Physics of System Biology

Course Instructor: Professor Jung Y. Huang

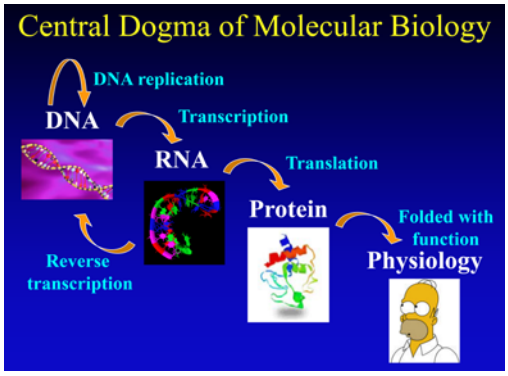
jyhuang@faculty.nctu.edu.tw

Course Website: http://www.jyhuang.idv.tw/Physics of System Biology.aspx

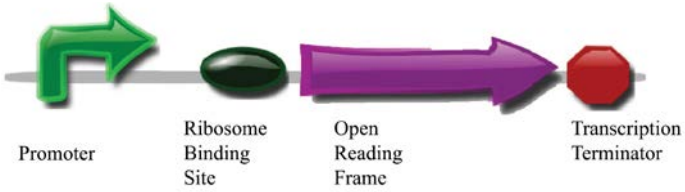
Chapter 1 Fundamental of Physics of System Biology

1.1 Overview

Two basic concepts of molecular biology have to be understood before endeavoring to engineer biological systems, which are how information flows in biological systems and how this information flow is controlled/regulated. With an understanding of these concepts we shall be able to apply engineering principles to the design and building of new biological systems.



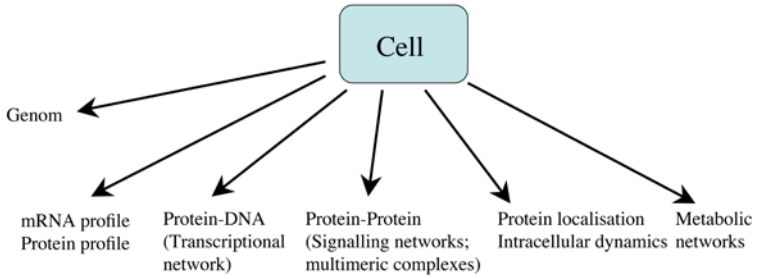
Genome is the information store, but the majority of functions within the cell are performed by proteins. The question thus arises: *how information leads to function*. The flow of information within biological systems forms what is known as the *central dogma*. When required, the message within DNA is transcribed into an intermediate molecule called messenger RNA (mRNA), before being translated into the final product, protein. The fundamental unit of hereditary information stored in the genome is known as a gene. A gene should also be considered to include the regulatory elements required to control it. The basic structure of a gene is depicted in the following:



When a gene is turned on, the protein (or other functional molecule) is produced (or called being expressed); when it is turned off it is not produced.

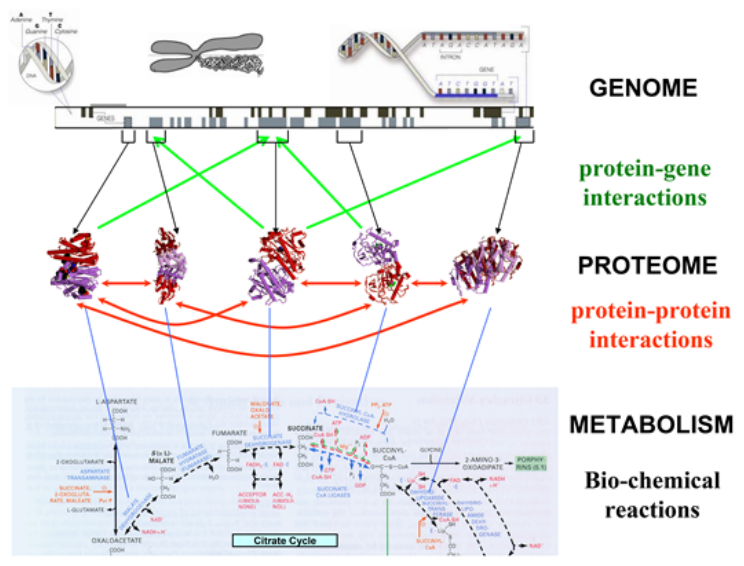
The availability of a fully sequenced human genome calls for more accurate understanding of the *fundamental properties of complex systems* residing inside the cell and elucidating the origins of complex human diseases.

The complex biological systems residing inside the cell



Cellular processes (such as signaling, gene regulation, and protein-protein interactions) are intertwined on many levels

Human genes and the proteins they encode in a biological system do not work in isolation but are connected at various levels, giving rise to a plethora of networks.

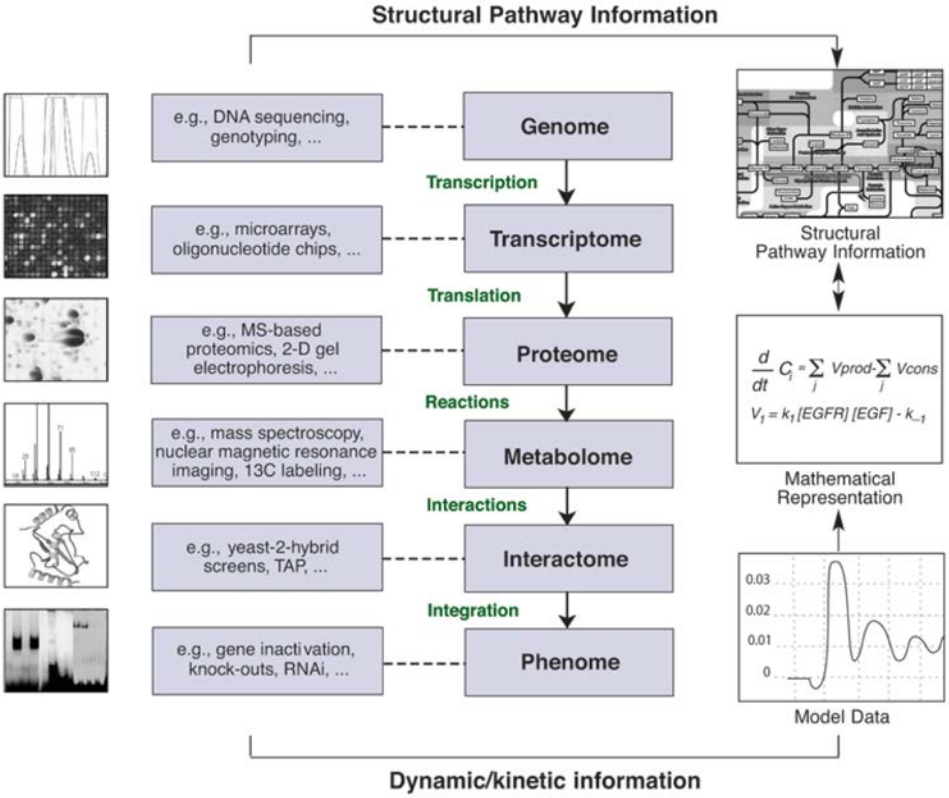


1.2 High-throughput Technologies

High-throughput technologies have recently led to a new perspective in biology, where the cell is interpreted as a large and complex system composed of highly integrated subsystems. The high-throughput methodologies, paired with computational approaches, can be used to infer *networks of interactions* and *causal relationships* within the cell.

The Omics technologies are the driving force behind the System Biology

With the development of the high-throughput technologies to identify and quantify (DNA, mRNA, protein, and metabolite at the molecular level, researchers are in a position to gather comprehensive genome-wide data of a given biological system. In addition, new techniques to manipulate cells in a directed manner allow researchers to perturb biological systems under controlled conditions (*e.g.*, single genes can be deactivated and the global response of the modified cell can be observed at the molecular level). Such comprehensive and accurate experimental data are critical for developing and testing models of biological processes.



The current challenge is to integrate the different information yielded from a variety of high-throughput methodologies to understand dynamical properties of cellular systems.

1.2.1 **Transcriptomic Data:** studies the *active genes* in a given cell at a given time

A. Gene Expression (GE) Profile

Analysis of gene expression (GE) can be done by several different methods including (reverse transcriptome PCR) RT-PCR, RNase protection assays, microarrays, as well as northern blotting (see **the knowledge base 1** and 2). Microarrays are commonly used, which yield consistent data with that obtained with northern blots, and can visualize thousands of genes at a time.

High Throughput Measures of Gene Expression can be used to

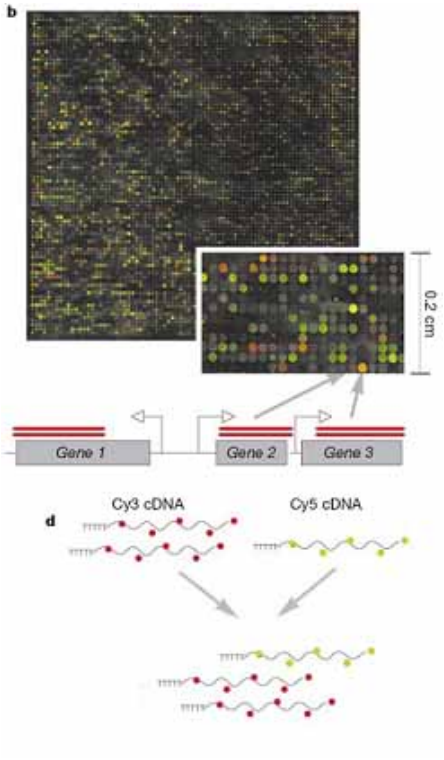
- Measure gene expression: quasi-estimate of the protein level and cell state
- High throughput: measure mRNA level of all active genes in the genome
- Checking the physiological status in many different situations

Typical Steps for the Use of Microarrays

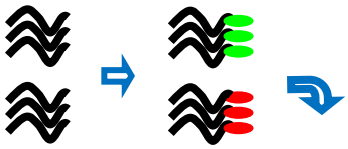
- Grow cells at certain condition, collect mRNA population, and label them;
- Microarray has high density sequence specific probes with known location for each gene/RNA;
- Sample hybridized to microarray probes by DNA base pairing (A-T, G-C), wash non-specific binding;
- Measure sample mRNA value by checking labeled signals at each probe location.

