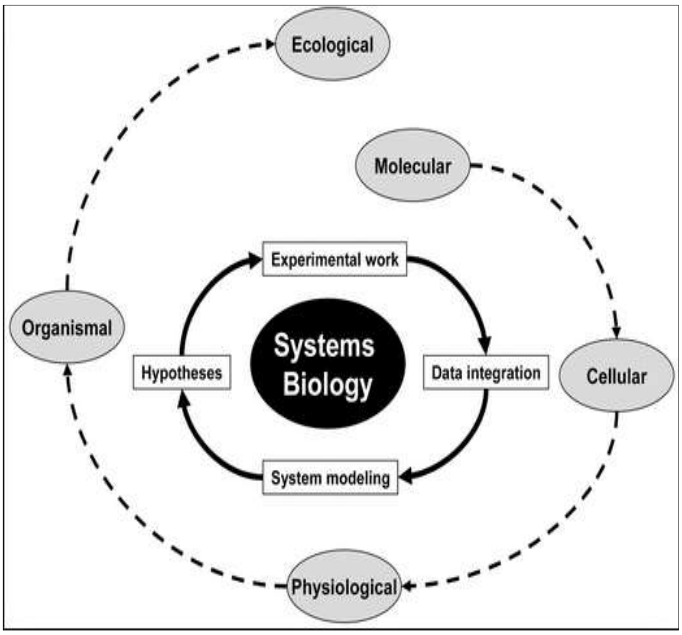
UNIT - I

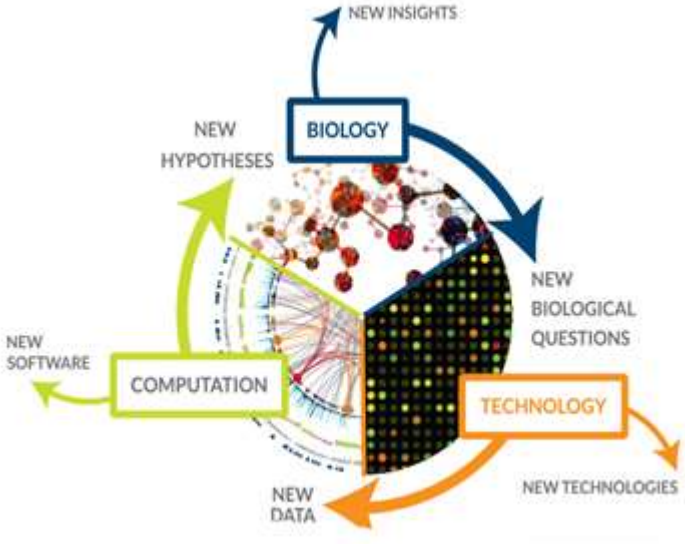
Introduction to Systems Biology-SBI1401

UNIT-I

GENERAL INTRODUCTION

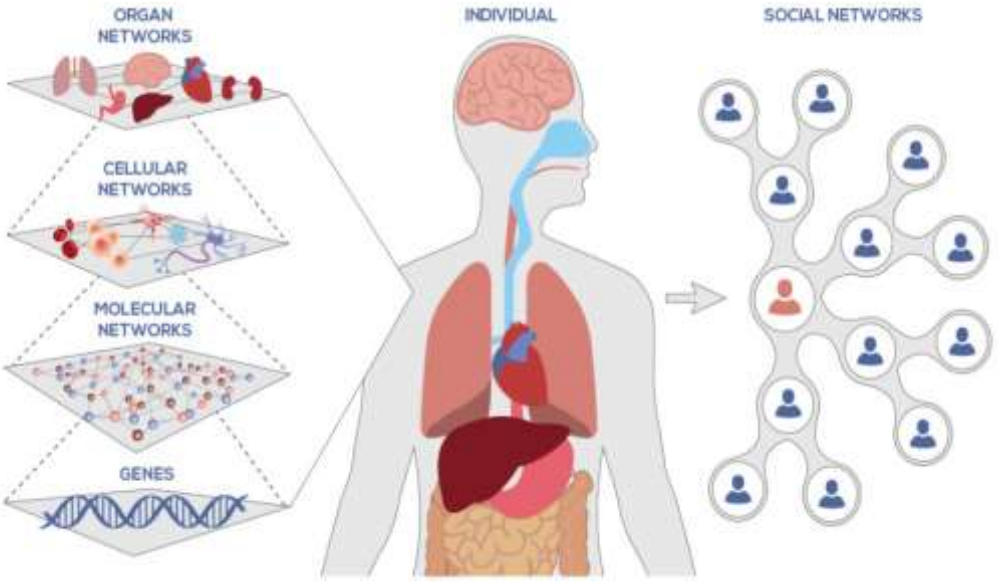
Systems biology is based on the understanding that the whole is greater than the sum of the parts.





- One of the tenets of systems biology we often refer to is the "Network of Networks."
- On a biological level, our bodies are made up of many networks that are integrated at and communicating on multiple scales.
- From our genome to the molecules and cells that makeup the organs in our bodies all the way out to ourselves in our world: we are fundamentally a network of networks.

- Systems biology looks at these networks across scales to integrate behaviors at different levels, to formulate hypotheses for biological function and to provide spatial and temporal insights into dynamical biological changes. It is not enough to understand only one part of a system when studying the complexity of biology.
- Therefore the framework of the "Network of Networks" provides meaningful insight into understanding how systems biology's approach is different, more integrated and more capable of analyzing and predicting state transitions in biological systems.



- Systems biology has been responsible for some of the most important developments in the science of **human health and environmental sustainability**.
- It is a **holistic approach** to deciphering the complexity of biological systems that starts from the understanding that the networks that form the whole of living organisms are more than the sum of their parts.
- It is **collaborative**, integrating many scientific disciplines – biology, computer science, engineering, bioinformatics, physics and others – to **predict** how these systems change over time and under varying conditions, and to develop solutions to the world's most pressing health and environmental issues.
- This ability to design predictive, **multiscale** models enables our scientists to discover new biomarkers for disease, **stratify** patients based on unique genetic profiles, and **target** drugs and other treatments.
- Systems biology, ultimately, creates the potential for entirely new kinds of exploration, and drives constant **innovation** in biology-based technology and computation.

Modeling in biology-Properties of models

What is a model?

Model is "a simplified or idealized description, representation or conception of a particular system, situation, or process, often in mathematical terms , that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions, etc."

Models should be as simple as possible, yet as complex as necessary to address a

given question of interest.

- ❑ "All models are wrong, but some of them are useful", George Box.
- ❑ "Everything should be made as simple as possible, but no simpler", Albert Einstein.
- ❑ "Entia non sunt multiplicanda praeter necessitatem" (entities must not be multiplied beyond necessity), William of Ockham.

In other words, a mathematical model is a representation of the essential aspects of an existing system (or a system to be constructed) which presents knowledge of that system in usable form.

Thus, models are not replicas of reality, they are simplified representations of it.

Simplification allows us to comprehend the essential features of a complex process without being burdened and overwhelmed by unnecessary details.

The modelling process is **considered successful** when the obtained model possesses the following characteristics:

1. Accurate: the model should attempt to accurately describe current existing observations.
2. Predictive: the model should allow to appropriately predict the behaviour of the system (through analysis or simulation) in situations not already observed.
3. Reusable: the model can be reused in another, similar case.
4. Parsimonious: the model should be as simple as possible. That is, given competing and equally good models, the simplest is preferred

Essential features of a modelling approach

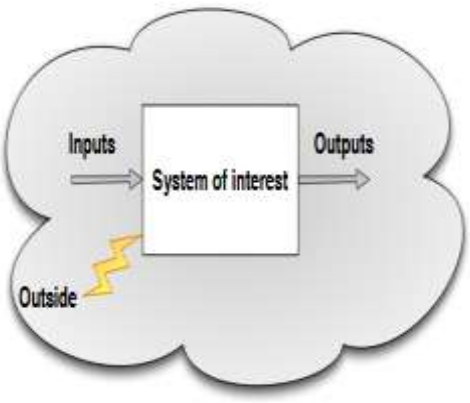
Isolate your system of interest.

•Identify what is important (and therefore what needs to be included in your model).

•List the quantities that can be observed/measured (they are the outputs).

•List the quantities that can be controlled/acted upon (they are the inputs).

•Define the environment and the constraints it places upon the system



Modelling of the system of interest

Typically, the model is composed of

- variables
  - independent, e.g., time t
    - 1 indep. var.: ODEs, e.g., time t
    - more than 1 indep. var.: PDEs, e.g., time t and space (x, y, z) (examples include: blood circulation, diffusion, growth)
  - dependent (on the independent variable(s)), e.g., concentrations functions of time $\{[E](t), [S](t), [P](t)\}$
- parameters
  - not dependent on independent variables
  - can be varied/changed under experimental conditions (this can lead to a qualitative change in the system behaviour)
- constants
  - fixed, e.g., Avogadro constant, gravitational constant

Based on these concepts, different types of models can be built.

Types of models

Continuous	Discrete
• the independent variables are continuous • ODEs, PDEs	• the independent variables are discrete • Difference equations
Deterministic	**Stochastic**
• var., param. and const. do not contain randomness • they are defined by a unique function	• dynamics contain an element of randomness (described by probabilities, e.g., the variables are random/stochastic processes) • e.g., SDEs
Linear	**Nonlinear**
• $\dot{x} - \frac{dx}{dt} - -kx$ • Linear ODE	• $\dot{x} - \frac{dx}{dt} - -kx + x^3$ • Nonlinear ODE
Autonomous	**Non-autonomous**
• Without control input: $\dot{x} - -kx$	• With control input: $\dot{x} - -kx + u$
Constructive	**Data-driven**
• mechanistic or deductive • also called "equation-based" or "(first) principle-based"	• phenomenological or inductive

Discrete models are typically used to model discrete events/discontinuous changes,

e.g., events/changes which occur at specific time instants (i.e., between two consecutive events nothing changes/happens). They can also be obtained, as we will see, as the result of the discretisation of continuous models.

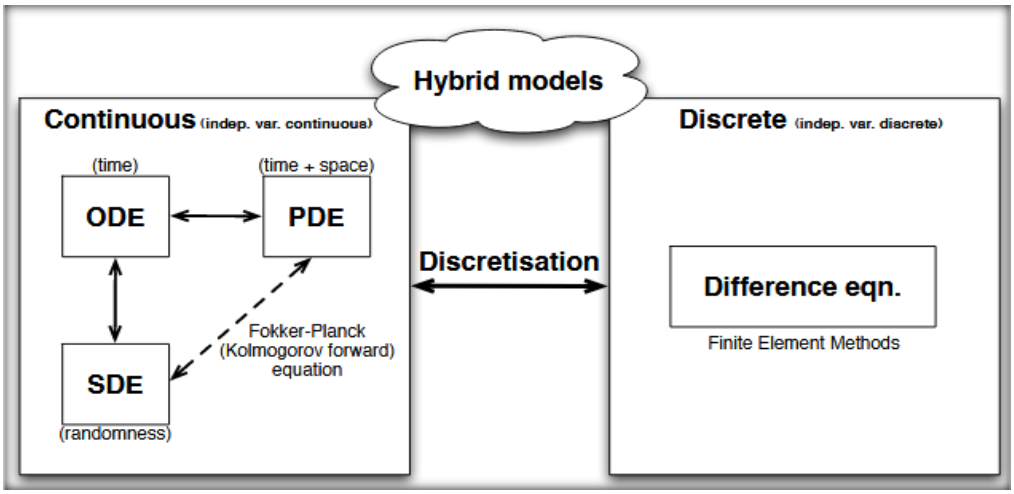
Stochastic models (e.g., SDEs) are typically used to model diverse phenomena such as fluctuating stock prices, physical systems subject to thermal fluctuations, or intrinsic noise/stochastic effects in cellular biology.

Continuous	ODEs PDEs	Deterministic (L or NL)
	SDEs	Stochastic (L or NL)
Discrete	Difference equations	Deterministic (L or NL)
		Stochastic (L or NL)

Nonlinear, stochastic models are almost unavoidable in biological modelling.

We will mostly deal with autonomous, deterministic models obtained through a constructive approach. But we will also briefly introduce stochastic models.

Linear deterministic models can be solved analytically. This is typically not the case for nonlinear or stochastic models, which, therefore, are often analysed using bifurcation and phase plane analysis tools (which we will cover in this course) and also through computer simulations, e.g., MATLAB



Transcription networks-Basic concepts

The cell is an integrated device made of several thousands of interacting proteins.Each protein is a nanometer size molecular machine that carries out a specific task with exquisite precision.

Eg:Micron long E.coli is a cell that contains few million proteins of about 4000 types (Refer fig in next slide)

Cells encounter different situation that require different protein .For example when sugar is sensed ,the cell begins to produce proteins that can transport the sugar into the cell and utilize it.

When damaged ,the cell produces repair proteins.

The cell therefore continuously monitors its environment and calculates the amount at which each type of protein is needed.This information processing function which determines the rate of production of each protein is largely carried out by transcription networks.

Typical Parameter Values for the Bacterial *E. coli* Cell, the Single-Celled Eukaryote *Saccharomyces cerevisae* (Yeast), and a Mammalian Cell (Human Fibroblast)

Property	*E. coli*	Yeast (*S. cerevisae*)	Mammalian (Human Fibroblast)
Cell volume	~1 μm^3	~1000 μm^3	~10,000 μm^3
Proteins/cell	~4 10^6	~4 10^7	~4 10^{10}
Mean size of protein	5 nm		
Size of genome	4.6 10^6 bp 4500 genes	1.3 10^7 bp 6600 genes	3 10^9 bp ~30,000 genes
Size of: Regulator binding site	~10 bp	~10 bp	~10 bp
Promoter	~100 bp	~1000 bp	~10^4 to 10^5 bp
Gene	~1000 bp	~1000 bp	~10^4 to 10^6 bp (with introns)
Concentration of one protein/cell	~1 nM	~1 pM	~0.1 pM
Diffusion time of protein across cell	~0.1 sec D = 10 μm^2/sec	~10 sec	~100 sec
Diffusion time of small molecule across cell	~1 msec, D = 1000 μm^2/sec	~10 msec	~0.1 sec
Time to transcribe a gene	~1 min 80 bp/sec	~1 min	~30 min (including mRNA processing)
Time to translate a protein	~2 min 40 aa/sec	~2 min	~30 min (including mRNA nuclear export)
Typical mRNA lifetime	2–5 min	~10 min to over 1 h	~10 min to over 10 h
Cell generation time	~30 min (rich medium) to several hours	~2 h (rich medium) to several hours	20 h — nondividing
Ribosomes/cell	~10^4	~10^7	~10^8
Transitions between protein states (active/inactive)	1–100 μsec	1–100 μsec	1–100 μsec
Timescale for equilibrium binding of small molecule to protein (diffusion limited)	~1 msec (1 μM affinity)	~1 sec (1 nM affinity)	~1 sec (1 nM affinity)
Timescale of transcription factor binding to DNA site	~1 sec		
Mutation rate	~10^{-9} /bp/generation	~10^{-10} /bp/generation	~10^{-9}/bp/year

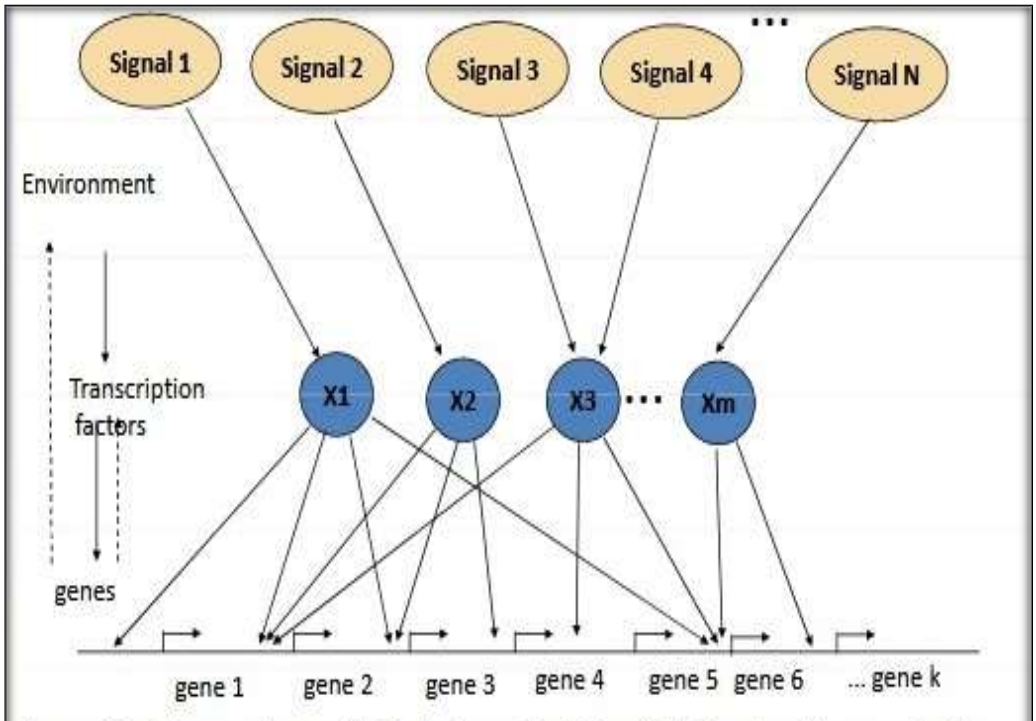
bp: base-pair (DNA letter).

Cognitive problem of the cell

- Cells live in a complex environment and can sense many different signals,including physical parameter such as temperature and osmotic pressure,biological signaling molecules from other cells,beneficial nutrients, and harmful chemicals.
- Information about the internal state of the cell,such as the level of key metabolites and intenal damage (eg;damage to DNA,membrane or proteins),is also important.
- Cells respond to these signals by producing appropriate proteins that act upon the internal or external environment.

To represent these environmental states,the cells uses special proteins called transcription factors as symbols.

- ✓ Transcription factors are usually designed to transit rapidly between active and inactive molecular states,at a rate that is modulated by a specific environmental signal(input).
- ✓ Each active transcription factor can bind the DNA to regulate the rate at which specific target genes are read.



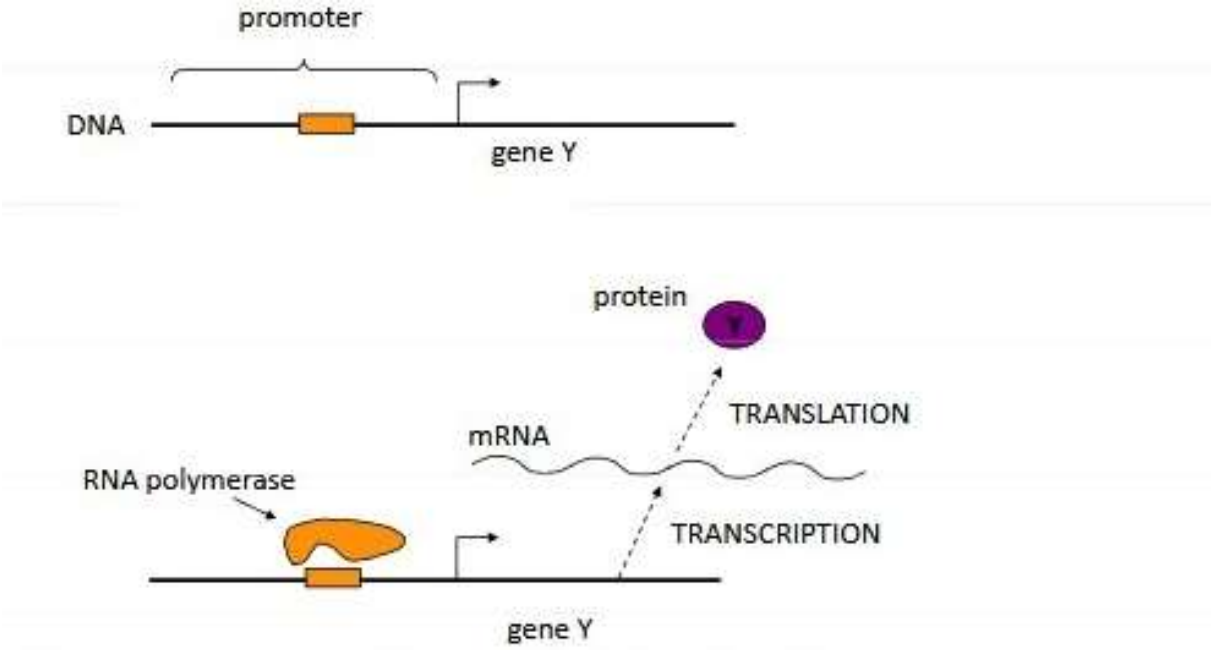
The mapping between environmental signals, transcription factors inside the cell and the genes that they regulate. The environmental signals activate specific transcription factor proteins. The transcription factors, when active, bind DNA to change the transcription rate of specific target genes, the rate at which mRNA is produced. The mRNA is then translated into protein. Hence, transcription factors regulate the rate at which the proteins encoded by the genes are produced. These proteins affect the environment (internal and external). Some proteins are themselves transcription factors, that can activate or repress other genes, etc.

- ✓ The genes are read (transcribed) into m RNA, which is then translated into proteins,which can **act on the environment.**The activities of the transcription factor in a cell therefore can be considered an **internal representation of the environment.**
- ✓ For example bacterium. **Ecoli has a internal represenation with about 300 degrees of freedom(transcription factors).**
- ✓ These regulate the rates of production of E.coli's 4000 proteins.
- ✓ Internal representation by a **set of transcripton factors is a very compact description of the myriad factors in the environment.**
- ✓ **It** seems that **evolution selected internal representations that symbolizes states that are most important for cell survival and growth**.
- ✓ Many different **are summarized by a particular transcription factor activity that signifies " I am starving ".Many other situations are summarized by a different tsituations ranscription factors activity that signifies " my DNA is damaged"**.These transcription factors regulate their target genes to mobilize the appropriate protein responses in each case.

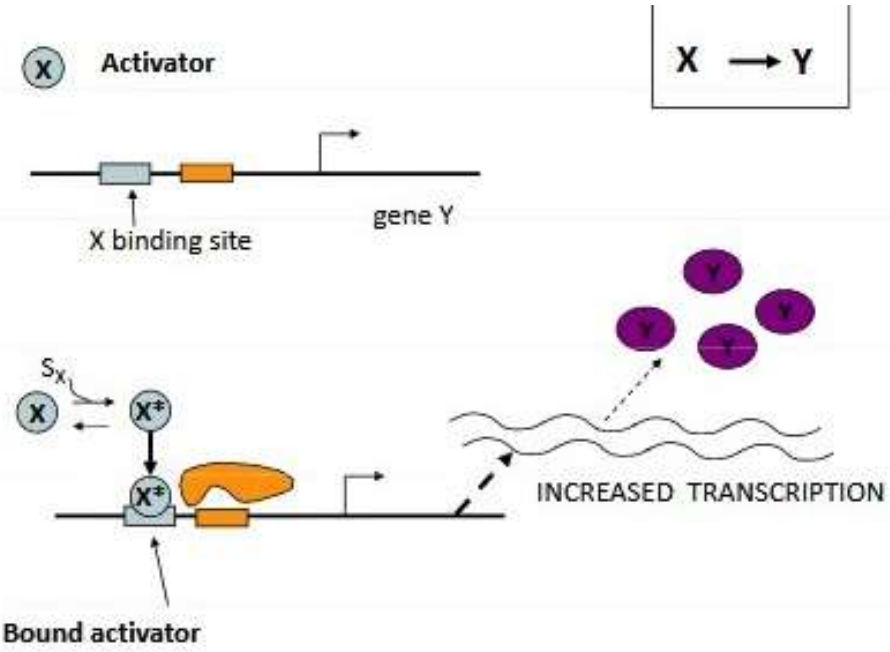
Elements of transcription networks

- ✓ Elements of transcription networks **are genes and transcription factors** (TFs). The interaction between TFs and genes is described by these networks .
- ✓ **Gene**:Stretch of DNA whose sequence encodes the information needed for production of a protein.
- ✓ Transcription of a gene is the process by which RNA polymerase (RNAp) produces m RNA that corresponds to that genes coding sequence.The **m RNA is then translated into a protein also called the gene product.**The rate at which the gene is transcribed ,the number of m RNA product per unit time is controlled by the **promoter**, a regulatory region of DNA that precedes the gene .
- ✓ RNAp binds a defined site (a specific DNA sequence) at the promoter .**The quality of this site specifies** the transcription rate of the gene.
- ✓ Whereas **RNAp acts on virtually all of the genes,changes in the expression of specific genes are due to transcription factors.**
- ✓ Each transcription factor modulates the transcription rate of a set of target genes.
- ✓ Transcription factors affect the transcription rate by binding specific sites in the promoters of the regulated gens.When bound ,they change the probability per unit time that RNAp binds promoter and produces an Mrna molecule.The transcription factors thus affect the rate at which RNAp initiates transcription of the gene.
- ✓ Transcription factors can act as activators that increase the transcription rate of a gene,or as repressors that reduce the trancription rate.
- ✓ The rate at which the gene is transcribed is controlled by the promoter, a regulatory region of the gene that precedes the gene.

