RNA Technologies

Nikolaus Rajewsky
Stefan Jurga
Jan Barciszewski
Editors

Systems Biology

Springer

RNA Technologies

More information about this series at http://www.springer.com/series/8619

Nikolaus Rajewsky • Stefan Jurga • Jan Barciszewski
Editors

Systems Biology

Springer

Editors
Nikolaus Rajewsky
Max Delbrück Center for Molecular Medicine
Berlin Institute for Medical Systems Biology
Berlin-Buch,
Berlin, Germany

Stefan Jurga
Nanobiomedical Center
Adam Mickiewicz University
Poznań, Poland

Jan Barciszewski
Institute of Bioorganic Chemistry
Polish Academy of Sciences
Poznań, Poland

ISSN 2197-9731 ISSN 2197-9758 (electronic)
RNA Technologies
ISBN 978-3-319-92966-8 ISBN 978-3-319-92967-5 (eBook)
https://doi.org/10.1007/978-3-319-92967-5

Library of Congress Control Number: 2018950402

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Systems Biology

The nineteenth and the twentieth century—a time during which our knowledge about how organisms function on a cellular and molecular level started to explode—witnessed the emergence of many new branches of biology such as cell biology, developmental biology, evolutionary biology, biochemistry, genetics, epigenetics, and molecular biology. Each of them focuses on a different aspect of the mechanisms and principles governing living organisms.

Systems biology brings the findings of these disciplines together with the aim to develop a holistic rather than reductionist understanding of cells, organisms, and ecosystems. Its goal is to understand the networks of individual biological components and to decipher how these networks and regulatory circuits interact to form living systems. A deep understanding of biological systems is achieved by gaining insight into their structure, dynamics, and control mechanisms. Systems biology represents a highly integrative and interdisciplinary approach. In addition to biology and medicine, it heavily relies on computer sciences and mathematics while also involving chemistry and physics.

The concept of systems biology emerged during the early twentieth century, when the notion became more and more accepted that biological systems follow physical and mechanical laws, elegantly outlined in D'Arcy Thompson's work "On Growth and Form," 1917. Other theories and discoveries contributed to the refinement of this concept during the course of the twentieth century. Notable examples include Conrad Waddington, who characterized networks of genes, cells, and tissues as decision-making dynamical systems; Ludwig von Bertallanffy with his "Outline of General Systems Theory" in 1950; Alan Lloyd Hodgkin and Andrew Fielding Huxley, who in 1952 spearheaded mathematical modeling of biological systems by describing how an action potential moves along the axon of a neuronal cell; Jacques Jacob and Francois Monod, who, when conducting their famous research on gene regulatory elements in the 1960s, concluded that mechanisms of gene regulation could form a variety of networks endowed with any desired degree

of stability; as well as Eric Davidson and Roy Britten, who in 1969 pioneered the concept of gene regulatory networks. The term systems biology is attributed to systems theorist Mihajlo Mesarovic. He coined it in 1966 when hosting the international symposium "Systems Theory and Biology" at the Case Institute of Technology in Cleveland, OH. With the *Institute for Systems Biology* in Seattle and the *Systems Biology Institute* in Tokyo, the first systems biology institutes were founded in the year 2000, and many others followed.

The rise of systems biology as a key biological discipline in the new millennium was fueled by the preceding and concurrent development of high-throughput technologies such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics. Omics technologies required novel specialized devices and experimental workflows as well as accompanying computational tools and mathematical models. The latter, which are needed to integrate the wealth of the generated data, were made possible thanks to the simultaneous vast expansion of computing power. Vice versa, systems biology continues to be a driving force behind the constant development and improvement both of experimental techniques and equipment to extract large amounts of qualitative and quantitative information from complex biological samples and of the bioinformatic pipelines necessary to obtain meaningful biological insights. An example of a more recent technological advancement in systems biology is represented by the development of single-cell omics technologies over the last decade, which now permit us to study the molecular make-up and dynamics of tissues and entire organisms at single-cell resolution.

A current challenge in systems biology is the integration of different regulatory levels such as genetic, epigenetic, and posttranscriptional gene regulation and the comprehension of the interplay between these levels. The long-term goal is the deduction of predictive models that enable us to foresee how cells and organisms change over time and in response to external stimuli or perturbations. Machine learning and artificial intelligence are going to be essential in the development of such multidimensional models that take spatial and temporal information into account. Being able to predict the fate of cells, tissues, organs, and organisms would be extremely powerful, since it would not only provide us with a fundamental understanding of how life works on a molecular and cellular level, but would also be a huge step forward for personalized medicine. It would allow us to foresee the course of human diseases and to choose the most effective therapies for each patient.

This book illustrates how systems biology is instrumental in advancing our knowledge about the principles of cellular and tissue organization. Themes covered include regulation of gene expression by genome structure, RNA-binding proteins, RNA–RNA interactions, noncoding RNAs, transcriptomics, epigenomics, metabolomics, posttranscriptional gene regulation, systems biology in health and disease, experimental and computational tools for systems biology research, computational methods for multidimensional data analysis, and integration as well as the deduction of predictive models.

The chapters will provide the reader with examples of how important scientific questions are addressed in systems biology and of bioinformatic tools designed to reach valuable conclusions from the abundance of the generated information.

Berlin, Germany — Nikolaus Rajewsky
Berlin, Germany — Verena Maier
Poznań, Poland — Stefan Jurga
Poznań, Poland — Jan Barciszewski

Contents

Systems Biology of Genome Structure and Dynamics

Zahra Fahmi, Sven A. Sewitz, and Karen Lipkow

Contents

Abstract Our view of the packed genome inside a nucleus has evolved greatly over the past decade. Aided by novel experimental and bioinformatic analysis techniques and detailed computational models, fundamental insights into the structure and dynamics of chromosomes have been gained. This has revealed that genome organisation has an essential role in controlling genome function during normal growth, cellular differentiation, and stress response, showing that, overall, 3D reorganisation is tightly linked to changes in gene expression. Chromatin, which is composed of DNA and a large number of different chromatin-associated proteins and RNAs, is often chemically modified, in patterns that affect gene expression. It has become clear that this highly interconnected system requires computational simulations to gain an understanding of the underlying system-wide mechanisms.

In this review, we describe different modelling approaches that are used to investigate both the linear and spatial arrangement of chromatin. We illustrate how dynamic computer simulations are used to study the mechanisms that control and maintain genome architecture and drive changes in this structure. We focus on

Z. Fahmi · S. A. Sewitz
Nuclear Dynamics Programme, Babraham Institute, Cambridge, UK

K. Lipkow (✉)
Nuclear Dynamics Programme, Babraham Institute, Cambridge, UK

Cambridge Systems Biology Centre, University of Cambridge, Cambridge, UK
e-mail: karen.lipkow@babraham.ac.uk

N. Rajewsky et al. (eds.), *Systems Biology*, RNA Technologies,
https://doi.org/10.1007/978-3-319-92967-5_1

models of the dynamics of epigenetic modifications, of protein–DNA interactions, and the polymer dynamics of chromosomes. These approaches provide reliable frameworks to integrate additional biological data; enable accurate, genome-wide predictions; and allow the discovery of new mechanisms.