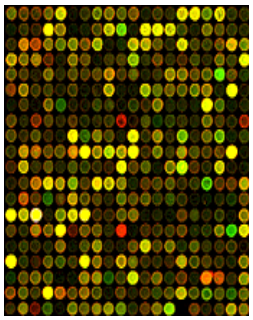
There are two different types of microarrays: i. Spotted cDNA microarrays
ii. Oligonucleotide arrays

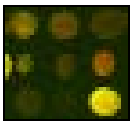
Spotted cDNA Arrays



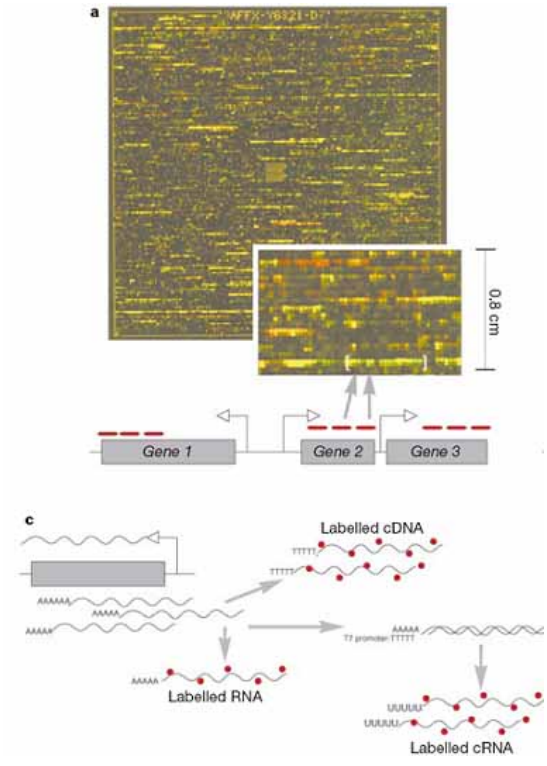
- Developed by Pat Brown Lab, Stanford University;
- Robotic spotting of cDNA (mRNA converted back to DNA without introns) ; typically containing several thousands of probes per array;
- One long probe per gene
- Competing hybridization: Control
  Treatment
- Detection: Green: high control
  Red: high treatment
  Yellow: equally high
  Black: equally low





Why Competing Hybridization? DNA concentration in probes is not the same because probes may not be spotted evenly. 

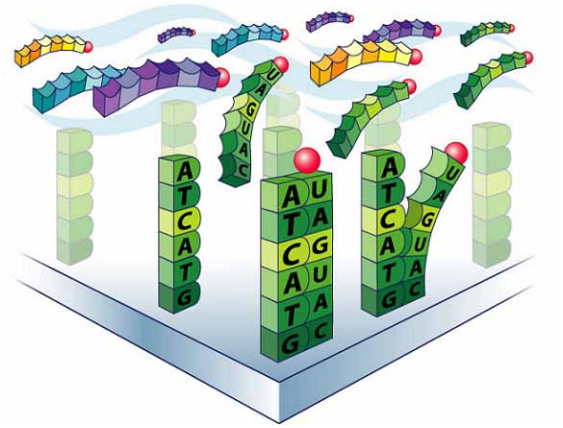
Oligonucleotide Arrays



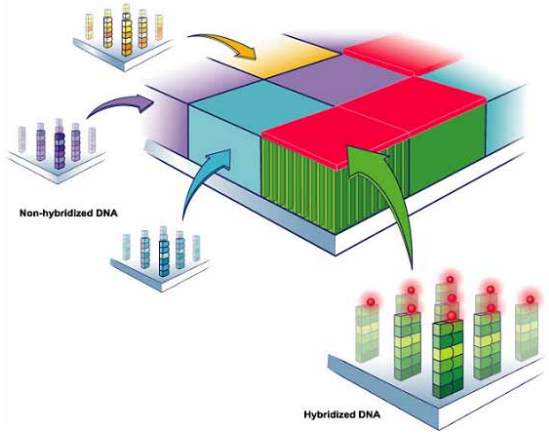
- GeneChip® provided by by Affymetrix;
- Parallel synthesis of oligonucleotide probes (25-mer) on a slide using photolithographic methods; typically containing millions of probes per microarray;
- Multiple probes per gene;
- One-color arrays;

Typical Steps for the Use of Oligonucleotide Arrays

Labeled Samples Hybridize to DNA Probes on GeneChip



Shining Laser Light Causes Tagged Fragments to Glow



B. **Next Generation Sequencing Technologies** There are several next generation sequencing technologies under development in order to bring down the sequencing cost to $1000 USD. We only introduce the principles of the most representative ones to give a brief overview of the field.

454 Life Sciences: Small fragments of DNA are mixed with small beads. The mixture is sufficiently dilute so that *each bead bind a single DNA molecule*. Next, the DNA-containing beads are dispersed on a silicon plate containing 400,000 regularly spaced wells. Each well captures a single bead. PCR is performed directly on the bead-tethered DNA to amplify each molecule. This is then used as a substrate for a second round of PCR that includes bioluminescent proteins as well as DNA polymerase.

The second round of DNA synthesis is performed *separately* with dATP, dGTP, dTTP, or dCTP, with a washing cycle between each pulse. The addition of one of the four deoxynucleotide triphosphates (dNTPs) (in the case of dATP we add dATPαS which is not a substrate for a luciferase) initiates the second step. DNA polymerase incorporates the correct, complementary dNTPs onto the template. This incorporation releases pyrophosphate (PPi) stoichiometrically. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5’ phosphosulfate. This ATP acts as fuel to the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in

amounts that are proportional to the amount of ATP. The light pulse represent the incorporation of a particular nucleotide. Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide. Thus in this way 400,000 DNAs are simultaneously sequenced (200-400 nucleotides per segment) to generate 100 Mb sequence per run.

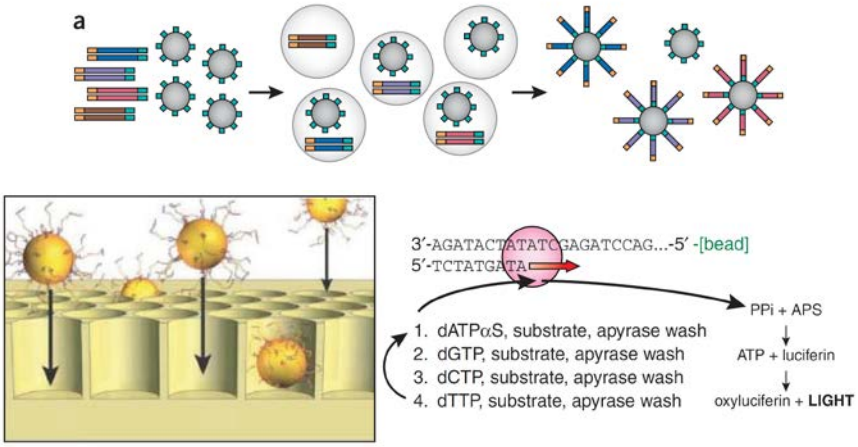
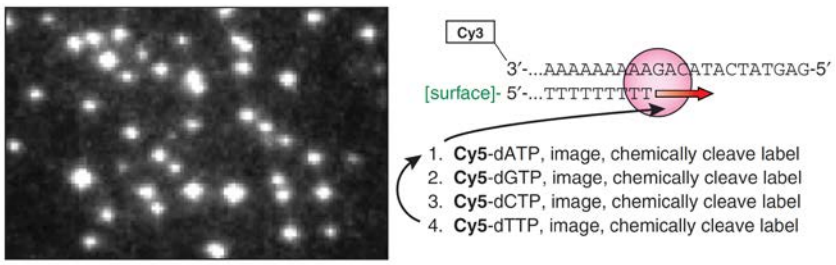


Figure Caption: (a) Clonally amplified 28-µm beads generated by emulsion PCR serve as sequencing features and are randomly deposited to a microfabricated array of picoliter-scale wells.

(b) With pyrosequencing, each cycle consists of the introduction of a single nucleotide species, followed by addition of substrate (luciferin, adenosine 5'-phosphosulphate dATPαS) to drive light production at wells where polymerase-driven incorporation of that nucleotide took place. This is followed by an apyrase wash to remove unincorporated nucleotide.

HeliScope platform: Single nucleic acid molecules are sequenced directly, that is, there is no clonal amplification step required. Poly-A–tailed template molecules are captured by hybridization to surface-tethered poly-T oligomers to yield a disordered array of primed single-molecule sequencing templates. Templates are labeled with Cy3, such that imaging can identify the subset of array coordinates where a sequencing read is expected.

