Jin GU[1]✉

1 MOE Key Laboratory of Bioinformatics and Bioinformatics Div, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China

✉ Corresponding author: Jin GU

MOE Key Laboratory of Bioinformatics and Bioinformatics Div, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China

Email: gujin_00@mails.tsinghua.edu.cn

Brief review: the frontiers in the computational explorings of the gene regulations

Abstract

Computational methods have greatly expanded our understandings of the complex gene regulations in a systematical view. The rapid progress in molecular biology and high-throughput bio-techniques is providing new opportunities and challenges for the computational studies.

Keywords: computational biology, bioinformatics, gene regulations, regulatory networks, complex networks

1 Introduction

After several thousand years, the human society developed from simply cooperation for food to the complex modern society. For much longer periods, several billion years, the living creatures never stop their “evolution” and may form the most mysterious and complex world on the earth. According to current knowledge, for higher eukaryote such as human and mouse contains ~35,000 and ~30,000 genes, respectively; two lower eukaryotes, worm contains ~18,000 genes and fly contains ~14,000 genes; even yeast, one kind of the unicellular eukaryote, contains ~6,000 genes [1]. Such a large number of genes make the biological systems very complex. In a single cell, millions of molecules are produced, processed, transported, interacted, and degraded at the same time. Any tiny error may cause cascade effect to the system. How are the molecular processes programmed? How are the complex systems evolved?

For decades, many efforts combining extensive experimental and computational studies are devoted to deciphering the “regulatory code” which defines the rules of the gene regulations – how these genes are organized to the complex regulatory networks (for example, see yeast regulatory networks in [2-4])? But this code may be much more complex and flexible than any well-known artificial code. Many important and interesting topics about gene regulations are still under poor investigations. This article is written for the experts in informatics who are also interested in the mysterious life, to help them enter the promising field and have a comprehensive understanding of the frontiers on gene regulations. Three related sections are written to describe the computational explorations of gene regulations (Fig. 1).

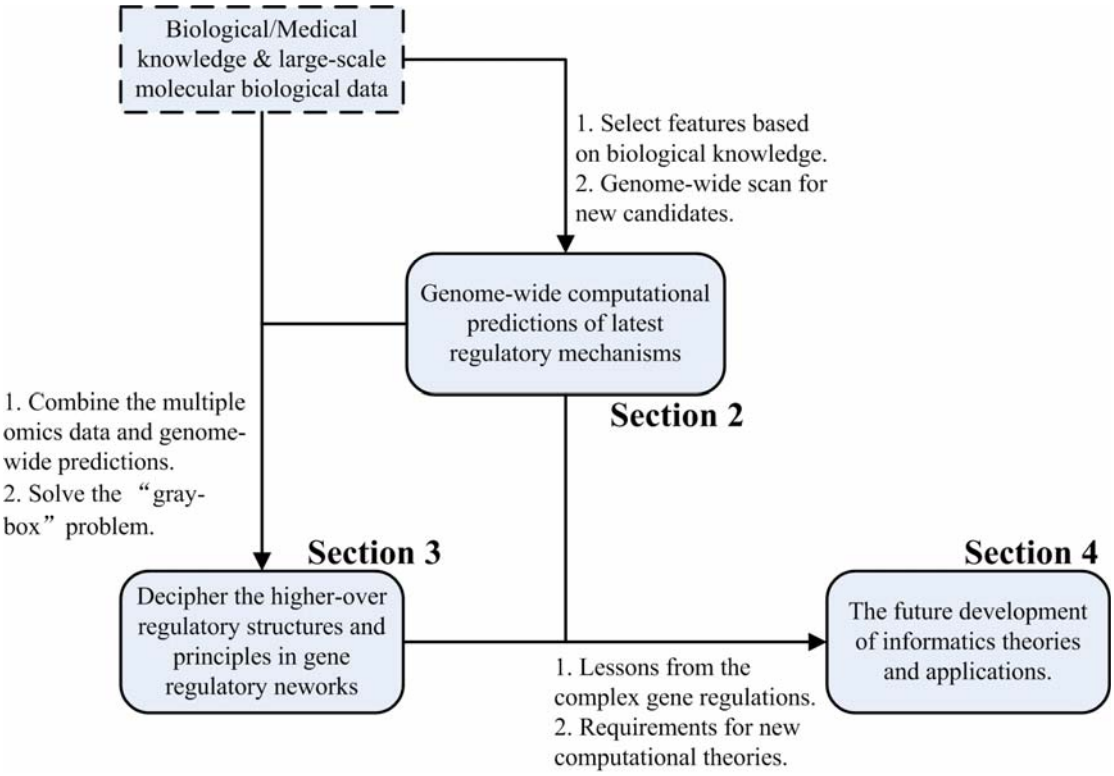
Biological/Medical knowledge & large-scale molecular biological data
1. Select features based on biological knowledge.
2. Genome-wide scan for new candidates.
Genome-wide computational predictions of latest regulatory mechanisms
Section 2
1. Combine the multiple omics data and genome-wide predictions.
2. Solve the "gray-box" problem.
Section 3
Decipher the higher-over regulatory structures and principles in gene regulatory neworks
Section 4
The future development of informatics theories and applications.
1. Lessons from the complex gene regulations.
2. Requirements for new computational theories.

Figure 1. The outline of the article.

