Statistical analysis of networks and biophysical systems of complex architecture

Olga Valba

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

UNIVERSITE PARIS-SUD

ECOLE DOCTORALE DE PHYSIQUE DE LA RÈGION PARISIENNE – ED 107

LABORATOIRE DE PHYSIQUE THÈORIQUE ET MODÈLES STATISTIQUES

DISCIPLINE: Physique

THÈSE DE DOCTORAT

soutenue le 15 Octobre 2013

par

Olga VALBA

Statistical analysis of networks and biophysical systems of complex architecture

---

Directeur de thèse: Dr. Sergei NECHAEV

Composition du jury:

Rapporteurs : Prof. Riccardo ZECCHINA, Prof. Martin WEIGT

Examinateurs : Prof. Olivier MARTIN, Dr. Denis GREBENKOV

Abstract

Complex organization is found in many biological systems. For example, biopolymers could possess very hierarchic structure, which provides their functional peculiarity. Understating such complex organization allows to describe biological phenomena and predict molecule functions. Besides, we can try to characterize the specific phenomenon by some probabilistic quantities (variances, means, etc), assuming the primary biopolymer structure to be randomly formed according to some statistical distribution. Such a formulation is oriented toward evolutionary problems.

Artificially constructed biological network is another common object of statistical physics with rich functional properties. A behavior of cells is a consequence of complex interactions between its numerous components, such as DNA, RNA, proteins and small molecules. Cells use signaling pathways and regulatory mechanisms to coordinate multiple processes, allowing them to respond and to adapt to changing environment. Recent theoretical advances allow us to describe cellular network structure using graph concepts to reveal the principal organizational features shared with numerous non-biological networks.

The aim of this thesis is to develop bunch of methods for studying statistical and dynamic objects of complex architecture and, in particular, scale-free structures which have no characteristic spatial and/or time scale. For such systems the use of standard mathematical methods relying on the average behavior of the whole system, is often incorrect or useless, while a detailed many-body description is almost hopeless because of the combinatorial complexity of the problem. Here we focus on two problems.

The first part address to the statistical analysis of random biopolymers. Apart from the evolutionary context, our studies cover more general problems of planar topology appeared in description of various systems, ranging from gauge theory

to biophysics. We investigate analytically and numerically a phase transition of a generic planar matching problem, from the regime where almost all the vertices are paired, to the situation where a finite fraction of them remains unmatched.

In the second part of this work focus on statistical properties of networks. We demonstrate the possibility to define co-expression gene clusters within a network context from their specific motif distribution signatures. We also show how a method based on the shortest path function (SPF) can be applied to gene interactions sub-networks of co-expression gene clusters, to efficiently predict novel regulatory transcription factors (TFs). The biological significance of this method by applying it on groups of genes with a shared regulatory locus found by genomics is presented. Finally, we discuss formation of stable patters of motifs in networks under selective evolution in context of creation of islands of "superfamilies".

Résumé

De nombreux systèmes biologiques présentent une organisation complexe. Par exemple, les biopolymères peuvent posséder une structure très hiérarchisée responsable de leur fonction particulière. Comprendre la complexité de cette organisation permet de décrire des phénomènes biologiques et de prédire les fonctions des molécules. En outre, en supposant que la structure primaire du polymère est formée aléatoirement, nous pouvons essayer de caractériser ce phénomène par des grandeurs probabilistes (variances, moyennes, etc). Cette formulation est propre aux problèmes d'évolution. Les réseaux biologiques sont d'autres objets communs de la physique statistique possédant de riches propriétés fonctionnelles. Pour décrire un mécanisme biologique, on utilise différents types de réseaux biomoléculaires. Le développement de nouvelles approches peut nous aider structurer, représenter et interpréter des données expérimentales, comprendre les processus cellulaires et prédire la fonction d'une molécule.

L'objectif de cette thèse est de développer des méthodes pour l'étude d'objets statiques ou dynamiques, ayant une architecture complexe. Ici, nous nous intéressons deux problèmes.

La première partie est consacrée l'analyse statistique des biopolymères aléatoires. Nous étudions une transition de phase présente dans les séquences aléatoires de l'ARN. On met alors en évidence deux modes : le régime o presque toutes les bases qui composent l'ARN sont couplées et la situation o une fraction finie de ces bases restent non complémentaires.

La deuxième partie de cette thèse se concentre sur les propriétés statistiques des réseaux. Nous développons des méthodes pour l'identification d'amas de gènes co-expressifs sur les réseaux et la prédiction de gènes régulateurs novateurs. Pour cela, nous utilisons la fonction du plus court chemin et l'analyse du profil des motifs formés par ces amas. Ces méthodes ont pu prédire les facteurs de

transcription impliqués dans le processus de longévité. Enfin, nous discutons de la formation de motifs stables sur les réseaux due une évolution sélective.

Contents

Introduction

"I propose to develop first what you might call a naive physicist's ideas about organisms,that is, the ideas which might arise in the mind of a physicist who, after having learnt his physics and, more especially, the statistical foundation of his science, begins to think about organisms and about the way they behave and function and who comes to ask himself conscientiously whether he, from what he has learnt,from the point of view of his comparatively simple and clear and humble science, can make any relevant contributions to the question. It will turn out that he can. The next step must be to compare his theoretical anticipations with the biological facts."

by Erwin Schrodinger,
"What is life?", 1944.

One of main features of biophysical systems is the presence of selective interactions between their elements which could have a very complex spatially distributed architecture. Developing new mathematical methods for studying statistical and dynamic properties for such complex systems and, in particular, for structures which have no characteristic spatial and/or time scale is highly demanded. For such systems use of standard mathematical methods based on describing of the average behavior of the whole system, is often incorrect or useless, while a detailed many-body description is almost hopeless because of the combinatorial complexity of the problem. Situations like that are typical both for a number of fundamental biophysical and bioinformatic problems, as well as for distributed systems like networks (not obliged biological).

In this thesis we consider two important complementary problems.

The first part of the thesis deals with investigation of random biopolymers.

Many molecular biological phenomena are associated with generic properties and characteristics of polymers: chain structure, flexibility, volume interactions, topological constraints, etc. At the same time, in a number of biologic processes (which are definitely the most fundamental in living world), an important role belongs to certain specific features of the structure of biopolymer molecules themselves. One can regard a protein or a DNA chain not only as a molecule with peculiar chemical and conformational properties, but also as a weird machine, or automaton, executing certain operations. From the physical viewpoint, biopolymer molecules must possess very uncommon (often, hierarchical) structure to function properly. The problem of describing any phenomenon in such a complex system can be formulated in two ways.

1. One may search for a precise algorithm to characterize specific and completely known primary structure with all interactions. Such a formulation is intended for comparative analysis of *existing* biopolymers.

2. One may try to characterize the phenomenon by some probability quantities (variances, means, and so on), assuming the primary structure to be randomly formed according to some statistical distribution. Such a formulation pursues predictive goals and mostly is oriented toward evolutionary problems. The primary structures of real biopolymers are quite complex, for which the correlations between nucleotides are often unknown. So, even in studies of existing biopolymers with no evolutionary problems in mind, the random primary structure is frequently regarded as a fairly adequate model for real complex primary structures, or at least as a reasonable nontrivial starting point.

This part is devoted to random DNA and RNA sequences. In contrast to linear DNAs, RNA sequences can form hierarchical cloverleaf-like secondary structures. We analyze the statistical properties of random RNA-type sequences, in particular, mean and fluctuation of ground state free energy. The dependence of structure topology on the alphabet used in random sequence is considered in detail. This analysis oriented to the evolutionary questions (such as "Why natural alphabet consists of four different nucleotide types?", or "If there are grounds for a RNA-world hypothesis?").

In addition, we consider a model, in which a random RNA-type sequences has spatial disorder in gaps between distributed along a chain monomers (nucleotides). Using the optimization procedure for a special class of concave-type potentials, borrowed from optimal transport analysis, we derive the local difference equation for the ground state free energy of the chain with the planar (RNA-like) architecture of paired links. We consider various distribution functions of intervals between neighboring monomers and demonstrate the existence of a topological crossover from sequential to essentially nested configurations of sequentially paired links.

The second part of thesis deals with the statistical analysis of networks. The investigation of complex networks constitutes a rapidly developing interdisciplinary area, which unites study of various types of experimentally observed biological, social and engineering networks, as well as artificial random graphs constructed by various probabilistic techniques. Many statistical properties, including the vertex degree distribution, clustering coefficients, "small world" structure and spectra of adjacency matrices have been studied. Here we show how the statistical properties of networks can be used in analysis of biological networks. For the connectivity network in *C.elegance* we consider different statistical approaches which allow to predict regulators. Connectivity networks have become increasingly useful for biology because of the expanding availability of data on the physical and functional links between individual genes and proteins. This connectivity data enables to expand our knowledge beyond the experimentally validated results. New functional interactions can be predicted and tested by means of analysis and theoretical expectations. In this work we demonstrate an application of several algorithms developed for ranking of potential gene-expression regulators within the context of an associated network.

In this part we consider as well in detail motif distributions on example of random Erdös–Rényi networks. It is known that the existence of different three–vertex motifs (triads) in a directed network are tightly correlated with the network function. Namely, all naturally observed directed networks can be split into four broad *superfamilies* according to their motif distribution, and the networks within the same superfamily tend to have similar function. However, to the best of our knowledge, there is still no common opinion on why this clusterization into superfamilies happens and why some particular motif distributions are preferred

in natural networks. In this part we put forward a hypothesis which may give at least a partial answer to this question.

The thesis is organized in eight chapters. In the first chapter, principal characteristics of random biopolymers are described. We consider the main structural properties of RNA and DNA molecules. The existing approaches for the secondary RNA structure prediction are discussed in detail. Finally, we present thermodynamic properties of random RNAs and discuss the corresponding matrix method in context of RNA folding problem.

Our own algorithm is introduced in Chapter 2. We reveal the similarities and differences between computations of the free energy of associating heteropolymer complexes and standard matching algorithms. A new statistical method of alignment of two heteropolymers which can form hierarchical cloverleaf-like secondary structures is proposed. This offers a new constructive algorithm for quantitative determination of binding free energy of two RNAs with arbitrary primary sequences. The proposed algorithm is based on two observations: i) the standard alignment problem is considered as a zero-temperature limit of a more general statistical problem of binding of two associating heteropolymer chains; ii) this last problem is generalized onto the sequences with hierarchical cloverleaf-like structures (i.e. of RNA-type).

The Chapter 3 focuses on discussion of statistical properties of linear complexes and RNA-type complexes of two pairing RNAs. First, we discuss mean energy and energy fluctuations as functions of the sequence lengths in random RNA-RNA complexes. Next, we describe the model used to estimate binding probability for random sequence polymers in RNA-like complexes. Finally, we report the results of the analysis of the loop length distribution in complexes and propose models describing these distributions.

The Chapter 4 deals with statistics of a single random RNA chain. In particular, we consider the fraction of nucleotides involved in the formation of a cloverleaf secondary structure as a function of the number of different nucleotide species. We demonstrate the existence of the morphological phase transition in this system. The different models for estimation critical alphabet size are considered. In particular, we formulate the problem as the perfect matching problem in a random Erdos-Renyi graph, which allows us to estimate analytically the transition point. The analytical estimate is done from naive combinatorics viewpoint and

by the use of matrix theory description of RNA folding. Finally, the relevance of the transition from the evolutionary point of view is discussed.

In the Chapter 5 we describe a new toy model of a heteropolymer chain capable of forming planar secondary structures typical for RNA molecules. In this model, called the "random interval model" the sequential intervals between neighboring monomers along a chain are considered as quenched random variables, and energies of nonlocal bonds are assumed to be concave functions of those intervals. We demonstrate the possibility to pass from the nonlocal recursion relation for the ground state free energy to the local recursion relation.

The Chapter 6 summarizes the statistical methods for investigation of networks. We present the tools which are widely used in the analysis of the network architecture. Three structural properties are introduced: the degree distribution, the clustering coefficient, and the measure of path lengths in networks. Another structural property, the motif distribution, is discussed separately, because it plays a particular role in network analysis. We demonstrate how the different properties of networks can be interpreted from biological point of view. And finally, we discuss network superfamilies, appeared in investigation of three-node motif distributions in different networks.

In the Chapter 7 we analyze how different types of connectivity between genes and proteins affect the topology of the integral functional network of the free-living nematode *C.elegans*. We also show how a method based on the shortest path function (SPF) can be applied to gene interactions sub-networks of co-expression gene clusters, to efficiently predict novel regulatory transcription factors (TFs). To demonstrate the usefulness of our methods we predict the regulators for a cluster of ribosomal/ mRNA metabolic genes and highlight their potential relevance to regulation of longevity.

In the Chapter 8 we consider random non-directed Erdös–Rényi networks subject to a dynamics conserving vertex degrees and study analytically and numerically equilibrium three-vertex motif distributions in the presence of an external field coupled to one of the motifs.

Chapter 1

Random biopolymers

Biopolymers — natural macromolecular compounds, which serve as the structural basis of all living organisms [1]. Biopolymers include proteins, nucleic acids and polysaccharides, there are also mixed biopolymers — glycoproteins, lipoproteins, glycolipids, etc. The nucleic acids in a cell operate genetic functions. The sequence of monomer units, nucleotides, in a DNA determines (in a form of the genetic code) the sequence of monomer units (residues) and thus, the structure of an organism and biochemical processes in it. Genetic information is transferred from a DNA to a RNA synthesized on the DNA as a template, this process is called transcription [2]. A so-called messenger RNA (mRNA) serves as a template for protein synthesis. Biological variability required for evolution occurs on a molecular level by changes in a DNA. Proteins perform a vast array of functions within living organisms, including catalyzing metabolic reactions, replicating DNAs, responding to stimuli, and transporting molecules from one location to another.

Many molecular biological phenomena are associated with quite ordinary properties and characteristics of polymers — a chain structure, flexibility, volume interactions, topological constraints, and so on. At the same time, in a number of biophysical processes (which are definitely the most fundamental in living world), an important role belongs to certain specific features of the structure of biopolymer molecules themselves. One can regard a protein or a DNA chain not only as a molecule of some substance, but also as a machine, executing certain operations. From a physical viewpoint, biopolymers possess often a very unusual hierarchical structure to function properly. First, it is well-known that

each chain of the biopolymer must have a definite sequence of links of different types; this sequence is formed during biological synthesis and is called the primary structure of the biopolymer. Second, there should exist an opportunity to form a short-range order in the spatial arrangement of chain elements dictated by the interaction of neighboring links in the chain. Usually, the short-range order in biopolymers manifests itself in form of a helix, or small pins. These elements of short-range order are called the secondary structure. Third, a biopolymer chain as a whole must possess a more-or-less defined spatial or tertiary structure, and this defines the long-range spatial order in the arrangement of links.

The problem of describing any phenomenon in such a system (helix formation, globule-coil transition, and so on) can be formulated in two ways. First, one may look for an algorithm to find the characteristics of an investigated phenomenon for each specific and completely known primary structure. Such a formulation is intended for research of biopolymers with perfectly known primary structure. For example, we may study the melting curve of a specific part of a DNA with known primary structure, or predict secondary and tertiary structures of globular proteins having known its primary structure. The picture of the registered phenomenon is often insensitive to particular details of the primary structure; and it is then relevant to resort to "linguistic" analysis to find the necessary rough characteristics of the "text" encoded in the primary sequence. Second, one may try to characterize the phenomenon by some probability quantities (variances, means, and so on), assuming the primary structure to be randomly formed according to some statistical distribution. Such a formulation is oriented toward evolutionary problems. It should be noted that the primary structures of real biopolymers are quite complex with hidden correlations between the "letters" constituting the sequences. Thus even in studies of existing biopolymers with no evolutionary problems in mind, the random primary structure is frequently regarded as a "playing ground" and fairly adequate model for real complex primary structures.

1.1 The ground state of a biopolymer: energy and structure

The problem of the secondary structure prediction is a central in bioinformatics. The most difficult this problem is for RNA molecules and proteins: their

secondary structures can have different elements, distinguished by energy and entropic contributions (Fig. 1.1) [3, 4, 5, 6]. For proteins, for example, the prob-

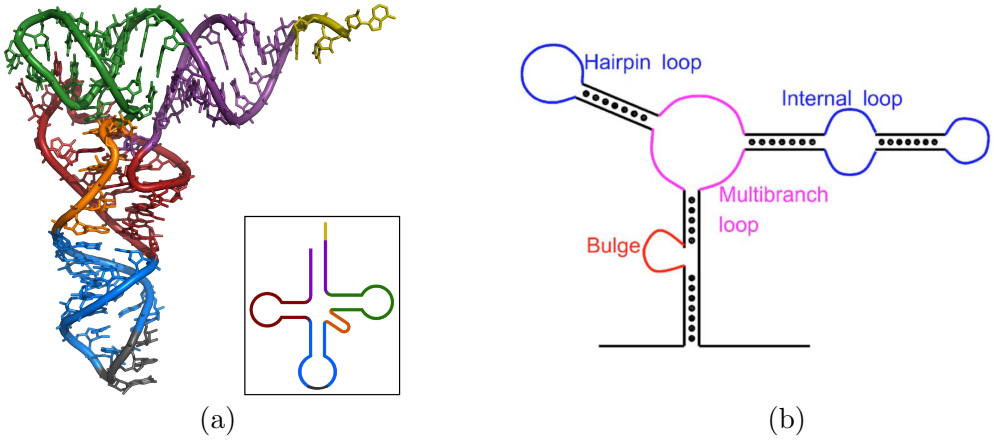


Figure 1.1. The secondary structure of RNA. (a) The X-ray structure of the tRNA from yeast (Image: Yikrazuul/wikipedia); (b) the structural elements of RNA molecule [7].

lem is even more complicated because there are no strong fixed hydrogen links (like base pairs in DNA and RNA), secondary structure is stabilized by the variety of interactions: covalent, ion, hydrogen and hydrophilic-hydrophobic bonds. Further, we will be interested in RNA-type structures, so we focus here on topological features of RNA structures and existing approaches of their prediction.

1.1.1 The secondary structure properties of RNA molecules

Ribonucleic acid is a ubiquitous family of large biological molecules that perform multiple vital roles in the coding, decoding, regulation, and expression of genes. Together with DNA, RNA comprises the nucleic acids. Cellular organisms use messenger RNA (mRNA) to convey genetic information (often notated using the letters G, A, U, and C for the nucleotides guanine, adenine, uracil and cytosine) that directs synthesis of specific proteins, while many viruses encode their genetic information using an RNA genome. Some RNA molecules play an active role within cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals.

RNAs are the specific examples of a wide class of so-called "associating" heteropolymers. Generally speaking, we call a polymer "associating" if, besides the strong covalent interactions responsible for the frozen primary sequence of

monomer units, it is capable of forming additional weaker reversible temperature-dependent ("thermo-reversible") bonds between different monomers. For associating polymers the variety of possible thermodynamic states ("secondary and ternary structures", in biological terminology) is determined by the interplay between the following three major factors: i) the energy gain due to the direct "pairing", i.e. formation of thermo-reversible contacts; ii) the combinatoric entropy gain due to the choice of which particular monomers (among those able to participate in bonds formation) do actually create bonds; iii) the loss of conformational entropy of the polymer chain due to pairing (and in particular, the entropic penalty of loop creation between two paired monomers) [8]. The hydrogen bonding and stacking interactions of the hydrophobic nucleobases are major contributors to the stable association of nucleotides within and between nucleic acid molecules. Hydrogen bonds are principally characterized by highly specific electrostatic interactions that stabilize the nucleic acid secondary structure [9]. Watson-Crick hydrogen bonds between the bases of the nucleosides — adenine (A) and uracil (U) or thymine, and guanine (G) and cytosine (C), and a multitude of non-canonical hydrogen bonds, play crucial roles in both the secondary and tertiary structures of nucleic acids and in their functions [10]. Non-canonical pairs are called wobble base pairs, the most common of them is U–G (Fig. 1.2). Wobble base pairs can essentially affect on the secondary structure of RNA and its function. In particular, it was shown, that these bonds play a role in a codon-anticodon binding process [2]. Theoretical calculations showed that the energy of U–G bond is comparable with the canonical base pair energies [11] (Tab. 1.1, however the geometry of wobble bond is quite different from Watson-Crick ones [12]. G–U and A–U base pairs differ from wobble G–U pairs in the type and location of functional groups that are projected into major and minor grooves. These pairs also differ in the orientation of the bases with respect to the phosphodiester backbone. Whereas the glycosidic angle is similar for all nucleotides in Watson-Crick pairs, both angles for G and U differ in the wobble pair.

The structural feature of RNA is that the system of formed bonds can be presented as a set of the nested arcs (see for the details Section 2.3). A pseudoknot containing two stem-loop structures in which half of one stem is intercalated between the two halves of another stem corresponds intersection in arc representation. Pseudoknots appear in RNA folding quite rarely (less then 1%),

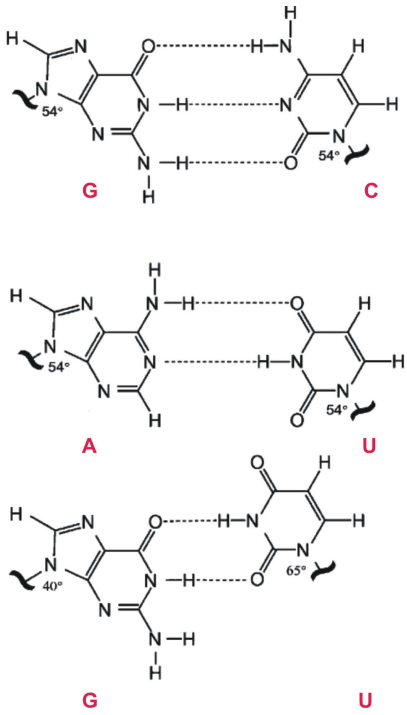


Figure 1.2. Watson-Crick base pairs and wobble pair.

Pair	$\Delta G^0_{300^\circ K}$, kcal/mol	d, Å
G–C	-5.53	2.94
A–U	-4.42	2.96
G–U	-4.45	3.75

Table 1.1. Energies and hydrogen bond distances for pairing of nucleotides [11].

however they bring an important functional role. For example, the telomerase RNA component contains a pseudoknot that is critical for activity [13], and several viruses use a pseudoknot structure to form a tRNA-like motif to infiltrate the host cell [14].

1.1.2 RNA structure prediction methods

Huge number of various algorithms have been developed for the prediction of RNA secondary structure from its primary sequence. In theory, the number of valid secondary structures for a given sequence is greater than 1.8^N, where N is the number of nucleotides [15]. Most folding programs fit into one or more of four classes:

(i) "Basic" algorithms predict hairpin and simple loop formation, but they exclude the prediction of multi-branched loops and perform very basic energy minimization. The first algorithms written were of this type, and most have been updated or are no longer in use.

(ii) "Combinatorial" methods generate lists of all possible secondary structure

elements and piece them together in all possible ways to find those with the lowest free energy.

(iii) "Recursive" algorithms build the secondary structure one nucleotide at a time while computing minimum energies along the way. Dynamic programs, which employ recursive algorithms, compute folding in time based on low energy paths of achieving secondary structure.

(iv) "Comparative sequence analysis" algorithms find conserved structure for a set of sequences using stochastic optimization on a population of tentative solutions.

The most popular method based on free energy minimisation [3, 5, 16, 4, 17] was proposed by M. Zuker and P. Stiegler [3]. The basic idea is that the true secondary structure must be thermodynamically stable and, thus have a minimum of free energy. To solve the free energy minimization problem the rules to calculate a pairing energy between nucleotides and effective algorithm of energy minimization are required. There were many attempts to build the rules for pairing energy calculation based on experimental data [5], the respective algorithms of energy minimization based on dynamical programming were developed [18]. The main equation for the partition function of RNA sequence is written as

$$g_{i,i+k} = g_{i+1,i+k} + \sum_{s=i+1}^{i+k} \beta_{i,s} g_{i+1,s-1}\, g_{s+1,i+k} \tag{1.1}$$

where $g_{i,j}$ describe statistical weight of the loop between i-th monomer and j-th, and $\beta_{i,j}$ is the Boltzmann weight of the contact between these nucleotides. Ground state energy is defined respectively as $f_{i,j} = k_B T \ln g_{i,j}$. Because of the base energy exceeds in ten times the room temperature $k_B T$ (Tab. 1.1), so-called zero temperature limit is applied very often. In this approximation, the ground state is defined by the number of base pairs, chain entropy is neglected. Expression (1.1) can be extended by different factors like the minimal loop length, the staking energy, the different entropies of structural RNA elements (Fig. 1.1(b)); a particular case is prediction of pseudoknots [19, 20]. The minimization energy methods nowadays are the most commonly used. But, unfortunately, these algorithms are not exhaustive, and their accuracy decreases strongly with increasing

the sequence length. It should also be noted that an approach quantitatively assessing the probability of erroneous modeling does not exist yet.

There is a set of other algorithms to find the optimal structure using stochastic optimization techniques, in particular genetic algorithms. One of the modern approaches based on the analysis of folding kinetics of RNA [21]. In contrast to the energy minimization methods not the most stable structure, but the structure of the kinetically accessible for folding is considered. The wide analysis was not performed for this approach, but despite the physical clarity, this method contains a lot of unaccounted factors.

Finally, there is a so-called "biological" approach based on the idea that biologically important secondary structures must be preserved in evolution [22]. In this approach, the set of sequences of the same biological role must be analyzed. However, for analyzing a plurality of polymers energy minimization algorithms are often used. To conclude, the RNA secondary structure prediction problem is still open question, and research in this area is ongoing [23, 24, 25, 26, 27].

A particular place among the prediction problems takes the problem of pairing of RNA with other biopolymers (DNAs, RNAs, proteins).

1.1.3 Pairing of biopolymers

Fig. 1.3(a) presents the basic interaction types between two biopolymers. The role of DNAs and RNAs in the mechanisms of cell regulation is well known. Their interaction is one of the necessary stages of the cell cycle associated with the storage and transmission of genetic information. Besides the well-known mechanisms of translation and transcription of genetic information, which involve DNA–RNA complexes (Fig. 1.3(b)), an exceptionally important role plays RNA–RNA interactions. These interactions have a crucial value for regulation of gene expression [28, 29, 30]. Schematically, the formation of an RNA–RNA complex occurs by complementary base pairing of an RNA with a messenger RNA (mRNA) or its segment, which precludes translation from the mRNA [28]. The RNA molecules participating in processes of this type are called noncoding RNAs (ncRNAs) because they are not themselves translated into proteins [29] and are therefore left out of translation process.

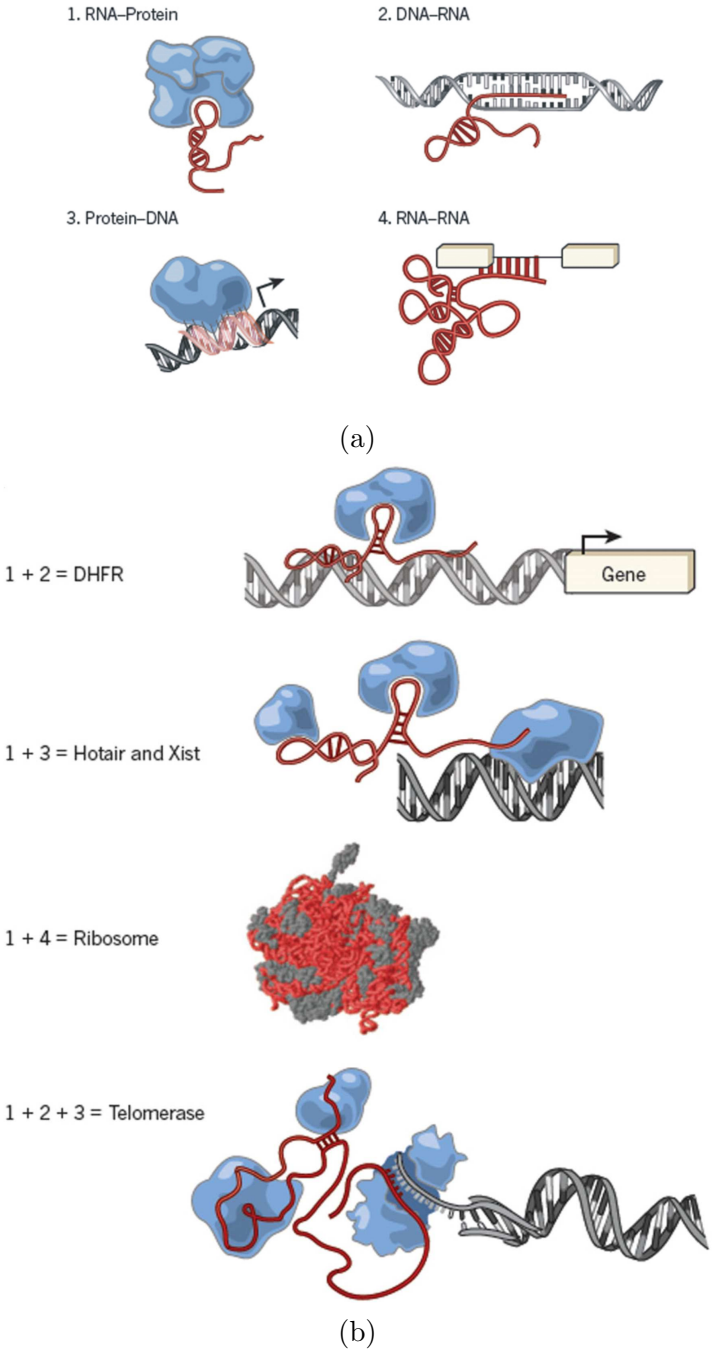


Figure 1.3. The four principles of nucleic acid and protein interactions (a) and their functions (b). (1) RNAprotein interactions, (2) DNA–RNA hybridization-based interactions, (3) DNA–protein interactions and (4) RNA–RNA hybridization based interactions [31]. (b) Each of these performs various functions in a cell. For example, RNA–protein interactions and DNA–RNA binding lead to localization of the protein in specific DNA and thus regulate transcription from this DNA [32, 33]; Interaction of RNA–protein complex with a complex protein-DNA leads to a set of proteins, which are specific to concrete DNA through the formation of RNA–DNA bonds; ribosome is a multifaceted complex of RNA–protein interactions, whose functionality provides by "correct" RNA–RNA binding; regulation of telomerase replication is an example of combining RNA–protein, RNA–DNA and protein–DNA interactions [34].

In view of the important role of RNA–RNA interactions in biological processes, an efficient algorithm is required for theoretically calculating the RNA–RNA binding energy given the primary sequences, as well as for predicting ncRNA secondary structure (i.e., thermodynamically optimal intra-chain bonding architecture). It is shown below that this problem is closely related to the alignment problem for two linear sequences (DNA-type). One important distinction of RNA alignment from the analogous problem for DNA is the existence of nontrivial secondary structure of RNA molecules (Fig. 1.1).

There are many different approaches for prediction energy and structure of RNA–RNA pairing complex [35, 36, 37, 38, 39]. However, they are applicable in narrow class of RNA molecules and "work" well just in specific samples. In analogy to the RNA structure prediction problem, the efficiency here also depends on a choice of conditions.

Certainly, the disadvantages of a particular method may be important for predicting the structure of a concrete RNA molecule and lead to incorrect conclusions about its function. However, for the study of statistical properties of random RNA sequences, i.e. chains with random primary structures, is sufficient to take into account the fundamental properties of the biopolymer. For RNA, primarily, it is a complex hierarchical cloverleaf-type structure formed according to the complementarity rules. And we can neglect that properties which affect on the structure (and function) in a definite molecule — pseudoknots, stacking interactions and others.

1.2 Statistical properties of biopolymers

In this work, we are not concerned with the structure formed by a specific sequence. Instead, we will consider statistical properties of secondary structures of long (more than 1000 monomers) random RNA sequences. This research plays an important role, for example, in understanding how close or far in evolutionary terms random RNAs from real molecules [40], or what properties are the most essential for performing a definite function and finally, how functional RNAs could appear from random ones [41]. Random RNAs are also quite traceable system for the study of thermodynamical properties. Investigations of phase transition [42, 43, 44], chain response on external force [45, 46] are based on the model of a random polymer. The essential advantage of random polymer is the opportunity to solve the problems as numerically as well analytically. In this section we focus on two problems in statistical physics of random RNAs: thermodynamical properties and description of RNA folding in terms of random matrix theory.

1.2.1 Thermodynamical properties of random RNAs

For random heteropolymer sequence the important stages were performed in the study of thermodynamical properties of RNA molecules. These results are widely used in the development of screening methods of experimental data for detection of genetic markers of diseases [47], sequencing of single nucleotide polymorphisms, selection of optimal condition in hybridization and cloning experiments [48, 49]. In addition, the development of DNA chips for rapid screening and sequencing is based on the ability to predict thermodynamical stability of complexes formed by oligonucleotide probes [50, 51].

From the pioneering works of R. Bundschuh and T. Hwa [42, 52, 43], the study of thermodynamical properties of random RNAs were considered by many researchers [44, 53, 54, 55]. To date, it is accepted to believe that random RNA undergoes a phase transition from the high temperature *molten* phase to the low temperature *glass* phase. Based on the replica analysis, M. Lassig and K.J. Wiese [56] and F. David and K.J. Wiese [57] formulated the problem of the transition in terms of field theory. We give below the arguments of R. Bundschuh and T. Hwa, proving the existence of a phase transition and discuss the properties of the different phases.

In the high-temperature phase the system remains in the molten phase, characterized by a flat free energy landscape and a homopolymer-like behavior. In the molten phase the disorder is irrelevant, and all the binding energies can be replaced by some effective value ε. Carrying out the two-replica calculation, authors [43] were able to prove that the system exhibits a phase transition from a high-temperature regime, in which the replicas are independent, to a lower temperature phase, in which the disorder is relevant and replicas are strongly coupled. The authors numerically characterized the transition to a glassy phase by imposing a pinch between two bases and measuring the corresponding energy cost (Fig. 1.4(a)).