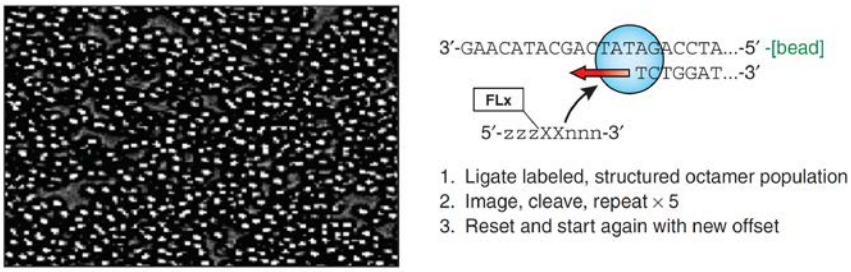
Each cycle consists of the polymerase-driven incorporation of a single species of fluorescently labeled nucleotide at a subset of templates, followed by fluorescence imaging of the full array and chemical cleavage of the label.



SOLiD Platform (Sequencing by Oligonucleotide Ligation and Detection, http://en.wikipedia.org/wiki/2_Base_Encoding)**:** Clonally amplified 1-µm beads are used to generate a disordered, dense array of sequencing features. Sequencing is performed with a *ligase*, rather than a polymerase. Each sequencing cycle introduces a partially degenerate population of fluorescently labeled *octamers* (8 bases in length). The population is structured such that *the label correlates with*

the identity of the central 2 bp in the octamer (the correlation with 2 bp, rather than 1 bp, is the basis of two-base encoding). These 8-base long probes have a free hydroxyl group at the 3' end, a fluorescent dye at the 5' end and a cleavage site between the fifth and sixth nucleotide. The first three bases *nnn* (starting at the 3' end) are complementary to the nucleotides being sequenced. Bases 4 through 5 are degenerate (*XX*) and able to pair with any nucleotides on the template sequence. Bases 6-8 (*zzz*) are also degenerate but are cleaved off, along with the fluorescent dye, as the reaction continues. In this manner positions *n*+1 and *n*+2 are correctly base-paired followed by *n*+6 and *n*+7 being correctly paired, *etc*. The composition of bases *n*+3,*n*+4 and *n*+5 remains undetermined until further rounds of the sequencing reaction.

After ligation and imaging in four channels, the labeled portion of the octamer is cleaved via a modified linkage between bases 5 and 6, leaving a free 5' end for another cycle of ligation. Several such cycles will iteratively interrogate an evenly spaced, discontiguous set of bases. The system is then reset (by denaturation of the extended primer), and the process is repeated with a different offset (e.g., a primer set back from the original position by one or several bases) such that a different set of discontiguous bases is interrogated on the next round of serial ligations. The sequencing step is basically composed of five rounds and each round consists of about 5-7 cycles. A complete reaction of five rounds allows the sequencing of about 25 base pairs of the template.



RNAseq: a revolutionary tool for transcriptomics

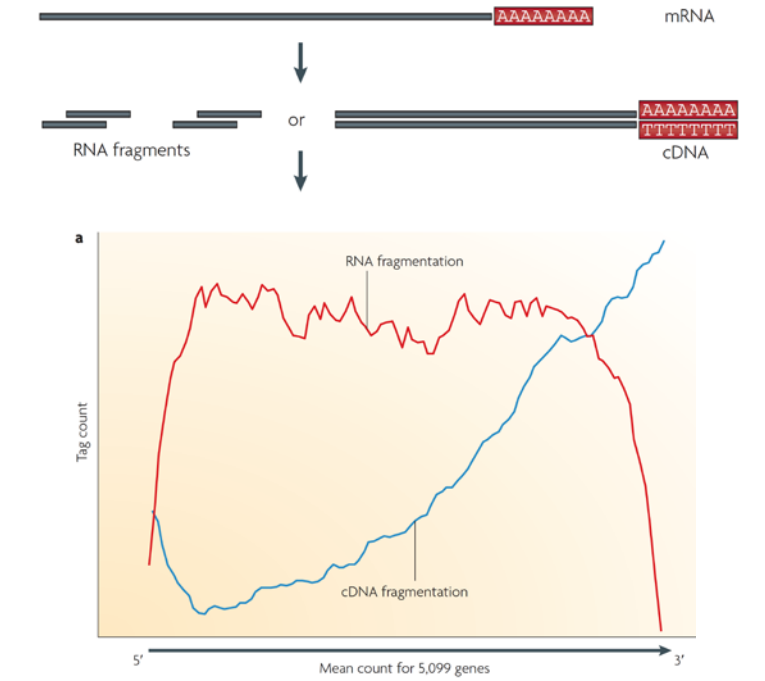
The recent development of next-generation massively parallel sequencing (MPS) technologies by companies such as Roche (454 GS FLX), Illumina (Genome Analyzer II), and ABI (AB SOLiD) has completely transformed the way in which quantitative transcriptomics can be done. These new technologies have reduced both the cost per reaction and time required by orders of magnitude, making the use of sequencing a cost-effective option for many experimental approaches. One such method that has recently been developed uses MPS technology to directly survey the RNA content of cells, without requiring any of the traditional cloning. This approach, called "RNA-seq", can generate quantitative expression scores that are comparable to microarrays, with the added benefit that the entire transcriptome is surveyed without the requirement of a priori knowledge of transcribed regions. The important advantage of this technique is that not only can quantitative expression measures be made, but transcript structures including alternatively spliced transcript isoforms, can also be identified.

RNA-Seq: Alternative to Microarrays with Unique Features

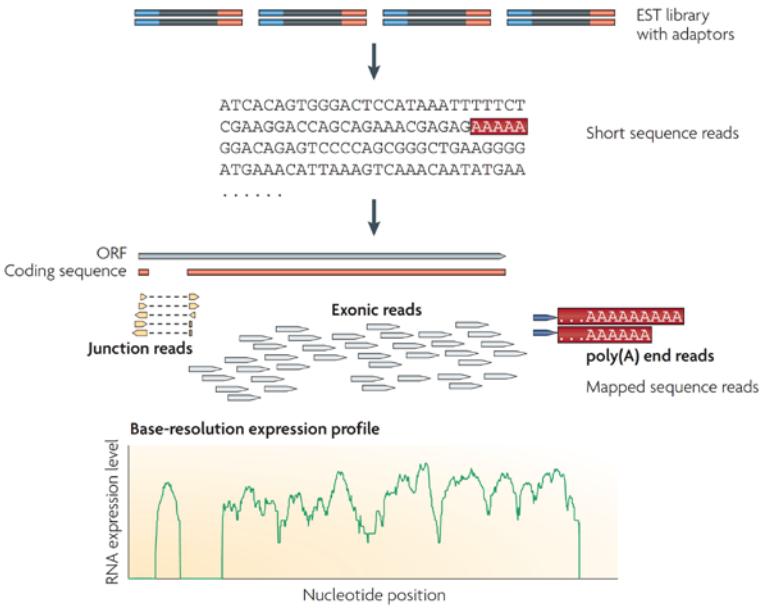
- Yielding general expression profiling

- Applicable for novel genes
- Providing information about alternative splicing
- Can detect gene fusion
- Can be used on any sequenced genome
- Better dynamic range
- Cleaner and more informative data
- But data analysis is challenging

Working Principle



(a)



(b)

Figure Caption: (a) Long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation

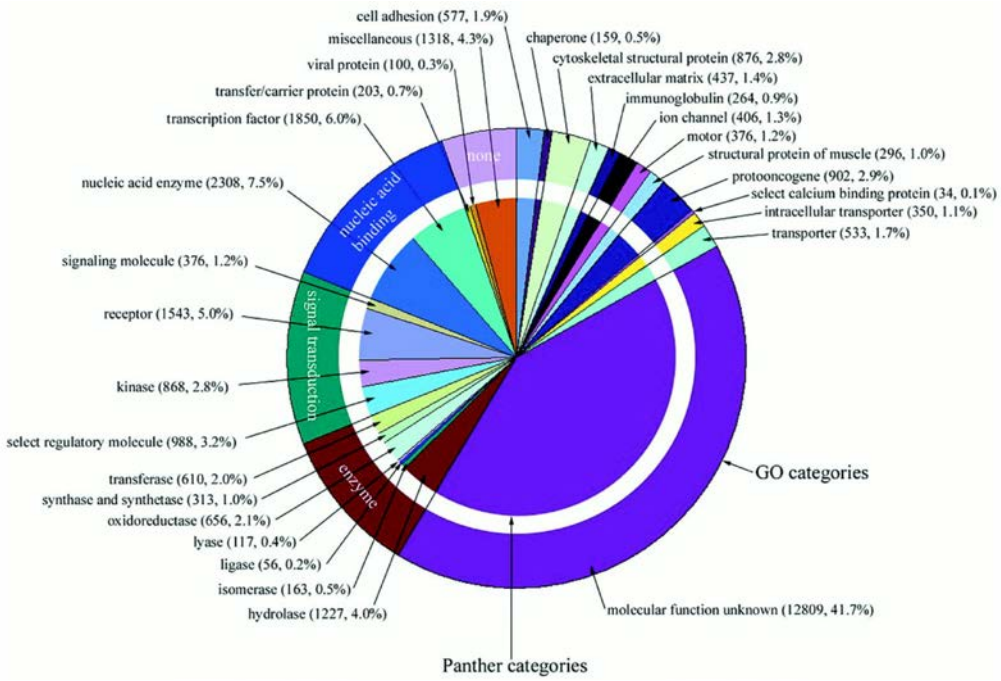
or DNA fragmentation. Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends.

(b) Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom (a yeast ORF with one intron is shown).