The "easiest" way to start the computational investigations about gene regulations would be to identify specific types of regulatory codes or molecules using statistical and machine learning methods. In recent years, many new regulatory mechanisms have been discovered: such as alternative splicing, which makes us re-think the definition of "gene" and the complex degree of genome [5-7], and microRNAs (miRNAs), which play important roles in embryonic development, aging and cancer suppression, promote a hot topic on non-coding genes [8-11]. Using biological experiments to identify all these regulatory codes is still time-consuming and costly. Combinations of biological features and the proper computational methods, genome-wide predictions can be easily made based on a few known samples directly identified by biological experiments. In this article, the author would first review a few new regulatory mechanisms and related resources in (Section 2).

Modeling the complex gene regulations is the central task for the computational studies. Experimental data are crucial for these approaches. Thanks to the new high throughput bio-techniques and huge financial investment, significantly more data are available than the year 2003 when the human genome project was finished [12]. These omics* (the terminologies denoted by * will be shortly explained in Appendix A) data provide a chance to decipher the regulatory codes in a systematical vision. Now, high throughput parallel sequencing technique can produce millions of reads in a single reaction [13-16]. Micro-arrays can profile the expressions of all genes or even the whole genomes for any organism on a few chips [17-20]. The protein-DNA interaction, protein-protein interaction and protein-RNA interaction can also be detected in large scale by specific techniques [21-25]. Many traditional statistical methods and machine learning methods have been introduced to analyze different types of data and have improved our understanding of the complex regulations. As the data are rapidly accumulating and more regulatory mechanisms are discovered, it is under great expectation to develop new computational methods considering the non-linear dynamic effects of molecular interaction and the local optimization property after the long-term natural adaptation process. Also, modeling and analyzing the structure and dynamics of the complex regulatory network needs more systematical approaches and more biological concerns. In (Section 3), the author investigates the most challenging problems in which new computational theories and methods are needed.

Computational methods greatly advance the studies in gene regulations, while the principles in biological systems also forward the studies in computational methods. Holland put up with Genetic Algorithm (GA) in the 1960s by introducing natural adaptation process into the optimization of artificial systems [26]. It is a splendid case of learning from the nature. Life science has progressed significantly in recent decades. These progresses, especially the organization of the molecular regulatory networks, may provide new guides for the future developments of informatics theories and applications: thousand genes are organized as scale-free network; sub-networks are enriched in specific types of network motifs; the systems are robust under fluctuations. The author would like to briefly discuss this topic in (Section 4).

2 Latest regulatory mechanisms in gene regulations and their computational predictions

Computational predictions of regulatory elements and events are the fundamental applications of computational efforts in biology. Although many high-throughput bio-techniques have been developed to monitor multiple biological molecular, these experiments are still very expensive and laborious. Additionally, many new regulatory mechanisms, which are discovered in recent years, are hardly to be investigated by current techniques. Computational predictions can quickly provide a preliminary but panoramic vision of these new hotspots which can give great help to biologists.

To accomplish the predictions, much attention should be paid to the latest and key regulatory mechanisms. The methods based on specific biological and statistical features should be developed to investigate these mechanisms. Many canonical computational methods, such as neural network, hidden-Markov model and support vector machine have been successfully adapted to the predictions. In the following paragraphs, the author will present a few latest discovered regulatory mechanisms. A few key references and the existing obstacles about the computational predictions of these mechanisms will also be given.

At first, an overview about basic gene regulatory steps is given to help the understandings (Fig. 2). DNAs, RNAs and proteins are the three essential molecules in gene regulations. At specific conditions, transcription factors (one type of proteins) and pol*II* complex bind to DNA to activate transcription*. In transcription, new RNA primary transcripts are produced according to the DNA template. The primary transcripts are spliced* into mature messenger RNAs (mRNAs) or processed to other noncoding RNAs* such as tRNAs, rRNA and miRNAs. Then the ribosome binds to mRNAs to produce proteins according to the well-known triplet genetic code. This process is called translation*. While, the noncoding RNAs will play their roles in multiple molecular processes. The latest regulatory mechanisms presented below are involved in any of the three essential gene regulatory steps: transcription, splicing and translation.