Keywords Chromatin organisation · Computational model · Histone modification · Facilitated diffusion · Polymer · Chromatin loop · Self-organisation

1 Background

Intensive studies over the past decades have revealed multiple levels of organisation in eukaryotic genomes. The DNA wraps around eight histone proteins to make a nucleosome, the fundamental subunit of the chromatin fibre (van Holde 1989; Ramakrishnan 1997; Sewitz et al. 2017b). In mammals, the chromatin then folds to build higher genomic structures of different scales such as sub-megabase topologically associated domains (TADs), megabase A and B compartments, and chromosomal territories (Bonev and Cavalli 2016; Sewitz et al. 2017b). The nucleus is a highly crowded environment with efficiently packed and organised chromatin and hundreds to thousands of protein species, engaged in various types of interactions, such as protein–protein, DNA–protein, chromatin–chromatin, and chromatin–lamina interactions. It is now known that these interactions play an important role in controlling the organised structure and regulating the transcriptional activity of the genome (Gómez-Díaz and Corces 2014; Long et al. 2016; Flavahan et al. 2016) and that the structure changes upon differentiation and internal and external conditions (Guidi et al. 2015; Javierre et al. 2016; Sewitz et al. 2017a; Lazar-Stefanita et al. 2017). However, a comprehensive view of the mechanisms that drive organisation and dynamics of this highly complex system remains elusive.

Many research projects have investigated the linear arrangement of DNA, identifying the local regulatory elements that modulate transcription, such as transcription factor binding sites and their consensus sequences (Levine and Tjian 2003), enhancers (Long et al. 2016), histone modifications (Smolle and Workman 2013), and sites of DNA methylation (Schübeler 2015). Activator and repressor proteins recruit enzymes, such as histone acetyltransferase or histone deacetylase, that modify histones. Histone modifications control gene expression by altering the local chromatin structure and inhibiting or attracting DNA-binding factors (Dindot and Cohen 2013). In addition, DNA methylation can repress transcription through blocking the binding of transcription factors or mediating the binding of repressors (Jaenisch and Bird 2003).

It has more recently become possible to quantitatively investigate 3D genome architecture using live-cell microscopy, and chromosome conformation capture techniques, such as 3C, 4C, 5C, Hi-C, and Capture Hi-C (Schmitt et al. 2016b). This has greatly enhanced our understanding of gene regulatory mechanisms, by showing how the three-dimensional organisation of the genome influences gene

regulation (Babu et al. 2008; Cavalli and Misteli 2013; Zuin et al. 2014; Lupiáñez et al. 2015; Dixon et al. 2016; Schmitt et al. 2016a). Many genes occupy preferred non-random positions within the nucleus: in mammals, gene-poor or transcriptionally inactive regions are located close to the nuclear envelope in most cell types, whereas gene-rich or transcriptionally active regions prefer to localise at the borders of chromosome territories, away from the nuclear periphery (Foster and Bridger 2005; Nagano et al. 2013). Manipulating the position of genes can also affect their activity; for human and mouse cells, it has been shown that relocating genes from their normal position to regions close to the nuclear periphery results in gene silencing (Reddy et al. 2008; Finlan et al. 2008). The single-celled eukaryote *S. cerevisiae* displays a mosaic arrangement of heterochromatin and euchromatin at the nuclear periphery, with active genes located close to nuclear pores (Casolari et al. 2004) and inactive genes associated with other parts of the nuclear periphery and the nuclear centre (Zimmer and Fabre 2011).

This organisation is achieved within a highly dynamic nucleoplasm (Misteli 2001; Vazquez et al. 2001; Lanctôt et al. 2007). For example, in mammalian cells, GFP-tagged proteins were measured to diffuse with diffusion coefficients of 0.24–0.53 $\mu m^2\ s^{-1}$, taking 24–54 s to travel 5 μm, a distance almost equal to the radius of the nucleus (Phair and Misteli 2000). Tagged chromosomal loci in living *S. cerevisiae* cells move more than 0.5 μm, equivalent to half of nuclear radius, within a few seconds (Heun et al. 2001). There is now evidence that the dynamics of the heterogeneous chromatin fibre contributes to thermodynamically driven 3D self-organisation (Sewitz et al. 2017a).

Investigation of chromatin organisation in space and time by novel experimental techniques has unravelled some of the key features of this intricate system of how genome structure relates to the function of the genome. To further study the dynamics of chromosome structures, particularly aspects that are not amenable to experimental analysis, scientists have adopted modelling approaches. Models provide the most direct way to explore mechanisms, as all components, interactions, reactions, and forces are defined, and any observed behaviour must be a consequence of these. During recent years, a wide range of models of the full or partial genome have been developed to analyse the interplay of genome structure and function. In this review, we categorise these models into three major groups: models of epigenetic modification dynamics, protein–DNA models, and polymer-based models.

2 Models of Epigenetic Modification Dynamics

Histone proteins can be covalently modified on several residues after translation (Allfrey et al. 1964), which leads to the recruitment of transcriptional regulatory proteins and structural proteins over a local chromatin region. For example, the combined deacetylation and methylation of the lysine at position 9 of histone H3 (H3K9) is required to create a binding site for the Swi6/HP1 silencing factor

(Nakayama et al. 2001; Shankaranarayana et al. 2003). Binding of silencing factors facilitates the modification of histones on adjacent nucleosomes, and sequential rounds of epigenetic modification and protein binding lead to the spreading of heterochromatin over a chromatin region (Grewal and Moazed 2003). Specialised boundary elements inhibit the heterochromatin extension and therefore separate silent and active chromatin domains (West et al. 2002; Labrador and Corces 2002).

To understand the mechanisms behind the epigenetic memory of monostable domains, predictive models have investigated the behaviour of H3K9 methylation domains (Hathaway et al. 2012; Hodges and Crabtree 2012; Müller-Ott et al. 2014; Erdel and Greene 2016). Simulations at single-nucleosome resolution showed that confined and heritable steady states of histone marks can be achieved by modelling linear propagation of histone modifications from nucleation sites to adjacent nucleosomes. Turnover of modified nucleosomes could also happen simultaneously (Hathaway et al. 2012; Hodges and Crabtree 2012). In contrast, another model assumed loop-driven spreading of histone marks with sparse nucleation sites. By adjusting parameters such as modification rates, the model was shown to be robust against replication (Erdel and Greene 2016), and the response towards transient perturbations was in line with experimental data (Müller-Ott et al. 2014).