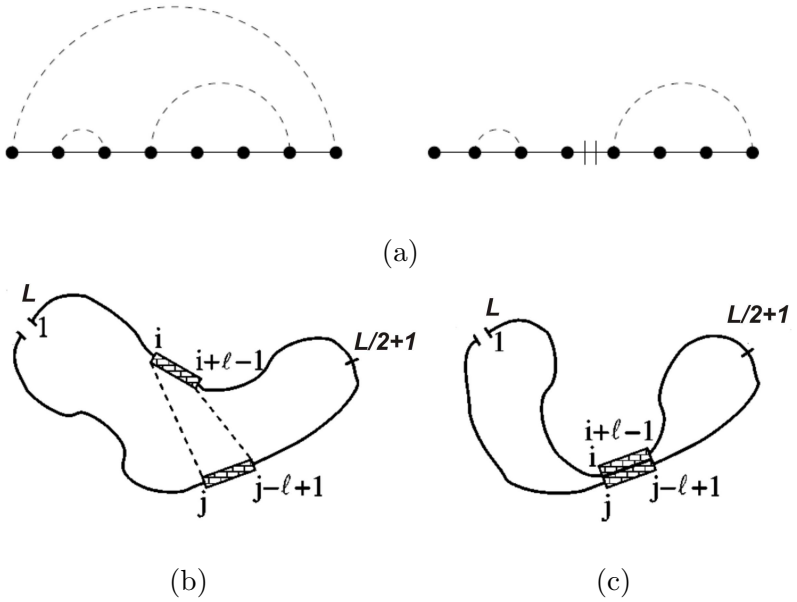


Figure 1.4. The pinch energy calculation procedure. (a) division of a chain into halves restrict the number of possible configurations; (b) shows the positions of two pieces with exactly complementary bases, one of which is between positions 1 and $L/2$ and the other of which is between positions $L/2+1$ and L. Such a piece of length $l \sim \ln L$ can be found for almost all sequences; (c) shows how restricting configurations to those in which the good match forms Watson-Crick base pairs splits the molecule into two loops, which can still form base pairs within the loops independently of each other [43]

Consider a pinch energy, which is $\Delta F(L) = k_B T \ln P_{1,L/2}$, where $P_{1,L/2}$ is the probability of binding between 1 and $L/2$ monomer of the chain of the length L. The pinch energy can be presented as

$$\Delta F(L) = F_{1,L} - (F_{1,L/2} + F_{L/2+1,L}) \tag{1.2}$$

In the high temperature regime the probability $P(L/2)$ of the contact depends only on the distance between the monomers, one just needs to recognize that $P(L)$ corresponds in the random walk analogy to the first-return probability of a random walk after L steps

$$P(L/2) = \frac{(L/2)^{-\frac{3}{2}}(L/2)^{-\frac{3}{2}}}{L^{-\frac{3}{2}}} \tag{1.3}$$

So the pinch energy is

$$\overline{\Delta F}(L) \sim \frac{3}{2} k_B T \ln L \tag{1.4}$$

The linearity of $\Delta F(L)$ is broken at some point T_g, which is the phase transition temperature. For determination of this temperature the following procedure is usually used. The function $\Delta F(L)$ is fitted by a linear law

$$\overline{\Delta F}(L) = a(T) \ln L + b(T) \tag{1.5}$$

Secondly, the dependence of the slope $a(T)$ on the temperature is built. In the high temperature phase with good accuracy $a(T) = \frac{3}{2}T$ (Fig. 1.5), as in low temperature, the pinch energy is defined by disorder in a polymer (Fig. 1.4(b,c)). The probability to have a contact between 1-st and $L/2$-th monomers is deter-

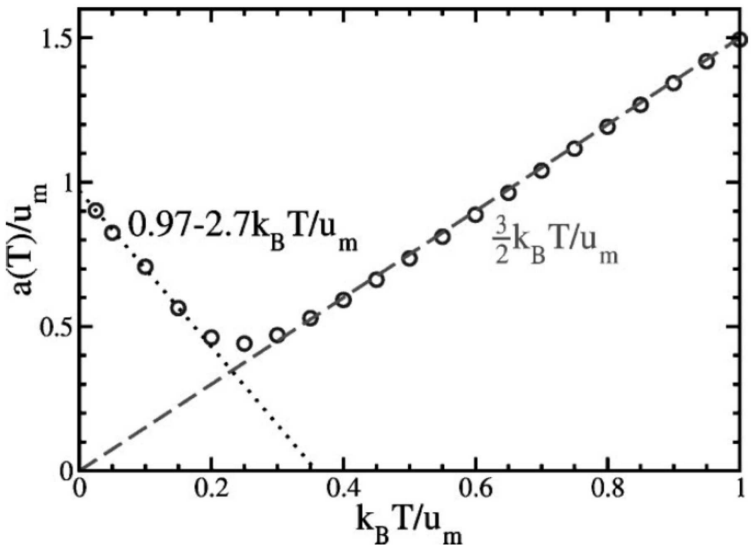


Figure 1.5. Prefactor a(T) of the logarithmic dependence (1.5) for random RNA sequences generated by the sequence disorder model. At high temperatures, the prefactor indicated by the circle is well described by the dashed line $(3/2)k_BT$ expected of the molten phase [43]

mined by energetic factors, such as a specific energy per a pair of nucleotides, average number of gaps. The linear dependence of $a(T)$ is broken (Fig. 1.5). It was proposed that the pinch energy in low temperature phase also have quadratic correction in the logarithm [54].

The transition between the high temperature molten phase and the low temperature glass phase belongs to the continuous phase transition of the second order [56]. It was shown that the transition temperature depends on the average number of gaps in the ground state structure [43]. But, analytical estimates of the transition point differs from numerical result by a order [43].

The high temperature and the low temperature phases have different scaling properties. One of the quantities that describes a secondary structures, is its "size profile". This value measures the maximal number of pairing one has to cross to go from a pair involving base 1 to the base with the maximal extent of nesting. This quantity can be very easily visualized as the "height" of the mountain representation of the secondary structure as shown in Fig. 1.6.

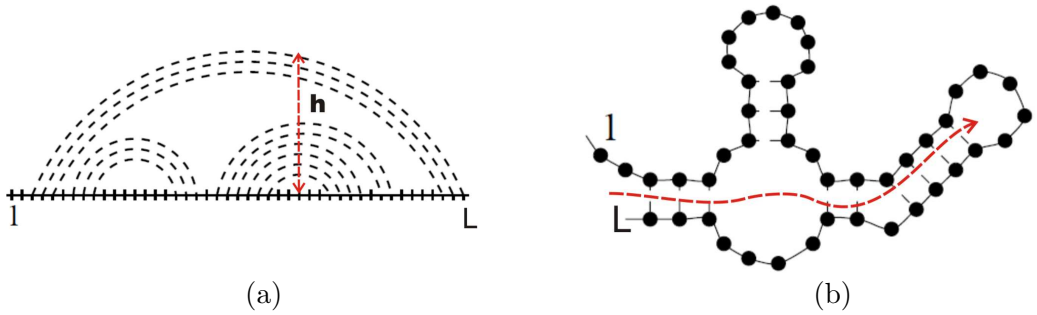


Figure 1.6. Characteristic size of RNA secondary structure. The height diagram (a) in arc representation is defined by the number of base pairs to be unzipped in the largest loop (b) [42].

It was shown numerically [43] and then proved analytically [57], that the low temperature phase is characterized by a power law dependence of h on the sequence length:

$$< h > \sim L^{\zeta} \tag{1.6}$$

with the exponent $\zeta \approx 0.64$, that is close to $\zeta_0 = 2/3$, and points to Kardar-Parizi Zhang class universality [58]. This scaling behavior is typical for correlated processes like surface growth or ballistic deposition [59]. High temperature regime is characterized by $\zeta \approx 0.54$ [43], which is in agreement with the expected

value $\zeta_0 = 1/2$, describing a size of a random polymer coil [60].

Several other works [53, 61] used an alternative, so-called "coupling" method, to investigate the nature and the scaling laws of the glassy phase, observing the effect of typical excitations imposed by a bulk perturbation. The authors argued that for the models with non-degenerate ground states, the low-temperature phase is not marginal, but is governed by a scaling exponent, close to $\theta = 1/3$. The explicit numerical studies of the specific heat demonstrate that the molten-glass transition is only a fourth order phase transition [44].

It is worth noting that an essential role in analytic study of the scaling properties deals with the matrix approach. The field theory analysis allows to suggest that the primary freezing occurs above some critical temperature, with local islands of stable folds forming within the molten phase [56]. The next subsection is devoted to matrix description of secondary structure of random RNA.

1.2.2 RNA folding and matrix theory

We formulate the RNA folding problem as a $N \times N$ matrix field theory. This matrix formalism allows us to give a systematic classification of the terms in the partition function according to their topological character. The theory is set up in such a way that the limit $N \to \infty$ yields the clover-leaf secondary structure (Hartree theory). Tertiary structure and pseudo-knots are obtained by calculating the $1/N^2$ corrections to the partition function.

For our purposes we could think of the nucleotides as beads on a flexible chain, with the beads to be "glued" together in pairs. As discussed in detail in [62], the partition function of a random chain can be represented as follows:

$$Z_L = 1 + \sum_{<ij>} V_{i,j} + \sum_{<ijkl>} V_{i,j}V_{k,l} + \sum_{<ikjl>} V_{i,j}V_{k,l} + ... \tag{1.7}$$

where $V_{i,j} = exp(-\beta\epsilon_{i,j})$ is the Boltzmann weight of the contact (i, j) with respective energy $\epsilon_{i,j}$; $< ij >$ denote all pairs with $i < j$, $< ijkl >$ all quadruplets with $i < j < k < l$ and so on. Each term in (1.7) corresponds to its own arc configuration. In this representation, the chain nucleotides are points, oriented from the 5‘ to 3‘ end of the molecule and each base pair in the structure is the arc between respective points. Diagram, consisting of non-intersecting arcs are called *planar*.

The planar diagrams play an important role in various branches of theoretical physics, such as matrix and gauge theories [63], many-body condensed matter physics[64], quantum spin chains [65] and random matrix theory [66].

The basic idea of matrix description is the following [62]. Consider the quantity:

$$Z_L(N) = \frac{1}{A_L(N)} \int \prod_{k=1}^{L} d\phi_k e^{-\frac{N}{2}\sum_{i,j}(V^{-1})_{i,j} tr(\phi_i \phi_j)} \frac{1}{N} tr \prod_{l=1}^{L}(1+\phi_l) \tag{1.8}$$

Here $\phi_i, i = 1, ..., L$ denote L independent $N \times N$ Hermitian matrices and $\prod_{l=1}^{L}(1+\phi_l)$ represents the ordered matrix product $(1+\phi_1)(1+\phi_2)...(1+\phi_L)$. The normalization factor $A(L)$ is defined by

$$L_L(N) = \int \prod_{k=1}^{L} d\phi_k e^{-\frac{N}{2}\sum_{i,j}(V^{-1})_{i,j} tr(\phi_i \phi_j)} \tag{1.9}$$

and V is symmetric $L \times L$ matrix with the elements $V_{i,j}$. The matrix integral defines a matrix theory with L matrices. The integral (1.8) evaluates precisely to the infinite series

$$Z_L(N) = 1 + \sum_{<ij>} V_{i,j} + \sum_{<ijkl>} V_{i,j}V_{k,l} + \frac{1}{N^2} \sum_{<ikjl>} V_{i,j}V_{k,l} + ... \tag{1.10}$$

The relation with the expansion in (1.7) is obvious. The two series coincide for $N = 1$, whereas for $N > 1$ the series in (1.7) contains topological information. All the planar structures are given by the $O(1)$ term of (1.7) and higher-order terms in $1/N^2$ correspond to RNA secondary structures with pseudoknots. The classification of pseudoknots induced by this expansion is reviewed in [67].

Generically for given RNA, described by the matrix $V_{i,j}$, the problem of determination of all possible configuration is quite complicated. In order to get exact results, one makes a set of additional simplifications. We can assume that any possible pairing between nucleotides is allowed (independently of the type of nucleotides and from their distance along the chain) and that all the pairings may occur with the same probability. In other words, we assume that the matrix $V_{i,j}$ has all entries equal with $\upsilon > 0$. In this case each term in (1.10) provides an information about topology and number of contacts in a structure (Tab. 1.2). The multidimensional integral (1.8) can be converted by the Hubbard–Stratonovich

transformation into the one-dimensional integral, including the spectral density of Gaussian random matrices [68]. The expression for the spectral density is known from random matrix theory [69] and, thus, the integral (1.8) can be calculated exactly. Let us illustrate this point by a simple example for $L = 4$. In this case the partition function $Z_4(N) = 1 + 6\upsilon + 2\upsilon^2 + \upsilon^2/N^2$ and all possible contact structures are listed in Fig. 1.7.

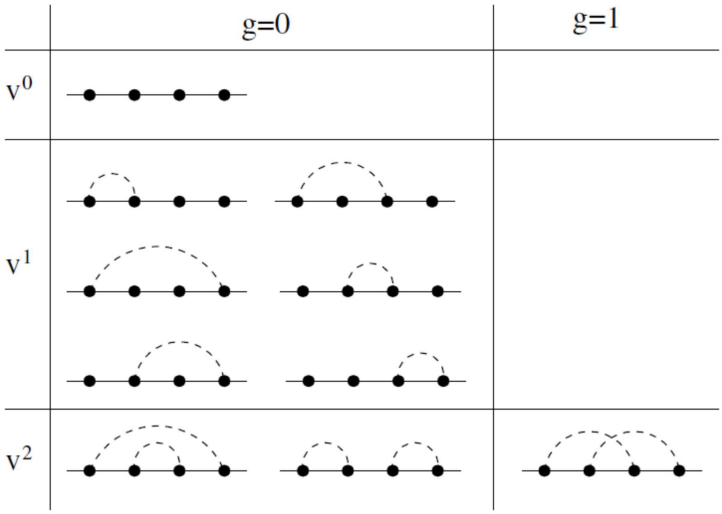


Figure 1.7. All possible arc diagrams with L = 4. Diagrams with i arcs are associated with the power v^i, and g is the genus.

In general case $1/N^2$ expansion of partition function $Z_L(N)$ can be written

$$Z_L(N) = \sum_{L=0}^{\infty} \frac{a_{l,g}\upsilon^l}{N^{2g}} \tag{1.11}$$

where the coefficients $a_{l,g}$ give exactly the number of diagrams at fixed length L and fixed genus g ($g = 0$ for planar diagrams, $g = 1$ for configurations with one pseudoknot and so on.) with l base pairs (arcs) (Tab. 1.2).

Expansion of $Z_L(N)$ for the orders higher than $1/N^2$ counts structures with complex pseudoknots ($g > 2$). The theory of these configurations is given in [71].

L	$Z_L(N)$
1	1
2	$1+\upsilon$
3	$1+3\upsilon$
4	$1+6\upsilon+2\upsilon^2+\upsilon^2/N^2$
5	$1+10\upsilon+10\upsilon^2+5\upsilon^2/N^2$
6	$1+15\upsilon+30\upsilon^2+5\upsilon^3+(15\upsilon^2+10\upsilon^3)/N^2$
7	$1+21\upsilon+70\upsilon^2+35\upsilon^3+(35\upsilon^2+70\upsilon^3)/N^2$
8	$1+28\upsilon+140\upsilon^2+140\upsilon^3+14\upsilon^4+$ $(70\upsilon^2+280\upsilon^3+70\upsilon^4)/N^2+21\upsilon^4/N^2$

Table 1.2. $1/N^2$ expansion of partition function (1.8) for different sequence length L [70].

Chapter 2

Algorithm for RNA energy and structure prediction

In this chapter we reveal the similarities and differences between computations of the free energy of associating heteropolymer complexes and standard matching algorithms. The matching (or "alignment") problem, even for linear structures is one of the key tasks of computational evolutionary biology. In particular, one of the most important applications of Longest Common Subsequence (LCS) search in linear structures is a quantitative definition of a "closeness" of two DNA sequences. Such a comparison provides information about how far, in evolutionary terms, two genes of one parent have deviated from each other. Also, when a new DNA molecule is sequenced *in vitro*, it is important to know whether it is really new or it is similar to already existing molecules [72, 73]. This is achieved quantitatively by measuring the LCS of the new molecule with other ones available from databases. The task of the present work consists of extending the statistical approach developed for alignment of linear sequences to the computation of pairing free energy of two RNA–type structures. First, we reformulate a pairwise sequence alignment problem as a problem of finding the ground state free energy (i.e., in the limit of $T \to 0$) in a statistical model of linear polymer associates (complexes). Second, we take into account the capability of each polymer to fold in a hierarchical RNA-like structures and derive an expression for the partition function of such a complex at nonzero temperature. Third, going back to the limit of $T \to 0$ in the final expression, we obtain nonlocal recursion relations for determining the ground state energy of a RNA-RNA complex. The developed

algorithm also allows to reconstruct the optimal secondary RNA structure (and the structure of RNA-RNA complex). We demonstrate the recovery procedure in detail and discuss on example of two real RNAs.

2.1 Alignment of linear sequences

The problem of finding the LCS of a pair of linear sequences drawn from the alphabet of c letters is formulated as follows. Consider two sequences $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ (of the length m) and $\beta = \{\beta_1, \beta_2, \ldots, \beta_n\}$ (of the length n). For example, let α and β be two random sequences of $c = 4$ base pairs A, C, G, T of a DNA molecule, e.g., $\alpha = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{C}, \mathrm{T}, \mathrm{A}, \mathrm{C}\}$ with $m = 6$ and $\beta = \{\mathrm{C}, \mathrm{T}, \mathrm{G}, \mathrm{A}, \mathrm{C}\}$ with $n = 5$. Any subsequence of α (or β) is an ordered sublist of α (and of β) entries which need not to be consecutive, e.g, it could be $\{\mathrm{C}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}$, but not $\{\mathrm{T}, \mathrm{G}, \mathrm{C}\}$. A common subsequence of two sequences α and β is a subsequence of both of them. For example, the subsequence $\{\mathrm{C}, \mathrm{G}, \mathrm{A}, \mathrm{C}\}$ is a common subsequence of both α and β. There are many possible common subsequences of a pair of initial sequences. The aim of the LCS problem is to find the longest of them. This problem and its variants have been widely studied in biology [74, 75, 76, 77], computer science [78, 79, 80, 81], probability theory [82, 83, 84, 85, 86, 87] and more recently in statistical physics [88, 89, 90, 42].

The basis of dynamic programming algorithms for comparing genetic sequences has been formulated for the first time in [91] (see also [92]). In general setting this algorithm takes into account the number of perfect matches and the difference between mismatches and gaps. Being formulated in statistical terms, it consists in constructing the "cost function", F, having a meaning of an energy (see, for example [93, 89] for details)

$$F = N_{\mathrm{match}} + \mu\, N_{\mathrm{mis}} + \delta\, N_{\mathrm{gap}} \tag{2.1}$$

Where N_{match}, N_{mis} and N_{gap} are respectively the numbers of matches, mismatches and gaps in a given pair of sequences, and μ and δ are respectively the energies of mismatches and gaps. Without the loss of generality, the energy of matches can be always set to 1. Besides (2.1) we have an obvious conservation

law

$$n + m = 2N_{\text{match}} + 2N_{\text{mis}} + N_{\text{gap}} \tag{2.2}$$

which allows one to exclude N_{gap} from (2.1) and rewrite it as follows:

$$F = N_{\text{match}} + \mu N_{\text{mis}} + \delta(n + m - 2N_{\text{match}} - 2N_{\text{mis}}) = \\ = (1 - 2\delta)N_{\text{match}} + (\mu - 2\delta)N_{\text{mis}} + \text{const} \tag{2.3}$$

In (2.3) the irrelevant constant $\delta(n + m)$ can be dropped out.

Adopting $(1 - 2\delta)$ as a unit of energy, we arrive at the following expression

$$\tilde{F} = N_{\text{match}} + \gamma N_{\text{mis}} \tag{2.4}$$

where

$$\gamma = \frac{\mu - 2\delta}{1 - 2\delta} \tag{2.5}$$

and $\gamma \leq 1$ by definition. The interesting region is $0 \leq \gamma \leq 1$, otherwise there are no mismatches at all in the ground state (i.e., there is no difference between $\gamma = 0$, which corresponds to simplest version of the LCS problem, and $\gamma < 0$).

It is known [89, 52] that the maximal cost function

$$\tilde{F}^{\max} = \max\left[N_{\text{match}} + \gamma N_{\text{mis}}\right] \tag{2.6}$$

can be computed recursively using the "dynamic programming"

$$\tilde{F}^{\max}_{m,n} = \max\left[\tilde{F}^{\max}_{m-1,n},\, \tilde{F}^{\max}_{m,n-1},\, \tilde{F}^{\max}_{m-1,n-1} + \zeta_{m,n}\right] \tag{2.7}$$

with

$$\zeta_{m,n} = \begin{cases} 1, & \text{match} \\ \gamma, & \text{mismatch} \end{cases} \tag{2.8}$$

The expressions (2.7)—(2.8) have the following meaning. Starting from the left ends of sequences, in each step we chose a such alignment, which contributes to the cost function F the maximum. The terms in (2.7) correspond to three possible matching situations: a gap in first sequence, a gap in second, and a matched pair (i,j) of nucleotides.

2.2 Matching vs pairing of two random linear heteropolymers

Consider the statistical model describing the formation of a complex of two heteropolymer linear chains with arbitrary primary sequences. Let these chains be of the lengths $L_1 = m\ell$ and $L_2 = n\ell$ respectively. In what follows we shall measure the lengths of the chains in number of monomers, m and n, supposing that the size of an elementary unit, ℓ, is equal to 1. Every monomer can be chosen from a set of c different types A, B, C, D, ... Monomers of the first chain could form saturating reversible bonds with monomers of the second chain. The term "saturating" means that any monomer can form a bond with at most one monomer of the other chain. The bonds between similar types (like A–A, B–B, C–C, etc.) have the attraction energy u and are called below "matches", while the bonds between different types (like A–B, A–D, B–D, etc.) have the attraction energy v and are called "mismatches". This general description covers both cases (DNA and RNA) by a straightforward redefinition of letters. In real DNA matches are the base pairs, forming according to the complementarity rules (see paragraph 1.1.1). In general case the energies of the base pairs can be different (Tab. 1.1). Suppose also that some parts of the chains can form loops. These loops obviously produce "gaps" since the monomers inside the loops of one chain have no matching (or mismatching) counterparts in the other chain. Schematically a particular configuration of the system under consideration for $c = 2$ is shown in Fig. 2.1.

Our aim is to compute the free energy of the described model at sufficiently low temperatures under the assumption that the entropic contribution of the loop formation is negligible compared to the energetic part of the direct interactions between chain monomers. Let $G_{m,n}$ be the partition function of such a complex; $G_{m,n}$ is the sum over all possible arrangements of bonds. In the low–temperature limit we can write $G_{m,n}$ recursively:

$$\begin{cases} G_{m,n} = 1 + \sum_{i,j=1}^{m,n} \beta_{i,j}\, G_{i-1,j-1} \\ \\ G_{m,0} = 1;\ G_{0,n} = 1;\ G_{0,0} = 1 \end{cases} \tag{2.9}$$

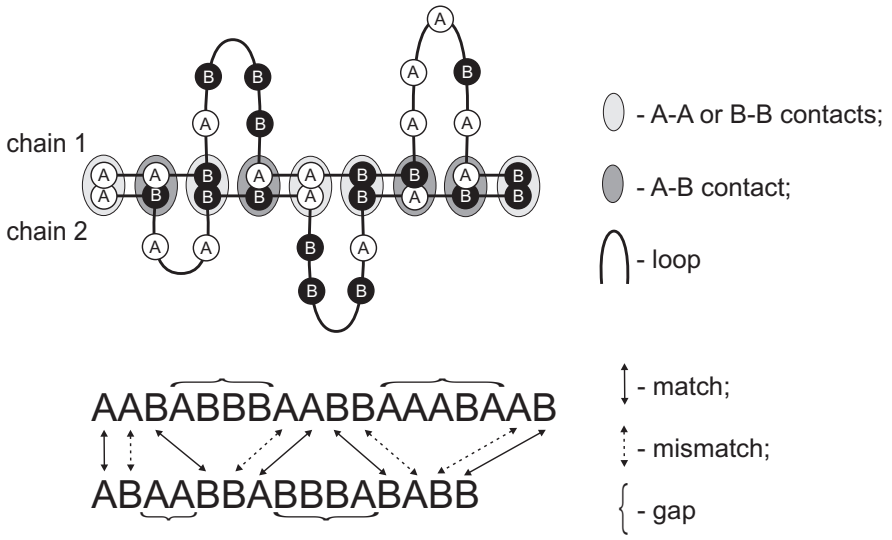


Figure 2.1. Linear pairing complex. Schematic picture of binding of two heteropolymer chains with two types of letters ($\mathbf{c = 2}$).

The meaning of the equation (2.9) is straightforward. Starting from, say, the left ends of the chains shown in Fig. 2.1 we find the first actually existing contact between the monomers i (of the first chain) and j (of the second chain) and sum over all possible arrangements of this first contact. The first term "1" in (2.9) means that we have not found any contact at all. The entries $\beta_{i,j}$ ($1 \leq i \leq m,\ 1 \leq j \leq n$) are the statistical weights of the bonds which are encoded in a contact map $\{\beta\}$(for simplisity we believe $T = k_B T$):

$$\beta_{m,n} = \begin{cases} \beta^+ \equiv e^{-u/T}, & \text{match} \\ \beta^- \equiv e^{-v/T}, & \text{mismatch} \end{cases} \tag{2.10}$$

The straightforward computation shows that the partition function $G_{m,n}$ (2.9) obeys the following exact local recursion

$$G_{m,n} = G_{m-1,n} + G_{m,n-1} + (\beta_{m,n} - 1)\, G_{m-1,n-1} \tag{2.11}$$

Note that if $\beta_{i,j} = 2$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$, the recursion relation (2.11) generates the so-called Delannoy numbers [94].

Let us point out that since we are working at finite temperatures, the account for "loop factors" is desirable. Under the "loop factor" we understand the entropic contribution to the free energy of the entire system coming from the fluctuations

of parts of heteropolymer chains between successive contacts. Obviously, in the zero–temperature limit these fluctuations vanish.

Write the partition function $G_{m,n}$ as $G_{m,n} = \exp\{F_{m,n}/T\}$, where $-F_{n,m}$ and T are the free energy and the temperature of the complex of two heterogeneous chains of the lengths m and n. Considering the $T \to 0$ limit of the equation (2.11), we get

$$F_{m,n} = \lim_{T\to 0} T \ln \left(e^{F_{m-1,n}/T} + e^{F_{m,n-1}/T} + (\beta_{m,n} - 1)\, e^{F_{m-1,n-1}/T} \right) \tag{2.12}$$

which can be regarded as an equation for the ground state energy of a chain. The expression (2.12) reads

$$F_{m,n} = \max \left[F_{m-1,n},\, F_{m,n-1},\, F_{m-1,n-1} + \eta_{m,n} \right] \tag{2.13}$$

where

$$\eta_{m,n} = T \ln(\beta_{m,n} - 1) = \begin{cases} \eta^+ = T \ln(e^{u/T} - 1) & \text{match} \\ \eta^- = T \ln(e^{v/T} - 1) & \text{mismatch} \end{cases} \tag{2.14}$$

Indeed, the ground state energy (2.13) may correspond either: (i) to the last two monomers connected, then the ground state energy equals $\tilde{F}^{\max}_{m-1,n-1} + \zeta_{M,N}$, or (ii) to the unconnected end monomer of the fist (or second) chain, then the ground state energy is $\tilde{F}^{\max}_{m,n-1}$ (or $\tilde{F}^{\max}_{m-1,n}$).

Taking η^+ as the unit of the energy, we rewrite (2.13) in a form, which is identical to the dynamic programming equation (2.7):

$$\tilde{F}_{m,n} = \max \left[\tilde{F}_{m-1,n},\, \tilde{F}_{m,n-1},\, \tilde{F}_{m-1,n-1} + \tilde{\eta}_{m,n} \right] \tag{2.15}$$

with

$$\tilde{\eta}_{m,n} = \begin{cases} 1, & \text{match} \\ a = \dfrac{\eta^-}{\eta^+}, & \text{mismatch} \end{cases} \tag{2.16}$$

(compare to (2.8)). In the low–temperature limit the parameter a has simple expression in terms of the coupling constants u and v:

$$a = \frac{\eta^-}{\eta^+} = \left. \frac{\ln(e^{v/T} - 1)}{\ln(e^{u/T} - 1)} \right|_{T\to 0} = \frac{v}{u} \tag{2.17}$$

The initial conditions for $\tilde{F}_{m,n}$ are transformed into $\tilde{F}_{0,n} = \tilde{F}_{n,0} = \tilde{F}_{0,0} = 0$.

Extrapolating the free energy of linear sequences to zero temperature we recover (for linear sequences only) the well–known standard dynamic programming algorithm described in (2.7)–(2.8).

2.3 Matching vs pairing of two random RNA-type heteropolymers

Having the applications to RNA molecules in mind, we assume that the structures formed by thermo-reversible bonds of each chain are always of a clover-leaf or a cactus–like type, as shown in Fig. 2.2(a). It means that we restrict ourselves to the situation in which the chain conformations with "pseudoknots" shown in Fig. 2.2(b) are prohibited. The difference between allowed and not allowed structures becomes more transparent, being redrawn in the following way. Represent a polymer under consideration as a straight line with active monomers situated along it in the natural order, and depict the bonds by dashed arcs connecting the corresponding monomers. Now, the absence of pseudoknots means the absence of intersection of the arcs – see Fig. 2.2(c,d).

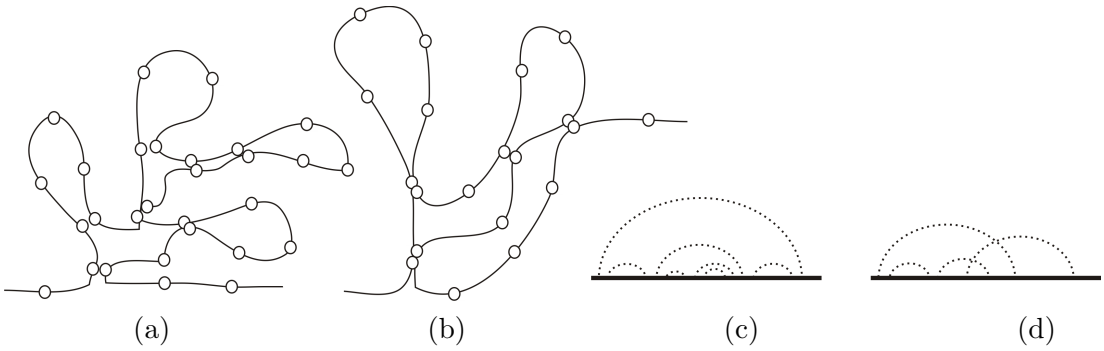


Figure 2.2. Arcs representation of RNA structure. (a,b) Schematic picture of allowed (a) cactus–like and prohibited (b) pseudoknot configurations of the bonds; (c,d): Arc diagrams corresponding respectively to configurations (a) and (b) (note the intersection of arcs in (d)).

We assume for simplicity, that besides the pseudoknots, all other bond configurations are allowed. This means, in particular, that at the moment we do not require any minimal loop length, as well as we do not yet take into account the cooperativity effect. The cooperativity means that if two links are connected

with each other, then the two adjacent links have larger probability to be also connected. These assumptions are known to be false for real RNA molecules (for example, there are no loops shorter than 3 monomers in RNA chains [95]). However, one can speculate that (see, for example, [96]) if the links of the chain are considered as renormalized quasi–monomers consisting of several "bare" units, these assumptions seem to be plausible. Nevertheless, in the last section of this chapter we consider the effect of the minimal loop length on an example of RNA sample.

Let us remind that one of the goals in this work consists in developing an algorithm for the computation of the cost function, which characterizes the similarity of two RNA–type random sequences. To succeed, we should incorporate in the conventional cost function discussed above the contribution coming from the entropy of different rearrangements of cactus–like conformations typical for RNAs. It is not obvious how to do that directly in the frameworks of the dynamic programming approach formalized in the recursion relation (2.7).

Following [97] we write the partition function $G_{m,n}$ of a complex of two heteropolymers capable of forming a cactus–like structure (compare with (2.9)):

$$\begin{cases} G_{m,n} = g^{(1)}_{1,m}\, g^{(2)}_{1,n} + \sum\limits_{i,j=1}^{m,n} \beta_{i,j}\, G_{i-1,j-1}\, g^{(1)}_{i+1,m}\, g^{(2)}_{j+1,n} \\ G_{m,0} = g^{(1)}_{1,m}; \;\; G_{0,n} = g^{(2)}_{1,n}; \;\; G_{0,0} = 1 \end{cases} \tag{2.18}$$

where $g^{(1)}_{i,j}$ and $g^{(2)}_{i,j}$ are the partition functions of individual chains. They satisfy the self–consistent equation [98, 99]:

$$\begin{cases} g^{(a)}_{k,n} = 1 + \sum\limits_{i=k}^{n-1} \sum\limits_{j=i+1+\ell}^{n} \beta'_{i,j} g^{(a)}_{i+1,j-1}\, g^{(a)}_{j+1,n}; \\ g^{(a)}_{0} = 1, \quad a = 1, 2. \end{cases} \tag{2.19}$$

This equation generates the topology of cactus–like RNA structures and it has frequently appeared in the RNA context (see, for example, [42, 95, 53, 97]). Here $g^{(a)}_{i,j}$ is the statistical weight of the loop from the nucleotide i till the nucleotide j in the 1-st ($a = 1$) or 2-nd ($a = 2$) sequence. The Boltzmann weights $\beta'_{i,j}$ are the constants of self–association, which are, similarly to $\beta_{m,n}$, encoded by the contact map. The summation over j runs from $i+1+\ell$ till n ensuring the absence of loops

of lengths smaller than ℓ monomers, in what follows we mostly consider $\ell = 3$. Note also, that since in this paper we are interested in the low–temperature behavior of the partition function, we neglect here the aforementioned "loop weights", i.e. entropic factors due to the formation of intra–chain loops loops.

Figure 2.3. Diagrammatic form of the Dyson–type equation. This equation is specific for the partition function of an individual chain g_n with cactus–like topology (2.19).

Equations (2.18)–(2.19) constitute the analytical basis of our numerical studies, for the problem of RNA–like matching (i.e. matching of sequences with RNA–type architecture). These equations replace the dynamic programming algorithm (2.15)–(2.16) valid for linear sequences.

