UNIT -II GENE NETWORKS

What is the need for gene network

- Many biological research areas **such as drug design** require gene regulatory networks to provide clear insight and understanding of the cellular process in living cells.
- This is because interactions among the genes and their products play an important role in many molecular processes.
- A gene regulatory network can act as a **blueprint for the researchers to observe the relationships among genes.**
- Gene X is said to be regulated by gene Y if a change in the expression of gene Y also induces change in the expression of gene X.
- This regulation can be either up-regulation or down-regulation.
- Up-regulation results in activation and down regulation results in inhibition.
- One of the **most famous paradigms in gene regulation is the lac operon in prokaryotes.**

Computational approaches for gene regulatory network construction, which include

1. Bayesian network,
2. dynamic Bayesian network,
3. Boolean network,
4. probabilistic Boolean network,
5. ordinary differential equation and
6. network

>Bayesian network methodology

- Bayesian network is a directed and acyclic graph where G=(X, A) with a set of local probability distribution p.
- X represents the nodes, {x1, …, xn}, which represent the gene variables.
- However, A indicates the directed edge that corresponds to probabilistic dependence interactions among genes.
- A DAG is a **network graph with directed edges and no cycles**.
- In a DAG, the edge pointing from one node to another represents the regulatory relationship between parent node and child node.

- For instance, if X1 is the parent of X2 and X2 is the parent of X3, the assumptions are that X1 is the ancestor of X3 and X3 is the descendant o

$$P\{x_1, x_2, \ldots x_n\} = \prod_{i=1}^{n} P\{x_i | \text{ parents } (x_i)\}$$

- where xi is a given gene node.
- n is the total number of genes involved.
- The parents (xi) is the parent gene that regulates gene xi.
- P(xi/ parent (xi)) for node i is denoted as Conditional Probability Distribution (CPD) or local distribution. CPD must be specified for the nodes that have parents.
- Bayesian network modelling involves two major steps:
- structure learning and parameter learning.
- Structure learning is also known as model selection.
- The structure of the network is constructed in the structure learning stage.
- However, the probability values of each network node are estimated during parameter learning.

>Applications of Bayesian network

Dataset	Description
• *Saccharomyces cerevisiae* mutants	• Extended Bayesian network to handle perturbation and applied discretisation procedure to dataset.
• *Saccharomyces cerevisiae* • Synthetic RAF-pathway	• Applied Bayesian network with the confluence of information from expression data with prior knowledge.
• *Saccharomyces cerevisiae*	• Clustered the genes according to gene ontology and employed the cross-correlation between co-clustered genes to provide time delay information to Bayesian network.
• *Escherichia coli* • RAF cell signaling	• Utilised SCS algorithm to reduce computational time of Bayesian network and obtained optimal network using the combination of search-and-score and constraint-based approaches.
• *Saccharomyces cerevisiae* • *Escherichia coli*	• Constructed gene networks using the combination of Bayesian network with hill-climbing algorithm and Efron's bootstrap approaches.

>Dynamic Bayesian network

- Constructing gene regulatory networks using dynamic Bayesian network (DBN) has become the major focus for discovering the relationships between genes.
- Dynamic Bayesian network is an extension of Bayesian network that is able to infer the interaction uncertainties among genes by using a probabilistic graphical model.
- Compared to a Bayesian network, Dynamic Bayesian network is capable of modelling cyclic interactions among genes, which is an important aspect for biological network modelling.
- In dynamic Bayesian network, **the nodes in the network are duplicated** in order **to model the cyclic interactions and dynamic behaviours of gene regulatory networks.**
- Dynamic Bayesian network is capable of **modelling the evolution of temporal processes** based on the time slices property.
- Temporal relationships among genes are represented by interconnecting time slices.
- Dynamic Bayesian network provides DAG as in Bayesian network. However, the edges in dynamic Bayesian network connect particular nodes between two consecutive time slices.
- Dynamic Bayesian network **supports both intra-slice connections and inter-slice connections.**
- Intra-slice connections are the **connections within time slices**, whereas interslice connections are the **connections between time slices**.