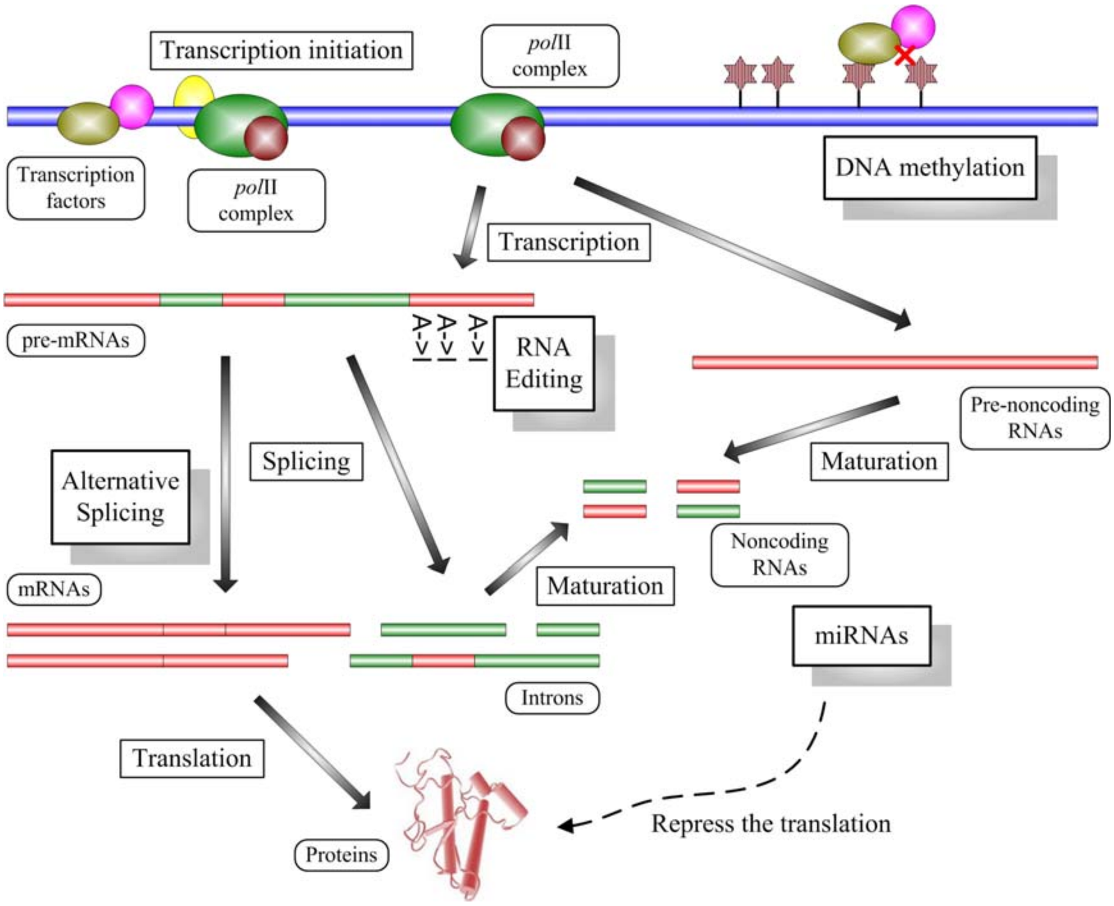

Figure 2. The cartoon of the essential molecular processes. The right-corner boxes indicate the molecular processes. The shadowed right-corner boxes show the four latest discovered regulatory mechanisms which will be discussed in the following paragraphs. The round-corner boxes indicate the molecules involved in corresponding processes.

2.1 DNA methylation

The DNA and histone methylation may change the chromatin structure and the binding affinity between transcription factors and their target sequences. Till to now, public DNA methylation data are still poor. According to authors' knowledge, Robins et al. published a dataset of about 4,000 DNA methlyation entries [27] and a public database, MethDB [28-30], contains about 20,000 entries. Based on these public data, several computational methods have been developed, such as Methylator [31], MethCGI [32] and another two methods provided by Das et al. [33] and Bock et al. [34]. In DNA methylation predictions, the CpG/TpG content, the distribution of Alu Y repeat sequences and the occurrence of certain TFBSs are the three useful features [32].

2.2 RNA editing

It is an important post-transcriptional modification of the RNA transcript that alters the original information of DNA sequence by substituting, inserting, or deleting single or di-nucleotides of mRNA after transcription and before splicing. RNA editing events include many different types: the C-to-U substitution editing mostly existing in higher plant mitochondria and chloroplasts and the A-to-I substitution editing found in the mammalian are the two main types. Three public databases are now available for RNA edits: EdRNA [35], dbRES [36] and REDIdb [37]. For the C-to-U editing, the two triplet nucleotides at upstream and downstream of the editing sites and the about 250-nt long flanking regions are two basic features for predictions [38-42]. For A-to-I editing, much complex model, including the local secondary structures of transcripts, sequence constraint and site clustering should be used [43].

2.3 Alternative splicing

During mRNA (or other types of RNAs) maturation, preliminary transcripts will be processed into mature mRNAs: introns are removed from transcripts and exons are spliced together into mRNAs. There are many splicing sites on a preliminary transcript. Different combinations of these splicing sites can result in different mRNA products. AltSplice Database is one of the most popular data source for the alternative splicing study [44, 45]. Although alternative splicing have been discovered many years before, the computational prediction of alternative splicing sites purely based genomic sequences is still a challenging task [46, 47]. Xia et al. introduced competitive mechanism of nearby spicing sites into the prediction model, which greatly improved the prediction performance [48].

2.4 MicroRNA (miRNA) regulations

MiRNAs are a class of ~22nt endogenous small RNAs which regulate target mRNAs by repressing the translation activity or directly degrading the mRNA transcripts. MiRNAs can guide gene silencing by partial sequence complementary to mRNA 3’-UTRs. About 500 miRNA genes have been discovered in human. In human, ~30% genes are under miRNA regulations. Due to their important roles in gene regulations, dozens of databases and computational studies about miRNA and their target sites predictions have been published. In them, miRBase is the authoritative database for miRNA sequences [49, 50], and TarBase provides comprehensive collections of experimental validated miRNA target sites [51]. The most popular miRNA prediction methods are reviewed in [52]; and the most target prediction methods are reviewed in [53, 54]. Till to now, miRNA target predictions without comparative genomic information remain a big problem due to too many false positives.

Except the latest regulatory mechanisms, the classic “well-known” gene regulation steps also need advanced computational investigations, such as the identifications and mappings of transcription factor binding sites, predictions of RNA motifs in mRNA untranslated regions, etc.

Prediction of these regulatory mechanisms in genome-wide is the first and the basic step to understand gene regulations in a systematical view. The same as other applications, computational predictions in biology consist of three steps: feature extraction, data-preprocessing and classifier design, and result evaluation. In these studies, the biggest challenging may not be in the classifier design but in the understanding of the corresponding biological process, which would help you reliably extract useful information from the original biological data.

3 Computational methods for deciphering the higher-order gene regulations

Beyond the computational predictions of single regulatory element or event, deciphering the higher-order gene regulations involving two or multiple regulatory elements is a more challenging task. Till to now, extensive computational investigations on a specific regulatory mechanism (such as transcription factor binding, splicing site recognition, miRNA target site binding, etc.) have been conducted in genome-wide scale. At the same time, high throughput bio-techniques, which can monitor multiple molecular processes in genome-wide scale, are also developing rapidly. These techniques not only produce significant more data than the traditional molecular biological experiments but also provide a systematical descript of molecular system. New methods which combine multiple experimental data sources and regulatory events need to be developed. In the following paragraphs, three challenging topics on deciphering the principles of the higher-order gene regulations will be discussed.