Now, the ground state free energy $F_{m,n}$ (i.e. the binding free energy at zero temperature) for RNA–like structures can be explicitly computed by extending the approach developed in Section 2.2. Indeed, taking the zero–temperature limit in (2.18) (compare to (2.12)–(2.13)) we get:

$$F_{m,n} = \max_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \left[f^{(1)}_{1,m} + f^{(2)}_{1,n}, Q^{m,n}_{i,j} \right] \tag{2.20}$$

where $f^{(a)}_{i,j} = \lim_{T\to 0} T \ln g^{(a)}_{i,j}$ $(a = 1, 2)$ are the free energies of individual subsequences from the nucleotide i till the nucleotide j, and $Q^{m,n}_{i,j}$ is the zero–temperature limit of the (i, j)-th term in (2.18):

$$Q^{m,n}_{i,j} = F_{i-1,j-1} + f^{(1)}_{i+1,m} + f^{(2)}_{j+1,n} + \tilde{\eta}_{i,j} \tag{2.21}$$

Clearly, $Q^{m,n}_{i,j}$ has a meaning of the ground state free energy of a complex which is forced to have a bond in position (i, j). In turn, the ground state energy of a single chain satisfies the following equation:

$$f^{(a)}_{i,j} = \max_{\substack{r=1,\dots,i \\ s=i+1+\ell,\dots,j}} \left[f^{(a)}_{r+1,s-1} + f^{(a)}_{s+1,j} + \tilde{\eta}'^{(a)}_{r,s} \right] \tag{2.22}$$

Here the values $\tilde{\eta}_{i,j}$ are the inter–sequence matching constants (the same in (2.16)),

while $\tilde{\eta}^{\prime(a)}_{i,j}$ are the intra–sequence matching constants.

The boundary conditions for the ground state free energy follow from the boundary conditions of the partition function (2.18):

$$\begin{cases} F_{0,0} = 0; \\ F_{i,0} = f^{(1)}_{1,i}; \quad 1 \le i \le m \\ F_{0,j} = f^{(2)}_{1,j}; \quad 1 \le j \le n \end{cases} \tag{2.23}$$

Thus, to compute the ground state free energy of a complex of two RNA–like sequences, we should first reconstruct the matrices $f^{(1)}$ and $f^{(2)}$ for individual chains by applying (2.22) and then find the matrix F using (2.21). The boundary conditions (2.23) together with (2.22) allow us to compute the elements of the matrices $Q^{1,j}$ for $m = 1$ and any $1 \le j \le n$. Knowing the corresponding matrix $Q^{1,j}$ we define the elements $F_{1,j}$ $(1 \le j \le n)$ of the free energy matrix by using (2.21). Then we proceed recursively and determine the matrices $Q^{2,j}$, compute $F_{2,j}$ $(1 \le j \le n)$, etc. Clearly, this algorithm can be completed in time of order $O(m^2 \times n^2)$.

Note, that expression (2.22) can be used for the determination of a structure of a single RNA chain. From statistical point of view (in contrast to biological meaning), it does not matter what is under consideration — a single chain or a RNA-RNA complex.

2.4 Structure recovery

In this Section we describe the implementation of the structure recovery algorithm for linear and cactus–like structures by the corresponding matrices of free energies F at zero temperature. Let us point out that due to degeneration, the restored sequence is one among the ensemble of sequences with the same free energy.

2.4.1 Finding the Longest Common Subsequence for linear chains

As we have already discussed, sequence matching problem for linear structures consists in finding the longest common subsequence (possible with gaps) of two

given sequences of nucleotides. Let us demonstrate on simple example how the algorithm works. Consider two sequences of $m = n = 6$ nucleotides:

C G U U C C –S1
G C G G A A –S2

Construct the incidence matrix η with $\eta_{i,j} = 1$ if monomers i of the 1st sequence and j of the second one match each other, and $\eta_{i,j} = 0$ otherwise – see Fig. 2.4(a). In Fig. 2.4(b) we have shown the matrix of ground state free energies, F, computed via the recursion algorithm (2.15)–(2.16).

$\eta =$

	G	C	G	G	A	A
C	1	0	1	1	0	0
G	0	0	1	0	0	0
U	0	0	0	0	1	1
U	0	0	0	0	1	1
C	1	0	1	1	0	0
C	1	0	1	1	0	0

(a)

$F =$

	G	C	G	G	A	A
C	1	1	1	1	1	1
G	1	2	2	2	2	2
U	1	2	2	2	3	3
U	1	2	2	2	3	4
C	1	2	3	3	3	4
C	1	2	3	4	4	4

(b)

	G	C	G	G	A	A
C	1	1	1	1	1	1
G	1	2	2	2	2	2
U	1	2	2	2	3	3
U	1	2	2	2	3	4
C	1	2	3	3	3	4
C	1	2	3	4	4	4

G C G G A A
| |
C G U U C C

(c)

	G	C	G	G	A	A
C	1	1	1	1	1	1
G	1	2	2	2	2	2
U	1	2	2	2	3	3
U	1	2	2	2	3	4
C	1	2	3	3	3	4
C	1	2	3	4	4	4

G C G G A A
| |
C G U U C C

(d)

Figure 2.4. Linear pairing algorithm.(a) Incidence matrix η, (b) ground state free energy matrix F; (c)-(d) structure recovery algorithm for linear chains.

In order to see which nucleotides form links, let us proceed as follows. Take the element $F_{i,j}$ of the matrix F and compare its value to the values of three neighboring matrix elements $F_{i-1,j-1}, F_{i-1,j}, F_{i,j-1}$. Now we take the following

decisions:

- If $F_{i-1,j-1} = \max\left[F_{i-1,j-1}, F_{i-1,j}, F_{i,j-1}\right]$ then i of the 1-st sequence is linked to j of the 2-nd one;
- If $F_{i-1,j} = \max\left[F_{i-1,j-1}, F_{i-1,j}, F_{i,j-1}\right]$ then we skip the element i in the 1-st sequence;
- If $F_{i,j-1} = \max\left[F_{i-1,j-1}, F_{i-1,j}, F_{i,j-1}\right]$ then we skip the element j in the 2-nd sequence.

This procedure begins with the element $F_{m,n}$.

This prescription for computing the matrix of ground state free energies shown in Fig. 2.4(b) gives (due to degeneration) many sequences with the same value of the free energy. Two possible realizations are depicted in Fig. 2.4(c,d).

2.4.2 Finding the secondary structure for interacting RNA–like chains

The structure recovery for the chains with cactus–like structures is much more involved problem, however it can also be described recursively. In this case the algorithm consists of the following successive steps:

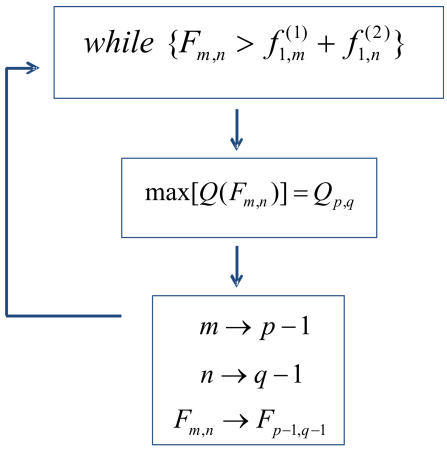


Figure 2.5. The structure recovery algorithm scheme for RNA-type complex.

Beginning with the element $F_{m,n}$, we look at the respective matrix $Q(F_{m,n})$ (2.23). The maximal element of this matrix $Q_{\max} = Q_{p,q}$ either

- $Q_{\max} = F$, and it means the presence of the contact (p, q) in the folding. We look for the next pair (s, r), substituting $p - 1$ and $q - 1$ in our procedure; or

- $Q_{\max} < F$, then (according to (2.23)) there are no any more pairs of linked nucleotides in the considered branching structure.

Knowing pairs of linked nucleotides, for example, (p, q) and (s, r), we reconstruct the structure of the loops between the paired nucleotides by the corresponding loop energies $f^{(1)}_{p,s}$ and $f^{(1)}_{q,r}$.

Below we demonstrate on simple example how this algorithm works. Take two sequences S1 and S2:

$$\begin{array}{l} \mathsf{A\ U\ C\ U\ C\ A\ C} \quad -\mathsf{S1} \\ \mathsf{G\ C\ C\ A\ G\ G\ G} \quad -\mathsf{S2} \end{array}$$

The respective incidence matrices η' (for intra–matching S1–S1), η'' (for intra–matching S2–S2), and η (for inter–matching S1–S2) are shown in Fig. 2.6(a,b,c) respectively.

The matrices of effective statistical weights $f^{(1)}$ and $f^{(2)}$ of first and second sequences, as well as the ground–state free energy matrix F, are shown in the Fig. 2.6 (d),(e) and (f) respectively. The elements $f_{m+1,j}$ and $f_{n+1,j}$, which formally present in the computations, are set to zero: $f_{m+1,j} = f_{n+1,j} = 0$ for all j.

According to our scheme, we consider the matrices Q for established contacts. For example, the maximal element $Q_{7,7}$ for $F_{7,7}$ shows that 7-th monomer of S1 and 7-th nucleotide of S2 form the base pair in optimal configuration. In the next step, the matrix Q for the element $F_{6,6}$ is considered. If the Q has several maximal elements, it indicates about the degeneracy of the ground state structure and the procedure must be performed for all possible states. The loop structures can be reconstructed by corresponding statistical weights – see Fig. 2.6(j). The two possible configurations for given RNA segments S1 and S2 are presented in Fig. 2.6(j).

The proposed algorithm is applied to a longer trial sequences shown in Fig. 2.7. Namely, we have performed the structure recovery for three different cases: for linear chains (a) (for them we use the algorithm described in 2.2), for cactus-like

chains (b) and for cactus-like chains with the restriction on the size of the minimal loop length (c) (there are no loops less than 4 nucleotides). These structures are depicted in the figures Fig. 2.7(a), (b) and (c) respectively.

It should be noted that the polymer complex structure strongly depends on details of the model. For instance, the difference between the structures depicted in Fig. 2.7(b) and (c) is entirely due to the value of the minimal loop length ℓ. Since the global topology of optimal structure is highly sensitive to microscopic details, results consistent with experimental data can be obtained only when reliable information about loop factor, bond energies, and cooperativity parameter is available. As noted above, these parameters can be taken into account when necessary within the framework of the proposed model.

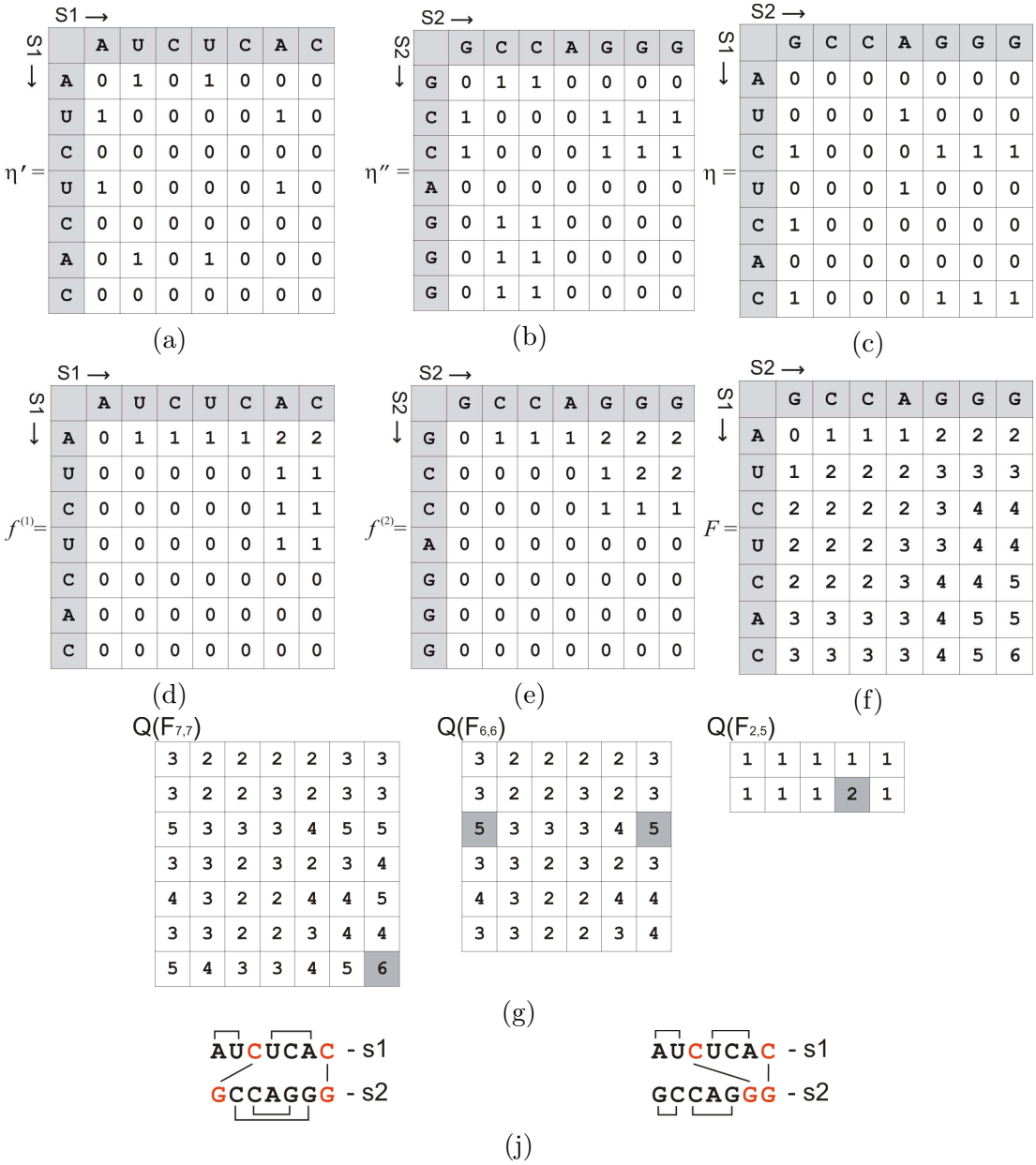


Figure 2.6. The algorithm description for RNA-type complex. Incidence matrices for pairs of chains with possible clover–leaf structures inside each sequence: (a) intra–matching S1–S1, (b) intra–matching S2–S2, (c) inter–matching S1–S2; the free energy matrices: (d) for the 1-st sequence, (e) for the 2-nd and for the complex f; the matrices Q (g), involved in the structure recovery procedure; (j) the predicted RNA-RNA complex structures.

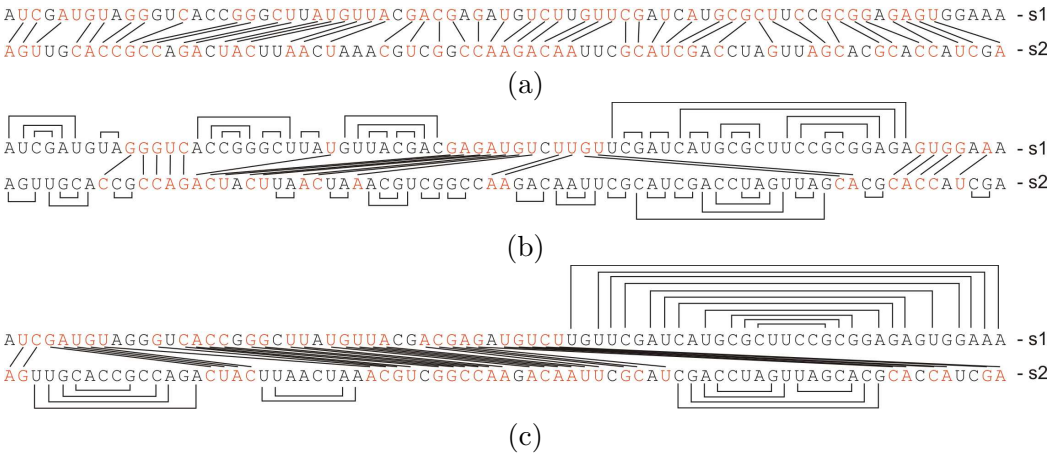


Figure 2.7. Structures recovered from the pair of sequences [100].
(a) Linear structure, (b) branching structure, (c) branching structure with the restriction on the size of the minimal loop (there are no loops less than 4 nucleotides).

Chapter 3

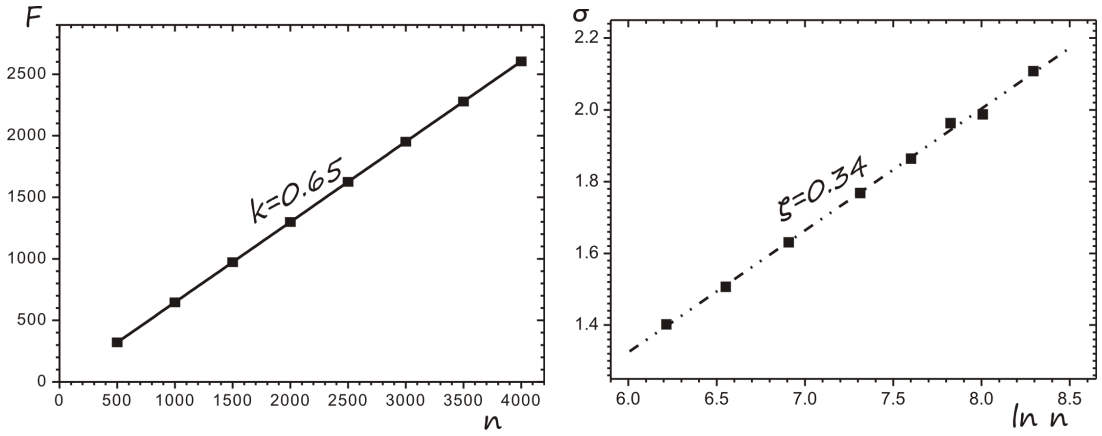
Statistical properties of random RNAs

This chapter is focused on discussion of statistical properties of complexes of two linear and/or RNA-type molecules. First, we consider the mean energy and energy fluctuations as functions of the sequence lengths in random RNA-RNA complexes. Next, we describe the model in which we estimate the binding probability for random sequence polymers in RNA-like complexes. Finally, we report the results concerning the analysis of the loop length distribution in complexes and propose models describing these distributions.

3.1 Mean energy and energy fluctuations for paired RNAs

The problem of finding the ground state energy of a polymer complex of two linear chains has been addressed in a number of studies in the framework of the Bernoulli matching model (e.g., see [101, 102]). In these works, the matrix elements in $\eta_{m,n}$ (2.14) are independent random variables taking on values 1 and 0 with probability $p = 1/c$ and $q = 1 - p$, respectively. It was shown in [102] that the ground state energy distribution for $n, m \gg 1$ is

$$\langle F \rangle = \frac{2\sqrt{pmn} - p(m+n)}{q} + \frac{(pmn)^{1/6}}{q}\left[(1+p) - \sqrt{\frac{p}{mn}}(m+n)\right]^{2/3}\chi \quad (3.1)$$



(a) (b)

Figure 3.1. Statistics for linear complex of two random RNAs. Free energy (a) and energy fluctuation (b) in dependence on the sequence length.

where χ is a random variable having the Tracy-Wisdom distribution with ($\langle\chi\rangle = -1.7711...$ and $\langle\chi^2\rangle - \langle\chi\rangle^2 = 0.8132...$)(e.g., see review [103] for a more detailed description of this distribution). When $m = n$, the energy of the complex can be represented as

$$\langle F_{m,n}\rangle \approx \frac{2}{1+\sqrt{c}}n + f(c)\,\langle\chi\rangle\, n^{1/3} \tag{3.2}$$

where $f(c) = c^{1/6}(\sqrt{c}-1)^{1/3}/(\sqrt{c}+1)$. According to [102], free energy fluctuations behave as

$$\sigma \equiv \sqrt{\langle F^2\rangle - \langle F\rangle^2} \approx \sqrt{\langle\chi^2\rangle - \langle\chi\rangle^2}f(c)n^{1/3} \tag{3.3}$$

The exponent 1/3 is typical for the stochastic growth in strongly correlated systems which belong to the so-called Kardar-Parisi-Zhang (KPZ) universality class [58]. Fig. 3.1 shows numerical results obtained for paired linear polymers. Here, the slope $k_1 \approx 0.65$ (Fig. 3.1(a)) is in good agreement with $k_1 = \lim\limits_{n\to\infty} \langle F\rangle / n \to 2/3$, predicted by (3.2). Furthermore, the slope $\zeta = 0.34$ for the fluctuation energy dependence is close to 1/3. Thus, the expression (3.2), obtained in Bernoulli approximation provide a satisfactory description of the behavior of the ground state energy of linear complex of two random RNAs.

A similar analysis has been performed for paired sequences forming a cloverleaf (RNA-like) secondary structure with the minimal loop length $\ell = 0$. The mean energy and energy fluctuations are plotted in Fig. 3.2. As in the case of a linear

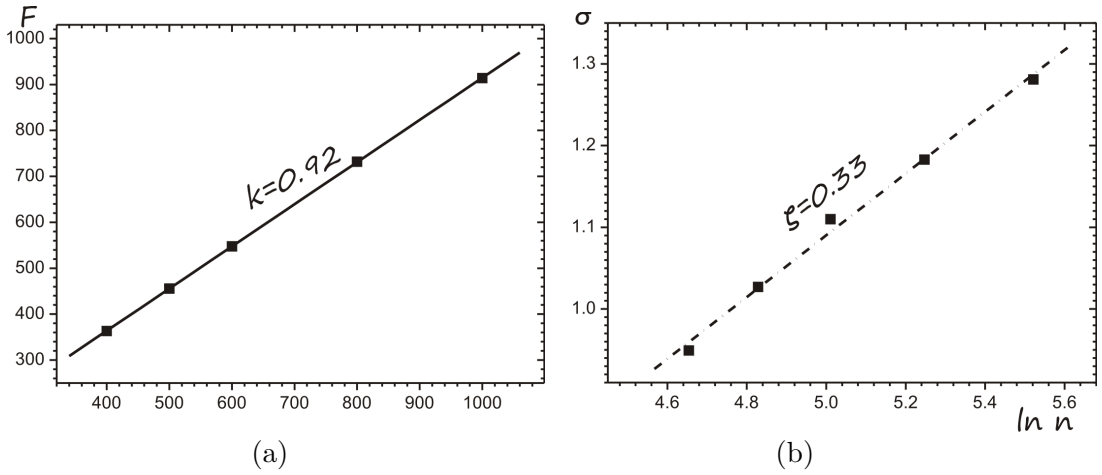


Figure 3.2. Statistics for RNA-type complex of two random RNAs. Free energy (a) and energy fluctuation (b) in dependence on the sequence length.

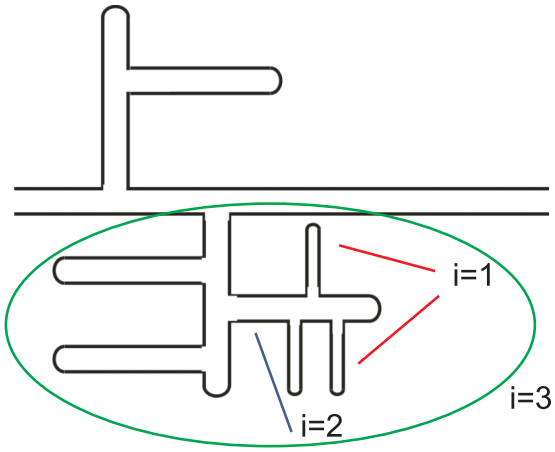


Figure 3.3. Hierarchical model of RNA-like complex. Loops of firs level ($i = 1$), second ($i = 2$) and third ($i = 3$) are depicted.

polymer complex, $\langle F(n) \rangle = k_c n$ for $n \gg 1$ (Fig. 3.2), but the slope is much steeper ($k_c \approx 0.92$) because of intraloop interactions between nucleotides. The behavior of the ground state energy fluctuations is the same, see Fig. 3.2(b).

We can estimate analytically the slope k_c as a function of sequence length borrowing the ideas from the renormalization group analysis. Namely, consider the RNA-RNA complex as a hierarchal structure, having loops of different "levels" (Fig. 3.3). Each i-th level loop can be treated as a complex consisting of the strands of the $(i-1)$-th level that form the loop of the previous level.

From the expression (3.1) it follows, that a maximum contribution to the total free energy is reached at $m = n$. Therefore, an upper bound for the free energy of a loop can be evaluated as the binding energy between the two halves

of such a hairpin. When an RNARNA complex can be represented as a nested structure, a renormalization group treatment can be applied [104]. This idea is formalized by assuming that each ith level of RNA-RNA complex consists of strands with energy of interacting monomers are renormalized to the $(i+1)$th level loop energy. Since the loop energy (3.2) is roughly proportional to the loop length, it can be represented as $F_L^{(i)} \approx k_r^{(i)} L$. Substituting the loop statistical weights $g_{i,i+L} = e^{k_r L/T}$ in (2.15), we can estimate the free energy of a random RNA-RNA complex as

$$F_{m,n}^{(i+1)} = \max\left[F_{m-1,n} + k_r^{(i)},\, F_{m,n-1} + k_r^{(i)},\, (F_{m-1,n-1} + u)\mathcal{P}(m,n)\right] \tag{3.4}$$

This expression should be interpreted as follows. First, find the free energy of a complex $F_{m,n}^{(2)}$, which includes only first level hairpins. Next, determine the binding energy coefficient in the second-level loops as

$$k_r^{(2)} = \frac{F_{m,n}^{(2)}}{m+n} \tag{3.5}$$

Substituting the result back to (3.4), calculate the energies for the third-level loops $k_r^{(3)}$, and so on. The factor $\mathcal{P}(m,n)$ takes account of the ith-level minimum loop length:

$$\mathcal{P}(m,n) = \begin{cases} 1, & m\ n \text{ match} \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

We suppose that, monomers m and n match when: i) the segment $[m-\ell, m-1]$ in the sequence S_1 has no bonds to the segment $[n-\ell, n-1]$ of S_2, where ℓ is the minimum nessasary loop length at a certain level (if $m < \ell$ and/or $n < \ell$, than $[1,m]$ and/or $[1,n]$ are taken in the respective sequences);
ii) $(m-1)$-th monomer of S_1 and $(n-1)$-th monomer of S_2 form the base pair and in the substitute $(m-1) \to m$, $(n-1) \to n$ i) (or ii)) holds.

Tab. 3.1 presents the values for the binding coefficients and the minimum loop lengths of the i-th level calculated for $m = n = 10^4$. The average binding coefficient weakly depends on the sequence lengths. However, longer sequences can be used to perform the estimation for a larger number of hierarchical levels. Note that the binding coefficient calculated in this manner approaches to 1 loga-

rithmically with increasing number of levels ($n \to \infty$) because the minimum loop length exponentially increases with i (see Tab. 3.1).

Thus, the binding coefficient determined numerically (Fig. 3.2(a)) varies with the loop length, and the value of $k_c \approx 0.92$ obtained here just indicates that sequences of the length of 1000 monomers have only two or three hierarchical levels.

Level, i	2	3	4	5	6	7
The minimum loop length	2	6	24	78	240	726
The binding coefficient	0.851	0.912	0.931	0.937	0.94	0.941

Table 3.1. The probability of pairing in the different level loops.

3.2 The loop length distribution in random RNA-RNA complex

We have analyzed the loop length distributions in paired linear polymers and RNA-like complexes. Fig. 3.4 shows the number of the loops $W(n)$ in a paired linear polymer complex of length n. It is clear that these results follow an exponential distribution, which is typical for systems with independent bindings (i.e. the probability that a monomer is bound to the opposite strand is independent of the binding status of its nearest neighbors). Indeed, when $n \gg 1$, the value of $k_1 = \langle F \rangle / n$ can be interpreted as the binding probability in a complex. If the pairing status of a monomer is independent of those of its nearest neighbors, then the number of the loops of the length s in a paired linear polymer complex of length n can be estimated as

$$W(s) = nk_1^2(1-k_1)^s \tag{3.7}$$

When $n \gg 1$, this loop length distribution obviously satisfies the relation $\sum_{s=1}^{n} sW(s) = (1-k_1)n$. The semi-log plot of numerical results in Fig. 3.4 is well fitted by the line $y(s) = a - bs$, where $a \approx \ln(nk_1^2)$ and $b \approx \ln(1-k_1)$. Thus, the loop statistics in linear complexes is well described by the model with independent binding of monomers.

However, it is worth noting that the model of independent binding gives good results for the alphabetic sequences with quite large alphabets $c \geq 4$. For two-letter and three-letter alphabetic sequences the binding is correlated and the expression (3.1) can not be used for description of linear complexes.

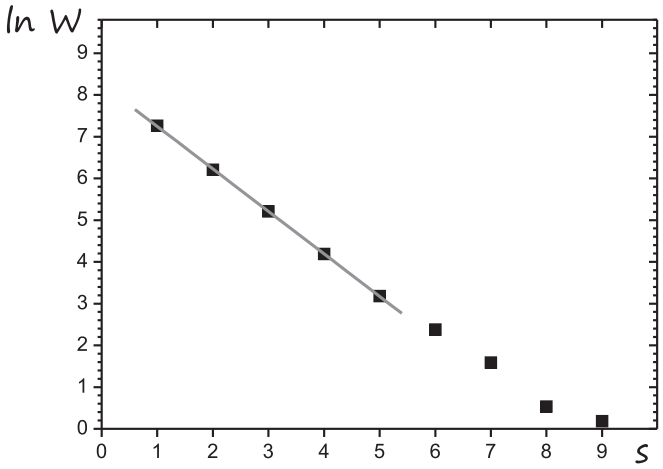


Figure 3.4. Loop length distribution in linear complex of two RNAs. The calculations were performed for random polymers of the length $n = 10^4$, the graph is averaged by the ensemble of 10^4 complexes.

Essentially different statistical behavior is observed for complexes with RNA-like cloverleaf structure. Fig. 3.5(a) shows the number of loops of length s in the ensemble of 10^3 random samples. Note the following properties of the distribution. First, it obeys a power law with the exponent varying within the interval $[1.38, 1.5]$ in RNAs of various lengths. Second, the distributions obtained for RNAs of the different lengths are similar, which makes it possible to restrict calculations to short sequences. Third, when s is small ($s \leq 5$), the number of loops with odd lengths is small, while the number of those with even lengths is high. The reason is that RNA-RNA complexes with $\ell = 0$ are characterized by a high energy per nucleotide pair ($k_c \approx 0.92$) due to intra-loop interactions, and the formation of a small loop of an odd number of nucleotides implies that at least one bond is lost within the loop. Thus, the formation of the odd-length loops energetically is less efficient. Finally, the distribution reaches a plateau at large values of s because of the finite size effect (e.g., see the theory developed in [105] for an analogous phenomenon in a slightly different system). Let us now discuss the numerical

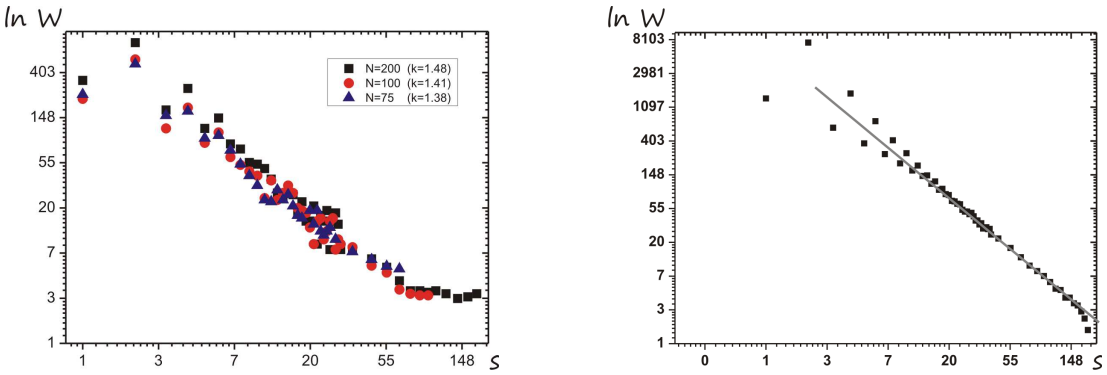


Figure 3.5. Loop length distribution in RNA-type complex. (a) The loop length distribution in a cactus-like complex of two RNAs (The simulations were performed for the ensembles of 10^3 random samples with the sequence lengths $n = m = 75, 100 \ \ 200$; for $s \geq 30$ the distribution function was smoothed over 10 adjacent values); (b) The Motzkin path length distribution (the random walk length is 200, the number od samples is 10^3, for $s \geq 30$ the function was smoothed over 10 adjacent values).

distributions above.

We can associate each secondary structure with a random walk on a $(1+1)$-dimensional lattice as follows (Fig. 3.6). Each monomer in a RNA-like loop is represented by a step in a random walk. If the monomer is the head or tail of a loop (paired with another one behind or ahead of it along the loop sequence, respectively), it corresponds to an upward or a downward steps. If the nucleotide is not paired, it is a horizontal step in random walk path. This obviously implies the correspondence between RNA-like secondary structures and Motzkin paths [106]. A return to the abscissa axis corresponds to the formation of a loop in a RNA-RNA complex. It is well known [107] that the number of distinct Motzkin paths of the length n with t horizontal steps is expressed by the Catalan numbers:

$$P(n,t) = \begin{pmatrix} n \\ t \end{pmatrix} C_{(n-t)/2} = \begin{pmatrix} n \\ t \end{pmatrix} \frac{1}{\frac{n-t}{2}+1} \begin{pmatrix} n-t \\ \frac{n-t}{2} \end{pmatrix} \tag{3.8}$$

where $\begin{pmatrix} n \\ t \end{pmatrix}$ are binomial coefficients, $C_{(n-t)/2}$ are the Catalan numbers. When $n \gg 1$ the function (3.8) has the asymptotic behavior $P(n,t) \sim \ n^{-3/2}$. We plotted the length distribution for random Motzkin paths under assumptions that the probability of an upward (as well as a downward) step is equal to $p_m \approx k_c/2 =$

0.46, where $k_c = 0.92$, which corresponds to the binding probability established numerically in a cactus-like complex of two RNAs. The probability of horizontal step is taken respectively $1 - 2p_m$. The result is represented in Fig. 3.5(b). This model distribution, having all features of the RNA-type complex distribution, describes very well numerically obtained dependence (Fig. 3.5(a)).