1.2.2 **Proteomic Studies:** discover which proteins are present and in what amounts

Challenges in Proteomics

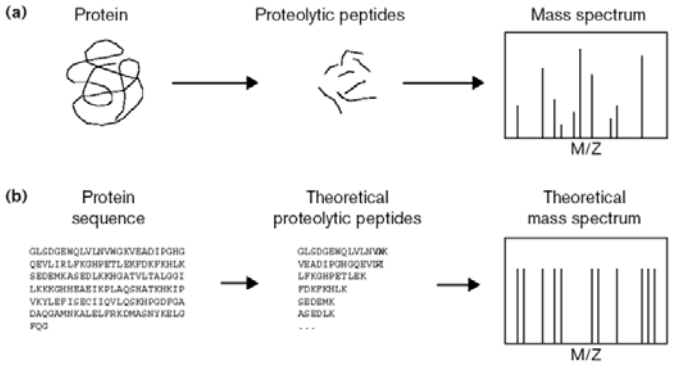
- Too many proteins:
  - ➔ 20 building blocks instead of 4;
  - ➔ 65% genes have splice and translation isoforms;
  - ➔ Many different post-translation modifications (PTM);
  - ➔ # Proteins > # mRNA > # genes
- Difficult to capture and study activities
  - ➔ Protein state changes are often faster than transcription regulation;
  - ➔ Need to be in functionally folded state, with right PTM;
  - ➔ Many are difficult to purify and study in vitro, *e.g.* membrane proteins
- Big dynamic range
  - ➔ Yeast 10^6, human 10^9, current method 10^2-10^4
  - ➔ No PCR for amplifying proteins, hard to profile low-abundance proteins
- Why bother then?
  - ➔ A lot of practical applications
  - ➔ Most drugs act on proteins not DNA/RNA
  - ➔ Enzyme / TFs are the master regulators
  - ➔ DNA ~ "the novel", mRNA ~ "the script", proteins ~ "the actors"



High-throughput methods for measuring protein-protein interactions

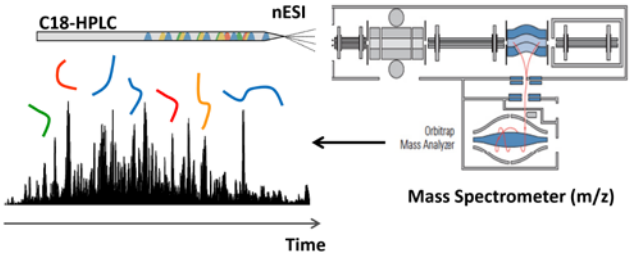
High-throughput methods for detecting PPIs at large-scale that have been introduced include yeast two-hybrid (Y2H) screening, affinity purification–mass spectrometry (AP–MS), and protein microarrays.

Mass Spectrometry: Identify proteins by comparing observed MS spectrum to computed spectrums from genome



Mass spectroscopies do not resolve too long proteins; so we need to cut protein to shorter peptides (~15 aa) by using trypsin, which cleaves at Arginine (R) and Lysine (K) but not before Proline (P).

Apparatus and Working Principle

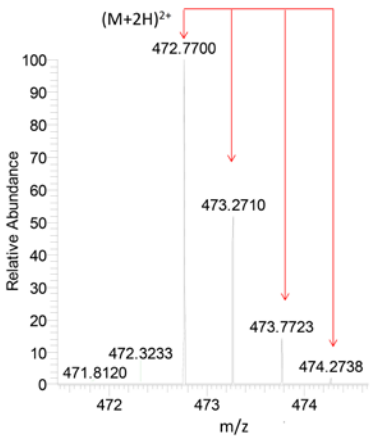


Tandem high performance liquid chromatography-mass spectrometry (HPLC-MS) is an analytical chemistry technique that combines the physical separation capabilities of HPLC with the mass analysis capabilities of mass spectrometry. HPLC-MS is a powerful technique used for many applications which has very high sensitivity and specificity. Generally its application is oriented towards the specific detection and potential identification of chemicals in the presence of other chemicals (in a complex mixture). HPLC separates complex protein mixtures into less complex sub-fractions based on ionic interactions with a charged column (ion exchange chromatography) and hydrophobic interactions with the column (reverse phase chromatography). An electrospray is used that employs high voltage to disperse a liquid through a fine glass or metal capillary resulting in a highly charged aerosol. Mass spectrometry (MS) is an analytical technique for the determination of the elemental composition of a sample molecule. It is also used for elucidating the chemical structures of molecules, such as peptides and other chemical compounds. The MS principle consists of ionizing chemical compounds to generate charged molecules or molecule fragments and measurement of their mass-to-charge ratios.

In a typical MS procedure:

1) A sample is loaded onto the MS instrument, and undergoes vaporization.
2) The components of the sample are ionized to form charged particles.
3) The positive ions are then accelerated by an electric field.
4) Computation of the mass-to-charge ratio (m/z) of the particles based on the details of motion of the ions as they transit through electromagnetic fields.
5) Detection of the ions, which in step 4 were sorted according to m/z.

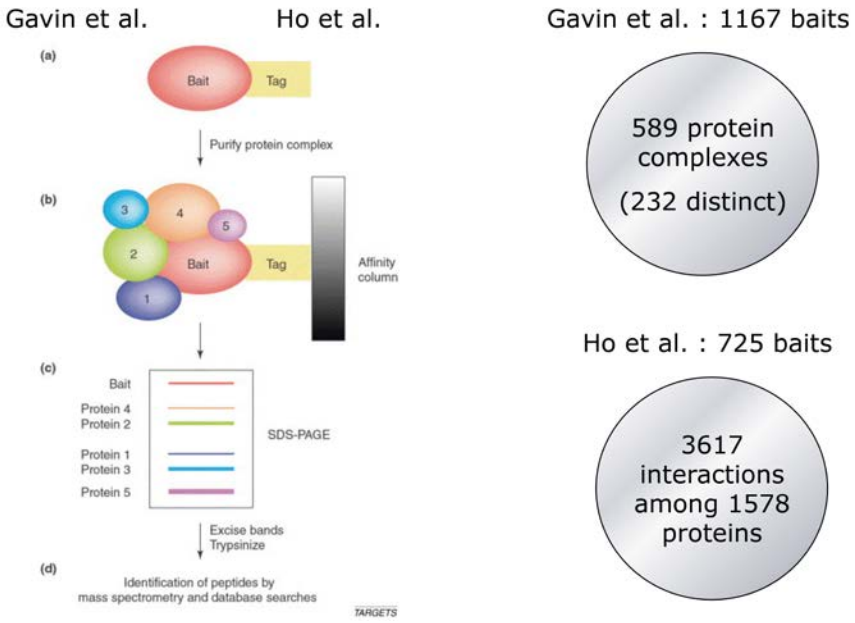
Peptide Charge and Mass



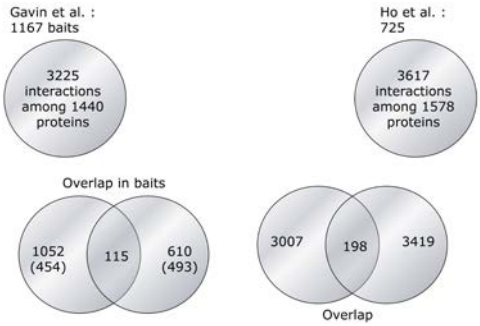
From the charge z=2 and $m = M + 2H^+$, the measured peak with $m/z = 472.7700 = (M + 2H^+)/2$

and $H^+ = 1.0073$, the peptide mass is then estimated to be $M = 943.5254$.

Two large-scale MS experiments

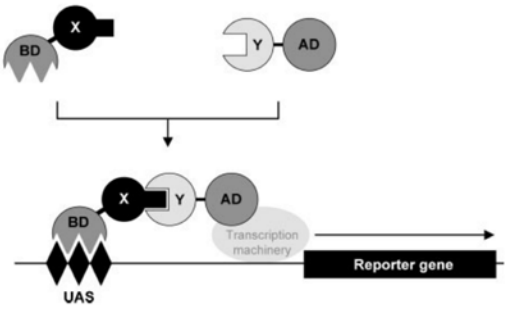


Reproducibility in AP/MS



Yeast Two-Hybrid (Y2H)

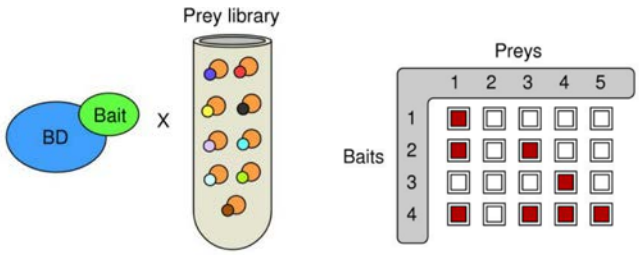
Two-hybrid screening (also known as yeast two-hybrid system or Y2H) is a molecular biology technique used to discover protein-protein interactions. The premise behind the test is the activation of downstream reporter gene(s) by the binding of a transcription factor onto an upstream activating sequence (UAS).