Genomic regions of high epigenetic dynamics are bistable states, characterised by the presence of both activating and repressive histone marks (Bernstein et al. 2006). They have been observed for confined chromatin domains in various cell types (Rohlf et al. 2012; Tee et al. 2014). To study the features and dynamics of these states, several computational models have been developed (Dodd et al. 2007; Sedighi and Sengupta 2007; David-Rus et al. 2009; Micheelsen et al. 2010; Mukhopadhyay et al. 2010; Angel et al. 2011; Dodd and Sneppen 2011; Berry et al. 2017). In these models, a region of chromatin is represented as a sequence of nucleosomes. At every time step, each nucleosome has a state or a rate of histone modification based on its histone marks, with rules that govern state transitions or changes in rates. These models have shown that nonlinear positive feedback loops are required for robust and heritable bistable epigenetic states. Positive feedback loops arise when modifications of one nucleosome stimulate the modifications of other nucleosomes. The required nonlinearity can be achieved in different ways: (1) via the cooperativity of two or more nucleosomes with the same histone marks, which recruit histone modifiers on other nucleosomes (Dodd et al. 2007; Sedighi and Sengupta 2007; David-Rus et al. 2009; Micheelsen et al. 2010; Mukhopadhyay et al. 2010; Angel et al. 2011; Dodd and Sneppen 2011); (2) through two-step feedback loops, where the switch of histone modification states of nucleosomes occurs via an intermediate state, i.e. the state first changes to the intermediate state and then to the favoured state (Dodd et al. 2007; Angel et al. 2011; Berry et al. 2017); (3) through the local transcription rate, which can be affected by silencing, in turn leading to a change in the local modification rate (Sedighi and Sengupta 2007); and (4) through interactions with non-neighbour nucleosomes (Dodd et al. 2007). Another mathematical model with a 1D array of nucleosomes has been formulated to study the dynamics of histone modification in bivalent domains, where active and repressive histone marks coexist on nucleosomes (Ku et al. 2013). These domains

are important elements in stem cells, and according to the model's prediction, their formation process is generally slow. The model also suggested that a coordinated set of parameters, such as recruitment and exchange rates of marks, leads to established and maintained bivalent domains over several cell cycles.

3 Protein–DNA Models

Transcription factors (TF) affect the transcriptional activity of specific genes through binding to specific DNA sequences (Ptashne and Gann 2002). It has been proposed that these proteins search for their target sequences through facilitated diffusion (Berg et al. 1981, 1982; Berg and von Hippel 1985), i.e. alternating rounds of 3D diffusion in the solution, sliding along the DNA, short-range excursions called hopping, and intersegmental transfer between DNA segments. The characteristics of this search mechanism have been widely studied, and computational models of different scales have brought new insights into its dynamics. All models discussed in this section have focused on facilitated diffusion of TFs.

At the most detailed, atomistic level, molecular dynamics (MD) simulations have been used to explain how, e.g. the *lac* repressor protein (LacI) moves along DNA (Marklund et al. 2013) and how it identifies its target site (Furini et al. 2013). LacI is modelled to take a helical path to probe the DNA, with its DNA-binding interface being insensitive to modest bends in DNA conformation. The hydrogen bonds formed between the DNA and the LacI interface are dynamic and flexible, allowing fast sliding of the protein (Marklund et al. 2013). This was found to enable the protein to probe the DNA quickly and reach the proximity of the target site. Once the specific DNA sequence is bound, it becomes significantly slower, resulting in the formation of a stable protein–DNA structure and a drop in enthalpy (Iwahara and Levy 2013; Furini et al. 2013). Another fine-grained MD simulation has proposed that binding of the CSL (CBF1/Suppressor of Hairless/LAG-1) protein to the DNA can transmit a signal through the protein structure according to the bound sequences. This influences the inter-domain dynamics of the protein and consequently its functional activities (Torella et al. 2014).

The effects of DNA conformation on the dynamics of TF proteins probing the DNA were explored via coarse-grained MD simulations, where proteins interact with the DNA via electrostatic interactions (Bhattacherjee and Levy 2014a, b). The geometry of DNA was tuned by two factors, curvature and the degree of helical twisting. Highly curved or highly twisted DNA was seen to lead to a decrease in sliding frequencies and an increase in hopping events (Bhattacherjee and Levy 2014a). In addition, introducing curvatures in the DNA conformation was found to increase the frequency of jumping events of a multidomain protein between distant DNA sites. However, curvature does not necessarily result in faster search kinetics as sliding happens less often (Bhattacherjee and Levy 2014b). Hence, an optimal DNA conformation can lead to a balanced number of searching events and maximal probing of DNA.

To investigate the role of nonspecific DNA–protein interactions during the search for specific target sites, Monte Carlo simulations were adopted (Das and Kolomeisky 2010; Tabaka et al. 2014; Mahmutovic et al. 2015). It was argued that the binding of the LacI repressor to nonspecific DNA is controlled by either activation or steric effects instead of being limited by diffusion (Tabaka et al. 2014; Mahmutovic et al. 2015). Furthermore, it was shown that for efficient and fast probing of DNA, moderate ranges of nonspecific binding energies and protein concentrations are required (Das and Kolomeisky 2010). The necessity for moderate DNA–protein binding strength has been indicated for proteins with different subdiffusive motions using simulations based on Brownian dynamics (Liu et al. 2017).

Large-scale computer simulations have been performed to study the search kinetics of transcription factors both in prokaryotic and eukaryotic cells. Software called GRiP (Gene Regulation in Prokaryotes) (Zabet and Adryan 2012a) provides a simulation framework for analysing the stochastic target search process of TF proteins. In GRiP the DNA is modelled as a string of base pairs, and TFs are highly diffusing components that interact with DNA sequences or with each other. This framework has been utilised to build a detailed model of facilitated diffusion, where TF orientation on the DNA, cooperativity of TFs, and crowding were incorporated (Zabet and Adryan 2012b). A similar model was adopted to dissect the effects of biologically relevant levels of mobile and immobile crowding on TF performance in a bacterial cell (Zabet and Adryan 2013): immobile crowding fixed on the DNA raises the occupancy of target sites significantly, whereas both mobile and immobile crowding have negligible impacts on the mean search time. Another model of the bacterial genome has taken two types of crowding molecules into account (Brackley et al. 2013). Proteins which bind to and move along DNA (1D crowding) do not change the search time significantly, even at very high densities. However, crowding molecules diffusing freely in 3D space increase the frequency of 1D sliding of TFs along DNA, while they enhance the robustness of the search time against any change in protein–DNA affinity.

A different approach based on the Gillespie stochastic simulation algorithm has been developed to analyse the influence of macromolecular crowding on gene expression in stem cells (Golkaram et al. 2017). The crowding was assumed to be correlated with the local chromatin density, which was calculated using Hi-C data. Diffusive TFs and RNA polymerases were only moving in the proximity of promoters, as crowding would not allow them to diffuse to other regions between rebindings. The model predicted that an increase in chromatin density during development leads to a rise in transcriptional bursting and subsequently heterogeneous expression of genes in a cell population.

Our lab has developed a computational model of TF motions in eukaryotes (Schmidt et al. 2014; Sewitz and Lipkow 2016) using the particle-based simulator Smoldyn (Andrews et al. 2010). This model has considered different types of movements for TFs: 3D diffusion, sliding, hopping, and intersegmental transfer. Among others, it showed the importance of intersegmental transfer, and it provided an explanation for the size of nucleosome-free regions on the DNA, which improve

the process of TFs binding to their targets. Similar to a prokaryotic model (Tabaka et al. 2014), inclusion of 1D diffusion reduced the time to find the target sites by one and two orders of magnitude.