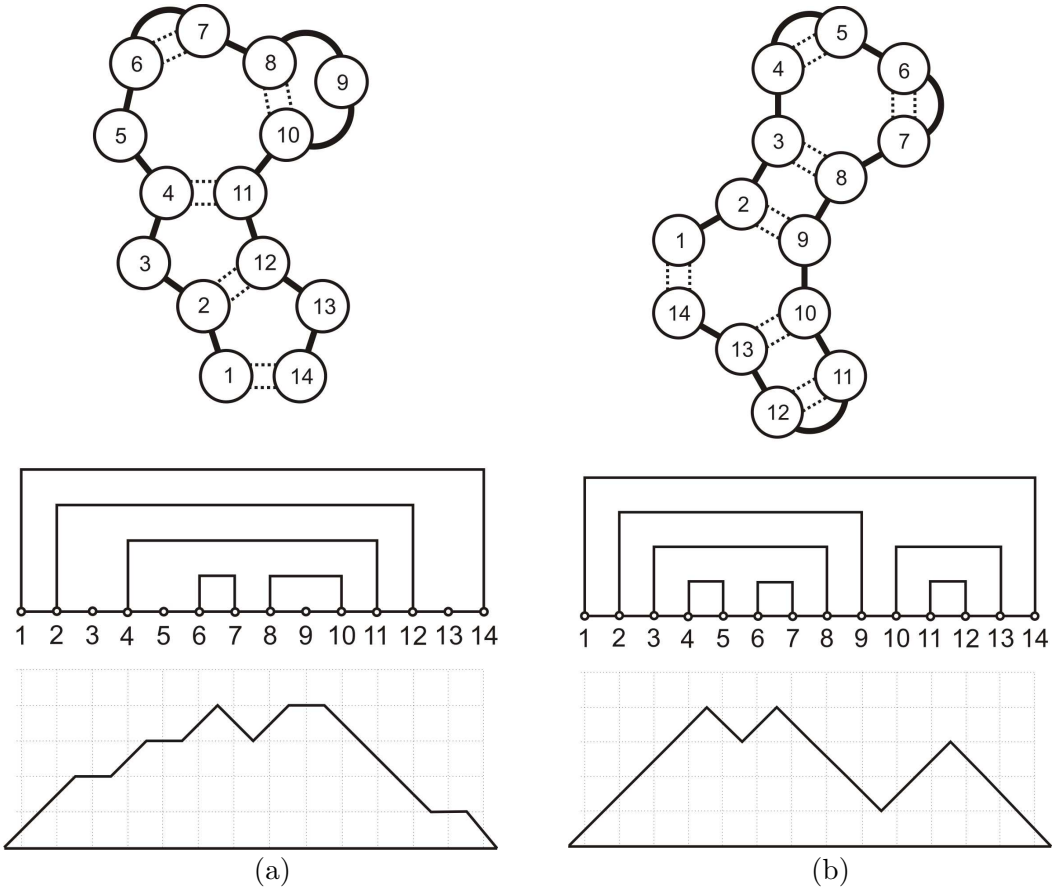


Figure 3.6. Random walk representation of RNA structure. Secondary structure of RNA with gaps and respective Motzkin path (a); perfect matching RNA structure without gaps and respective Dyck path (b).

The representation of RNA structures in terms of Motzkin paths with well known statistical properties leads us to an interesting observation. The RNA-type structure is very sensible to the alphabet (the number of different monomer species) used in construction of a random sequence. As it is shown in the next Section, there is an alphabet-dependent phase transition between two essentially different states, distinguished by the fraction of paired monomers.

Chapter 4

Random RNA-type polymer with the different alphabet

In contrast to the previous chapter we are focused here on the study of a single random RNA-type chain (a loop in random RNA-RNA complex). We consider the fraction f of nucleotides involved in the formation of a cactus-like secondary structure as a function of the number c of different nucleotide species. We show, that with changing c, the secondary structures of random RNAs undergo a morphological transition: $f(c) \to 1$ for $c \leq c_{cr}$ as the chain length goes to infinity, signaling the formation of a virtually perfect gapless secondary structure; while $f(c) < 1$ for $c > c_{cr}$, which means that a non-perfect structure with gaps is formed. The strict upper and lower bounds $2 \leq c_{cr} \leq 4$ are proven, and the numerical evidence for c_{cr} is presented. We propose different model for determination of the transition point c_{cr}. In particular, we formulate the problem as the perfect matching problem in a random Erdos-Renyi graph and give the analytical estimate for the transition point. The relevance of the transition from the evolutionary point of view is discussed.

4.1 Statistics of alphabetic RNA sequences

Consider a random polymer of the length L, forming a secondary structure typical for RNA molecules (Fig. 2.2(a)). We are interested in the dependence of the binding probability $f_\infty(c)$ in such a secondary structure of random RNA on the alphabet size, used in it in thermodynamical limit, i.e. for the infinite chains.

First of all, we provide simple arguments proving an existence of a phase transition and giving rough low and upper bound for the critical alphabet. There is $c = c_{cr}$ such that the limit value of the fraction of paired nucleotides approaches unity if $c \leq c_{cr}$ and remains less than unity if $c > c_{cr}$. This can be demonstrated as follows.

Consider $c_{\rm cr}^{\rm min} = 2$. It turns out that matching with $f(c = 2) \to 1$ as $n \to \infty$ is possible not only on average but for any given primary structure. Indeed, consider a random heteropolymer RNA constituted of A and B monomers, forming saturating bonds of type A–A and B–B and construct the optimal structure as follows. Take the left end of the chain as a starting point, and move along a sequence until meeting the first pair of two sequential letters AA or BB. Connect these two letters with a bond and erase them from the sequence. Iterating this procedure, one arrives finally to an alternating sequence of the type ABAB... (we have assumed that the starting letter is A). Connect now the first letter A from this sequence to the last one, the next B to the B before the last A, etc. It is clear that this algorithm results in a nested secondary structure which leaves unmatched at most two letters (one – in the middle of the ABAB...–sequence and, possibly, another one in the very end). The fraction of mismatched letters decreases as n^{-1} with n, proving the conjecture. A similar algorithm for alternating (A–B) bonding can be easily constructed (though the fraction of mismatches decreases as $n^{-1/2}$ in this case). Note, that this lower bound is already nontrivial: in the celebrated "longest common subsequence" used for the comparison of two *linear* DNA sequences (3.1), the fraction of matches equals $f_{\rm lin}(c = 2) = 2(\sqrt{2} - 1) < 1$, and the "critical" alphabet size, at which $f_{\rm lin} = 1$, is $c_{\rm cr}^{\rm lin} = 1$.

To construct the upper bound for $c_{\rm cr}$, recall the bijection between cactus–like RNA secondary structures and discretized Brownian excursions, known as Motzkin paths [106]. By this bijection, shown in Fig. 3.6(b), the gapless ("perfect") secondary structures correspond to excursions with no horizontal steps, the Dyck paths. The total number $D(n)$ of the Dyck paths of the even length n is given by the Catalan number $C_{n/2}$:

$$D(n) = C_{n/2} \equiv \frac{\Gamma[n+1]}{\Gamma\left[\frac{n}{2}+1\right]\Gamma\left[\frac{n}{2}+2\right]} \sim \frac{2^n}{n^{3/2}} \tag{4.1}$$

where $\Gamma[n]$ is the Γ-function, and the asymptotic expression is valid for $n \gg 1$.

Consider a set of random sequences of the length n. Each of these sequences (there are c^n of them) must correspond to a certain perfect match, i.e. a Dyck path. Meanwhile, if one particular Dyck path corresponds to a perfect match of some particular sequence, it simultaneously corresponds to perfect matches of many others. Indeed, each "up–down" pair of steps in a Dyck path can be realized in c different ways (A–A, B–B, etc...) independently of all others, leading to a degeneracy of order $c^{n/2}$. Thus, the number of different primary sequences which can have perfect secondary structures is at most

$$W(c,n) = c^{n/2}\, D(n) \sim \frac{(2\sqrt{c})^n}{n^{3/2}} \tag{4.2}$$

One primary sequence can be represented by several Dyck paths, thus this is an estimate from above. Comparing the value $W(c,n)$ to the total number of primary sequences, $W_0(c,n) = c^n$, we have for $n \gg 1$:

$$\begin{cases} \lim\limits_{n\to\infty} \dfrac{\ln W(c,n)}{n} > \lim\limits_{n\to\infty} \dfrac{\ln W_0(n)}{n} & \text{for } c < c_{\rm cr}^{\max} = 4 \\ \lim\limits_{n\to\infty} \dfrac{\ln W(c,n)}{n} < \lim\limits_{n\to\infty} \dfrac{\ln W_0(n)}{n} & \text{for } c > c_{\rm cr}^{\max} \end{cases} \tag{4.3}$$

One can follow this reasoning to develop the upper bound also for $f(c)$ at $c > c_{\rm cr}^{\max} = 4$. In this case the fraction of random primary sequences admitting a perfect match among all $W_0(c,n)$ of them is exponentially small. Therefore, the ground states of almost all of sequences should correspond to matchings with gaps, i.e. to Motzkin paths. The Motzkin paths with finite fraction of gaps (horizontal steps) produce much more possibilities for the RNA ground states than Dyck paths of the same length. The number of n–step Motzkin paths with m gaps is $M(m,n) = \frac{n!}{m!(n-m)!} D(n-m)$ and

$$\frac{\ln M(f,n)}{n} = -(1-f)\ln(1-f) - f\ln\frac{f}{4} + o\left(\frac{\ln n}{n}\right) \tag{4.4}$$

where $f = \frac{n-m}{n}$ and (4.4) works for $n \gg 1$ and even $m - n$.

How many different primary structures can have a given Motzkin path as a ground state? Each pair of "up–down" steps is bound to belong to the same species, as for the Dyck paths, while each horizontal step can be chosen indepen-

dently. The total degeneracy Z is thus

$$Z(c,n,f) = c^{(fn)/2} c^{(1-f)n} = c^{n(2-f)/2}. \tag{4.5}$$

As f decreases, the total number of structures which can have ground states with the fraction of matches more than f increases and is given by

$$W(c,n,f) = \sum_{j=0}^{(1-f)n} Z(c,n,j/n) M(j,n), \tag{4.6}$$

At some $\bar{f}$ it becomes equal to the total number of possible primary structures $W_0(c,n) = c^n$, giving the estimate for the typical value of $f(c)$. For $n \gg 1$ the sum in (4.6) can be evaluated up to the leading order using the saddle–point approximation. One can introduce value

$$\Delta w(f,c) = \lim_{n\to\infty} \frac{1}{n} \ln \frac{W(c,n,f)}{W_0(c,n)} \tag{4.7}$$

which is from (4.5)-(4.7)

$$\Delta w(f,c) = \begin{cases} -f \ln \dfrac{\sqrt{c} f}{2} - (1-f)\ln(1-f); & f > f_{\rm m} \\ \ln\left(1 + \dfrac{\sqrt{c}}{2}\right) > 0; & f < f_{\rm m} \end{cases} \tag{4.8}$$

where $f_{\rm m} = \frac{2}{2+\sqrt{c}}$. For $f > f_{\rm m}$ the sum in (4.8) is dominated by contribution from the upper boundary, while for $f < f_{\rm m}$ it is given by the maximum at $f_{\rm m}$ and is, therefore, independent of the upper summation limit. The desired value of $\bar{f}(c)$ is defined by the solution of the equation $\Delta w(f,c) = 0$ and is plotted in Fig. 4.1(b) with a black line [1].

We have analyzed numerically the statistical properties of the ground state free energy $f(c)$, applying (2.20) to the ensemble of random sequences with different numbers of letters (nucleotide types) $c = 3, 4, ..7$. Fig. 4.1(a) illustrates that the specific free energy indeed tends to some average value f_∞, which is a function of the alphabet. In Fig. 4.1(b) we plotted the dependence of this function $f_\infty(c)$, the

[1] It may seem that $\bar{f}$ is an estimate for the "typical smallest", not average value of f. However, since $\bar{f} > f_{\rm m}$, it belongs to the regions where the sum in (4.8) is dominated by the upper bound and thus the average and "typical largest" values of f converge in thermodynamic limit.

numerical values of $f_\infty(c)$ lay lower then the upper bound given by the theory.

More detailed analytic analysis of random alphabetic sequences shows that the critical alphabet is set by stronger inequality: $2+\epsilon_1 < c_c < 3-\epsilon_2$ [108], going beyond the integer values. In the next section we consider a model of a random polymer with the effective non-integer alphabet size.

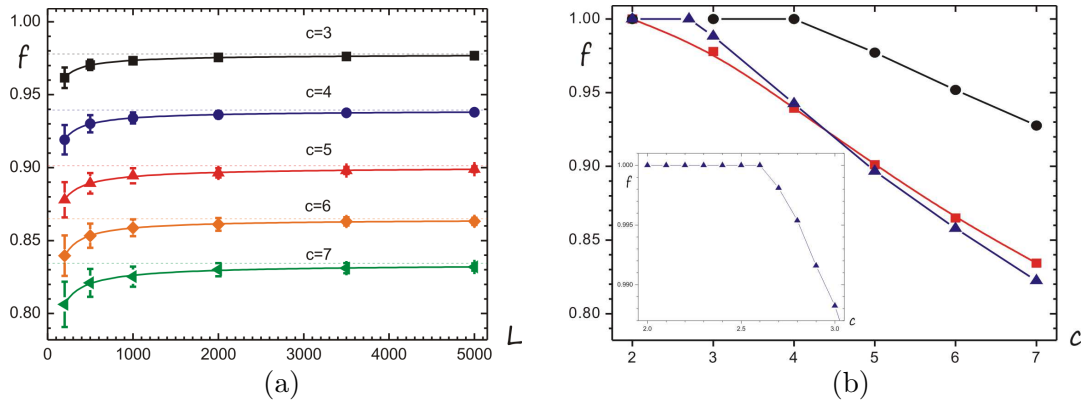


Figure 4.1. Energy dependence for random RNA-type polymers.(a) The dependence of specific free energy on the sequence length at different alphabet c. (b) The dependence of the limiting energy $f_\infty(c)$ on the alphabet for alphabetic sequences (the red curve), for the Bernoulli polymer (the blue curve) and upper analytical estimate (the black curve).

4.2 Bernoulli model of a random RNA polymer

As we have mentioned, the Bernoulli approximation, in which the contact map $\eta_{m,n}$ (2.14) is taken as a random matrix of zeros and ones, gives a good agreement with numerical result for DNA alignment problem. For our purposes, we consider the matrix $V = \eta'$ in (2.20) consisting of independent identically distributed random variables, equal to one with probability p for any $i \neq j$, and equal to zero otherwise. This defines the uniform distribution on the entries of the matrix V:

$$\mathrm{Prob}(V_{ij}) = p\delta(V_{ij}-1) + (1-p)\delta(V_{ij}) \tag{4.9}$$

where $\delta(x) = 1$ for $x = 0$, and $\delta(x) = 0$ otherwise. Obviously that the alphabet in this model is

$$c_{eff} = \frac{1}{p} \tag{4.10}$$

The matrix V can be regarded as an adjacency matrix of a random Erdös-Rényi graph $G(V)$ without self-connections. In these terms, we can describe a phase transition, as follows (typical for onstraint satisfaction problem). The number of constraints per node imposed by the matrix V is varied and cross a certain critical value, the instances pass from satisfiable with high probability to almost surely unsatisfiable in the large L limit [109]. In other words, we show that there is a critical value of the bond formation probability, p_c, such that for any large ($L \gg 1$) instance of the matrix V, for $p > p_c$ it is always possible to find at least one "gapless" planar diagram, which involves in its formation almost all vertices and only $\sim o(L)$ vertices are missing, while for $p < p_c$ a finite fraction of missing vertices of order $\sim O(L)$ exists.

Fig. 4.1(b) illustrates the dependence of $f_\infty(c)$ for the Bernoulli polymers. Firstly note, that the values f_∞ in the Bernoulli model distinguish from the respective values for discrete alphabetic sequences less than by 1%, that justifies the applicability of the model. Two different phases are observed: for $p > p_c$ one has a gapless perfect matching with all nucleotides involved in planar binding, while for $p < p_c$ there is always a finite fraction of gaps in the best possible matching.

To get numerically the critical transition point, we look for the fraction, $\eta_L(p)$, of sequences, which allow perfect matchings, in the whole ensemble of random sequences, one has $\eta_\infty(p) = 1$ for $p > p_c$, and $\eta_\infty(p) = 0$ for $p < p_c$. The corresponding dependencies are shown in Fig. 4.2(a) for the different polymer lengths, $L = 500, 1000, 2000$. As $L \to \infty$, the function $\eta_L(p)$ tends to a step function. The scaling analysis allows us to determine the phase transition point as $p_c \approx 0.379$, which corresponds to effective alphabet:

$$c_c \approx 2.64$$

Fig. 4.2(b) shows that curves with different L collapse, demonstrating the scaling behavior $\eta\left((p - p_c)/L^\nu\right)$, giving the transition width in form of power-law decay $L^{-\nu}$, with $\nu = 0.5$.

The convergence of the function f_L to a limiting value $f_\infty(p)$ in the perfect

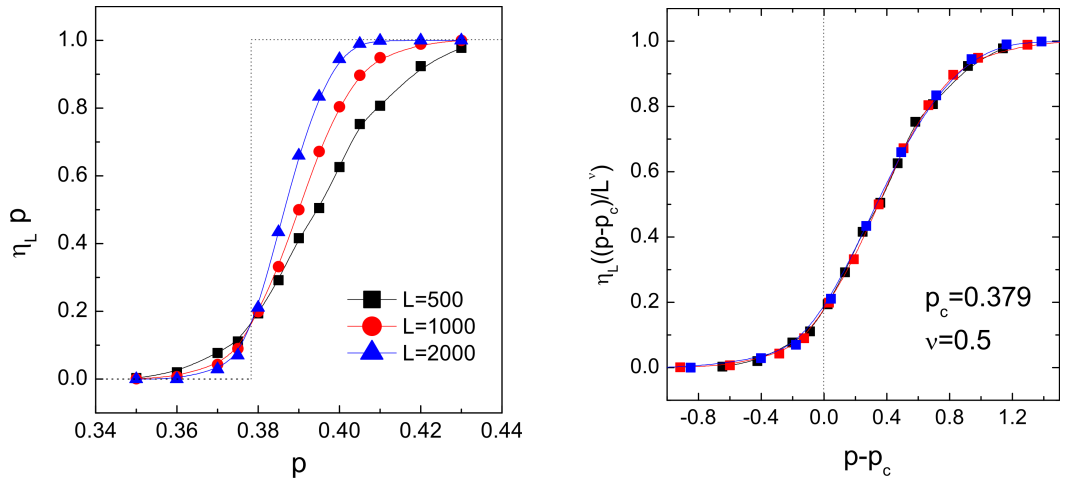


Figure 4.2. Numerical results for Bernoulli random polymer model.(a) The fraction of perfect matchings $\eta_L(p)$ as a function of the density p of ones in the contact matrix V for chain lengths $L = 500, 1000, 2000$, averaged over 10^4 realisations. The dashed line corresponds to the thermodynamic limit $L \to \infty$, yielding the critical value $p_c = 0.379$. (b) The scaling analysis of curves, corresponding to the different chain lengths L. The tting procedure gives the exponent of the transition width $\nu = 0.5$.

and imperfect phases has, respectively, an exponential and a power–law tails:

$$\begin{cases} f_\infty(p) - f_L(p) \sim e^{-L/\ell(p)} & \text{for } p > p_c \\ f_\infty(p) - f_L(p) \sim L^{-\alpha(p)} & \text{for } p < p_c \end{cases} \tag{4.11}$$

where the screening length $\ell(p)$ diverges at the point $p = p_c$ (two examples for $p = 0.38$ and $p = 0.4$ are shown on Fig. 4.3(a) in the semi-logarithmic scale), and finite-size scaling analysis gives $0.8 \leq \alpha(p) \leq 1$ (see Fig. 4.3(b) for two examples, $p = 0.32$ and $p = 0.34$ on the log-log plot). Note that the exponential scaling in the perfect phase may not be universal (with respect to other models) and is likely to be a feature of the Bernoulli model, while the power-law behavior in the imperfect phase appears in other models, e.g. for integer-valued "alphabet".

Perfect and imperfect phases are different also by their fluctuational behavior. Perfect regime is characterized by fast exponential fall of the fluctuation with increasing L, as well as in imperfect phase the fluctuations increases as $L^{1/2}$ (compare with Fig. 3.2(b)).

Naturally to expect, that asymptotic and fluctuation behavior of $f(L)$ depends on the details of random polymer model, for example the complementarity rules.

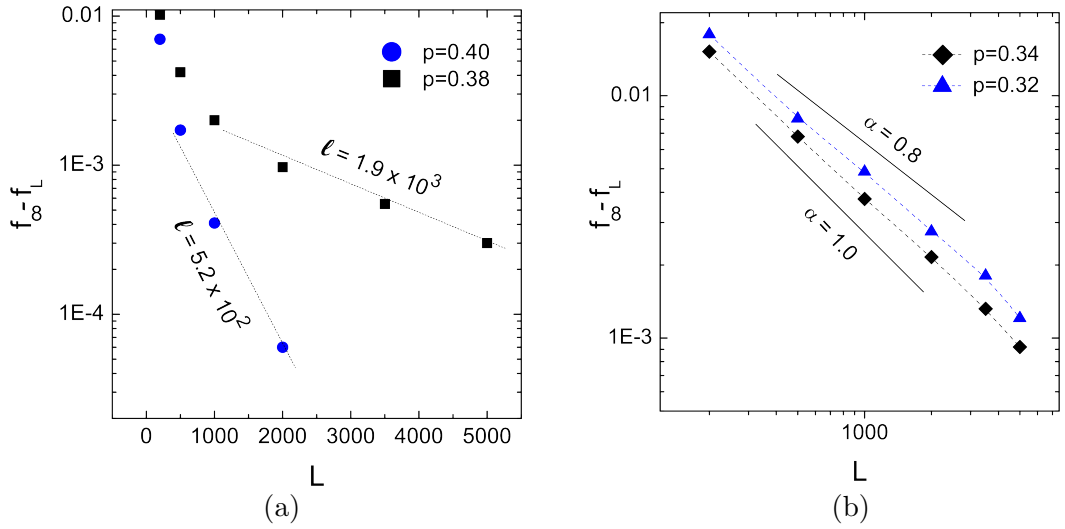


Figure 4.3. Convergence of fraction of links, involved in planar binding, f_L to the limiting value f_∞ in two regimes, $p > p_c$ and $p < p_c$. (a) In the perfect phase, the exponential convergence is demonstrated for $p = 0.38$ and $p = 0.4$ in the semi-logarithmic scale. The screening length $\ell(p)$ diverges as p approaches the critical value p_c. (b) In the imperfect phase, the power-law behavior is shown for $p = 0.32$ and $p = 0.34$ in the log-log scale. The exponent $\alpha(p)$ as a function of p takes values between 0.8 and 1. The data points are averaged over 1000 realisations.

However, the existence of the transition between two different phases is irrelevant to these details.

We can draw an analogy of this transition with the transition taking place in percolation theory [110]. A representative question is as follows. Put a liquid on top of a porous material. Will the liquid be able to percolate through the sample, reaching the bottom? This physical question is modelled mathematically as a three-dimensional network of $n \times n \times n$ vertices, usually called "sites", in which the edge or "bonds" between each two neighbors may be open (allowing the liquid pass through) with probability p, or closed with probability $1 - p$ (they all are assumed to be independent). Therefore, for a given p, what is the probability that an open path exists from the top to the bottom? The behavior for large n is of primary interest. There is a critical p below which the probability is 0 and above which the probability is 1 [110].

4.3 Analytical estimates of the critical point

In this section we focus on analytical consideration of the phase transition in the framework of the Bernoulli polymer model.

4.3.1 The mean-field estimate

We formulate the problem in terms of planar diagrams (Fig. 1.7). Consider a random graph with numbered nodes and adjacency matrix V. The problem is reduced to the question, how to choose among the allowed by V contacts $L/2$ arcs, providing complete planar structure. That is all nodes are involved in the structure exactly once and any pair of connections (i_1, j_1) and (i_2, j_2) is satisfied by [97]:

$$(j_1 - i_1)(j_2 - i_1)(j_1 - i_2)(j_2 - i_2) > 0 \tag{4.12}$$

A naive estimation of p_c fully consistent with the consideration in the previous section, can be easily obtained via the following mean-field-like argument. Since each arc in the diagram is present with a probability p, the probability that the whole configuration is allowed, is given by $p_1 = p^{L/2}$. One can introduce the probability p_k to have k allowed configurations from #. For example, for $k = 2$

$$p_2 = p^{L/2} p^{L/2} p^{-n_{1\cap 2}} = p^L p^{-\kappa_2 L} \tag{4.13}$$

where $n_{1\cap 2} \equiv \kappa_2 L$ is the number of shared arcs for two randomly chosen planar diagrams, averaged by ensemble #. The probability p_3 is defined as:

$$p_3 = (p^{L/2})^3 p^{-n_{1\cap 2\cap 3}} = p^{3L/2} p^{-C_3^2 \kappa_2 L} p^{\kappa_3 L} \tag{4.14}$$

The values κ_k can be calculated with any accuracy, so κ_2 lays strictly in the interval $[1/15, 1/14.8]$. The probability $\mathcal{P}$ to have *at least* one matching diagram for given density p of the matrix V is:

$$\mathcal{P} = \#p_1 - \frac{\#(\# - 1)}{2} p_2 + C_\#^3 p_3 + \dots \tag{4.15}$$

Naively assuming that planar diagrams in the fully-connected ensemble are *statistically independent*, we get the probability to have at least one perfect planar matching configuration:

$$\mathcal{P} = 1 - (1 - p^{L/2})^{C_{L/2}} = 1 - \exp\left(-p^{L/2} C_{L/2}\right) \tag{4.16}$$

where the last equality is valid for $L \to \infty$. In this limit, the probability $\mathcal{P}$ is equal to one for $p > p_c$, and to zero for $p < p_c$. The perfect-imperfect naive

mean-field threshold p_c is then given by the condition

$$\lim_{L\to\infty} p_c \left[C_{L/2}\right]^{2/L} = 1, \tag{4.17}$$

yielding $p_c = 1/4$. This consideration coincides with naive estimates in Section 4.1, just reformulated in terms of planar diagrams. Therefore, it provides only a lower bound to the true value of p_c. A careful account for correlations leads to a natural generalization of the critical condition (4.17):

$$\lim_{L\to\infty} \xi(p_c) \left[C_{L/2}\right]^{2/L} = 1,\ \xi(p_c) = 1/4, \tag{4.18}$$

where $\xi(p)$ is some weight of correlated diagrams to be determined.

4.3.2 Combinatorics of "corner counting"

An estimation of $\xi(p)$, and therefore of p_c, can be obtained by exploiting the combinatorial properties of Dyck paths. The consideration below provides an intuitive understanding of the statistical reasons beyond the shift of the transition probability from the mean-field value $p_c = 1/4$.

Our estimation is relied on the following observation: the probabilities to find different arcs in a perfect matching structure crucially depend on the lengths of arcs. Consider a perfect structure consisting of $L/2$ arcs connecting L points. In the limit $L \to \infty$ the *local statistics* of "up" and "down" steps in a corresponding Dyck path becomes independent on the global constraint for the random walk to be a Brownian excursion. Using the bijection between Dyck paths and arc diagrams, we see that the arc is drawn between i-th and j-th steps if and only if the i-th step is $\nearrow$ ("up") and j-th step is the first step $\searrow$ ("down") at the same height after i. Therefore, the probability to find an arc from i to j in a randomly chosen diagram can be formally written as a "correlation function":

$$P(i,j) = \frac{\langle \nearrow |\mathcal{D}_{i+1,j-1}| \searrow \rangle}{2^{j-i+1}}. \tag{4.19}$$

In this expression, the denominator represents the total number of possible sequences from i-th to j-th step; $\mathcal{D}_{i+1,j-1}$ is a Dyck path between $(i+1)$-th to $(j-1)$-th steps: this part of the walk should be a Dyck path itself to return to the same spatial coordinate for the *first time* at j-th step. The number of such

Dyck paths are given by the Catalan numbers $C_{(j-i-1)/2}$. Thus, $P(i,j)$ depends only on $k = j - i$ and equals to

$$P(i,j) = \frac{C_{(k-1)/2}}{2^{k+1}}, \tag{4.20}$$

they are non-zero for odd k only: $P(i,i+1) = 1/4$, $P(i,i+3) = 1/16$, $P(i,i+5) = 1/32$, *etc.* The whole set of $P(i,j)$ sums to $\sum_{k=1}^{\infty} P(i,i+k) = 1/2$, which has a meaning of a probability that the i is a starting (rather than ending) point of an arc.

Thus, the fraction of short arcs, in particular the shortest arcs of length $k = 1$, represented by "up corners" $\bigwedge$ in a Dyck path, is exceptionally high. Indeed, in a typical fully connected diagram one half of the arcs ($L/4$ out of $L/2$) correspond to such corners. Moreover, while a fraction of long arcs chosen in each particular diagram is decaying at $L \to \infty$ (indeed, the number of possible long arcs is of order L^2 , so the fraction of those chosen in each structure is of order L^{-1}), the fraction of chosen corners converges to $1/4$ (there are $L-1$ possible corners, $L/4$ of them are chosen in a typical structure). Therefore, the values of quenched weights $V_{i,i+1}$ assigned to the short arcs in our model influence the existence of a perfect arc structure in a crucial way. In what follows we quantitatively estimate how this exceptional role of the sub-diagonal values $V_{i,i+1}$ influences p_c.

Assume that the typical arc structure is constructed as follows: i) take $1/4$ corners (from $L-1$ possible places) such that none of them touch each other, ii) select remaining $L/2 - L/4 = L/4$ arcs at random from ensemble of *any longer* arcs. Since the total number of longer arcs is of order of $L^2 \gg L/4$, we assume that the quenched disorder in the entries $V_{i,j}$ away from the sub-diagonal can be ignored, and the contribution from the longer arcs into $\xi(p)$ remains as it is in the mean-field case (each arc is allowed with a probability p independently of others), thus

$$\xi^{L/2}(p) = \underbrace{p^{L/4}}_{\text{long arcs}} \underbrace{\mathcal{P}_{\wedge}(p)}_{\text{corners}}, \tag{4.21}$$

The contribution of corners, $\mathcal{P}_{\wedge}(p)$, is to be determined. It has a meaning of a probability to take $L/4$ corners at random (respecting the non-touching constraint) in a way that all of them belong to the set of pL allowed ones.

To estimate this probability, note that due to the non-touching constraint

the problem of distributing corners can be mapped onto a problem of choosing $L/4$ objects (corners) out of $3L/4$ ones ($L/4$ corners plus $(L-2)\times L/4 \simeq L/2$ unmatched vertices, see Fig. 4.4). The number of corresponding partitions $\mathcal{Z}$ is

$$\mathcal{Z} = C_{3L/4}^{L/4} = \frac{\frac{3L}{4}!}{\frac{L}{4}!\frac{L}{2}!} \tag{4.22}$$

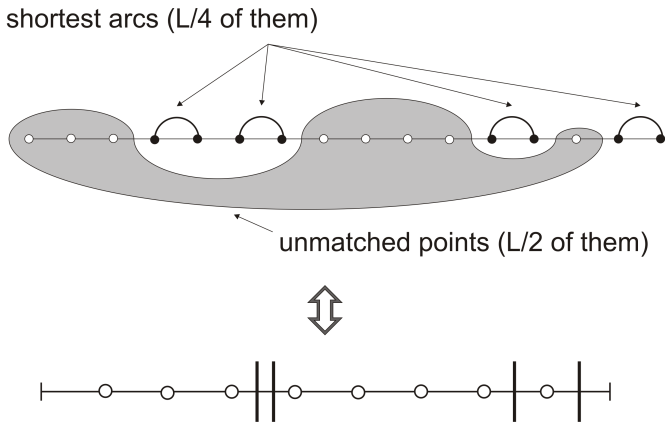


Figure 4.4. Computation of $\mathcal{Z}$ and $\mathcal{Z}(p)$. Selection of $L/4$ non-touching arcs on the set of L points ($L/2$ black nodes remain unmatched) is reformulated as a partitioning of vertical segments (arcs) between black dots (unmatched points). A certain number of partitions are forbidden by the matrix of contacts V.

In the Bernoulli model, only the fraction p of all arc positions is allowed. Because of the non-touching constraint, it is natural to assume that of $3L/4$ positions in the "point-and-stick" representation in Fig. 4.4(b) only $p(L-L/4) = 3pL/4$ are allowed on average (i.e., correspond to unity weights in the connectivity matrix V).

Thus, the number of allowed partitions can be estimated as

$$\mathcal{Z}(p) = C_{3pL/4}^{L/4} = \frac{\frac{3pL}{4}!}{\frac{L}{4}!(\frac{3pL}{4}-\frac{L}{4})!}. \tag{4.23}$$

Here $\mathcal{Z}(p)$ is the average number of possibilities to distribute shortest non-touching arcs at a given fraction p of allowed arcs, and

$$\mathcal{P}_{\wedge}(p) = \frac{\mathcal{Z}(p)}{\mathcal{Z}} \tag{4.24}$$

is a probability, given p, to pick an allowed set of short arcs at random. Substituting (4.23) into (4.24) one gets in the limit $L\to\infty$ the following result for $\xi(p)$

(see (4.18)):

$$\ln \xi(p) = \frac{1}{2}\ln p + \frac{3p}{2}\ln\frac{3p}{2} - \frac{3p-1}{2}\ln\frac{3p-1}{2} - \frac{3}{2}\ln\frac{3}{2}. \tag{4.25}$$

Being substituted into (4.17), (4.25) gives an estimate for the transition point

$$\begin{aligned} &\ln \xi(p_c) = -\ln 4; \\ &p_c \approx 0.35 \ (c_c = 2.87) \end{aligned} \tag{4.26}$$

We see therefore that the transition point shifts significantly from its mean–field value due to the special role of a quenched disorder in the sub-diagonal entries $V_{i,i+1}$ of the connectivity matrix V.

