学　位　論　文

A Memetic Algorithm for Reconstructing Gene Regulatory Networks from Expression Profile (Memetic アルゴリズム を用いた遺伝子制御ネットワークの推定)

指導教員　伊庭　斉志 教授

平成１８年１２月博士（科学）申請

東京大学大学院　新領域創成科学研究科　基盤情報学専攻

ノマン　ナシムル

論文内容の要旨

近年の分子生物学の急速な進歩は，多くの新領域，複合領域の応用研究の発展を促した．新しい研究領域の一つに，生物有機体の全てをシステマティックに説明しようとするシステム生物学がある．システム生物学は，新しい生物システムの設計と，現存生物の制御を可能とする事を研究の目的としている．目的を達成するためには，現存生物の構造の特徴と，その振る舞いを明らかにするという基礎研究が必要になる．そのような特徴と振る舞いは，遺伝子の携帯で必要な全ての情報を含むゲノムと相互作用をする，分子の構成要素間の複雑な制御ネットワークによって決定される．この制御ネットワークとネットワークの働きのメカニズムを解明することが，システム生物学における重要な課題である．

現在の所，多くの生体内作用の中心である遺伝子の制御関係は，プロモーター遺伝子によるシス作用制御と，その他の遺伝子の転写因子によるトランス作用制御の組合せであると考えられている．現象論的モデルを用いてシス作用を単純化すれば，遺伝子間の全てのトランス作用は「遺伝子制御ネットワーク」としてモデル化することができる．しかしながら，この生体制御ネットワークは動的かつ高次の非線形的性質を持っており非常に複雑であるため,このネットワークで表現される生体内要素やその依存関係，分子レベルでの相互作用の解明が進んでいなかった．

ところが最近，DNA マイクロアレイやオリゴヌクレオチドチップなどの測定のための新技術が開発され，複数の遺伝子発現の時系列変化や定常状態データを高速かつ同時に測定する事が可能になった．ゲノム規模でトランスクリプトームを観測することにより，ゲノム活性の構造と動的な変化について大域的な視点から考察を行うことができる．観測によって得られた大量の実験データにより，特定の生化学プロセスにおける特定遺伝子の活動の包括的な理解が可能になるのみならず，遺伝子制御ネットワークの構造も理解することが可能になると考えられている．従って，大量の生物学な観測データをもとに計算モデルをたて同定を行うシミュレーションにより、遺伝子ネットワークの転写制御を明らかにする手法の開発が必要となっている．

用いられる計算モデルはネットワークの接続関係や速度定数，生化学濃度のようなパラメータを考慮にいれたものである．GRN のモデルとして利用できるものに生化学系理論－モデル化と生化学系の分析のための一般化されたフレームワークに基づく S-system モデルがある．このモデルは生物学的妥当性と数学的柔軟性の双方の利点を持つ．高次元問題を扱うために、元のモデルは結合した形態からネットワークの推定において有効である分離問題に分解される．この

問題に対して進化論的計算手法を用いると，モデルの動的な性質と観測されたシステム応答のノイズの優位水準に優れていることが判明している．そのため進化論的計算手法を S-system を用いた GRN 推定に適用することが期待されている．

現在研究されている S-system を用いた GRN 推定での課題は，(i)疎なトポロジーアーキテクチャを検知すること，(ii)有意水準のノイズにより破損している限られた量の遺伝子発現データから kinetic パラメータを推定すること，(iii)GRN を同定するために大域的最適解を探索する能力のある効率的な最適化方法を考案すること，の三つである．本論文では，これらの課題に対処するために分解S-systemモデルを利用したGRN推定のためのミメティック方法を提案している．

最初に，有用であるとされる進化アルゴリズム(EA)である標準的な Differential Evolution (DE) を，交叉に基づく局所探索を用いて性能強化を行った．また DE に近傍探索を組み込み性能向上を狙った．局所探索のためのヒューリスティクスを組み込むことは大域的最適化のための効果的な進化アルゴリズムにとって非常に有効であることが広く知られており，近傍探索を組み込んだ EA は Memetic Algorithm (MA) として有名である．

広範囲のベンチマーク問題で実験を行った結果，提案手法である拡張版 DE は従来の DE より良いか，同等の性能を持つことが分かった．また，他の有名な進化アルゴリズムとの性能比較も行った．連続ランドスケープ発生器によって生成されるランダム問題で，元のアルゴリズムに対する優越性が確認された．さらに，実問題に対する適合性を，高次非線形，多峰性，だまし要素を含む S-system 最適化問題に適用することによって調査した．その結果，実問題の最適化では元の DE と比較して優越性があることが示された．

次に，この効率的かつロバストな最適化アルゴリズムを用いて，分離 S-system 形態で表される生化学ネットワーク内の転写制御関係を推定するためのミメティックアルゴリズムを設計した．提案アルゴリズムは，ロバストな転写制御関係の同定，正確なキネティックパラメータの推定，ネットワークのアーキテクチャの獲得，そして効率的な計算などの課題の解決を目標として設計した．アルゴリズムの核である，最適化エンジンは memDE と呼ばれる前述の拡張した最適化アルゴリズムを用いて実装した．さらに，拡張アルゴリズムでは骨格ネットワークを効率的に獲得するための局所探索を組み込んでいる．局所探索では，より疎なネットワークを獲得するために各世代からの最良個体とランダムに選ばれた個体の周辺の山登り探索を行う．また，ネットワーク内の最もロバストな制御関係と正確なキネティックパラメータを同定するために，アルゴリズムを 2 倍の最適化用に設計した．

ターゲットとするネットワークの最適パラメータの組合せを探索するときに，解候補を評価するための測度が必要となる．最もよく用いられている評価基準は数値計算によって算出される遺伝子発現レベルと観測されたシステムのダイナミクスの差である平均二乗誤差(MSE)である．ネットワークの骨格構造を獲得するために，適合度基準に基づく MSE のための新しいペナルテ

ィー項の提案を行う．本提案では，ネットワークのパラメータの組み合わせ最適化の性能向上のために，伝統的なMSEの代りに，適合度基準に基づくAICを用いる．実際には，適合度評価関数でAICのペナルティー項にComplexity項の追加を行い，疎なネットワークの構造を有するモデルの選択を容易にした．

提案手法の評価として，推定に用いるデータの量，ノイズのレベルなど様々な次元・特徴において，提案方法の適合性を調べる実験を行った．どの推定方法もネットワークトポロジーと制御パラメータを獲得することができたが，その精度は発現データの量とノイズのレベルに制限されてしまう．既存の適合度関数と比較して，提案手法は正しい発現データ量やノイズのレベルに対してロバスト性が高く，どのような状態でもネットワークトポロジーの同定と正確なパラメータを推定するのに適していることが分かった．提案アルゴリズムの異なる要素が，ロバストさ，効率，正確な制御パラメータの推定に必要であることを示すためにアルゴリズムの実験分析を行った．さらに既存のGRN推定アルゴリズムに比べて，提案した推定アルゴリズムは，計算量について効率的，ノイズに対してロバスト，より少ないサンプルサイズで推定可能であることが分かった．最後に，提案方法を2つの実データに適用したところ，主要な制御遺伝子の基本的なネットワークが獲得された．

今回提案した拡張アルゴリズムによる遺伝子ネットワークアーキテクチャとパラメータの推定の性能評価は，数十の遺伝子から構成される中型ネットワークを推定する事で行った．しかし数百，数千の遺伝子を含むネットワークを推定するにはさらなる強化，拡張，改良が必用となると考えられる．本論文では今後の課題についての方向付けを与えて結論を出す．

Abstract

With increasing number gene expression data being available, the field of the systems biology is targeting the genome-wide identification of the structure of biomolecular interactions. However, the limited amount of gene-expression data and the significant amount of noise from the measurement technology place the greatest challenges for such reconstruction processes. This dissertation addresses the challenges of reverse engineering molecular pathways of gene regulation from gene expression data using the decoupled S-system model. Again, modeling gene regulatory networks using S-system imposes additional difficulties for the reconstruction algorithms such as identifying the sparse network architecture and efficient learning of the model parameters. This work also deals with these issues.

In order to design an efficient and robust optimizer, first, the standard Differential Evolution (DE) algorithm was hybridized with a crossover based local search operation to improve its neighborhood exploration capability. This improved optimizer was used in the core of the reconstruction algorithm for inferring the transcriptional regulations in a biochemical network. Besides, a hill-climbing local search method was embedded in the developed algorithm for obtaining the sparse network structure efficiently. For identifying the skeletal structure of the target network, enhancements of the conventional Mean Squared Error (MSE) based fitness function and a new Information Criteria based fitness function have been proposed.

The suitability of the method is tested in gene circuit reconstruction experiments, varying the network dimension and/or characteristics, the amount of gene expression data used for inference and the noise level present in expression profiles. The reconstruction method inferred the network topology and the regulatory parameters with high accuracy. The proposed fitness functions have been found more suitable for evaluating the candidate network models compared to the existing ones. The proposed algorithm ascertained higher computational efficiency compared to other algorithms. Finally, the methodology was applied for analyzing two real gene expression profiles to reconstruct the underlying networks.

Acknowledgements

I would like to express my deepest gratitude and profound thankfulness to Professor Hitoshi Iba for providing me the concept, plan and overall supervision of this work. Dr. Iba initiated me into the application of evolutionary computation to the problems in bioinformatics, with particular emphasis in gene network reconstruction. I had the glorious opportunity to learn many important concepts of theoretical and practical evolutionary computation from a true master. Dr. Iba taught me how to use this powerful methodology to deal with the challenges of reverse engineering genetic networks. His novel ideas, thoughtful discussions, judicious criticism and prudential recommendations added greatly to reach the final result. I also owe to him for his constant encouragement, direction and suggestions which were a source of inspiration for the progress of the work.

I am endlessly indebted and grateful to my parents and to my wife. I know they are the three persons who sacrificed most so that I can complete this work. Thank you for your unconditional love, persistent encouragement and immeasurable support without which this achievement could not be a reality.

I am grateful to the Japanese Government (Monbukagakusho) for the financial support throughout the tenure of my postgraduate study which was indispensable for the successful completion of my degree. I am indebted to Iba Laboratory and to the Department of Frontier Informatics, Tokyo University for their facilities and services that I enjoyed during the course of my studies. Last but not the least, I will like to express my appreciation to the current and former members of the laboratory for their never failing support, help and encouragement that made my stay in a foreign country most comfortable, trouble-free and pleasing.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

At the beginning of the new millennium, with the completion of human genome sequencing, a new era of genomic research has begun which will eventually change many of our long-existing concepts about us and other species on this earth. Formally initiated in 1990, the Human Genome Project (HGP) was a 13-year effort resulted from the high-throughput sequencing technology and developments in functional genomics. The primary goals of the HGP were to determine the sequences of the three billion nucleotides that make up human DNA and to identify the portions of the whole genome that constitute functional genes.

