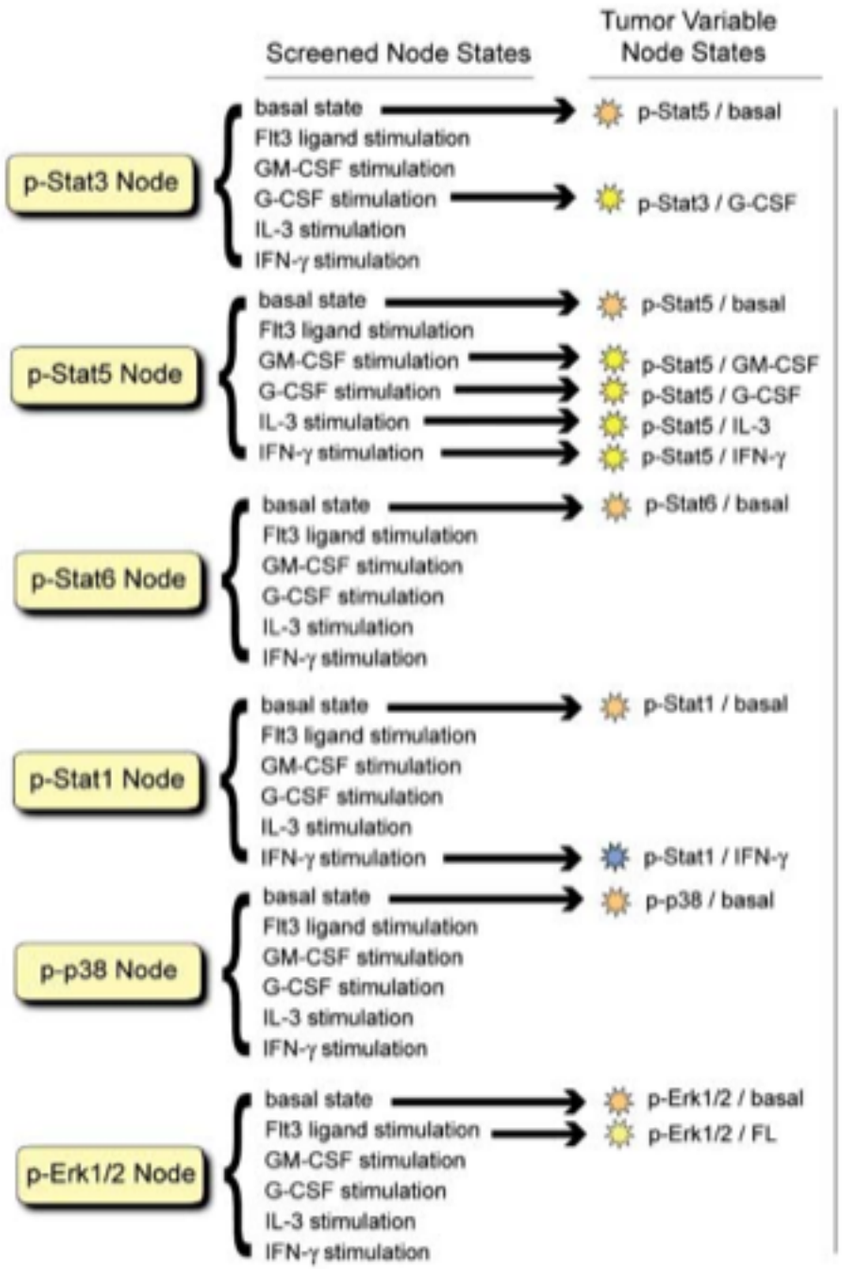

Variance of Node States Across 30 AML Samples
Variance (σ²) of Node States
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Basal FL GM-CSF G-CSF IL-3 IFN-γ
p-Stat3
p-Stat5
p-Stat6
p-Stat1
p-p38
p-Erk1/2