>Dynamic bayesian network techniques in gene regulatory network construction:

Dataset	Description
• *Saccharomyces cerevisiae*	• Improved the quality of inferred network by incorporating dynamic Bayesian network with perturbations.
• *Saccharomyces cerevisiae*	• Enhanced the performance of dynamic Bayesian network using GSR and MCMC.
• *Arabidopsis thaliana* • *Saccharomyces cerevisiae* • Synthetic RAF-pathway	• Improved the convergence over RJMCMC by introducing a dynamic programming scheme into a multiple change-point process.
• *Saccharomyces cerevisiae*	• Proposed a dynamic Bayesian network-based model to improve the prediction efficiency and reduce the computation time.
• *Escherichia coli* • *Cyanothece* • Glucose homeostasis	• Combined the dynamic Bayesian network approach with deterministic global optimisation for gene regulatory network construction from time course expression data.

>Boolean network:

- Boolean network is said to be the simplest models for regulatory networks construction.
- It is easy to simulate and popular in capturing the global dynamical behaviour of genetic regulatory networks and interactions of genes .
- Boolean Network can describe different biological phenomena in a system, such as **oscillations, multi-stationarity, switch-like behaviour stability and hysteresis**
- Two types of inferring approaches are used in the Boolean network:
- correlation and inferring
- For the correlation approach, gene relationship information is obtained by different approaches, which are then **used in modelling the topology of the connections between genes.**
- **The inferring approach is a machine learning approach** and the most commonly used algorithm in genetics
- A Boolean is defined as variable that can only assume two values; the values are usually represented by 1 and 0 or true and false. The operators of logic are and, or and not.

Dataset	Description
• *Schizosaccharomyces pombe*	• Boolean network model with network connectivity graph was used to model fission yeast cell cycle.
• *Saccharomyces cerevisiae*	• O(log n) state transition pairs were used to determine the original Boolean network.
• *Caenorhabditis elegans*	• Temporal Boolean network model was constructed and decision output tree was represented by 0 or 1.
• *Saccharomyces cerevisiae*	• Gene network data from multiple gene disruption was analysed.
• *Drosophila melanogaster*	• The results indicate that if the net effect is preserved, the kinetic details of the interactions do not matter.

>Ordinary differential equation:

- This is a popular tool to model dynamic system gene regulation.
- In order to analyse network dynamics, locate limit cycles or investigate bifurcation behaviour, ODE are the best analysed approach for non-linear systems

$$\frac{dxi}{dt} = fi(x1,\ x2,\ \ldots,\ xN,\ p,\ u) \quad (4)$$

in which x is the expression level of gene i at time t (the independent variable), n represents the number of genes, u is the an external perturbation to the system and p indicates the parameter set of the system. The initial condition of X is where variable x at time t is 0. In ODE models, which use continuous time variables with constraints, there is no negative value allowed in the cellular concentration, such as for protein and mRNA molecules. The degradation of mRNA or proteins is assumed to not be regulated.

$$\text{Transcription}: \quad \frac{dr_i}{dt} = F(f_i^R(p_1),\ (f_i^R(p_2),\ \ldots,\ f_i^R(p_n)) - \gamma_i r_i$$

$$\text{Translation}: \quad \frac{dp_i}{dt} = f_i^p(r_i) - |\delta_i p_i \quad (6$$

i: any given gene;
r_i : rate of change of the concentration of the transcribed mRNA;
p_i : rate of change of the concentration of translated protein.

Advantages and disadvantages of computational approaches in gene regulatory network construction.

Computational approaches	Strength	Weakness
Boolean network	• Capable to analyse large regulatory networks • Easier to interpret due to its simplicity • Phenomena of biological realistic complex can represent by Simplistic Boolean formalism [66] • Large set of algorithms is provided which is available in already supervised learning in the binary domain	• Require synchronous update • Deterministic in nature • Unable to handle incomplete regulatory network data only involves two representative states for gene expression level • High computing time is needed • Most of BN can only use with a small number size of genes
Probabilistic Boolean network	• Cope with uncertainties • Two or more transition function for each variable is allowed the use of positive feedback and probabilities can make the model work more effectively [67] • Compared to DBN, PBN can explain more details in the regulatory roles of different sets of gene [68]	• Difficult to apply for large scale of network • High computational complexity • Cannot cope with instantaneous interactions between variables
Bayesian network	• Ability of handling noisy • Handle with uncertainty • Able to work on the logically interacting components with small number of variables • Integrate the prior knowledge to strengthen the causal relationship • Infer the structure of network statistically	• Hard to distinguish between the origin and the target of an interaction • Feedback loops not allowed • Failure to capture temporal information of time series microarray data • Support small sized gene regulatory networks • Combinatorial learning of Bayesian network