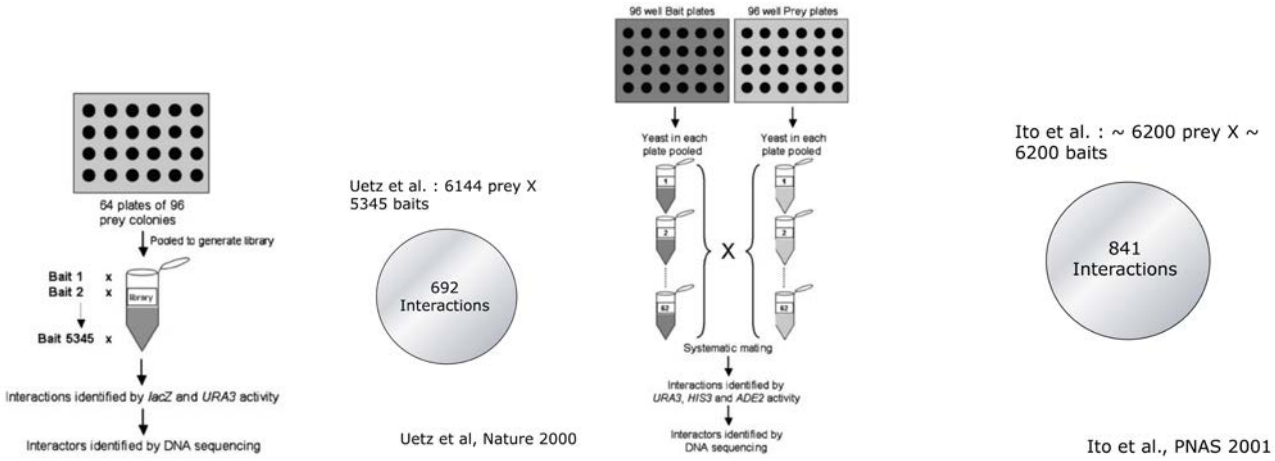
For two-hybrid screening, the transcription factor is split into two separate fragments, called the binding domain (BD) and activating domain (AD). The BD is the domain responsible for binding to the UAS and the AD is the domain responsible for the activation of transcription. The most common

screening approach is the yeast two-hybrid assay. This system often utilizes a genetically engineered strain of yeast in which the biosynthesis of certain nutrients is lacking. When grown on media that lack these nutrients, the yeast fail to survive. This mutant yeast strain can be made to incorporate foreign DNA in the form of plasmids. In yeast two-hybrid screening, separate bait and prey plasmids are simultaneously introduced into the mutant yeast strain. Plasmids are engineered to produce a protein product in which the DNA-binding domain (BD) fragment is fused onto a protein while another plasmid is engineered to produce a protein product in which the activation domain (AD) fragment is fused onto another protein. The protein fused to the BD may be referred to as the bait protein, and is typically a known protein the investigator is using to identify new binding partners. The protein fused to the AD may be referred to as the prey protein and can be either a single known protein or a library of known or unknown proteins.

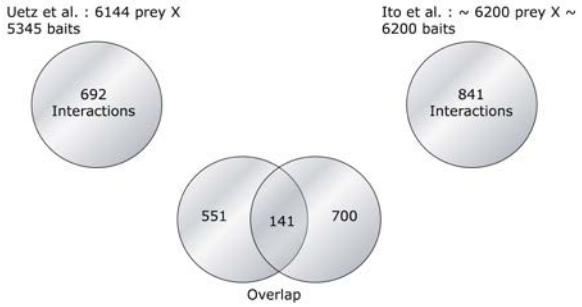
A library may consist of a collection of protein-encoding sequences that represent all the proteins expressed in a particular organism or tissue. If the bait and prey proteins interact, then the AD and BD of the transcription factor are indirectly connected, bringing the AD in proximity to the transcription start site and transcription of reporter gene(s) can occur. If the two proteins do not interact, there is no transcription of the reporter gene. In this way, a successful interaction between the fused proteins is linked to a change in the cell phenotype.



Two large-scale Y2H studies: Uetz *et al*. and Ito *et al*.

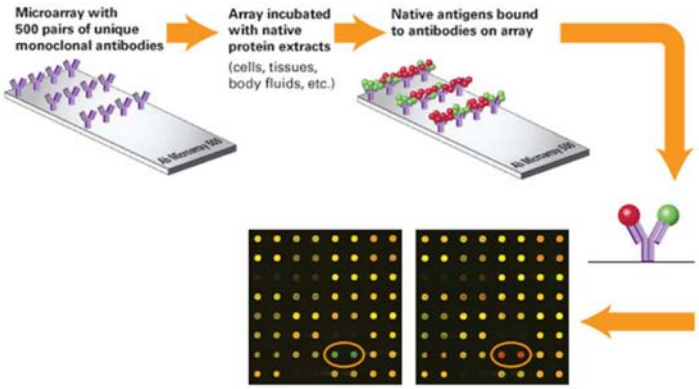


Reproducibility in Y2H



Antibody Microarrays

An antibody microarray is a specific form of protein microarrays, a collection of capture antibodies are spotted and fixed on a solid surface, such as glass, plastic and silicon chip for the purpose of detecting antigens. Antibody microarray is often used for detecting protein expressions from cell lysates in general research and special biomarkers from serum or urine for diagnostic applications. The covalent labeling of all proteins in a complex mixture provides a means for detecting bound proteins after incubation on an antibody microarray. If proteins are labeled with a tag, such as biotin, the signal from bound proteins can be amplified.



Limitations: Not so widely used, Need for reliable antibodies, usually relatively few proteins (100s) are measured, requires a lot of protein extract (no PCR amplification!)

Could also consider global proteomics using mass spectrometry but this still requires optimization to be broadly useful and is very expensive

1.2.3 Metabolome: examine which metabolic processes occur under different conditions

Biological systems, which exploit suitable energy sources, can achieve spontaneous self-organization (order) and allow them to reach high levels of diversity and complexity by means of adaptive processes. From the thermodynamic point of view, the actual decrease of entropy of the system, relative to its organization, is balanced by the entropy increase of the surrounding environment.

Metabolism is a set of biochemical reactions by which the cells can extract energy and materials from its environment, and uses them to produce different metabolites necessary for its survival and function. Metabolomics is the scientific study of chemical processes involving metabolites. The metabolome represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes. Thus, metabolic profiling can give an instantaneous snapshot of the physiology of that cell.

Comparison of High Throughput Analytical Platforms

	Detection limit	Advantages	Disadvantages
NMR	> 10^{-6} M	Min sample prep Structural info	Sensitivity Spectral interpretation
Mass Spectrometry	**> 10^{-18} M**	**Sensitivity Molecular diversity**	**Indirect structural info Sample prep**
Capillary electrophoresis	> 10^{-15} M	Sensitivity	Indirect structural info Molecular diversity Sample prep

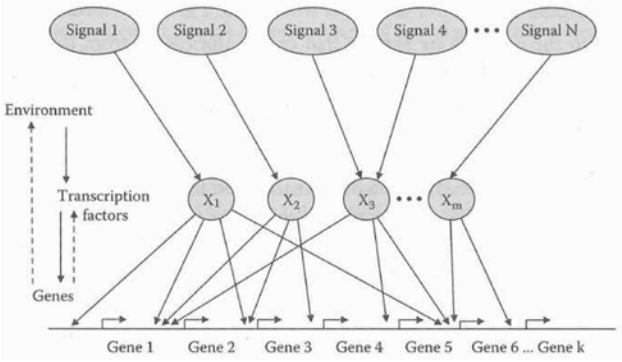
1.3 Biological Networks

Human genes and the proteins they encode in a biological system do not work in isolation but are connected at various levels, giving rise to a plethora of networks. Depending on the kind of biological network, the edges and nodes of the graph have different meaning.

1.3.1 Transcriptional Regulatory Networks (protein-DNA interactions)

The regulation of gene expression ensures the correct expression of specific gene at the right time and at the right location. The regulation of gene expression gives a cell control over its structure and function. In eukaryotes, the transcription of each gene is regulated by transcription factors (TFs). Transcriptional regulation is the process by which genes regulate the transcription of other genes. A gene X directly regulates a gene Y, if the protein that is encoded by X is a transcriptional factor for gene Y.

In a transcriptional regulation network (TRN), nodes represent genes and directed edges between them represent interactions through which the products of one gene affect those of another. Thus, the network structure is an abstraction of the system's biochemical dynamics that is responsible for regulating the expression of genes in the cell.

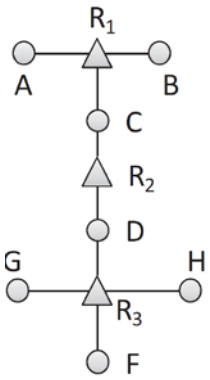


The mapping between environmental signals $\{S_i\}$, transcription factors $\{X_i\}$ inside the cell, and the genes that they regulate. The transcription factors, when active, bind DNA to change the transcription rate of specific target genes.