Finally, the complexity of gene regulation in higher eukaryotes has motivated the study of evolutionary dynamics of the TF repertoire and their binding preferences. A stochastic model based on duplication and mutation of genes suggested that more complex organisms with higher number of genes have higher levels of redundancy of TF binding (Rosanova et al. 2017).

4 Polymer-Based Models

The dynamic nature of the chromatin fibre lends itself to simulating chromatin as an extended, highly mobile polymer. Several studies have extended concepts developed in physics and applied them to the analysis of chromatin (Tark-Dame et al. 2011; Koslover and Spakowitz 2014; Shukron and Holcman 2017). This has led to an understanding of genome-wide data of chromosome folding and their interactions with each other and with other nuclear elements. In all models presented here, the chromatin fibre is a diffusing and self-avoiding chain of beads arranged in 3D space.

4.1 Models Based on Chromatin Loops

Chromatin loops have been observed in both eukaryotes and prokaryotes (Hofmann and Heermann 2015), and their vital regulatory impact has been demonstrated. A number of these models have suggested that chromatin loops are formed mainly by interactions between specific protein complexes like condensin (Cheng et al. 2015) or CTCF (Tark-Dame et al. 2014). These models have successfully reproduced the experimentally observed genome compaction. In addition, the importance of balance between short-range and long-range loops for controlling the changes in chromosomes structure has been revealed (Tark-Dame et al. 2014). It has furthermore been indicated that the dynamic bridges between condensin complexes bring about the intrachromosomal interactions during both interphase and mitosis in budding yeast (Cheng et al. 2015).

Other models have explored the general effects of protein interactions on chromatin structure. A heteropolymer model incorporated proteins implicitly, by mapping different epigenetic states onto the beads. Specific interactions between beads of the same state were differentiated from nonspecific interactions between any pair of beads (Jost et al. 2014). The model predicted that inter-TAD interactions are highly dynamic, which was in line with Hi-C results. It also predicted the fast formation of TADs, followed by a slow and long process of compaction (Jost et al. 2014). The lattice version of this model (Olarte-Plata et al. 2016), and another heteropolymer model (Ulianov et al. 2016) with active or inactive

epigenomic states for beads, confirmed stronger self-attraction for inactive domains (Ulianov et al. 2016; Olarte-Plata et al. 2016) and an increase in their compaction as the domain size grows (Olarte-Plata et al. 2016). Other models based their assignment on levels of gene activity, with highly active or less active states assigned according to their expression levels (Jerabek and Heermann 2012). Highly active chromatin sections had low interaction strength, while less active ones had higher interaction affinity. The average distances between genomic loci, the average volume ratio between highly active and less active regions, and the positioning of highly active loci close to the boundary of chromosome territories were all in line with experimental measurements. In another work the polymer model was informed by protein binding sites and histone modifications (Brackley et al. 2016) and produced a population of genome conformations, which predicted the 3D distances between selected genomic sites on the globin locus in mouse ES cells.

In addition, polymer models based on protein interactions and without relying on predetermined information for the state of chromatin beads were developed (Giorgetti et al. 2014; Tiana et al. 2016; Chiariello et al. 2016). Using iterative Monte Carlo simulations and comparisons to the measured contact frequencies, the parameters of the models were optimised, and ensembles of chromatin configurations were achieved (Giorgetti et al. 2014; Tiana et al. 2016; Chiariello et al. 2016). These models correctly estimated the contact frequencies of TADs (Giorgetti et al. 2014; Chiariello et al. 2016) and the mean 3D distances between labelled loci upon perturbations of specific sites (Giorgetti et al. 2014). Combined with live-cell measurements, it has been suggested that changes in TAD conformations happen fast enough (in a much shorter time frame than the cell cycle) to facilitate dynamic interactions between regulatory elements, such as enhancer–promoter interactions (Tiana et al. 2016). A homopolymer model (Doyle et al. 2014), which implemented chromatin loops in the proximity of enhancer and promoter elements, indicated that the loops can either facilitate or insulate the enhancer–promoter interactions significantly. It was shown that the regulatory effect of the loop was dependant on the relative positions of loop anchors. To minimise the reliance on specific biological data, a heteropolymer model was built based on hierarchical folding and statistical physics of disordered systems (Nazarov et al. 2015). This model has two types of monomers that can interact with each other. By tuning the 1D sequence of monomers and the temperature controlling the folding, the simulated contact maps achieved a resemblance to Hi-C data.

Besides the notion that direct interactions between bound proteins shape chromatin loops, another mechanism, called loop extrusion, has been proposed (Nasmyth 2001; Alipour and Marko 2012; Sanborn et al. 2015; Fudenberg et al. 2016). This model calls for the action of extruding machines, possibly condensin or cohesin complexes, to bind and move along the DNA in opposite directions (Nasmyth 2001; Alipour and Marko 2012; Sanborn et al. 2015; Fudenberg et al. 2016). This leads to the extrusion of DNA loops until domain boundaries, occupied by CTCF proteins, are reached (Sanborn et al. 2015; Fudenberg et al. 2016). This mechanism can account for the compaction and folding of mitotic chromosomes (Nasmyth 2001; Alipour and Marko 2012). Furthermore, in combination with polymer physics,

the model reproduced the observed decay of contact probabilities with increasing genomic distance, leading to simulated contact maps consistent with Hi-C data. It also predicted the changes in contact frequencies and 3D distances between loci due to CTCF and cohesin perturbations (Sanborn et al. 2015; Fudenberg et al. 2016).

4.2 Models Based on Supercoiling

Different levels of unconstrained supercoiling have been observed for chromatin (Kouzine et al. 2013; Naughton et al. 2013), and it has been reported that transcription leads to supercoiling (Wu et al. 1988; Kouzine et al. 2008; Papantonis and Cook 2011). To explore the effects of supercoiling on genome organisation in both eukaryotic (Benedetti et al. 2014) and prokaryotic (Le et al. 2013) cells, detailed polymer models have been employed. In a eukaryotic model, borders of TADs were mapped to the chromatin fibre, and strong supercoiling was imposed to the intervening chromatin (Benedetti et al. 2014). This led to the formation of TADs and contact maps broadly consistent with 3C data. In a bacterial model, chromatin was simulated as a dense array of plectonemes that were attached to a back bone (Le et al. 2013). By inserting plectoneme-free regions in the model at the positions of highly expressed genes, the contact frequencies observed for chromosomal interaction domains were reproduced. Overall, supercoiling is essential for creating chromosomal interaction domains (Le et al. 2013) and topologically associated domains (Benedetti et al. 2014). Intriguingly, a recent model investigated the role of supercoiling introduced by the transcribing RNA polymerase (Racko et al. 2017): when both CTCF and cohesin were included in the simulation, cohesin rings were seen to accumulate at CTCF sites demarking TAD borders. These observations are also seen experimentally (Uusküla-Reimand et al. 2016). Under these conditions, supercoiled DNA loops were extruded, and the supercoiling was the driving force for extruding the DNA loops. This is interesting because until now it was unclear how the energetically expensive loop extrusion could be achieved. Now, RNA polymerase-generated supercoiling provides a credible and testable hypothesis.