4.3.3 Self-consistent field theory for planar arc counting

A different way to attack the planar matching problem consists in using the matrix model approach (see Paragraph 1.2.2) proposed in [70] and $1/N$-expansion, a standard technique of the quantum field theory. For the set of L vertices in our problem, associate to a vertex i an Hermitian matrix $(\phi_i)_{N\times N}$. The L-point generating functional Z_L can be written as follows:

$$Z_L(N;V) = \frac{\int d\phi_1 ... d\phi_L e^{-H_0} \frac{1}{N}\mathrm{tr}\,(\phi_1 ... \phi_L)}{\int d\phi_1 ... d\phi_L e^{-H_0}} \equiv \langle \phi_1 ... \phi_L \rangle_{H_0} \tag{4.27}$$

where

$$H_0 = \frac{N}{2}\sum_{i,j}(V^{-1})_{ij}\mathrm{tr}(\phi_i\phi_j) \tag{4.28}$$

Since $\mathrm{tr}(\phi_i\phi_j) = \sum_{a,b}\phi^i_{ab}\phi^j_{ba} = \sum_{a,b,c,d}\delta_{ad}\delta_{bc}\phi^i_{ab}\phi^j_{cd}$, every propagator enters with a $1/N$-factor, while every loop gives a N–factor. Due to the Wick theorem, one has:

$$\langle \phi_1 \dots \phi_L \rangle_{S_0} = \left\langle \sum_{\text{pairs } k,k'} \prod \phi_k \phi_{k'} \right\rangle_{H_0} \tag{4.29}$$

where each non-planar configuration comes with a factor $1/N^2$ to some power and therefore vanishes in the $N \to \infty$ limit. Thus, the generating function $Z_L(N;V)$ counts in the limit $N \to \infty$ the number of planar diagrams with exactly $L/2$ arcs

(on genus $g = 0$ surface) compatible with a specific realization of the disorder defined by the matrix V. In the absence of any disorder, one can set $V_{ij} \equiv \alpha$ for any (i, j), where α is some constant (it corresponds to the $p = 1$ limiting case). In this case the multi-dimensional integral (4.27) can be reduced by a series of Hubbard-Stratonovich transformations to a one-dimensional integral involving the spectral density of a Gaussian matrix, which is a well-known result of the Random Matrix Theory (RMT). We will refer to this realization of A as to the fully-connected case. If we set $\alpha = 1$, we get [70]

$$\lim_{N\to\infty} Z_L(N; V) = C_{L/2}, \tag{4.30}$$

where $C_{L/2}$ is the Catalan number, as it should be. However, for a generic disordered matrix V, the calculations are intractable. Still, we show below that by averaging over the matrix distribution (4.8) and by applying the self-consistency arguments, we are able to treat the partially-connected system with $0 < p < 1$ as an effective fully-connected system with α different from one, thus obtaining a correction to the naive mean-field result.

According to the consideration above, the function $\xi(p)$ defined by (4.18) can be calculated within the matrix approach by averaging $Z_L(N; V)$ over the distribution (4.9). To this end, we use the standard Hubbard-Stratonovich transform and integrate over V with the weight (4.9):

$$\begin{aligned} &\int dV\ P(V)\, Z_L(N; V) = \\ &C \int \prod_{k=1}^{L} d\phi_k \frac{1}{N} \mathrm{tr}\,(\phi_1 \dots \phi_L) \int \prod_{m=1}^{L} dh_m e^{iN \sum_i \mathrm{tr}(h_i \phi_i)} e^{S} \end{aligned} \tag{4.31}$$

where C is a constant, $S = S_0 + U$, and

$$S_0 = -\frac{pN}{2} \sum_{ij} \mathrm{tr}(h_i h_j), \tag{4.32}$$

$$\begin{aligned} U = &\frac{p(1-p)N^2}{8} \sum_{ij} [\mathrm{tr}(h_i h_j)]^2 \\ &- \frac{p(1-p)(1-2p)N^3}{48} \sum_{ij} [\mathrm{tr}(h_i h_j)]^3 + \dots \end{aligned} \tag{4.33}$$

Up to this point, all the calculations are exact. The S_0 term (4.32) corresponds

to a fully-connected matrix with an additional factor p behind. If this term is the only present, then, performing the inverse Hubbard-Stratonovich transformation and returning to the functional of the type (4.18), we get $\xi(p) = p$, recovering the value $p_c = 1/4$ given by the critical condition (4.17).

The correction to p_c due to the rest of the series U (4.32) can be estimated as follows. The series defined by the action S can be thought of as a Gaussian theory with the interaction U. Since U contains an infinite number of terms, it is impossible to treat it perturbatively. Still, we can use a self-consistent non-perturbartive approach reminiscent of the Feynman's variational principle [111] in the field theory: as all the fields $\{h_i\}_{i=1,\dots,L}$ in (4.33) are equivalent, we assume that the average $\langle N\mathrm{tr}(h_i h_j)\rangle_{S_0} \equiv \bar{U}$ is independent on (i,j). Within the adopted mean-field approximation, the replacement $e^S = e^{S_0} e^{\langle U\rangle}$ is supposed, where

$$\begin{aligned}\langle U\rangle = \; & \frac{p(1-p)N}{8}\bar{U}\sum_{ij}\mathrm{tr}(h_i h_j) \\ & -\frac{p(1-p)(1-2p)N}{48}\bar{U}^2\sum_{ij}\mathrm{tr}(h_i h_j) + \dots\end{aligned} \tag{4.34}$$

Resumming the series (4.34), we obtain the following self-consistent equation for the "propagator" $\bar{U}$:

$$\frac{1}{\bar{U}} = -\frac{2}{\bar{U}}\log\left[1 - p + p\exp\left(-\frac{\bar{U}}{2}\right)\right]. \tag{4.35}$$

The equation (4.35) yields $\bar{U} = -2\log\left[1 - (1 - 1/\sqrt{e})/p\right]$. Hence, finally, we can write

$$S = -\frac{\xi(p)N}{2}\sum_{ij}\mathrm{tr}(h_i h_j) \tag{4.36}$$

where

$$\xi(p) = \left(-2\log\left[1 - \frac{1 - 1/\sqrt{e}}{p}\right]\right)^{-1}. \tag{4.37}$$

Substituting (4.37) into (4.18), we get an estimation for the critical value

$$p_c^* = 0.455$$

Although the self-consistent approximation (4.35) seems to be rather crude (the numerical estimation of p_c for large matrices is $p_c \approx 0.38$), it leads to the

correct direction of the shift of p_c from the naive mean–field value $p_c = 0.25$. It would be interesting to understand how to treat the interaction term U (4.34) more properly.

4.4 Matching vs freezing

In this section we discuss the relation of the perfect/imperfect transition with the transition, described in Paragraph 1.2.1.

4.4.1 Glassy phase transition in Bernoulli random polymer

Regardless of particular details of models mentioned in Paragraph 1.2.1, it is clear that the existence of the glassy phase is possible only in a sufficiently disordered and frustrated system. Besides the planarity constraint, the property shared by all simple models of random RNA, the Bernoulli model is described by a unique disorder parameter, p, that controls the density of allowed contacts. It is also clear that in this model, the appearance of the glassy phase is impossible above a certain threshold, p^*. Indeed, it is well-known that for $p = 1/2$ (corresponding to an effective alphabet $c = 1/p = 2$), there is no transition to the glassy phase at all, and the system remains always in the molten phase [44, 43]. Below, we present arguments, supporting the hypothesis that p^* is equal to the critical value p_c discussed above.

To identify the dependence of the molten-glass transition temperature on the effective alphabet (defined as $c = 1/p$), we follow the procedure suggested in [43] and described in Paragraph 1.2.1. In the high-temperature regime the disorder is irrelevant and the entries of the adjacency matrix can be replaced by constants, $V_{ij} = \alpha$ (this corresponds to a homopolymer-like behavior in polymer language). In this regime the free energy of the chain of the length L scales linearly with L, up to a logarithmic correction which is just the logarithm of the power-law multiplier in the Catalan number (4.2) enumerating all possible structures: $F(L,T) = f(T)L - (3T/2)\ln L$, where $f(T)$ is some non-universal function of the temperature. In particular, the energy cost of imposing a bond connecting two monomers at distance L from each other equals in the high temperature

regime

$$\Delta F(L,T) = F(L,T) - 2F(L/2,T) = \frac{3}{2}T \ln \frac{L}{4}. \tag{4.38}$$

The violation of this behavior indicates [43] the appearance of the glassy phase. This fact can be used to detect the transition temperature in the Bernoulli model. Namely, use the following fit for $\Delta F(L,T)$ (see (1.6))

$$\Delta F(L,T) = a(T) \ln L + b(L), \tag{4.39}$$

and plot the temperature dependence of the coefficient $a(T)$, see Fig. 4.5. As the coefficient deviates from $a(T) = 3T/2$, the glass transition occurs. Note, that the logarithmic fit (4.39) for the free energy does not have to give a true asymptotic at low temperatures for this procedure to be valid (indeed, the true asymptotics is known to include power-law and logarithm-squared terms [54]).

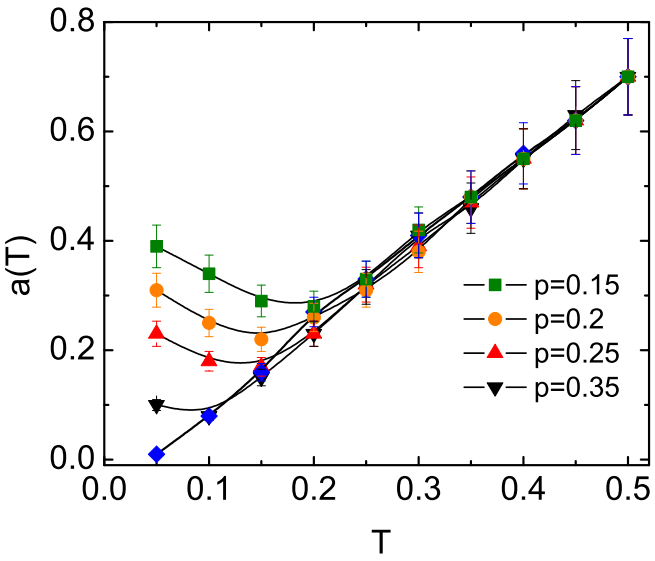


Figure 4.5. The dependence of the logarithmic proportionality coefficient of the pinching energy cost on the temperature. The simulations were done for $p = 0.15, 0.2, 0.25, 0.35, 0.5$. For p larger then some threshold p^* $(0.35 < p^* < 0.5)$, $a(T)$ seems to follow the $a(T) = 3T/2$ law, corresponding to the molten phase, up to very low temperatures. For $p > p^*$, the $a(T)$-dependence deviates from the high-temperature behavior at some temperature, which we identify as a critical temperature of transition to the glassy phase. The data points are averaged over 10^4 realisations.

As it follows from Fig. 4.5, the high-temperature behavior (4.39) is indeed observed at high temperatures, and is violated at a certain temperature T_c. Following [43] we identify this regime change with the molten-glass transition. We

see that with the increase of the parameter p, the critical temperature T_c shifts to lower values, approaching the zero value for some $0.35 < p^* < 0.5$. At low temperatures, the numerical computations become very time consuming and demand the consideration of extremely long sequences, which leads to the loss of precision in the vicinity of p^*. However, it seems that the hypothesis $p^* = p_c$ still holds: the sequences corresponding to $p > p_c$ remain in the molten phase, the pinching free energy (4.40) has the same dependence even for very low temperatures.

4.4.2 The relation of perfect matching transition with glassy phase transition

The results presented in this work suggest the following generic picture. Fig. 4.6 represents the phase diagram of the Bernoulli model for random RNA chains. We have studied the perfect-imperfect transition at zero temperature, separating two matching regions: with and without gaps. Analytically, we proved the existence of the transition from the perfect matching region to the imperfect one, and provided estimates for the values of the transition point, p_c. Using the exact dynamical programming algorithm (2.22), we found that this critical value to be $p_c \approx 0.379$, highlighted by a thick dashed line (B-C) in Fig. 4.6. The previous studies have been mostly concentrated on the description of the finite-temperature molten-glass transition for a sufficiently frustrated model with a fixed alphabet (a fixed p in the Bernoulli model). An example of such a phase transition point is marked by a thin dashed line in Fig. 4.6, and corresponds to an intensively studied case of the 4-letter alphabet ($p = 0.25$). The ensemble of critical points for different values of p forms a critical curve (A-B) in the (T, p) plane.

As already discussed in the previous section, the computational cost increases drastically for temperatures close to zero (and, hence, in the vicinity of p_c). However, we can still try to carry out the analysis of the pinching free energy $\Delta F(L, T)$ at zero temperature, using the exact dynamical programming algorithm (2.22). Indeed, the glassy phase does not exist if $\Delta F(\infty, 0) = 0$. This happens for $p > p^*$, where p^* is defined as the density of constrains, for which the critical temperature is zero: $T_c(p^*) = 0$. The corresponding plot is shown in the inset of Fig. 4.6. According to (4.11), the pinching free energy decreases with growth of L in the imperfect matching phase, while increases (with growth of L) in the

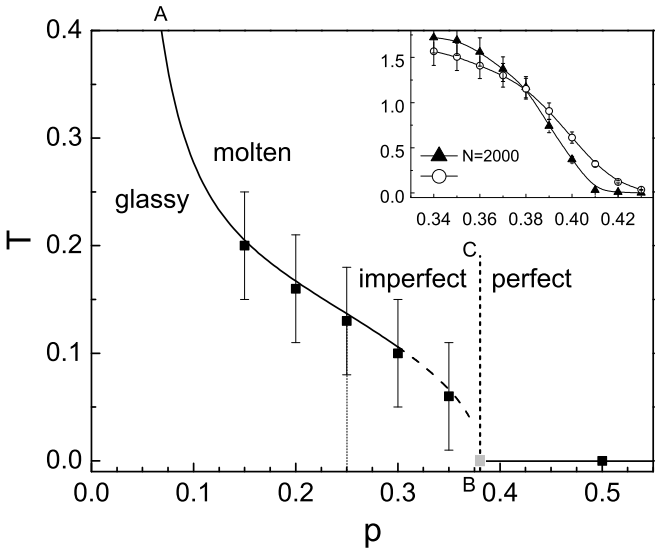


Figure 4.6. The phase diagram of Bernoulli model on the $(\mathbf{T}, \mathbf{p})$ plane. The data points correspond to the critical temperature T_c of the molten-glass transition for different values $p = 0.15, 0.2, 0.25, 0.3, 0.35, 0.5$. An intensively studied case, corresponding to a 4-letter alphabet ($p = 0.25$), is highlighted by a thin dashed line. The critical curve (A-B) separates glassy and molten phases. We conjecture that at zero temperature, the critical curve's endpoint B position p^* coincides with the critical point p_c for the perfect-imperfect transition. The thick dashed line (B-C) separates the perfect and imperfect regions of the phase diagram, and the glassy phase lies entirely inside the region, characterized by gaps. Inset – an evidence for the conjecture $p^* = p_c$: study of the pinching free energy $\Delta F(L, T)$ at zero temperature. In the limit of large L, the glassy phase is absent for $p > p^*$, characterized by $\Delta F(\infty, 0) = 0$. The point p^* can be identified as a crossing point for different $\Delta F(L, 0)$ curves, presented here for $L = 1000$ and $L = 2000$, and it's value is found to be very close to $p_c = 0.379$.

perfect matching regime. Hence, the value of p^* in the large L limit can be identified as a crossing point of $\Delta F(L, 0)$ curves for different L. The crossing point for $L = 1000$ and $L = 2000$ is indeed found to be very close to the value $p_c = 0.379$, strongly supporting the hypothesis $p^* = p_c$. The aforementioned results indicate that the critical curve $T_c(p)$ crosses zero at the critical value p_c. Hence, the perfect-imperfect transition point seems to lie at the critical line, separating molten and glassy regions, and coincides with its limiting $T = 0$ value. We see that although the glassy phase exists only in the region where the gaps are present, the molten phase lies in both, perfect and imperfect, matching regions.

Because of the one-parameter dependence, the Bernoulli model is probably the simplest model for modeling the secondary structure of the RNA, that captures the essential physical properties of the process. Being applied to the studies of the thermodynamic properties of random RNAs, the problem introduced in this

work provides some enlightenment on the nature of molten-glass transition at zero temperature. Starting from the Bernoulli model, one could directly generalize our approach to investigate more sophisticated and realistic models of the RNA secondary structure, for example, by introducing the minimal allowed hairpin length [44, 43, 53], taking into accounts the pseudoknots [70] and different binding probabilities [70, 61].

4.5 Other approaches to non-integer alphabetic sequences

The principal disadvantage of a Bernoulli model consists of the absence of explicit correspondence of a Bernoulli contact matrix V and a primary polymer structure. We cannot distinguish different types of monomers and, thus, the transitivity is completely broken. In this section we describe other methods to generate a random sequence with effectively non-integer alphabet, in which the transitivity is preserved entirely or partially.

4.5.1 Correlated alphabet

One of the approaches is to distribute three types of monomers (A,B and C) in a chain not randomly, but correlated. We can arrange a chain like Markov process[107], i.e, the probability of future step (monomer type) depends on the present nucleotide:

	A	B	C
A	$1-2\epsilon$	ϵ	ϵ
B	ϵ	$1-2\epsilon$	ϵ
C	ϵ	ϵ	$1-2\epsilon$

The (i,j)-th matrix element determine the probability to have the monomer of the j-type after the i-th monomer. Changing ϵ from 0 to $\frac{1}{3}$ we mimic alphabets with effective number of letters continuously varying from 1 to 3. We chose the probability matrix in this form to make it symmetrical in respect all three monomer types. The relation between the parameter ϵ and alphabet c can be

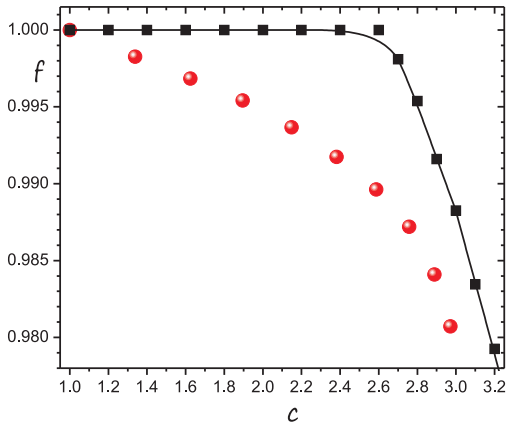


Figure 4.7. The model of correlated alphabet. The dependence of binding probability on the alphabet (the red points). Numerical results are obtained for ensemble of 10^3 samples of the length $L = 5000$. For comparison the respective dependence for the Bernoulli model (the black curve) is depicted.

easily received, using the Shannon information entropy [112]:

$$c = \left(\frac{1}{\epsilon} - 2\right)^{2\epsilon} \frac{1}{1-2\epsilon} \tag{4.40}$$

As a result, we have a polymer with a block structure, the typical block size depends on ϵ. Fig. 4.7 illustrates the dependence of the specific free energy on the alphabet size in this model. There is no transition as in the Bernoulli model. Even for the alphabets $c < 2$, the perfect matching structures are not formed. The explanation of this effect is quite simple. After the erasing procedure (see Section 4.1), the tree-letter correlated polymer is reduced to three-letter sequence with a random monomer distribution (we erase all the blocks of the same monomers). But, a three-letter random sequence form the structure with $O(L)$ gaps. The length of the sequence rest after erasing is proportional to the sequence length L. In this way, for any ϵ we have finally a three-letter random sequence, which belongs to imperfect regime. However, note, that the dependence (Fig. 4.7) has a sharp decline in the vicinity of the Bernoulli critical point.

4.5.2 Rational alphabet

Other model, in which the transitivity remains partially can be formulated as follows. Consider a random polymer with P different monomer types, but allow

each of them to bind with Q others monomer types according to some specific complementarity rules. For example, the alphabet $c = 5/2$ can be presented as a random five-letter sequence (A,B,C,D,E are the monomer types) with the complementary rules organizing by the pentagon (Fig. 4.8). These rules can be constructed for any rational alphabet and the model is sensitive to the parameter Q. For example, the alphabets $c = 14/5$ and $c = 28/10$ give slightly little different results (Fig. 4.9). It is obvious that in the limit $P \to L$ we are back to the Bernoulli model. From the biological point of view, this model seems to be rather natural. Indeed, recall that in addition to the Watson-Crick base pairs the Wobble pairs can be formed (see paragraph 1.1.1).

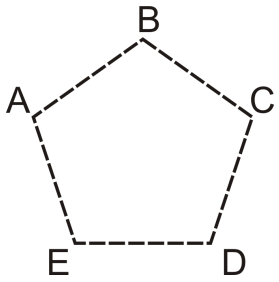


Figure 4.8. The scheme of allowed contacts for the alphabet c = 5/2

	A	B	C	D	E
A	0	1	0	0	1
B	1	0	1	0	0
C	0	1	0	1	0
D	1	0	1	0	1
E	1	0	0	1	0

Table 4.1. The matrix of allowed contacts for the alphabet c = 5/2

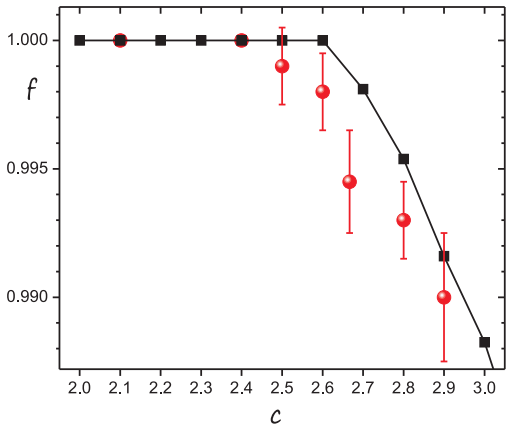


Figure 4.9. The model of rational alphabet. The dependence of binding probability on the alphabet (the red points). Numerical results are obtained for ensemble of 10^3 samples of the length $L = 1000$. For comparison the respective dependence for the Bernoulli model (the black curve) is depicted.

Summing up, we demonstrate here that alphabets with different number of

letters, c, are nonequivalent if one considers the matching problem of long random RNA. This nonequivalence is tightly coupled to the restrictions on the morphology of allowed secondary structures. Indeed, the existence of two regimes (for $c \leq c_{\rm cr}$ and $c > c_{\rm cr}$) is a peculiarity of RNAs and is due to the additional freedom in the formation of the complex cactus–like secondary structures typical for messenger RNAs. For linear matching problem used in DNA comparison, the fraction of nucleotides in the optimal alignment is less than 1 for any alphabet with $c > 1$. In our model the transition between two regimes occurs at $2 < c_{\rm cr} < 4$. The exact value of the critical alphabet size should be sensitive to the microscopic details of the model, and one can enumerate factors which are neglected in our model and which could shift the transition point to bigger or lower values away from the observed critical value.

On one hand, the presence of stacking energies and minimal loop sizes in real RNA leads to the bonds being effectively formed not by single nucleotides, but by blocks of them, increasing the effective alphabet size for given c, thus, decreasing $c_{\rm cr}$ in terms of the size of a "bare" alphabet. On the other hand, one would not expect any real–life random RNA to have a completely random structure with exactly equal concentrations of letters and no short–range correlations between them. Any such correlations reduce the information entropy of the sequence, and, therefore, lead to the decrease of the effective alphabet size, and thus, push $c_{\rm cr}$ to higher values. The exact value of $c_{\rm cr}$ is non-universal. However our analysis shows: (i) the existence of two different morphological regimes, depending on the number of nucleotide types in the alphabet, and (ii) the fact that this transition point can plausibly be rather close to a natural alphabet.

This particular number, obviously, sounds suggestive since it is exactly the number of nucleotide types in the alphabet used in real–world RNAs. The criticality on alphabet size, observed only for RNAs thus nicely rhymes with the modern opinion that the life originates from the template–directed replication of random RNA molecules (the so-called "RNA world" hypothesis) [113, 114]. Can it be indeed advantageous to have the alphabet of critical or close-to-critical size? For RNA to have a biological function it should: i) fold predictably, and ii) form a robust structure not too sensitive to thermal noise. Short nucleotide alphabets with $c < c_{\rm cr}$ tend to produce structures which have many different ground states, also compare with similar reasoning for proteins [115, 116]. On the other hand,

long alphabets correspond to loosely bound ground states with many unpaired nucleotides, which is disadvantageous in terms of stability of the structure. The critical alphabets, thus, seem to be optimal for biological purpose.

Chapter 5

Optimal transportation problem and RNA-like random interval model

In the works on investigation of thermodynamic properties of random RNAs [56, 53, 43, 62] it has been supposed that the energies, $u_{i,j}$, are quenched uncorrelated random variables, which depend on i and j, with a Gaussian distribution. The quenched randomness in the primary sequence affects on the height diagram (Fig. 1.5). It was found numerically that in glassy state of random RNA the roughness exponent γ takes the value close to $\gamma = 2/3$. Recent analytic estimates by field-theoretic arguments and RG analysis [56] give $\gamma \simeq 5/8$. Despite the essential progress in the field, the question about the value of roughness exponent for random heteropolymer RNAs is still open.

In this chapter we describe a new toy model of a heteropolymer chain capable of forming planar secondary structures typical for RNA molecules. In this model the sequential intervals between neighboring monomers along a chain are considered as quenched random variables, and energies of nonlocal bonds are assumed to be concave functions of those intervals. Several factors are neglected: the contribution of loop factors to the partition function, the variation in energies of different types of complementary nucleotides, the stacking interactions, and constraints on the minimal size of loops. However the model captures well the formation of folded structures without pseudoknots in an arbitrary sequence of nucleotides. Using the optimization procedure for a special class of concave–

type potentials, borrowed from optimal transport analysis, we derive the *local* difference equation for the ground state free energy of the chain with the planar (RNA–like) architecture of paired links. We consider various distribution functions of intervals between neighboring monomers (truncated Gaussian and scale–free) and demonstrate the existence of a topological crossover from sequential to essentially nested configurations of paired links.

5.1 Optimal transportation problem

In the classical transportation problem, let μ be a distribution of iron mines throughout the countryside (latter supposed to be one-dimensional), and ν the distribution of factories that require iron ore. We are interested in the following question: which mines should supply ore to each factory in order to minimize the total transportation costs. The cost per ton of ore transported from the mine at x to the factory at y is given by a function $c(x, y)$, so the problem can be formulated as a linear programming. Indeed, the similar question has been considered in the "optimal transportation theory" developed by Kantorovich [117] and Koopmans [118], though its origins date much further back to Monge (1781). The transportation problem belongs to NP class complexity. There are many different approaches to solve this problem, for example the simplex method, the north-west corner method, the method of following stone [119]. For mathematical simplicity other formulations of the optimal transportation problem are used [119, 120].

The optimal transportation problem can be formulated as a perfect matching problem on graphs [121, 122]. Recall some notions from combinatorial optimization on graphs. A matching in an undirected graph is any set of its mutually disjoint edges: no two edges from such set can share a vertex. A matching is called perfect if it involves all vertices of the graph (the number of vertices is then necessarily even). Depending on the structure of the graph, there may exist many perfect matchings. Sup pose that edges of a graph are endowed with real weights; then it makes sense to look for a perfect matching composed from a set of edges with a minimum sum of weights. We can treat a particular case of this minimum-weight perfect matching problem where the graph is complete, all its vertices are located on a line, and edge weights are related to distances between

the vertices along the line.

Following [123], we call the function w a *cost function of concave type* if for any $x_1, x_2, y_1, y_2 \in \mathbb{R}$ the inequality

$$w(x_1, y_1) + w(x_2, y_2) \leq w(x_1, y_2) + w(x_2, y_1) \tag{5.1}$$

implies that the intervals connecting x_1 to y_1 and x_2 to y_2 are either disjoint or one of them is contained in the other. Examples are: $w(x, y) = |x - y|^\alpha$ with $0 < \alpha < 1$, or $w(x, y) = \ln |x - y|$ extended to the diagonal $x = y$ by $-\infty$. In fact, whenever a cost function w of concave type is spatially homogeneous and symmetric, i.e., $w(x, y) = g(|x - y|)$, the function g must be strictly increasing and strictly concave [123].

Although somewhat idealized, the setting just described provides a reasonable model for applications in which shipping occurs along a single route: a railway line or highway, or along one coast of North America. Concavity of g reflects a shipping tariff that increases with the distance, even while the cost per mile shipped goes down. Despite its economic relevance, transportation with concave costs has received much less attention than the same problem for convex costs. The latter enjoys a sizable literature and, at least in one dimension, has been completely understood (see [124] or [125] for reviews). For concave costs on the other hand, it was only recently observed [126] that the solutions will not be smooth, but display an intricate structure which was unexpected; it seems equally fascinating from the mathematical and the economic point of view. To describe one effect in economic terms: the concavity of the cost function favors a long trip and a short trip over two trips of average length; as a result, it can be efficient for two trucks carrying the same commodity to pass each other travelling opposite directions on the highway: one truck must be a local supplier, the other on a longer haul. In optimal solutions, such "pathologies" may nest on many scales, leading to a natural hierarchy among the regions of supply (where $\mu \geq \nu$) and of demand (where $\mu \leq \nu$).

Let now $x_1 < x_2 < \cdots < x_{2n}$ be an even number of points on the real line $\mathbb{R}$. Consider the complete graph K_{2n} on these points, each of whose edges (x_i, x_j) is equipped with a weight $w(x_i, x_j)$. We look for a minimum-weight perfect matching in the graph K_{2n}, i.e., for a set of n edges such that the sum of their

weights is minimal.

Taking weights $w(x_i, y_i) = \ln|x_i - y_i|$, we can straightforwardly check that the minimal value of the total cost function $\Omega(x_1, y_1; ...; x_n, y_n)$,

$$\Omega(x_1, y_1; ...; x_n, y_n) = \sum_{\{\text{arcs}\}} \ln|x_i - y_i|,$$

is achieved at some planar configuration of pairings (see Fig. 5.1).

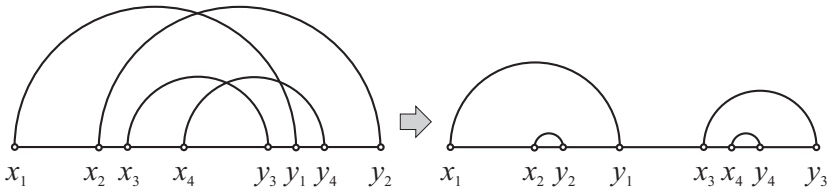


Figure 5.1. Optimization procedure. The use of concave-type cost function leads to the planar pairing.

5.2 Random Interval Model

Now we are in position to formulate our toy Random Interval Model (RIM) of a quenched heteropolymer RNA, in which the paired monomers interact with the energy $\varepsilon_{i,j}$, which is a concave function of the distance between monomers along the chain. In particular, we choose $\varepsilon_{i,j}$ of the form

$$\varepsilon_{i,j} = u \ln|x_i - x_j|; \qquad (j \neq i) \tag{5.2}$$

where u is some positive constant, and x_i, x_j are the coordinates of monomers i and j along the chain. The distances $d_i = |x_{i+1} - x_i|$ along the chain between sequential monomers capable to form pairs are quenched random variables taken independently from some distribution $P(d_i = d)$. Schematically, a typical realization of a RIM is depicted in Fig. 5.2 by arcs (a) and by a height diagram (b).

Let us emphasize that the key feature of the RIM consists in the fact that the interaction energy between paired monomers, $\varepsilon_{i,j}$, is a concave function of distance. In principle, one could take $\varepsilon_{i,j}$ in the form $\varepsilon_{i,j} = u|x_j - x_i|^{\alpha_1}$, where $0 < \alpha_1 < 1$, or $\varepsilon_{i,j} = -u|x_j - x_i|^{-\alpha_2}$, where $\alpha_2 > 0$ $(j \neq i)$. The main conclusions will hold, though the details are model-dependent.

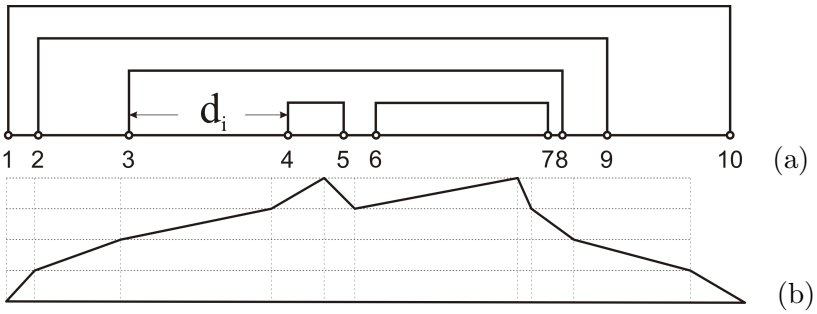


Figure 5.2. Typical configuration of a random interval RNA. Structure is shown by arcs (a), and by a height diagram (b).

Supposing that every monomer in the ground state structure is involved in binding, after some simplifications we get from (2.20):

$$F_{i,i+k} = \min_{s=i+1,i+3,\ldots,i+k} \Big[\varepsilon_{i,s} + F_{i+1,s-1} + F_{s+1,i+k}\Big] \tag{5.3}$$

with the "boundary conditions" $F_{i+1,i} = 0$ for any i. Note that it is enough to extend the min in s over values with odd increments with respect to i: no arc can cover an odd number of points, because otherwise some of them would be excluded from the structure due to planarity.

1. It is easy to see that the recursion (5.3) enumerates all planar arc structures on points $x_i, x_{i+1}, \ldots, x_{i+k}$. In particular it implies that

$$F_{i,i+k} \le \varepsilon_{i,i+k} + F_{i+1,k-1} \tag{5.4}$$

for all i and all odd $k \ge 1$ and that

$$F_{i,i+k} \le F_{i,i+\ell} + F_{i+\ell+1,i+k} \tag{5.5}$$

for all i and $1 \le \ell < k$ with k, ℓ odd. The latter property can be described as *sub-additivity* of the functional F: for two non-overlapping configurations of points $x_1 < x_2 < \cdots < x_{i+\ell}$ and $x_{i+\ell+1} < x_{i+\ell+2} < \cdots < x_{i+k}$, the value $F_{i,i+k}$ for the united configuration is not greater than the sum of the values $F_{i,i+\ell}$ and $F_{i+\ell+1,i+k}$ on the two partial configurations.