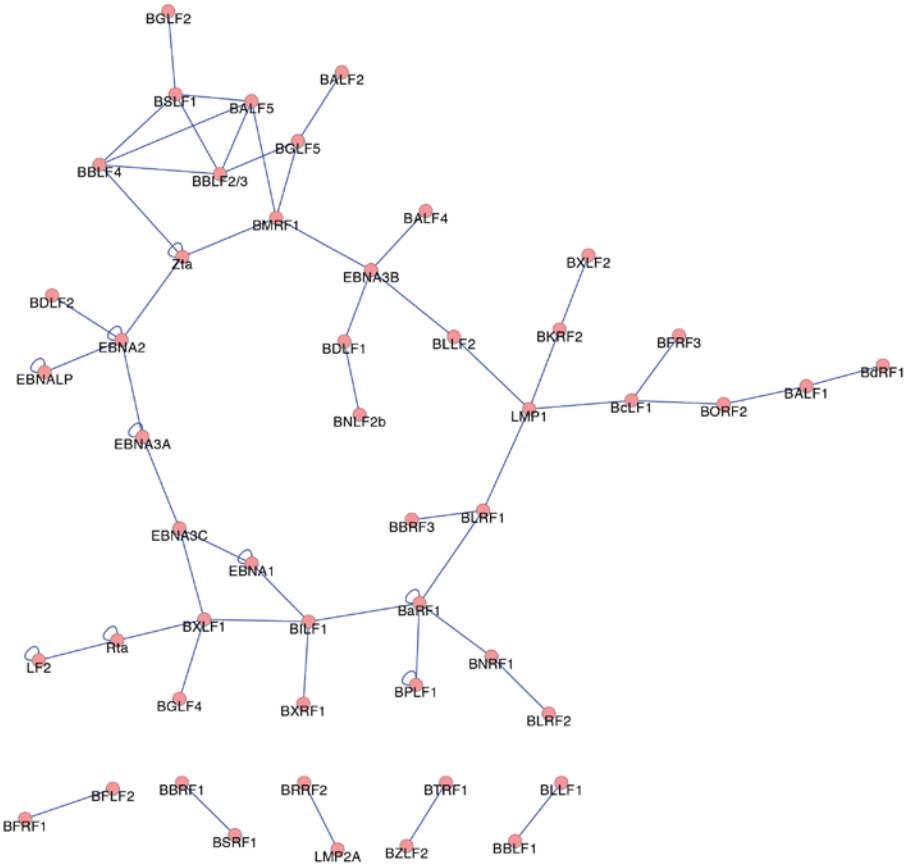
The genome of the unicellular yeast encodes 200 predicted TFs, which is 3% of all protein-coding genes; the relatively simple metazoan nematode C. elegans contains 934 predicted TFs (5% of all protein-coding genes); humans may devote up to 10% of their coding potential to

regulatory TFs. TRN identification today is mainly accomplished with the combination of microarray techniques with chromatin immune precipitation (**ChIP**, see the knowledge base file).

1.3.2 Protein-Protein Interaction (PPI) Networks



A vast number of biological processes in an organism are dependent on precise physical interactions among many individual proteins. PPI networks reflect this interconnected nature of biological processes. In PPI networks, nodes represent proteins and undirected edges between them represent physical interactions.



A PPI network showing the Epstein-Barr Virus (EBV) pathogenesis

1.3.3 Metabolic Networks (MN)

A metabolic network is an abstraction to represent cellular metabolism. Feist *et al*. proposed a general four-step process for a metabolic network reconstruction. The first step consists of generating a network draft from the genome annotation of the organism of interest. The draft network should be manually curated in the second step to remove incorrectly included reactions, add missing reactions for filling gaps, or assign enzymes localization to different cellular compartments. In the third step, the curated network should be translated into a mathematical model that facilitates the analysis of the network and its validation. The final step of metabolic network reconstruction process is to deploy the metabolic network model to produce new insights about the organism's metabolism.

For graph-theoretic analysis, metabolic networks are usually represented by a bipartite graph with two sets of nodes: one corresponding to metabolites and the other corresponding to reactions (or

enzymes). Edges connect reactions (or enzymes) with their participating metabolites.

$$\text{Reaction 1:} \quad A + B \xrightarrow[E_2]{E_1} C$$

$$\text{Reaction 2:} \quad C \xrightarrow{E_3} D$$

$$\text{Reaction 3:} \quad D + F \xrightarrow{E_4} G + H$$

1.3.4 Signaling Networks (SN)

Signaling networks are used by cells to constantly monitor a wide array of external and internal stimuli. External stimuli can include nutrient levels, growth factors, hormones, and so on; whereas DNA damage, protein misfolding, and so on, can be considered as internal stimuli.108 Both stimuli are sensed by receptors and the information is transmitted through a series of biochemical reactions. Ultimately, appropriate decisions are taken.

In a signaling network, nodes represent proteins and protein complexes. However, nucleotides, lipids, and chemical compounds involved in the information flow process can also be included. A signaling network may contain a wide variety of relationships with different edges to represent phosphorylation/dephosphorylation, activation/inhibition, binding, association, ubiquitination, and so on. Some of these relationships are directional, such as phosphorylation, activation, or ubiquitination, where a protein acts on another protein to pass the information. However, in relations such as binding, and association, protein complexes are formed and no directionality of information flow can be ascertained.

1.4 Mathematical Modeling for Understanding Complex Biological Processes

It is important to note that biological networks usually include all known interactions in a cellular system. However, only a subset of these interactions may be active in a particular cellular and environmental context at a particular time. To discover the activated regions of the networks and their biological implications, networks have to be integrated with experimental data that represent the physiological state of the system in that particular condition. Then an iterative process is met with genome-scale data constraining and driving the development of models. The hypotheses used for the model can be tested experimentally to allow further refinement of the models.

Different mathematical frameworks have been developed for modeling the behavior of different types of biological systems. Two examples are described below:

Modeling Metabolic Processes

A metabolic network can be represented as a stoichiometric model, which is a mathematical representation of the mass conservation law (or mass balance) applied to the metabolites of the network. The mass balance for all metabolites in a network can be expressed in matrix notation by mapping the information contained in the network into the stoichiometric matrix (S). The stoichiometric matrix is of dimensions $n \times m$, where n is the number of metabolites and m is the

number of reactions in the network.

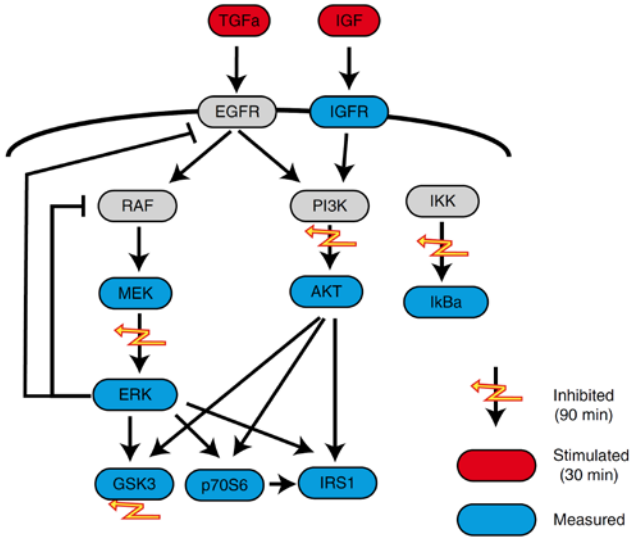
Using the stoichiometric matrix, the mass balance for the metabolites can be written as

$$\frac{dc}{dt} = S \cdot v = \sum_{j=1}^{m} S_{ij} v_j \ ,$$

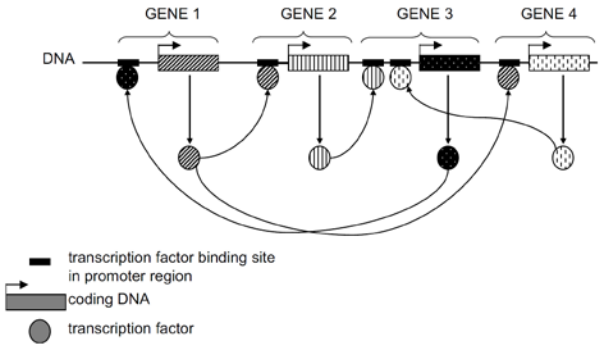
where $c(n \times 1)$ is the vector of metabolite concentrations and $v(m \times 1)$ is the vector of reaction rates.. The model can be solved for steady state conditions: $S \cdot v = 0$, which is often underdetermined because the number of reactions is typically greater than the number of metabolites in a metabolic network. This problem is addressed in constraint-based methods that analyze metabolic networks by characterizing the vector of reactions rates (also called flux distribution) that satisfy $S \cdot v = 0$. The method (called **F**lux **B**alance **A**nalysis **FBA**) estimates flux distributions that optimize a biological objective while satisfying $S \cdot v = 0$ and additional constraints regarding the reversibility and carrying capacity of the reactions. Commonly used objectives that have proven to produce meaningful flux distributions include the maximization (or minimization) of the biomass production rate, ATP production, or the synthesis of a particular metabolite.

Modeling Gene regulatory Networks

Signaling pathways serve as the cell's central control machinery, which tightly regulates the cell's response to external and internal stimuli. Many signaling pathways are triggered by the binding of extracellular biomolecules (*e.g.*, hormones or growth factors such as TGF, IGF) to a docking molecule (*i.e.*, receptor, such as EGFR, IGFR) embedded in the membrane surrounding the cell. If a signaling molecule binds to the extracellular region of the receptor, the receptor's three-dimensional structure may change that can trigger cascades of biochemical reactions within the cytoplasm. These cascades often involve specialized signaling molecules such as enzymes known as *kinases* (such as Raf, PI3K, Akt, MEK, ERK), which transfer phosphate groups from one molecule (the donor) to a specific target molecule (the substrate). The addition of the phosphate group changes the substrate protein's biochemical behavior so that it, in turn, can modify additional signaling molecules in the signaling cascade. Ultimately, this chain reaction results in the activation of proteins called transcription factors that bind in the cell nucleus to DNA, triggering expression of distinct sets of target genes.



Genetic and biochemical experiments in the 1960s demonstrated the presence of regulatory sequences in the proximity of genes and the existence of proteins that are able to bind to those elements and to control the activity of genes by either activation or repression of transcription. These regulatory proteins are themselves encoded by genes (see the following Figure for the representation of a fictional **G**ene **R**egulatory **N**etwork (**GRN**)). This forms a complex regulatory network with positive and negative feedback loops.