3.1 Non-linear and non-parameter methods to RECONSTRUCT tissue-specific regulatory structures of gene expressions

Multiple regulatory elements are organized into specific regulatory networks in different tissues. Current experimental methods are hardly to directly capture the complex regulatory structures. In an informatics or statistical vision, the “gray-box model” can be used to reconstruct the regulatory structures between multiple regulatory elements (Fig. 3). The inputs of the “gray-box” are the potential regulatory elements of a specific gene and the outputs are the gene expression levels* in different conditions (see review in [55]). Many computational methods (such as MEME [56, 57], MDScan [58], STORM [59], MCS [60] and PCS [61]) and new high-throughput bio-techniques have already developed to determine the inputs and outputs. How to reconstruct the regulatory structures in the box based on its inputs and outputs?

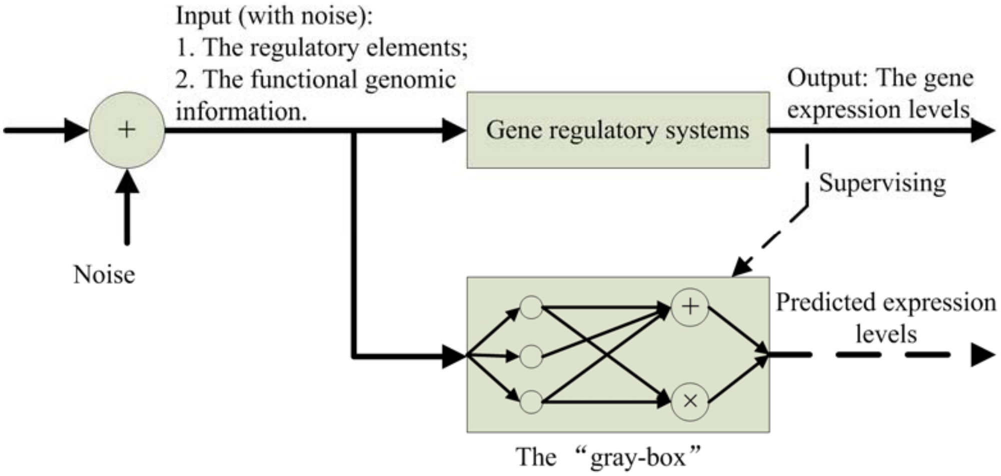
Input (with noise):
1. The regulatory elements;
2. The functional genomic information.
+
Noise
Gene regulatory systems
Output: The gene expression levels
Supervising
Predicted expression levels
+
×
The "gray-box"

Figure 3. The “gray-box” model to describe the high-over gene regulatory structures.

To resolve the structures in the “gray-box”, several regression models have been proposed: such as linear regression [62], multiple multivariate adaptive regression splines algorithm [63, 64], logistic regression [65] and regularized sliced inverse regression [66]. Following is a very simple regression model which considers the maximal two-order motif interactions:

$$G = e_0 + \sum_{i_1=0}^{n} \alpha_{i_1} m_{i_1} + \sum_{i_1=0}^{n} \sum_{i_2=0}^{n} \alpha_{i_1 i_2} (m_{i_1} m_{i2})$$

G is the gene expression level, *e₀* is the basic expression level, *mᵢ* is the count of matched *i*-th motifs, *n* is the number of different types of motifs and α is the contribution of the motifs or the motif-pairs. Regression models can explain the contributions of different regulatory elements or their combinations to the expressions of their target genes in the give condition (see review in [67]). The significant elements and their combinations in the regression models form the final resolved regulatory structures. Although regression methods can capture the regulatory structures in a clear “cause-and-result” relationship, the used regression models need pre-defined model-assumptions which are not always true in the real world. For example, the linear regression model assumes that the relationship between the regulatory elements’ total binding strength and the expressions of their target genes is linear [62]. Nonlinear regression methods with relaxed model assumptions should be developed to overcome this problem.

Except regression methods, several non-parametric methods may also fulfill this task. Till to now, clustering methods or machine learning methods are rarely used to decipher the above “gray box”. Recently, a few databases collecting experimentally validated enhancers are available, such as VISTA Enhancer Browser [68] and REDfly [69]. Based on these data, our lab-mates are devoted to introducing machine learning methods to reconstruct the regulatory structures. To realize the applications, many aspects of these methods should be revised and improved: 1) because the available data are highly sparse, the risk of the predictor should be under careful control; 2) the expression data are highly noisy, so the false discovery rate of the predictor should also be taken into serious consideration; 3) the aim of the “predictor” is not to predict but to decipher the contribution of each motif and combination to the gene expressions in specific tissues, so the optimal principles may not be simply the same with minimizing the errors.

3.2 Unification view for gene regulation by combining multiple regulatory mechanisms

Several latest discovered regulatory mechanisms have been introduced in Section 2. These events interact with each other not only in the time scale but also in the space scale to maintain the gene expressions. Different regulatory mechanisms interact with each other may significantly increase the flexibility and the complexity of the gene regulations.