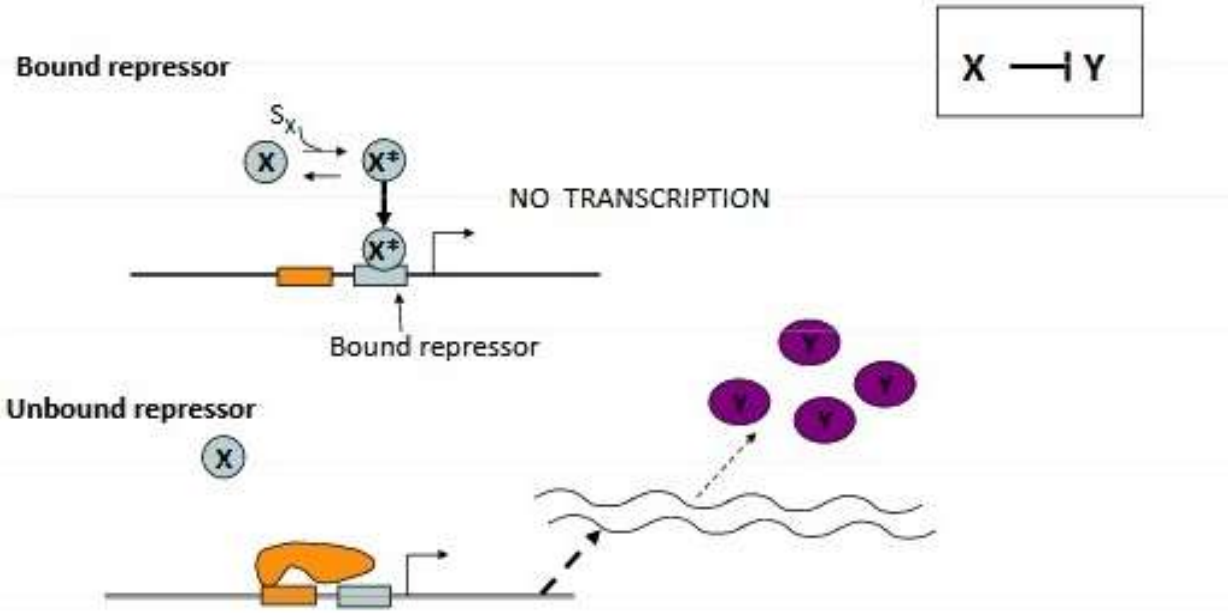
- ✓ When TFs are bound to the promoter region, they change the probability per unit time that RNAp binds the promoter and produces an mRNA molecule.
- ✓ TFs can act as activators that increase the transcription rate of the gene, or as repressors that reduce the transcription rate.



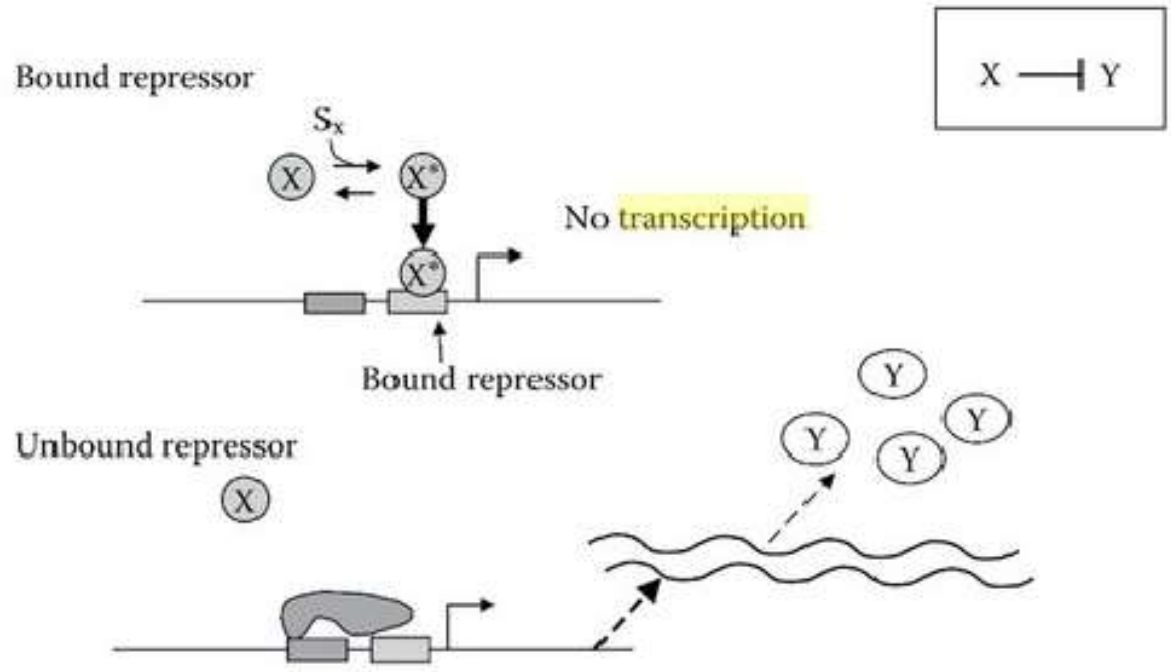
Each gene is usually preceded by a regulatory DNA region called the promoter. The promoter contains a specific site (DNA sequence) that can bind RNA polymerase (RNAp), a complex of several proteins that forms an enzyme That can synthesize mRNA that is complementary to the genes coding sequence. The process of forming the mRNA is called transcription. The mRNA is then translated into protein.



Activator X, is a transcription- factor protein that increases the rate of mRNA transcription when it binds the promoter. The activator transits rapidly between active and inactive forms. In its active form, it has a high affinity to a specific site (or sites) on the promoter. The signal S_X increases the probability that X is in its active form X*. Thus, X* binds the promoter of gene Y to increase transcription and production of protein Y.

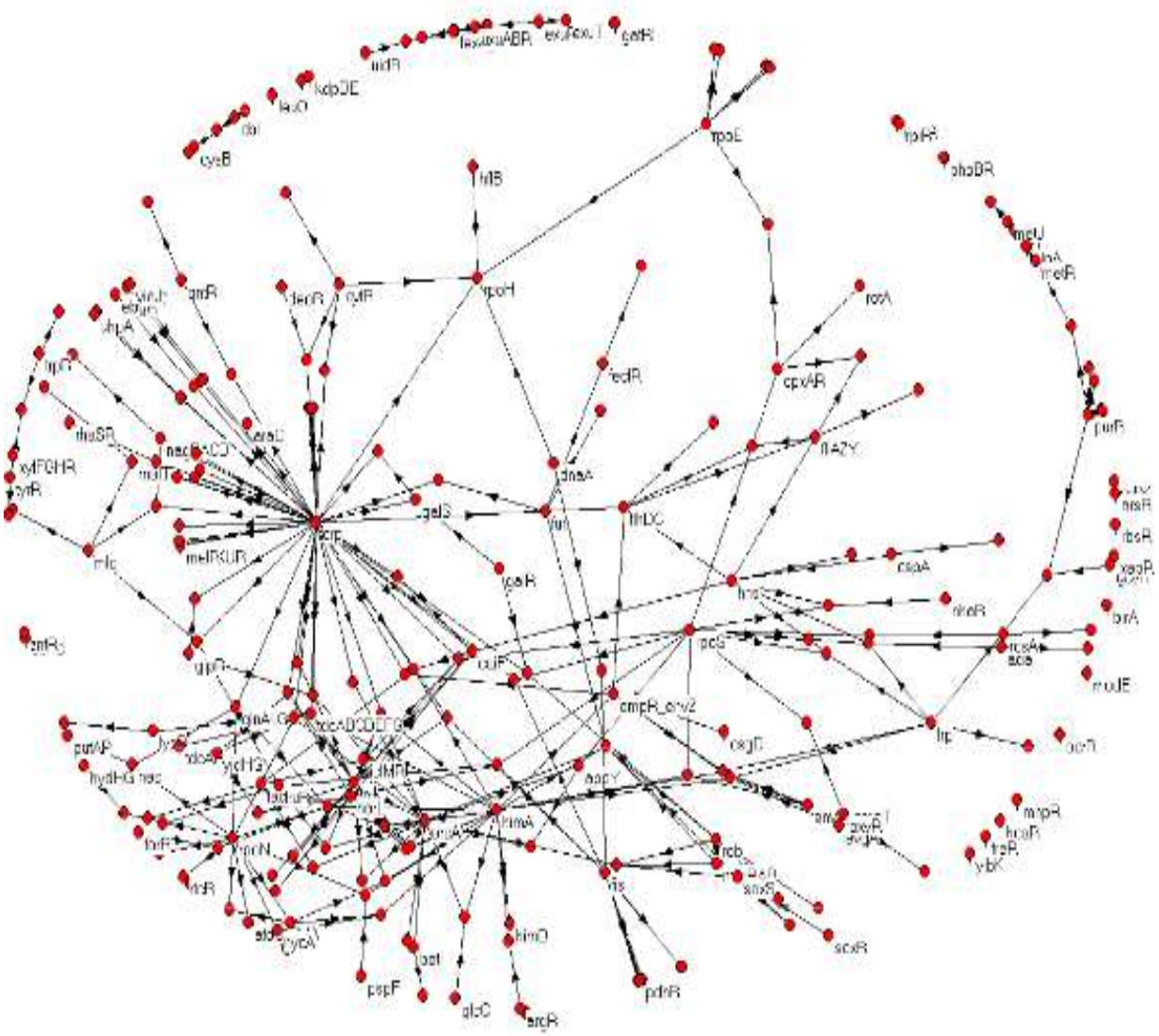


A repressor X, is a transcription- factor protein that decreases mRNA transcription when it binds the promoter. The signal S_X increases the probability that X is in its active form X*. X* binds a specific site in the promoter of gene Y to decrease transcription and production of protein Y. Many genes show a weak (basal) transcription when repressor is bound.



- Transcription factors themselves encoded by genes,which are regulated by other transcription factors,which in turn may be regulated by yet other transcription factors and so on.This set of interactions forms a **trancription network.**
- Trancription network **describes all of the regulatory transcription interactions in a cell.**

- In a network ,the **nodes** and **edges** represents transcription regulation of one gene by the protein product of another gene.
- A directed **edge X->Y** means that the product of **gene X is a transcription factor protein that binds the promoter of gene Y** to control the rate at which gene Y is transcribed.
- The **inputs to the network are signals that carry information from the environmen**t.Each **signal is a small molecule,protein modification or molecular partner that directly affects the activity of one of the transcription factors**.
- Often,external stimuli activate **bio chemical signal transduction pathways that culminate in a chemical modification of specific transcription factors.**



- In other systems the signal can be a **simple as as sugar moelcule** that enters **the cells and directly binds the transcription factors**.The signals usually **cause a physical change** in the **shape of the transcription factor protein**,causing it to assume an **active molecular state**.Thus signal **Sx** can cause **X to rapidly shift to its active state X^*** ,bind the **promoter of gene Y** and **increase the rate of transcription leading to increased production of protein Y.**

- The network thus represents **a dynamic system: after** an input signal arrives, **transcription factor activities change,leading to change in the production rate of proteins**.
- Some of **the proteins are transcription factors that activate additional genes and so on**.The rest of the **proteins are not transcription factors,but rather carry out the diverse functions of the living cells,such as buiding structures and catalyzing reactions.**

- Strong **separation of time scales.**
- **Modularity** of transcription networks.
- **Step-function** approximations of Hill rate functions.
- **Network motifs:** basic building blocks of biological networks/circuits.
- **Robustness:** biological circuits have robust designs such that their essential function is nearly independent of biochemical parameters.
- **Kinetic proofreading:** How can a biochemical recognition system pick out a specific molecule in a sea of similar molecules?
- **Optimal gene circuit design**: Are bio-circuits designed in some optimal way for a given environment?
- Similarity between **men made/engineered and evolved biological systems.** Are there some deeper explanations for these similarities?

Separation of timescale

- Transcription networks are designed with a strong seperation of timescales.
- The input signals usually change transcription factor activities on a sub second timescale.
- Binding of the active transcription factor to its DNA sites often reaches equilibrium in seconds.Transcription and translation of the target gene take minutes,and the accumulation of the protein product can take many minutes to hours.Thus different steps between the signal and the accumulation of the protein products have very different time scales.

Binding of a small molecule (a signal) to a transcription factor, causing a change in transcription factor activity	~1 msec
Binding of active transcription factor to its DNA site	~1 sec
Transcription + translation of the gene	~5 min
Timescale for 50% change in concentration of the translated protein (stable proteins)	~1 h (one cell generation)

- Transcription networks are designed with a strong separation of time scales: the input signals usually activate TFs on a sub-second time scale.
- Binding of an active TF to its DNA reaches equilibrium in seconds.
- Transcription and translation takes minutes.
- Accumulation of the protein takes many minutes to hours.
- Typical approximate time scales for *E. coli*:
  1. Binding signaling molecule to a TF $\sim$ 1 msec.
  2. Binding active TF to its DNA site $\sim$ 1 sec.
  3. Transcription + translation of the gene $\sim$ 5 min.
  4. 50% change of protein concentration $\sim$ 1 h.
- Hence, 1 and 2 can be considered instantaneous when studying transcription networks.

The Human Interactome

- The **human interactome** is the set of protein–protein interactions (the interactome) that occur in human cells.
- The physiology of a cell can be viewed as the product of thousands of proteins acting in concert to shape the cellular response.
- Coordination is achieved in part through networks of protein–protein interactions that assemble functionally related proteins into complexes, organelles, and signal transduction pathways.
- Understanding the architecture of the human proteome has the potential to inform cellular, structural, and evolutionary mechanisms and is critical to elucidating how genome variation contributes to disease
- BioPlex 2.0 (Biophysical Interactions of ORFeome-derived complexes), which uses robust affinity purification–mass spectrometry methodology to elucidate protein interaction networks and co-complexes nucleated by more than 25% of protein-coding genes from the human genome, and constitutes, to our knowledge, the largest such network so far
- With more than 56,000 candidate interactions, BioPlex 2.0 contains more than 29,000 previously unknown co-associations and provides functional insights into hundreds of poorly characterized proteins while enhancing network-based analyses of domain associations, subcellular localization, and co-complex formation.

Large-scale analysis of disease pathways in the human interactome

- Discovering disease pathways, which can be defined as sets of proteins associated with a given disease, is an important problem that has the potential to provide clinically actionable insights for disease diagnosis, prognosis, and treatment.
- Computational methods aid the discovery by relying on protein-protein interaction (PPI) networks.

- They start with a few known disease-associated proteins and aim to find the rest of the pathway by exploring the PPI network around the known disease proteins

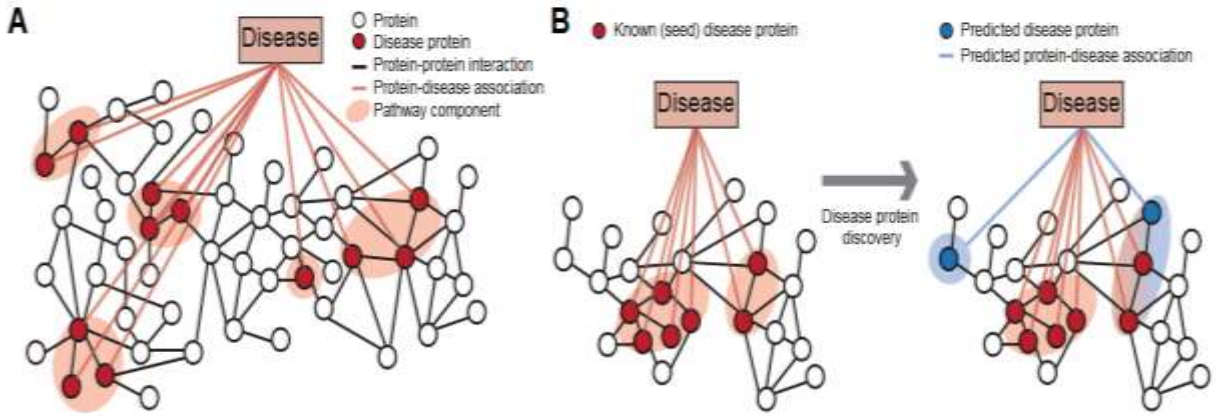


Fig. 1. **Network-based discovery of disease proteins. A** Proteins associated with a disease are projected onto the protein-protein interaction (PPI) network. In this work, *disease pathway* denotes a (undirected) subgraph of the PPI network defined by the set of disease-associated proteins. The highlighted disease pathway consists of five pathway components. **B** Methods for disease protein discovery predict candidate disease proteins using the PPI network and known proteins associated with a specific disease. Predicted disease proteins can be grouped into a disease pathway to study molecular disease mechanisms.

Inborn errors of metabolism and the human interactome: a systems medicine approach

1. Curation of disease genes associated with IEM
2. Constructing the IEM interactome (IEMi) and the expanded IEM interactome (eIEMi)
3. Calculation of *z*-score
4. Enrichment analysis
5. Comparison to non-IEM diseases
6. Drug target information

UNIT - II

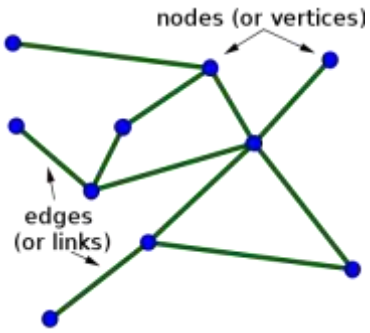
Introduction to Systems Biology-SBI1401

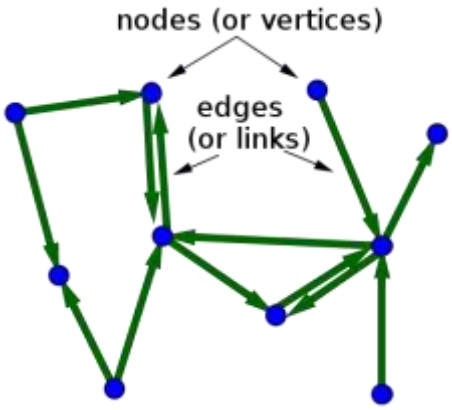
Patterns, Randomized Networks and Network Motifs in Biology

- Basic dynamics of single interaction in a transcription network.
- Actual transcription network with many interaction edges.
- Example:Transcription network of E.Coli included 20% of its total genes.
- Thus any organism transcription network is complex.
- Understandable patterns of connections that serve as building blocks of the network;will be helpful in understanding the entire network based on the dynamics of the individual building blocks.
- Detecting building blocks patterns in complex networks called network motifs.

Graph definition

- The term *graph* can refer to two completely different things. Students usually first learn of a graph as plot of a function, or a function graph. Here, we refer to a different definition of graph, in which a graph is another word for a network, i.e., a set of objects (called vertices or nodes) that are connected together. The connections between the vertices are called edges or links.
- A network is simply a collection of connected objects. We refer to the objects as nodes or vertices, and usually draw them as points. We refer to the connections between the nodes as edges, and usually draw them as lines between points.





An undirected network with 10 nodes (or vertices) and 11 edges (or links).
A directed graph with 10 vertices (or nodes) and 13 edges (or links).

- Networks can be classified as '**directed**' or '**undirected**' based on the nature of the interaction.

Directed Network:

In such networks as shown in fig the interaction between any two nodes has a well defined direction like the direction of signalling from a transcription factor to a gene or the direction of material flow from a substrate to a product in a metabolic reaction

Undirected graphs have edges that do not have a direction. The edges indicate a two-way relationship, in that each edge can be traversed in both directions. This figure shows a simple undirected graph with three nodes and three edges.

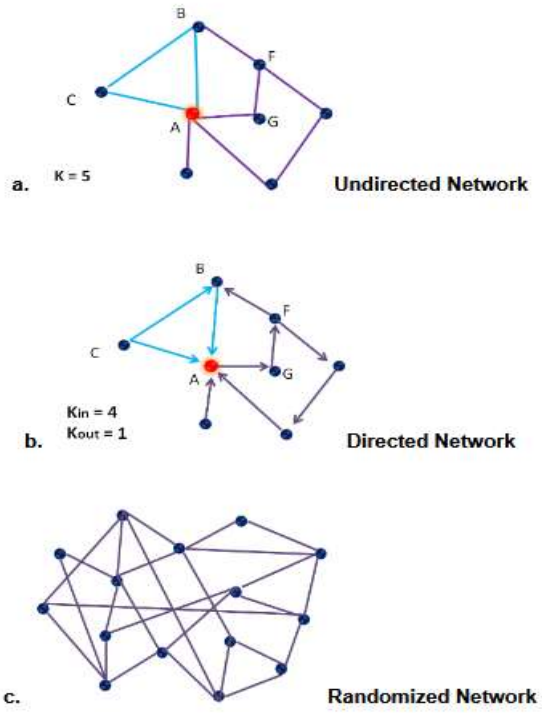
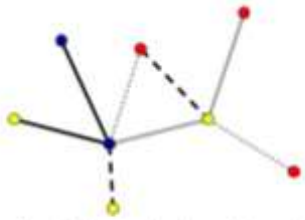


Fig 1: Representation of a biological network

In some networks, not all nodes and edges are created equal. For example, in metabolic networks, nodes may indicate different enzymes which have a wide variety of behaviors, and edges may indicate vastly different types of interactions. To model such difference, one can introduce different types of nodes and edges in the network, as illustrated by the different colors and edge styles, above.

In networks where the differences among nodes and edges can be captured by a single number that, for example, indicates the strength of the interaction, a good model may be a weighted graph. One can represent a weighted graph by different sizes of nodes and edges.



An undirected network where the nodes and edges have different types, as indicated by their colors and line styles.

A directed network where the edges and nodes have different weights, as indicated by their sizes.