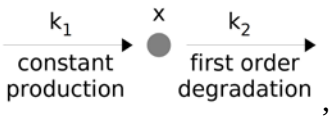
Representation of a simple, fictional transcription factor network. All genes shown encode transcription factors that control the activity of genes encoding transcription factors.

The dynamics of the resulting GRNs can be described with systems of differential equations

$$\frac{dx_1}{dt} = f_1(x_1, x_2, ..., x_n; k_1, k_2, ..., k_n$$
$$\frac{dx_2}{dt} = f_2(x_1, x_2, ..., x_n; k_1, k_2, ..., k_n$$
$$.....$$
$$\frac{dx_n}{dt} = f_n(x_1, x_2, ..., x_n; k_1, k_2, ..., k_n$$

These equations are typically highly nonlinear. Linear differential equations can be solved analytically; nonlinear ones cannot and a different perspective on nonlinear system dynamics is needed.

Let us consider a qualitative linear dynamics

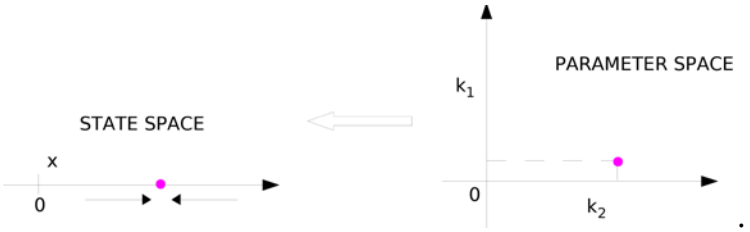


,

which can be properly described by

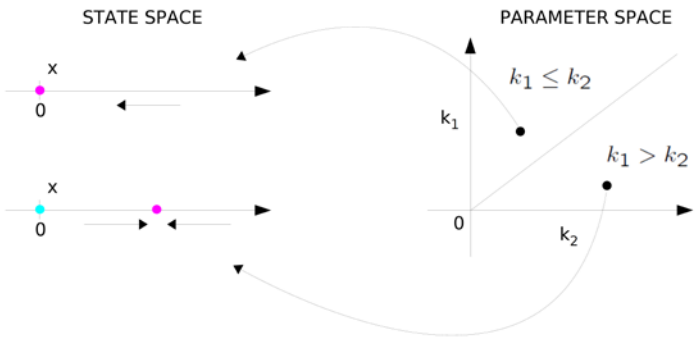
$$\frac{dx}{dt} = k_1 - k_2 x .$$

The equation has a solution $x(t) = \frac{k_1}{k_2} + (x(0) - \frac{k_1}{k_2})\exp(-k_2 t)$. For a given pair of parameter values (k_1 and k_2)), there is a unique stable steady state in the state space, which is globally attracting. Any other pair of parameter values has the same qualitative features and the dynamics only differs in a quantitative way.



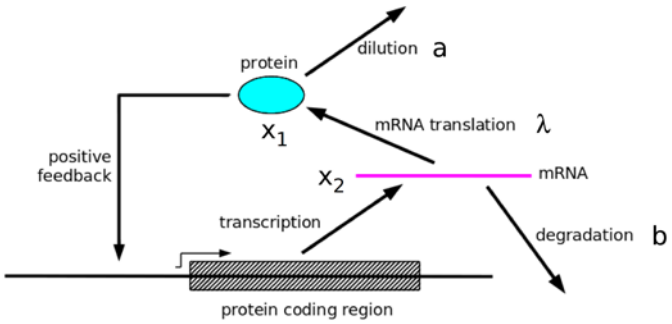
.

Now let us introduce a nonlinearity into the system $\frac{dx}{dt} = k_1 x(1-x) - k_2 x$. The parameter space breaks up into two regions: in the region $k_1 \leq k_2$ each pair of parameter values gives a unique, stable steady state at $x = 0$, which is globally attracting; in the region $k_1 > k_2$ each pair of parameter values has two steady states, one unstable at $x = 0$ and one stable at $x = 1 - k_2/k_1$ to which all positive initial conditions are attracted. A bifurcation takes places at $k_1 = k_2$



For a general system of nonlinear differential equations, parameter space can break up into multiple regions. Dynamical features shift abruptly through bifurcations as the boundary between two parameter regions is crossed. In each of these regions the state-space dynamics has the same qualitative features (the # of steady states and their stabilities); different regions have different qualitative dynamics.

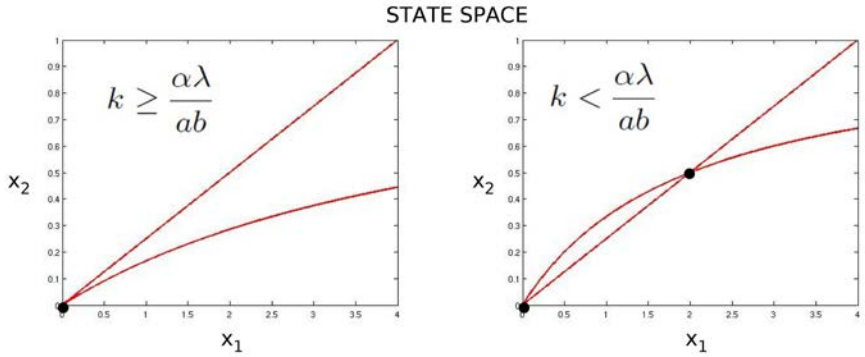
To investigate the dynamics in the state space, the first thing to calculate is the steady states – they are the skeleton around which the dynamics takes place. As an example, let us consider a gene transcription system with a positive feedback on itself. The schematic is shown as follows:



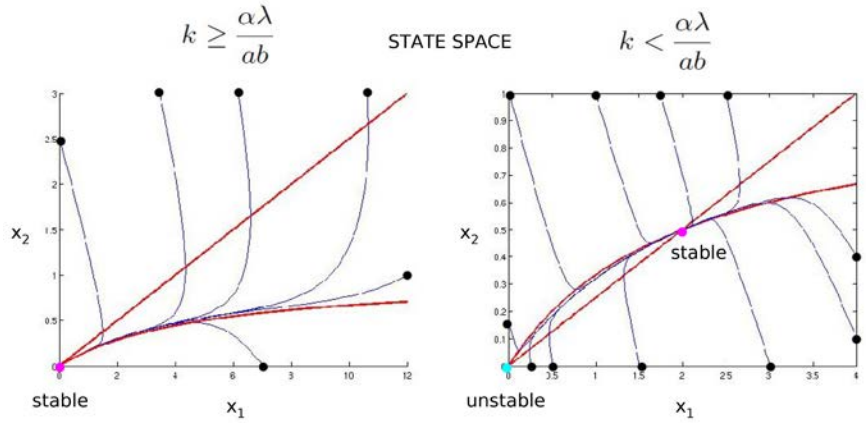
The system dynamics can be described by

$$\frac{dx_1}{dt} = \lambda x_2 - a x_1$$
$$\frac{dx_2}{dt} = \frac{\alpha x_1}{k + x_1} - b x_2 .$$

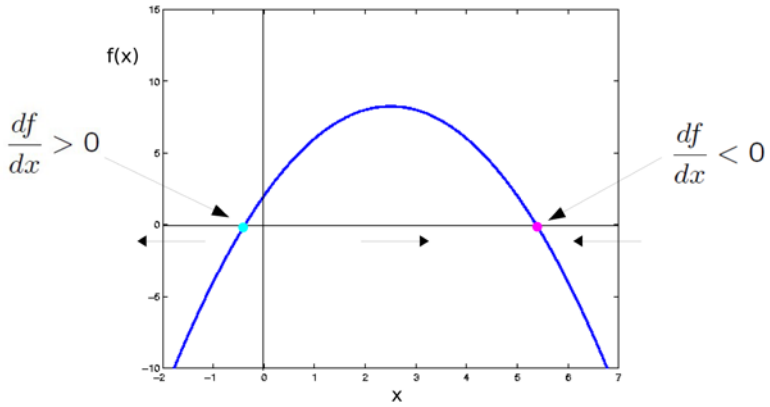
Here we introduce the method of nullclines for the 2D systems. The x_1 nullcline is the locus of points satisfying $\frac{dx_1}{dt} = 0 \rightarrow x_2 = \frac{a}{\lambda} x_1$. Similarly, the x_2 nullcline is the locus of points satisfying $\frac{dx_2}{dt} = 0 \rightarrow x_2 = \frac{\alpha}{b} \frac{x_1}{k + x_1}$. The steady states are the intersections of the nullclines.



The dynamics of the system can then be revealed with the trajectories in the state space. The following figure presents the trajectories starting from different points (the black spots) in the state space. The parameters used are $\lambda = 0.08\text{sec}^{-1}$; $a = 0.02\text{sec}^{-1}$; $b = 0.1\text{sec}^{-1}$; $\alpha = 0.1\mu M\ \text{sec}^{-1}$ and $k = 5\mu M$ for the left figure and $k = 2\mu M$ for the right figure. The system is found to dynamically approach to the nearby stable state (the pink color spot) no matter what it locates initially in the state space. The trajectories are attracted to and then hug the x_2 nullcline because α and b are 10-fold larger than λ and a. This time scale separation (i.e., x_2 is a fast variable and x_1 the slow variable) allows us to eliminate x_2 after the initial transient and reduces the system to a 1-dimensional dynamics (the Tykhonoff's Theorem).