Dynamic Bayesian networks	• Able to model cyclic interaction among gene • Capable to handle stochastic components • Well-suited for handling time-series gene expression data • Model indirect or direct causal relationships • Handle perturbation or structural modification of networks	• Excessive computation time and cost • Performance restrict by the missing values o gene expression data • Support small sized gene regulatory network
Ordinary differential Equations	• Produced directed signed graphs • Suited for steady-state and time series expression profiles • Can work entirely in classical category	• Only applicable to smaller network • Difficult to find appropriate parameter value that fit with the data exhibit increasing in finite time
Neural network	• Able to recognize input pattern • Able to model any functional relationships and data structure • Captures the nonlinear and dynamic interactions • Noise resistant • Biological plausible	• Difficult to obtain efficient training since learning rate must be defined for different d situation • High computational complexity therefore ca only apply to very small systems

- 10 genes linked to a particular disease (for the sake of example say cancer).
- I want to build a gene network for these 10 genes.
- Any web based tools available which can do the job?

GeneMANIA. If you open up the advanced options, there's hundreds of networks to choose from. You can save the network from the file menu.

STRING. This is the oldest and probably most widely used web tool for this purpose.

Other web-based network-construction tools for human include:

FunCoup -- one of the few tools with regulatory interactions. Also has some nice display and clustering options

FunctionalNet -- by one of the founders of the field of functional interaction networks.

Cytoprophet, that reconstructs networks based on protein and domain interaction databases. You just have to load a list of SwissProt homologs, and then the plugin does the rest.

Reactome Functional Interaction (FI) Network plugin. It has been developed for analyzing cancer and other disease datasets.

Gene expression profiling

Gene expression profiling is the determination of the pattern of genes expressed, at the level of transcription, under specific circumstances or in a specific cell to give a global picture of cellular function. Techniques to measure this include DNA microarrays which measure the relative activity of previously identified target genes, or sequencing technologies that allow profiling of all active genes.

- In genetics, gene expression is the most fundamental level at which the genotype (the genetic makeup of a cell, an organism, or an individual) gives rise to the phenotype (any observable characteristic of a gene).
- Every cell in the body contains a full set of chromosomes and identical genes, with a few exceptions. However, only a fraction of these genes are turned on – or "expressed" – and it is this subset of genes that confers the unique properties of each cell type. This biological mechanism acts as both an "on/off" switch to control which genes are expressed within a cell, as well as a "volume control" that increases or decreases the level of expression of particular genes.
    - Disruptions or changes at any step of gene expression are responsible for many genetic diseases. Through the use of microarrays, scientists can determine – in a single experiment – the expression levels of hundreds or thousands of genes within a cell.

Mechanisms of Gene Regulation

Gene expression workflow.

Researchers may perform gene expression analysis at any one of several different levels at which gene expression is regulated: transcriptional, post-transcriptional, translational, and post-translational protein modification.

Transcription, the process of creating a complementary RNA copy of a DNA sequence, can be regulated in a variety of ways. Transcriptional regulation processes are the most commonly studied and manipulated in typical gene expression analysis experiments.

The binding of regulatory proteins to DNA binding sites is the most direct method by which transcription is naturally modulated. Alternatively, regulatory processes can also interact with the transcriptional machinery of a cell. More recently, the influence of epigenetic regulation, such as the effect of variable DNA methylation on gene expression, has been uncovered as a powerful tool for gene expression profiling. Varying degrees of methylation are known to affect chromatin folding and strongly affect accessibility of genes to active transcription.

Following transcription, eukaryotic RNA is typically spliced to remove noncoding intron sequences and capped with a poly(A) tail. At this post-transcriptional level, RNA stability has a significant effect on functional gene expression, that is, the production of functional protein. Small interfering RNA (siRNA) consists of double-stranded nucleic acid molecules that are participants in the RNA interference pathway, in which the expression of specific genes is modulated (typically by decreasing activity). Precisely how this modulation is accomplished is not yet fully understood. A growing field of gene expression analysis is in the area of microRNAs (miRNAs), short RNA molecules that also act as eukaryotic post-transcriptional regulators and gene silencing agents.

Gene Expression Profiling and Quantitation: Methods and Techniques:

- Researchers studying gene expression employ a wide variety of molecular biology techniques and experimental methods.
- Gene expression analysis studies can be broadly divided into four areas: RNA expression, promoter analysis, protein expression, and post-translational modification.