To better understanding of the functions and characteristics of regulatory mechanisms’ interaction, attentions should be paid for the intersections event among different regulatory mechanisms. For example, in an older view, spliced introns are useless and to be quickly degraded. But recent studies on small regulatory RNAs show that many spliced introns enter other processing pathways and produce miRNAs [70, 71]. The chimeric structure of mRNAs and miRNAs can greatly increase the regulatory complexity in a single unit of regulon. For another example, the RNA editings in miRNA seed regions are reported to redirect target recognitions. The edited miR-376 targets an important gene which contributes the tight and tissue-specific regulation of uric-acid levels [72]. Current genome-wide computational studies mostly focus on a single mechanism. The methods which can scan for the interactions of different regulatory mechanisms should be presented. For different mechanisms, computational methods may encounter the different accuracies or different data scales. Genome-wide scans and statistical studies should be more carefully implemented to eliminate false positives.

3.3 Analysis of the complex network

From the early 2000s, gene regulatory networks have been established in several model organisms, such as yeast [2-4], worm [73, 74] and fly [75]. These studies mainly rely on the genome-wide mapping of the protein-DNA and protein-protein interaction. Besides reconstructing different networks, analysis of the complexities (such as the structures and dynamics) of the regulatory networks is another challenging topic.

Sontag et al. have tried to study the a known gene regulatory network by analyzing the dynamics of the node states [76]. In the network, each protein-coding gene can be regarded as a node, the expression level of the gene can be assigned as the status of the node, the regulations of the gene expression (such as transcriptional regulations and mRNA degradation) can be regarded as the inputs of the node and the protein products can be regarded as the output of the node. The status (expression level) of the node (gene) changes dynamically as a function of the inputs and the output (protein level) changes dynamically as another function of the status. In a complex gene regulatory network, multiple steady statuses may exist. Each steady status may represent specific phenotype [76]. The conditions and probabilities of the steady status transmission can greatly help us understand and control the behaviors of the complex networks (Fig. 4). In a recent article, Kitano discussed the concept biological robustness [77]. Not the same as the traditional concept of robustness – the systems remain at their steady state under perturbations; the biological networks may intend to change their structures and states to actively adapt environmental perturbations. New computational theories and methods are needed to investigate these adaptive actions in the complex biological networks.

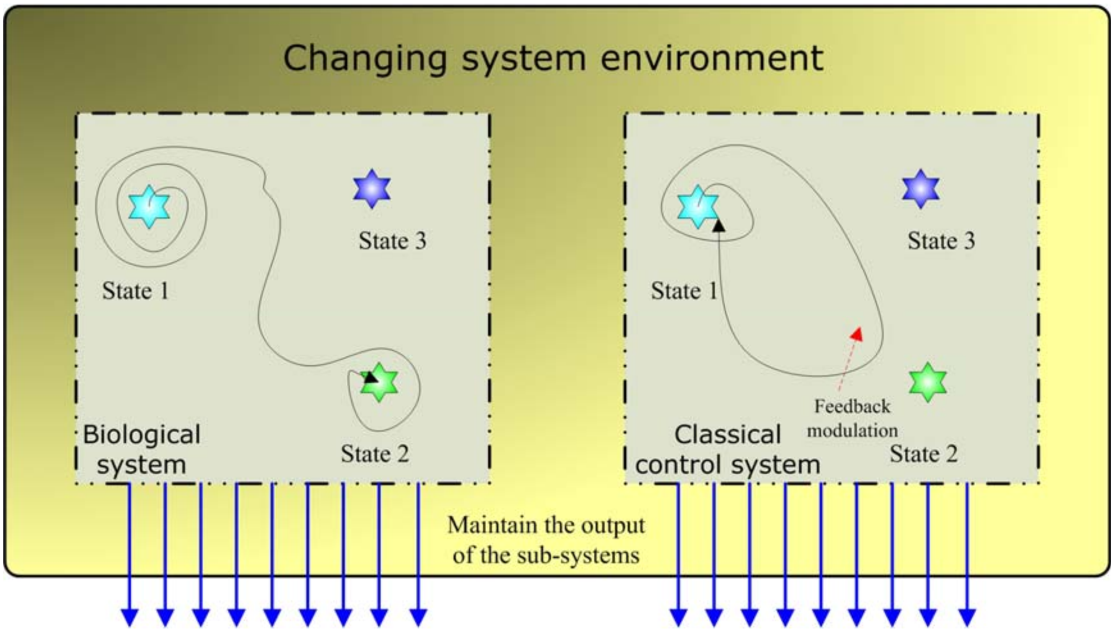

Figure 4. The different system reactions in changing environment. In the biological systems (left box), systems maintain their functions by shifting between states. In the classical control systems (right box), systems maintain their functions by keeping their states through the feedback modulation.

Identifications of hub-genes and network motifs are another two useful ways to analyze the complexity of gene regulatory networks. In a review article about complex network [78], authors mentioned that many cellular networks, such as many metabolism networks [79] and yeast protein-protein interaction network [80], are found to be like scale-free networks, with a few hub-genes having significant more connections. Milo et al. and Shen-Orr et al. introduced a statistical method to evaluate the significances of the different types of sub-networks which are defined as network motifs [81, 82]. However, these statistics are still unable to capture all key nodes. To increase the performance of current methods, more types of information, such as the expression fluctuations of the genes and the states of the neighbor genes, should be taken into consideration.

From above paragraphs, it can be clearly found that computational studies become more and more important in life science. The combinations of traditional molecular experiments and computational efforts will unravel more mysteries of life in a systematical scale.

4 Hints for informatics theories and applications from gene regulations

Genetic Algorithm starts from the simulation of the natural adaptation process [83]. An Artificial Neural Network inspired by the way biological nervous systems, such as the brain, process information [84]. They are two big successes by learning from life science. As our understandings of the complex gene regulations are progressing rapidly, more lessons can be learned from living creatures to the future development of informatics theories and applications. In this part, two principles derived from gene regulatory networks will be briefly discussed.