2. For the cost function $w(x_i, x_j) = \varepsilon_{ij}$ of concave type, the free energy functional is not only sub-additive, but has a stronger property: for all i, odd

$1 < \ell < k$ and even j with $j \leq \ell + 1$, F verifies the inequality

$$F_{i,i+k} + F_{i+j,i+\ell} \leq F_{i,i+\ell} + F_{i+j,i+k} \tag{5.6}$$

of which (5.5) is a particular case corresponding to $j = \ell + 1$. This property of F is called *submodularity*. It suffices to establish submodularity for $j = 2$ and $\ell = k - 2$:

$$F_{i,i+k} \leq F_{i,i+k-2} + F_{i+2,i+k} - F_{i+2,i+k-2} \tag{5.7}$$

the general case (5.6) is recovered by induction. Indeed, it was established in [122] that F satisfies a recursion

$$F_{i,i+k} \quad = \quad \min\bigl[\varepsilon_{i,i+k} \; + \; F_{i+1,k-1}; \quad F_{i,i+k-2} \; + \; F_{i+2,i+k} \; - \; F_{i+2,i+k-2}\bigr] \tag{5.8}$$

that combines (5.3) and (5.7). In other words, F is the *maximal* submodular functional that satisfies also (5.3).

Thus, the function $F_{i,i+k}$ for concave–type potentials satisfies not only the standard *nonlocal* (2.20), but also a *local* (5.8). For completeness a derivation of (5.8) (which coincides with minor modifications with [122]) is included in Appendix.

5.3 Topological properties of Random Interval Model

The random interval model defined above has some interesting topological features. Namely, the height diagram, h, which can be regarded as a quantitative characteristics of the "nesting degree" of planar arcs, displays for the Gaussian distribution of intervals a topological crossover from sequential pairing of monomers to essentially embedded (i.e. nested) one. Another interesting behavior of h is observed for a power–law (i.e. a scale–free) distribution of intervals, where the dependence of the height on the the exponent in the distribution has a well–defined maximum.

5.3.1 Numerical results

The truncated Gaussian distribution

Consider a random chain, in which the distances between nearest–neighboring monomers, $d_i = |x_{i+1} - x_i|$, are distributed with the truncated Gaussian distribution:

$$f(d,\sigma) = \begin{cases} \dfrac{C}{\sqrt{2\pi}\sigma} e^{-\frac{(d-\mu)^2}{2\sigma^2}}, & d_{\min} < d < d_{\max} \\ 0, & \text{else} \end{cases} \tag{5.9}$$

where $C = 2\left[\mathrm{erf}\left(\frac{d_{\max}-\mu}{\sqrt{2}\sigma}\right) + \mathrm{erf}\left(\frac{\mu-d_{\min}}{\sqrt{2}\sigma}\right)\right]^{-1}$ is the constant determined by the normalization condition $\int_{d_{\min}}^{d_{\max}} f(x,\sigma)\,dx = 1$. To avoid any possible misunderstandings, require all energies in (5.4) to be positive. Without loss of generality we can choose the following values of the parameters of the distribution function in (5.9): $\mu = 2$; $d_{\min} = 1$; $d_{\max} = 3$. The distribution function (5.9) is depicted in Fig. 5.3 for different dispersions σ.

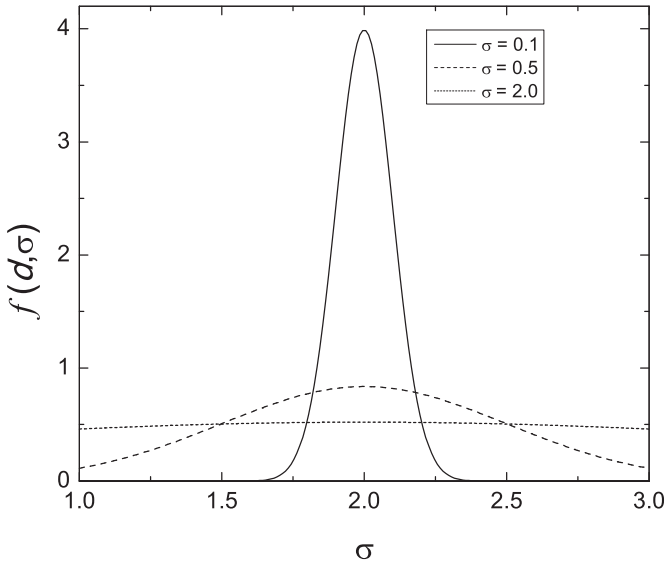


Figure 5.3. Truncated Gaussian distribution. $f(\sigma)$ of distances between nearest-neighboring monomers for $\sigma = 0.1; 0.5; 2.0$.

Our numerical analysis shows the existence of a crossover for random interval RNAs in topology of monomer pairings (planar diagrams) from sequential to essentially nested one. The parameter which controls this behavior is the dispersion σ of the distribution $f(d,\sigma)$.

For $\sigma < \sigma_{\mathrm{cr}}$, i.e. for essentially peaked distributions, the ground state of a random RNA chain has a height equal to 1. This means that only sequential

pairs of nearest neighboring monomers do form bonds. The value $\sigma_{\rm cr}$, at which the height diagram exceeds 1, we call the topological crossover point. The value $\sigma_{\rm cr}$ is computed for finite chains and depends on its total length, N; when N is increasing, the point of transition shifts towards smaller values and, apparently, reaches zero when N tends to infinity. Fig. 5.4 presents our numerical results for random interval chain with $N = 250, 500, 1000$ monomers. Above the crossover point, i.e. for $\sigma > \sigma_{\rm cr}$ the height diagram monotonically increases with σ and reaches some averaged stationary value for the RIM with uniform distribution of intervals ($\sigma \to \infty$). We prefer to use the term "crossover" instead of "transition" since we expect that it is not a true phase transition, the width of which shrinks to zero in the thermodynamic limit [127].

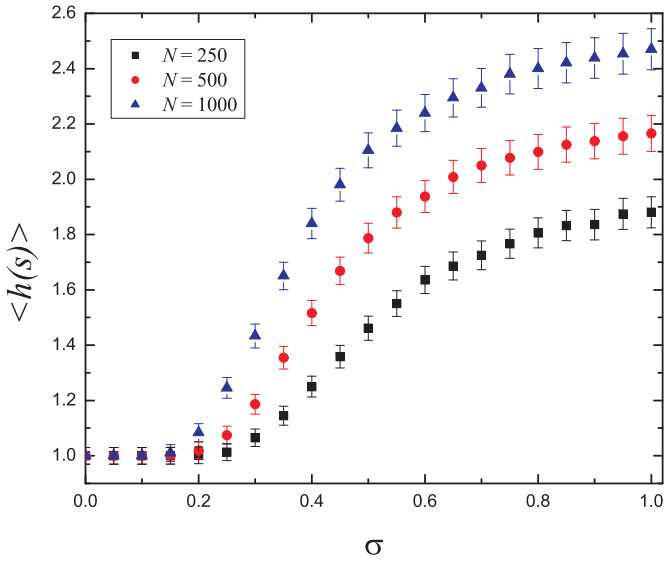


Figure 5.4. Dependence of the average height, $\langle h \rangle$ on the control parameter σ for the Gaussian truncated distribution. The simulations were done for ensemble of 10^4 polymers.

The power-law distribution

The truncated Gaussian distribution considered above is good for testing the key features of the RIM of RNA–like chains, however itself this distribution is rather artificial. It is much more natural to consider scale–free distributions of distances between neighboring monomers. In this case the intervals d_i have the following probability density function:

$$f(d, \gamma) = \frac{C}{1 + d^{\gamma}} \tag{5.10}$$

with $\gamma > 0$ and $d_{\min} < d < d_{\max}$. The normalization constant $C \equiv C_\gamma(d_{\max}, d_{\min})$ is

$$\begin{aligned} C(d_{\max}, d_{\min}) &= [A_\gamma(d_{\max}) - A_\gamma(d_{\min})]^{-1}; \\ A_\gamma(x) &= {}_2F_1\left(1, \gamma^{-1}, 1+\gamma^{-1}, -x^\gamma\right) x \end{aligned} \tag{5.11}$$

where ${}_2F_1(...)$ is the hypergeometric function. In what follows we take the following numerical values: $d_{\min} = 1$; $d_{\max} = 20$. In contrast to the truncated Gaussian distribution, in the truncated scale-free distribution the probability of very long distances between neighboring monomers is not exponentially small.

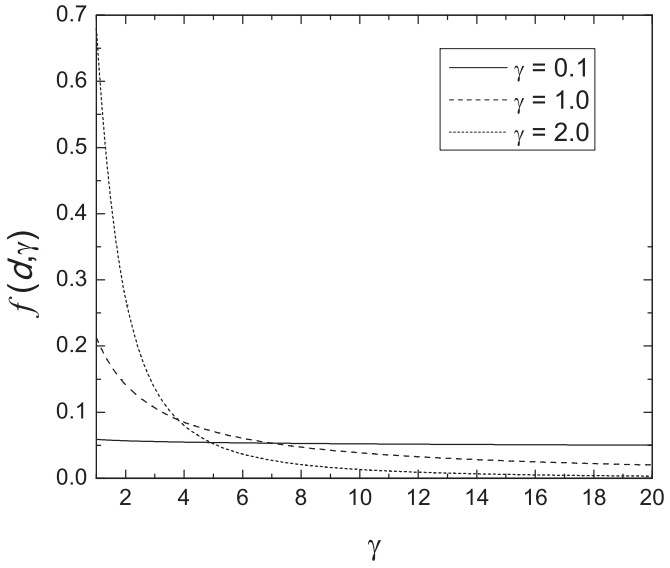


Figure 5.5. Power-law distribution function. $f(d, \gamma)$ of distances between nearest-neighboring monomers for $\gamma = 0.1; 1.0; 2.0$.

The presence of "heavy tails" in the distribution affects the topology of the ground state of the RNA RIM in a nontrivial way. Indeed, when γ in (5.10) is increasing from zero, the "nesting degree", h, behaves non-monotonically: at small $\gamma > 0$ it increases up to some maximal value (at $\gamma = 1$) and then decreases, tending to 1 (for $\gamma \to \infty$), see Fig. 5.6.

It is worth to note that the presence of "heavy tails" in the distribution releases the creation of nested configurations in an optimal pairing. For large values of γ the height diagram decreases which, as in the case of Gaussian distribution, corresponds to weakly random (practically equidistant) RNAs with sequential optimal pairing.

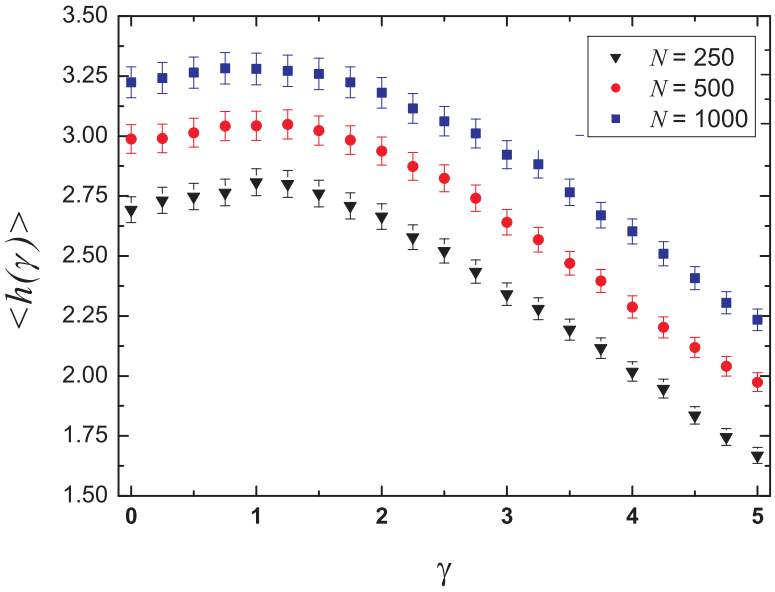


Figure 5.6. Dependence of the height, $\langle h \rangle$ on the control parameter γ for the truncated power-law distribution. The simulations were performed for ensemble of 10^4 random samples.

5.3.2 Analytical estimates

The nesting in an optimal configuration of RIM is affected by two complementary factors. On one hand, the nesting becomes favorable under some condition (explicitly written below) on lengths of three sequential intervals $d_{i-1},\ d_i,\ d_{i+1}$. On the other hand, the creation of a covering arc between two distant monomers i and j could be favorable if below this arc all pairs of *neighboring* monomers have formed bonds. Creation of a covering arc involves a global reorganization of linked pairs in a RIM. To the contrary, the nesting discussed above, is the local property of the RIM due to the special arrangement of sequential triples.

Let us focus on the nesting in an optimal configuration dealing with *local* properties of a RIM. The nested configuration of two arcs is favorable with respect to the sequential pairing, if the following inequality for the weight values holds:

$$\omega_{i-1,i+2} + \omega_{i,i+1} < \omega_{i-1,i} + \omega_{i+1,i+2} \tag{5.12}$$

Taking into account that $\omega_{i,j} = u \ln |x_i - x_j|$, we can easily transform (5.12) into

the condition on three sequential intervals $d_{i-1},\ d_i,\ d_{i+1}$:

$$\begin{cases} d_{i-1} > d_i \\ d_{i+1} > \dfrac{d_i(d_{i-1}+d_i)}{d_{i-1}-d_i} \end{cases} \tag{5.13}$$

or in a more symmetric form

$$d_i < \frac{d_{i-1}+d_{i+1}}{2}\left(\sqrt{1+\frac{4d_{i-1}d_{i+1}}{(d_{i-1}+d_{i+1})^2}}-1\right) \tag{5.14}$$

Having the distribution $f(d)$ in $[d_{\min}, d_{\max}]$, we can compute the probability P that inequalities (5.14) hold. Since the intervals $d_{i-1},\ d_i,\ d_{i+1}$ are distributed independently, the desired probability P is determines by the integral

$$P = \int_{d_{\min}}^{d_{\max}} f(x)\,dx \int_{d_{\min}}^{d_{\max}} f(y)\,dy \int_{d_{\min}}^{\frac{x+y}{2}\left(\sqrt{1+\frac{4xy}{(x+y)^2}}-1\right)} f(z)\,dz, \tag{5.15}$$

where integration over x corresponds to d_{i-1}, over y, to d_{i+1}, and over z, to d_i.

The equation (5.15) describes appearance of 1st level nesting ($h = 2$). Moreover, it is present as a multiplier in the probability of the 2nd level nesting ($h = 3$). So, we can expect that numerical curves for $h(\sigma)$ or $h(\gamma)$ have the same features as the function.

Gaussian truncated distribution

Substituting the truncated Gaussian distribution $f(d, \sigma)$ (5.9) with the parameters ($\mu = 2;\ d_{\min} = 1;\ d_{\max} = 3$) we get the function P plotted in Fig. 5.7. Note that $P(\sigma)$ repeats the profile of $\langle h(\sigma)\rangle$ displayed in Fig. 5.4 for the average height of the arc diagram. However our analytic approach does not take into account the slight dependence of the transition point on the polymer length since this effect has "global" property and is beyond the precision of our method. It should be also emphasized that the appearance of the 2nd level nesting (i.e. of the diagrams with the heights $h > 2$) deals exclusively with global reorganization of pairing in the RIM. Indeed, in order to have the 2nd level nesting, the condition (5.12) should be valid for the intervals $d_{i-2},\ d^{(1)},\ d_{i+2}$, where we substitute for the middle interval $d^{(1)}$ the combination of neighboring triples, $d_{i-1}+d_i+d_{i+1}$,

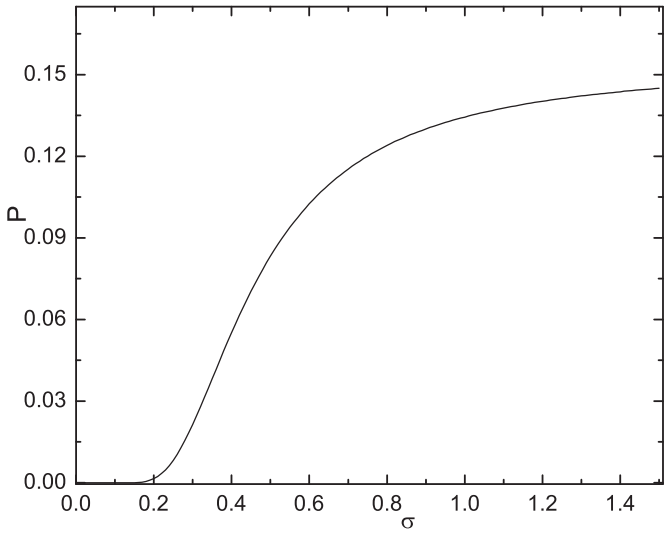


Figure 5.7. Analytical estimate for the truncated Gaussian distribution. Dependence of the probability P (5.15) on the control parameter σ.

which itself is nested. The minimal value for the middle interval $d^{(1)}$, as it follows from (5.12), is $d^{(1)} = 2(\sqrt{2}+1)d_{\min} + d_{\min}$. For the parameters of our distribution, we can conclude, that $d^{(1)} > d_{\max}$, what contradicts with the definition of the model. It means that all the configurations with the $h > 2$ have at least one long "global" arc.

Power-law truncated distribution

The same analysis can be performed for the RIM with the power-law distribution $f(d,\gamma)$ (5.10). We see that the function $P(\gamma)$ (Fig. 5.8) has the maximum at the point $\gamma = 1$. At $\gamma \gg 1$ the probability P tends to zero. Contrary to the truncated Gaussian distribution, the 2nd level nesting is allowed since $d^{(1)} < d_{\max}$, however the 3rd level nesting is forbidden, because $d^{(2)} = 2(\sqrt{2}+1)d^{(1)} + d^{(1)} > d_{\max}$. So, in the configurations with $h > 3$ the nesting is again due to "global" factors.

We have shown that for truncated Gaussian distribution $f(d,\sigma)$ of intervals, the height diagram exhibits a topological transition in pairing of monomers from sequential to essentially nested one. The parameter which controls this behavior is the dispersion, σ, of the distribution $f(d,\sigma)$.

In contrast to the truncated Gaussian distribution, for the truncated scale-free distribution $f(d,\gamma)$ the probability of very long distances between neighboring monomers is not exponentially small. The presence of such "heavy tails", or, in other words, of the "intermittent behavior" (i.e. very long tails mixed with

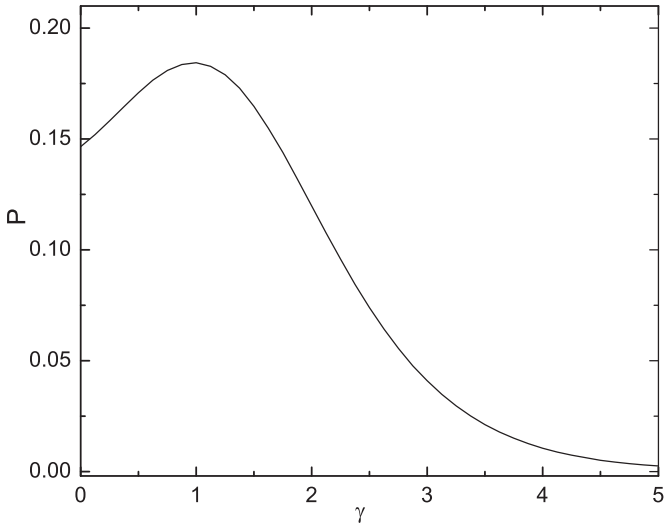


Figure 5.8. Analytical estimate for the truncated power-law distribution. Dependence of the probability P (5.15) on the control parameter γ.

very short ones) non-trivially affects the topology of the ground state of the RNA Random Interval Model.

The important result deserving attention, concerns the possibility to pass from the *nonlocal* recursion relation for the ground state free energy (2.22) to the local recursion relation (7.4) if and only if the interaction energy between paired monomers, $\varepsilon_{i,j}$, is a concave function of distance. So, for any potential (even random) of concave form, the equation (4.20) can be essentially simplified resulting in shortening the computational time if these equations are implemented for numeric analysis of secondary structures of polymer chain with RNA-type architecture.

The final remark concerns the possible interplay between optimization problems and some particular results of Random Matrix Theory (RMT) for RNA folding, addressed in [62, 128]. Let us recall that our basic result relies on the theorem which proves that optimal pairings on the line with the concave transport function are non-intersecting (i.e. planar). Being formulated in RMT terms, this means that optimization leads to the extraction of a special subclass of planar diagrams in the large-N random matrix ensemble, namely, the so-called *rainbow diagrams* (see, for example, [129]). To this end it would be interesting to formulate our Random Interval Model as a matrix model for finite N in order to check how the optimization algorithms allow extract planar diagrams of special topology in matrix models.

Chapter 6

Statistical analysis of networks: review of methods

Genes and gene products interact on several levels. At the genomic level, transcription factors can activate or inhibit the transcription of genes to give mRNAs. Since these transcription factors are themselves products of genes, the ultimate effect is that genes regulate each other's expression as part of gene regulatory networks. Similarly, proteins can participate in diverse post-translational interactions that lead to modified protein functions or to formation of protein complexes that have new roles; the totality of these processes is called a protein-protein interaction network. The biochemical reactions in cellular metabolism can likewise be integrated into a metabolic network whose fluxes are regulated by enzymes catalyzing the reactions. In many cases these different levels of interaction are integrated for example, when the presence of an external signal triggers a cascade of interactions that involves both biochemical reactions and transactional regulation.

A system of elements that interact or regulate each other can be represented by a mathematical object called a graph. At the simplest level, the systems elements are reduced to graph nodes (also called vertices) and their interactions are reduced to edges connecting pairs of nodes (Fig. 6.1). Edges can be either directed, specifying a source (starting point) and a target (endpoint), or non-directed. Directed edges are suitable for representing the flow of material from a substrate to a product in a reaction or the flow of information from a transcription factor to the gene whose transcription it regulates. Non-directed edges are

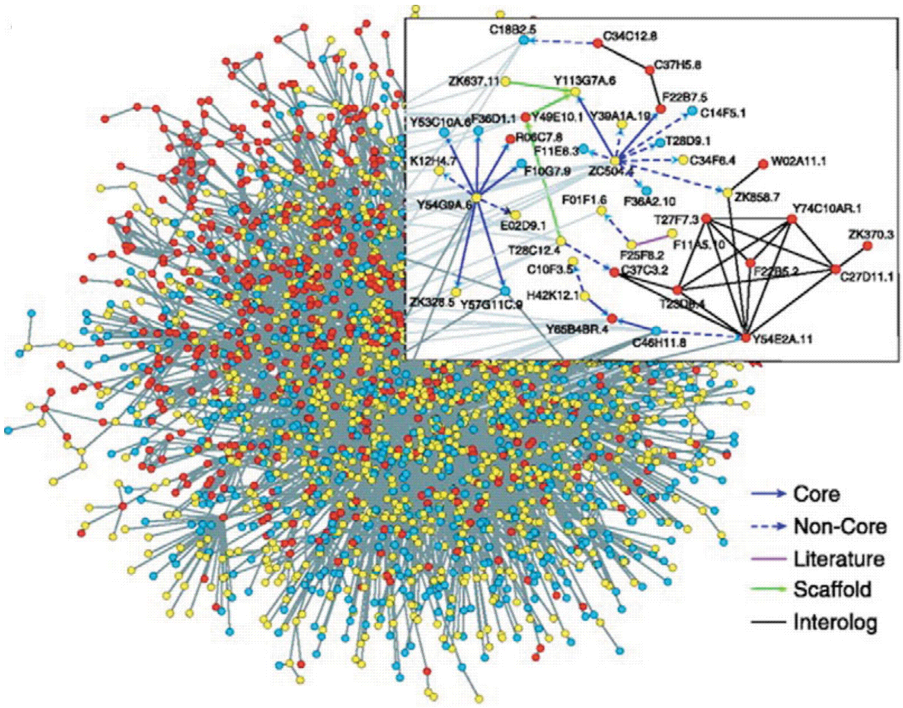


Figure 6.1. ***C.elegans*** **protein interaction network.** The nodes are colored according to their phylogenic class: ancient, red; multicellular, yellow; and worm, blue. The inset highlights a small part of the network [130]

used to represent mutual interactions, such as protein-protein binding. Graphs can be augmented by assigning various attributes to the nodes and edges; multipartite graphs allow representation of different classes of node, and edges can be characterized by signs (positive for activation, negative for inhibition), confidence levels, strengths, or reaction speeds. In this chapter it is shown how graph representation and analysis can be used to gain biological insights through an understanding of the structure of cellular interaction networks.

6.1 Structural properties

Statistical structural properties, considered below, are traditionally used to describe topologies of the graphs corresponding to real networks and a large amount of literature on these properties is available. Reviews of Albert and Barabasi [131], Newman [132] and Dorogovtsev and Mendes [133] present these properties and give their interpretations depending on the real systems studied. A completely mathematical treatment of such topics can be found, for example, in the monograph by Distel [134] or in Phd thesis by P. Kaluzo [135]. The definitions, which

are used by us to characterize the networks that we study, are based on these references.

Degree distribution

The degree d of a node in a network is the number of connections to other nodes. If the connections of a network are directed, it is possible to define separately the in-degree d_{in} and the out-degree d_{out} of each node, as the number of input and output connections that come to or go from the node, respectively, $d = d_{in} + d_{out}$. Usually, nodes of a network have different degrees. By counting the number of nodes with the same degree d in a graph, one obtains its degree distribution $P(d)$. In the case of directed networks, there are the in-degree, $P_{in}(d)$, and the out-degree, $P_{out}(d)$, distributions. Additionally, we can construct joint distributions $P(d_{in}; d_{out})$ which indicate the numbers of nodes with d_{in} input and d_{out} output connections in a network. For example, for random Erdos and Renyi networks these distributions are binomial. In regular networks, degrees of all nodes are equal and thus the distributions are singular.

The degree distributions of numerous networks, such as the Internet, human collaboration networks and metabolic networks, follow a well-defined functional form $P(d) = Ad^{-\gamma}$ called a power law. Here, A is a constant that ensures that the $P(d)$ values add up to 1, and the degree exponent γ is usually in the range $2 < \gamma < 3$ [131]. This function indicates that there is a high diversity of node degrees and no typical node in the network that could be used to characterize the rest of the nodes. The absence of a typical degree (or typical scale) is why these networks are described as "scale-free" (Fig. 6.2).

Clustering coefficient

This local property of networks is a measure of the density of interactions among a set of nodes. In an undirected network, a node i with k neighbors has clustering coefficient C_i defined as the ratio between the actual number of connections E_i among its k neighbors and their maximum possible number of connections $k(k-1)/2$,

$$C_i = \frac{2E_i}{k(k-1)} \tag{6.1}$$

For a network, the clustering coefficient C is the average of the clustering coefficients C_i of its nodes. In other words, the clustering coefficient quantifies how close the local neighborhood of a node is to being part of a *clique*, a region of the graph (a subgraph) where every node is connected to every other node. Various

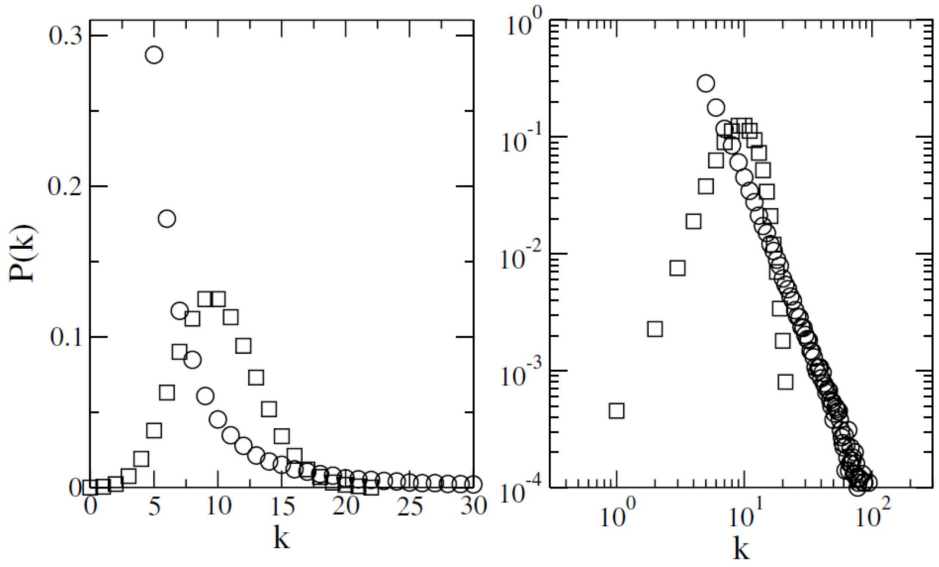


Figure 6.2. Scale-free and random networks. Comparison between the degree distribution of scale-free networks (circles) and random graphs (squares) having the same number of nodes and edges: the same two distributions are plotted both on a linear (left) and logarithmic (right) scale. The bell-shaped degree distribution of random graphs peaks at the average degree, by contrast, the degree distribution of the scale-free network follows the power law $P(k) = Ak^{-3}$ without any characteristic size [136].

networks, including protein interaction and metabolic networks [137, 138], display a high average clustering coefficient, which indicates a high level of redundancy and cohesiveness. Averaging the clustering coefficients of nodes that have the same degree d gives the function $C(d)$, which characterizes the diversity of cohesiveness of local neighborhoods. Several measurements indicate a decreasing $C(d)$ in metabolic networks [139] and protein interaction networks [138], following the relationship $C(d) = B/d^{\beta}$ (where B is a constant and β is between 1 and 2). This suggests that low-degree nodes tend to belong to highly cohesive neighborhoods whereas higher-degree nodes tend to have neighbors that are less connected to each other.

Most of networks have directed connections and, therefore, one cannot use the definition (6.1) to determine this property. In structural investigations of directed networks, such as in the study of the WWW by Adamic and Huberman [140], networks are often made bidirectional by discarding information about their directions. Although we can use the definition (6.1), the networks loose one of their main characteristics, their directionality. Moreover, many different directed

networks will then be transformed to the same undirected one.

Path

A path in a network is the set of sequential connections that link two nodes. Its length is the number of connections that are in the path. The distance between two nodes is the length of the shortest path among all available ones. When two nodes cannot be connected, the distance between them is taken as infinite. In undirected graphs, the distance between two nodes is independent of the directions in which we move. If such a network is connected (without isolated nodes), we can reach any node of the network starting from any other one. In directed networks, the situation is similar. However, now we have to consider the directions of the connections. A path between two nodes must connect them following the directions of its links. In consequence, two nodes can be connected only in one direction, or the distances between them can be different, depending on the node from which we start to move.

If edges are characterized by the speed or efficiency of information propagation along them, the concept can be extended to signify, for example, the path with shortest delay [141]. In most networks observed, there is a relatively short path between any two nodes, and its length is in the order of the logarithm of the network size [131, 132]. This small world property appears to characterize most complex networks, including metabolic and protein interaction networks. If a path connects each pair of nodes, the graph is said to be connected; if this is not the case, one can find connected components, graph regions (subgraphs) that are connected.

The connectivity structure of directed graphs presents special features, because the path between two nodes i and j can be different when going from i to j or vice versa. Directed graphs can have one or several strongly connected components, subgraphs whose nodes are connected in both directions; *in-components*, which are connected to the nodes in the strongly connected component but not vice versa; and *out-components*, which can be reached from the strongly connected component but not vice versa. It is important to note that this topological classification reflects functional separation in signal transduction and metabolic networks. For example, the regulatory architecture of a mammalian cell [142] has ligand-receptor binding as the in-component, a central signaling network as the strongly connected component and the transcription of target genes and phe-

notype changes as part of the out-component.

6.2 Motif distributions

The statistical properties, which we have described, give general characteristics of networks. Analyzing real networks, it has been shown [143] that they can have similar degree distributions, short path lengths between the nodes and high clustering coefficients, and nonetheless be very different with respect to their local structures. This suggests that such statistical properties are not enough to characterize the network architecture. Milo et. al. [144, 145, 146] have proposed to investigate local structures by considering the relative frequency of appearance of small subgraphs, or *motifs*, as compared to the randomized versions of the same networks.

In the case of directed subgraphs of size three, there are 13 different possible patterns of connections or motifs, which are all shown in Fig. 6.3. Although some of them have the same structural properties, such as the clustering coefficient, they are very different from the point of view of dynamics. For example, motifs 7 and 8, representing the feed-forward and the feedback loops, have the same clustering coefficient $C = 0.5$. However, only dynamical feedback loops can play a role in the robust adaptation of signal transduction networks to the variation of biochemical parameters [147]. Thus, analyzing the appearance of these motifs in a network allows us to obtain a more detailed structural characterization. Because we want to study the local structure of relatively small networks, we will consider only the motifs with three nodes.

Each of these motifs has a relative appearance Z_i in a network G, telling whether it is expressed more ($Z_i > 0$) or ($Z_i < 0$) less frequently than in the randomized versions of the same network G. After normalization, the relative appearances Z_i form a vector Z , the normalized Z score, which gives us the motif distribution of a given network. The normalized Z score is determined as following. The statistical significance of a motif i in a network G is given by its z_i score:

$$z_i = \frac{N_i - \langle N_i \rangle}{\sigma_k}; \quad i = [1...m] \tag{6.2}$$

is calculated, where N_i is the amount of i-th subgraphs in the initial network; and $\langle N_i \rangle$ and σ_i correspondingly the mean and the standard deviation of subgraphs

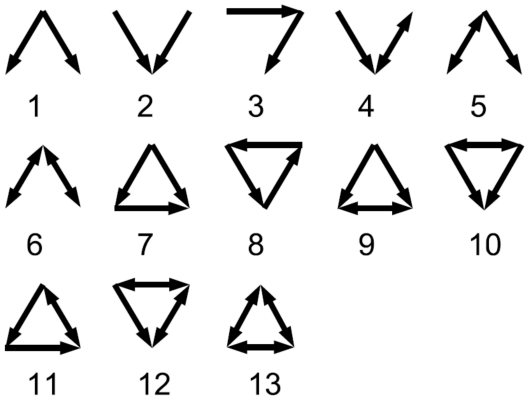


Figure 6.3. All possible three-node motifs. Nodes are the vertices where the links start and to which they arrive.

of given type in the randomized networks; m is the total number of considered subgraphs. Such randomized networks have the same degree distributions and the degree sequence (i. e. the number of nodes with a specific input and output degrees) as the analyzed network G, but they represent new random patterns of connections. The distribution of motifs in the network under consideration is characterized by the significance profile which is a normalized vector $Z = \{Z_1, ..., Z_m\}$ of statistical significance of all subgraphs of given size. The components of the vector Z are

$$Z_i = \frac{z_i}{\sqrt{\sum_{i=1}^{m} z_i^2}}; \quad i = [1...m] \tag{6.3}$$

6.3 Interpretation of network properties

The architectural features of molecular interaction networks are shared by other complex systems ranging from technological to social networks. While this universality is intriguing and allows us to apply graph theory to biological networks, we need to focus on the interpretation of graph properties in light of the functional and evolutionary constraints of these networks.