The HGP was a landmark genome project and some people believe that the era of genomics is one of the fundamental advances in human history. With the steady decrease in genome-sequencing cost and availability of newer technologies many genome projects were undertaken for many model organisms of different complexity e.g. from bacteria to chimpanzee. Many of these genome projects have already been completed which in turn triggered new genome projects to decipher the code hidden in genomes.

In fact, all the genome projects are aimed to understand the complex organisms in terms of their constituent components, more specifically genes. It is hypothesized that the mechanism behind all the complex processes of life are hidden behind the interactions among the genes. In experimental reality, the endeavor from organism's cell to DNA sequence results from the radical reduction from higher level to lower level representation. Without questioning the amount of information lost due to such reduction it is commonly accepted that all the necessary information for building and working of cells are encoded in the genomic sequences.

Therefore, determination of the complete genome sequence simply marks the beginning of a new chapter where the reconstruction process will be attempted from

the molecular components. Consequently, the focus is now shifting to the accurate annotation of genomic sequences, to the interplay between genes and proteins, and to the genetic variability of species. The genome annotation process is increasingly based on comparative approaches involving evolutionary considerations and model organisms. The interplay between DNA and proteins is the most fundamental of biological interactions and has pervasive implications in biology, medicine, and pharmacology. Genetic variability is the source of phenotypic variation, pathogen susceptibility, environmental factor susceptibility, and individual differences in drug response.

1.1 Functional Genomics in Post Genome Era

Understanding the function of genes and other parts of the genome is the task of functional genomics. The genome projects are just the first step in understanding organisms at the molecular level. After the sequencing phase is complete, the work to determine the function of the identified genes begins. Efficient interpretation of the functions of all genes and other DNA sequences in a genome requires that resources and strategies be developed to enable large-scale investigations across whole genome [50]. A technically challenging first priority is to generate complete sets of full-length cDNA clones and sequences for model-organism genes. Other functional-genomics goals include studies into gene expression and control, creation of mutations that cause loss or alteration of function in organisms, and development of experimental and computational methods for protein analyses.

Understanding, not only the function of each gene in isolation but the complexity of functional networks and control systems is of particular importance for discovery of novel and valid drug targets. Answering complex biological questions in this context necessitates high- throughput gene functional characterization using an array of genomic, proteomic and in silico- based tools and technologies.

With the technological advances, as the genomic research keep generating enormous amount of data, the integration of informatics in the field has become indispensable. Computational Biology, Bioinformatics, Systems Biology are terms for interdisciplinary fields joining information technology, biology, medicine and engineering that has skyrocketed in recent years. These fields are located at the interface between the two scientific and technological disciplines that can be argued to drive a significant if not the dominating part of contemporary innovation.

The goal of these fields is to provide computer-based methods for coping with

and interpreting the genomic data that are being uncovered in large volumes within the diverse genome sequencing projects and other new experimental technology in molecular biology. The field presents one of the grand challenges of our times. It has a large basic research aspect, since we cannot claim to be close to understanding biological systems on an organism or even cellular level. At the same time, the field is faced with a strong demand for immediate solutions, because the genomic data that are being uncovered encode many biological insights whose deciphering can be the basis for dramatic scientific and economical success[59]. With the pre-genomic era that was characterized by the effort to sequence the human genome just being completed, we are entering the post-genomic era that concentrates on harvesting the fruits hidden in the genomic text. In contrast to the pre-genomic era which, from the announcement of the quest to sequence the human genome to its completion, has lasted less than 15 years, the post-genomic era can be expected to last much longer, probably extending over several generations.

However, the role of informatics in genomic research has significantly changed from pre-genomic era to genomic era. The pre-genomics informatics or the genome informatics was developed to manage the huge volume of data generated by the genome projects. Its primary role was to support experimental projects. In contrast, the post-genome informatics will focus on knowledge discovery through computational and statistical analysis of sequence and genetic data and the mathematical modeling of complex biological interactions, which are critical to the accurate annotation of genomic sequences, the study of the interplay between genes and proteins, and the study of the genetic variability of species.

1.2 Reconstructing Gene Regulatory Networks

Starting with genomic sequences, the past few years have provided gene expression data on the basis of ESTs (expressed sequence tags) and DNA microarrays (DNA chips). These data have given rise to a very active new subfield of computational biology called expression data analysis. These data go beyond a generic view on the genome and are able to distiniguish between gene populations in different tissues of the same organism and in different states of cells belonging to the same tissue. For the first time, this affords a cell-wide view of the metabolic and regulatory processes under different conditions. Therefore, these data are believed to be an effective basis for new revealing the mechanism behind gene regulation and protein interaction.

The advent of novel, cutting-edge technologies permit the rapid and parallel

measurement of gene expression as either time series or steady-state data. Monitoring transcriptomes on a genome-wide scale, scientists are forming global views of the structural and dynamic changes in genome activity during different phases in a cell's development and following exposure to external agents. Interpretation of this vast amount of experimental data not only capable of providing comprehensive understanding of the activity of a particular gene in a specific biochemical process, but can facilitate greater understanding of the regulatory architecture also. Therefore, with the availability of these massive amounts of biological data the researchers are trying to unravel the underlying transcriptional regulations in gene circuits using model-based identification methods.

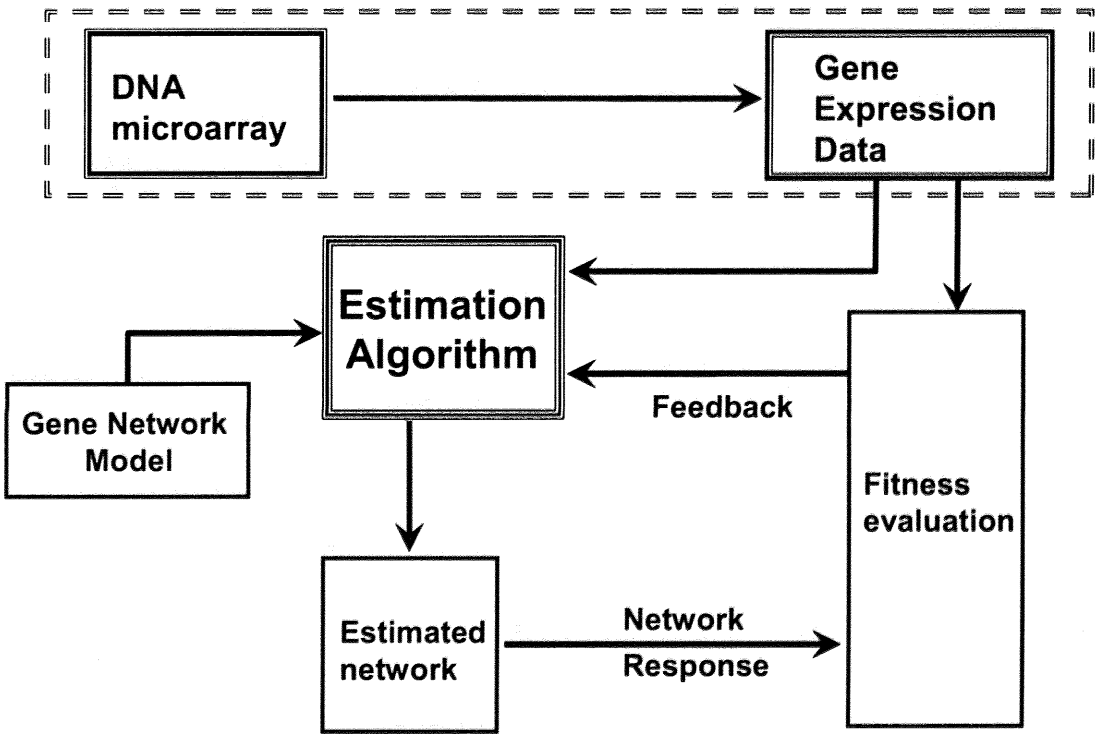

Figure 1.1: Model Based Estimation of Gene Regulatory Networks

Given a dynamic model of gene interactions, the problem of gene network inference is equivalent to learning the structural and functional parameters from the time series representing the gene expression kinetics, i.e. the network architecture is reverse engineered from its activity profiles. It is often wondered whether it is at all possible to reverse engineer a genetic network from gene expression data. Though reverse engineering is possible in principle, the success depends on the characteristics of model involved, the availability of the gene expression data and the level of noise present in the data.

Reverse engineering, of an extremely complex system like genetic networks, needs the use of a reliable, robust and expert methodology. Moreover, the poor understanding of the molecular constituents, limited availability of the information, poor and limited amount of dynamic responses make the reconstruction task even more difficult. Evolutionary Algorithms (EAs) have established them as a suitable approach for working in such an environment and hence the field of genetic network inference has seen a surge of application of EAs. EAs are population based search technique that look for the optimum solution of a problem through repeated creation of new solutions and refining of them. A schematic diagram of the EA based gene network reconstruction is shown in Fig. 1.1.

As shown in Fig. 1.1, the other components involved in the process are *model* of genetic regulation, *evaluation* criteria for candidate networks and the *gene expression data*. Each of these components is of fundamental importance in this research work. Therefore, each of them will be discussed with great detail in different chapters.

1.3 Objectives of the Research

The ultimate goal of the current investigation is to develop an effective and efficient evolutionary methodology for automatic reconstruction of gene regulatory networks from the gene expression data. The challenge consists of many pivotal objectives which may be states as follows

- To study the fundamentals and the principles of gene regulation mechanism and different approaches for modeling the genetic regulation with their basic characteristics, requirements and implementation parameters.

- To make a comparative assessment among the existing algorithms for reconstructing gene regulatory networks using S-system model in terms of their efficiency, scalability and the accuracy of the inferred network structure and parameter values.

- To enhance and accelerate an evolutionary optimizer to work reliably in a highly multimodal and deceptive environment like the problem in hand.

- To modify and/or improve the existing criteria for evaluating the candidate network model to facilitate the identification of skeletal architecture of biological networks using an evolutionary algorithm.

- To design and implement an efficient, robust and reliable evolutionary algorithm for estimating the gene regulatory network architecture and kinetic parameter values.

- To study the performance of the new method applying into the problem of gene network inference both from synthetic and real gene expression data sets and validate the efficiency of the proposal comparing with the existing techniques.

1.4 Scope and Methodology of the Research

In an attempt to make a critical comparison among the different modeling approaches for gene regulatory networks a thorough survey of the related literature was made and the relative advantages and disadvantages were analyzed. Based on this investigation the decoupled form of the S-system formalism [112] was chosen as the fundamental model for this study. The S-system model offers an excellent compromise between accuracy and mathematical flexibility.

Then a detailed study was carried out on different reconstruction processes using the S-system model for pinpointing the shortcomings. It was found that due to the model flexibility the reconstruction algorithms often converge to some local minima and fail to identify the sparse network architecture which is the hallmark of biological networks. Therefore, some modifications and/or extensions of the existing fitness evaluation criteria were suggested in order to limit the search space and assist the search process thereby.