4.1. Enrichment of specific types of network motifs

Milo et al. (2002) published their systematical research about network motifs in the transcriptional regulatory networks of E. coli and yeast [81]. Since then, dozens of studies have been reported. Shen-Orr et al. reported that three types of network motifs are enriched in the E. coli's transcriptional regulatory networks: 1) feed-forward loop; 2) single input module (SIM) and 3) dense overlapping regulons (DOR) [82]. Coherent feed-forward loops can act as a circuit that ignore the transient activation from the general transcription factors and responds only to long-term effect, while allowing a rapid system shutdown. The SIM motif may correspond to the cases that a group of genes involved in the same pathway are activated by a single transcription factor but with different activation thresholds. Shen-Orr et al. also observed that feedback loops occur rarely in E. coli transcriptional regulatory network. But the feedback loops, especially negative feedback loops are important for systems stability. That many feedbacks may be achieved by post-transcriptional regulations may explain the lacking of feedback loops in transcriptional regulatory networks. This assumption now has strong supports from the current studies on miRNA-directed translational repression [85]. In an exhaustive computational analysis, Prill et al. demonstrated that the robustness to small perturbations is highly correlated with the relative abundance of network motifs [86].

4.2 Stronger system robustness with more system flexibility

As the definition by Kitano, “robustness is a property that allows a system to maintain its functions against internal and external perturbations” [87]. Biological system can shift from on one steady state to another state to maintain its functions if the environmental conditions have changed [77, 88, 89]. In the classic control theory, how to maintain the system state under perturbations is the fundamental task. Many artificial systems and networks are designed on the classic control theory. To better adapt in the fluctuations, systems can be designed with multiple steady states. If only shifting between these steady states, the systems can always work normal although with different system states [77] (Fig. 4).

5 Summary: Understanding the gene regulations in system level

Computational explorations of gene regulations are interesting and promising. Now it is in a very challenging period. Most well-known computational methods have already been applied on gene regulations successfully. However, as new biological knowledge and experimental data are bursting in recent years, current methods cannot fulfill the systematical studies in gene regulations. New theories and methods are needed for manipulating, integrating and analyzing these omics data in a systematical view. Years ago, the concept of systems biology was introduced. It intends to understand a biological system in 1) system structures; 2) system dynamics; 3) the control methods and 4) the design method [90]. At the same time, the studies on gene regulations are developing from the single element analysis to the large-scale analysis. Systems biology puts up with new research scopes of gene regulations and the complex gene regulatory networks provide a good platform to study molecular biological systems. In the future, combinations of the abstract concepts in systems biology and concrete systems in gene regulations will be worthy to be expected.

At last, it must be mentioned that due to the limitation of journal space and our knowledge, this article cannot cover all aspects of this rapidly developing field, although the author intends to give a broad review of the computational investigations on gene regulations.

Appendix A. Terminology

Gene expressions/expression levels: in this article, it means the amount of RNA molecules derived from specific gene region which is the DNA template of the RNA molecules.

Non-coding RNAs (ncRNAs): the RNA transcripts that do not encode a peptide function directly as regulators or modifiers, such as microRNAs (miRNAs) and small interfering RNAs (siRNAs).

Omics: Omics is a general term for a broad discipline of science and engineering for analyzing the interactions of biological information objects in various omes. These include genomics, proteomics, metabolomics, expressomics and interactomics [91].

Splice: messenger RNA (mRNA) is first transcribed as a continuous long primary transcript and then the long primary transcript is processed into mature mRNA. RNA splicing is a process that removes introns and joins exons in a primary transcript.

Transcription: in genetics the process by which a DNA sequence is converted to a corresponding RNA sequence based on the Watson-Crick complement rule (A->U, T->A, C->G, G->C).

Translation: synthesis of proteins (or peptides) from mRNA based on the triplet genetic code.

Acknowledgements

The author is grateful to thank Professor Yanda Li for the detail discussions of the outlines and writings. The author thanks Tao He and Pufeng Du for providing key references about RNA editing and Tao Peng for his help on the part of alternative splicing. The author also wants to thank all our lab-mates who provide us great help to have a comprehensive vision on the computational studies of the gene regulations. This work is supported by the National Natural Science Foundation of China 60572086 and 60775002.

References

1. Lewin B. Gene VIII. Pearson Prentice Hall 2004.
2. Cliften P, Sudarsanam P, Desikan A *et al.* Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science (New York, NY 2003; **301** (5629):71-76.
3. Harbison CT, Gordon DB, Lee TI *et al.* Transcriptional regulatory code of a eukaryotic genome. Nature 2004; **431** (7004):99-104.
4. Lee I, Date SV, Adai AT *et al.* A probabilistic functional network of yeast genes. Science (New York, NY 2004; **306** (5701):1555-1558.
5. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annual review of biochemistry 2003; **72**:291-336.
6. Blencowe BJ. Alternative splicing: new insights from global analyses. Cell 2006; **126** (1):37-47.
7. Lopez AJ. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annual review of genetics 1998; **32**:279-305.
8. Ambros V. The functions of animal microRNAs. Nature 2004; **431** (7006):350-355.
9. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004; **116** (2):281-297.
10. Baulcombe D. RNA silencing in plants. Nature 2004; **431** (7006):356-363.
11. Kim VN, Nam JW. Genomics of microRNA. Trends Genet 2006; **22** (3):165-173.
12. Pennisi E. Human genome. Reaching their goal early, sequencing labs celebrate. Science (New York, NY 2003; **300** (5618):409.
13. Brenner S, Johnson M, Bridgham J *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature biotechnology 2000; **18** (6):630-634.
14. Brenner S, Williams SR, Vermaas EH *et al.* In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proceedings of the National Academy of Sciences of the United States of America 2000; **97** (4):1665-1670.
15. Margulies M, Egholm M, Altman WE *et al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005; **437** (7057):376-380.
16. Ng P, Tan JJ, Ooi HS *et al.* Multiplex sequencing of paired-end ditags (MS-PET): a strategy for

the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic acids research 2006; **34** (12):e84.