4.3 Integrative Models and Self-Organisation

With significant amounts of genome-wide datasets becoming available, computational models of chromatin are becoming more sophisticated and feature-rich. Computational models have explored the role of this heterogeneity in self-organisation of the genome structure.

In budding yeast, highly expressed genes are less occupied by chromatin-associated proteins, whereas genes that show lower overall expression are bound more extensively (Sewitz et al. 2017a). Protein occupancy can affect the local physical properties of the chromatin segment by means of a range of parameters

such as changes in mass, diameter, local viscosity (Jirgensons 1958; Oldfield and Dunker 2014), diffusion speed (Jerabek and Heermann 2012; Phillip and Schreiber 2013; Wollman et al. 2017), and electrical charge of chromatin. This has led to the development of heteropolymeric models which incorporate some of the underlying complexity and points towards protein occupancy being a causal factor in determining self-organisation of genome structure in yeast (Sewitz et al. 2017a).

A significant challenge in this area is to continue to develop physical models of heteropolymeric motion applicable to chromatin. In many instances, insights are mainly qualitative and require physical parameters that are known to be unphysiological. As an example, it was shown that two chromosomes that differed in temperature-driven mobility would separate via a process akin to phase separation (Loi et al. 2008). Chromatin segments that harboured more active genes were given a higher temperature. This model reproduced the experimentally observed chromosomal territories (Ganai et al. 2014), but only if a temperature difference of 20-fold was assumed. Using much longer chromosomal segments, similar phase separations could already be observed with much smaller differences in temperature, bringing the model in closer proximity to real-life biological systems (Smrek and Kremer 2017). Still, current models are not yet fully able to deal with the structural complexity that is the hallmark of chromatin.

5 Conclusion and Outlook

It is now evident that the study of chromatin structure is at a stage where computational models are not just an accessory but a required component of any thorough investigation. The advent of pervasive high-performance computing has made it possible to attempt whole genome simulations at moderate resolutions, or smaller genomes at higher resolutions. Two future strands of development are now visible. Firstly, an ever-increasing amount of relevant genomic data is making its way into computational simulations. This will lead to more complex models that incorporate genome-wide protein binding data, extended epigenetic data, and measures of local chromatin conformation. This will also push the theoretical descriptions in polymer physics, where we foresee that increased and intensive collaboration and exchange is necessary. This will be mutually beneficial, as both fields will fundamentally improve their understanding of an area of biological physics that underpins questions of gene regulation during development, in response to external changes, and, in cases of misregulation, disease. These efforts are just at the beginning and will require the combined expertise of computational scientists, physicists, and experimental biologists to fully unravel the complex dynamics that lead to chromatin self-organisation.

References

Alipour E, Marko JF (2012) Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic Acids Res 40:11202–11212

Allfrey VG, Faulkner R, Mirsky AE (1964) Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proc Natl Acad Sci U S A 51:786–794

Andrews SS, Addy NJ, Brent R, Arkin AP (2010) Detailed simulations of cell biology with Smoldyn 2.1. PLoS Comput Biol 6:e1000705

Angel A, Song J, Dean C, Howard M (2011) A polycomb-based switch underlying quantitative epigenetic memory. Nature 476:105–108

Babu MM, Janga SC, de Santiago I, Pombo A (2008) Eukaryotic gene regulation in three dimensions and its impact on genome evolution. Curr Opin Genet Dev 18:571–582

Benedetti F, Dorier J, Burnier Y, Stasiak A (2014) Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. Nucleic Acids Res 42:2848–2855

Berg OG, von Hippel PH (1985) Diffusion-controlled macromolecular interactions. Annu Rev Biophys Biophys Chem 14:131–160

Berg OG, Winter RB, von Hippel PH (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. Biochemistry 20:6929–6948

Berg OG, Winter RB, von Hippel PH (1982) How do genome-regulatory proteins locate their DNA target sites? Trends Biochem Sci 7:52–55

Bernstein BE, Mikkelsen TS, Xie X et al (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125:315–326

Berry S, Dean C, Howard M (2017) Slow chromatin dynamics allow polycomb target genes to filter fluctuations in transcription factor activity. Cell Syst 4:445–457.e8

Bhattacherjee A, Levy Y (2014a) Search by proteins for their DNA target site: 1. The effect of DNA conformation on protein sliding. Nucleic Acids Res 42:12404–12414

Bhattacherjee A, Levy Y (2014b) Search by proteins for their DNA target site: 2. The effect of DNA conformation on the dynamics of multidomain proteins. Nucleic Acids Res 42:12415–12424

Bonev B, Cavalli G (2016) Organization and function of the 3D genome. Nat Rev Genet 17:661–678

Brackley CA, Cates ME, Marenduzzo D (2013) Intracellular facilitated diffusion: searchers, crowders, and blockers. Phys Rev Lett 111:108101

Brackley CA, Brown JM, Waithe D et al (2016) Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. Genome Biol 17:59

Casolari JM, Brown CR, Komili S et al (2004) Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. Cell 117:427–439

Cavalli G, Misteli T (2013) Functional implications of genome topology. Nat Struct Mol Biol 20:290–299

Cheng TMK, Heeger S, Chaleil RAG et al (2015) A simple biophysical model emulates budding yeast chromosome condensation. eLIFE 4:e05565

Chiariello AM, Annunziatella C, Bianco S et al (2016) Polymer physics of chromosome large-scale 3D organisation. Sci Rep 6:29775

Das RK, Kolomeisky AB (2010) Facilitated search of proteins on DNA: correlations are important. Phys Chem Chem Phys 12:2999–3004

David-Rus D, Mukhopadhyay S, Lebowitz JL, Sengupta AM (2009) Inheritance of epigenetic chromatin silencing. J Theor Biol 258:112–120

Dindot SV, Cohen ND (2013) Epigenetic regulation of gene expression: emerging applications for horses. J Equine Vet Sci 33:288–294

Dixon JR, Gorkin DU, Ren B (2016) Chromatin domains: the unit of chromosome organization. Mol Cell 62:668–680