Hubs

In a scale-free network, small-degree nodes are the most abundant, but the frequency of high-degree nodes decreases relatively slowly. Thus, nodes that have degrees much higher than average, so-called hubs, exist. Because of the

heterogeneity of scale-free networks, random node disruptions do not lead to a major loss of connectivity, but the loss of the hubs causes the breakdown of the network into isolated clusters [131]. The validity of these general conclusions for cellular networks can be verified by correlating the severity of a gene knockout with the number of interactions the gene products participate in. Indeed, as much as 73% of the *S. cerevisiae* genes are non-essential, i.e. their knockout has no phenotypic effects [148]. This confirms the cellular networks robustness in the face of random disruptions. The likelihood that a gene is essential (lethal) or toxicity modulating (toxin sensitive) correlates with the number of interactions its protein product has [149, 150]. This indicates the cell is vulnerable to the loss of highly interactive hubs. Among the most well-known examples of a hub protein is the tumor suppressor protein p53, which has an abundance of incoming edges, interactions regulating its conformational state (and thus its activity) and its rate of proteolytic degradation, and numerous outgoing edges in the genes it activates.The protein p53 is inactivated by mutation in 50% of human tumors, which is in agreement with the vulnerability of cellular networks to their most connected hubs [151].

Modularity

Cellular networks have long been thought to be modular, composed of functionally separable subnetworks corresponding to specific biological functions [152]. Since genome-wide interaction networks are highly connected, modules should not be understood as disconnected components but rather as components that have dense intra-component connectivity but sparse inter-component connectivity. Several methods have been proposed to identify functional modules on the basis of the physical location or function of network components [153] or the topology of the interaction network [154, 155, 156]. The challenge is that modularity does not always mean clear-cut subnetworks linked in well-defined ways, but there is a high degree of overlap and crosstalk between modules [157]. As Ravasz et al. [139] recently argued, a heterogeneous degree distribution, inverse correlation between degree and clustering coefficient (as seen in metabolic and protein interaction networks) and modularity taken together suggest hierarchical modularity, in which modules are made up of smaller and more cohesive modules, which themselves are made up of smaller and more cohesive modules, etc.

Motifs and cliques

Growing evidence suggests that cellular networks contain conserved interaction motifs, small subgraphs that have well-defined topology. Interaction motifs such as auto-regulation and feed-forward loops have a higher abundance in transcriptional regulatory networks than expected from randomly connected graphs with the same degree distribution [158, 146]. Protein interaction motifs such as short cycles and small completely connected subgraphs are both abundant [154] and evolutionarily conserved [159], partly because of their enrichment in protein complexes. Triangles of scaffolding protein interactions are also abundant in signal transduction networks, which also contain a significant number of feedback loops, both positive and negative [142]. Yeger-Lotem et al. [160] have identified frequent composite transcription/protein interaction motifs, such as interacting transcription factors co-regulating a gene or interacting proteins being co-regulated by the same transcription factor. As Zhang et al. [161] have pointed out, the abundant motifs of integrated mRNA/protein networks are often signatures of higher-order network structures that correspond to biological phenomena. Conant and Wagner [162] found that the abundant transcription factor motifs of *E. coli* and *S. cerevisiae* do not show common ancestry but are a result of repeated convergent evolution. These findings, as well as studies of the dynamical repertoire of interaction motifs, suggest that these common motifs represent elements of optimal circuit design [142, 163, 164].

Path

Any response to a perturbation requires that information about the perturbation spreads within the network. Thus the short path lengths of metabolic, protein interaction and signal transduction networks (their small world property) is a very important feature that ensures fast and efficient reaction to perturbations. Another very important global property related to paths is path redundancy, or the availability of multiple paths between a pair of nodes [165]. Either in the case of multiple flows from input to output, or contingencies in the case of perturbations in the preferred pathway, path redundancy enables the robust functioning of cellular networks by relying less on individual pathways and mediators. The frequency of node participation in paths connecting other components can be quantified by their betweenness centrality, first defined in the context of social sciences [166]. Node betweenness, adapted to the special conditions of signal transduction networks, can serve as an alternative measure for identifying

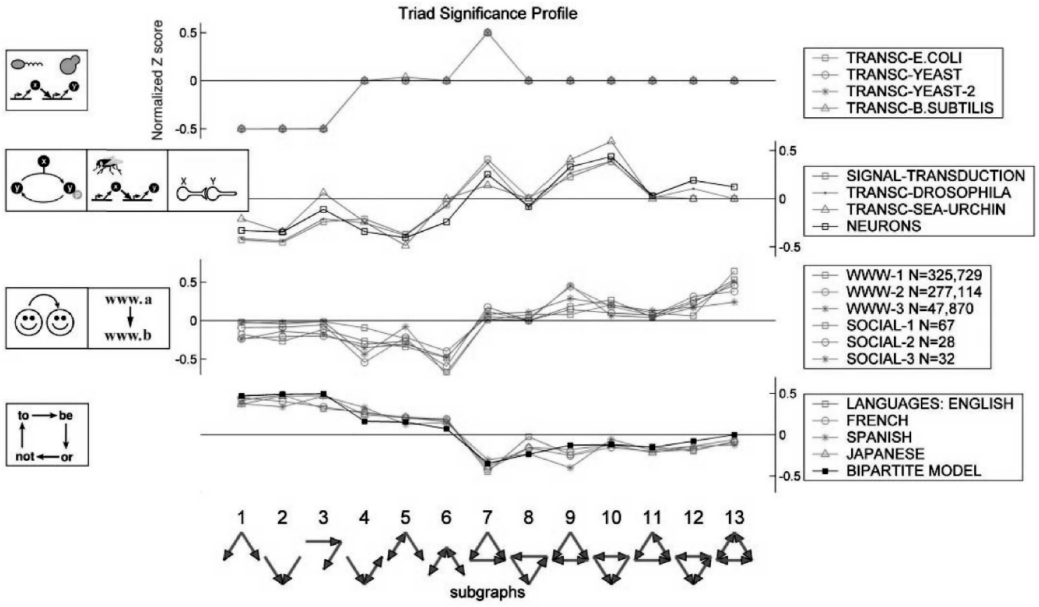


Figure 6.4. Three node motif profiles of networks from various disciplines. Networks with similar characteristic distributions are grouped into superfamilies [144].

important network hubs.

6.4 Network superfamilies

Analyzing three-node motif distributions of many networks of different fields and origins, Milo et. al. [144] have found that there are remarkable similarities between such distributions and it is possible to identify four main superfamilies. Fig. 6.4 shows different motif distributions for each superfamily. In the first superfamily, several sensory transcription networks that control gene expression in bacteria and yeast in response to external stimuli are found. In Fig. 6.4 motif distributions of three microorganisms, the bacteria *Escherichia coli* and *Bacillus subtilis* and the yeast *Saccharomyces cervisiae*, are given. Such motif distributions show increases in the appearance of motif 7, the "feed-forward loop". This loop is related to signal-processing tasks, such as persistence detection, pulse generation, and acceleration of transcription responses. The motif 3, the 3-chain, is underrepresented, because of the shallow architecture of these networks which have only few long cascades. These networks are "sensory networks" which need to respond within minutes to transient signals such as stresses and nutrients. The minimal time required for a response (for the first proteins to be expressed) is in-

deed on the order of minutes. If the information needs to pass additional steps (a regulator protein needs to be expressed and cross its activation threshold to turn on a gene), then the response time becomes much longer. Thus, these networks can be understood as "rate-limited networks", where the desired response times are often as short as the response times of the network components [144]. In the rate-limited network superfamily, long cascades and feedback loops are rare.

The second superfamily includes three kinds of information-processing networks in biological macroorganisms: signal-transduction in mammalian cells, development transcription networks that guide the development in fruits and sea-urchin, and synaptic connections between neurons in *Caenorhadbitis elegans*. In these networks, motifs 7, 9 and 10 are enhanced, whereas by motifs 1, 2, 4 and 5 are suppressed. In contrast to the networks of the previous superfamily, these networks include two-node feedbacks that regulate or are regulated by a third node (motif 10 and 9) and are less biased against cascades (motif 3). The common feature to this superfamily of information-processing networks is that the response time of each step in the network is usually much shorter than the response time required for the biological function of the network. For example, protein signal-transduction networks often need to respond within an hour or longer, but each interaction can take minutes or less. Cascade steps in developmental networks can have response times of tens of minutes, but the processes they control are much slower, on the order of animal cell-division times that can take several hours. It has been therefore suggested [144] that this superfamily characterizes biological information-processing networks which are not rate-limited.

Several WWW networks of hyperlinks between Web pages and some social networks show similar three node motif distributions and they are combined in the third superfamily. In this superfamily, motifs 9, 10, 12 and 13, that represent triangle transitive interactions are overrepresented, whereas motifs 4, 5 and 6 have low frequencies. Finally, the last superfamily is formed by networks of word-adjacency. These networks are constructed by taking as nodes the words of a text. A directed connection between two nodes exists if these two words appear consecutively in the text. In Fig. 6.3 motif distributions of several languages (English, Frensh, Spanish, and Japanese) taken from several texts are shown. The main characteristic of these motif profiles is that motifs from 7 to 13 are underrepresented. According to [144], the lower significances of the motifs with

triangles can be explained by the structures of the languages, where words have different categories and a word from a certain category tends to be followed by one from a different category (for example, a preposition is usually followed by nouns or articles). The bipartite networks (also shown in Fig. 6.3), where there are two classes of nodes and the connections are allowed only between nodes of different class, yield a similar motif distribution.

Chapter 7

Analysis of functional networks in *C.elegances*

Connectivity networks have become increasingly useful for biology because of the expanding availability of data on the physical and functional links between individual genes and proteins [167]. This connectivity data enables expanding our knowledge beyond the experimentally validated results. New functional interactions can be predicted and tested by means of analysis and theoretical expectations.

7.1 Materials and Methods

7.1.1 Data preparation

Microarray data are adopted from [168] where two parental *C.elegans* strains N2 (Bristol) and CB4856 (Hawaii) and recombinant inbred lines were used to measure gene expression in a 16 ^{0}C and 24 ^{0}C environment, representing the two genetic and ecological extremes of *C. elegans* [169]. Their genetic distance amounts to about one polymorphism per 873 base pairs [170]. Both strains and their offspring have contrasting gene expression levels [168, 171, 172, 173, 174], phenotypes [175, 176, 177, 178] and differ strikingly in their response to a temperature change [179, 180]. In [168, 179, 181, 182] the strains were exposed to 16 ^{0}C and 24 ^{0}C, temperatures which are known to strongly affect gene expression [168, 182] and phenotypic characteristics such as body size, lifespan and reproduction [180, 179, 183]. Gene expression patterns were measured by hy-

bridization to oligonucleotide micro-arrays. All micro-array data was retrieved from NCBI's Gene Expression Omnibus (GEO [184]) under GSE5395. By means of the Mev4 application [185] we have performed clustering of the gene expression profiles represented by measurements of absolute mRNA values in a set of conditions. K-means clustering algorithm based on estimation of Euclidian distances between different probe profiles has been applied to produce a predefined number of 50 gene clusters. The resulting clusters were used as "co-expression" clusters in our analysis. For eQTL-hotspot gene selection we used the eQTL data from [171], which we retrieved from WormQTL[186]. This experiment was done at three age groups. In each age group genes with a shared regulatory locus were selected by taking all the genes having an eQTL with a $log_{10}(p)$-value above 3 at the same locus (see Tab. 7.1). String software [187, 188] has been used to reconstruct graphical networks from the sets of *C.elegans* genes. WormBase database 220 [189] has been used for retrieval of ID, gene name, associated functional annotations and ontological categories. WormNet [190] has been used as a source of information on pair-wise interactions between genes. We also used WormNet to retrieve data on separated genetic interactions of *C.elegans* genes and co-expression links in *C.elegans*.

EQTL-hotspot	Chromosome	Left marker	Right marker	Number of genes
Juvenile worms				
1	I	4	6	261
2	V	98	100	183
Reproducing worms				
3	IV	61	63	131
4	V	95	100	194
Old worms				
5	II	37	40	144
6	IV	61	65	164
7	IV	68	68	92
8	V	95	100	215

Table 7.1. EQTL-hotspots for three different age groups.

7.1.2 Statistical analysis of network connectivity

To investigate the WormNet connectivity properties of the selected gene clusters and to establish potential regulators for these gene clusters we used the following

approach: denote by $d_{i,M}$ the shortest path along the network from a given vertex (regulatory gene) M to some other vertex (cluster gene) i. Consider the shortest path function (SPF) determined as follows:

$$k_M^{SPF} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{d_{i,M}} \tag{7.1}$$

where the summation is performed over all cluster genes (N is the number of the genes in the cluster, i.e. the cluster size). The connectivity of the gene cluster on the whole network and/or on its subnetwork is described by the SPF defined for all cluster genes

$$k_{cl}^{SPF} = \frac{2}{N(N-1)} \sum_{i,j=1}^{N} \frac{1}{d_{i,j}} \tag{7.2}$$

here N is the number of the genes in the cluster, $d_{i,j}$ is the shortest path between the nodes i and j. Thus defined, the SPF has a very transparent meaning, since it gives the averaged reciprocal paths between pair of cluster genes. If i and j are not linked on the network, the contribution to the SPF from this pair (i, j) equals to zero, while the maximal contribution is reached for directly linked pairs. So, the SPF can be used to characterize quantitatively the connectivity of a gene cluster in a given network. For comparison we also use another method for determining the cluster connectivity. The so-called "connectivity coefficient" is defined as a ratio between the number of *inner* links (which connect only cluster genes) to the number of *all* links which the cluster has in the network:

$$f_{cl} = \frac{N_{in}}{N_{all}} \tag{7.3}$$

We used WormNet and its parts as an unweighted undirected network. Note that the SPF analysis can be easily extended for weighted and directed networks as well by adding to (8.2)-(7.2) a coefficient for the strength of the interaction.

Another approach to characterize connectivity (and topology) of a cluster is to investigate its motifs distribution. The procedure described by (6.2)-(6.3) (see Section 6.2 for details) is used to characterize the statistical properties of the gene clusters.

7.1.3 Prediction of regulators for gene clusters

We apply the method of the SPF coefficients k_M^{SPF}, determined in (8.2) to search potential regulators of co-expressed gene clusters. Note, that potential regulators can belong to the cluster or lie outside of it. For the search of the potential cluster regulators it is important that the regulator is specific to it. Moreover, it should have many connections to the genes in the cluster but at the same time only a weak connection to other genes in the network. In other words, the potential regulator must have enough links to control the cluster not being herewith the hub of the network. We introduce the value which takes into account the connectivity of the regulatory gene in the whole network (G):

$$K_M^{SPF} = \frac{\sum_{i=1}^{N} \frac{1}{d_{i,M}}}{\sum_{i=1}^{G} \frac{1}{d_{i,M}}} \tag{7.4}$$

By definition, the K_M^{SPF} is the ratio of the sum of the shortest paths from regulator to the cluster genes, and the sum of the shortest path from regulator to all genes in the network. One sees, that K_M^{SPF} measures how specific a regulator is to the cluster. The function (8.2) can be considered as a first term in expansion of the function (7.4). We use this value for the search of potential regulators to the gene clusters in the network and its subparts. We compare this method with some other methods described below.

Namely, for any potential regulator we define the fraction of the cluster genes, which are connected to it:

$$f_M = \frac{N_M}{N} \tag{7.5}$$

Analogously to (7.4) we introduce the value, which takes into account the "interaction" of the potential regulator with outer part of the cluster genes:

$$F_M = \frac{N_M}{N_M(G)} \tag{7.6}$$

Here, $N_M(G)$ is the number of all links, which the potential regulator M has in the whole network G.

7.2 Results

7.2.1 Statistical properties of co-expression clusters

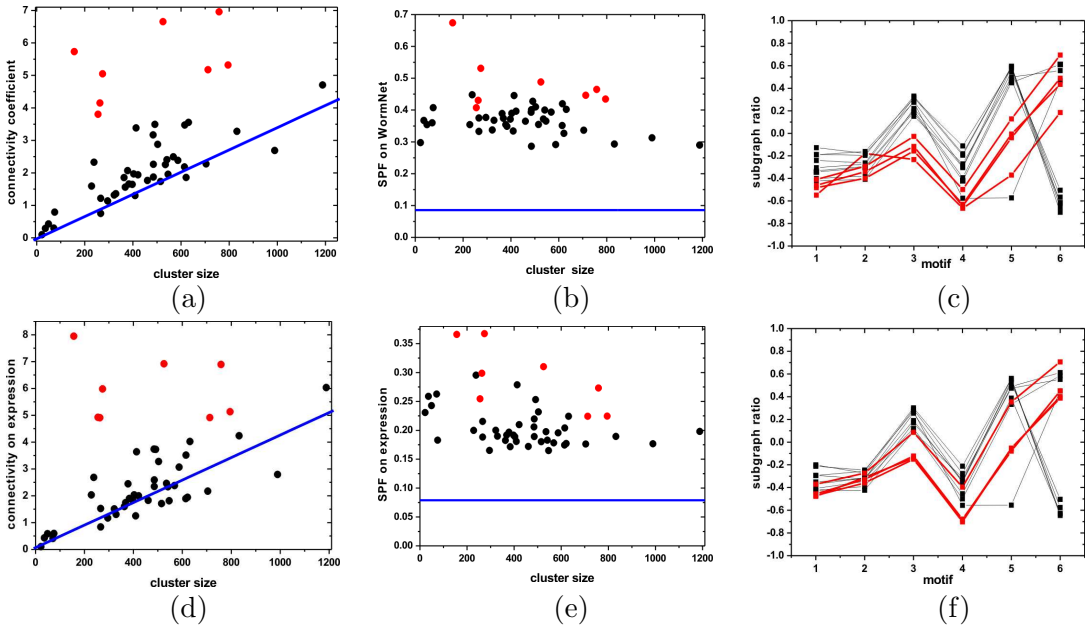


Figure 7.1. Statistical properties of co-expression clusters. The connectivity coefficient ((a) and (d)), the SPF coefficients ((b) and (e)) and motif distribution ((c) and (f)) in whole WormNet (up) and in its expression sub-part (below); the respective values for random set of genes are shown by the blue line.

Our first goal was to define the properties of the extracted gene clusters from the whole WormNet network. Fig. 7.1 presents the dependencies of the cluster's connectivity and the SPF on the size of a cluster (the number of genes included in the cluster), as well as the motif's distribution for some clusters. The connectivity coefficient is proportional to a cluster size. Generally, an expression cluster consists of genes that show a stronger connection in WormNet than any group of randomly selected genes — see Fig. 7.1(a). The SPF coefficients for all gene clusters lie much higher than the SPF coefficients for random clusters, as shown in Fig. 7.1(b) (here and below: a random cluster consists of randomly selected nodes in WormNet). The most connected clusters are mainly bound by co-expression links in WormNet (Fig. 7.1(d)). We have selected the clusters with very high connectivity coefficients. Such clusters are well-connected and are depicted by red points in the graphs. These clusters are also characterized by a motif's distribution where linear 4-nodes chains prevail (Fig. 7.1(c)). The gene clusters can be divided into two groups, according to their distribution of motifs.

The first group includes the gene clusters, in which the fully connected subgraphs dominate. All well-connected clusters belong to this group. The second group is formed by the clusters with a small number of fully connected motifs. It is worth noting that protein structure networks are also characterized by the same motif distribution [144].

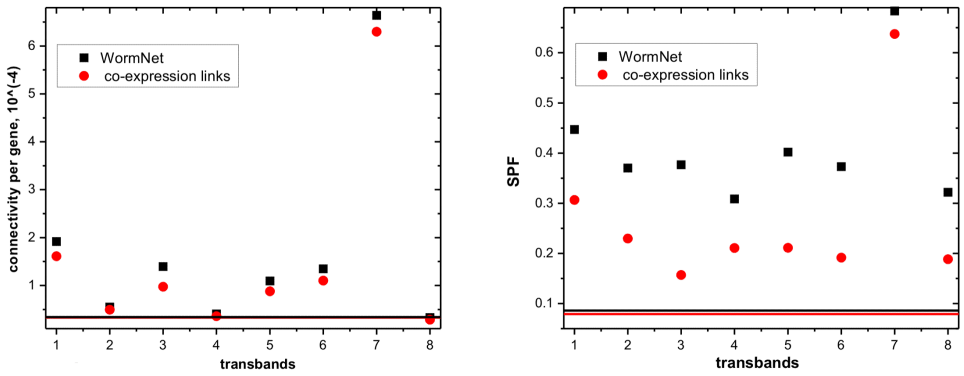
The number of co-expression and physical connections are much more abundant than genetic interactions in WormNet. This illustrates the necessity for additional experimental analysis, or *in-situ* prediction of potential regulatory modules for *C.elegans*. Our analysis shows that the motif's profiles found for expression sub-network of gene clusters resemble the motif profiles in WormNet itself (Fig. 7.1(c)). This means that in our clusters the co-expression links are more abundant than other links and provides proof that genes of well-connected clusters are mainly connected to each other by co-expression. However the SPF scores are much lower when only co-expression considered, while the connectivity coefficient remains approximately the same. Also, our analysis demonstrates that co-expression links are responsible for the existence of the cycles in well-connected clusters (Fig. 7.1(f)).

The analysis of trans-regulatory hotspots (or trans-bands) has also shown that these trans-bands have more connections than a random set of genes (Fig. 7.2). This is a strong indication that they have a shared biological function or indeed shared regulation. Fig. 7.2 demonstrates the advantage of the SPF method. In contrast to the connectivity coefficients, the SPF coefficients for trans-bands are more different from random and so offer a better prediction of a shared gene. The difference between WormNet and expression sub-network is also more clearly demonstrated by the SPF analysis.

7.2.2 Prediction of expression cluster regulators

We have used the most connected cluster, cluster 1, to find an optimal algorithm to predict potential regulators. Two different methods have been tested.

The first method is based on ranking the network nodes according to the fraction of cluster genes (FCG) they are directly connected to. This method requires a dense connectivity matrix to be efficient. The predicted regulators are expected to have a strong and specific involvement in modulation of expression of a particular gene cluster. However, the ranked regulator list needs manual



(a) (b)

Figure 7.2. Statistical properties of eQTL-hotspots. The connectivity coefficient per gene in an eQTL-hotspot group (a) and the SPF coefficients (b) in whole WormNet (black) and in its expression sub-part (red). The black and red lines are the respective functions for equally sized random set of genes.

filtering to eliminate function that unlikely affect gene expression directly. The second method is based on ranking nodes according to the shortest path function (SPF) which uses the shortest average distance to all cluster genes. As in the first method, the obtained gene regulator lists need filtering when applied to integrated (whole) network.

The different connectivity subnetworks of the cluster have different topologies, with the genetic interactions subnetwork as the least connected (Fig. 7.1). The power of the SPF method is that it compensates for the absence of knowledge about the regulation of many genes by generation of quantitatively validated predictions of the potential sharing of known regulators of few cluster genes by all the co-expressed gene clusters. Also, the SPF method does not require a matrix to be dense and can be applied to a subnetwork of genetic interactions. Both the FCG and SPF method were applied to the total WormNet and to the co-expression connectivity subnetwork with similar outcomes. The top of the predicted regulators is presented in Tab. 7.2.

Seq. IDs	Gene	Function
F57B9.6	*inf-1*	Transl.initiation/ RNA transport
T05G5.10	*iff-1*	Transl.initiation/ NMD
Y71G12B.8	Y71G12B.8	RNA helicase/ RNA transport
T10C6.14, T10C6.12, T10C6.11, F45F2.3, F45F2.4, F45F2.12, ZK131.4, ZK131.6, ZK131.8, ZK131.10, K06C4.10, K06C4.11, K06C4.4, K06C4.3, K06C4.12, ZK131.1, K06C4.2, F35H10.1, F17E9.12, F17E9.13, C50F4.7, K03A1.6, C50F4.5, F08G2.2, B0035.9, B0035.7, F07B7.9, F07B7.10, F07B7.4, F07B7.3, F07B7.11, F54E12.3, F54E12.5, F55G1.11, F55G1.10, F22B3.1, H02I12.7, T23D8.5, T23D8.6	38 His genes	Histones
C41D11.2	*eif-3.H*	Transl.initiation
F32E10.1	*nol-10*	Nucleolar protein, polyglut.binding
F54H12.6	*eef-1B.1*	Elongation factor
C01F6.5	*aly-1*	RNA export
M163.3	*his-24*	Histones
B0564.1	*tin-9.2*	Decay/ NMD
Y18D10A.17	*car-1*	Decay/decapping
F56D12.5	*vig-1*	RISC component/miRNA binding
F26D10.3	*hsp-1*	Splicing
R04A9.4	*ife-2*	Transl.initiation

Table 7.2. Top of the predictable regulators for test cluster 1 by the FCG and the SPF method.

Seq. IDs	Gene	Function
Y55D5A.5,B0334.8,Y116F11B.1	*daf-2*, *age1*, *daf-28*	Insulin/aging
F35H8.5	*exc-7*	mRNA processing
W10D5.1	*mef-2*	TF
C17D12.2	*unc-75*	Splicing
C47G2.2	*unc-130*	TF
F30F8.8	*taf-5*	Transl.initiation
R74.3	*xbp-1*	TF, histone modulation
F33A8.1	*cwc22*	Splicing
C41C4.4	*xre-1*	(RNA processing) decay/ processing
C37H5.8	*hsp-6*	Decay
C26D10.2	*hel-1 (helicase)*	DNA helicase
C07H6.5	*cgh-1 (decapping)*	Decay/ decapping
F02E9.4	*sin-3 (HDAC)*	Histone modulation
M163.3	*his-1*	Histone
212312 C25A1.10	*dao-5*	rRNA transcription/aging
ZC247.3	*lin-11*	TF
R107.8	*lin-12*	TF
C05D9.5	*ife-4*	Transl.initiation
R11E3.6	*eor-1*	TF
F43G9.11	*ces-1*	TF
ZK909.4	*ces-2*	TF

Table 7.3. Top of the predictable regulators for the test cluster 1 by the SPF method on genetic subnetwork.

The SPF method has also been applied to the gene regulatory network. As a result of this, several predicted regulators are well integrated both in the top cluster networks and in the network reconstructed from the genes with a longevity phenotype retrieved from WormBase. Among the most promising predicted regulators that could connect the top clusters with a longevity regulation are: *daf-2*, *iff-1*, *cgh-1*, *tin9.2*, *car-1*. They all are related to mRNA processing/translation/decay and are in a cross-talking relationship (Tab. 7.2). According to their position in the network, they may play a role of linkers between the top connected co-expression clusters related to ribosomal biogenesis, proteosome and central metabolic functions (Fig. 7.3). Note that the SPF method allows us to predict some regulators which we could not detect by the FCG method, and the power of this method had been enforced by its application to a gene regulatory connectivity subnetwork, as it is clearly demonstrated in Tab. 7.3. We could identify a number of TFs that may be considered for a role of transcription regulators of genes in the top co-expression cluster 1, such as: *taf-5* — transcription initiation factor TFIID subunit 5, *xbp-1* — heat-shock transcription factor, *sin-3* — histone deacetylase subunit and pre-mRNA-splicing factor *cwc-22*. This method also greatly increased the ranking position of *daf-2* and genes upstream *daf-2* (C25A1.10) or being directly affected by *daf-2* mutation (C05C8.3) which are immediate potential connections to a group of genes with longevity phenotype that have a strong overlap with our cluster 1 (Fig. 7.3). Our statistical analysis demonstrates that the most connected components of WormNet are included in test cluster 1, so the exclusion of the hub-regulators by the procedures (7.4) for the SPF method and (7.6) for the FCG method give about the same list of potential regulators.

Fig. 7.4 illustrates a typical position of the predicted potential regulators for the cluster 1. Nodes predicted by the FCG method (purple frame) are proximal to the cluster or even inside the cluster. The nodes predicted by the SPF method can be significantly distant from the many nodes in the cluster (*ces-1*, *eor-1*, orange frames on Fig. 7.4). Though the connections between the SPF-predicted node and the cluster may include several intermediate steps, the majority of these steps do contain the nodes that can translate signals at the level of mRNA pool regulation, potentially representing complexes of proteins with a joint regulatory performance. As one sees, the type of connectors utilized by different algorithms

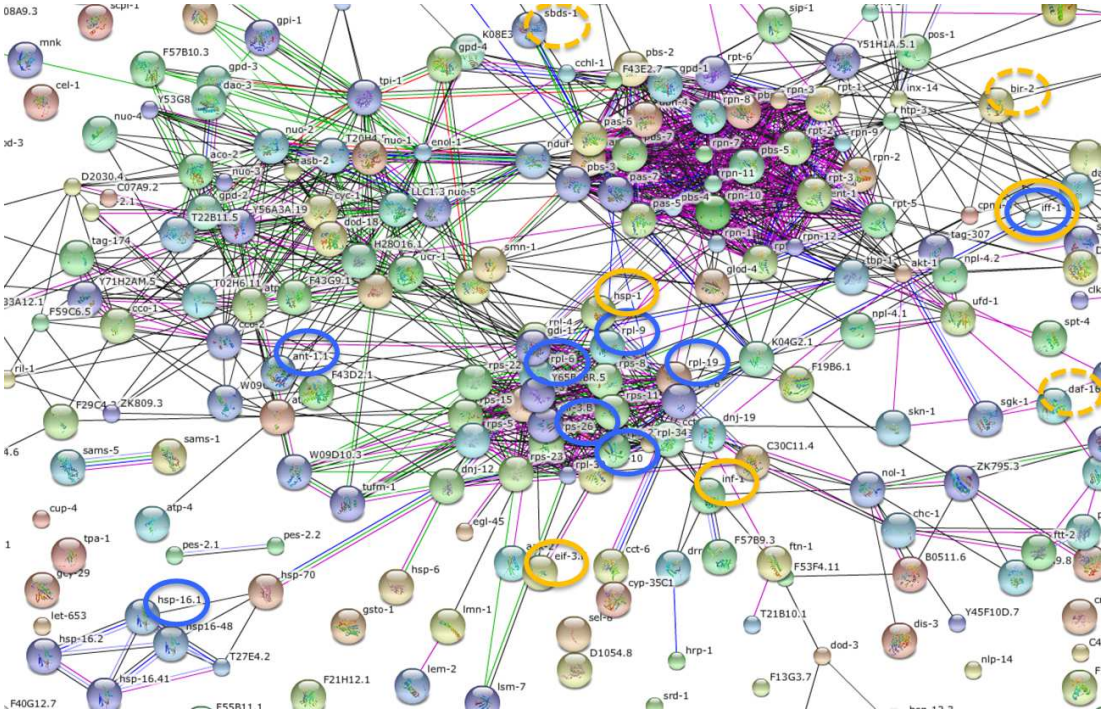


Figure 7.3. Network reconstructed from the *C.elegans* genes with an adult life span phenotype from WormBase 220. Three main distinguished clusters can be seen: in the center — ribosomal, top left —metabolic, top right — proteosome and exosome functions. Blue circles indicate the test cluster 1 genes. Orange— predicted regulators, dashed borders — functionally associated regulators discussed in the manuscript. (Not all aging-related functions related to the cluster 1 are shown on this figure).

differs essentially: experimental, regulatory connections are fundamental for the SPF and more dense, co-expression connections, for the FCG.

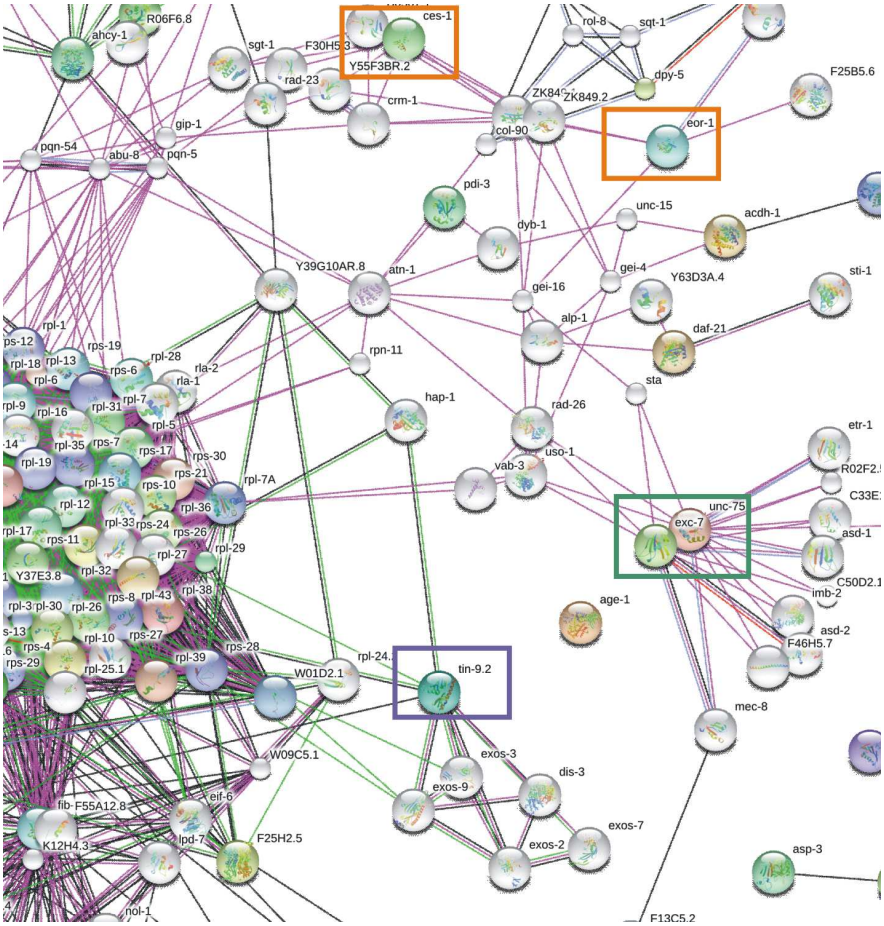


Figure 7.4. Connectivity between the predicted regulators and the cluster 1 in STRING Network browser. Evidence view for high confidence (0.700) connections. Pink connectors - Experimentally derived interactions (pink), co-expression (black), and co-localization on the genomes (green), and co-occurrences in the genomes (blue). Colored circles represent input genes, white circles — the most associated additional nodes (set number of 200) automatically added by a STRING software on a request to increase a connectivity between uploaded functions. Predicted potential regulators are shown in frames: orange — the SPF method, purple — the FCG method, green node excluded in hub-exclusion SPF method.