Different Levels of System Representation

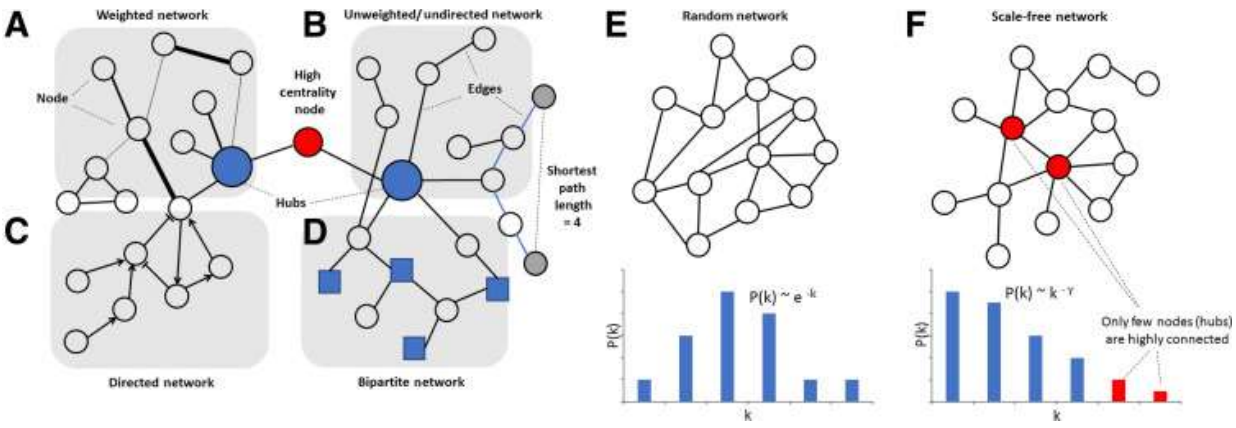
A- gene ontology

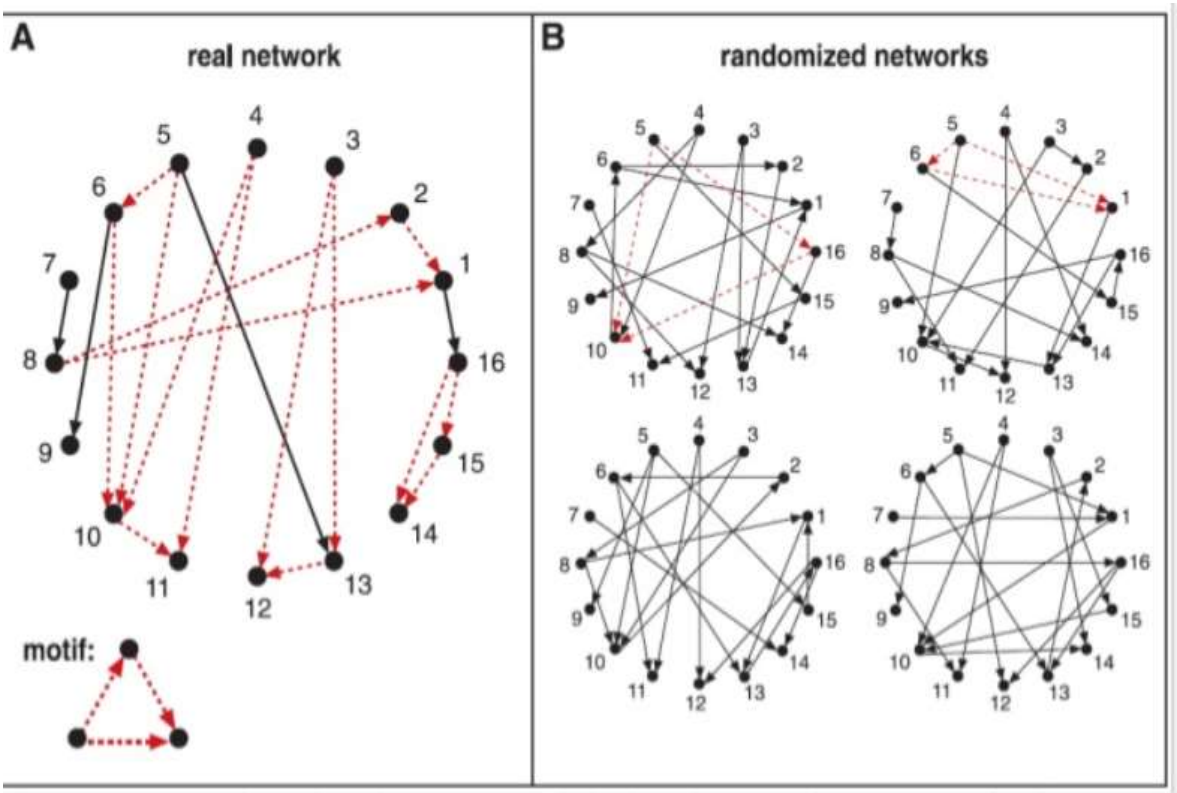
B- protein-protein interactions (undirected graphs)

C- signaling network diagrams (mixed graphs, directed/undirected)

D- ODE modeling of signaling pathways (directed and weighted)

E- PDE modeling of signaling pathways considering space (directed, weighted and nodes can move or be at different compartments)

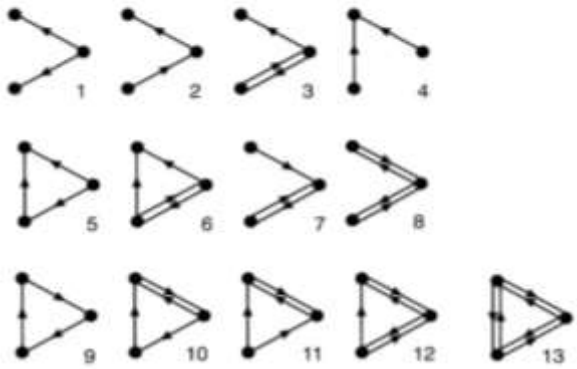




Network Motif Detection

- The "network motifs" are those patterns for which the probability P of appearing in a randomized network an equal or greater number of times than in the real network is lower than a cutoff value (P = 0.01).
- Patterns that are **functionally important but not statistically significant** could exist, which would be missed by this approach.
- Qualitative measure of statistical significance: $Z_{score} = \frac{N_{real} - N_{rand}}{SD}$

13 types of three-node connected subgraphs.



Network motifs are simple building blocks of complex networks and are statistically over represented sub structures or sub graphs in a network. Since the number of sub graphs in biological networks increases exponentially with the network or motif size, it is difficult to detect larger network motifs in a biological network.

Hence network motifs can be defined as recurring patterns of interactions that are significantly over represented in a biological system. This over representation of the sub network indicates the functional importance of such motif

Therefore it becomes important to explore these abundant motifs in biological networks. Milo et al, (*Science 2002) in one of their break through explorations ,work analyzed 18 different networks from*

1. Transcription networks of E.coli and S.cerevisiae

2. Synaptic connections between neurons in c. elegans

3. Throphic interactions between predator and prey in ecological systems

Each of these networks the number of **nodes is represented by 'n'.** In each of the above mentioned cases all possible motifs of size **n=3 and n=4 were enumerated and compared to an average count over thousand random networks.**

Randomized networks were generated without compromising on certain properties of the original network.

1. **In-degree, out- degree and mutual degree**: this is done by swapping edges to generate random graphs.

2. The number of appearances of all n-1 node sub graphs for n>3. This is done to ensure that high significance is assigned to highly significant sub pattern also.

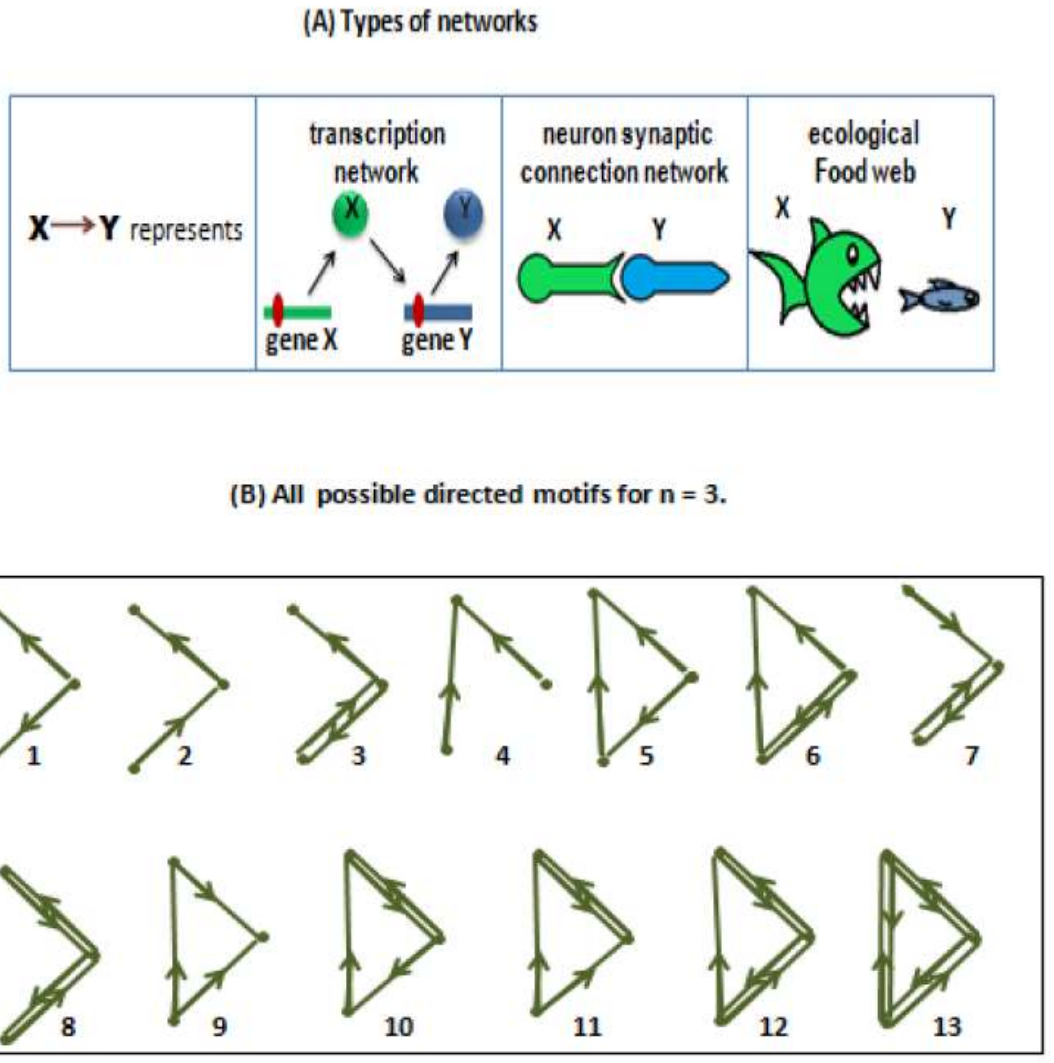


Fig 3: (a) Types of network investigated in Milo et. al (Science 2002)
(b) Directed motifs for n=3 where n refers to the number of nodes.

Identification of Network Motifs

The following rules help us identify network motifs in biological systems

1. Identification all sub graphs of n nodes in the network

2. Randomization of the network without changing the number of nodes, edges and degree distribution.

3. Identification of all sub graphs of n nodes in the randomized network.

4. Comparison of more frequently occurring significant sub graphs with randomized ones in the network and designating them as motifs.

Real live transcription networks of organisms like *E.coli show numerous patterns of nodes and edges but it is important to look for meaningful patterns that are statistically significant to derive biological information that brings us to discuss the concept of randomized networks.*

Randomized networks are a type of networks that posses the same characteristics of a real network. **These have the same numbers of nodes and edges as in biological systems**. But in these networks **random connections are made between nodes and edges**.

Network **motifs are patterns that occur more significantly and more often in real networks than in randomized networks**. Recurrent patterns give us the basic idea that these must have been evolutionarily conserved against mutation.

The best way to illustrate this is the fact that edges are easily lost in a transcription network and a single mutation abolishes transcription factor binding hence facilitating the loss of edge in the network.

In the same way mutations which **generate a binding site for transcription factor X in promoter region of gene Y can help add new edges to the network**.

Mutations, duplication events that reposition pieces on the genome or insertion events can generate new binding sites and hence add new edges to the network.

This clearly demonstrates that the occurrence of the **network motifs more often than in randomized network proves that this selection offers an advantage to the organism.**

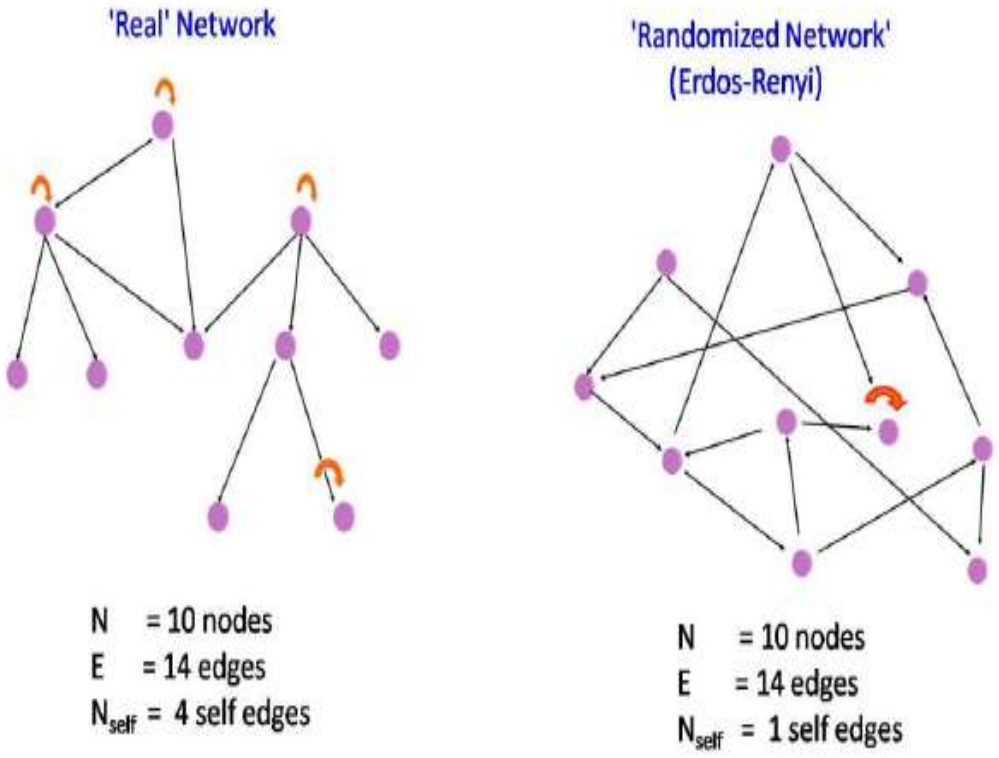
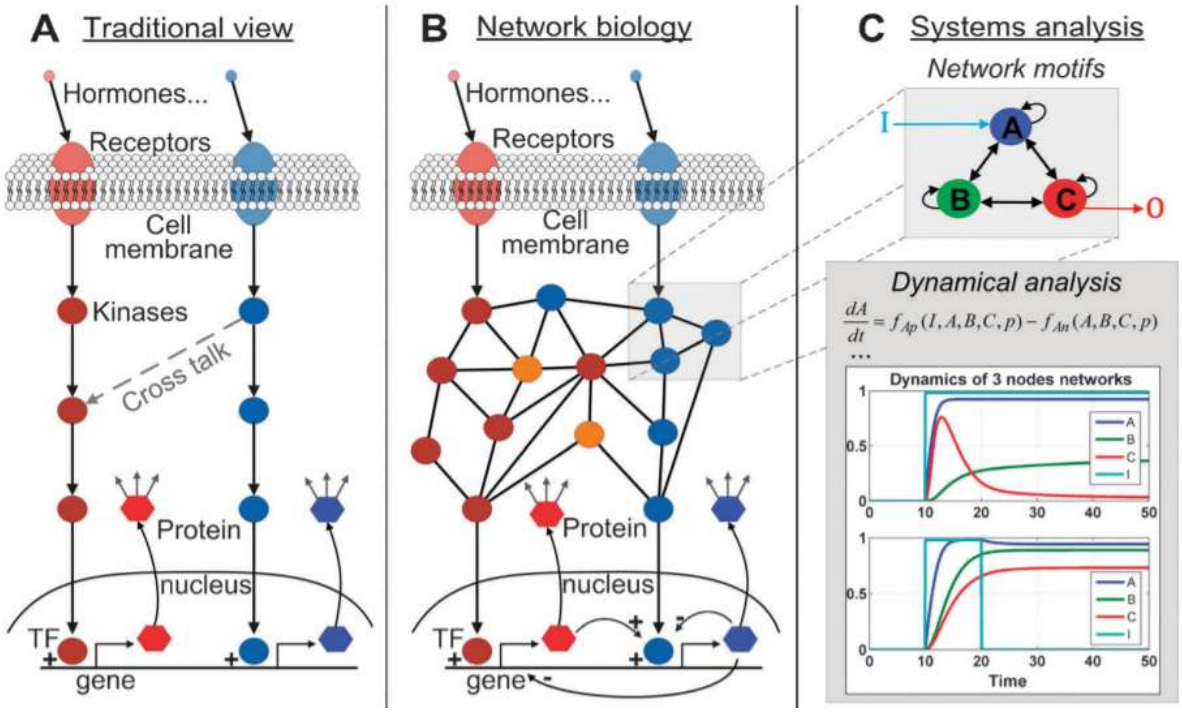


Fig 4: Representation of Real and Randomized Network with nodes, edges and self edges

What questions to ask?

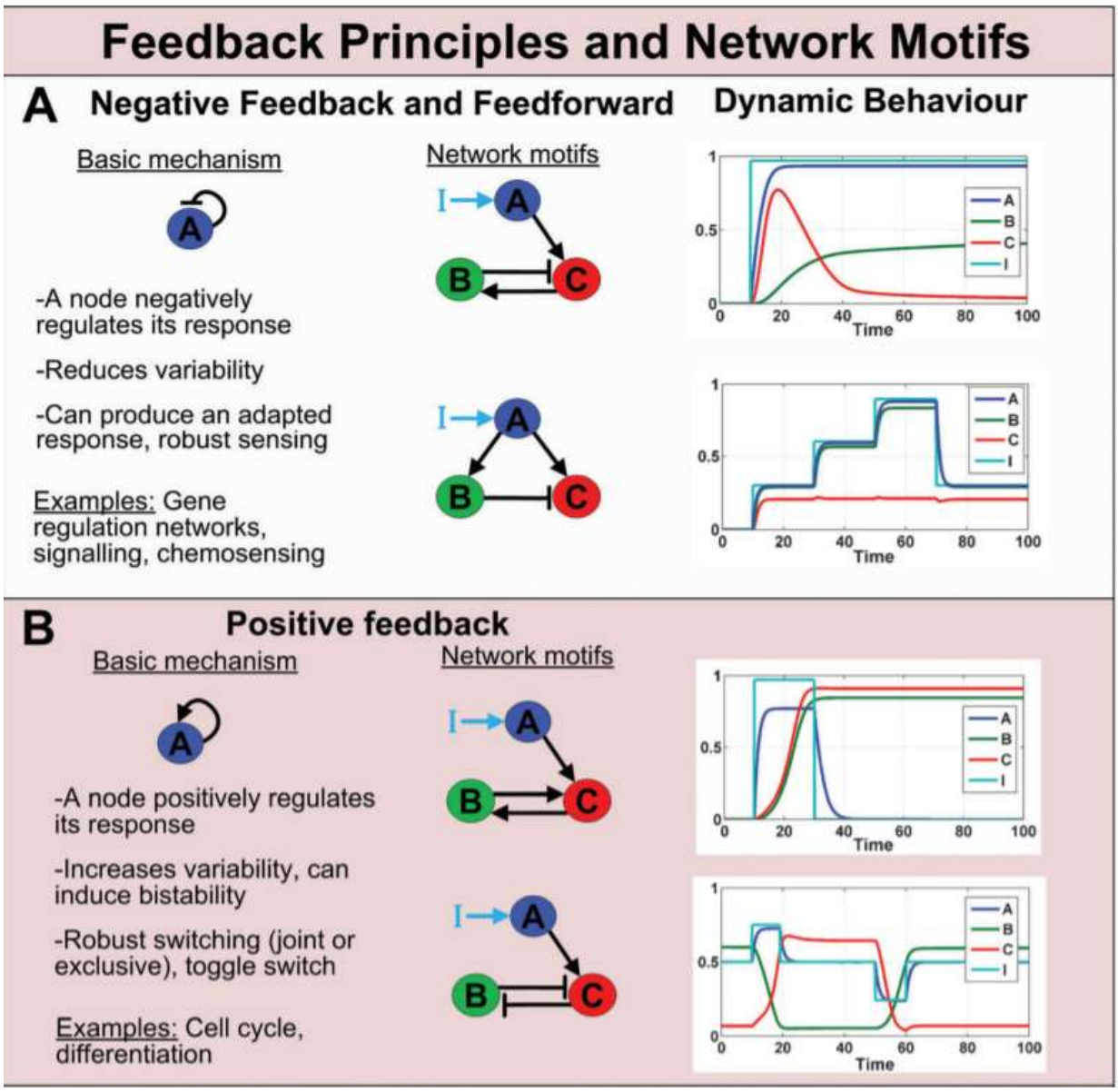
- How many elements? How many interactions? (*n*,*e*)
- Directed/ undirected, weighted/ unweighted?
- What kind of connectivity? What kind of connectivity distribution?
- Is there any special importance to a node or a group of nodes or a link?
- Is it important for the network to be **together**? Good (information dissemination) and bad (disease spread).
- How are the degree-degree correlations?
- Are there special **motifs** for performing specific functions?
- How to identify structural and functional modules?



(A) Traditional, pathway specific view with receptors activating linear signaling cascades leading to TFs activation, gene expression and protein.

(B) Network view where biological information from high-throughput experiments is used to build an unbiased network where components are not necessarily linked to one pathway. Also shown in (B) is the possible feedback loops within gene expression networks.

C: Systems analysis of biological networks components. Here a 3 nodes network is analyzed with ODEs, showing 2 recurrent cases: adaptation and bistable switching

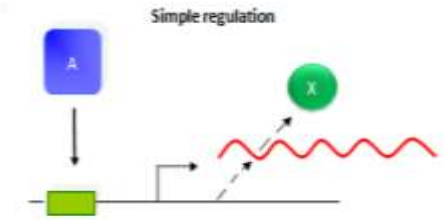


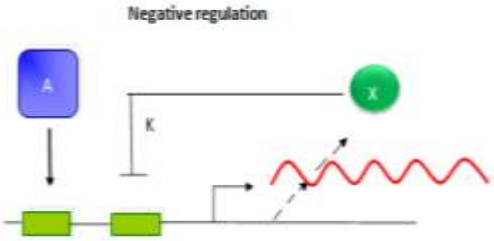
Self edges are those edges that originate and end at the same node.

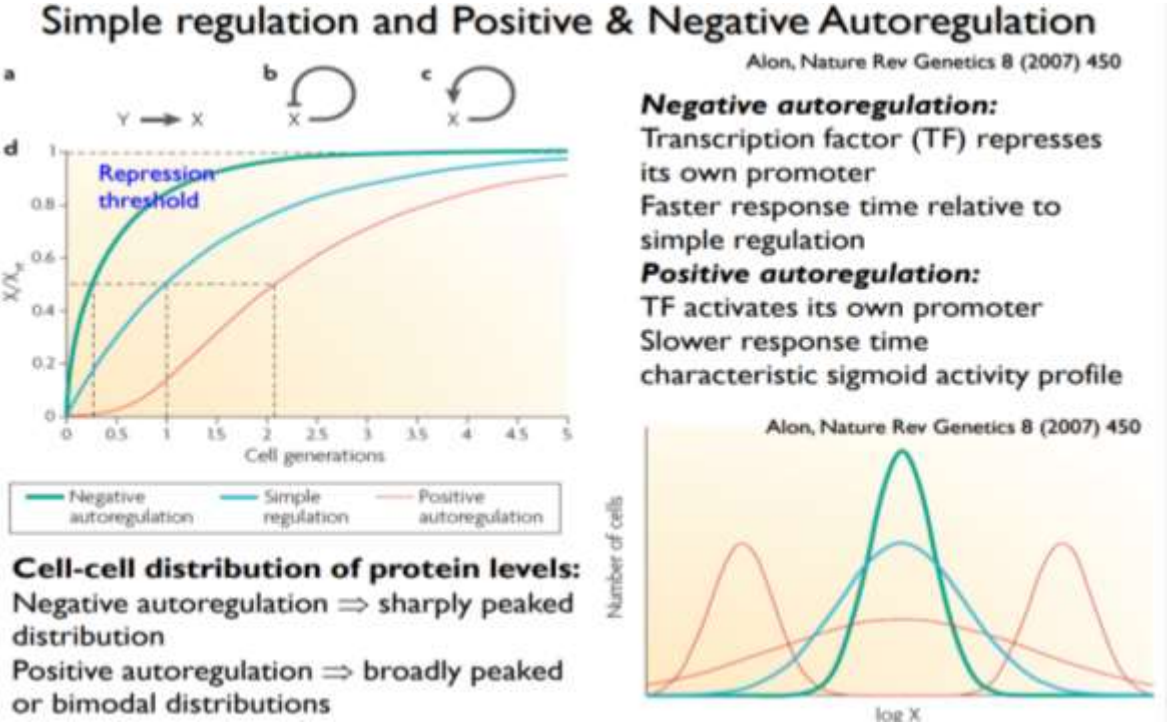
The E.coli network has approximately 40 self edges each of which correspond to transcription factors that regulate the transcription of their own genes. This type of regulation of a gene by its own gene product is called autogenous control or auto regulation.

34 of the auto regulatory proteins in the E.coli network have been found to repress their own transcription. This process is referred to as negative auto regulation as referred to in fig Negative auto regulation is a network motif and occurs at higher numbers than expected in random networks.