C

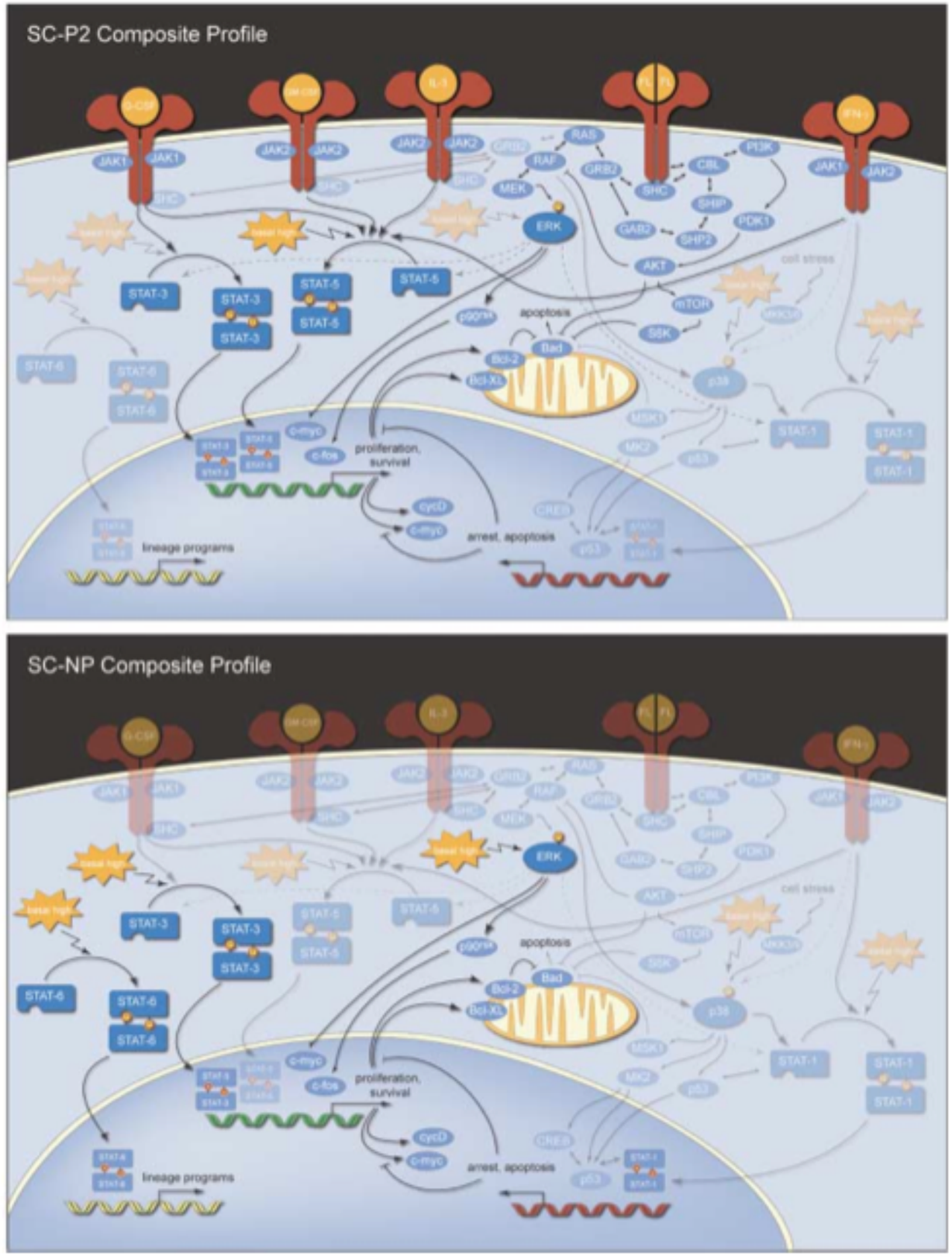

Absolute Median of Node States
10.00
1.00
0.10
0.01
0.001
(log scale)
0.001
0.01
0.10
1.00
10.00
Variance (σ²) of Node States
Significant Node State
Background
-3 Fold
No Change
+3 Fold
Phosphorylation Scale (log₂)


A
Clustered
AML Biosignature
AML-P05
AML-P07
AML-P08
AML-P01
AML-P10
AML-P23
AML-P26
AML-P27
AML-P24
AML-P17
AML-P30
D01-D06
AML-P03
AML-P16
AML-P06
AML-P02
AML-P12
HL-60
AML-P19
AML-P09
AML-P18
AML-P22
U937
AML-P13
AML-P14
AML-P11
AML-P04
AML-P20
AML-P15
AML-P28
AML-P25
AML-P21
AML-P29
p-Stat1 / IFN-γ
p-Stat5 / GM-CSF
p-Stat5 / IL-3
p-Stat3 / G-CSF
p-Stat5 / G-CSF
p-Stat5 / IFN-γ
p-Erk1/2 / FL
p-Stat5 / basal
p-Stat1 / basal
p-Erk1/2 / basal
p-Stat3 / basal
p-Stat6 / basal
p-p38 / basal
-2 Fold No Change +2 Fold
Phosphorylation Scale
Remiss.
Resist.
Untreat.
Chemotherapy Response
WT
Mutant
N.D.
Flt3 Mutation
High
Low
N.D.
CD15 / Lewis X
B
Clinical Parameters
Chemotherapy Response
Flt3 Mutation
CD15 / Lewis X
Cytogenetics
α = 0.05
p = 0.002
p = 0.02
p = 0.03
p < 0.001
Favorable
Diploid
Intermediate
Unfavorable
N.D.
Cytogenetics
SC-NP (36.7%)
SC-PB
SC-P1
SC-P2 (33.3%)