RNA Expression

Northern blotting — steady-state levels of mRNA are directly quantitated by electrophoresis and transfer to a membrane followed by incubation with specific probes. The RNA-probe complexes can be detected using a variety of different chemistries or radionuclide labeling. This relatively laborious technique was the first tool used to measure RNA levels

DNA microarrays — an array of oligonucleotide probes bound to a chip surface enables gene expression profiling of many genes in response to a condition. Labeled cDNA from a sample is hybridized to complementary probe sequences on the chip, and strongly associated complexes are identified optically. Gene expression profiling is often a first step in a gene expression analysis workflow, investigating changes in the expression profile of a whole system or examining the effects of mutations in biological systems

Real-Time PCR

Steady-state levels of mrna are quantitated by reverse transcription of the RNA to cdna followed by quantitative PCR (qpcr) on the cdna. The amount of each specific target is determined by measuring the increase in fluorescence signal from dna-binding dyes or probes during successive rounds of enzyme-mediated amplification.

This precise, versatile tool is used to investigate mutations (including insertions, deletions, and single-nucleotide polymorphisms (snps)), identify DNA modifications (such as methylation), confirm results from northern blotting or microarrays, and conduct gene expression profiling. Expression levels can be measured relative to other genes (relative quantification) or against a standard (absolute quantification).

Real-time PCR is the gold standard in nucleic acid quantification because of its accuracy and sensitivity. Real-time PCR can be used to quantitate mrna or mirna expression following conversion to cdna or to quantitate genomic DNA directly to investigate transcriptional activity

Promoter Analysis

Expression of reporter genes/promoter fusions in host cells — promoter activity (transcription rate) is measured in vivo by introducing fusions of various promoter sequences with a gene encoding a product that can be readily measured to monitor activity levels

In vitro transcription (nuclear run-on assays) — transcription rates are measured by incubating isolated cell nuclei with labeled nucleotides, hybridizing the resultant product to a membrane (slot blot), and then exposing this to film or other imaging media

Gel shift assays — also called electrophoretic mobility shift assays, these are used to study protein-DNA or protein-RNA interactions. DNA or RNA fragments that are tightly associated with proteins (such as transcription factors) migrate more slowly in an agarose or polyacrylamide gel (showing a positional shift). Identifying the associated sequences provides insight into gene regulation

Chromatin immunoprecipitation (ChIP) — protein-binding regions of DNA can be identified in vivo. In living cells, DNA and protein are chemically cross-linked, and the resulting complex is precipitated by antibody-coated beads (immunoprecipitation). Following protein digestion and DNA purification, the sequences of the precipitated DNA are determined

Let's look at obesity as an example of how gene expression can correlate with disease risk:

Obesity is a major health risk in worldwide that threatens children and adults alike. It can lead to heart disease, high blood pressure, and diabetes, especially as people age. A complex medical condition, obesity is influenced by diet, exercise, metabolism, and genetics.

Shan, age 17, is more than 100 kg overweight relative to his height. His parents and grandparents are all overweight as well.

Allen is similar to Shan with respect to age, height, diet, and exercise habits, but he is not overweight. Furthermore, no one in Allen's family is overweight.

Both Shan's and Allen's families volunteer to participate in a university study to identify genes that play a role in obesity. How will the researchers approach this question?

Three generations of family members provide cell samples (liver and fat cells) to the researchers. Liver and fat cells were chosen because they are important in metabolism and making fats.

The researchers will use an approach called gene expression profiling to identify active and inactive genes in a cell or tissue.

Expression profiling then tells the scientist which genes may play a role in obesity.

Scientists run similar expression profile studies on all the family members, as well as on those of other study participants.

Once the scientists compare the results from everyone in the study, they have a good idea which genes play a role in obesity. This information can be used in several future applications:

1. Creating diagnostic tests to predict whether a patient has a genetic predisposition to obesity.

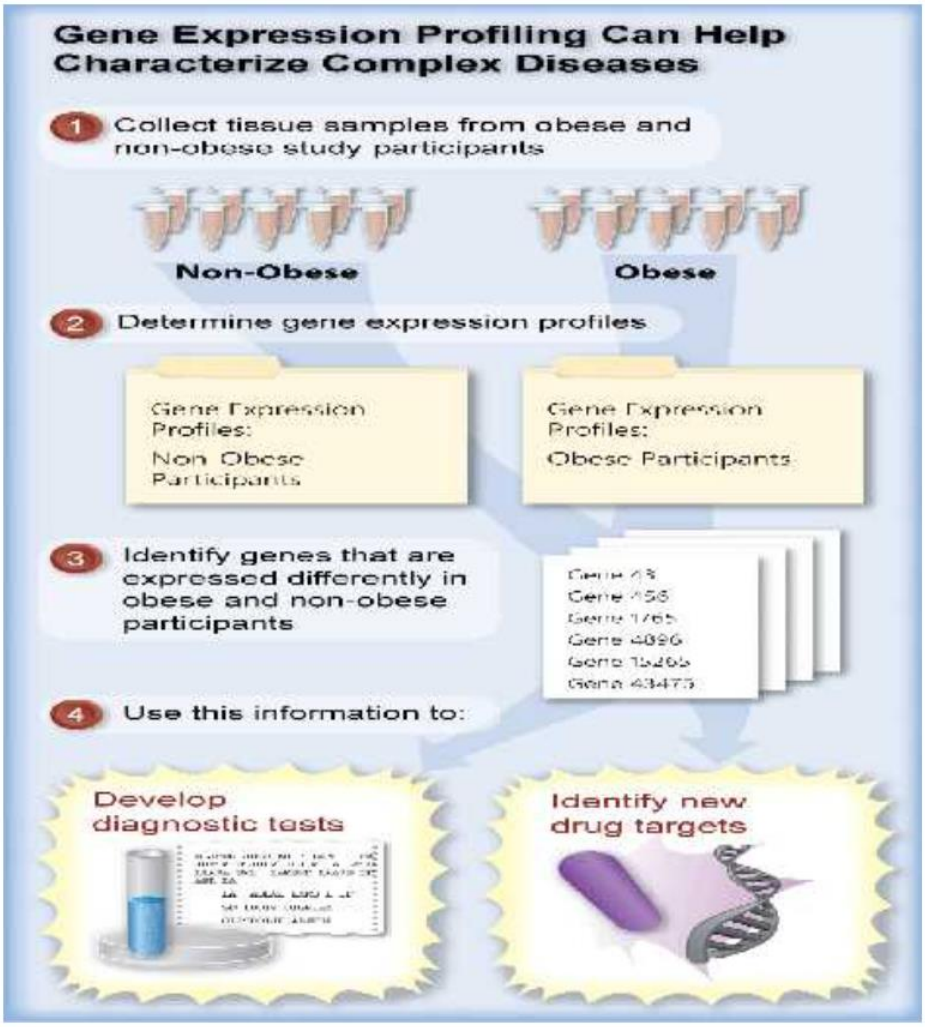
One test might examine the DNA sequence of a person's obesity-related genes, in order to detect genetic signatures that predict a predisposition to obesity.

Another test might examine a tissue sample for abnormal gene expression patterns that indicate a predisposition to obesity.

2. Designing Drugs

Designing drugs intended to treat or prevent obesity. This could be done by isolating the protein products of the identified obesity genes, determining their molecular structures and functions, and making drugs to inhibit them.

Designing drugs to control expression of obesity genes. These drugs would interact directly with DNA in key cells and tissues to prevent genes that are activated in obesity from being turned on—or, conversely, to prevent genes that are inactivated during obesity from being turned off.



4.3.2 Example 2: Computational analysis of signal specificity in yeast

Yeast is well recognized as an excellent model organism for systems level analysis [45]. Their ability to undergo efficient homologous recombination is particularly useful for studying the functional role of proteins in vivo, through gene disruption or gene replacement. Because of this property, the yeast pheromone response system is arguably the best-characterized signaling pathway of any eukaryote. This pathway bears strong similarities to signaling networks in mammals. In particular, the MAPK components share extensive sequence similarity with their human counterparts [46]. Another feature common to the yeast pheromone response pathway and response pathways of higher organisms is the sharing of signaling proteins among multiple systems. This property makes the pheromone response pathway an excellent system for studying signal specificity.