Then as the first step of designing an evolutionary algorithm for gene network estimation, a survey was carried out on the existing evolutionary optimizers. Among the existing algorithms Differential Evolution (DE) [123] was found promising with reliable and robust performance for real world applications. But the convergence velocity of the algorithm still can not meet the requirements for expensive function optimization like the case under study. Therefore, an attempt was made to accelerate the optimizer using local search heuristics and improve the performance of the algorithm.

Then a memetic algorithm was designed using the modified DE as the core optimizer for identifying not only the network structure correctly but also for estimating the parameter values precisely. During the design several issues were taken in consideration, such as: double optimization applied for selecting robust parameter values, hill-climbing local search included for accelerating the identification of the skeletal network topology and a mutation-phase embedded to maintain the diversity in the

population for finding the global optimal solution.

Finally, the developed algorithm was validated by applying it in different gene network reconstruction problems. Artificial networks of different dimensions and characteristics were created and were simulated for generating synthetic gene expression data and these artificial gene expression data were used for reconstructing the underlying gene circuits. Capability of the methodology was verified introducing different levels of noise in the expression profile and varying the amount of expression data used for reconstruction. Comparisons with existing algorithms and previous fitness criteria were carried out and analysis of some real microarray data was also performed.

1.5 Layout of the Thesis

The research work conducted for the achievement of the stated objectives is presented in this dissertation in several chapters organized in a way that the steps involved in the study may properly delineate the methodology. A brief description of the contents of each chapter is as follows:

The introductory concepts of this research work such as background, objectives, scope, methodology and an outline of the thesis are presented in this chapter. **Chapter 2** lays a tutorial on the modeling concepts of gene regulation and introduces many modeling approaches for genetic networks.

Chapter 3 introduces the S-system formalism in detail which is the fundamental model of this study. Both the canonical form and the decoupled forms of the model are discussed and then different model evaluation criteria for the model are analyzed. Then two new evaluation criteria are presented for the evaluation of the candidate network models and these proposals were based on [88, 90]. Finally a brief review of the existing reconstruction algorithm is presented.

Chapter 4 attempts to design an evolutionary optimizer with increased velocity for use in the algorithm of genetic network inference. A crossover based local search strategy called *Fittest Individual Refinement* (FIR), is used for increasing the convergence velocity and robustness of Differential Evolution (DE). The proposed memetic version of DE (augmented by FIR) is expected to obtain an acceptable solution with a lower number of fitness evaluations particularly for higher dimensional functions. Using two different implementations, DEfirDE and DEfirSPX, it was shown, the proposed FIR enhances the convergence rate and robustness of DE for well known benchmark functions, random problems from landscape generator

and real world problems. Contents of this chapter appears as part of [85, 86].

Chapter 5 extends the concepts of previous chapter by introducing the adaptability in the local search. Determining a single local search length that can serve for a wide range of problems is a critical issue. In this chapter, a local search technique is proposed to solve this problem by adaptively adjusting the length of the search, using a hill-climbing heuristic. The emphasis of this paper is to demonstrate how this local-search scheme can improve the performance of DE. Experimenting with a wide range of benchmark functions, it was shown that the proposed new version of DE, with the adaptive local search, performs better, or at least comparably, to classic DE algorithm. Performance comparisons with other local-search heuristics and with some other well known evolutionary algorithms from literature are also presented. This chapter is based on [83].

Chapter 6 presents the inference algorithm developed for the inference of the gene network. After the detail description of the algorithm it was applied for reconstructing the network structure and estimating the model parameters. The performance of the algorithm was analyzed with different artificial networks, with different noise levels and with different amount of expression profiles for reconstruction. The contents of this chapter are based on the work in [84, 92]. **Chapter 7** validates the effectiveness of the methodology by applying it for analyzing the real microarray data. The memetic algorithm was employed for predicting the interaction among the genes in SOS DNA repair network in *Escherichia coli*. The results of this chapter make the partial content of [89].

In **Chapter 8** the newly proposed algorithm was employed in the experiment of inferring regulations in cell-cycle network of budding yeast. The results of this chapter are presented in [84].

Chapter 9 is devoted for overall discussion of the experimental results. It also presents some empirical analysis of the reconstruction algorithm and compares it with other algorithms. **Chapter 10** summarizes the general conclusions from this study and also identifies the topics which warrants further investigation.

Chapter 2

Genetic Networks: In Vivo, In Silico

The biological process of gene expression is a rich and complex set of events that leads from DNA through many intermediates to functioning proteins. This chapter presents the notion of regulation of gene expression which is behind the concept of genetic networks both from the molecular biological background and mathematical background. Then different approaches for modeling gene regulatory networks are presented.

Cells use some mechanism to manage the information contained in their DNA which is the blue print to construct the whole organism. Management systems are necessary because of the immense volume of information in the genome and because different cell types require different information at different times in their development. The DNA in the human genome contains about 30,000 genes, and if all the genes were expressed equally, then all cells would have been the same. It is estimated that a typical higher eukaryotic cell expresses 10~20% of the total gene complement of the cell. Each particular cell type has unique function(s), structure and enzyme(s) and thus needs to express only certain genes and not the others. Consequently, complex organisms have evolved mechanisms to regulate the expression of genes so that a unique, specific set of genes is expressed in each cell type. The general phenomenon of specific gene expression in a certain cell type or during a specific stage of development is known as *differential gene expression* [10]. The mechanisms controlling differential gene expression are a critical part of each species' genome.

2.1 Gene Expression is Regulated at Different Levels

Many of the regulatory mechanisms for controlling gene expression in eukaryotes and prokaryotes are similar but there are some very different, in part because of the very different environments in which they live. There are multiple points in the steps between gene expression and protein synthesis at which gene expression can be controlled in both prokaryotic and eukaryotic cells. These points of control can be separated into two general areas: transcriptional control and post-transcriptional control [10].

Transcriptional Control, which is the primary control point, is responsible for the regulation of RNA synthesis from a DNA template. There is a large set of sequence-specific DNA binding proteins in all cells that controls the transcription of RNA by turning genes on or off. DNA binding proteins possess precise structures that recognize and bind to specific DNA sequences.

Post-transcriptional controls are secondary mechanisms for controlling gene expression after transcription. These types of controls encompass (1) RNA processing control (2) translational control (3) mRNA degradation control and (4) protein activity control. RNA processing control, only applicable to eukaryotes, determines how and when the primary transcript is spliced or otherwise processed to form a usable mRNA. Translational control determines the time and particular type of mRNA that to be translated into proteins. The stability of certain mRNA types are managed by mRNA degradation control. Protein activity control selectively activates, inactivates, modifies or compartmentalizes specific protein molecules within the cell or within a certain cell type, thereby affecting how and when the proteins act.

Among all of these regulatory mechanisms, transcriptional regulation is the most critical control over gene expression having the greatest impact on the biochemical properties of the cell [16]. Therefore, the main focus of this study is limited to the modeling and inferring the transcriptional regulations from gene expression data.

2.2 Basics of Transcription

Transcription is a complicated set of events by which an RNA copy of one of the strands in the DNA double helix is made. The anti-sense strand of the DNA directs the synthesis of a complementary RNA molecule. The RNA molecule produced is

therefore identical to the sense strand of the DNA - except that it contains U (uracil) instead of T (thymine).

Consider a given gene, as shown in Fig. 2.1(a). The gene consists of a *protein coding sequence*, which might be contiguous or broken up into a series of exons and introns, and which begins with a START codon (ATG) and concludes with a STOP codon (TAA, TAG or TGA). This coding region is the part of the gene that will be transcribed into mRNA and translated into a finished protein. Apart from this, a gene must have *regulatory sequences* associated with it that contributes the control of the gene. These are stretches of DNA which do not themselves code for protein but which act as binding sites for RNA polymerase (RNAP) and its accessory molecules as well as a variety of transcription factors (TF). Together, the regulatory sequences with their bound proteins act as molecular switches that determine the activity state of the gene - e.g. OFF or FULL-ON or, more often, something in between. The regulatory sequences include the *promoter* region together with *enhancer* elements.

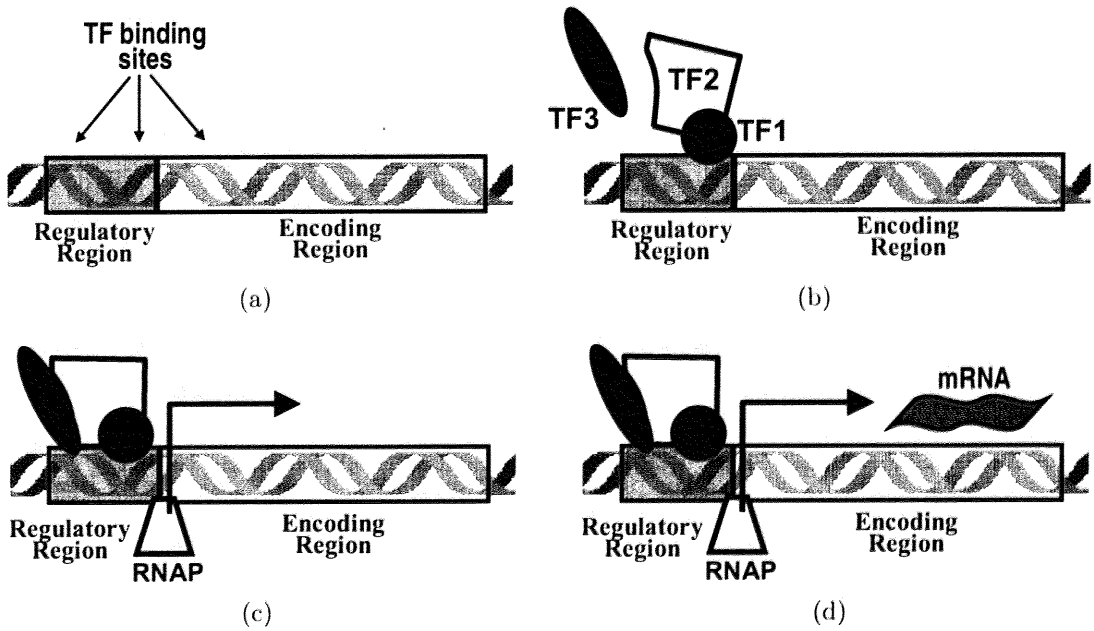

Figure 2.1: Gene Transcription (a) transcription factor binding (b) formation of transcriptional complex (c) RNAP binding and (d) transcription initiation (adapted from [14])

In simple prokaryotes, the regulatory region is typically short (10-100 bases) and contains binding sites for a small number of TFs. In eukaryotes, the regulatory region can be very long (up to 10,000 or 100,000 bases), and contains binding sites of multiple TFs. TFs may act either positively or negatively; that is, an increase in

the amount of transcription factor may lead to either more or less gene expression, respectively [14].