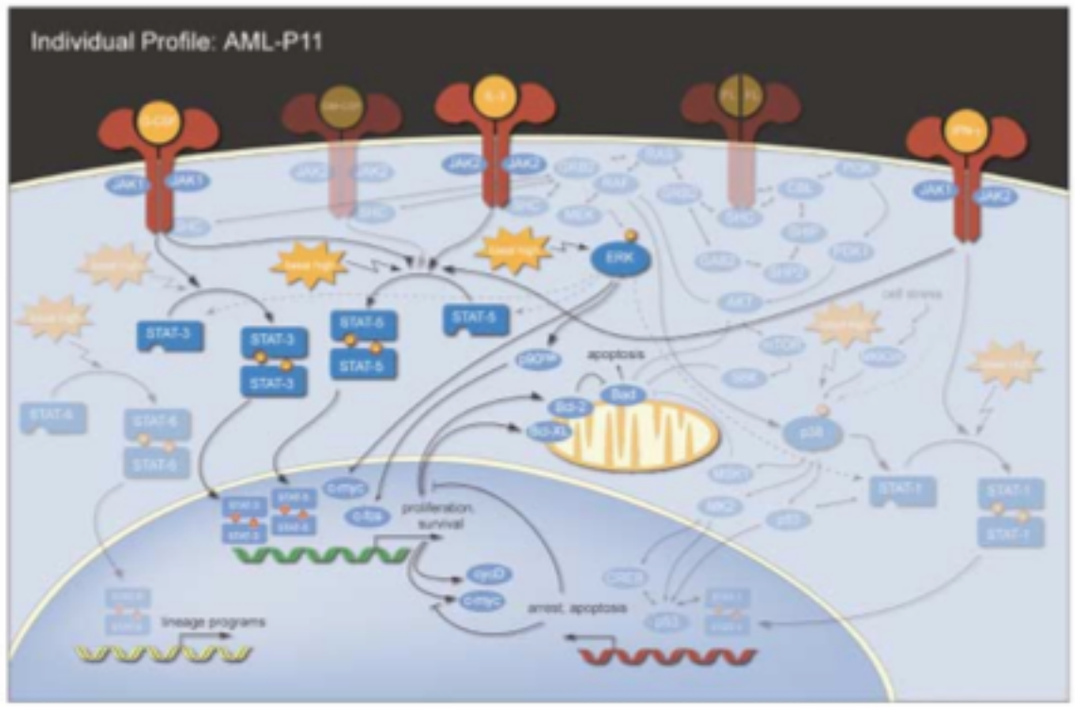
A


Screened Node States
Tumor Variable Node States
p-Stat3 Node
basal state
Flt3 ligand stimulation
GM-CSF stimulation
G-CSF stimulation
IL-3 stimulation
IFN-γ stimulation
p-Stat5 / basal
p-Stat3 / G-CSF
p-Stat5 Node
basal state
Flt3 ligand stimulation
GM-CSF stimulation
G-CSF stimulation
IL-3 stimulation
IFN-γ stimulation
p-Stat5 / basal
p-Stat5 / GM-CSF
p-Stat5 / G-CSF
p-Stat5 / IL-3
p-Stat5 / IFN-γ
p-Stat6 Node
basal state
Flt3 ligand stimulation
GM-CSF stimulation
G-CSF stimulation
IL-3 stimulation
IFN-γ stimulation
p-Stat6 / basal
p-Stat1 Node
basal state
Flt3 ligand stimulation
GM-CSF stimulation
G-CSF stimulation
IL-3 stimulation
IFN-γ stimulation
p-Stat1 / basal
p-Stat1 / IFN-γ
p-p38 Node
basal state
Flt3 ligand stimulation
GM-CSF stimulation
G-CSF stimulation
IL-3 stimulation
IFN-γ stimulation
p-p38 / basal
p-Erk1/2 Node
basal state
Flt3 ligand stimulation
GM-CSF stimulation
G-CSF stimulation
IL-3 stimulation
IFN-γ stimulation
p-Erk1/2 / basal
p-Erk1/2 / FL
Phospho-protein Biosignature of AML