Such structures display engineering advantages and are more prone to evolutionary selection in their appearances network motif.







- In order to understand the general features of such networks and to extract useful information from them, **we dissect them at hierarchical levels- into modules and motifs which can explain their functionality, evolution and dynamic behavior.**
- Over the process of **<u>evolution, these networks show information processing functions.</u>**
- Interesting investigations on **network behaviour have shown that simple switching circuits, amplifiers or oscillators can map to the core process of biological decision making**.
- These have been implemented by **two or three gene network motifs and are characterized by how they behave around fixed points in the system.** Here the steady state of the system as well as the process of achieving equilibrium in the system reflects the characteristic function performed by the genetic circuit.
- Network motifs appear at frequencies much higher than those expected at random and hence imply information processing roles for these motifs.

- To arrive at such significant patterns, one first identifies the different patterns of these motifs in real and randomized networks and then calculates the number of appearances of these patterns in the real and random networks.
- The discussion that follows focuses on patterns with 3 nodes (forming a triangle), **Fig.There are 13 possible 3-node patterns in such arrangement.**
- **Of these, only one of them qualifies to be a network motif called the Feed Forward Loop.**

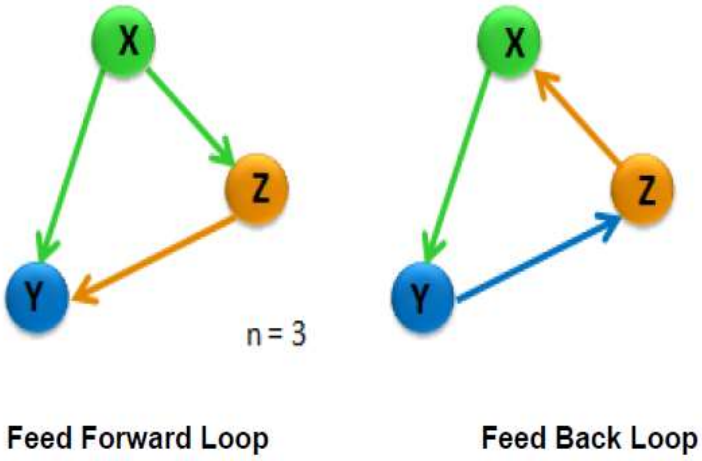


Fig 1 Representative Feed Forward and Feed Back Loops with nodes n=3 forming a triangle

-

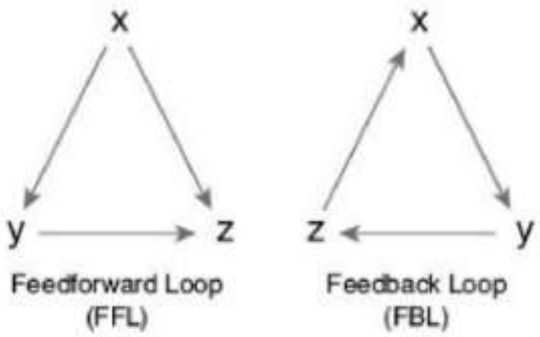
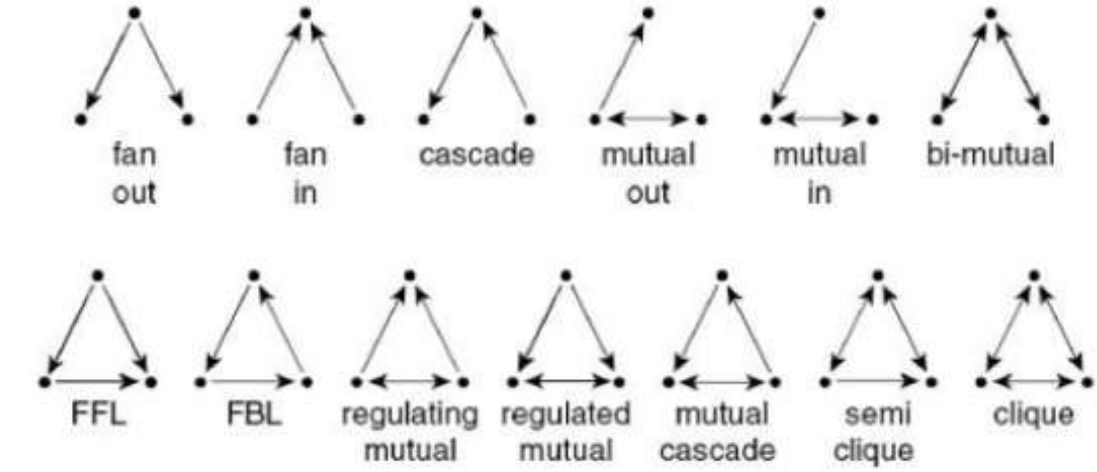
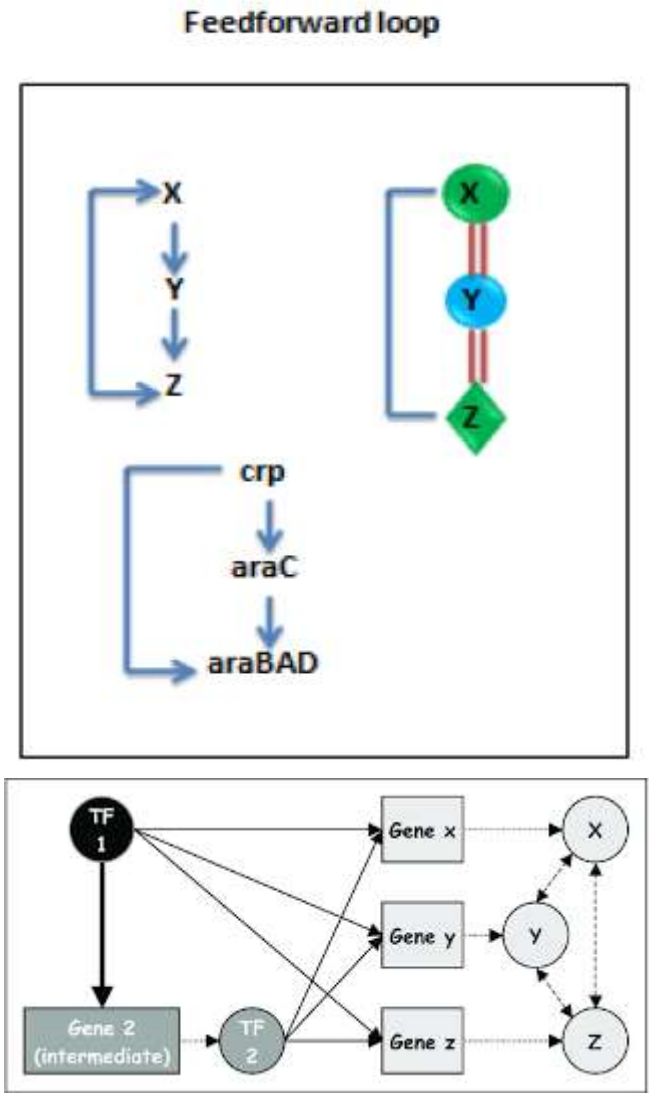


FIGURE 3.1



- The most significant of the network motifs in E. coli and yeast is the Feed Forward Loop which is defined by a transcription factor X that regulates a second transcription factor Y.
- X and Y both jointly regulate an operon Z by binding to its regulatory region.
- Here X is called the general transcription factor, Y the specific transcription factor and Z the effector operon.
- As described in Fig , this type of motif occurs in the L-arabinose utilization system where Crp is the general transcription factor and Ara-C is a specific transcription factor. Such a motif characterizes 40 effector operons in 22 different systems in the network database and accommodates 10 different transcription factors



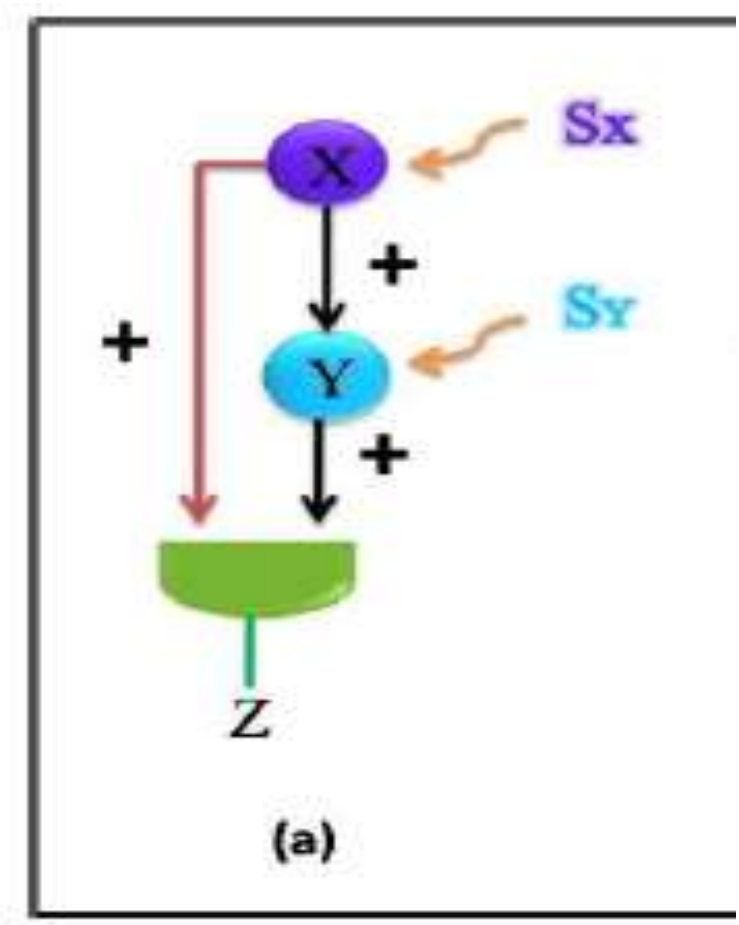
Coherent Feed Forward Loops

A Feed Forward Loop is **termed 'coherent' if the direct effect of the general transcription factor X on the effector operons has the same sign (positive or negative) as its net indirect effect through the specific transcription factor.**

In other words, **if X regulates Y positively and if X and Y both positively regulate Z, the Feed Forward Loop is coherent**. i.e. the sign of the direct path of regulation (X to Z) is the same as the overall sign of the indirect regulation path (X to Z through Y).

The overall sign of a path **is determined by the multiplication of the sign of each arrow on the path.**

In **Fig 3 (a) we see that the sign of the indirect path (Xà YèZ) is plus x plus =plus, while the direct path (XèZ) is already plus. Since both the direct and indirect paths have the same positive sign, this loop is called a Coherent Feed Forward Loop.**



Incoherent Feed Forward Loops

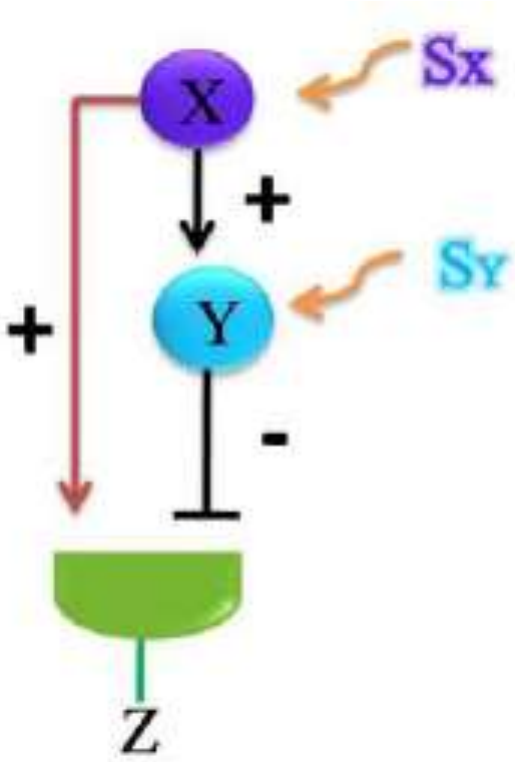
The other type of FFL (as in Fig) is called Incoherent FFL in which the sign of the indirect path of regulation is opposite to that of the direct path.

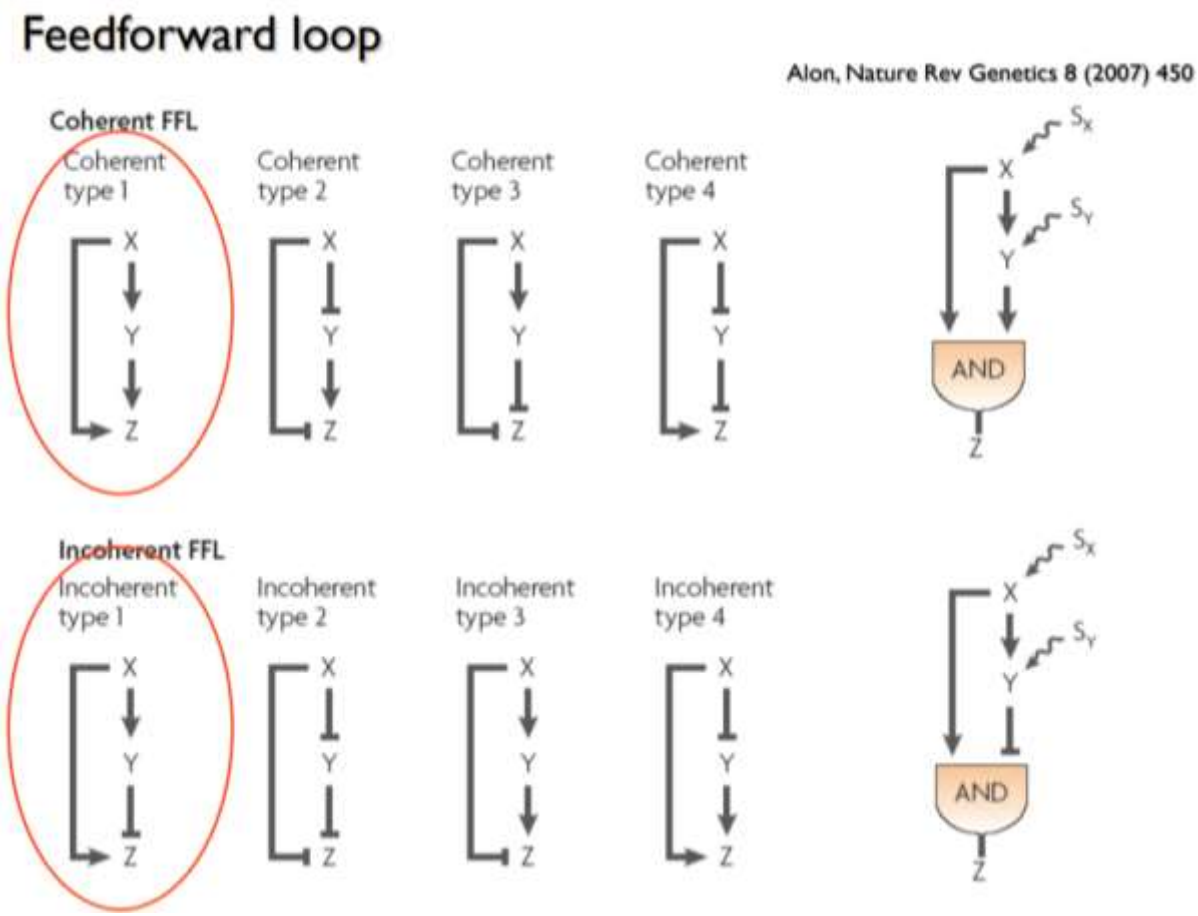
In type-1 Incoherent FFL as denoted the direct path is positive and the indirect path is negative. The Incoherent FFLs show odd number of minus edges.

In both the coherent and incoherent loops, the effects of the general and specific transcription factors X and Y are integrated at the promoter region of gene Z.

The expression profile of Z is modulated by the concentrations of X and Y bound to their inducers.

The cis regulatory input function of Z describes this modulation. cis regulatory input functions include logic gates like AND which require both X and Y to express Z and OR gates in which either X or Y is sufficient to express Z.





Both Coherent and Incoherent Feed Forward Loops are sign sensitive.

Type 1 coherent FFLs (in which all three regulations are positive) are the most abundant type of Feed Forward Loops.

The Incoherent Feed Forward loop type-1 is the second most abundant type of FFL among biological networks.

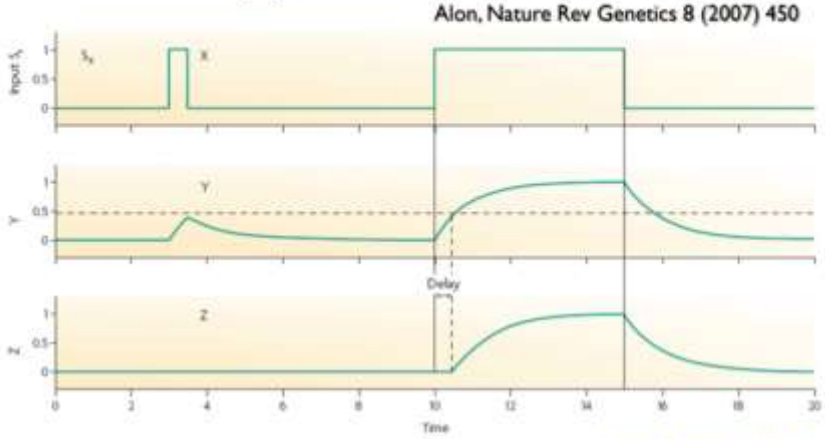
The other types of feed forward loop do not appear more frequently than CFFL I and ICFFL I.

Coherent Feedforward loop: persistence detector

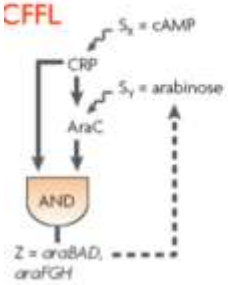
Alon, Nature Rev Genetics 8 (2007) 450

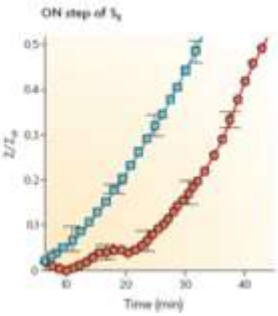
The CFFL shows a delay after stimulation starts but no delay after stimulation stops:

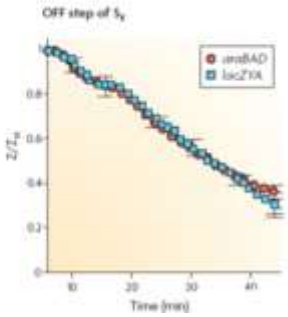
A 'sign-sensitive' delay element for filtering out brief spurious signal pulses



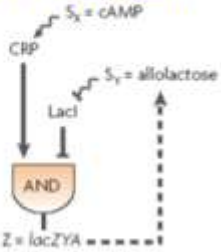
CFFL





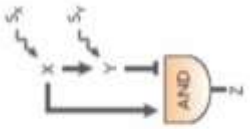


Simple regulation



Incoherent Feedforward loop: pulse generator

Alon, Nature Rev Genetics 8 (2007) 450

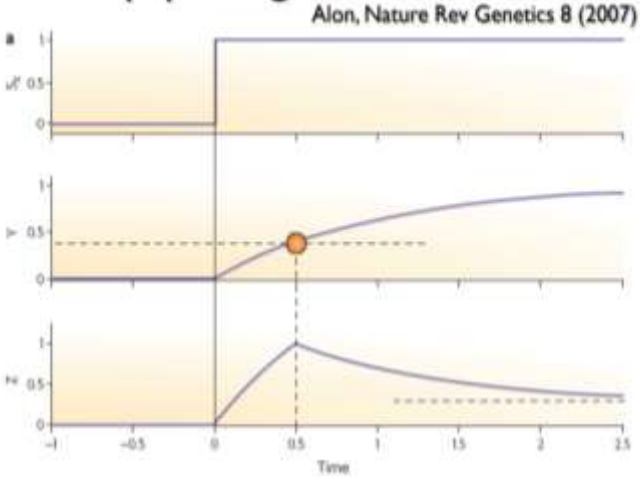


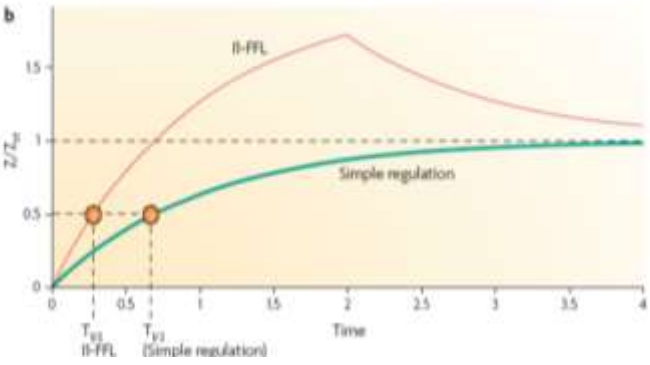
The two branches of IFFL act in opposition

X activates Z, but indirectly represses Z by activating its repressor Y

Initially, signal activating X causes rapid production of Z Later Y levels accumulate to repression threshold for Z, decreasing its production

Pulse-like dynamics and response acceleration relative to simple regulation



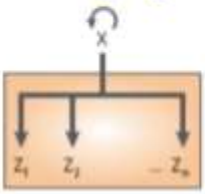


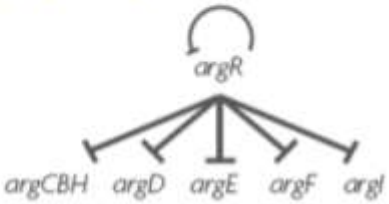
Single-input module

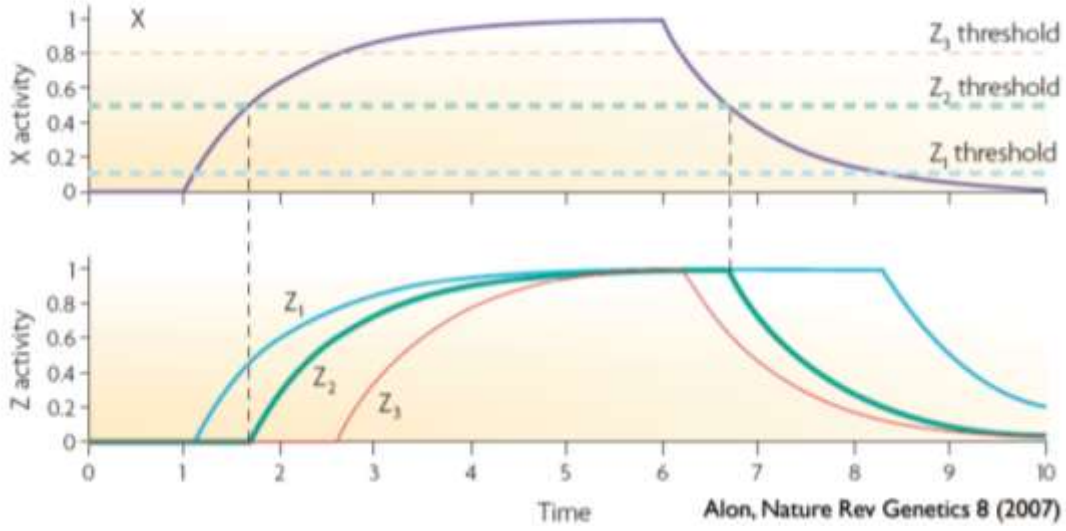
Allows coordinated expression of a group of genes with shared function

Can generate a temporal expression program with a defined sequence of activation of each target by using different thresholds

A single regulator X controls a group of target genes $Z_1, Z_2, \ldots Z_n$







Alon, Nature Rev Genetics 8 (2007) 450

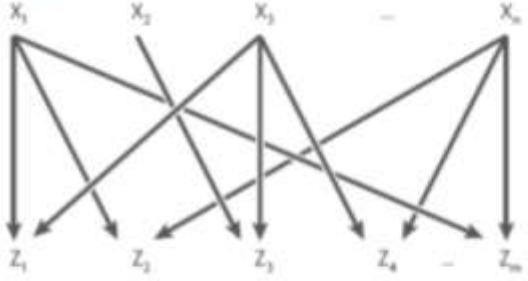
Multi-input motifs

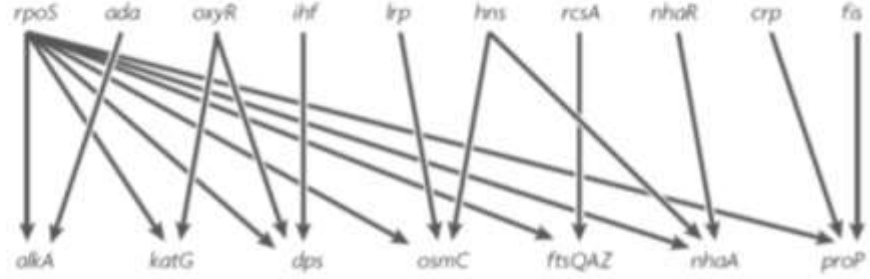
Many inputs regulate many outputs

Alon, Nature Rev Genetics 8 (2007) 450

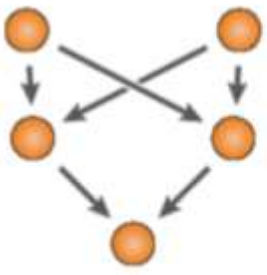
A set of regulators combinatorially control a set of output genes

Can be responsible for a broad function, e.g., carbon utilization, stress response, anaerobic growth (E Coli), etc.