Typically, TFs do not bind singly, but in complexes as shown in Fig. 2.1(c). Once bound to the DNA, the TF complex allows RNAP to bind to the DNA upstream of the coding region, called *promoter*. RNAP forms a transcriptional complex that separates the two strands of DNA, thus forming an open complex, then moves along one stand of the DNA, step by step, and transcribes the coding region into mRNA. Like DNA replication, transcription occurs in three phases - initiation, elongation and termination. Initiation of transcription usually occurs to the 3′ side of the promoter, and termination occurs at specific sites downstream of the coding sequence of gene.

There are fundamental differences in the ways in which genes are transcribed in prokaryotes and eukaryotes. Most protein coding genes in prokaryotes are transcriptionally active by default, i.e. in the absence of other factors, RNAP can bind to the promoter of a gene produce RNA. Transcriptional control is brought to bear on the gene by repressor proteins that occludes RNAP binding or prevents a bound RNAP from transcribing by binding itself to regulatory region of the gene. On the other hand the eukaryotic genes are transcriptionally inactive because eukaryotic RNAPs are unable to recognize promoter sequences themselves [104].

2.3 *cis*-acting vs *trans*-acting

'cis' and *'trans'* are two important terms relevant to the study of gene regulation. The transcription initiation complex is composed of promoter sequences and DNA binding proteins. These two components of transcription are normally described as *cis*-acting elements and *trans*-acting factors.

cis-acting elements are the DNA sequences in the vicinity of the structural portion of a gene that are required for gene expression. In other words we can say a locus is *cis*-acting on a second locus if it must be on the same DNA molecule in order to have an effect. The operator is a *cis*-acting element because it works only when physically attached to the gene whose expression it regulates.

Other *cis* elements include enhancer, promoters, terminators, attenuators, translation initiation sites, mRNA splicing signals, mRNA degradation signals, protein localization sequences, and protein degradation tags.

Genes code for diffusible products, meaning they act in *trans* and their ultimate function is not dependent on their location in the genome once they are expressed.

A locus is *trans*-acting if it can affect a second locus even when on a different DNA molecule. Structural genes code for a product that has a structural or enzymatic function while regulatory genes are a subset of structural genes whose function is to regulate the expression of other genes. Regulatory genes include transcription factors, repressors, activators, antiterminators and translational regulatory proteins.

Those factors which bind to consensus module sequences can bind to any promoter that contains the sequence. The binding of multiple factors, for example, multiple *trans*-acting factors each with one of the four properties mentioned above, may be essential for transcription initiation. Enhancers, which normally have a consensus 72 bp repeat sequence, have sites for multiple *trans*- acting factors to bind. Thus genes with enhancers may require several complexes to be constructed for gene expression to be initiated.

To a molecular biologist, a *cis*-acting regulatory element is usually a target site for a DNA-binding protein, upstream of the gene whose expression is being regulated. A *trans*-acting element is the regulatory protein itself, which can diffuse through the cell from its site of synthesis to its DNA-binding site [16].

2.4 Positive and Negative mode of Control

The molecular mechanism of regulation usually classified in to two broad categories: *negative regulation* and *positive regulation* [34].

In a negative control system the default status is "on" and transcription takes place until it is switched off by a *repressor* protein. Typically a repressor protein either binds to DNA to prevent RNA polymerase from initiating transcription, or binds to mRNA to prevent a ribosome from initiating translation.

In contrast to negative regulation the default status of a positively regulated system is "off" and expression is possible only when an active regulatory protein is present. The regulatory protein interferes with DNA and with RNA polymerase to assist the initiation event. Such a regulatory protein is called *activator* protein.

A negatively (positively) regulated system may be either *inducible* or *repressible* depending on how the active repressor (active activator) is formed. Inducible systems function only in the presence of the small-molecule *inducer*. Repressible systems function only in the absence of the small-molecule *co-repressor*. In other words we can say induction is achieved when an inducer inactivates a repressor protein or activates an activator protein. And repression is accomplished when a co-repressor activates a repressor protein or inactivates an activator protein. Fig.

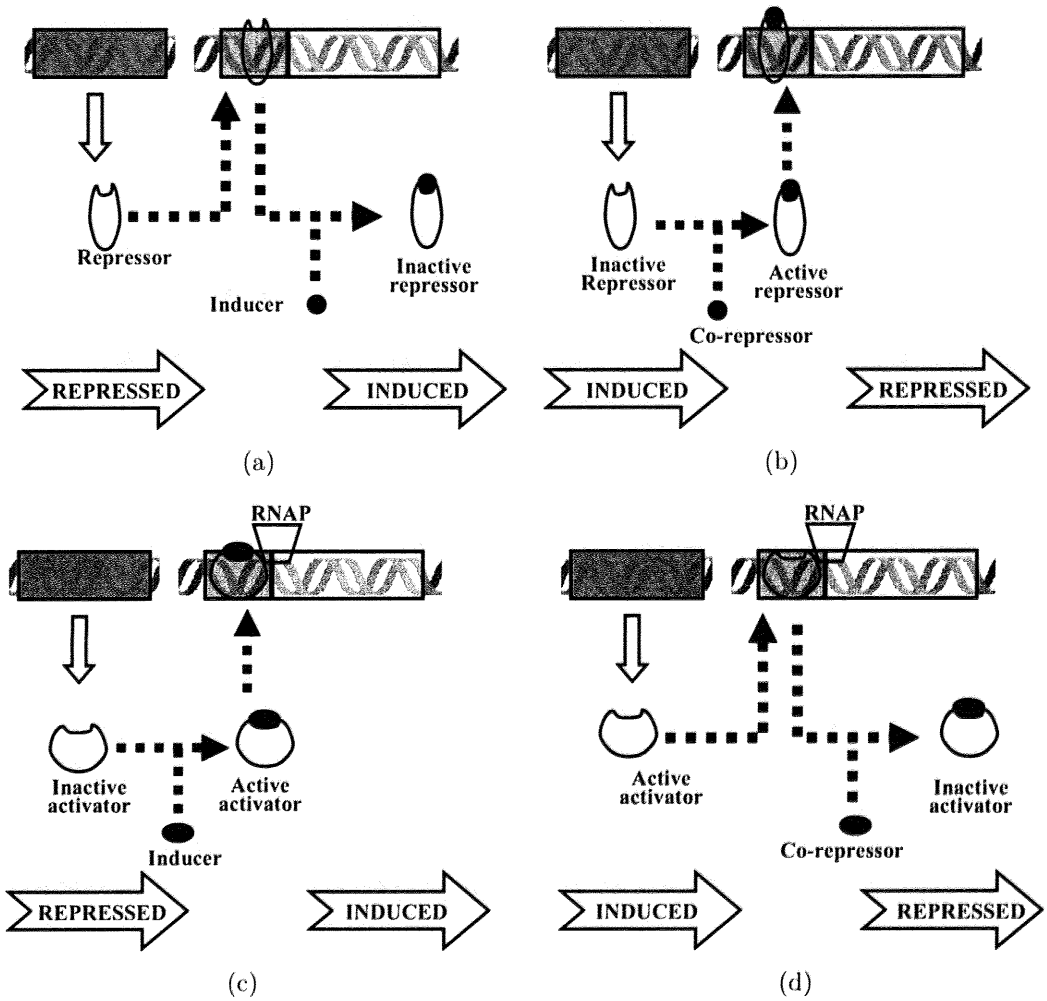

Figure 2.2: Control Circuits (a) Negative control of induction (b) Negative control of repression (c) Positive control of induction (d) Positive control of repression (adapted from [62])

2.2 shows induction and repression under positive and negative mode of controls.

2.5 Genetic Network

Gene regulation, the core of many biological processes, may be thought of as a combination of *cis*-acting regulation by the extended promoter of a gene, and of *trans*-acting regulation by the transcription factor products of other genes. If we simplify the *cis*-action by using a phenomenological model that can be tuned to data, then the full *trans*-acting interaction between multiple genes can be modeled as a network which is commonly known as *gene circuits*, *gene regulatory network* or *genetic network* [76]. However, the mechanism behind such biological networks, which are dynamic and highly nonlinear, is excessively complicated. Because of poor understanding of the biological components, their dependencies, interaction and nature of regulation grounded on molecular level, study of such systems had been impeded until very recent.

Several cutting-edge technologies such as DNA microarrays, oligonucleotide chips have opened the door of surveying thousands of genes under hundreds of varying conditions. In order to draw meaningful inference, such data sets may be analyzed using a range of methods with increasing depth. Beginning with cluster analysis and determination of mutual information content, it is possible to capture the control processes shared among genes. However, the ultimate goal of analysis of expression data is the detailed identification of the molecular mechanism of gene regulatory networks. Nevertheless, the success of such analysis efforts largely depends on the breadth, sensitivity and precision of experimental data to accurately identify the underlying biological system.

2.6 Cluster Analysis of Gene Expression Data

With the advent of the microarray technology large amounts of gene expression data are being routinely generated. Analysis of these data offers potential insight into gene function and regulatory mechanism. A key step in the analysis of gene expression data is the detection of groups of genes that manifest similar expression patters. The corresponding algorithmic problem is known as cluster of gene expression data [13].

Many clustering algorithm has been developed for discovering knowledge from gene expression data. Tavazoie *et al.* [128] have used k-means algorithm for cluster-

ing gene expression data of yeast cell cycle gathered by Cho *et al.* [18] and grouped the ORFs into 30 clusters on the basis of their common expression pattern. Tamayo *et al.* used self-organizing maps (SOM), a type of mathematical cluster analysis, for analyzing data collected from yeast and human [127]. Wen *et al.* used the FITCH hierarchical clustering software for grouping mRNA expression of 112 genes during rat central nervous system development [138]. Eisen *et al.* used a system for hierarchical cluster analysis for genome-wide expression data from budding yeast [26]. Ben-Dor *et al.* have proposed a clustering algorithm based on random graph theory and using features of hierarchical clustering and k-means algorithm [13].

Because of the difference in their working principles, different clustering algorithms often generate very different clustering of the genes for the given same data. Therefore, choosing an appropriate clustering algorithm for a data is often difficult because little guidance is available right now. However, some methodologies for assessing the clustering algorithms from their results are being developed [147].

2.7 Modeling Genetic Networks

Cluster analysis of gene expression data can help elucidate the regulation (or co-regulation) of individual genes, which provides valuable information and insights, but often fails to identify system-wide functional properties of a given network [33]. Computational modeling and simulation can provide a much more detailed and better understanding of the regulatory architecture such as network connections, rate constants and biochemical concentrations etc.