B

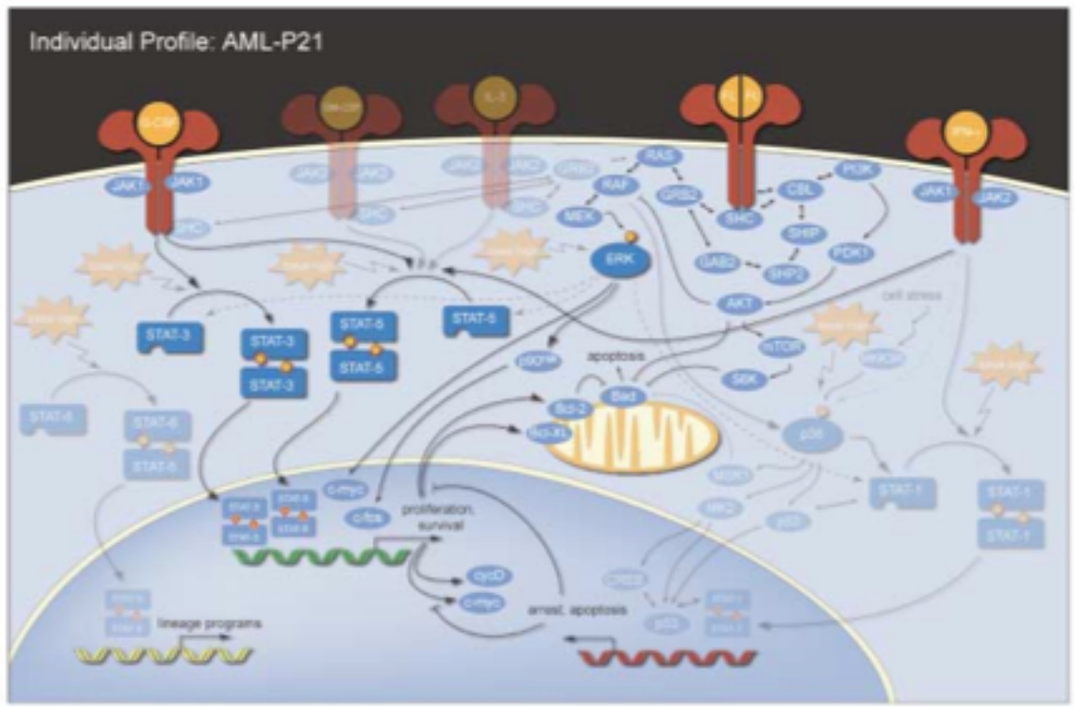

SC-P2 Composite Profile
SC-NP Composite Profile
STAT-3
STAT-5
STAT-6
ERK
AKT
Bcl-2
Bad
Bcl-XL
c-myc
c-fos
proliferation, survival
lineage programs
arrest, apoptosis
p38
STAT-1

A

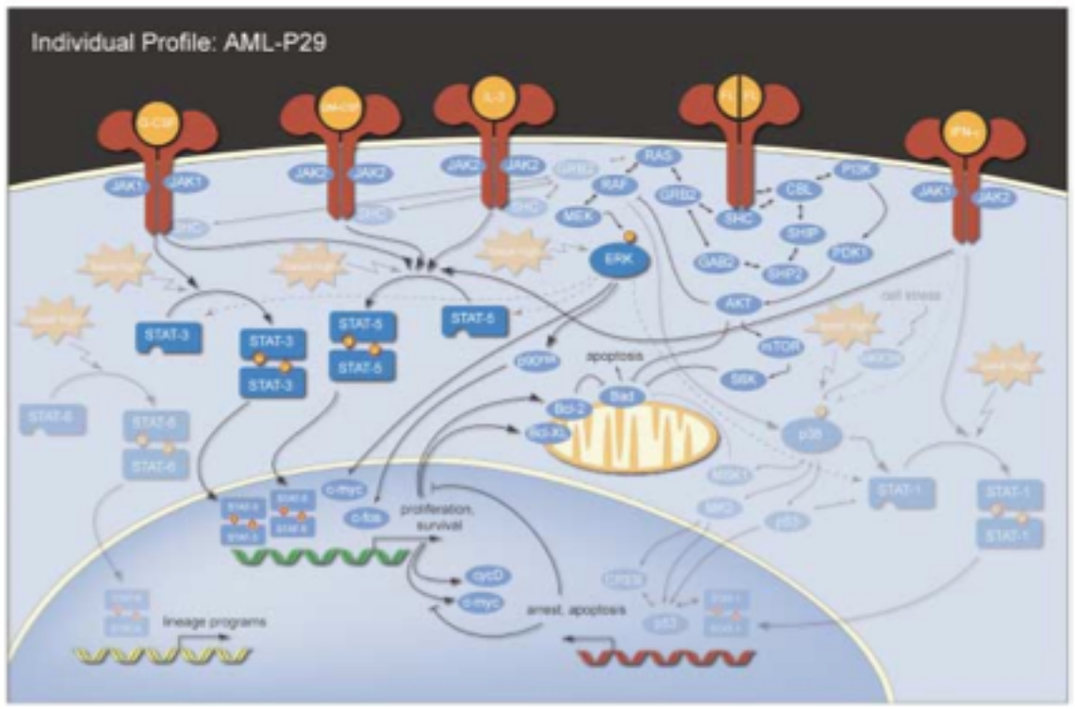


Signaling profile summaries for individual AML patients profiled as SC-P2. Each had detectable Flt3-ITD and resisted course 1 chemotherapy.