Multi-layer perceptrons



Similar to multi-layer perceptron model of neural networks – but only one layer !

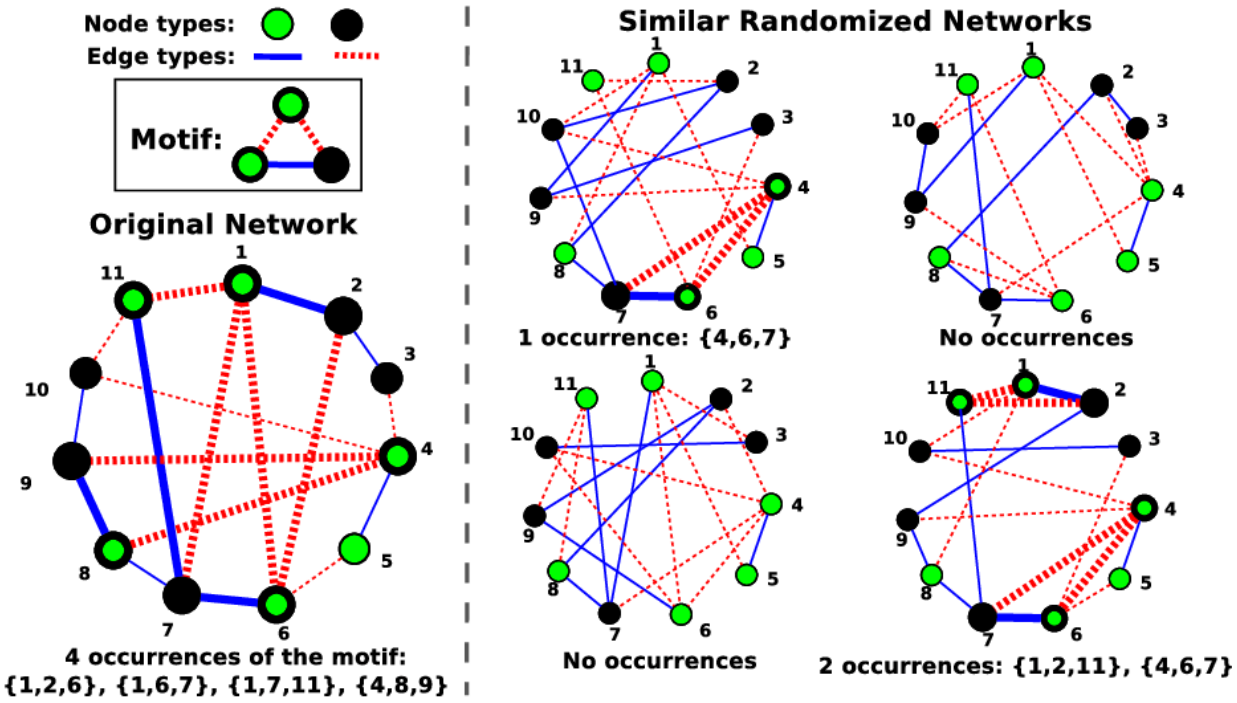
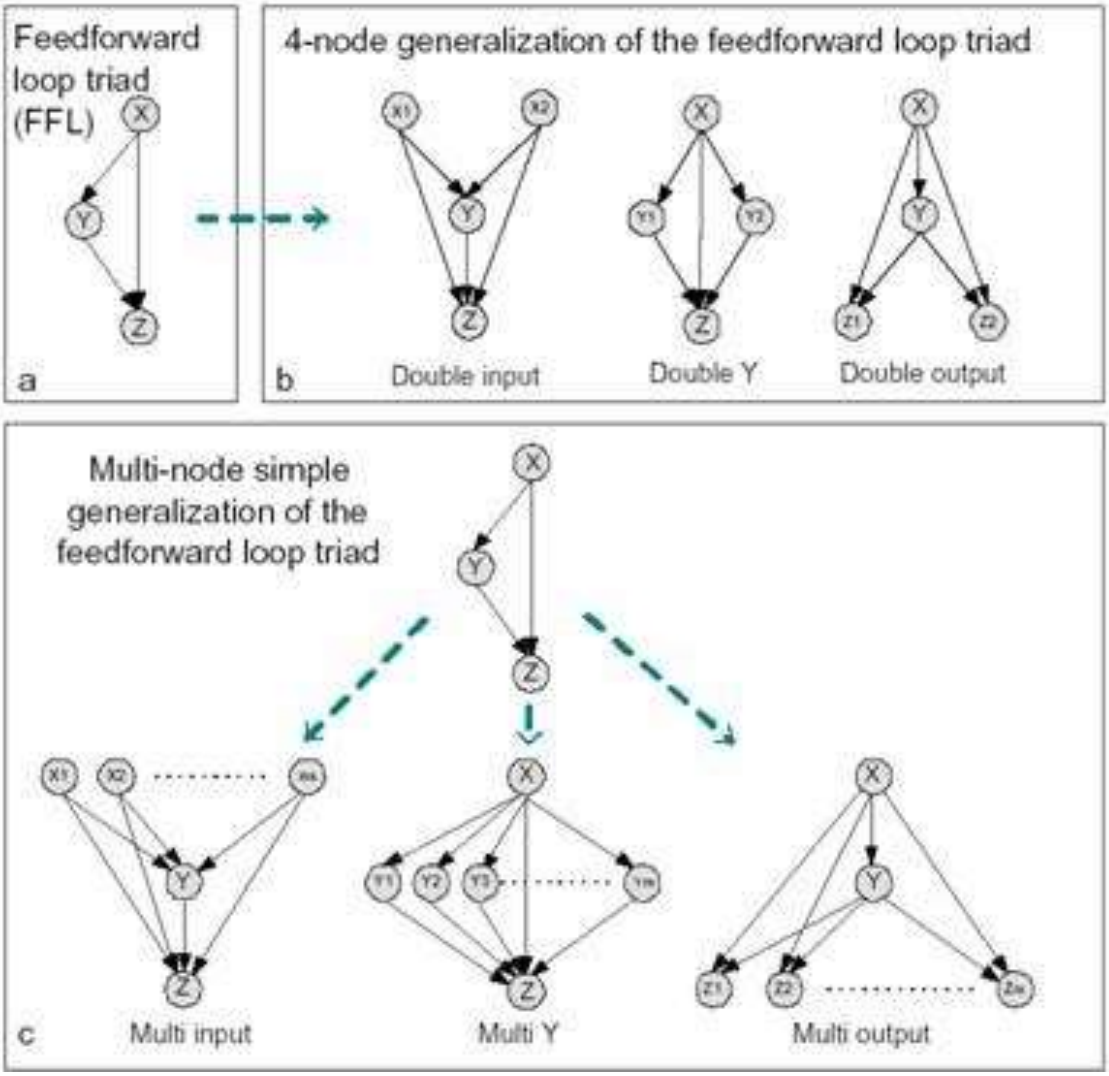
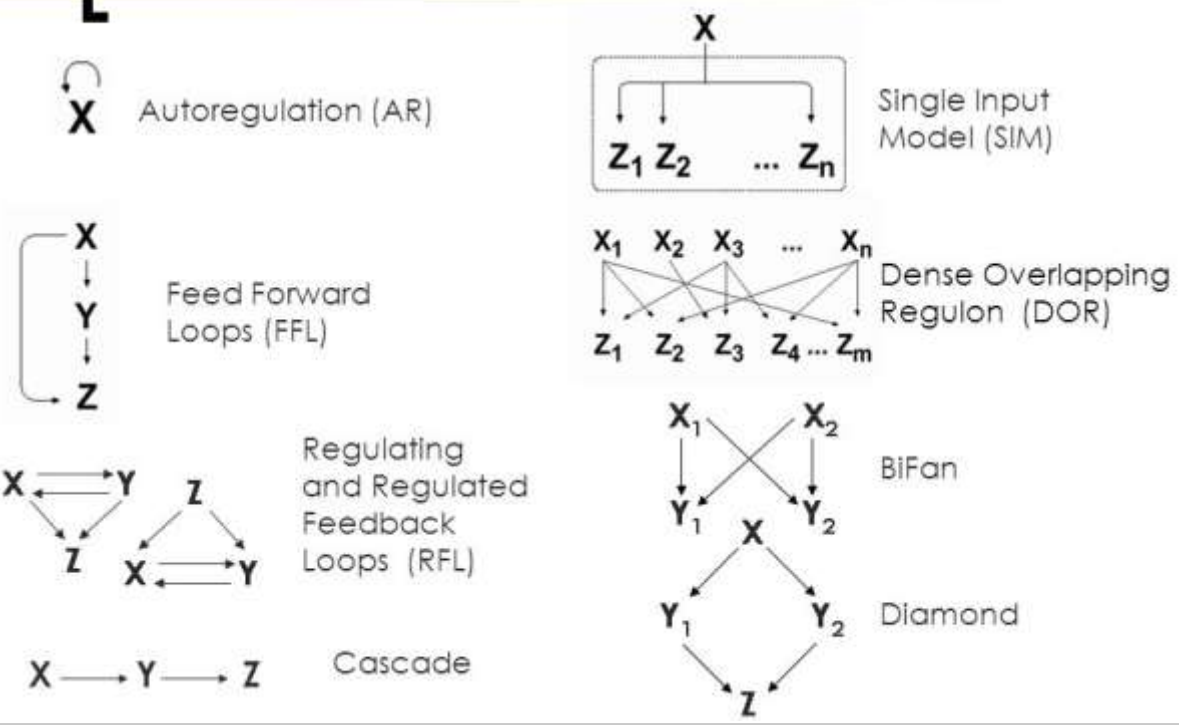


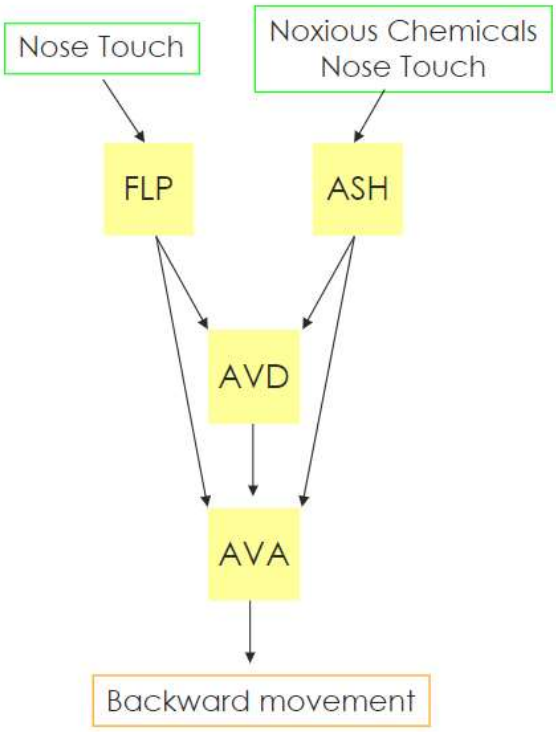
Fig. 1 An example colored network motif with 3 nodes.

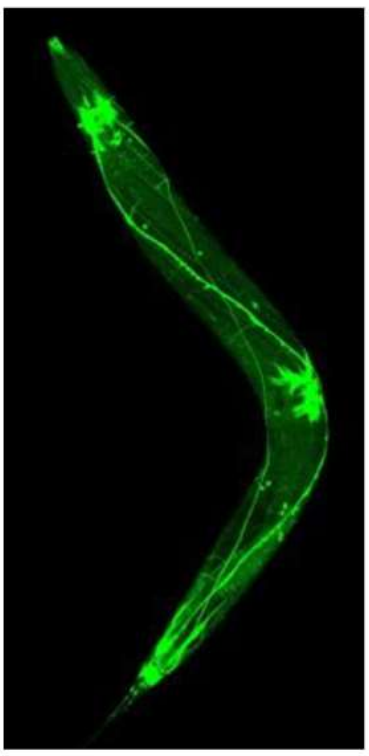


Biological Network motifs

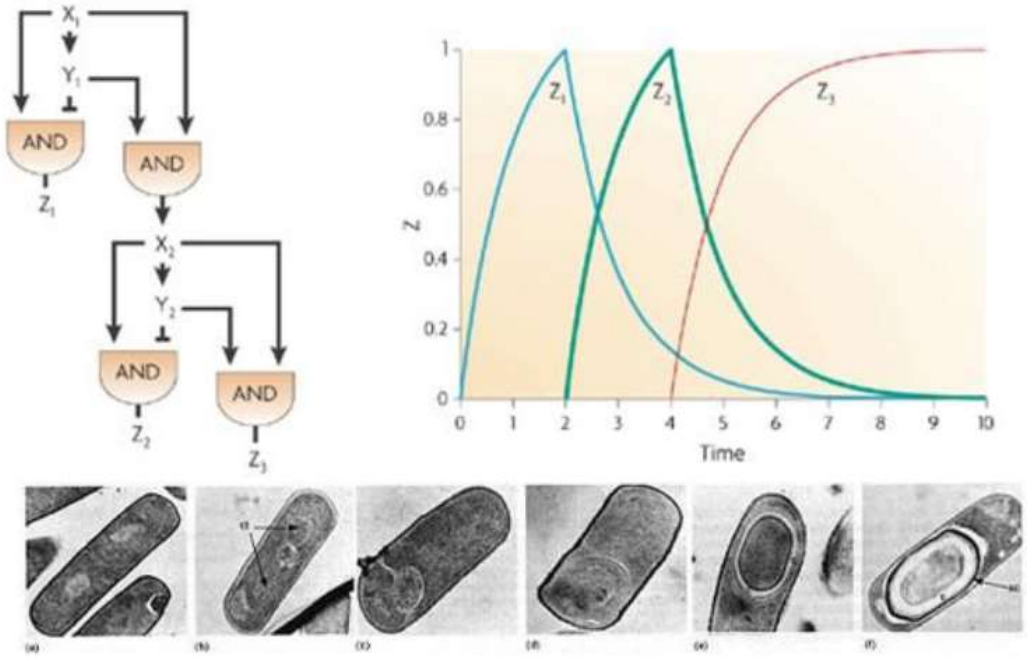


Multi-input FFL in Neuronal Networks



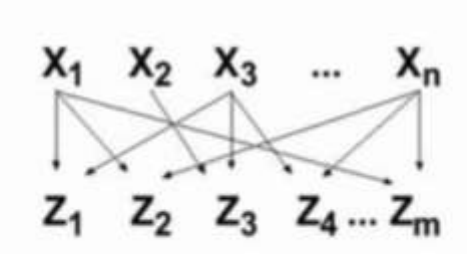


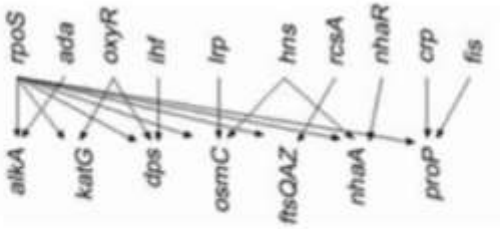
Interlocking Feed forward loops



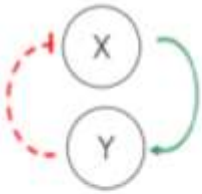
Bacillus Subtilis sporuation process

Dense Overlapping Regulon (DOR)

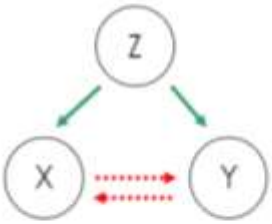




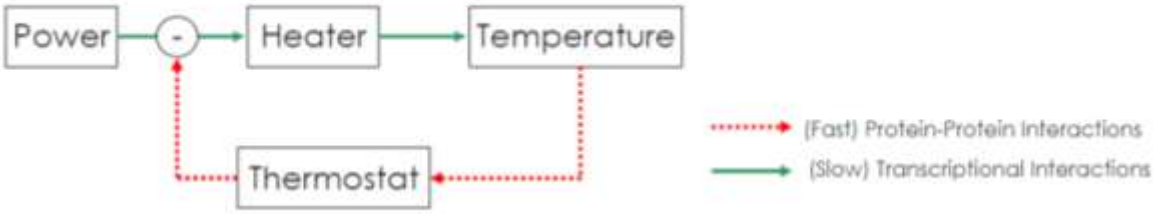
Feed Back Loops



- X transcriptionally activates X.
- Y inhibits X.



- Z transcriptionally activates X and Y
- X forms a complex with Y.
- X phosphorylates Y.



Developmental Transcription Netwroks

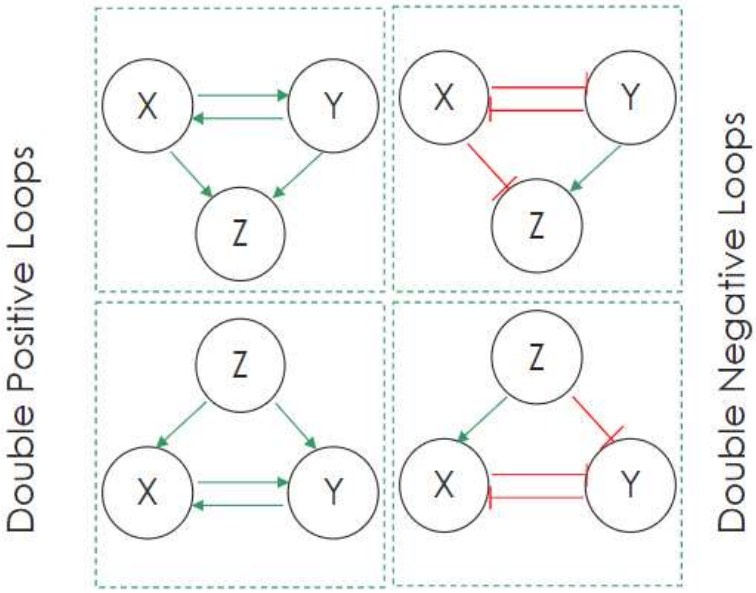
Two-node Feedback Loops

- Both X **AND** Y are ON at the same time.
- Genes regulated by X and Y belong to the same tissue (or strip).

- X **OR** Y is ON at a given time.
- Genes regulated by X and Y belong to different tissues (strips).

Developmental Transcription Netwroks

Regulating Feedback Loops



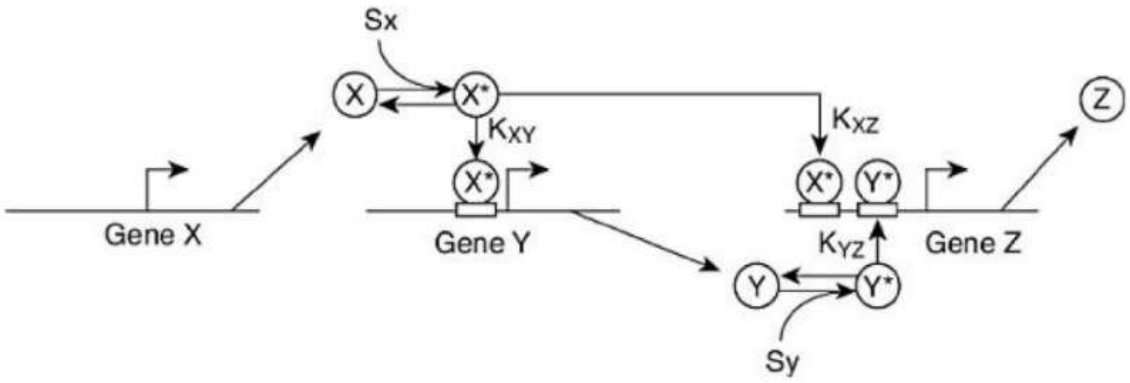
Double Positive Loops

Double Negative Loops

Regulated Feedback Loops

DYNAMICS OF THE COHERENT TYPE-1 FFL WITH AND LOGIC

suppose that the cell expresses numerous copies of protein X, the top transcription factor in the FFL. The input to X is the signal S_x (Figure 3.7). Without the signal, X is in its inactive form. Now, at time $t = 0$, the signal S_x appears and triggers the activation of X. This is known as a **step-like stimulation** of X. As a result, the transcription factor X rapidly transits to its active form X^*. The active protein X^* binds the promoter of gene Y, initiating production of protein Y, the second transcription factor in the FFL. In parallel, other copies of X^* bind the promoter of gene Z. However, since the input function at the Z promoter is AND logic, X^* alone cannot activate Z production.



3.5 THE C1-FFL IS A SIGN-SENSITIVE DELAY ELEMENT

Production of Z requires binding of both X^* and Y^*. Z activation thus requires that the second input signal, S_y, is present, so that Y is in its active form, Y^* (Figure 3.7). Moreover, the concentration of Y^* must build up to sufficient levels to cross the activation threshold for gene Z, denoted K_{YZ}. This results in a delay in Z production.

We will now mathematically describe the FFL dynamics, in order to see how a simple model can be used to gain an intuitive understanding of the function of a gene circuit. We'll use logic input functions. Production of Y occurs at rate β_Y when X^* exceeds the activation threshold K_{XY}, as described by the step function θ:

$$\text{production rate of } Y = \beta_Y \theta(X^* > K_{XY}) \tag{3.4.1}$$

When the signal S_x appears, X rapidly shifts to its active conformation X^*. If the signal is strong enough, X^* exceeds the activation threshold K_{XY} and rapidly binds the Y promoter to activate transcription. Thus, Y production begins shortly after S_x. The accumulation of Y is described by our familiar dynamic equation with a term for production and another term for removal:

$$\frac{dY}{dt} = \beta_Y \theta(X^* > K_{XY}) - \alpha_Y Y \tag{3.4.2}$$

The promoter of Z is governed by an AND-gate input function. The AND gate can be described by a product of two step functions, because both regulators need to cross their activation threshold:

$$\text{production of } Z = \beta_Z \theta(X^* > K_{XZ}) \theta(Y^* > K_{YZ}) \tag{3.4.3}$$

Thus, the C1-FFL gene circuit has three activation thresholds (numbers on the arrows). The dynamics of Z are the balance of a production term with an AND input function and a removal term:

$$\frac{dZ}{dt} = \beta_Z \theta(X^* > K_{XZ}) \theta(Y^* > K_{YZ}) - \alpha_Z Z \tag{3.4.4}$$

We now have the equations needed to study the C1-FFL.

3.5 THE C1-FFL IS A SIGN-SENSITIVE DELAY ELEMENT

To analyze the dynamics of the C1-FFL, we will consider the response to steps of S_x, in which the signal S_x is absent and then saturating S_x suddenly appears (ON steps). We will also consider OFF steps, in which S_x is suddenly removed. For simplicity, we assume throughout that the signal S_y is present, so that the transcription factor Y is in its active form:

$$Y^* = Y \tag{3.5.1}$$

3.5.1 Delay Following an ON Step of S_X

Following an ON step of S_x, Y begins to be produced at rate β_Y. Hence, as we saw in Chapter 1, the concentration of Y begins to exponentially converge to its steady-state level $Y_{st} = \beta_Y/\alpha_Y$ (Figure 3.8):

$$Y^* = Y_{st}(1 - e^{-\alpha_Y t}) \tag{3.5.2}$$

$$Y^*(TON) = Yst(1 - e^{-\alpha Y TON}) = KYZ \tag{3.5.3}$$

which can be solved for T_{ON}, yielding:

$$TON = \frac{1}{\alpha Y} \log\left(\frac{1}{1 - KYZ/Yst}\right) \tag{3.5.4}$$

This equation describes how the duration of the delay depends on the biochemical parameters of the protein Y (Figure 3.9). These parameters are the removal rate of the protein, α_Y, and the ratio between its steady-state level Y_{st} and its activation threshold K_{YZ}. The delay can, therefore, be tuned over evolutionary timescales by mutations that change these biochemical parameters.