17. Biemar F, Nix DA, Piel J *et al.* Comprehensive identification of Drosophila dorsal-ventral patterning genes using a whole-genome tiling array. Proceedings of the National Academy of Sciences of the United States of America 2006; **103** (34):12763-12768.

18. Biemar F, Zinzen R, Ronshaugen M *et al.* Spatial regulation of microRNA gene expression in the Drosophila embryo. Proceedings of the National Academy of Sciences of the United States of America 2005; **102** (44):15907-15911.

19. Cheng J, Kapranov P, Drenkow J *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science (New York, NY 2005; **308** (5725):1149-1154.

20. Manak JR, Dike S, Sementchenko V *et al.* Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nature genetics 2006; **38** (10):1151-1158.

21. Gerber AP, Luschnig S, Krasnow MA *et al.* Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America 2006; **103** (12):4487-4492.

22. Iyer VR, Horak CE, Scafe CS *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 2001; **409** (6819):533-538.

23. Lieb JD, Liu X, Botstein D *et al.* Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nature genetics 2001; **28** (4):327-334.

24. Ren B, Robert F, Wyrick JJ *et al.* Genome-wide location and function of DNA binding proteins. Science (New York, NY 2000; **290** (5500):2306-2309.

25. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nature biotechnology 2000; **18** (12):1257-1261.

26. Holland HJ. Adaptation in Natural and Artificial Systems. MIT Press 1992.

27. Rollins RA, Haghighi F, Edwards JR *et al.* Large-scale structure of genomic methylation patterns. Genome research 2006; **16** (2):157-163.

28. Amoreira C, Hindermann W, Grunau C. An improved version of the DNA Methylation database (MethDB). Nucleic acids research 2003; **31** (1):75-77.

29. Grunau C, Renault E, Roizes G. DNA Methylation Database "MethDB": a user guide. The Journal of nutrition 2002; **132** (8 Suppl):2435S-2439S.

30. Grunau C, Renault E, Rosenthal A *et al.* MethDB--a public database for DNA methylation data. Nucleic acids research 2001; **29** (1):270-274.

31. Bhasin M, Zhang H, Reinherz EL *et al.* Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS letters 2005; **579** (20):4302-4308.

32. Fang F, Fan S, Zhang X *et al.* Predicting methylation status of CpG islands in the human brain. Bioinformatics (Oxford, England) 2006; **22** (18):2204-2209.

33. Das R, Dimitrova N, Xuan Z *et al.* Computational prediction of methylation status in human genomic sequences. Proceedings of the National Academy of Sciences of the United States of America 2006; **103** (28):10713-10716.

34. Bock C, Paulsen M, Tierling S *et al.* CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS genetics 2006; **2** (3):e26.

35. Hung JH, Wang WC, Huang HD. Systematic identification and repository of RNA editing site in human genome. International Computer Symposium 2006:1386-1391.

36. He T, Du P, Li Y. dbRES: a web-oriented database for annotated RNA editing sites. Nucleic acids

research 2007; **35** (Database issue):D141-144.

37. Picardi E, Regina TM, Brennicke A *et al.* REDIdb: the RNA editing database. Nucleic acids research 2007; **35** (Database issue):D173-177.

38. Cummings MP, Myers DS. Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. BMC bioinformatics 2004; **5**:132.

39. Mower JP. PREP-Mt: predictive RNA editor for plant mitochondrial genes. BMC bioinformatics 2005; **6**:96.

40. Thompson J, Gopal S. Correction: genetic algorithm learning as a robust approach to RNA editing site site prediction. BMC bioinformatics 2006; **7**:406.

41. Thompson J, Gopal S. Genetic algorithm learning as a robust approach to RNA editing site prediction. BMC bioinformatics 2006; **7**:145.

42. Du P, He T, Li Y. Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features. Biochemical and biophysical research communications 2007; **358** (1):336-341.

43. Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS biology 2004; **2** (12):e391.

44. Stamm S, Riethoven JJ, Le Texier V *et al.* ASD: a bioinformatics resource on alternative splicing. Nucleic acids research 2006; **34** (Database issue):D46-55.

45. Thanaraj TA, Stamm S, Clark F *et al.* ASD: the Alternative Splicing Database. Nucleic acids research 2004; **32** (Database issue):D64-69.

46. Ohler U, Shomron N, Burge CB. Recognition of unknown conserved alternatively spliced exons. PLoS computational biology 2005; **1** (2):113-122.

47. Wang M, Marin A. Characterization and prediction of alternative splice sites. Gene 2006; **366** (2):219-227.

48. Xia H, Bi J, Li Y. Identification of alternative 5'/3' splice sites based on the mechanism of splice site competition. Nucleic acids research 2006; **34** (21):6305-6313.

49. Griffiths-Jones S. miRBase: the microRNA sequence database. Methods in molecular biology (Clifton, NJ 2006; **342**:129-138.

50. Griffiths-Jones S, Grocock RJ, van Dongen S *et al.* miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids research 2006; **34** (Database issue):D140-144.

51. Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA (New York, NY 2006; **12** (2):192-197.