B



C



Phospho-proteomics analysis with Cytoscape

B-CLL phospho-set

Example (5): analysis of metabolite network

Molecular Systems Biology 6; Article number 441; doi:10.1038/msb.2010.93
Citation: ***Molecular Systems Biology*** **6:441**

www.molecularsystemsbiology.com

molecular systems biology

REPORT

Metabonomic, transcriptomic, and genomic variation of a population cohort

Michael Inouye[1,2,14,*], Johannes Kettunen[2,14], Pasi Soininen[3,4], Kaisa Silander[5], Samuli Ripatti[6], Linda S Kumpula[4], Eija Hämäläinen[2], Pekka Jousilahti[7], Antti J Kangas[4], Satu Männistö[7], Markku J Savolainen[4,8], Antti Jula[9], Jaana Leiviskä[10], Aarno Palotie[2,5,11,12], Veikko Salomaa[7], Markus Perola[6], Mika Ala-Korpela[3,4,8,13] and Leena Peltonen[2,5,11,12,✠]

[1] Immunology Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia, [2] Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK, [3] NMR Metabonomics Laboratory, Laboratory of Chemistry, Department of Biosciences, University of Eastern Finland, Kuopio, Finland, [4] Computational Medicine Research Group, Institute of Clinical Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland, [5] Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland, [6] Institute for Molecular Medicine (FIMM), University of Helsinki and Unit of Public Health Genomics, National Institute for Health and Welfare, Helsinki, Finland, [7] Unit of Chronic Disease Epidemiology and Prevention, National Institute for Health and Welfare, Helsinki, Finland, [8] Department of Internal Medicine and Biocenter Oulu, Clinical Research Center, University of Oulu, Oulu, Finland, [9] Population Studies Unit, National Institute of Health and Welfare, Turku, Finland, [10] Disease Risk Unit, National Institute for Health and Welfare, Helsinki, Finland, [11] Department of Medical Genetics, University of Helsinki and the Helsinki University Hospital, Helsinki, Finland, [12] The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA and [13] College of Chemistry and Chemical Engineering, Central South University, Changsha, Hunan Province, China
[14] These authors contributed equally to this work
✠ Deceased
* Corresponding author. Immunology Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia. Tel.: +61 3 9345 2555; Fax: +61 3 9347 0852; E-mail: inouye@wehi.edu.au

Received 8.6.10; accepted 17.10.10

Comprehensive characterization of human tissues promises novel insights into the biological architecture of human diseases and traits. We assessed metabonomic, transcriptomic, and genomic variation for a large population-based cohort from the capital region of Finland. Network analyses identified a set of highly correlated genes, the lipid–leukocyte (LL) module, as having a prominent role in over 80 serum metabolites (of 134 measures quantified), including lipoprotein subclasses, lipids, and amino acids. Concurrent association with immune response markers suggested the LL module as a possible link between inflammation, metabolism, and adiposity. Further, genomic variation was used to generate a directed network and infer LL module's largely reactive nature to metabolites. Finally, gene co-expression in circulating leukocytes was shown to be dependent on serum metabolite concentrations, providing evidence for the hypothesis that the coherence of molecular networks themselves is conditional on environmental factors. These findings show the importance and opportunity of systematic molecular investigation of human population samples. To facilitate and encourage this investigation, the metabonomic, transcriptomic, and genomic data used in this study have been made available as a resource for the research community.
Molecular Systems Biology **6**: 441; published online 21 December 2010; doi:10.1038/msb.2010.93
Subject Categories: functional genomics; metabolic and regulatory networks
Keywords: bioinformatics; biological networks; integrative genomics; metabonomics; transcriptomics

Introduction

Our understanding of the genetic basis of complex disease has recently been transformed by genome-wide association studies, leading to the identification and cataloging of hundreds of genomic loci associated with human disease (Hindorff *et al*, 2009). In parallel, current technologies have brought systematic functional investigation of the underlying disease pathways within reach. Building upon previous work to integrate genetic and transcriptional profiles to uncover disease genes (Hubner *et al*, 2005; Mehrabian *et al*, 2005), Chen *et al* (2008) and Emilsson *et al* (2008) constitute two recent large-scale studies, which led to the identification of novel candidate genes for obesity and the characterization of the macrophage-enriched metabolic network module, a subnetwork enriched for genes in inflammatory processes and metabolic syndrome. These studies added to the growing body of evidence linking inflammation and the immune