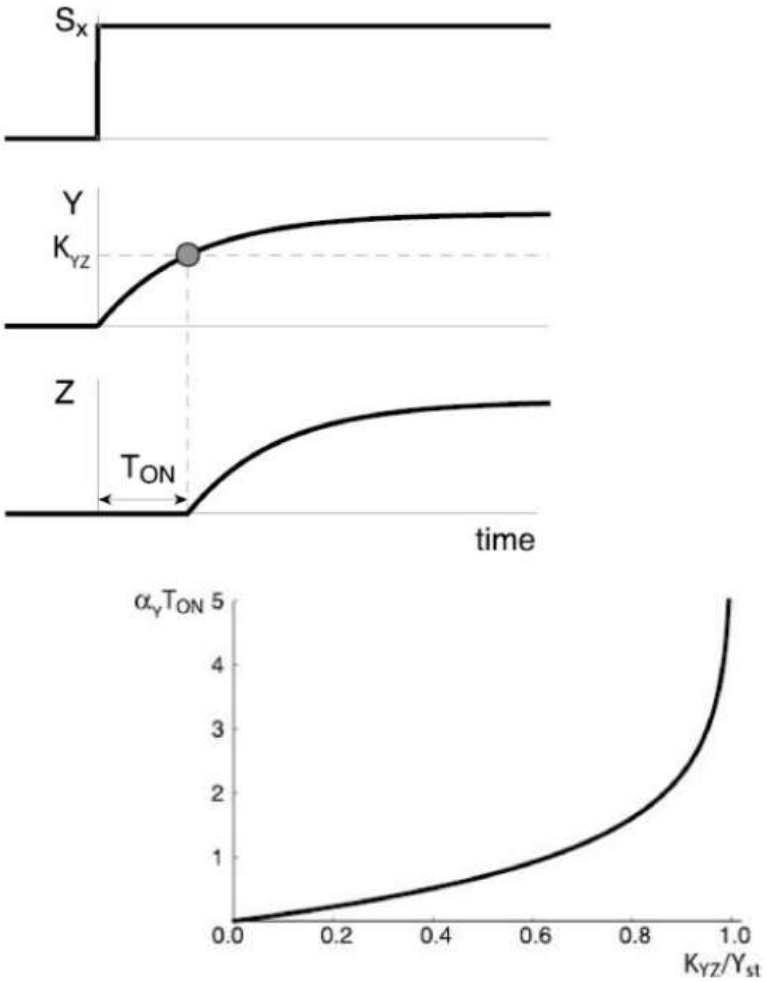


FIGURE 3.9

Note that the delay T_{ON} diverges when the activation threshold K_{YZ} exceeds the steady-state level of Y, because protein Y can never reach its threshold to activate Z (Figure 3.9). Recall that Y_{st} is prone to cell–cell fluctuations due to variations in protein production rates. Hence, a robust design will have a threshold K_{YZ} that is significantly lower than Y_{st}, to avoid these fluctuations. In bacteria, typical parameters provide delays T_{ON} that range from a few minutes to a few hours.

UNIT - III

Introduction to Systems Biology-SBI1401

Network Motifs and Graphlets

- To analyse the structure of biological networks
- They can roughly and historically be divided into local and global network properties.
- degree of a node (the number of edges that the node participates in),
- degree distribution (the distribution of degrees over all nodes of a network),
- clustering coefficient of a node (the number of edges between the neighbours of a node as a percentage of the maximum possible number of edges between them),
- average clustering coefficient of the network over all its nodes etc.

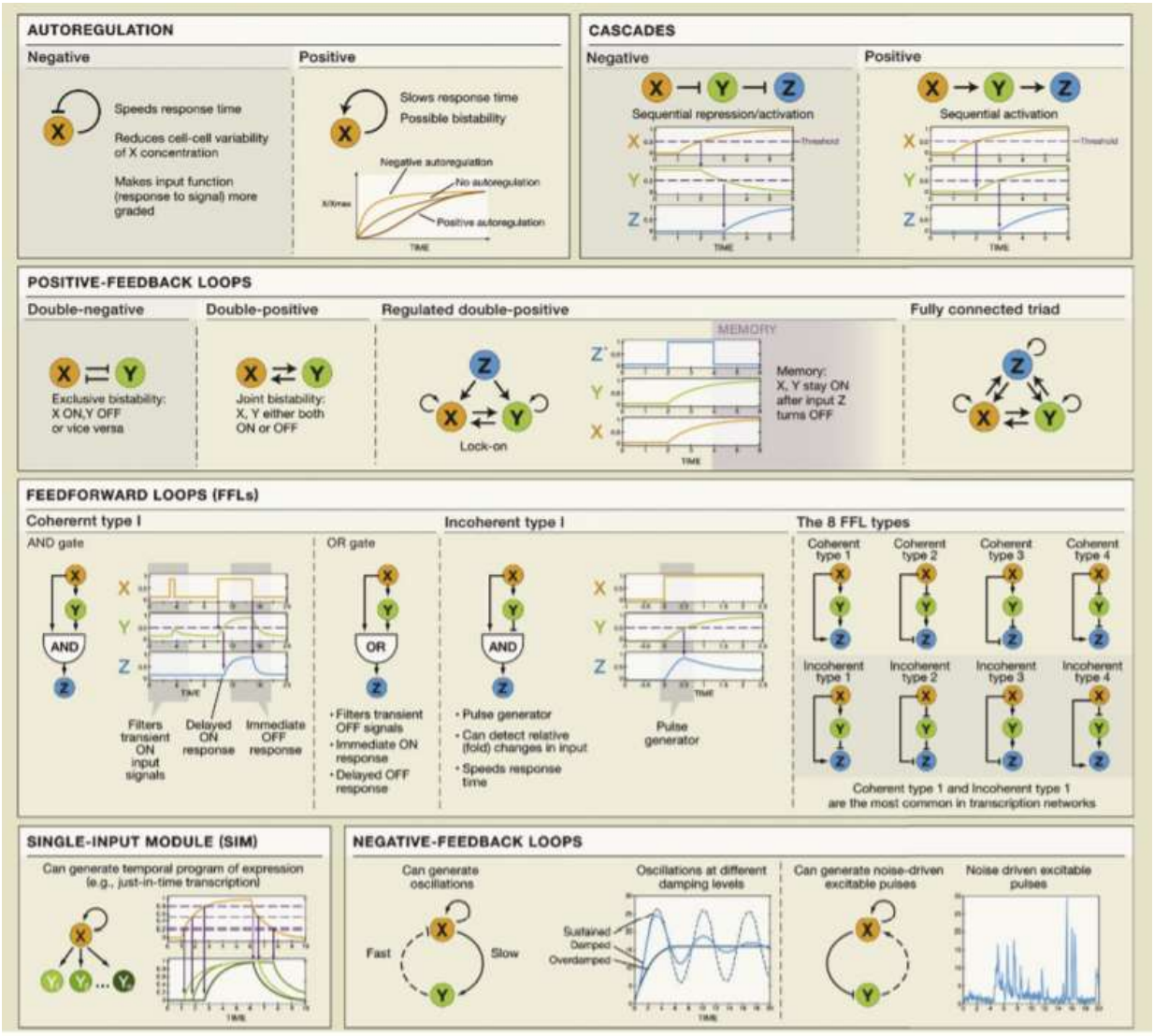
- Network motifs have been introduced by the group of Uri Alon (Milo et al., 2002).

- They are defined as small patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks (Milo et al., 2002).

- They were used to study the transcriptional regulation networks of well-studied microorganisms (Alon and Mangan, 2003; Mangan et al., 2003), as well as of higher order organisms (Charney et al., 2017; Datta et al., 2017).

- It was shown that these networks appear to be made up of a small set of recurring regulation patterns, captured by network motifs.

- Example motifs include positive and negative autoregulation, positive and negative cascades, positive and negative feedback loops, feedforward loops (FFLs), single input modules, and combinations of these, illustrated in Fig (Shoval and Alon, 2010).
- They were linked with biological function. Since the same network motifs have been found in diverse organisms from bacteria to humans, it has been suggested that they serve as basic building blocks of transcription networks.
- To understand the difference in the definition of motifs and graphlets, we need to introduce the following simple graph theoretic terms.
- A partial subgraph is a subgraph of a larger network in which once we pick the nodes that form the subgraph, we can pick any subset of edges between the chosen nodes of the larger network.
- An induced subgraph is a subgraph in which we must pick all the edges between the chosen nodes of the larger network to form the subgraph.
- Network motifs are partial subgraphs that are significantly overrepresented in the data compared to a chosen random graph model that is assumed to fit well the data



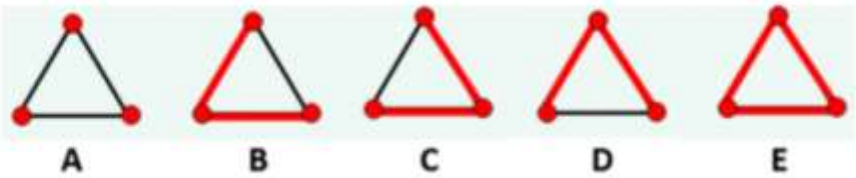


Fig. 2 Partial *vs* induced subgraphs. In the fully connected network with 3 nodes illustrated in panel A, if we pick all 3 of its nodes to make connected subgraphs, there exist three partial subgraphs corresponding to 3-node paths, illustrated in panels B, C and D, but only one induced subgraph, a triangle, illustrated in panel E.

However, as network comparison is computational intractable, determining a well-fitting network model is hard, as it involves comparing the data network with model networks.

A random network model that is usually used as well-fitting is that of a random graph constructed to have the same degree distribution as the data network, while edges are drawn at random.

Also, note that motifs are partial subgraphs, while characterizing the structure of any graph class is based on induced subgraphs

Hence, analyses of biological networks involving network motifs have been criticised, since the definitions of motifs and anti motifs heavily depend on the choice of a random graph (network)

model (Artzy-Randrap et al., 2004; Milo et al., 2004), since they are partial subgraphs (Przulj et al., 2004) and since they can exhibit a whole range of dynamic behaviours (Ingram et al., 2006).

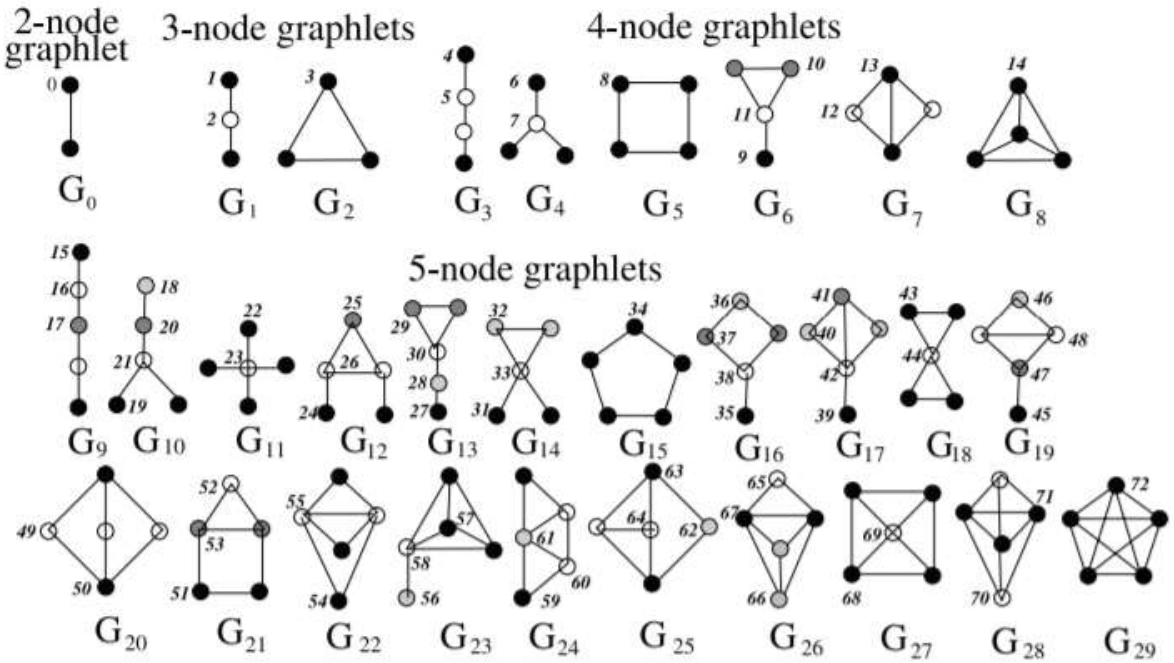
Furthermore, it is computationally hard to identify network motifs (and the same holds for graphlets), as the number of possible sub-graphs increases exponentially with the network and motif size (node and edge counts).

Graphlets are small, connected, induced subgraphs of large networks.

They can appear in the data network at any frequency and hence their definition does not depend on any assumed random graph model that fits well the data.

Also, they are induced, so they are a suitable tool for designing various algorithms for mining domain relevant new information from the structures (topologies) of biological and other network data.

All 2- to 5-node undirected graphlets are illustrated in Fig



Graphlet frequency distribution is introduced to be superior to the degree distribution (which is the first in the spectrum of 73 graphlet degree distributions), the clustering coefficient (the measure of "cliquishness" of the network) and network diameter (which measures how "far spread" the nodes of the network are) by imposing a large number of similarity constrains on the networks being compared.

Basically, it compares sequences of numbers over 73 pairs to compare two networks.

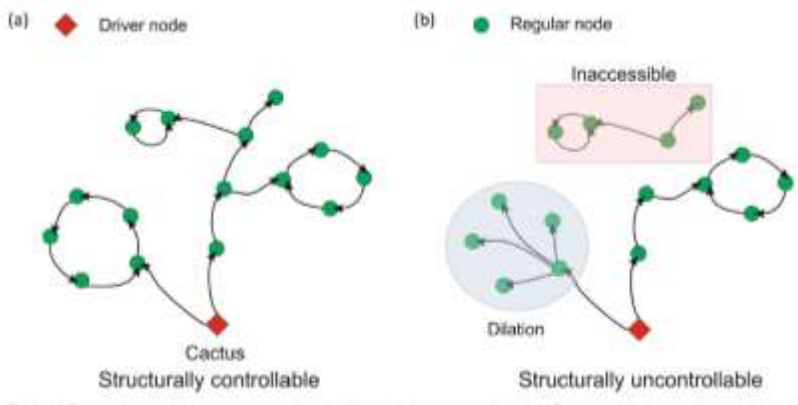
It can be classified into both types of network comparison heuristics, global and local, as it does comparison globally over the entire network, but uses local network features in the comparisons

Control of Neuronal Network in Caenorhabditis elegans

Caenorhabditis elegans, a soil dwelling nematode, is evolutionarily rudimentary and contains only*300 neurons which are connected to each other via chemical synapses and gap junctions. This structural connectivity can be perceived as nodes and edges of a graph.Controlling complex networked systems (such as nervous system) has been an area of excitement for mankind. Various methods have been developed to identify specific brain regions, which when controlled by external input can lead to achievement of control over the state of the system. But in case of neuronal connectivity network the properties of neurons identified as driver nodes is of much importance because nervous system can produce a variety of states (behaviour of the animal). Hence to gain insight on the type of control achieved in nervous system we implemented the notion of structural control from graph theory to C. elegans neuronal network. We identified 'driver neurons' which can provide full control over the network. They studied phenotypic properties of these neurons which are referred to as 'phenoframe' as well as the 'genoframe' which represents their genetic correlates. Hear they find that the driver neurons are primarily motor neurons located in the ventral nerve cord and contribute to biological reproduction of the animal. Identification of driver neurons and its characterization adds a new dimension in controllability of C. elegans neuronalnetwork.study suggests the importance of driver neurons and their utility to control the behaviour of the organism.

Control of complex networks is an emerging topic in the areas of network science. One such example network in which control of physiological activities/state of the network is of crucial importance is that of neuronal connectivity network. Controllability naturally raises two key questions: what are the points of control and what is to be controlled. Determination of such points of control can be achieved with the help of various graph theoretical measures such as degree, betweenness centrality, closeness and using importance of nodes identified by evolutionary algorithm. The idea of control of brain states is aligned with the studies on control of behaviour (state) of an organism by identifying and controlling a few important regions (nodes) via external inputs (impulses of electric or magnetic fields). From a connectionist paradigm, brain could be thought of as a network of neurons, a complex dynamical system, the state of which is to be controlled. This aspect has been studied as 'structural control' in a network aimed to be achieved with the help of a few 'driver nodes/neurons'.

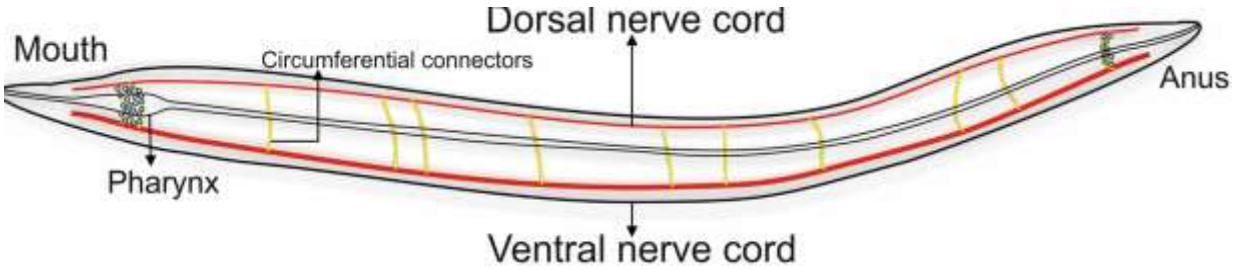
It has been proposed that networks possessing cacti structure (without having inaccessible nodes or dilations) are controllable as shown in Fig



A structural network with linear time
invariant dynamical system could be represented as Eq (1), where x(t) represents the state of the system at time t, A is the state matrix, B input matrix and u(t) is input signal.

$$x(t) = Ax(t) + Bu(t)$$

The state of such a system is proven to be controllable only if it possess full rank



C. elegans neuronal network

Caenorhabditis elegans (C. elegans), a nematode, is a model biological organism whose neuronal network is fully charted. This hermaphrodite animal has rudimentary nervous system consisting of 302 neurons and can process complex information of senses, behaviour and even memory. The neurons are divided into various subtypes and are classified based on their functional roles, location within the body of the animal and span of the neuron axons.

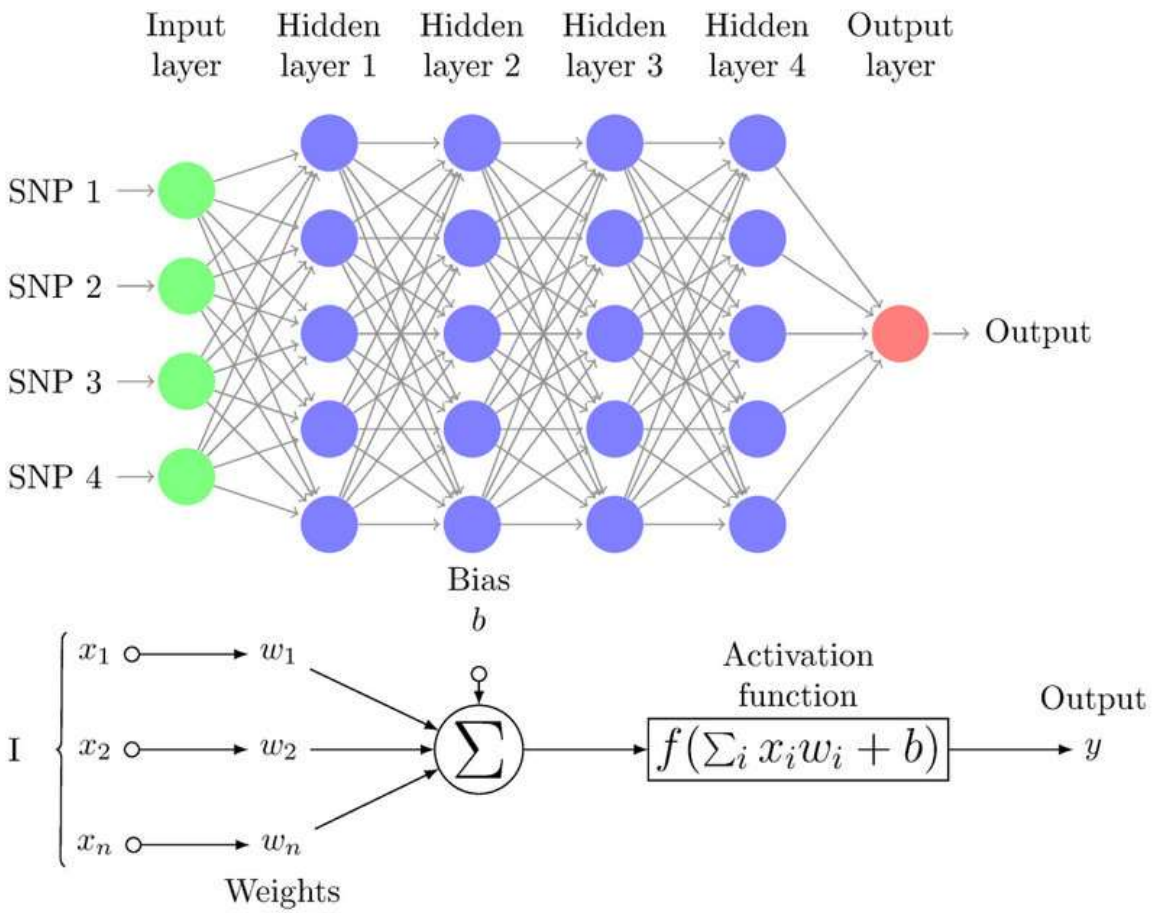
According to functional roles, neurons are primarily of three types viz. sensory neurons, motor neurons and inter neurons. Sensory neurons pick up external signals to which the animal responds by sending motor signals to effector organs through motor neurons which connect to command inter-neurons on dendritic side and neuro-muscular junction on the axonal side. Motor neurons are distributed mainly over the ventral nerve cord (VNC) with ganglia at each end. some of which

extend their processes circumferentially to form a dorsal nerve cord (DNC)as shown in Fig. Both VNC and DNC control locomotion of the animal.

In accordance with definition of driver nodes, these critical neurons control the state of neuronal network when provided with external input. To investigate this state space and what kind of changes one can bring by controlling Dn in C. elegans state we examined phenotypic properties of these neurons. Study of properties such as location, functional type and span of neurons provided us with the potential functional association of driver neurons. Further we investigated specific biological functions underlying these neurons with the help of gene ontological enrichment studies

Information processing using multi layered perceptron

Multilayer perceptron network is one of the most popular NN architectures, which consists of a series of fully connected layers, called input, hidden, and output layers. The layers are connected by a directed graph.



Multi-Layer Perceptron (MLP) diagram with four hidden layers and a collection of single nucleotide polymorphisms (SNPs) as input and illustrates a basic "neuron" with n inputs. One

neuron is the result of applying the nonlinear transformations of linear combinations (xi, wi, and biases b)

The multilayer perceptron network (MLP) is one of the most popular DL architectures, which consists of a series of fully connected layers called input, hidden, and output layers . In the context of genomic prediction, the first layer receives the SNP genotypes (x) as input and the first layer output is a weighted nonlinear function of each input plus the ”bias” (i.e., a constant).