52. Berezikov E, Cuppen E, Plasterk RH. Approaches to microRNA discovery. Nature genetics 2006; **38 Suppl**:S2-7.

53. Rajewsky N. microRNA target predictions in animals. Nature genetics 2006; **38 Suppl**:S8-13.

54. Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. Nature methods 2006; **3** (11):881-886.

55. Bussemaker HJ, Foat BC, Ward LD. Predictive modeling of genome-wide mRNA expression: from modules to molecules. Annual review of biophysics and biomolecular structure 2007; **36**:329-347.

56. Grundy WN, Bailey TL, Elkan CP. ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. Comput Appl Biosci 1996; **12** (4):303-310.

57. Grundy WN, Bailey TL, Elkan CP *et al.* Meta-MEME: motif-based hidden Markov models of protein families. Comput Appl Biosci 1997; **13** (4):397-406.

58. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications

to chromatin-immunoprecipitation microarray experiments. Nature biotechnology 2002; **20** (8):835-839.

59. Schones DE, Smith AD, Zhang MQ. Statistical significance of cis-regulatory modules. BMC bioinformatics 2007; **8**:19.

60. Xie X, Lu J, Kulbokas EJ *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 2005; **434** (7031):338-345.

61. Gu J, Fu H. The Pairwise Conservation Scores - An Algorithm to Identify Conserved K-mers. Available from: http://bioinfo.au.tsinghua.edu.cn/member/~gujin/pcs/.

62. Roven C, Bussemaker HJ. REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. Nucleic acids research 2003; **31** (13):3487-3490.

63. Das D, Banerjee N, Zhang MQ. Interacting models of cooperative gene regulation. Proceedings of the National Academy of Sciences of the United States of America 2004; **101** (46):16234-16239.

64. Das D, Nahle Z, Zhang MQ. Adaptively inferring human transcriptional subnetworks. Molecular systems biology 2006; **2**:2006 0029.

65. Keles S, van der Laan MJ, Vulpe C. Regulatory motif finding by logic regression. Bioinformatics (Oxford, England) 2004; **20** (16):2799-2811.

66. Zhong W, Zeng P, Ma P *et al.* RSIR: regularized sliced inverse regression for motif discovery. Bioinformatics (Oxford, England) 2005; **21** (22):4169-4175.

67. Jaqaman K, Danuser G. Linking data to models: data regression. Nature reviews 2006; **7** (11):813-819.

68. Visel A, Minovitsky S, Dubchak I *et al.* VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic acids research 2007; **35** (Database issue):D88-92.

69. Gallo SM, Li L, Hu Z *et al.* REDfly: a Regulatory Element Database for Drosophila. Bioinformatics (Oxford, England) 2006; **22** (3):381-383.

70. Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Drosha processing. Nature 2007; **448** (7149):83-86.

71. Kim YK, Kim VN. Processing of intronic microRNAs. The EMBO journal 2007; **26** (3):775-783.

72. Kawahara Y, Zinshteyn B, Sethupathy P *et al.* Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science (New York, NY 2007; **315** (5815):1137-1140.

73. Vermeirssen V, Barrasa MI, Hidalgo CA *et al.* Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network. Genome research 2007; **17** (7):1061-1071.

74. Deplancke B, Mukhopadhyay A, Ao W *et al.* A gene-centered C. elegans protein-DNA interaction network. Cell 2006; **125** (6):1193-1205.

75. Davidson EH, Rast JP, Oliveri P *et al.* A genomic regulatory network for development. Science (New York, NY 2002; **295** (5560):1669-1678.

76. Chaves M, Albert R, Sontag ED. Robustness and fragility of Boolean models for genetic regulatory networks. Journal of theoretical biology 2005; **235** (3):431-449.

77. Kitano H. Towards a theory of biological robustness. Molecular systems biology 2007; **3**:137.

78. Albert R, Barabasi AL. Statistical mechanics of complex networks. Rev Mod Phys 2002; **74** (1):47-97.

79. Jeong H, Tombor B, Albert R *et al.* The large-scale organization of metabolic networks. Nature 2000; **407** (6804):651-654.

80. Jeong H, Mason SP, Barabasi AL *et al.* Lethality and centrality in protein networks. Nature 2001;

411 (6833):41-42.

81. Milo R, Shen-Orr S, Itzkovitz S *et al.* Network motifs: simple building blocks of complex networks. Science (New York, NY 2002; **298** (5594):824-827.

82. Shen-Orr SS, Milo R, Mangan S *et al.* Network motifs in the transcriptional regulation network of Escherichia coli. Nature genetics 2002; **31** (1):64-68.

83. Holland JH. Adaptation in natural and artificial system. The MIT Press 1992.

84. Haykin S. Neural Networks: A Comprehensive Foundation. Prentice Hall; 2nd edition (July 6, 1998) 1998.

85. Tsang J, Zhu J, van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Molecular cell 2007; **26** (5):753-767.

86. Prill RJ, Iglesias PA, Levchenko A. Dynamic properties of network motifs contribute to biological network organization. PLoS biology 2005; **3** (11):e343.

87. Kitano H. Biological robustness. Nat Rev Genet 2004; **5** (11):826-837.

88. Kitano H. The theory of biological robustness and its implication in cancer. Ernst Schering Research Foundation workshop 2007 (61):69-88.

89. Kitano H. Biological robustness in complex host-pathogen systems. Progress in drug research Fortschritte der Arzneimittelforschung 2007; **64**:239, 241-263.

90. Kitano H. Systems biology: a brief overview. Science (New York, NY 2002; **295** (5560):1662-1664.

91. Omics.org. an openfree wiki site for -omes and -omics in Biotechnology and Bioscience. Available from: http://omics.org/index.php/What_is_omics.