The 1-D dynamical system becomes $dx/dt = f(x)$. The procedure to discover the 1D stability is:



1) find a steady state $x = x_{st}$, so that $\left.\frac{dx}{dt}\right|_{x=x_{st}} = f(x_{st}) = 0$;

2) calculate the derivative of f at the steady state $\left.\frac{df(x)}{dt}\right|_{x=x_{st}}$;

3) if the derivative is negative, then x_{st} is stable;
4) if the derivative is positive, then x_{st} is unstable;
5) if the derivative is zero, then nothing can be said.

But notice that derivatives give only local information, we cannot tell the size of the region from the deriviative. The stability region may be vanishingly small.

For n-dimensional systems with $dx/dt = f(x)$, the procedure can be modified as follows:

1) find a steady state $\boldsymbol{x} = \boldsymbol{x}_{st}$, so that $\left.\frac{dx}{dt}\right|_{x=x_{st}} = f(x_{st}) = 0$;

2) calculate the Jocobian matrix of f at the steady state $A = (Df)\big|_{x=x_{st}}$;

3) if all the eigenvalues of A have negative real part, then $\boldsymbol{x}_{st}$ is stable;
4) if none of the eigenvalues of A have zero real part and at least one of them has a positive real part, then $\boldsymbol{x}_{st}$ is unstable;

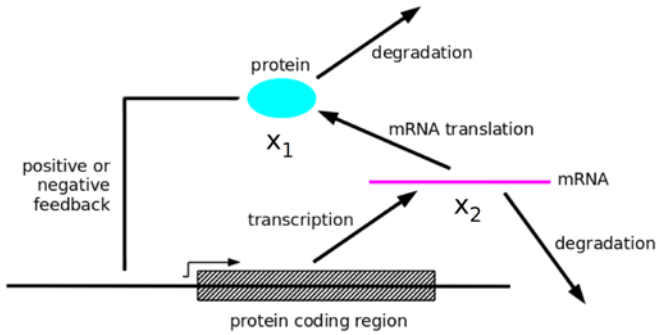
5) if one of the eigenvalues of *A* has zero real part, then nothing can be said.

There exists a homeomorphic mapping between the nonlinear state space and the linear state space that preserves the dynamical trajectories (qualitative similarity). So it is useful to linearize the nonlinear dynamic system, find the steady states of the corresponding linear system, and then study the dynamics nearby the steady states

$$\frac{dx}{dt} = f(x) \quad \xrightarrow[x = x_{st}]{\text{linearise around}} \quad \frac{dy}{dt} = \left[(Df)|_{x=x_{st}}\right] y$$

.

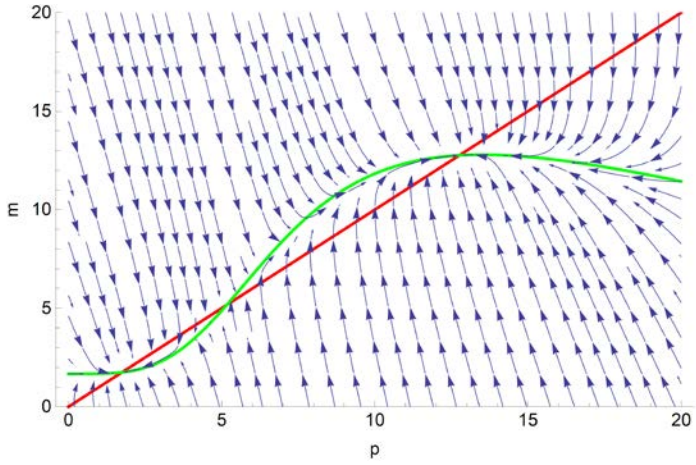
A Case Study of Auto-Regulation (AR) of a Single Gene



$$\frac{dx_1}{dt} = \frac{dm}{dt} = translation - \deg radation = r(p) - am$$
$$\frac{dx_2}{dt} = \frac{dp}{dt} = transcription - \deg radation = bm - cp \qquad \text{with} \quad a, c > 0 .$$

To find the nullcline for mRNA, set the equation dm/dt to 0: $m = r(p)/a$. Do the same for the equation for dp/dt: $m = cp/b$.

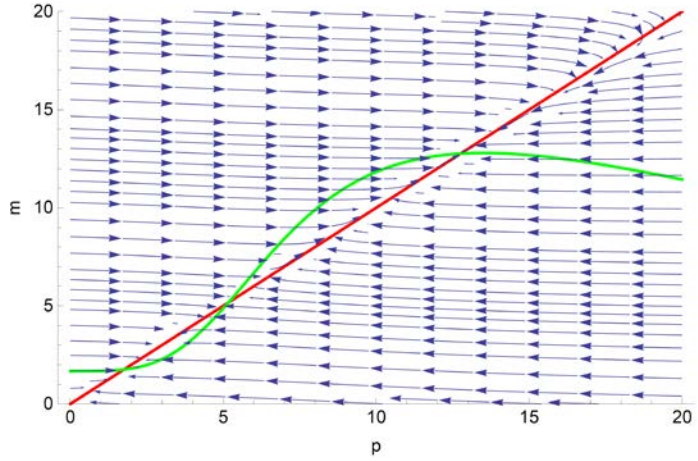
A vector field can be plotted in Mathematica using the StreamPlot command; and overlay plots using the Show command. Make a plot of the nullclines for *m* and *p*, overlaid with a vector field plot. Use *a* = 0.6, *b* = 0.1, *c* = 0.1.



The red curve is the protein nullcline; the green curve is the mRNA nullcline. The three intersections between the two nullclines represent the fixed points. The first fixed point (small *p*, small *m*) is a stable steady state because it attracts neighboring trajectories; the last fixed point

(large *p*, large *m*) is also a stable steady state. The intermediate steady state is unstable because trajectories tend to deflect away from it towards one of the two other steady states. The trajectories rapidly converged to the green curve (mRNA nullcline) and then traveled along the green curve to reach one of the two stable steady states (intersection with the red curve). This means that for these parameters, the mRNA rapidly reached steady state and stopped changing, followed by a slower change in the protein concentration.

Make a plot of the nullclines and the vector field again but using $a = 0.6$, $b = 10$, $c = 10$.



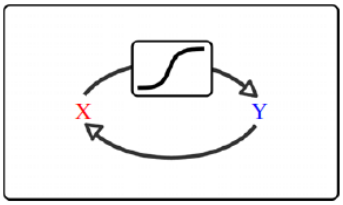
In this plot, we see the reverse phenomenon: the trajectories rapidly converge to the red curve (protein nullcline) and then travel more slowly to the steady states. This indicates that protein equilibrates more rapidly than mRNA for these parameters.

Assume that transcription and mRNA degradation are fast processes relative to protein translation and degradation. With this assumption, we can make a timescale separation and reduce the two-variable system to a one-variable system. We set the equation for $dm/dt = 0$ and solve for *m*: $m = r(p)/a$. Substitute your expression into $dp/dt = bm - cp$, $dp/dt = bm - cp = b\,r(p)/a - cp = \alpha r(p) - cp$. Here α represents a scaling factor that determines how the protein production rate is proportional to the transcription rate, *r*(*p*). A useful way to depict a one-dimensional system is to separate it into two parts: synthesis and degradation. Suppose that we have an equation of the form $dx/dt = S(x) - D(x)$, where *S* is a synthesis term and *D* is a degradation term. The criterion for a fixed point $dx/dt = 0$ then becomes $S(x) = D(x)$. If we draw $S(x)$ and $D(x)$ on the same plot, the fixed points are just the points at which $S(x)$ and $D(x)$ intersect. This allows us to easily see where the fixed points are, and how they are influenced by changes in the shape of the synthesis and degradation curves, giving us a powerful way in which to determine the qualitative properties of the system over a range of parameters.

A Case Study of a Bistable Network Motif

Complex dynamics of molecular signaling networks arise collectively from interactions among individual components. Cellular functions, such as proliferation, differentiation, homeostasis, mobility, metabolism and rhythmic behaviors, require proper integration of the dynamical properties of networks. Here we will use bistability as an example to illustrate the principle that can be invoked to produce the desired dynamics with a properly structured network motif.

Many cellular responses, including proliferation, differentiation, lineage specification and apoptosis, are all-or-none, in which cells choose between two discrete outcomes. Once cells commit to one fate over the other, the state transition is usually irreversible under physiological conditions. To deepen our insight into the behavior, let us consider a simple two-variable system in which genes *X* and *Y* activate each other transcriptionally, with linear degradation of each gene product. *Y* activates *X* in a simple Michaelis–Menten fashion, whereas *X* activates *Y* with the Hill function:



$$\frac{dX}{dt} = k_2 \frac{Y}{K_M + Y} - k_1 X$$
$$\frac{dY}{dt} = k_4 \frac{X^n}{K_H{}^n + X^n} - k_3 Y \quad .$$

The possible steady states of this system appear as intersection points of the *X* and *Y* nullclines (*i.e.*, the curves deduced from d*X*/dt=0 and d*Y*/dt=0): $X = \frac{k_2}{k_1} \cdot \frac{Y}{K_M + Y}$, and $Y = \frac{k_4}{k_3} \cdot \frac{X^n}{K_H{}^n + X^n}$. For the system to be bistable, the two nullclines must intersect each other three times, corresponding to two stable steady states and one unstable steady state in between. Given that the *X* nullcline bends upward or is at best a straight line, the *Y* nullcline has to be sufficiently 'twisted' in a certain way in order to cross the *X* nullcline back and forth multiple times. This behavior can be readily achieved when the *Y* nullcline is sigmoid. In fact, a certain degree of ultrasensitivity ($n>1$) in either of the two arms of a positive feedback loop is essential for bistability to arise.