The first layer output is then:

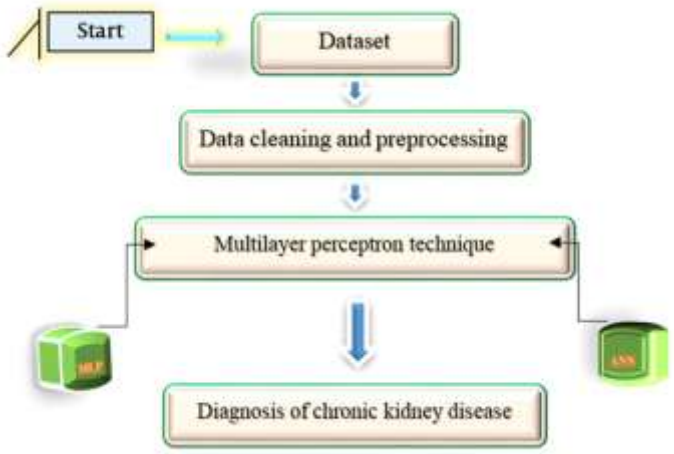
$$z^{(1)} = b_0 + W^{(0)} f^{(0)}(x)$$

$$z^{(k)} = b_{k-1} + W^{(k-1)} f^{(k-1)}\left(z^{(k-1)}\right)$$

where, x contains the genotypes of each individual, b is called the “bias” and is estimated together with the rest of weights W(0), and f is a nonlinear function (available activation functions in Keras are in https://keras.io/activations/). In successive layers, the same expression as above is used except that neuron inputs of a given layer are the outputs from the previous layer

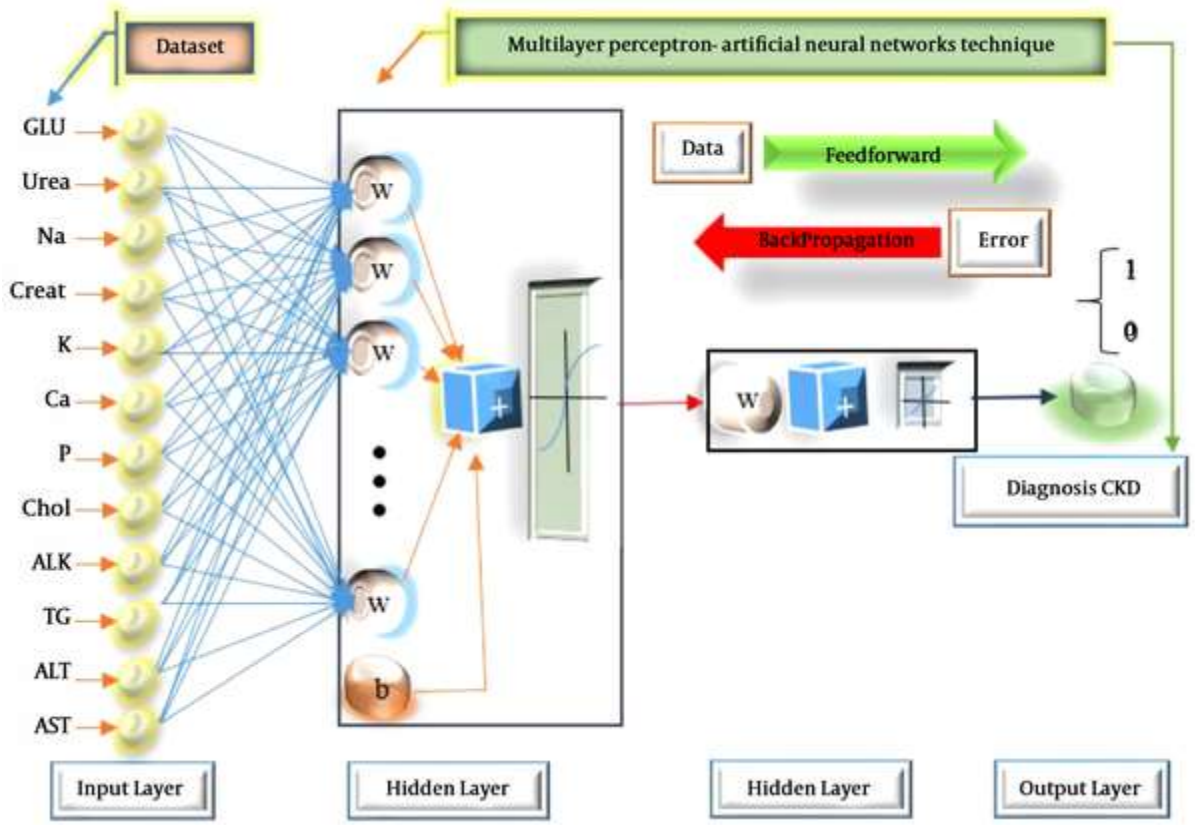
The final layer produces a vector of numbers if the target is a real-valued phenotype, or an array with probabilities for each level if the target is a class (i.e., a classification problem). Although MLPs represent a powerful technique to deal with classification or regression problems, they are not the best option to manage spatial or temporal datasets. To face these issues, other DL techniques such as convolutional neural networks, recurrent neural networks or deep generative networks have been proposed in recent years.

A Novel Classification Method Based on Multilayer Perceptron-Artificial Neural Network Technique for Diagnosis of Chronic Kidney Disease(CKD)



The dataset included blood and urine samples collected from 50 healthy people and 90 patients. Informed consent was obtained from all subjects before enrolment in the study. Samples were kept at -20°C until use. For each person, we gathered the data of the concentrations of glucose (GLU), urea, creatinine, sodium (Na), potassium (K), calcium (Ca), phosphorus (P), cholesterol (Chol), triglycerides (TG), alkaline phosphatase (Alk), alanine aminotransferase (ALT), and aspartate aminotransferase (AST) to use as inputs to the ANN.

ANN architecture with a hidden layer and different numbers of neurons. It had 12 input nodes and one output node and the problem was a binary classification. The output was either 0 or 1 where 0 indicated a healthy case and 1 stood for CKD. To train the network, the weights between hidden-output and input-hidden layers were randomly initialized with a small value ranging from 0 to 1. Figure indicates the framework of ANN for the diagnosis of CKD. The input layer contained 12 neurons. In the hidden layers, there were 10 neurons. The output layer had only one neuron, representing CKD.



UNIT - IV

Introduction to Systems Biology-SBI1401

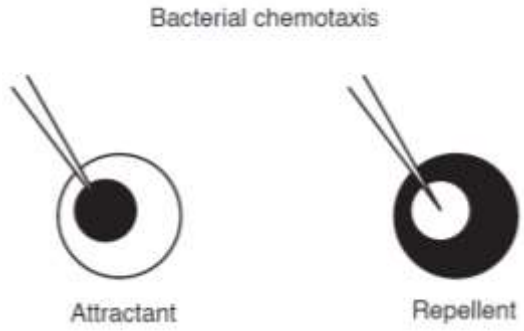
Introduction

We saw how bifunctional proteins can make the input-output relation of a signaling circuit precise despite variation in protein levels. But not all signaling circuits need to simply transduce the signal level. Some circuits are built to make more sophisticated computations, and to do so robustly. To see this, we will now consider the robustness of a remarkable protein circuit called the **bacterial chemotaxis** circuit, that allows bacteria to navigate. Bacterial chemotaxis is so well-characterized on the level of both molecules and behavior that it is a testing ground for important ideas in systems biology, including robustness. We will describe the biology of bacterial chemotaxis, and models and experiments that demonstrate how the computation performed by this protein circuit is made robust to changes in protein levels. We will see that the principle of robustness can help us to rule out many plausible mechanisms and to home in on the correct design.

Bacterial chemotaxis, or how bacteria think

Chemotaxis behaviour

When a pipette containing nutrients is placed in a plate of swimming *Escherichia coli* bacteria, the bacteria are attracted to the mouth of the pipette and form a cloud (Figure 7.1). When a pipette with noxious chemicals is placed in the dish, the bacteria swim away from the pipette. This process, in which bacteria sense and move along gradients of specific chemicals, is called **bacterial chemotaxis**.



Chemicals that attract bacteria are called **attractants**. Chemicals that drive the bacteria away are called **repellents**. *E. coli* can sense a variety of attractants, such as sugars and the amino acids serine and aspartate, and repellents, such as metal ions and the amino acid leucine. Most bacterial species show chemotaxis, and some can sense and move toward stimuli such as light (phototaxis) and even magnetic fields (magnetotaxis).

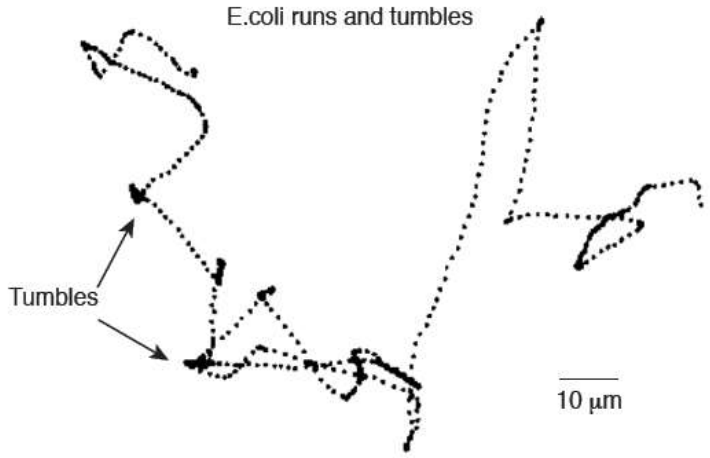
Bacterial chemotaxis achieves remarkable performance despite the great physical limitations faced by the bacteria. Bacteria can detect concentration gradients as small as a change of one molecule per cell volume per micron and function in background concentrations spanning over five orders of magnitude. All this is done while being buffeted by Brownian noise, such that if the cell tries to swim straight for 10 sec, its orientation is randomized by 90° on average.

How does *E. coli* manage to move up gradients of attractants despite these physical challenges?

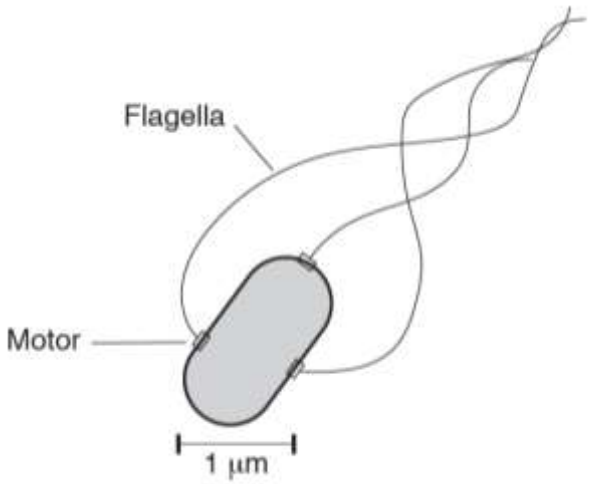
It is evidently too small to sense the gradient along the length of its own body.

The answer was discovered by Howard Berg in the early 1970s: *E. coli* uses **temporal gradients** to guide its motion. It uses a biased-random-walk strategy to sample space and convert spatial gradients to temporal ones. In liquid environments, *E. coli* swims in a pattern that resembles a random walk. The motion is composed of **runs**, in which the cell keeps a rather constant direction, and **tumbles**, in which the bacterium stops and randomly changes direction (Figure).

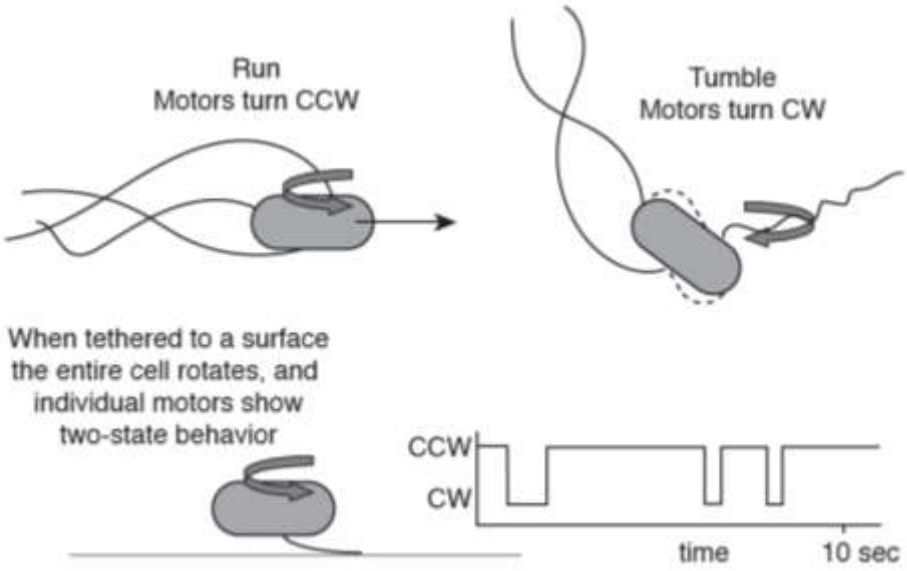
The runs last about 1 sec on average and the tumbles about 0.1 sec.



To sense gradients, *E. coli* compares the current attractant concentration to the concentration in the past. When *E. coli* moves up a gradient of attractant, it detects a net positive change in attractant concentration. As a result, it reduces the probability of a tumble (it reduces its **tumbling frequency**) and tends to continue going up the gradient. The reverse is true for repellents: if it detects that the concentration of repellent increases with time, the cell increases its tumbling frequency, and thus tends to change direction and avoid swimming toward repellents. Thus, chemotaxis senses the temporal derivative of the concentration of attractants and repellents. It follows a simple strategy: If life is getting better, keep going, and if life is getting worse, change direction.



The runs and tumbles are generated by different states of the motors that rotate the bacterial flagella. Each cell has several flagella motors that can rotate either clockwise (CW) or counterclockwise (CCW). When the motors turn CCW, the flagella rotate together in a bundle and push the cell forward. When one of the motors turns CW, its flagellum breaks from the bundle and causes the cell to tumble about and randomize its orientation. When the motor turns CCW, the bundle is reformed and the cell swims in a new direction



Response and exact adaptation

The basic features of the chemotaxis response can be described by a simple experiment. In this experiment, bacteria are observed under a microscope swimming in a liquid with no gradients. The cells display runs and tumbles, with an average **steady-state tumbling frequency** f, on the order of $f \sim 1\ sec^{-1}$

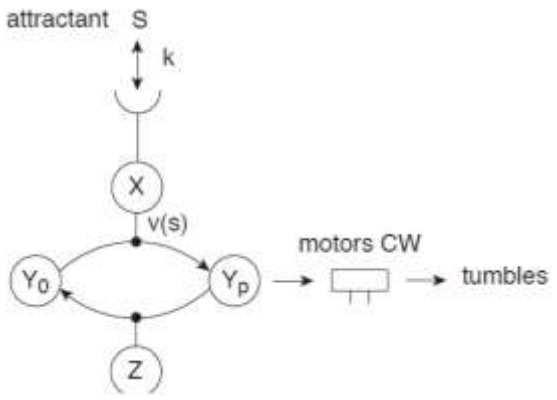
We now add an attractant such as aspartate to the liquid, uniformly in space. The attractant concentration thus increases at once, but no spatial gradients are formed. The cells sense an

increase in attractant levels, no matter which direction they are swimming. They think that things are getting better and suppress tumbles: the tumbling frequency of the cells plummets within about 0.1 sec

After a while, however, the cells realize they have been fooled. The tumbling frequency of the cells begins to increase, even though attractant is still present This process, called **sensory adaptation**, is common to many biological sensory systems. For example, when we move from light to dark, our eyes at first cannot see well, but they soon adapt to sense small changes in contrast. Adaptation in bacterial chemotaxis takes several seconds to several minutes, depending on the size of the attractant step Bacterial chemotaxis shows **exact adaptation**: the tumbling frequency in the presence of attractant returns to the same level as before attractant was added. In other words, *the steady-state tumbling frequency is independent of attractant levels*. If more attractant is now added, the cells again show a decrease in tumbling frequency, followed by exact adaptation. Changes in attractant concentration can be sensed as long as attractant levels do not saturate the receptors that detect the attractant. Exact adaptation poises the sensory system at an activity level where it can respond to multiple steps of the same attractant, as well as to changes in the concentration of other attractants and repellents that can occur at the same time. It prevents the system from straying away from a favorable steady-state tumbling frequency that is required to efficiently scan space by random walk.

The chemotaxis protein circuit

We now look inside the *E. coli* cell and describe the protein circuit that performs the response and adaptation computations. The input to this circuit is the attractant concentration, and its output is the probability that motors turn CW, which determines the cells' tumbling frequency (Figure). The chemotaxis circuit was worked out using genetics, physiology, and biochemistry, starting with J. Adler in the late 1960s, followed by several labs,including those of D. Koshland, S. Parkinson, M. Simon, J. Stock, and others. The broad biochemical mechanisms of this circuit are shared with many signaling pathways in all types of cells.



Attractant and repellent molecules are sensed by specialized detector proteins called **receptors**. Each receptor protein passes through the cell's inner membrane, and has one part outside of the cell membrane and one part inside the cell. It can thus pass information from the outside to the inside of the cell. The attractant and repellent molecules bound by a receptor are called its **ligands**.

E. coli has five types of receptors, each of which can sense several ligands. There are a total of several thousand receptor proteins in each cell. They are localized in a cluster on the inner membrane, such that ligand binding to one receptor appears to affect the state of neighboring receptors. Thus, a single ligand binding event is amplified, because it can affect more than one receptor (Bray, 2002), increasing the sensitivity of this molecular detection device (Segall et al., 1986; Jasuja et al., 1999; Sourjik and Berg, 2004).

Inside the cell, each receptor is bound to a protein kinase called CheA.[3] We will consider the receptor and the kinase as a single entity, called X. X transits rapidly between two states, active (denoted X*) and inactive, on a timescale of microseconds. When X is active, X*, it causes a modification to a response-regulator protein, CheY which we will denote Y, which diffuses in the cell.

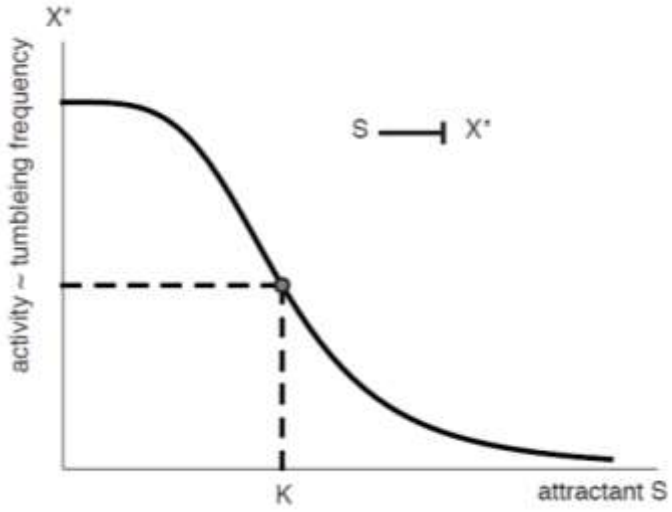
This modification is the addition of a phosphoryl group (PO_4) to Y to form Yp. This type of modification, called **phosphorylation**, is used by most types of cells to pass bits of information among signaling proteins, as we saw in Chapter 6. Yp can bind the flagella motor and increase the probability that it switches from CCW to CW rotation. Thus, the higher the concentration of Yp, the higher the tumbling frequency (Cluzel et al., 2000). The phosphorylation of Yp is removed by the phosphatase CheZ, denoted Z. At steady-state, the opposing actions of X* and Z lead to a steadystate Yp level and a steady-state tumbling frequency. Thus, the main pathway in the circuit is phosphorylation of Y by X*, leading to tumbles We now turn to the mechanism by which attractant and repellent ligands can affect the tumbling frequency.

Attractants lower the activity of X

When a ligand S binds receptor X, it changes the probability[4] that X will assume its active state X*. The concentration of X in its active state is called the **activity of X**. Binding of an attractant *lowers* the activity of X. Therefore, attractants reduce the rate v(S) at which X phosphorylates Y, and levels of Yp drop, resulting in fewer tumbles. These responses occur within less than 0.1 sec. The response time is mainly limited by the time it takes Yp to diffuse over the length of the cells to the motors that are distributed all around the cell membrane. The pathway from X to Y to the motor explains the initial response in Figure 7.5, in which attractant leads to reduction in tumbling. The reduction in activity X* due to the binding of attractant S is well described by a Hill function. Where Xmax is the maximal activity.

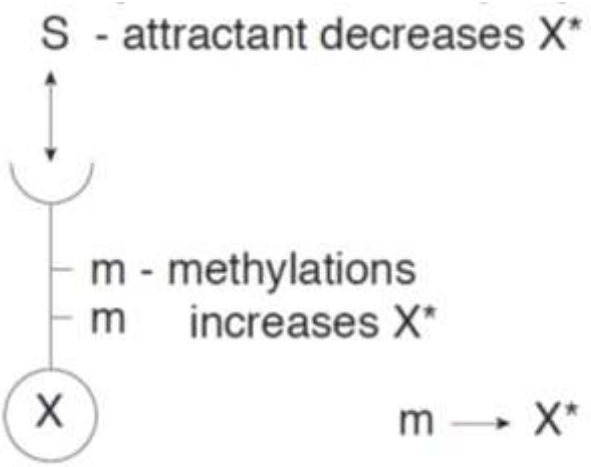
The halfway point for reduction in activity is K, the binding constant of the attractant to the receptor. The Hill coefficient n is due to clusters of n receptors that show cooperativity: binding of ligand to one receptor

in the cluster changes the conformation of the other receptors in the cluster and raises the affinity of ligand to the other receptors.



Adaptation is due to slow modification of X that increases its activity

The chemotaxis circuit has a second pathway devoted to adaptation. As we saw, binding of ligand reduces the activity of the receptor X. However, each receptor has several biochemical "buttons" that can be pressed to increase its activity and compensate for the effect of the attractant (Fig). These buttons are **methylation** modifications, in which a methyl group (CH_3) is added to four or five locations on the receptor. Each receptor can thus have between zero and five methyl modifications. The more methyl groups that are added, the higher the activity of the receptor.



The methylation buttons work by changing the binding constant K of the receptor to attractants[5]. The more methylated the receptor the higher is K (lower chemical affinity to the attractant), (Fig). Therefore, the less attractant it binds, so that there is less inhibition of X activity, X*. In this way, methylation increases receptor activity. Mathematically, we can describe the effect of methylation on K using the concept of free energy ΔG.

The binding constant K is given by the exponential of the free energy of binding the ligand $K = e^{67}$ (the

Boltzmann constant k_9T is included in ΔG) . Each methylation adds some free energy γ to the bound state of the receptor, making it less favorable, so that $\Delta G = \Delta G_< + \gamma m$, where m is the number of methylations. As a result, K increases with methylation, $\sim e_{>^*}$, raising the half-way-point ligand level needed for inhibition of activity, Figure 7.9. The higher the methylation, the higher the half-way point for binding K, and more ligand is needed to reduce the activity X*.

Methylation of the receptors is catalyzed by a protein called CheR and is removed by an a protein called CheB, denoted R and B. Methyl groups are continually added and removed by these two antagonistic proteins, regardless of whether the bacterium senses any ligands (Fig). This seemingly wasteful cycle has an important function: it allows cells to adapt.

Adaptation is carried out by a negative feedback loop through B. This protein removes methyl groups only from receptors in their active conformation, X*. Thus, reduced X activity means that B is less active, causing a reduction in the rate at which methyl groups are removed by B. Methyl groups are still added, though, by R at an unchanged rate. Therefore, methylation increases. Methylation makes the receptor more active, the tumbling frequency increases. Thus, the receptors X first become less active due to attractant binding, and then methylation level gradually increases, restoring X activity. This is a negative feedback loop with a slow arm in which X* reduces methylation, and a fast arm in which methylation raises X* (Fig).

Methylation reactions are indeed much slower than the reactions in the main pathway from X to Yp to the motor (the former are on the timescale of seconds to minutes, and the latter on a sub-second timescale). The protein R is present at low amounts in the cell, about 100 copies, and appears to act at saturation (zero-order kinetics). The slow rate of the methylation reactions explains why the recovery phase of the tumbling frequency during adaptation is much slower than the initial response.

The Barkai-Leibler model of exact adaptation

Early models of chemotaxis used equations to describe the reactions just described and showed response to attractant and exact adaptation. However, in these models, exact adaptation depended on setting specific values for parameters such as the numbers of R and B proteins per cell. These parameters had to be tuned so that methylation could exactly compensate for the reduction in activity caused by attractant. Changing the protein level parameters ruined exact adaptation (Fig).