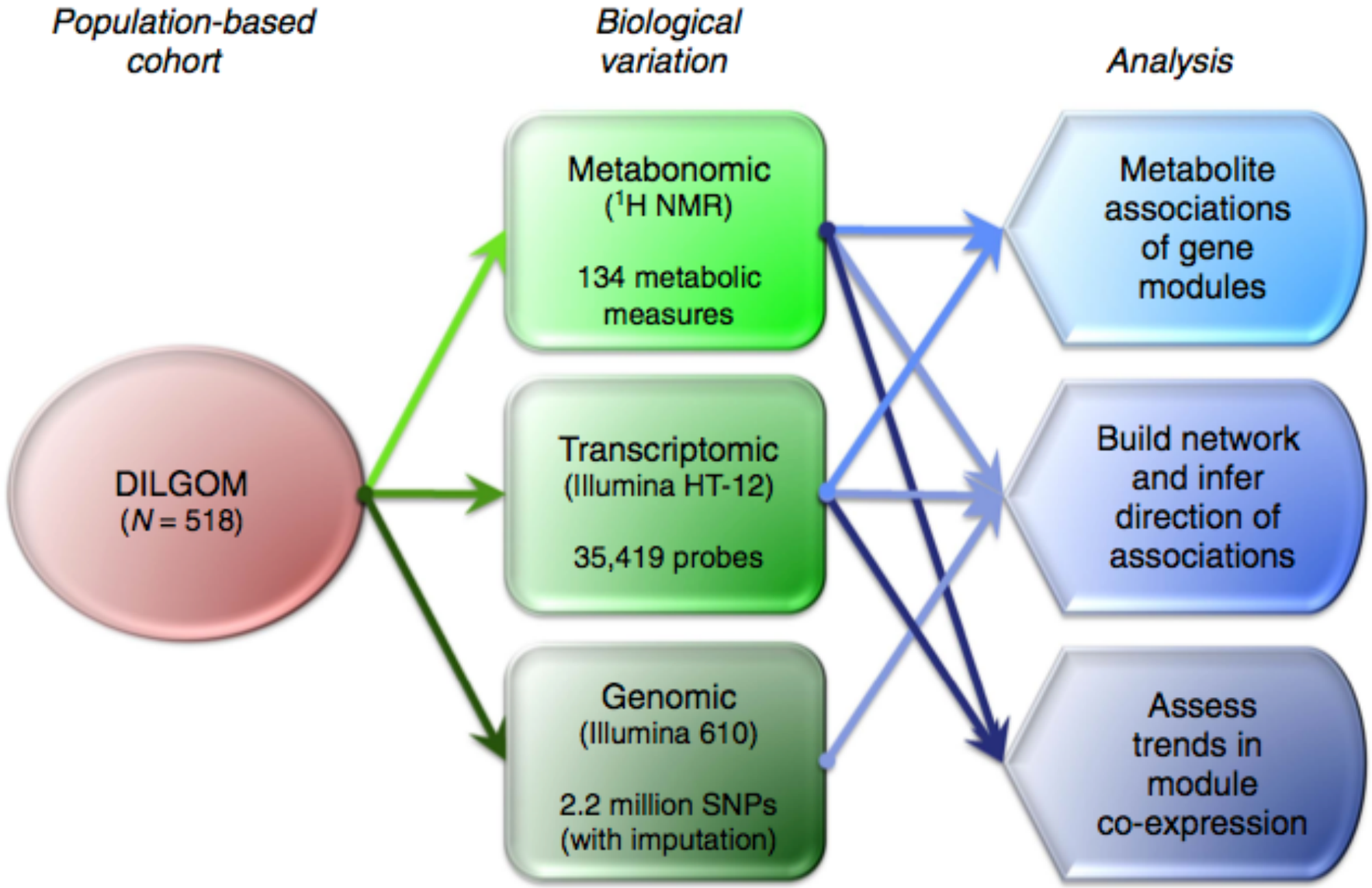


Figure 1 Overview of data integration and analyses. A high-level view of the study design and analysis employed.

(Supplementary Figure 6) (Horvath and Dong, 2008; Langfelder and Horvath, 2008). A Spearman's rank correlation of the LL module expression profile (first principal component) with metabolite distributions offered fine-scale detail of potentially causative/reactive effects (Table I, see Supplementary Figures 7–16). After correction for the 21 modules tested (a Bonferroni-corrected significance level of 2.4×10^{-3}), expression of LL module was, as expected from univariate analysis, positively associated with glycoprotein (P=1.83×10^{-8}) and creatine levels (P=9.02×10^{-4}), and negatively associated with isoleucine (P=4.03×10^{-16}). Our previous findings showed association of LL module expression with total HDL and APOB but not with total low-density lipoprotein (LDL) (Inouye *et al*, 2010). The deeper lipoprotein phenotypes obtained by serum NMR metabonomics allowed for division of the APOB–lipoprotein cascade into 10 subclasses: six for the very-low-density lipoprotein (VLDL) fraction, the intermediate-density lipoprotein (IDL), and three LDL subclasses. All VLDL subclasses had a strong correlation with LL module expression, indicating dominance of the VLDL fraction in the APOB association (Table I), as could be expected (Lusis *et al*, 2008). The current results further validated that, at current detection power, there remained little evidence for association with total LDL-C (P=0.07), while a few measures related to the medium and small LDL subclasses did associate weakly with LL module expression (Supplementary Figure 14).

Of particular interest were the opposite metabolic correlations observed for HDL subclasses both in terms of LL module association and overall (Supplementary Figure 3). For example, the smallest HDL subclass behaved similarly to the VLDL subclasses and had a negative correlation with the larger HDL subclasses. This finding has also been observed by others (Chasman *et al*, 2009) and suggests that the serum HDL fraction does not have a coherent physiological function and

Table I Significant LL module associations with serum metabolite concentrations and immune response markers