Various types of gene-network models have been proposed, which integrate biochemical pathway information and expression data to trace genetic regulatory interactions [51, 113, 8, 24, 71]. The modeling spectrum ranges from abstract Boolean descriptions to detailed Differential Equation based models each having own strengths, weaknesses and domain of applicability. In general, the modeling spectrum varies in terms of details of biochemical interactions incorporated, discrete or continuous gene expression level used, deterministic or stochastic approach applied, etc. [25]. And these criteria define how closely the model can represent genetic interactions. Generally, detailed biochemical modeling is very useful for capturing the precise mechanism in common regulatory pathways. However, as we try to approach from more abstract to more real representation, the complexity of the model increases accordingly. And with the increase in the model complexity the data requirement, data precision and computational effort for learning the model parameters also in-

creases. The rest of this section gives brief introduction to some popular models for gene regulatory networks.

2.7.1 The Boolean Network Model

Boolean networks, introduced by Kauffmann [51] and explored in [117, 2, 66, 44, 61] offer an attractive discrete time, Boolean model for gene expression. In this model each gene is either fully expressed (ON) or not expressed at all (OFF) at any given time. Following Akutsu *et al.* [2] a Boolean network is specified by pair $G(V, F)$, where the $V = v_1, \cdots, v_n$ is a set of nodes representing the genes of the network and the $F = (f_1, \cdots, f_m)$ is a list of Boolean functions. Each node has an associated expression value v_i that is either 1 (for expressed) or 0 (for not expressed). A Boolean function $f_i(v_{i1}, \cdots, v_{ik}) \in F$ with inputs from the specified nodes $v_{i1}, \cdots, v_{ik}$ is assigned to each node which correspond to the genes that influence the expression of the gene associated with the node v_i and f_i represents the exact functional dependence.

Other advantages of these models are lower time complexity of the algorithm for identifying the network from limited amount of expression data. Data requirement for inferring a Boolean genetic network is $\Sigma(2^K(K + log(N)))$, where K is the maximum connectivity of the network and N is the network dimension. And the time complexity of the algorithm for inferring such a N dimensional network from M point expression profile is $O(N^{K+1}M)$.

Though it is evident from the real gene expression data that the gene expression levels tend to be continuous rather than binary, such coarse representation of the gene state has certain advantage in terms of complexity reduction and computation. These models are typically used to obtain a first representation of network organization and dynamics.

2.7.2 Bayesian Network Models

Murphy and Mian [80] and Friedman *et al.* [30] have suggested using Bayesian network models of gene expression networks. The advantages of Bayesian networks models are that 1) they explicitly relate the directed acyclic graph model of the causal relations among the gene expression levels to a statistical hypothesis; 2) they include all of the aforementioned models, and Hidden Markov Models, as special cases; 3) there are already well developed algorithms for searching for Bayesian networks from observational data; 4) they allow for the introduction of a stochastic

element and hidden variables; 5) they allow explicit modeling of the process by which the data are gathered.

A Bayesian network consists of two distinct parts: a directed acyclic graph (DAG or belief-network structure) and a set of parameters for the DAG [55]. The DAG in a Bayesian network can be used to represent the causal relationships among a set of random variables (such as gene expression levels). A DAG represents the causal relations in a given population with a set of vertices V when there is an edge from A to B if and only if A is a direct cause of B relative to V.

There are two main approaches to searching for Bayesian network models. The first approach (as exemplified in the PC algorithm [99]) performs a series of tests of conditional independence on the sample, and uses the results to construct the set of DAGs that most closely implies the results of the tests. The second approach to searching for Bayesian networks assigns a score to each DAG based on the sample data, and searches for the DAG with the highest score. The scores that have been assigned to DAGs for variables that are discrete or distributed normally include posterior probabilities, the Minimum Description Length, and the Bayesian Information Criterion. A variety of methods of search for DAGs with the highest score have been proposed, including hill-climbing, genetic algorithms, and simulated annealing [36, 99].

Many extensions and improvements of the original model, methodology and algorithm have been done since the first proposal [93, 97, 46, 126].

2.7.3 State Space Models

A state-space description of a gene expression dynamic model has been proposed by Wu *et al.*, where gene expression levels are viewed as the observation variables of a cellular system, which in turn are linear combinations of the internal variables of the system [141]. In fact state space models are a class of dynamic Bayesian networks which assume that the observed measurements depend on some hidden state variables which evolve according to Markovian dynamics.

Compared to other competitive models, this model has the following characteristics. First, gene expression profiles are the observation variables rather than the internal state variables. Second, from a biological angle, the model can capture the fact that genes may be regulated by internal regulatory elements. Finally, although it contains two groups of equations (one is a group of difference equations and the other, algebraic equations), the parameters in this model are identifiable from existing microarray gene expression data without any assumptions on the connectivity

degrees of genes and the computational complexity to identify them is simple.

Though this model is very new in the family but has grown much interest among the researchers resulting many studies some of which are found in [142, 140, 102, 64, 148, 144, 37]

2.7.4 Petri Net based Models

Petri Net (PN) is a description method for modeling concurrent systems mainly used so far to model artificial systems such as manufacturing systems and communication protocols. The first attempt to use PNn for modeling biological pathways was made by Reddy *et al.* [103] which gave a method of representation of metabolic pathways. Hofestädt expanded this method to model metabolic networks [38].

Functional Petri Net (FPN) and Hybrid Petri Net (HPN) are the extensions of the PN formalisms that allow the quantitative modeling of regulatory biochemical networks. Kitagawa and Iba used FPN for identifying the topology and the parameters for typical test-case metabolic pathways [58]. In their work, Matsuno *et al.* demonstrated that by using HPNs, it is possible to translate biological facts into HPNs in a natural manner [71].

Recently, by extending the notion of HPN, Fujita *et al.* [31] introduced Hybrid Functional Petri Net (HFPN) in order to give more intuitive and natural modeling method for biological pathways. They have demonstrated that biological pathways can be modeled with some basic HFPN components. And using these components they have modeled fission yeast cell cycle [31] and gene regulation mechanisms of *Drosophila melanogaster* (fruit fly) circadian rhythm [72].

2.7.5 Linear and Non-linear Differential Equations

In the differential equation based approach to modeling gene regulation, the concentration levels of different reactants are assumed to be continuously changing according to differential equations [14]. If a network consists of N genes, an ordinary differential equation model will represent the change in i-th gene using a equation of the form

$$\frac{dX_i}{dt} = f_i(X_1, \cdots, X_N) \tag{2.1}$$

The function f_i describes how the transcription rate of i-th gene is directly affected by that of other genes $1, \cdots, N$. The concentration level of i-th gene will increase, decrease or remain unchanged depending on whether f_i is positive, negative

or zero respectively [17]. Typically, f_i will be positive for some combinations of $X_1, \cdots, X_N$ and will be negative for some others. The biological interpretation of this activity is that in some states the other genes of the network are acting to switch on the i-th gene and in other states they are switching it off.

The linear or additive models of genetic networks [132, 23, 21] evolve from the simplest form of the function f_i, i.e. the linear form

$$\frac{dX_i}{dt} = \sum_{j=1}^{N} a_{ij} X_j + a_0 \tag{2.2}$$

The quantity a_{ij} is the direct interaction effect of gene-j on the transcription rate of gene-i. In particular if the j-th gene does not directly affect the transcription rate of the i-th gene the $a_{ij} = 0$. Typically, in biological networks very few genes interact with a particular gene [9]. Therefore, many of the a_{ij} terms will be zero in Eq. 2.2. In this context, the reconstruction of the network means estimating the parameters a_{ij} that can reproduce the dynamics observed in the gene expression data.

Similar considerations apply for nonlinear models with the exception that the function f_i of Eq. (2.1) is non-linear. A non-linear model is more desirable because naturally occurring gene regulatory networks contains significant nonlinearities [35]. In reality the concentration level for a gene and its transcription rate must eventually saturate at some maximum - which is sufficient for being the function f_i nonlinear. However, reconstructing f_i may involve estimating a much larger number of parameters for the non-linear models and hence will need a larger volume of data [121]. Among the non-linear models for genetic networks, S-system [112] is a very flexible and popular one which offers an excellent compromise between accuracy and mathematical flexibility. Since this model has been used for reconstructing gene circuits throughout the study, the details of the model is presented in the next chapter. Many other models for gene networks have been proposed using special cases of the rate equation of (2.1) such as *piecewise-linear differential equation (PLDE)*, *qualitative differential equation (QDE)*, *partial differential equation (PDE)*. A nice review of many such models can be found in [20].

2.7.6 Other Modeling Approaches

Besides, many other models are available for gene regulatory networks and many new are coming out. These models range in a wide continuum which contains many other modeling frameworks such as stochastic models, spatial models, particle-based models etc. There is also a large number of modeling approaches which combine

aspects of different approaches described earlier. These intermediate models can be classified as hybrid models. Since real genetic networks are subject to considerable noise they should be modeled using stochastic differential equations [121]. But it should be kept in mind that though in-depth biochemical models are very useful in representing the precise interactions in the gene circuits, but their complexity and the currently available gene expression data restrict their application to very small systems.

In this chapter the major modeling approaches to genetic regulation are reviewed. In the next chapter the focus will be on the specific model that is used in this study for inferring genetic networks.

Chapter 3

Reconstructing Genetic Network - EA Approach

This chapter describes the art of the reverse engineering gene regulatory networks using evolutionary algorithms (EAs). First a detail review of the chosen model, the S-system, is presented. Then the EA based reconstruction theme is reviewed with different criteria used for S-system based model evaluation. Finally, new criteria for model evaluation are presented which are used for inferring the genetic networks in the subsequent chapters.

3.1 Biochemical Systems Theory (BST)

Biochemical Systems Theory (BST) [110, 111] is the mathematical basis of well-established methodological framework for analyzing networks of biochemical reactions and provides a general framework for modeling and analyzing nonlinear systems of genetic networks. It is based on the generic approximation of kinetic rate laws with multivariate power-law functions. This representation results from Taylor's theorem in logarithmic coordinates. As BST provides straightforward recipes for setting up model equations from kinetic and regulatory information and for analyzing them, this type of mathematical approach is called *canonical modeling.* BST and canonical modeling have been discussed and reviewed numerous times [112, 114, 134] which allow a minimized description of the mathematical background and theoretical aspects of the analysis.

In generic terms, canonical models are constructed as follows. Each metabolite that changes over time is represented by a dependent variable, whose concentration or value at every given time point is governed by an ordinary differential equation.

This equation relates the dynamic changes in the metabolite to influxes and effluxes and accounts for all constituents of the system that directly influence the fluxes, for instance, as substrates or through inhibition.

$$\frac{dX_i}{dt} = \sum_{j=1}^{N+M} V_{ji} - \sum_{j=1}^{N+M} V_{ij} \qquad (i = 1, 2, \cdots, N) \tag{3.1}$$

where an individual metabolite concentration is represented by X_i, an individual flux from X_i to X_j is represented by V_{ij}, and the numbers of dependent and independent concentration variables are given by N and M. Exact mathematical formulations for the fluxes are unknown, but a large body of evidence demonstrates that products of power law functions are often valid and effective representations. This type of representation is compatible with the observations of biological systems and has been proven to be capable of describing biological systems adequately [134]. Two most important variants within BST are the Generalized Mass Action (GMA)-system representation and the Synergistic (S)-system representation.