After adding attractant, the cells responded, but then returned to a different basal activity than before the attractant step. We say that exact adaptation in these models is **fine tuned.** A fine-tuned model is described in solved exercise X.X.robust mechanism for exact adaptation was proposed by Naama Barkai and Stan Leibler. In this mechanism, changing parameters such as R and B protein levels changes the steady-state activity. But changing parameters does not ruin exact adaptation: after a step of attractant, activity first drops but then returns to the pre-step level

In summary, the bacterial chemotaxis circuit has a design such that a key feature (exact adaptation) is robust with respect to variations in protein levels. Other features, such as steady-state activity and adaptation times, are fine-tuned. These latter features show variations within a population due to intrinsic cell–cell variations in protein levels. Because of the robust design, the intrinsic variability in the cell's protein levels does not abolish exact adaptationAs a theorist, one can usually write many different models to describe a given biological system, especially if some of the biochemical interactions are not fully characterized. Of these models, only very few will typically be robust with respect to variations in the components. Thus, the robustness principle can help narrow down the range of models that work on paper to the few that can work in the cell. Robust design is an important factor in determining the specific types of circuits that appear in cells. In the next chapter, we will study how robustness constraints can shape the circuits that guide pattern formation in embryonic development.

References:

Alon, U., Surette, M.G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, 397: 168–171.

Barkai, N. and Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, 387: 913–917.

Berg, H.C. (2003). *E. coli in Motion.* Springer.

Berg, H.C. and Brown, D.A. (1972). Chemotaxis in *Escherichia coli* analyzed by three-dimensional tracking. *Nature*, 239: 500–504.

Berg, H.C. and Purcell, E.M. (1977). Physics of chemoreception. *Biophys. J.*, 20: 193–219.

Knox, B.E., Devreotes, P.N., Goldbeter, A., and Segel, L.A. (1986). A molecular mechanism for sensory adaptation based on ligand-induced receptor modification. *Proc. Natl. Acad. Sci. U.S.A.*, 83: 2345–2349.

Kollmann, M., Lovdok, L., Bartholome, K., Timmer, J., and Sourjik, V., (2005). Design principles of a bacterial signalling network. *Nature*, 438: 504–507.

Spudich, J.L. and Koshland, D.E., Jr. (1976). Non-genetic individuality: chance in the single cell. *Nature*, 262: 467–471.

Yi, T.M., Huang, Y., Simon, M.I., and Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc. Natl. Acad. Sci. U.S.A.*, 97: 4649–4653.

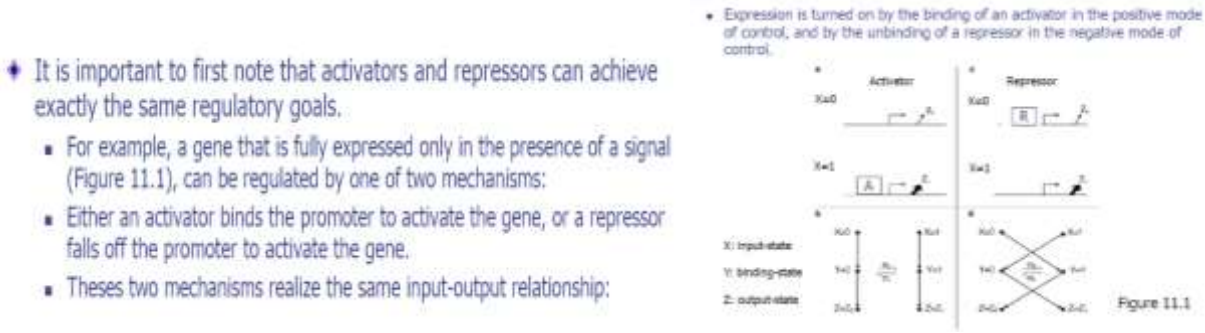
UNIT - V

Introduction to Systems Biology-SBI1401

Introduction:

KINETICS AND GENE REGULATION

- Kinetic proof reading of the genetic code – Recognition of self and non - self by the immune system - Proof reading of diverse recognition processes in the cell
- Demand rules for gene regulation - Savageau demand rule -Rules based on minimal error load- Demand rules for multi regulator systems - Simplicity in biology



Proofreading/editing in protein synthesis is essential for accurate translation of information from the genetic code. Kinetic proofreading is the theory proposed to rationalize the known lack of errors in biological synthesis. In biochemical reactions, enzymes not only enhance the rate of reaction, but also selectively choose the correct substrate leading to the desired product. Many biological processes, like protein synthesis or DNA replication, exhibit high specificity towards the selection of the correct substrates in presence of many other structurally or chemically analogous substrates.

Due to the similar binding energy of both the right and wrong substrates and the size/shape analogue to the enzyme, the error rate (the ratio of the rate of wrong product formation to that of the desired product formation) is expected to be high. To the contrary, the error rate is extremely low in selection of amino acid in protein synthesis (10^{-4}) and DNA replication (10^{-9}).

The molecular reason for such high selectivity is still not fully understood from a quantitative theory. This important problem has remained a debated subject for several decades, with the original Hopfield formulation of repeated activation found inadequate in several biosyntheses.

Recent experimental studies in several enzyme catalytic reactions reveal that the decomposition of the intermediates occurs through hydrolysis reaction. Several alternative editing mechanisms have been proposed and found to be satisfactory in different cases, outlining the fact that more than one mechanism could be operating. One of these mechanisms, proposed first by Fersht, employs hydrolysis of the wrong substrate as the main discriminatory step.

Each amino acid is brought into the ribosome connected to a specific tRNA molecule I The tRNA has a three-letter recognition site is complementary, and pairs with, the codon sequence for that amino acid on the mRNA I There is a tRNA for every codon that maps to an amino acid

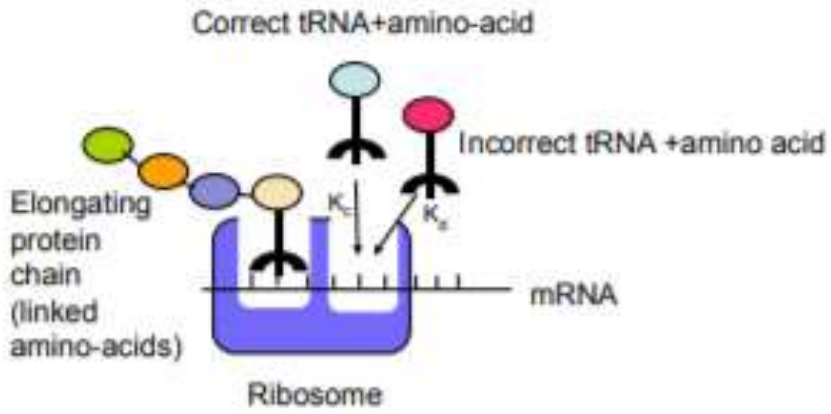


Fig 9.1: Translation of a protein at the ribosome. The mRNA is read by tRNAs that specifically recognize triplets of letters on the mRNA called codons. When a tRNA binds the codon, the amino acid that it carries links to the elongating protein chain. The tRNA is ejected and the next codon is read. Each tRNA competes for binding with the other tRNA types in the cell. The correct tRNA binds with dissociation Constant K_c, whereas the closest incorrect tRNA binds with $K_i > K_c$.

The codon must recognize and bind the correct tRNA I Since this is a molecular process working under thermal noise, it has an error rate I The wrong tRNA can attach to the codon, resulting in a translation error where the wrong amino acid is incorporated I Translation errors occur at a frequency of about 10−4 I A much higher error rate would be disastrous, because it would result in the malfunction of an unacceptable fraction of the cell's proteins

- We first analyze the situation of equilibrium binding of tRNAs to the codons. We will see that equilibrium binding alone cannot explain the observed error rate
- Consider codon C on the mRNA
- Consider the correct tRNA c
- Codon C binds c with an on-rate k_c
- The tRNA unbinds from the codon with off-rate k_c'
- When the tRNA is bound, there is a probability v per unit time that the amino acid attached to the tRNA will be covalently linked to the growing, translated protein chain. The freed tRNA then unbinds form the codon and the ribosome shifts to the next codon in the mRNA

- The process is

$$c + C \underset{k_c'}{\overset{k_c}{\rightleftharpoons}} [cC] \xrightarrow{v} \text{correct amino acid}$$

- The rate v is much smaller than k_c and k_c' and can be neglected for calculation of steady-state
- Hence, at steady-state, we have

$$[cC] = \frac{cC}{K_c},$$

where $K_c = k_c'/k_c$ is the dissociation constant

- The incorporation rate of the correct amino acid is equal to the concentration of the bound complex times the rate at which the amino acid is linked to the elongating protein chain. Hence

$$R_{\text{correct}} = v[cC] = \frac{vcC}{K_c}$$

- We now consider the incorrect tRNA with the highest probability of yielding false recognition by binding codon C and leading to the incorporation of the wrong amino acid
- Let d be this incorrect tRNA
- For d, we have the process

$$d + C \underset{k_d'}{\overset{k_d}{\rightleftharpoons}} [dC] \xrightarrow{v} \text{incorrect amino acid}$$

- At steady-state, we have

$$[dC] = \frac{dC}{K_d},$$

where $K_d = k_d'/k_d$

- The linking rate v is the same for both processes (the correct and incorrect tRNA), so

$$R_{\text{wrong}} = v[dC] = \frac{vdC}{K_d}$$

- Since d is the incorrect tRNA, it has a larger dissociation constant for binding C than the correct tRNA c, i.e. $K_d > K_c$, so

$$R_{\text{wrong}} < R_{\text{correct}}$$

  - The *error rate* F_0 is the ratio of the rates of incorrect and correct amino acid incorporation:

$$F_0 = \frac{R_{\text{wrong}}}{R_{\text{correct}}} = \frac{vdCK_c}{vcCK_d} = \frac{dK_c}{cK_d} \approx \frac{K_c}{K_d},$$

    since the tRNA concentrations for c and d are approximately equal
  - Generally, it is the off-rate k'_d that distinguishes the correct codon from the incorrect one, since the wrong tRNA unbinds more rapidly than the correct one because weaker chemical bonds hold it in the bound complex
  - The on-rates are roughly equal
  - Hence

$$F_0 = \frac{R_{\text{wrong}}}{R_{\text{correct}}} = \frac{K_c}{K_d} = \frac{k'_c}{k'_d}$$

DEMAND RULES FOR GENE REGULATION

- A critical feature of all living organisms is the ability to tune behavior in response to stimuli.1–5 The most widespread and well-understood mode of this tuning is transcription, which enables cells to modulate gene expression in response to cues.
- Looking at the simplest transcription network, where a regulator R, in the presence or absence of a signal, controls the expression of a target T – different possibilities emerge.
- Control of the target might be via positive or negative regulation.
- When we consider the fact that most transcription factors in E. coli are also autoregulators, six possible topologies emerge

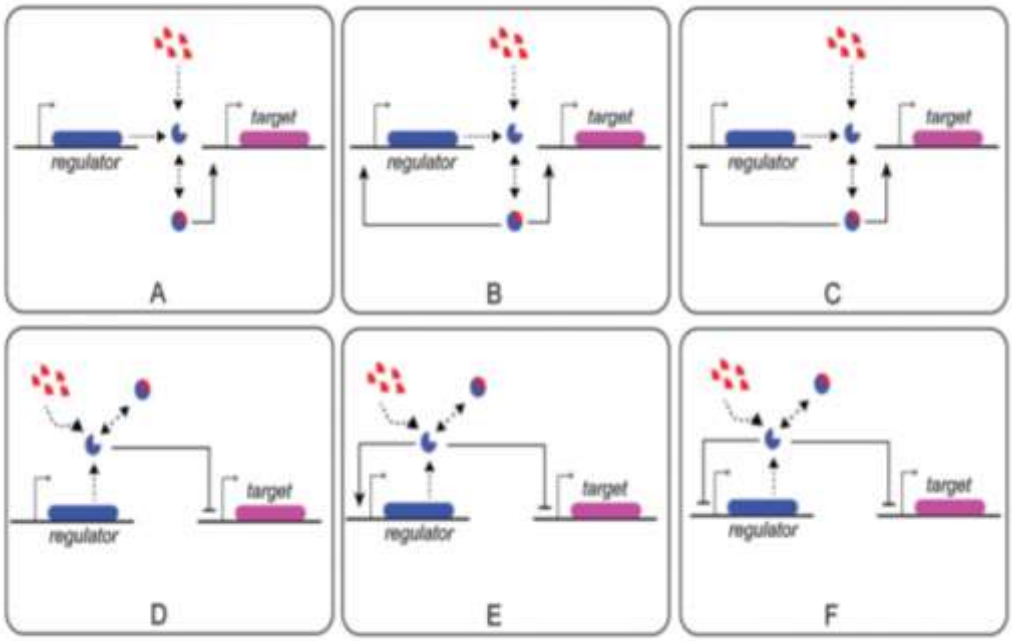


Fig. 1 Showcase of six topologies derived by interactions between a regulator R (blue) and a target, T (pink). In topologies A–C, in the presence of the appropriate environmental or cellular signal, the expression of target T is controlled positively by an induced regulator (R*). In contrast, target expression in topologies D–F is under repression by an un-induced regulator (R), and the repression is relieved under appropriate conditions.

- In a series of papers in the 1970s, Savageau proposed ‘‘demand rules for gene regulation’’, according to which, a target T is positively regulated, if, in the organism’s natural habitat, T is required for a high fraction of time.
- On the other hand, if the target is only required sporadically, it tends to be regulated negatively.
- Evidence for demand rules was provided as conformity in the regulation of sugar utilization enzymes in E. coli with the demand rules.

- The question for gene regulation was raised by M.A. Savageau in his pioneering study of transcriptional control (Savageau 1974, 1977, 1983).
  - Savageau found that the mode of control is correlated with the demand, defined as the fraction of time in the natural environment that the gene product is needed near the high end of its regulatory range.
  - High-demand genes, in which the gene product is required most of the time, tend to have positive (activator) control.
  - Low-demand genes tend to have negative (repressor) control.

- M.A Savageau noted a strong correlation between the mode of bacterial gene regulation and the probability that the gene is fully expressed in the environment.
- To formulate this rule, Savageau defined the **demand** for a gene system as follows:

  "When a system operates close to the high end of its regulatable range most of the time in its natural environment it is said to be a high-demand system. When it operates at the low end of its regulatable range most of the time in its natural environment it is said to be a low-demand system".

- Demand corresponds to the frequency at which the function carried out by the gene system is needed within the ecology of the organism.
  - For example, a system that degrades a certain sugar for use as an energy source is in low demand if the sugar is rare in the environment.
  - The system is in high demand if the sugar is often available.
  - A system that synthesizes an amino-acid is in low demand if that amino-acid is commonly available in significant amounts in the environment.

- Similar results shown in Table 11.2 for a number of biosynthesis systems that produce a compound in the cell.

Biosyntehtic System (induced in ab… of substrate)	Mode of regulation	Regulator	Demand for expression
Arginine	Negative	ArgR	Low
Cysteine	Positive	CysB	High
Isoleucine	Positive	IlvY	High
Leucine	Positive	Lrp, LeuO	High
Lysine	**Positive**	**LysR**	**Low**
Tryptophan	Negative	TrpR	Low
Tyrosine	Negative	TyrR	Low

Relatively abundant in the natural environment

  - The rule also successfully predicts that systems with antagonistic functions, such as biosynthesis and degradation of a compound, tend to have opposite modes of regulation.

- Some exceptions to the rule are also found.
  - One possible example is the biosynthesis system of lysine.
  - The definition of demand is often tentative, because we lack information on the ecology of the cells for many systems.

- To test the demand rule, one needs to have knowledge of the mode of regulation and of the demand for the gene system in question.
  - For this purpose, Savageau collected data on the natural environment of the bacterium *E.coli*.
  - A comparison of the mode of control on **inducible systems** that degrade nutrients is shown in Table 11.1.

Table 11.1 Molecular Mode of Regulation and Demand for Degradation Gene System in the Environment of E.coli

Degradation system (induced in pre… of substrate)	Mode of regulation	Regulator	Demand for expression
Arabinose	positive	AraC	High
Fucose	positive	FucR	High
Galactose	negative	GalR, GalS	Low
Glycerol	negative	GlpR	Low
Lactose	negative	LacI	Low
Lysine	positive	CadC	High
Maltose	positive	MalT	High
Rhamnose	positive	RhaS	High
Xylose	positive	CylR	High
Proline(degradation)	negative	putA	Low

Found at high concentrations

Seldom present at high concentrations

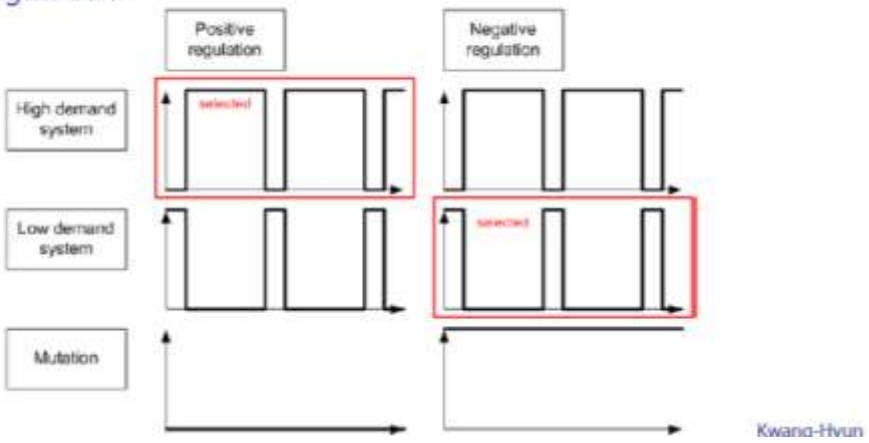
- The demand rule was deduced by Savageau based on the effects that mutations have on the two modes of regulation.
  - This theory first assumes that there are no inherent functional differences between the two modes of regulation.
  - This assumption suggests that one should focus on the behavior of mutations that are altered in the regulatory mechanism.

  - The theory next uses the fact that most mutations in highly evolved structures are detrimental, and very few mutations are beneficial.
  - Consequently, most mutations in a regulatory mechanism lead to loss of regulation.

- The result of these considerations is that the two modes will fare differently in a given environment.
  - The positive mode of regulation is more stable against mutations in a high demand environment, and the negative mode is more stable in a low demand environment.

- Consider a positively regulated gene in a high demand environment.
  - The wild-type organism will induce the gene to high levels most of the time. Mutants who have lost the regulation, will not express the gene. As a result they will be at a disadvantage most of the time, and will be lost from the population.

- The predictions are just the opposite when one considers a negative mode of regulation.



  - The main assumption is that in many regulation systems, **DNA sites that are bound tightly** to their regulatory protein are more **protected from error** than free DNA sites.
  - This leads to the proposal that in order to **minimize errors**, such systems will evolve positive control in high-demand environments and negative control in low-demand environment.
  - There are at least two sources of errors connected with the free site.
    - The first source of errors is **cross-talk** with the other transcription regulators in the cell, in which the wrong transcription factor binds to the site.
    - A second source of error arises from residual binding of the designated regulator in its **inactive form** to its own site.

- Consider a gene regulated by an activator, and the same gene regulated by a repressor, such that the two regulatory mechanisms lead to the same input-output relationship.
- The average reduction in fitness for a repressor, taking into account only errors from the free site (corresponds to high expression),

$$E_R = p\Delta f_1$$

P : demand (the fraction of time that the gene product is needed)
$\mathbf{f_1}$: relative fitness reduction by errors in high expression state
('1' denotes the high expression state.)

- The average reduction in fitness for an activator, taking into account only errors from the free site (corresponds to low expression),

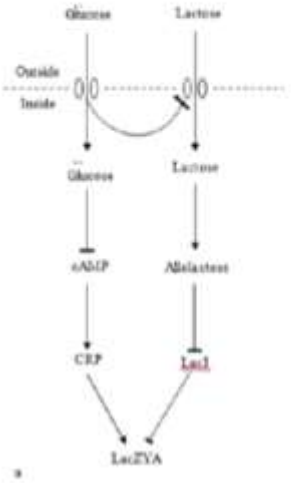
$$E_A = (1-p)\Delta f_0$$

The fraction of time that the gene is not in demand
$\mathbf{f_0}$: relative fitness reduction by errors in low expression
('0' denotes the low expression state.)

- In this simplest case, a repressor will have a fitness advantage over an otherwise equivalent activator when it has a lower error-load:

$$E_R < E_A$$

- Let us now turn to systems with multiple regulators. We will consider in detail the *lac* system of *E.coli*.
  - *Lac* system has two input stimuli, lactose and glucose.
  - This input-output relationship is implemented by two regulators, the repressor LacI , and the activator CRP (starvation sensor of cAMP).

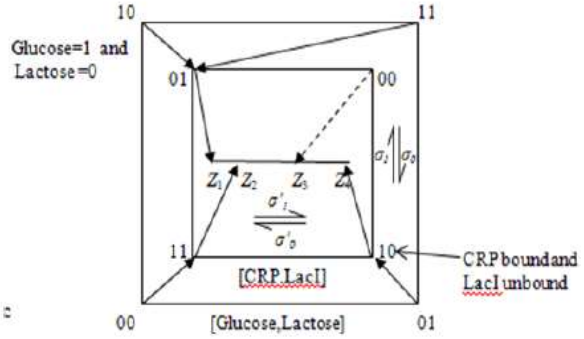


- The *lac* system has an additional mechanism that inhibits expression in the presence of glucose, which is called **inducer-exclusion:** When glucose is pumped into the cell, lactose entry is blocked, preventing the induction of the *lac* system.
(since glucose is always a better source than lactose).

Input State		Internal State		Output state
Glucose	Lactose	CRP	LacI	LacZYA
1	0	0	1	$Z_1 \approx 3 \cdot 10^{-4}$
1	1	0	1(0)	$Z_1 \approx 3 \cdot 10^{-4}$ ($Z_2 \approx 0.13$)
0	0	1	1	$Z_2 \approx 6 \cdot 10^{-4}$
0	1	1	0	$Z_2 \approx 1$

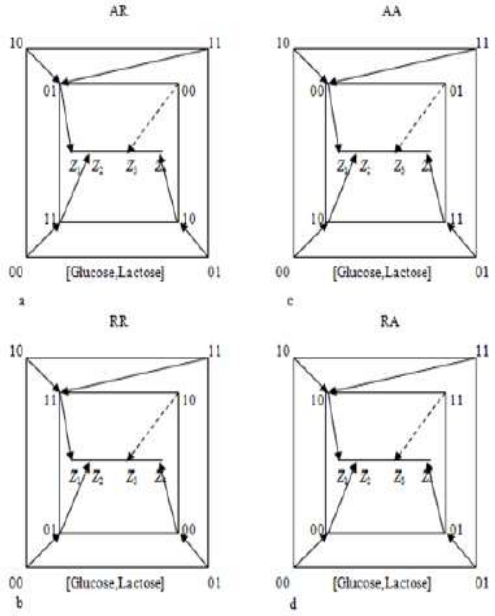
- The relation between the input-states, the DNA binding-states and the output-levels of the *lac* system.

- There are four possible binding-states, depending on whether the CRP and LacI sites are bound or free.



- These binding states are denoted [CRP, LacI]=[0,0], [0,1], [1,0], [1,1]
- The binding-state [0,0] does not correspond to any input-state, **excluded state.**

- The naturally occurring mechanism, with a glucose-responsive activator and a lactose responsive repressor, is only one of the four possible mechanisms in which the two regulators can have either mode of control.



- The four mechanisms can be denoted RR, RA, AR and AA where the first letter denotes glucose regulator and lactose regulator.
- All four mechanisms have inducer exclusion, and thus have an excluded state: the excluded state is [CRP, LacI]= [0,0], [0,1], [1,0], [1,1] in the AR, AA, RR, RA mechanisms respectively.

- Why the AR mechanism minimizes the error-load in the natural environment of *E. coli*?
  - The most frequent input-state, (glucose, lactose)=(0,0) corresponds to the binding-state [CRP,LacI]=[1,1], where both regulators bind their DNA sites. Thus, the AR mechanism keeps the DNA sites protected from errors most of the time.
  - In addition, the AR mechanism has another error-minimizing feature: the most error-prone binding-state [CRP, LacI]=[0,0], in which both sites are free, is concealed by inducer exclusion.
  - Hence, not only does the AR mechanism map the most frequent input-state onto the error-free binding-state [1,1], but also it excludes the most error-prone binding-state [0,0] and prevents it from ever being reached by any input-state.
  - This is in contrast to the three other possible mechanisms that make the error-prone binding-state [0,0] accessible to environmental conditions.