Dodd IB, Sneppen K (2011) Barriers and silencers: a theoretical toolkit for control and containment of nucleosome-based epigenetic states. J Mol Biol 414:624–637
Dodd IB, Micheelsen MA, Sneppen K, Thon G (2007) Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell 129:813–822
Doyle B, Fudenberg G, Imakaev M, Mirny LA (2014) Chromatin loops as allosteric modulators of enhancer-promoter interactions. PLoS Comput Biol 10:e1003867
Erdel F, Greene EC (2016) Generalized nucleation and looping model for epigenetic memory of histone modifications. Proc Natl Acad Sci U S A 113:E4180–E4189
Finlan LE, Sproul D, Thomson I et al (2008) Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet 4:e1000039
Flavahan WA, Drier Y, Liau BB et al (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature 529:110–114
Foster HA, Bridger JM (2005) The genome and the nucleus: a marriage made by evolution. Genome organisation and nuclear architecture. Chromosoma 114:212–229
Fudenberg G, ImakaevM LC et al (2016) Formation of chromosomal domains by loop extrusion. Cell Rep 15:2038–2049
Furini S, Barbini P, Domene C (2013) DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. Nucleic Acids Res 41:3963–3972
Ganai N, Sengupta S, Menon GI (2014) Chromosome positioning from activity-based segregation. Nucleic Acids Res 42:4145–4159
Giorgetti L, Galupa R, Nora EP et al (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell 157:950–963
Golkaram M, Jang J, Hellander S et al (2017) The role of chromatin density in cell population heterogeneity during stem cell differentiation. Sci Rep 7:13307
Gómez-Díaz E, Corces VG (2014) Architectural proteins: regulators of 3D genome organization in cell fate. Trends Cell Biol 24:703–711
Grewal SIS, Moazed D (2003) Heterochromatin and epigenetic control of gene expression. Science 301:798–802
Guidi M, Ruault M, Marbouty M et al (2015) Spatial reorganization of telomeres in long-lived quiescent cells. Genome Biol 16:206
Hathaway NA, Bell O, Hodges C et al (2012) Dynamics and memory of heterochromatin in living cells. Cell 149:1447–1460
Heun P, Taddei A, Gasser SM (2001) From snapshots to moving pictures: new perspectives on nuclear organization. Trends Cell Biol 11:519–525
Hodges C, Crabtree GR (2012) Dynamics of inherently bounded histone modification domains. Proc Natl Acad Sci U S A 109:13296–13301
Hofmann A, Heermann DW (2015) The role of loops on the order of eukaryotes and prokaryotes. FEBS Lett 589:2958–2965
Iwahara J, Levy Y (2013) Speed-stability paradox in DNA-scanning by zinc-finger proteins. Transcription 4:58–61
Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33(Suppl):245–254
Javierre BM, Burren OS, Wilder SP et al (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 167:1369–1384
Jerabek H, Heermann DW (2012) Expression-dependent folding of interphase chromatin. PLoS One 7:e37525
Jirgensons B (1958) Optical rotation and viscosity of native and denatured proteins. X. Further studies on optical rotatory dispersion. Arch Biochem Biophys 74:57–69
Jost D, Carrivain P, Cavalli G, Vaillant C (2014) Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res 42:9553–9561
Koslover EF, Spakowitz AJ (2014) Multiscale dynamics of semiflexible polymers from a universal coarse-graining procedure. Phys Rev E Stat Nonlinear Soft Matter Phys 90:013304

Kouzine F, Sanford S, Elisha-Feil Z, Levens D (2008) The functional response of upstream DNA to dynamic supercoiling in vivo. Nat Struct Mol Biol 15:146–154
Kouzine F, Gupta A, Baranello L et al (2013) Transcription-dependent dynamic supercoiling is a short-range genomic force. Nat Struct Mol Biol 20:396–403
Ku WL, Girvan M, Yuan G-C et al (2013) Modeling the dynamics of bivalent histone modifications. PLoS One 8:e77944
Labrador M, Corces VG (2002) Setting the boundaries of chromatin domains and nuclear organization. Cell 111:151–154
Lanctôt C, Cheutin T, Cremer M et al (2007) Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nat Rev Genet 8:104–115
Lazar-Stefanita L, Scolari VF, Mercy G et al (2017) Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. EMBO J 36:2684–2697
Le TBK, Imakaev MV, Mirny LA, Laub MT (2013) High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342:731–734
Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424:147–151
Liu L, Cherstvy AG, Metzler R (2017) Facilitated diffusion of transcription factor proteins with anomalous bulk diffusion. J Phys Chem B 121:1284–1289
Loi D, Mossa S, Cugliandolo LF (2008) Effective temperature of active matter. Phys Rev E Stat Nonlinear Soft Matter Phys 77:051111
Long HK, Prescott SL, Wysocka J (2016) Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell 167:1170–1187
Lupiáñez DG, Kraft K, Heinrich V et al (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161:1012–1025
Mahmutovic A, Berg OG, Elf J (2015) What matters for lac repressor search in vivo—sliding, hopping, intersegment transfer, crowding on DNA or recognition? Nucleic Acids Res 43:3454–3464
Marklund EG, Mahmutovic A, Berg OG et al (2013) Transcription-factor binding and sliding on DNA studied using micro- and macroscopic models. Proc Natl Acad Sci U S A 110:19796–19801
Micheelsen MA, Mitarai N, Sneppen K, Dodd IB (2010) Theory for the stability and regulation of epigenetic landscapes. Phys Biol 7:026010
Misteli T (2001) Protein dynamics: implications for nuclear architecture and gene expression. Science 291:843–847
Mukhopadhyay S, Nagaraj VH, Sengupta AM (2010) Locus dependence in epigenetic chromatin silencing. Biosystems 102:49–54
Müller-Ott K, Erdel F, Matveeva A et al (2014) Specificity, propagation, and memory of pericentric heterochromatin. Mol Syst Biol 10:746
Nagano T, Lubling Y, Stevens TJ et al (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502:59–64
Nakayama J, Rice JC, Strahl BD et al (2001) Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. Science 292:110–113
Nasmyth K (2001) Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. Annu Rev Genet 35:673–745
Naughton C, Avlonitis N, Corless S et al (2013) Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat Struct Mol Biol 20:387–395
Nazarov LI, Tamm MV, Avetisov VA, Nechaev SK (2015) A statistical model of intra-chromosome contact maps. Soft Matter 11:1019–1025
Olarte-Plata JD, Haddad N, Vaillant C, Jost D (2016) The folding landscape of the epigenome. Phys Biol 13:026001
Oldfield CJ, Dunker AK (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem 83:553–584
Papantonis A, Cook PR (2011) Fixing the model for transcription: the DNA moves, not the polymerase. Transcription 2:41–44