3.1.1 GMA-system Representation

In this representation within the power-law formalism elementary fluxes are grouped into aggregate fluxes that pass through reactions with rate laws given by V_{ik}^{+} and V_{ik}^{-} and Eq. (3.1) can then be written as

$$\frac{dX_i}{dt} = \sum_{k=1}^{e_i} V_{ik}^{+} - \sum_{k=1}^{l_i} V_{ik}^{-} \qquad (i = 1, 2, \cdots, N) \tag{3.2}$$

where the numbers of reactions entering and leaving the pool X_i are given by e_i and l_i. If each of these rate laws is represented in the power-law formalism, then the standard form of the GMA-system of equations becomes

$$\frac{dX_i}{dt} = \sum_{k=1}^{e_i} \alpha_{ik} \prod_{j=1}^{N+M} X_j^{g_{ijk}} - \sum_{k=1}^{l_i} \beta_{ik} \prod_{j=1}^{N+M} X_j^{h_{ijk}} \qquad (i = 1, 2, \cdots, N) \tag{3.3}$$

where the first N metabolite concentrations are dependent variables and the last M are independent variables. The *kinetic orders* in these equations are given by

$$g_{ijk} = \left(\frac{\partial V_{ik}^{+}}{\partial X_j}\right)_0 \left(\frac{X_j}{V_{ik}^{+}}\right)_0 \quad \text{and} \quad h_{ijk} = \left(\frac{\partial V_{ik}^{-}}{\partial X_j}\right)_0 \left(\frac{X_j}{V_{ik}^{-}}\right)_0$$

the *rate constants* are given by

$$\alpha_{ik} = (V_{ik}^{+})_0 \prod_{j=1}^{N+M} (X_{j0})^{-g_{ijk}} \quad \text{and} \quad \beta_{ik} = (V_{ik}^{-})_0 \prod_{j=1}^{N+M} (X_{j0})^{-h_{ijk}}$$

and the subscript "0" indicates that the results are evaluated at the nominal steady state. The kinetic orders g_{ijk} and h_{ijk} are real numbers, whereas the metabolite concentrations X_i, the aggregate fluxes V_{ik}^{+} and V_{ik}^{-}, and the rate constants α_{ik} and β_{ik} are non-negative real numbers. One can use the irreversible strategy of aggregation, but only if the aggregate fluxes are always positive. This representation becomes inappropriate if any of the aggregate fluxes goes to zero or changes direction.

The number of parameters involved in a gene network modeled by GMA-system model is $(e_i + l_i)\mathcal{N}(\mathcal{N}+1)$ where $\mathcal{N} = N + M$. The value of $\mathcal{N}$ can be estimated from the system and the environment, but in most system the values for e_i and l_i is very critical to estimate. If we consider the value of e_i and f_i in the order of $O(N)$, the number of GMA-system parameter becomes in the order of $O(n^3)$ which is too high for application of GMA-system model for optimizing real biological systems.

3.1.2 The S-system Model

The Synergistic (S)-system model [112] is actually a more computationally and analytically tractable specialization of GMA-system model. For this representation within the power-law formalism elementary fluxes are grouped into aggregate fluxes that pass into and out of metabolic pools. These aggregate fluxes have rate laws given by V_{+i} and V_{-i} and Eq. (3.1) can then be written as

$$\frac{dX_i}{dt} = V_{+i} - V_{-i} \qquad (i = 1, 2, \cdots, N) \tag{3.4}$$

If each of these rate laws is represented in the power-law formalism, then the canonical representation for the S-system of equations is given by

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{N+M} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{N+M} X_j^{h_{ij}} \qquad (i = 1, 2, \cdots, N) \tag{3.5}$$

where X_i is the concentration level of i-th metabolite. The first N metabolite concentrations are dependent variables and the last M are independent variables. The terms g_{ij} and h_{ij} represent interactive affectivity of metabolite-j to metabolite-i. For each dependent concentration X_i in this biochemical model there exists an aggregate production function and an aggregate consumption function. The first

term in right-hand side of (3.5) represents all influences that increase X_i, whereas the second term represents all influences that decrease X_i.

The signs of g_{ij} and h_{ij} determine the network structure. An exponent of zero for any X_j means that that variable has no direct influence on the rate of the corresponding aggregate process, a positive exponent means that they are positively correlated, and a negative exponent means that they are negatively correlated. More specifically, if g_{ij} is positive, metabolite j induces the synthesis of metabolite i. If g_{ij} is negative metabolite j suppresses the synthesis of metabolite i. If g_{ij} is zero, metabolite j has no effect on the synthesis process of metabolite i. Similarly if h_{ij} is positive, negative or zero then the metabolite j induces, suppresses or is irrelevant to the degradation process of substance-i respectively. The parameters that define the S-system are: $\Omega = \{\alpha, \beta, g, h\}$. In a biochemical engineering context, the non-negative parameters α_i , β_i are called *rate constants*, and real-valued exponents g_{ij} and h_{ij} are referred to as *kinetic orders*. It is known that biological networks are sparse, which means the number of regulators that have effect on a single gene is relatively small; so many of the *kinetic orders* are zero in real condition.

The specialization of GMA-system into S-system reduces the number of parameters to $2\mathcal{N}(\mathcal{N}+1)$, where $\mathcal{N}$ is the number of independent and dependent variables. If we assume the system is free of independent variables (i.e. $M = 0$) then the number of parameter becomes $2N(N+1)$ which makes S-system a more tractable model to design, analyze and optimize compared to the GMA-system model.

Since the details of the molecular mechanisms that govern interactions among system components are neither substantially known nor well understood, the description of these processes requires a representation that is general enough to capture the essence of the experimentally observed response. The S-system model is organizationally rich enough to reasonably capture various dynamics and mechanisms that could be present in complex system of genetic regulation. The strength of S-system model is its structure which is rich enough to satisfy these requirements and to capture all relevant dynamics; an observed response (dynamic response) may be monotone or oscillatory, it may contain limit cycles or exhibit deterministic chaos [129]. Furthermore, the simple homogeneous structure of S-system has a great advantage in terms of system analysis and control design, because the structure allows analytical and computational methods to be customized specifically for this structure [47].

However, the problem of reconstructing genetic network using S-system has the difficulty of high-dimensionality, since $2N(N+1)$ S-system parameters must be de-

termined in order to solve the set of differential equations (3.5). And estimation of parameters for a $2N(N+1)$ dimensional function optimization problem often causes bottlenecks and fitting the model to experimentally observed responses (time course of relative state variables or reactants) is never straightforward and is almost always difficult. Therefore, the application of S-system model has been limited to inference of small-scale gene networks only.

3.2 The Challenge of High Dimensionality in S-system

In order to deal with the problem of high-dimensionality, inherent in the S-system model based reconstruction, different decoupling approaches have been applied to the canonical model. This section presents the most popular decoupling techniques.

3.2.1 Decoupling with Linear Programming

Using the idea of linear programming (LP) Akutsu *et al.* [3] developed a simple method called SSYS-1 for inference of S-systems. Assuming $\frac{dx_i(t)}{dt} > 0$ at time t and taking log of each side of $\alpha_i \prod X_j^{g_{ij}}(t) > \beta_i \prod X_j^{h_{ij}}(t)$ the following inequality is obtained

$$\log\alpha_i + \sum_{j=1}^{N} g_{ij}\log X_j(t) > \log\beta_i + \sum_{j=1}^{N} h_{ij}\log X_j(t) \tag{3.6}$$

Since $X_j(t)$ are known data, Eq. (3.6) is a linear inequality, if $\log\alpha_i$ and $\log\beta_i$ are treated as parameters. In case of $\frac{dx_i(t)}{dt} < 0$ a similar inequality is obtained and solving these inequalities using LP the relative ratios of the parameters can be obtained.

The method is faster compared to the methods for solving canonical problem and therefore, can be applied to larger networks with hundreds of nodes. However, the approach can not determine unique parameters and reported to be vulnerable to noise.

3.2.2 Decoupling in Algebraic Equations

Voit and Almedia [133, 4] have shown that a tightly coupled system of non-linear differential equations can be validly decoupled as a set of algebraic equations. For this they substituted the derivatives on the left sides of differential equations with

estimated slopes which are to be estimated directly from the observed data. And then the approximated algebraic equations can be processed efficiently in parallel or sequentially. The estimation of slopes for time series of the metabolites is accomplished with a "universal function" that is computed directly from data by cross-validated training of an artificial neural network (ANN).

Although the ANN-derived "universal function" is obtained directly from metabolic profile data and without a predefined mathematical structure, it may not reflect the "true" underlying function and its resolution may need additional subject area information which is not always available. However, according to the reported results, the combination of methods in this form of decoupling speeds up the inverse problem considerably. But the authors verified their method applying only to very small scale networks [133, 4]. However, according to the reported results, though the method speeds up the reconstruction process its accuracy of the estimation was not very precise.

3.2.3 Decoupling by Problem Decomposition

Maki *et al.* [70] have used a problem decomposition strategy for decoupling the canonical S-system model and facilitating its application to larger gene network inference problem. Using the suggested decomposition strategy the original optimization problem is divided into N sub-problems [70, 56]. In each of these sub-problems the parameter values of gene i are estimated for realizing the temporal profile of gene expression. In other words, this disassociation technique divides a *2N(N+1)* dimensional optimization problem into N sub-problems of *2(N+1)* dimension. In i-th sub-problem for gene i $X_i^{cal}(t)$ is calculated by solving the following differential equation instead

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{N} Y_j^{g_{ij}} - \beta_i \prod_{j=1}^{N} Y_j^{h_{ij}} \tag{3.7}$$

For solving the differential equation (3.7) we need the concentration levels Y_j ($j = 1, \cdots, N$). In i-th sub-problem corresponding to gene i the concentration level $Y_{j=i}$ is obtained by solving the differential equation whereas the other expression levels $Y_{j\neq i}$ to be estimated directly from the observed time-series data. The optimization task for the tightly coupled S-system model is not trivial because Eq. (3.5) is non-linear in all relevant cases, thus requiring iterative optimization in a larger parameter space, where 95% of the total optimization time is expended in numerical integration of the differential equations [133]. Therefore, such disassociation could

be very useful in reducing the computational burden which will become clear from its applications presented in subsequent chapters. Moreover, the experimental results showed their usefulness in estimating the network parameters [70, 57]. In this work this particular decoupled form of S-system model has been used and any subsequent reference to decoupled form will indicate this particular form of S-system. And for direct estimation of expression levels $Y_{j\neq i}$ linear spline interpolation (described in Appendix B) [100, 28] was chosen.