Depending on specific external cues, yeast cells initiate either a mating response or an invasive growth program. Mating is initiated when haploid cell types **a** and α secrete and respond to type-specific pheromones, which act through G protein-coupled receptors on cells of the opposite mating type [47]. Alternatively, invasive growth occurs in nutrient-poor conditions [48]. Combined genetic and biochemical studies revealed that both mating and invasive growth require a protein kinase cascade comprised of Ste20 (MAP4K), Ste11 (MAP3K), and Ste7 (MAP2K) [Figure 4.5(a)]. The pathways diverge at the level of the MAP kinase. Whereas deletion of one MAP kinase gene (KSS1) blocks invasive growth, deletion of a second MAP kinase gene (FUS3) impairs pheromone-induced cell-cycle arrest. Deletion of FUS3 leads to enhanced activity of Kss1 [49]. However, the mechanism by which this cross inhibition occurs was unknown.

We recently combined mathematical modeling with experimental analysis to investigate how Fus3 limits the activity of Kss1 [50]. Six mathematical models were developed to describe different hypothetical mechanisms of cross inhibition. All six models were fit to the time courses for Fus3 and Kss1 activation obtained from wild-type cells as well as

Computational Analayis of Signal Specificity in Yeast:

From stains containing various genetic alterations,The experiments yielded a dataset of

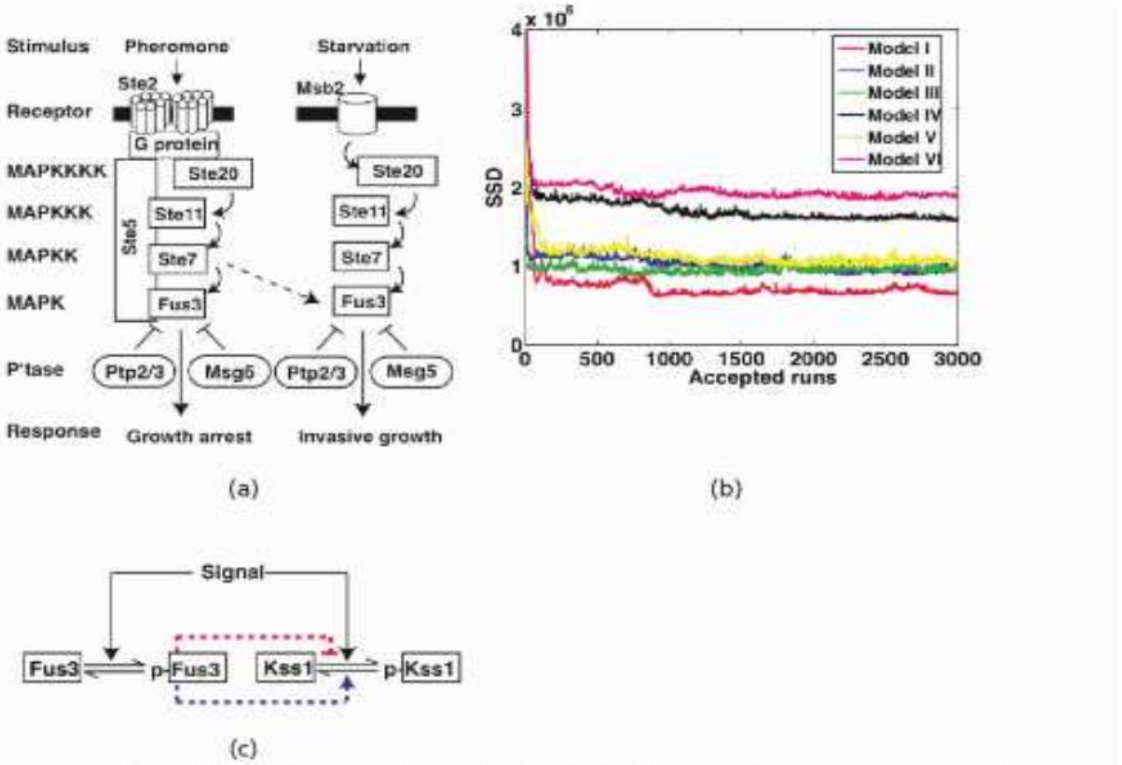


Figure 4.5 Data-driven modeling of signal specificity in yeast. (a) Components of the mating and invasive-growth pathways. Activation steps are indicated with arrows, and inhibition steps are indicated with a T-shaped line. (b) The sum of the squared differences (SSD) between the experimental data and output of the six models versus the number of accepted realizations in the Monte Carlo optimization routine. (c) A simple model that incorporates two mechanisms of cross-inhibition: Fus3 inhibits the rate of Kss1 phosphorylation (red dashed line), and Fus3 increases the rate of Kss1 dephosphorylation (blue dashed line).