Phair RD, Misteli T (2000) High mobility of proteins in the mammalian cell nucleus. Nature 404:604–609
Phillip Y, Schreiber G (2013) Formation of protein complexes in crowded environments – from in vitro to in vivo. FEBS Lett 587:1046–1052
Ptashne M, Gann A (2002) Genes & signals. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
Racko D, Benedetti F, Dorier J, Stasiak A (2017) Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes. Nucleic Acids Res 46:1648–1660
Ramakrishnan V (1997) Histone structure and the organization of the nucleosome. Annu Rev Biophys Biomol Struct 26:83–112
Reddy KL, Zullo JM, Bertolino E, Singh H (2008) Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature 452:243–247
Rohlf T, Steiner L, Przybilla J et al (2012) Modeling the dynamic epigenome: from histone modifications towards self-organizing chromatin. Epigenomics 4:205–219
Rosanova A, Colliva A, Osella M, Caselle M (2017) Modelling the evolution of transcription factor binding preferences in complex eukaryotes. Sci Rep 7:7596
Sanborn AL, Rao SSP, Huang S-C et al (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A 112:E6456–E6465
Schmidt HG, Sewitz S, Andrews SS, Lipkow K (2014) An integrated model of transcription factor diffusion shows the importance of intersegmental transfer and quaternary protein structure for target site finding. PLoS One 9:e108575
Schmitt AD, Hu M, Jung I et al (2016a) A Compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep 17:2042–2059
Schmitt AD, Hu M, Ren B (2016b) Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol 17:743–755
Schübeler D (2015) Function and information content of DNA methylation. Nature 517:321–326
Sedighi M, Sengupta AM (2007) Epigenetic chromatin silencing: bistability and front propagation. Phys Biol 4:246–255
Sewitz S, Lipkow K (2016) Systems biology approaches for understanding genome architecture. Methods Mol Biol 1431:109–126
Sewitz SA, Fahmi Z, Aljebali L et al (2017a) Heterogeneous chromatin mobility derived from chromatin states is a determinant of genome organisation in *S. cerevisiae*. bioRxiv:106344. https://doi.org/10.1101/106344
Sewitz S, Fahmi Z, Lipkow K (2017b) Higher order assembly: folding the chromosome. Curr Opin Struct Biol 42:162–168
Shankaranarayana GD, Motamedi MR, Moazed D, Grewal SIS (2003) Sir2 regulates histone H3 lysine 9 methylation and heterochromatin assembly in fission yeast. Curr Biol 13:1240–1246
Shukron O, Holcman D (2017) Transient chromatin properties revealed by polymer models and stochastic simulations constructed from chromosomal capture data. PLoS Comput Biol 13:e1005469
Smolle M, Workman JL (2013) Transcription-associated histone modifications and cryptic transcription. Biochim Biophys Acta 1829:84–97
Smrek J, Kremer K (2017) Small activity differences drive phase separation in active-passive polymer mixtures. Phys Rev Lett 118:098002
Tabaka M, Kalwarczyk T, Hołyst R (2014) Quantitative influence of macromolecular crowding on gene regulation kinetics. Nucleic Acids Res 42:727–738
Tark-Dame M, van Driel R, Heermann DW (2011) Chromatin folding – from biology to polymer models and back. J Cell Sci 124:839–845
Tark-Dame M, Jerabek H, Manders EMM et al (2014) Depletion of the chromatin looping proteins CTCF and cohesin causes chromatin compaction: insight into chromatin folding by polymer modelling. PLoS Comput Biol 10:e1003877

Tee W-W, Shen SS, Oksuz O et al (2014) Erk1/2 activity promotes chromatin features and RNAPII phosphorylation at developmental promoters in mouse ESCs. Cell 156:678–690

Tiana G, Amitai A, Pollex T et al (2016) Structural fluctuations of the chromatin fiber within topologically associating domains. Biophys J 110:1234–1245

Torella R, Li J, Kinrade E et al (2014) A combination of computational and experimental approaches identifies DNA sequence constraints associated with target site binding specificity of the transcription factor CSL. Nucleic Acids Res 42:10550–10563

Ulianov SV, Khrameeva EE, Gavrilov AA et al (2016) Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Res 26:70–84

Uusküla-Reimand L, Hou H, Samavarchi-Tehrani P et al (2016) Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. Genome Biol 17:182

van Holde KE (1989) Chromatin. Springer series in molecular and cell biology. Springer, New York

Vazquez J, Belmont AS, Sedat JW (2001) Multiple regimes of constrained chromosome motion are regulated in the interphase *Drosophila* nucleus. Curr Biol 11:1227–1239

West AG, Gaszner M, Felsenfeld G (2002) Insulators: many functions, many mechanisms. Genes Dev 16:271–288

Wollman AJ, Shashkova S, Hedlund EG et al (2017) Transcription factor clusters regulate genes in eukaryotic cells. eLIFE 6:e27451

Wu HY, Shyy SH, Wang JC, Liu LF (1988) Transcription generates positively and negatively supercoiled domains in the template. Cell 53:433–440

Zabet NR, Adryan B (2012a) GRiP: a computational tool to simulate transcription factor binding in prokaryotes. Bioinformatics 28:1287–1289

Zabet NR, Adryan B (2012b) A comprehensive computational model of facilitated diffusion in prokaryotes. Bioinformatics 28:1517–1524

Zabet NR, Adryan B (2013) The effects of transcription factor competition on gene regulation. Front Genet 4:197

Zimmer C, Fabre E (2011) Principles of chromosomal organization: lessons from yeast. J Cell Biol 192:723–733

Zuin J, Dixon JR, van der Reijden MIJA et al (2014) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc Natl Acad Sci U S A 111:996–1001

A Systems Perspective of Complex Diseases: From Reductionism to Integration

Khushdeep Bandesh, Pawan K. Dhar, and Dwaipayan Bharadwaj

Contents

Abstract Complex systems exist across all levels of biological organization ranging from the simplest (subatomic realm) to most complex (individual organism to whole populations and beyond). This complex nature of both the common diseases

K. Bandesh
Academy of Scientific and Innovative Research, CSIR-Institute of Genomics and Integrative Biology, New Delhi, India
e-mail: khushdeep@igib.in

P. K. Dhar
Synthetic Biology Group, School of Biotechnology, Jawaharlal Nehru University, New Delhi, India
e-mail: pawandhar@mail.jnu.ac.in

D. Bharadwaj (✉)
Systems Genomics Laboratory, School of Biotechnology, Jawaharlal Nehru University, New Delhi, India
e-mail: db@jnu.ac.in

N. Rajewsky et al. (eds.), *Systems Biology*, RNA Technologies,
https://doi.org/10.1007/978-3-319-92967-5_2

and the human beings has kept researchers far from a holistic understanding of underlying biological processes. Over the past decade, there has been a rapid and vast accumulation of large scale high-throughput biological data at physiological, cellular, molecular, and submolecular levels. It includes genetic association studies of complex human diseases and traits, quantification of genome-wide RNA expression patterns, comprehensive profiling of cellular proteins and metabolites, gene regulatory information (DNA methylation, histone modifications, chromatin accessibility, evolutionary constraint, etc.), and characterization of networks of molecular interactions. The clinical utility of such enormous data demands interpretation and understanding at the biological level to reveal mechanistic insights of molecular etiology. An important element of this task is to complement the detailed pieces of biological information with new advanced methods of system integration and reconstruction. This requires conversion of actual biological systems into computational models to make reliable predictions of biological responses following targeted manipulation under untested conditions. The frequency at which signals are presently being discovered mandates a systematic and integrative "omics" approach to bridge the "genotype to phenotype" gap. The chapter highlights the fundamental ways to integrate high-quality biological data that await systemic interpretations.

Keywords Complex systems · Common diseases · High-throughput data · Computational models · Systematic interpretation

1 Introduction to Systems Biology

1.1 Establishing the Need

The classic reductionist approach in biological sciences, generally known by the terms like molecular biology and biochemistry, has led to generation of enormous "parts-data." The collection of data has been aided by the parallel development of sequencing, structural, and expression measurement technologies. From low-throughput data collection, the community has reached high-throughput data collection, storage and analytical technologies.