3.3 Parameter Estimation by Reverse Engineering

After selecting a model for representing gene regulatory systems it is time to turn towards the problem of system identification. The methodology of estimating the interacting mechanisms among system components by using experimentally observed dynamic responses of the system is generally known as "inverse problem". This involves two sub-problems: (1) Writing equations with unknown parameters and (2) Estimating the values of those unknown parameters that generates system response closest to the observed dynamics. Since the equations for all bio-chemical reactions are obtained from the model description, we need only to consider the second sub-problem.

The general problem of parameter estimation is hugely important across different scientific disciplines. Different communities have different names for parts of the process, different problems with which they are concerned, and different techniques. However, this study is particularly focused on the art of reverse engineering.

Reverse engineering, as the name implies, is the reverse of engineering; in other words, the attempt to recapture the top level specification by analyzing the product - "attempt" because it is not possible in practice, or even in theory, to recover everything in the original specification purely by studying the product. Reverse engineering can be viewed as the process of analyzing a system to: (1) Identify the system's components and their interrelationships, (2) Create representations of the system in another form or a higher level of abstraction and (3) Create the physical representation of that system. Reverse engineering has a long history of use in different fields of engineering and recently it is being widely used in biochemical engineering. Perhaps because of the poor understanding of the biological components at molecular level, inadequate knowledge about the types of their dependencies and scare information about the nature of their interactions reverse engineering has

become the most feasible option for learning the parameters for such systems.

Several techniques exist for training parameters from trajectory (behavior) data which is often formulated as an optimization problem. Among the approximation schemes for non-linear systems a few deserve special mention.

Neural Networks

For feed-forward neural networks with two-way connections or higher-order ones, one can develop efficient gradient descent algorithms for training the network to match a training set of input/output patterns. The best known of these algorithms is the batch mode version of backpropagation in which an error signal propagates backward through the network layers and alters each connection.

Simulated Annealing

Another form of stochastic gradient descent is simulated annealing. This technique is typically much less efficient, but may still work when other data-fitting optimization procedures become ineffective due to large number of local minima. Simulated annealing can avoid local optimum traps by occasionally taking down-ward steps. Simulated annealing is a very powerful method, although it can be quite difficult to select appropriate parameters.

Evolutionary Computation (EC)

Evolutionary Computation is another class of population based stochastic optimization methods that work reliably for solving complex optimization problems in different fields starting from engineering to medicine. The advantages of this approach include its conceptual simplicity, broad applicability, ability to outperform classical optimization procedures on real-world problems and ease of hybridization with existing methods and data structures such as neural networks, finite state machines and fuzzy systems. Since this work makes extensive use of EC for reconstructing genetic networks much about this approach will be presented in subsequent chapters.

3.4 Genetic Network Inference using S-system

For reverse engineering a genetic network we need to simulate the network for generating the system response. In other words, we can say that we can obtain the gene

expression profile for a S-system model parameter set by simulating the S-system model (3.5) using the set of model parameters. Simulating a genetic network, represented by differential equation, is equivalent to solving the differential equation under given initial condition which is commonly known as initial value problems for differential equations. While there are many general techniques for analytically solving classes of ODEs, the only practical solution technique for complicated equations that can not be solved analytically, is to use numerical methods. Since S-system belongs to that particular class of differential equations without analytical solution we have to satisfy ourselves with an approximation to the solution.

The most popular of these is the Runge-Kutta method, but many others have been developed, including the collocation method and Galerkin method. A vast amount of research and huge numbers of publications have been devoted to the numerical solution of differential equations as a result of their importance in fields as diverse as physics, engineering, economics, and electronics. Among these Runge-Kutta method is the most popular and widely used numerical integration method for differential equations. This simple but powerful integration technique can be found very useful for many precise scientific calculations especially when an adaptive step-size algorithm is combined with it. The general fourth-order Runge-Kutta method was used for numerical integration of the differential equations in this work. Besides Irvine and Savageau have developed ESSYNS (Evaluation and Simulation of Synergistic Systems) for faster integration of canonical S-system [47].

Reconstruction of gene regulatory networks from gene expression profile using S-system is formalized as an optimization problem in which appropriate system parameters of S-system must be found so that the difference between the time course data calculated by the S-system model and the time course data observed in experiments becomes minimum. Tominaga *et al.* [129] used Genetic Algorithm (GA) for searching the set of parameters for S-system that produces the dynamics closest to experimental results. Since methods for finding analytic solution for this problem is almost impracticable, use of *Evolutionary Computation* (EC) has become more feasible and popular method among researchers [7, 54, 77, 56, 119]. Since this work also apply an evolutionary algorithm for reconstructing the genetic networks a review of these and many other contemporary works in the field are presented in a later section.

3.5 Model Evaluation Criteria

For inferring a genetic network a set of optimal parameters for the network model that generates closes response to the observed dynamics is searched. Such a computational problem in which the best of all possible solution is searched for is commonly known as optimization problem. For global optimization of a problem, a set of parameters that minimizes/maximizes a system's desirable properties is searched. The desirable properties to be minimized/maximized are often formulated as a function - commonly known as *objective function*. In this particular problem of genetic network reconstruction we search for the model parameters that minimize the difference between the simulated responses and the experimental observations. While searching for the set of optimal parameter for the target network we need some measure for evaluating different candidate models. Here different fitness criteria found in literature for model evaluation are reviewed briefly.

3.5.1 Mean Squared Error (MSE)

The most commonly used evaluation criterion is the discrepancy between the numerical solution of the differential equation and the observed system dynamics. Tominaga *et al.* gave the sum of mean squared error (MSE), between the resulting gene expression for the estimated parameter set and the measured gene expression, as the fitness evaluation function which should be minimized by Genetic Algorithm (GA) [129]. In other words the fitness of each set of estimated parameters for the target system is evaluated using the following function [129]

$$f^{MSE} = \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \left(\frac{X_i^{cal}(t) - X_i^{exp}(t)}{X_i^{exp}(t)} \right)^2 \right\} \tag{3.8}$$

where $X_i^{cal}(t)$ is gene expression level of gene X_i at time t calculated numerically by solving the system of differential equation of (3.5) for the estimated parameter set, and $X_i^{exp}(t)$ represents the experimentally observed gene expression level of X_i at time t, T is the number of sampling points of the experimental data. In this form of optimization problem the search algorithm tries to find a set of parameters that minimizes f^{MSE}.

In the decoupled form, the mean error for the expression levels of each gene is considered individually for evaluating the candidate set of parameters for that particular gene. In other words, the sum of squared relative errors between experimental and calculated gene expression levels of gene-i is used as the fitness function

in subproblem-i. So the objective function of the subproblem corresponding to i-th gene becomes

$$f_i^{MSE} = \sum_{t=1}^{T} \left\{ \left(\frac{X_i^{cal}(t) - X_i^{exp}(t)}{X_i^{exp}(t)} \right)^2 \right\} \tag{3.9}$$

And in subproblem-i the parameters $\Omega_i = \{\alpha_i, \beta_i, g_{ij}, h_{ij} (j = 1, \cdots, N)\}$ for gene-i that minimizes f_i^{MSE} are estimated.

3.5.2 Akaike's Information Criterion

Information criteria provide a simple method to choose from a range of competing models. Akaike's Information Criteria (AIC) [1] is most commonly used in statistical modeling to show disparity between the true model and the estimated one. Suppose $\varepsilon_i(t)$ is the error between the experimental and calculated expression level of gene-i at instant t, i.e. $\varepsilon_i(t) = (X_i^{cal}(t) - X_i^{exp}(t))$. If we assume $\varepsilon_i(t)$ is normally distributed with mean $\mu = 0$ and standard deviation σ, which are constant for all genes and over time, then the probability density function of $\varepsilon_i(t)$ is given by

$$p.d.f_{\varepsilon} = \frac{1}{\sqrt{2\pi\sigma^2}} exp \left\{ - \frac{(X_i^{cal}(t) - X_i^{exp}(t))^2}{2\sigma^2} \right\} \tag{3.10}$$

The log-likelihood Λ of the expression data for a set of model parameters Ω is

$$\Lambda(\Omega, \sigma) = -\frac{1}{2\sigma^2} \sum_{1}^{T} \sum_{i=1}^{N} (X_i^{cal}(t) - X_i^{exp}(t))^2 - \frac{TN}{2} ln(2\pi\sigma^2) \tag{3.11}$$

and the maximum likelihood estimate of σ^2 is obtained from

$$\hat{\sigma}^2 = \frac{1}{TN} \sum_{1}^{T} \sum_{i=1}^{N} (X_i^{cal}(t) - X_i^{exp}(t))^2 \tag{3.12}$$

The log-likelihood of the estimated model is obtained by substituting (3.12) into (3.11).

Different information criteria are formulated as a penalized log-likelihood and particularly AIC is defined as [1]

$$AIC = -2\Lambda + 2\Phi \tag{3.13}$$

where Φ is the is the number of parameters included in the model. When AIC is used for selecting among the alternative models then the model with lowest AIC

value is chosen. This original form of AIC has been used in [6, 5] for model selection.

3.6 New Fitness Criteria for Skeletal Network Structure

Generally, very few genes or proteins interact with a particular gene in biological networks [9]. But one major difficulty in the S-system based network inference process is detecting the skeletal system architecture that generates the experimentally observed dynamics. Because of the high degree-of-freedom of the model, there exist many local minima in the search space that mimic the time-courses very closely. Therefore, any method attempting to reproduce the time dynamics only, often gets stuck to some local optimum solution and fails to obtain the skeletal structure [54].

One of the main objectives of this work is to design a fitness evaluation function that can effectively evaluate the candidate network models considering their skeletal structures. This section presents the new fitness evaluation criteria proposed in this work for identifying the sparse network structure which is most common in biological network.

3.6.1 MSE based Fitness Criterion for Canonical Model

In literature, there exist some early efforts in this regard. Kikuchi *et al.* suggested to penalize the fitness function by using all the *kinetic orders* (i.e. g_{ij} and h_{ij}) of the network [54] as follows

$$f_i = \sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{X_i^{cal}(t) - X_i^{exp}(t)}{X_i^{exp}(t)}\right\}^2 + cNT\left\{\sum_{i,j}|g_{ij}| + \sum_{i,j,i\neq j}|h_{ij}|\right\} \tag{3.14}$$

where c is the weighted coefficient that balances the two evaluation terms. The first on the right-hand side of Eq. (3.14) is the same as in Eq. (3.8). And the second term, called the *pruning term* is added for identifying the skeletal structures. However, as reported in [54], using the above fitness function, their method could not identify the exact structure of regulations for a small network of five genes from noise-free gene expression data.