	Spearman's correlation	*P*-value
Metabolite		
Concentration of chylomicrons and extremely large VLDL particles	−0.464	4.46E−29
Triglycerides in medium VLDL	−0.442	3.46E−26
Concentration of large VLDL particles	−0.441	4.51E−26
Free cholesterol in large VLDL	−0.441	4.89E−26
Triglycerides in large VLDL	−0.441	5.20E−26
Triglycerides in VLDL	−0.439	7.25E−26
Concentration of very large VLDL particles	−0.439	8.88E−26
Total cholesterol in large VLDL	−0.438	1.00E−25
Triglycerides in very large VLDL	−0.438	1.09E−25
Phospholipids in large VLDL	−0.437	1.43E−25
Total lipids in large VLDL	−0.436	1.98E−25
Triglycerides in VLDL (Lipido)	−0.434	3.77E−25
Concentration of medium VLDL particles	−0.43	8.80E−25
Cholesterol esters in large VLDL	−0.43	9.68E−25
Total lipids in very large VLDL	−0.429	1.32E−24
Total lipids in medium VLDL	−0.428	1.59E−24
Serum total triglycerides	−0.427	2.53E−24
Phospholipids in medium VLDL	−0.423	6.17E−24
Triglycerides in small VLDL	−0.421	1.27E−23
Total triglycerides	−0.418	2.87E−23
Mobile lipids $-CH_2-$	−0.417	3.25E−23
Free cholesterol in medium VLDL	−0.417	3.26E−23
Phospholipids in very large VLDL	−0.414	6.36E−23
Ratio of triglycerides to phosphoglycerides	−0.41	2.14E−22
Total cholesterol in medium VLDL	−0.405	7.78E−22
Concentration of small VLDL particles	−0.394	1.18E−20
Cholesterol esters in medium VLDL	−0.383	1.36E−19
Phospholipids in chylomicrons and extremely large VLDL	−0.381	2.56E−19
Total lipids in chylomicrons and extremely large VLDL	−0.376	7.91E−19
Triglycerides in small HDL	−0.375	1.00E−18
Total lipids in small VLDL	−0.374	1.18E−18
Triglycerides in chylomicrons and extremely large VLDL	−0.372	2.03E−18
Phospholipids in small VLDL	−0.362	1.70E−17
Mobile lipids $-CH_3$	−0.35	2.11E−16
Isoleucine	−0.347	4.03E−16
Unsaturated lipids	−0.343	9.41E−16
Free cholesterol in small VLDL	−0.341	1.43E−15
Triglycerides in very small VLDL	−0.337	2.97E−15
ω-9 and saturated fatty acids	−0.329	1.61E−14
Total fatty acids	−0.316	1.66E−13
Total cholesterol in IDL (Lipido)	−0.307	9.13E−13
Apolipoprotein B by apolipoprotein A-I	−0.299	3.90E−12
Total cholesterol in small VLDL	−0.278	1.18E−10
Free cholesterol in large HDL	0.273	2.59E−10
ω-6 and -7 fatty acids	−0.269	5.02E−10
Apolipoprotein B	−0.263	1.17E−09
Average number of methylene groups per a double bond	−0.263	1.24E−09
Ratio of bisallylic groups to total fatty acids	0.263	1.29E−09
Ratio of bisallylic groups to double bonds	0.255	4.16E−09
Total cholesterol in large HDL	0.253	5.15E−09
Total lipids in large HDL	0.245	1.57E−08
Phospholipids in large HDL	0.244	1.77E−08
Glycoproteins	−0.244	1.83E−08
Cholesterol esters in large HDL	0.243	2.26E−08
Concentration of large HDL particles	0.241	2.87E−08
Average number of double bonds in a fatty acid chain	0.24	3.39E−08
Leucine	−0.236	5.58E−08
Total cholesterol in HDL2 (Lipido)	0.235	5.94E−08
Ratio of ω-9 and saturated fatty acids to total fatty acids	−0.234	6.82E−08
Phospholipids in very large HDL	0.221	4.00E−07
Total cholesterol in HDL3 (Lipido)	−0.211	1.30E−06
Concentration of very large HDL particles	0.205	2.56E−06
Triglycerides in IDL	−0.199	5.24E−06
Total lipids in very large HDL	0.198	5.49E−06
Ratio of ω-6 and -7 fatty acids to total fatty acids	0.194	8.94E−06
Total cholesterol in HDL	0.191	1.17E−05
Free cholesterol in very large HDL	0.186	1.96E−05
Concentration of very small VLDL particles	−0.182	3.23E−05
Cholesterol esters in very large HDL	0.177	5.08E−05
3-Hydroxybutyrate	0.171	9.48E−05
Concentration of small LDL particles	−0.17	1.02E−04
Total cholesterol in very large HDL	0.168	1.25E−04
18:2, Linoleic acid	−0.154	4.48E−04
Concentration of small HDL particles	−0.154	4.48E−04
Total phosphoglycerides	−0.151	5.78E−04
Total lipids in small LDL	−0.147	7.87E−04
Other polyunsaturated fatty acids than 18:2	−0.146	8.30E−04
Creatine	0.145	9.02E−04
Total lipids in very small VLDL	−0.145	9.13E−04
Description of average fatty acid chain length (not carbon number)	0.141	1.26E−03
Phospholipids in medium LDL	−0.141	1.31E−03
Triglycerides in very large HDL	−0.141	1.33E−03
Immune response marker		
Interleukin-1 receptor antagonist	−0.203	3.14E−06
High-molecular-weight adiponectin	0.189	1.55E−05
C-reactive protein	−0.16	2.62E−04