The enormous success of reductionist approach has helped to determine the composition of the system and individual correlation of parts with a given phenotype, in a large number of situations. However, it has also thrown up a major challenge, i.e., to understand collective behavior of thousands of parts working together to maintain the functioning and robustness of a cell and an organism. The big challenge is to construct a large virtual matrix where biological components interact virtually and help understand biological decisions various scales and granularity.

Back in 1944, Norbert Weiner foresaw the need for a new approach that focused on stitching individual parts to describe collective response and coined the term "Systems Biology." Though the idea was novel and path-building, the time was not

yet ripe for launching a new approach, due to scarcity of data and computational resources.

The idea of systems approach again picked up in the mid-1960s and 1970s, when concepts like metabolic flux and control analysis gained traction. The aim was to study the flow of metabolites through a certain path/pathway and identify choke points that controlled the flux. A large body of literature during this time led to emergence of a new Biochemical Systems Theory.

The situation remained somewhat unchanged for the next few decades, till a new high-throughput technology of gene sequencing and expression measurement arrived. Biological sciences suddenly changed the stick shift and went into a higher gear of data gathering, management and analysis. The paradigm shift was greatly helped by parallel technological advancement in the computer industry. The storage got cheaper, processes got faster and algorithms were written to swim through oceans of data to find patterns.

The speed, scale and variety of data breathed life into Nobert Weiner's work of 1940s and "Systems Biology" as a formal discipline was launched. For many years the community debated on the concept, definition, scope and tools of the new systems approach. However, what emerged as a common thread was the acceptance that (a) collective behavior of biological parts was different than the sum-of-its-parts and (b) modeling in biology was essential to understand biological decisions, narrow down the range of experiments and generate hypothesis.

The biological community was beginning to sense the power of mathematics and computation that played a major role in the origin and evolution of engineering from physics. The need for modeling was also felt for the reasons that, on one hand, not enough experiments could be performed to collect all kinds of data in all kinds of contexts. On the other hand a lot of data in the published literature domain was inaccurate.

Here it may be relevant to introduce a few definitions.

1.2 What Is a Model?

A model is a representation of a system in a certain form that looks closest to the real life situation. The skeletal system of a model is made of components and their interactions. It is somewhat easy to define a static system in terms of components and interactions. However, the real challenge arrives when one moves from a static to a dynamic description, i.e., creating a movie out of snapshots arranged along a certain time series.

Modeling itself is an iterative process that goes on and on till experimental results match the modeling predictions. A model may be rigorous with mathematical representation or simply a sketch of nodes and arrows. It may depict a flow of information (as in metabolic pathways) or direction invariant (as in protein–protein interaction networks).

Furthermore, mathematical models may be deterministic (responses are predictable) or stochastic (responses are determined by probability distribution). Watching a model grow over an *x*-axis of time is called simulation. Adding mathematical muscles to a bare bone model is both an art and a science. One needs to be convinced of the flow of information in a certain way to adopt a certain modeling approach. Also, the choice of modeling method is governed by the kind of question one asks, the availability of data (qualitative to quantitative) and computational resources.

1.3 Steps in Building a Model

1. Make a parts list data from literature and annotate every part by including measurements, protocols, perturbations, constraints, and error bar. Here it is important to know if the data were independently confirmed.
2. Draw a parts-interaction map in the form of pathways. The map may represent translocation (ion channel), transformation (substrate–enzyme reaction), and binding events (transcription factor) in the form of nodes (molecules) and edges (interactions).
3. Use appropriate qualitative or quantitative methods to empower the power of conversation. Build conceptual, analytical models for simulation.

Apply perturbations at predefined points where phenotypic assays are possible and generate novel observations and hypothesis (Fig. 1).

1.4 Modeling Methods and Tools

Ever since the first conference of Systems Biology was held in Tokyo in 2000, a large number of tools have come up addressing various needs of the modeling community. Some of the most common resources and tools used are:

1. ***Pre Constructed Pathway Maps***
Kyoto Encyclopedia of genes and genomes http://www.genome.ad.jp/kegg/
BioCyc http://www.biocyc.org
BioCarta https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways

2. ***Enzyme Databases***
BRENDA http://www.brenda-enzymes.info/
ExPASy https://www.expasy.org/

3. ***Tools for Constructing, Simulating, and Analyzing Pathways***
http://sbml.org/SBML_Software_Guide/SBML_Software_Summary

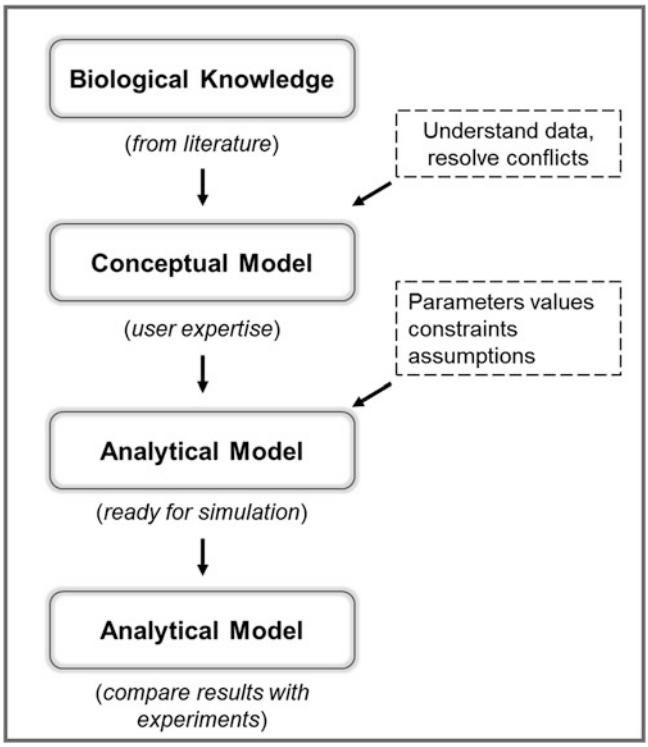

Fig. 1 A general strategy of building pathway models

In general, for modeling metabolic pathways, where large number of molecules interact (and data are frequently available) one uses ordinary differential equation based approach. For modeling gene expression events, where numbers are very less (transcription factor molecules) and fluctuations are high, the method of choice is stochastic. Some people also use ODE approach, as it comes with less computational cost. In situations where the large scale networks need to be modeled, rule based, fuzzy logic based, Boolean based and petri net approaches have been used with success. As the scale of the network increases in size the computational costs soar. To find the right balance, one may use a combination of qualitative (Boolean, rule based) and quantitative (ODE and stochastic) methods. Some of the issues that often emerge in quantitative modeling approaches are parameter estimation and optimization.

The need for a good parameter estimation method is felt more when the network data is incomplete, i.e., there is a space of unknown that needs to be considered and computed in the model. Several ODE and stochastic methods to estimate missing parameters are available. However, none of the methods can absolutely guarantee the accuracy of the output. One needs to feed in predicted data over and over again for optimization purposes. A fully parameterized and computationally optimized