Here, a new fitness function is proposed for reverse engineering the canonical

S-system model using a more effective penalty term as follows [87]

$$f_i = \sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\frac{X_i^{cal}(t) - X_i^{exp}(t)}{X_i^{exp}(t)}\right\}^2 + \frac{1}{N}\left\{\sum_{i,j}|g_{ij}| + \sum_{i,j}|h_{ij}|\right\} \tag{3.15}$$

The penalty term of (3.15) will force all the kinetic orders (g_{ij} and h_{ij}) towards zero. Therefore, while searching, the first term (the original fitness function) will try to find a set of parameters which will reproduce the time course, on the other hand the second term will try to find a set of parameters which will minimize it. And because of their join activity search will be directed to the sets of parameters which will have many zero values for g_{ij} and h_{ij}, representing the skeletal structure. In this new fitness function of (3.15) the novelty is the use of the reciprocal of network dimension as coefficient in penalty term. The reason for using this coefficient is to reduce the effect of the penalty term in total fitness as the network dimension grows. As the dimension of the genetic network increases the penalty term as well as the fitness value will also increase. Since we search for the minimum value of the fitness function, the search may be misguided because of the presence of large value of penalty term. Therefore, to balance the effect of the penalty term with the increase of network components the fitness function of (3.15) is proposed.

Kikuchi *et al.* exclude $h_{i\neq j}$ from their fitness function of (3.14) and thus biase it towards identifying $h_{i\neq j}$ parameters. The fitness function proposed here is free from such bias. Moreover, the inclusion of the number to samples (T) in the penalty term of (3.14) seems not to be reasonable because it supposed not to be related with it. Furthermore, the fitness function of (3.15) was successful to identify more correct structure under the same experimental condition [87].

Use of such *pruning term* or *penalty term*, based on Laplacian regularization term, in the basic fitness function of (3.8) was useful for finding a sparse network architecture in the canonical optimization problem [54, 87]. But because of high dimensionality these fitness functions have been applied to small scale networks only.

3.6.2 MSE based Fitness Criterion for Decomposed Model

Based on the same notion, Kimura *et al.* added another more effective *penalty term* to the objective function of (3.9) for obtaining sparse network structure in the

decoupled form of the problem [56, 57]

$$f_i = \sum_{t=1}^{T} \left\{ \frac{X_{k,i}^{cal}(t) - X_{k,i}^{exp}(t)}{X_{k,i}^{exp}(t)} \right\}^2 + c \sum_{j=1}^{N-I} (|G_{ij}| + |H_{ij}|) \tag{3.16}$$

where G_{ij} and H_{ij} are given by rearranging g_{ij} and h_{ij}, respectively, in ascending order of their absolute values (i.e., $|G_{i1}| \leq |G_{i2}| \leq \cdots \leq |G_{iN}|$ and $|H_{i1}| \leq |H_{i2}| \leq \cdots \leq |H_{iN}|$). And I is the maximum allowed cardinality (in-degree) of the network and c is the penalty constant. The superiority of this *penalty term* lies in including the maximum cardinality of the network. And thereby, this *pruning term* will penalize only when the number of genes that directly affect the i-th gene is higher than the maximum allowed in-degree I, thereby will cause most of the genes to disconnect when this penalty term is applied.

However, very few genes affect both activation and repression of a specific gene. Therefore, designing the penalty term considering both synthetic and degradative regulations together rather than separately will be more effective. Because such penalty will penalize whenever total number of regulators (whether synthetic or degradative) is greater than maximum allowed cardinality. Therefore, a further modification to the penalty term of (3.16) is suggested as follows [88]

$$f_i = \sum_{t=1}^{T} \left\{ \frac{X_{k,i}^{cal}(t) - X_{k,i}^{exp}(t)}{X_{k,i}^{exp}(t)} \right\}^2 + c \sum_{j=1}^{2N-I} (|K_{ij}|) \tag{3.17}$$

where K_{ij} are the *kinetic orders* (i.e. g_{ij} and h_{ij}) of gene i sorted in ascending order of their absolute values. Use of (3.17) instead of (3.16) as fitness function can identify the zero valued parameters increasingly and thus obtain the skeletal network structure more precisely [88].

3.6.3 AIC based Fitness Criterion for Decomposed Model

Ando and Iba have shown that AIC can be a useful measure for evaluating candidate models during the evolutionary search in a noisy environment [6, 5]. They estimated the network structure in canonical problems but did not try to estimate the kinetic parameters for the target network. However, in this study it was found that though the original AIC can estimate the network topology, is unable to estimate skeletal network structure and precise network parameters, as will be presented in some later chapter. Moreover, earlier studies [56, 54, 87] have shown the usefulness of penalty

term in identifying skeletal network structure. Therefore, a new AIC based fitness evaluation has been proposed in this work to overcome the weakness of the original AIC in evaluating alternative network models and making selection among them.

The new fitness evaluation criterion extends AIC by adding another penalty term for facilitating the task of identifying the sparse network architecture. The second term of basic AIC is the *penalty term* which penalizes for addition of model parameters. However, many modification or extension of the *penalty term* has been suggested resulting in various modified forms of AIC such as BIC, HQ, MCp, GCV, FPE etc. And such an extension of AIC was attempted for obtaining a network model with sparse connectivity among the components. The proposed fitness evaluation criterion for subproblem-i corresponding to gene-i is as follows

$$f_i^{AIC} = -2\Lambda_i + 2\Phi_i + c\sum_{j=1}^{2N-I}(|K_{ij}|) \tag{3.18}$$

As mentioned in section 3.6.2, this additional penalty term in (3.18) was designed to penalize a model only if the number of regulators included is higher than the maximum allowed for the network. Therefore, as long as the number of regulators is smaller than the maximum in-degree allowed, this additional penalty term will have zero effect in model selection. But it will interfere with the regular AIC fitness function only when the number of genes that directly influence the gene under consideration is higher than maximum allowed in-degree and will assist it in finding a sparse network architecture. This penalty term also introduces another parameter c in the fitness function but the value of this parameter can be chosen in a very easy empirical way as will be explained later. Using the fitness function given in (3.18) both the discrepancy in the expression levels and degree of freedom is considered for model selection as well as the sparse network structure is searched. Furthermore, in experimentation it was found that without this penalty term the pure AIC alone cannot identify the precise skeletal network structure as will be shown in later chapters.

3.7 Evolutionary Reconstruction Algorithms

As mentioned earlier, because of EAs reliable and robust performance, the inference of gene regulatory networks using S-system formalism has seen a surge of application of EAs. This section reviews some of the most prominent evolutionary algorithms that have been developed for solving this problem.

At first, Tominaga *et al.* [129] used a classic genetic algorithm (GA) for the inverse problem of parameter estimation of S-system model. Using uniform crossover and Gaussian mutation with roulette-wheel selection they designed a GA that was capable to estimate 12 parameters with reasonable accuracy among the 60 parameters in a 5 dimensional network in a noise-free environment.

Morishita *et al.* [77] developed the Network Structure Search Evolutionary Algorithm (NSS-EA) for inference of genetic network using S-system. Using a 5 dimensional gene circuit it was shown that NSS-EA can efficiently find multiple network structures that can generate responses similar to target dynamics. However, their method neither attempted to estimate the parameter values nor gave any direction to choose the best among the alternative solutions.

Sakamoto and Iba [107] proposed the use of Genetic Programming (GP) along with Least Mean Square (LMS) for inferring genetic networks. In presence of noise their method exhibited robust performance in estimating the network structure and parameter values for small scale networks.

Kikuchi *et al.* [54] enhanced the initial proposal of Tominaga *et al.* [129] by using a more robust real coded genetic algorithm (RCGA) called PEACE1. Using gradual optimization strategy they were successful to estimate the skeletal structure of a five dimensional network from the noise free time series data. And the inferred parameter values by their method were also pretty accurate. However PEACE1 was found computationally very expensive for applying to larger networks.

Ando *et al.* [7] used an hybrid evolutionary method of GP and statistical analysis for identifying the concise form of regulation between the metabolites from a given set of time series. Although this method may be robust in statistical terms, the algorithm was only tested on small Gene Regulatory Networks (GRNs) (ten genes) and the authors detected important scalability limitations when applied to more complex data.

Iba and Mimura [75] presented an iterative inference approach based on GA whose learning process was guided by a molecular biologist. One of the most important drawbacks of this methodology is that it requires that the biologist to have a good understanding of the dynamics of the GA for selecting leaning parameters.

Ando and Iba [6] used a divide-and-conquer approach for bottom up reconstruction of the genetic network using an assemblage of UNDX+MGG and messy GA. They also made use of available genomic knowledge in their reconstruction procedure. However, they focus on structure identification rather kinetic parameter estimation for the target models.

Speith *et al.* [119, 120] showed in their work that memetic algorithms (MAs) are more suitable for inferring genetic networks compared to standard evolutionary algorithm. Using a local search by evolutionary strategy with in the global search framework of GA, they showed their proposed algorithm is more suitable for optimization and structure identification for genetic networks. However, their methods were successful to identify the exact network, in terms of topology and parameter values, in 10% or less experimental runs from noise free gene expression data [119, 120].

Kimura *et al.* [56] used a memetic algorithm called GLSDC (Genetic Local Search with distance independent Diversity Control) for reconstructing genetic networks using decomposed S-system formalism. Their method, including a parameter estimation technique for initial gene expression level, was able to reconstruct medium scale genetic networks (30 nodes) with accurate parameters from the noise corrupted data. In a follow up work they showed that use of cooperative coevolutionary algorithm can even improve the prediction accuracy [57].

Wang *et al.* [137] used Lamarckian GA for reconstructing genetic networks using a differential equation based model. Their methodology reconstructed the regulation network of 27 yeast cell cycle genes from a real microarray dataset.

Noman and Iba [87] used differential evolution (DE) in gradual optimization framework for reconstructing GRN and identifying skeletal structure. In [88] it was shown that use of local search in the general frame work of evolutionary algorithm can improve the efficiency of the reconstruction algorithm in inferring the model parameters in decoupled formalism. The reconstruction capability of the algorithm was verified using both artificial and real microarray data analysis.

Tsai *et al.* [130] used hybrid differential evolution (HDE) for identifying the model structure and parameters. Using small scale networks they showed that there developed method outperformed PEACE1 and GLSDC. However, their method using the canonical model representation can not be scaled for very large networks and exhibited relatively poor performance in case of noise corrupted data.

Nakatsui *et al.* [81] extended the work of Morishita *et al.* [77] by using an analytical method for extracting common core binomial genetic interaction from different candidate network models. Using small scale artificial network they showed that their method can identify most of the core interaction in a genetic network.

Almost all of these evolutionary reconstruction algorithms tried to optimize some fitness function based on MSE or AIC for estimating an optimal set of network parameters. But the search space is notoriously multimodal and easily traps a search

algorithm in some local optimum. Moreover, because of the numerical integration involved in the method the fitness evaluation is the most expensive part of the search. Therefore, we need to use a very powerful global optimizer for the optimization task that can work reliably in a multimodal search space, exhibit robust performance in presence of noise, and possesses a very high convergence velocity for locating the global optimum. Hence, the first attempt in this work was to improve some existing evolutionary algorithm by increasing its convergence velocity and robustness. The next chapter presents the work where such performance enhancement of classic differential evolution algorithm was attempted.