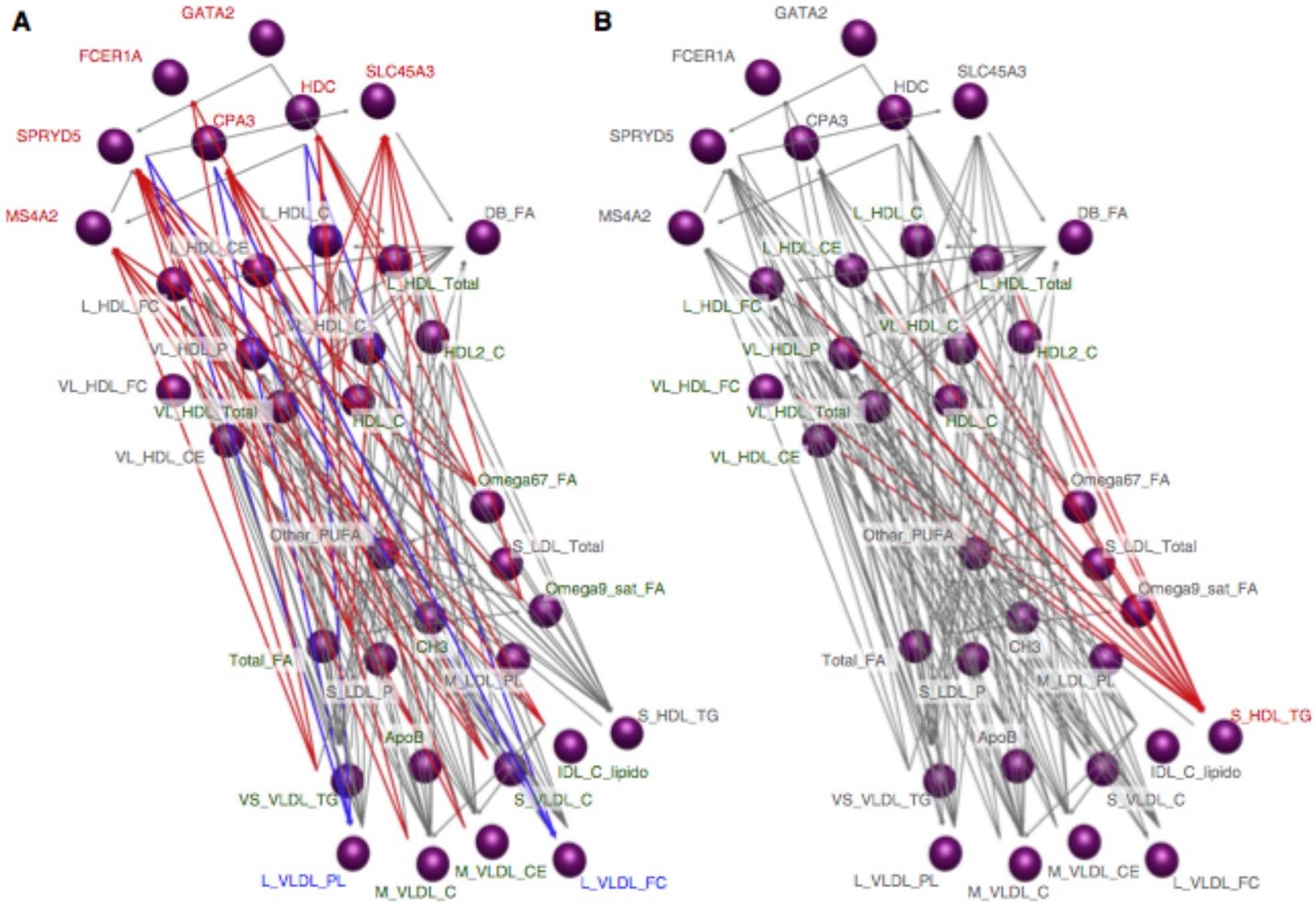


Figure 2 Edge-directed network of LL module and serum metabolites. Genetic variation was used to infer a causal network of core LL module expression and metabolites (see Supplementary Methods and Supplementary Table 1 for abbreviations). Core LL module genes are represented by purple nodes in the upper left, whereas metabolites significantly associated with genetic variation and LL module are denoted by all other nodes. Arrows denote directed edges. (**A**) Shows core LL module (red) reactivity to most lipoprotein subclass levels (green); however, two components of large VLDL (blue) were predicted downstream of *CPA3*, *SPRYD5*, and *HDC* expression. (**B**) Shows that triglycerides in the small HDL subclass (red) are predicted to be largely reactive to concentrations of the larger HDL subclasses (green). Source data is available for this figure at www.nature.com/msb.