EQTL-hotspot	Chromosome	Gene	Function
Juvenile worms			
1	I	K09H9.2	
1	I	R12E2.2	Endocytosis/regulation of growth rate
1	I	*clec-53*	
1	I	W01B11.1	
1	I	*sep-1*	Cell division
1	I	*mis-12*	Cell division
1	I	Y54E10BR.3	TF/Zn ion binding
1	I	Y71F9B.6	
2	V	*fbxa-192*	Protein interaction
2	V	*str-92*	
2	V	T10C6.7	Protein interaction
2	V	Y59A8A.3	
Reproducing worms			
3	IV	Y55F3BL.2	Metal ion transport
3	IV	Y69A2AR.16	Metabolism/oxidoredutase
3	IV	Y69A2AR.21	Embrionic development
4	V	Y32B12A.5	
4	V	Y43F8B.13	
4	V	Y43F8B.14	
4	V	Y51A2B.4	Lipid metabolism
4	V	Y70C5B.1	
4	V	*srh-296*	Integral membrane component
Old worms			
5	II	*moe-3*	RNA binding/iRNA modification
5	II	Y17G7B.18	Positive regulation of growth rate/development
5	II	*cpt-1*	Acetyl-transferase/histone modification
5	II	*csp-1*	Caspase/apoptosis
5	II	*pqn-87*	Prion/protein modification
6	IV	F15E6.4	
6	IV	F28F9.3	
6	IV	T08B6.4	
6	IV	Y9C9A.1	Structural element of vitelline membrane
7	IV	C17H12.12	Protein binding
7	IV	C17H12.5	Tyrosine phosphatase
7	IV	C31H1.1	
7	IV	F36H12.5	
7	IV	F38A5.6	
7	IV	ZK354.3	
8	V	Y38H6C.15	
8	V	Y38H6C.18	
8	V	*tgt-2*	Queuine tRNA-ribosyltransferase activity modification
8	V	T26E4.10	Lipid storage
8	V	T26F2.2	
8	V	*sri-7*	Integral membrane component
8	V	*nhr-218*	TF,steroid hormon receptor
8	V	*str-151*	Integral membrane component
8	V	*nhr-269*	TF,steroid hormon receptor

Table 7.4. Top of the predictable regulators for eQTL-hotspot gene groups by the SPF method in WormNet.

The developed method has been applied to eQTL data on age-associated gene expression. The eight groups of genes sharing an age dependent regulatory locus (Tab. 7.1) were used to test the performance of our algorithm in the prediction of regulators underlying eQTL hotspots (Tab. 7.4). All the eight groups of genes had more connections than a random set of genes (Fig. 7.2). This shows that the genes within those groups have a shared biological function. Application of the SPF method to a genetic network led to promising regulatory predictions for several eQTL-hotspots. For the eQTL-hotspot on the far right arm of chromosome V, found in all three age groups, no regulator could be identified even though the genes in this eQTL-hotspot are highly linked in Wormnet. This could mean a relatively less well studied gene might be the regulator. For the eQTL-hotspot on the left arm of chromosome I in the juvenile (L4) group four regulators could be predicted when selection includes the position of the eQTL-hotspot locus. Interestingly, 3 from top 4 suggested regulators (*Pop-1*, *xnp-1*, *lin-17* and *lin-44*) are related to WNT pathway (Tab. 7.5). *Pop-1* also associates with the L4 chromosome V eQTL-hotspot but can not be the first-order causal regulator of this QTL-hotspot as it is not located on the chromosome V locus. Both *daf-2* and *daf-16* associate with the two juvenile eQTL-hotspots possibly linking these regulatory loci together. For the eQTL-hotspot on chromosome II, specifically found in old worms, *age-1* was suggested by the analysis of the genetic newtork (Tab. 7.4).

EQTL-hotspot	Chromosome	Gene	Function
Juvenile worms			
1	I	*pop-1*	TCF/LEF TF, WNT pathway
1	I	*xnp-1*	DNA helicase, stress response
1	I	*lin-17*	Wnt signaling
1	I	*lin-44*	Wnt signaling
Old worms			
5	II	*age-1*	PI3K, daf-2 Insulin pathway

Table 7.5. Top of the predictable regulators for eQTL-hotspot gene groups by the SPF method in genetic subpart of WormNet.

Application of the SPF method to the whole network gave more diverse results presented in a Tab. 7.4. Besides a long list of candidate genes with unknown function there are obviously promising predictions of steroid-hormone receptors *nhr-218* and *nhr269* for old worms Chromosome V eQTL-hotspot and RNA

binding protein modulator *moe-3* for old worms chromosome II eQTL-hotspot.

7.3 Biological significance of statistical analysis of functional networks

Characterizing the degree of connectivity for a given gene to a specific set of genes in the network as a normalized sum of inverse distances in a network takes us to the applications of "harmonic means" for graph analysis. The main conjecture behind the application of harmonic mean to networks is as follows; if a vertex has multiple links to other vertices, the information is sent "in parallel", i.e. concurrently along the network. Thus, one can define the "efficiency", $e_{i,j}$ in communication between vertices i and j as the inverse of the shortest distance, i.e. $e_{i,j} = 1/d_{i,j}$, see [191]. The average efficiency is straightforwardly related to the definition of the SPF. It should be noted that the interpretation of the SPF function as the efficiency of communication could be very useful in further dynamic analysis of the networks. Actually, let v be the velocity, with which the information travels along the network, then the amount of information sent from the node i to the node j per unit of time is just $v/d_{i,j}$. The performance P is the total amount of information propagating over the network per unit of time [191]. In our forthcoming works we plan to analyze the clusters taking into account their limited speed of information propagation. The concept of performance seems very appropriate for that.

Well-connected clusters of co-expressed genes described in this paper largely represent protein functional complexes, and they can be distinguished by the presence of a specific well-connected-6 link (unoriented) motif. This highly-connected motif can be used for detection of protein functional complexes (islands) in integral networks. These islands, in turn, serve in prediction of new regulatory nodes. In this study we used gene clusters derived from gene absolute expression values data that probably increase detection of true protein complexes expressed from indeed highly co-regulated genes [192]. Among the most interconnected clusters are the ones for ribosomal proteins and the regulation of translation, proteasome, respiratory complex 1 and several central metabolic functions. Using the most interconnected cluster 1 we tried to detect potential regulators from the associated network context. Due to a non-directional nature of the edges in WormNet and an

absence of directions in the co-expression network we were unable to distinguish between a cause, consequence or undirected physical interaction in a connected pair of proteins/functions and may only suggest the presence of the functional linkage between the expressed genes and the regulators. However, additional data from literature mining will likely help to vectorize the predicted interactions.

The cluster, which we used for the method validation, contains a large number of genes involved in the translational machinery as well as several genes with a central metabolic role, among which we found a large number of genes associated with a longevity phenotype in WormBase database (Tab. 7.2). The protein translation processes indeed have been recently considered for a central role in the regulation of aging processes [193, 194] and we assumed that the regulators predicted in this study may also be linked to the processes underlying aging and the control of longevity.

The role of regulation of translational machinery by the insulin pathway in aging has been widely discussed in literature [195, 196], however, the regulatory modules affecting the expression of the related genes downstream of *daf-2* have not been clearly defined. The candidates suggested by the FCG algorithm, *iff-1* and *bir-2* were shown to depend on *daf-16*-insulin response [197] and *iff-1* has also been detected in gene expression screen for the longevity phenotype in *C.elegans* [194]. *iff-1* is a eIF-5A homolog [198], and eIF-5A links processes of mRNA translation to the nonsense-mediated mRNA decay (NMD) [199]. Activation of eIF-5A requires posttranslational modication of one of the protein‘s lysines into hypusine, where spermidine is used as a substrate for the modifying NAD-dependent enzyme deoxyhypusine synthase. Spermidine is known to be involved in life span regulation and reproduction in a range of different organisms [200, 201, 202], though the mechanism of this action is not clear. We suggest that its stimulatory role in NMD via regulation of eIF-5A may be of importance in regulating translation and as a consequence the life span of an organism. Interestingly, the ribosome maturation as well as the mRNA binding SBDS protein [203, 204] that is linked in a network to *iff-1* and *tin9.2*, are both required for the longevity phenotype of *daf-2* [205]. We schematically simplified the suggested functional relationship between the predicted regulators and longevity Fig. 7.5. Our analysis points to a potential role of mRNA decay processes downstream of the insulin-dependent pathway in regulating translation and longevity.

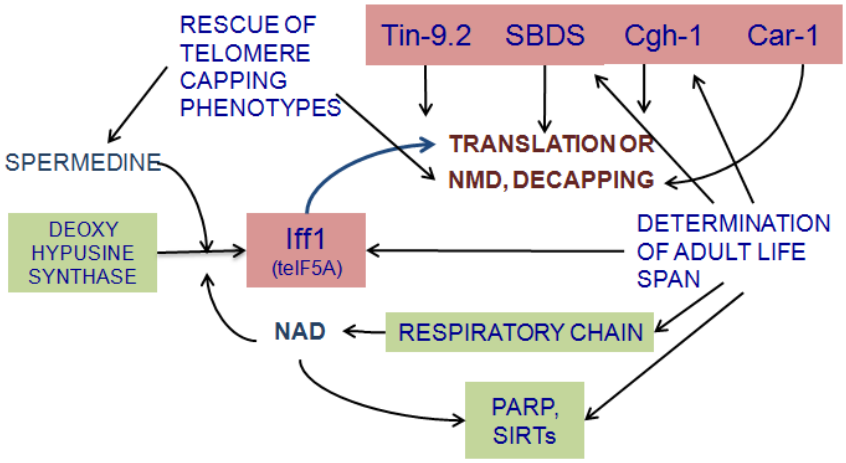


Figure 7.5. Potential involvement of the suggested regulators of protein translation in longevity determination.

We suggest that elements of translational machinery, that are regulated via insulin/caloric restriction, may be indeed non-responsive to temperature changes as we see on the example of the cluster 1 genes. Such persistence of expression may require specific mechanisms of adjustment to altered kinetics of biochemical reactions and may indeed involve a regulated mRNA decay process. Homeostasis of pathways regulated by nutrient-supply regardless of the temperature may lead to a very species-specific dynamics of cell survival and growth and an organism‘s life span adapted to the specific ecological dynamics of nutrients flow. The transcription factors predicted in our study by the SPF method may also occur to be involved in aging and regulation of longevity . The genes *cgh-1* [206], *dao-5* [207], *hel-1* [208] were already linked to aging processes downstream *daf-2*, *daf-16*, and in case of *dao-5* – to a *daf-16* independent pathway associated with determination of the adult life span GO-term in WormBase database. Analysis of genotype-phenotype relationships [209] when more data for the listed genes are available would allow deeper understanding of the direction of the defined links and more narrow prediction of their function.

This may especially be useful in finding the causal genes fo gene expression QTLs in genomics studies. As genes sharing an eQTL are very likely to have a common regulator as well as a biological function, the candidate regulators are still numerous when only based on genomic position a method for refinement is needed. Our methods predict the most likely regulator based on hundreds of previously published experiments, e.g. those used to generate the original

network [190](Fig. 7.6). Finding *age-1* as a possible regulator for a eQTL-hotspot expressed in old age worms is especially interesting as this gene has been suggested to be also responsible for the affect of heat-shock on lifespan [180]. *Pop-1*, a predicted regulator of the chromosome I juvenile eQTL-hotspot, is a TF that functions as a component of WNT signaling pathways [193]. Both longevity-related DAF-2 and DAF-16 interact with POP-1 binding partner, beta-catenine (armadillo, zemphBar-1) (String database of protein-protein interactions), and may crosstalk with WNT signaling. It was shown, for instance, that DAF-2, DAF-16 and BAR-1 synergistically affect the immune response to infections in *C.elegans* [210]. Even in cases when a regulator cannot be predicted by the

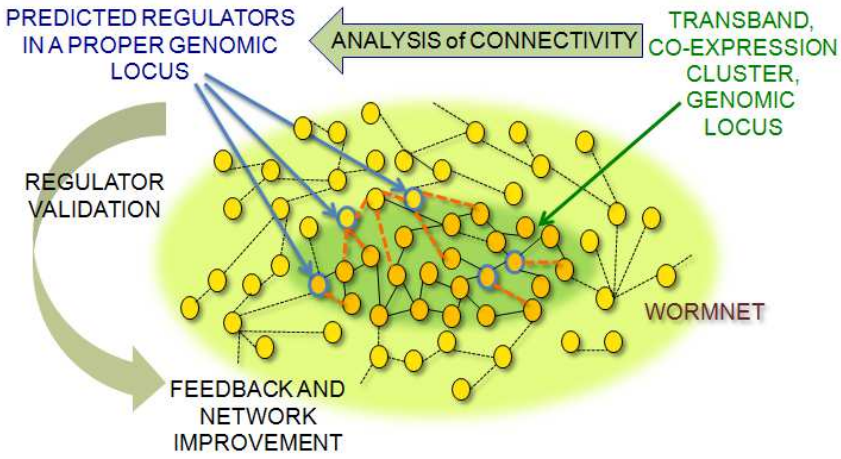


Figure 7.6. Overview of the in- and outputs of the SPF method. Dashed lines — regulatory interactions; solid lines — co-expression; orange lines — the SPF-top ranked connections; orange circles — group of co-expressed genes; yellow circles — genes in Wormnet; blue borders — predicted regulators.

SPF or FCG value ranking, the descriptive topological values (the number of connections, etc) may indicate if genes in an eQTL-hotspot are linked by biology or by a possible technical issue. Our method can be used as a validation for the biological linkage of groups of genes found in an eQTL-hotspot or by any other experiment. As the number of genomic experiments with this method increases and more species like yeast [211, 212], Arabidopsis [213, 186, 214, 215], worm [168, 171, 173, 216, 217], mice [218], chicken [219], and human [220] are investigated in such a way, an efficient way of candidate gene selection becomes urgent. This work provides new insights to the structure of biological functional networks and highlights the aspects that need to be considered in prediction of

regulatory nodes, protein complexes and regulatory modules from a multilevel network context. In our opinion, it could be useful for improvement of analytic methods and software in network-based applications in biology and other scientifc areas.

Chapter 8

Motif distributions of random networks

Here we consider random graphs, which construction was proposed by Erdös–Rényi [221]. To obtain a random network, each pair of nodes is connected with the same probability p.

8.1 Evolution in motif space

8.1.1 The law of mass action

The microscopic state of a non-directed N–vertex network can be defined by $N(N-1)/2$ Boolean variables denoting presence/absence of edges. Description of the network in terms of motif concentrations corresponds to a mapping of this high-dimensional space onto a low-dimensional motif space, resembling construction of a macroscopic description from a microscopic one in statistical mechanics. A notion of an entropy, as a number of microscopic realizations corresponding to a given macroscopic state, naturally emerges from this mapping. We argue that the *entropic landscape* of a network influences the observed motif distribution in a crucial way. Entropically favorable motif distributions, i.e. those corresponding to local maximum of possibilities to construct a network from a given set of subgraphs, should be more stable than others, and can be considered as effective traps for a network dynamics and evolution. Such entropically favorable states correspond to *islands of stability* in a sea of motif distributions, as conjectured in [222].

The motif distribution is characterized by a vector of normalized triad concentrations c_i (4 of them in non-directed graphs, and 16 in directed ones, although not all of these concentrations are independent). Note, that in [144] the authors use renormalized densities of the motifs. In our notation, the components of their motif vector $\mathbf{Z}$ (6.3) are equal to

$$Z_i = (c_i - c_i^0)/\chi_i = \Delta c_i/\chi_i \tag{8.1}$$

where $c_i^0 = c_i(h = 0)$, and χ_i is the susceptibility of ith concentration to an external field h_i coupled to it, taken at $h_i = 0$. This linear shift does not change the results presented in our work. In this research we concentrate on non-directed networks, whose four possible triads are shown in Fig. 8.1. Studying the entropy of a network as a function of motif distribution from first principles seems to be an overwhelmingly difficult task. Instead, we introduce an auxiliary external field, h, which couples to concentration of some motif, and then run simulations to study the equilibrium behavior of the network in this external field. Such a technique, reminiscent of the biased molecular dynamics used, for example, in [223, 224], allows us, by varying h, to skew the motif distribution and thus sample the states of the network which are otherwise unaccessible. As a result, we obtain a full free energy landscape of the network as a function of motif distribution. For thermal equilibration of a network in a given external field we use the standard randomization procedure [144], in which multiple permutations of network links (see Fig. 8.1) are allowed, but the node degree distribution is conserved.

For $h = 0$ the system is stabilized in the entropically largest basin of attraction corresponding to some equilibrium distribution of motifs. As h is increased, the motif distribution gets gradually more and more skewed away from equilibrium. In the limit $h \to \infty$ the entropic effects become irrelevant, and the motif's vector coupled to the field approaches the largest possible value. In this scenario two qualitatively different behaviors are possible: (i) if the entropy is a convex function of a motif, the absolute value of the motif vector grows smoothly with h, while (ii) if the entropy, as a function of a motif, has a concave region, there exists a value $h_{\rm cr}$ at which the motif distribution undergoes an abrupt shift into a new localized state (a stability island). The latter behavior constitutes a first-order phase transition, and it is exactly what we observe in our simulations.

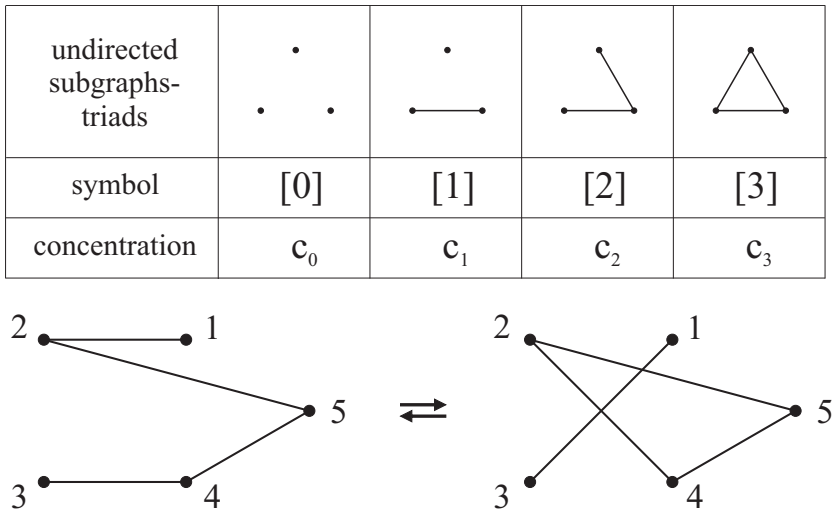


Figure 8.1. Possible triads in a non-directed network and an example of a single permutation step. The links (12) and (34) are removed, and the links (13) and (24) are added instead, thus conserving degrees of all the nodes involved. The network has 5 nodes, so only one elementary reaction occurs: one triad of type 0 ($\{135\}$) and three triads of type 2 ($\{125, 245, 345\}$) are deleted, and one triad of type 3 ($\{245\}$) and three triads of type 1 ($\{125, 135, 345\}$) are created.

The randomization procedure consists of repeated application of permutations (see Fig. 8.1) to randomly chosen pairs of links. Each permutation changes the number of triads of different types: there are $N - 4$ "elementary reactions" (one for every node that does not form the permuted edges), each of them removes 4 triads and creates 4 new ones. An example of such an elementary reaction is shown in Fig. 8.1. One can check explicitly that for a non-directed network there is only one possible non-trivial elementary reaction:

$$[0] + 3[2] \rightleftarrows [3] + 3[1] \tag{8.2}$$

Other reactions either do not change the concentrations of triads (e.g., $2[0] + 2[1] \to 2[1] + 2[0]$) or are forbidden by the rules of the randomization process. Equation (8.2) establishes a connection between the time derivatives of the triads concentrations:

$$3\dot{c}_0(t) = -\dot{c}_1(t) = \dot{c}_2(t) = -3\dot{c}_3(t), \tag{8.3}$$

so only one of them is actually independent. Therefore, three independent conservation laws (integrals of motion) control the triad dynamic. Two of them are

trivial:

$$\sum_{i=0}^{3} c_i = 1; \qquad \sum_{i=0}^{3} i c_i = 3p, \tag{8.4}$$

where p is the fraction of the links in the network (the average degree of the node is $n = p(N-1)$). The third, "hidden"conservation law can be chosen, for example, in the form

$$I_3 = \frac{1}{2}(c_0 + c_3) = \text{const} \tag{8.5}$$

The existence of hidden conservation laws is due to the particular form of randomization rules. Similarly, in directed networks, there are 16 possible triads, 3 trivial conservation laws (conservation of vertices, directed and bilateral links), and 6 hidden conservation laws, so that the dynamics of a directed network in the motif space is effectively 7-dimensional, see supplementary online materials of [144] for a more detailed discussion. We take $c(t) = \frac{1}{2}(c_3(t) - c_0(t))$ as a single independent variable describing the evolution of motifs.

Let us now introduce an external field h coupled to c. This coupling means that each randomization step gives rise to a change in energy of the system:

$$\Delta E = -\frac{1}{2} h(\Delta M_3 - \Delta M_0) = -Mh\Delta c, \tag{8.6}$$

where ΔM_0 and ΔM_3 are the changes in the number of triads of types 0 and 3 due to a particular randomization step, and $M = \frac{N!}{3!(N-3)!}$ is the total number of triads in the network. Equilibrating the system in the external field h is equivalent to making the ratio of the probabilities of forward and backward randomization steps equal $\frac{p_+}{p_-} = e^{-\Delta E}$. Without loss of generality we set the temperature of the system equal to unity, $k_B T = 1$. In the Metropolis algorithm we use this is achieved by accepting permutation step with the probability 1 if it decreases energy, and with probability $e^{-\Delta E}$ otherwise.

The energy change due to one elementary reaction of type (8.2) equals to $\Delta E_{react} = \pm h$. However, a randomization step implies many simultaneous reactions, which makes ΔE step–dependent and means that elementary reactions do not, generally speaking, happen independently. Nevertheless, as a first approximation we can still assume them to be independent, and apply the machinery of chemical kinetics to (8.2). The equilibrium reaction constant, K, should be set to $K = e^h$ ($K = 1$ in the absence of the external field, h) and the law of mass

action [225, 226] provides us with the following implicit $c(h)$–dependence

$$K \equiv e^h = \frac{c_3 c_1^3}{c_0 c_2^3} = \frac{(I_3 + c)(2 - 3p - I_3 + 3c)^3}{(I_3 - c)(3p - 1 - I_3 - 3c)^3}, \tag{8.7}$$

where in the right hand side of (8.7) we have used (8.4)–(8.5). This expression describes the change of c with changing h in the presence of conservation laws (8.4)–(8.5).

We have applied this consideration to Erdös–Rényi (ER) networks with link formation probability, p. Equilibrating ER network at $h = 0$, we get the average densities of triads

$$\bar{c}_0 = (1-p)^3;\ \bar{c}_1 = 3(1-p)^2 p;\ \bar{c}_2 = 3(1-p)p^2;\ \bar{c}_3 = p^3. \tag{8.8}$$

These concentrations, as expected, satisfy (8.7) with $K = 1$, which provides a self–consistency check of our approach.

8.1.2 Statistics of subgraphs far from equilibrium

Next, we have performed a Metropolis randomization procedure for $h = \ln K > 0$. The results are demonstrated in Fig. 8.2. The dashed line shows the $\Delta c(h) = c(h) - c(h = 0)$–dependence as given by (8.7) with $c(h = 0) = \frac{1}{2}(\bar{c}_3 - \bar{c}_0) = \frac{1}{2}(p^3 - (1-p)^3)$. The saturation of this dependence at $\Delta c \approx 0.08$ for high h (as shown in the inset) is due to depletion of triads of type 2 with growing c. We compare this behavior with the numeric results for the same p and different network sizes, N (the main plot in the figure). In the vicinity of $h = 0$ the numerical results are in good agreement with the law of mass action, but increasing of h leads to an abrupt change in Δc to a value of $\Delta c \approx 0.05$, which is not predicted by the law of mass action.

This disagreement suggests that correlations between elementary reactions become important far away from equilibrium point ($\Delta c = 0$), which causes the violation of the law of mass actions. To check this, we have studied distributions of elementary reactions (8.2) corresponding to one randomization step for different fixed values of Δc. The resulting distributions are plotted in Fig. 8.3 for varying Δc and fixed $N = 40$ and $p = 0.35$. Technically, the calculation was done as follows. To obtain a network with given $c = c^*$ we equilibrated ER network in

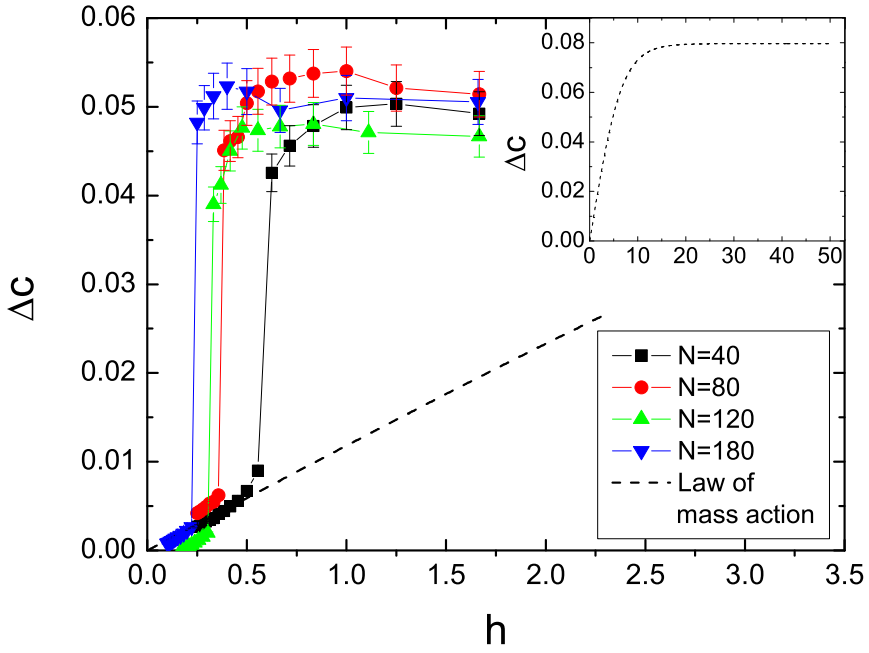


Figure 8.2. The motif distribution $\Delta c(h) = c - c(h = 0)$ in ER networks with $p = 0.35$ re-equilibrated at different h. Solid lines – numerical results for networks of sizes $N = 40, 80, 120, 180$, dashed line – prediction of the law of mass action (8.7). Inset: law of mass action extended to a region of large h.

a potential $H_c = \exp(a|c - c^*|^2)$ with sufficiently large a, so that the network becomes strongly localized around $c = c^*$. Then, for each attempted step of randomization (regardless of whether it is accepted or rejected), the difference, $\mathcal{N} = \mathcal{N}_f - \mathcal{N}_b$ between "forward", $\mathcal{N}_f$ and "backward", $\mathcal{N}_b$, elementary reactions corresponding to this step was calculated.

One sees that at $\Delta c \approx 0$ the distribution $P(\mathcal{N})$ is slightly shifted to the left, but still is nearly Gaussian. This signals that different backward and forward reactions occur independently from each other and the law of mass actions is valid in this limit. However, as Δc gets progressively larger, the $P(\mathcal{N})$–distribution becomes substantially non-Gaussian, developing a bimodal shape in the transition region $0.01 < \Delta c < 0.05$. This indicates that the elementary reactions are no longer independent, and all the permutations can be roughly divided into two classes: (i) those which do not change the motif distribution much (the right peak in the distribution), and (ii) those which lead to an essential reduction in the number of triads of type 3, i.e. pushing the system towards the equilibrium motif distribution. As motif concentration approaches the saturation value, the forward reactions $[0] + 3[2] \to [3] + 3[1]$ get suppressed, as there are almost no subgraphs

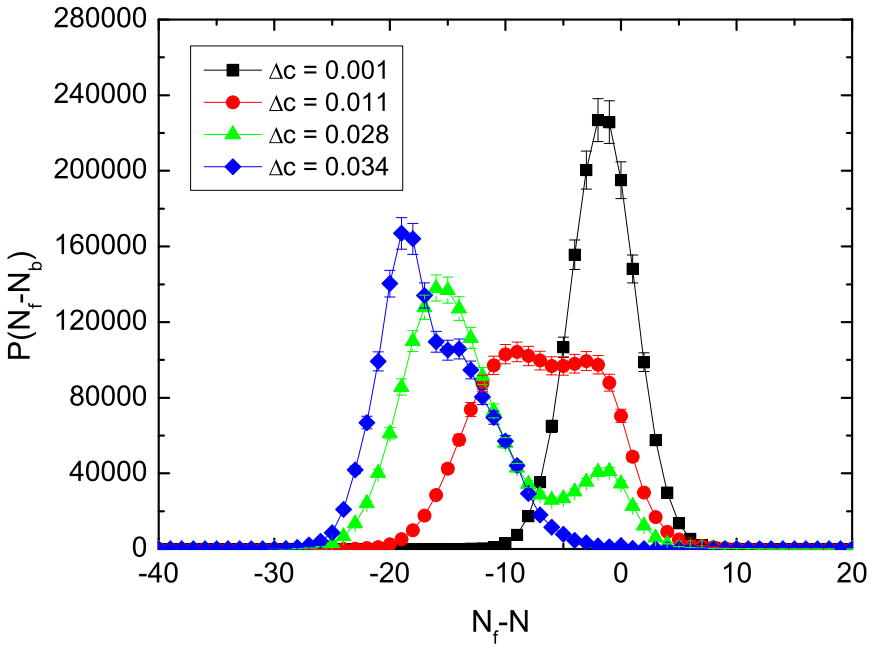


Figure 8.3. Distribution P(N) for different values of the order parameter $\Delta c = 0.001, 0.011, 0.028, 0.034$**.** Y-axis scale is arbitrary.

of type 2 left in the system, and the backward reactions $[0] + 3[2] \leftarrow [3] + 3[1]$ become dominant.

8.2 Description of phase transition in the space of subgraphs

The abrupt change of Δc with changing h, depicted in the Fig. 8.2, is reminiscent of a first order phase transition from the low–field phase with the motif distribution close to that of the equilibrium Erdös–Rényi network, to the high–field phase with strongly skewed motif distribution. It seems natural to describe this transition in the frameworks of the phenomenological mean–field Landau–type theory [226]. Assume the excess of the motif concentration $\Delta c = \eta$ to play the role of an order parameter. Clearly, there is no $\eta \leftrightarrow -\eta$ symmetry in the problem, so the Landau expansion of the free energy $H(\eta)$ should include both odd and even powers of η, and up to the 4th order term it reads

$$\begin{aligned} \mathcal{H} &= M(H_0 - h\,\eta); \\ H_0 &= \tfrac{\chi}{2}\eta^2 + \tfrac{b(N)}{3}\eta^3 + \tfrac{g(N)}{4}\eta^4 + o(\eta^4) \end{aligned} \tag{8.9}$$

Here H_0 is a specific free energy in the absence of an external field, with an equilibrium point $\eta = 0$. This H_0 is a purely combinatorial object, it is temperature– and field–independent. Since both the mean-field theory, and the numerical results at low h, do not depend on the size of the system N, the same is true for the susceptibility χ, while we expect the higher orders in the expansion of H_0 to be N–dependent. If the third-order term in (8.9) is large enough, a first order phase transition may occur: as h increases, the $H(\eta)$–dependence gets tilted, for $b^2 > 3g\chi$ it eventually exhibits two competing minima at different values of η. When the values of H match in these minima, a transition occurs. Hence, the equilibrium motif, $\bar{\eta}(h)$, is defined by a minimization of (8.9) $\frac{\partial H(\eta)}{\partial \eta} = 0$, giving

$$\chi\,\bar{\eta} + b\,\bar{\eta}^2 + g\,\bar{\eta}^3 = h \tag{8.10}$$

Equation (8.10) has either one or three solutions for a given h. The zero–field susceptibility, χ, can be calculated in the standard way: $\chi \equiv \left.\frac{\partial^2 H(\eta)}{\partial \eta^2}\right|_{h=0} = \left.\frac{\partial h(c)}{\partial c}\right|_{h=0}$. Expanding (8.7) up to the linear term in h, we get

$$\chi = \frac{1}{\bar{c}_0} + \frac{9}{\bar{c}_1} + \frac{9}{\bar{c}_2} + \frac{1}{\bar{c}_3} = \frac{1}{p^3(1-p)^3} \tag{8.11}$$

For an ER network with $p = 0.35$ (8.11) gives $\chi \approx 85$. Choosing the parameters $\chi, b(N), g(N)$ in (8.9) as $\chi = 85$, $b = 4.55 \times 10^3$ and $g = 6.5 \times 10^4$, we demonstrate in the Fig. 8.4 that the phenomenological Landau theory matches the experimental $\Delta c(h)$–dependence with a reasonable accuracy. Moreover, the Landau theory predicts a hysteretic behavior of $\Delta c(h)$, which is actually seen in the simulations for few samples (A,B,C), as shown in the Fig. 8.4. The hysteresis has been recorded in the $\Delta c(h)$–curve when the strength of the field h is increased adiabatically from zero up to the maximal value $h = 2$ (in dimensionless units) and then adiabatically decreased back to $h = 0$.

To summarize, a mapping from microscopic (in terms of configuration of network connections) to macroscopic (in terms of triads concentration) description gives rise to a notion of motif distribution entropy. In the presence of an external field, h, the equilibrium value of the motif concentration, Δc (which is effectively one-dimensional for the non-directed case considered here) is determined by the balance of the energy imposed by h, and this entropy induced by a mapping from microscopic to macroscopic description. If the entropic landscape is concave, a

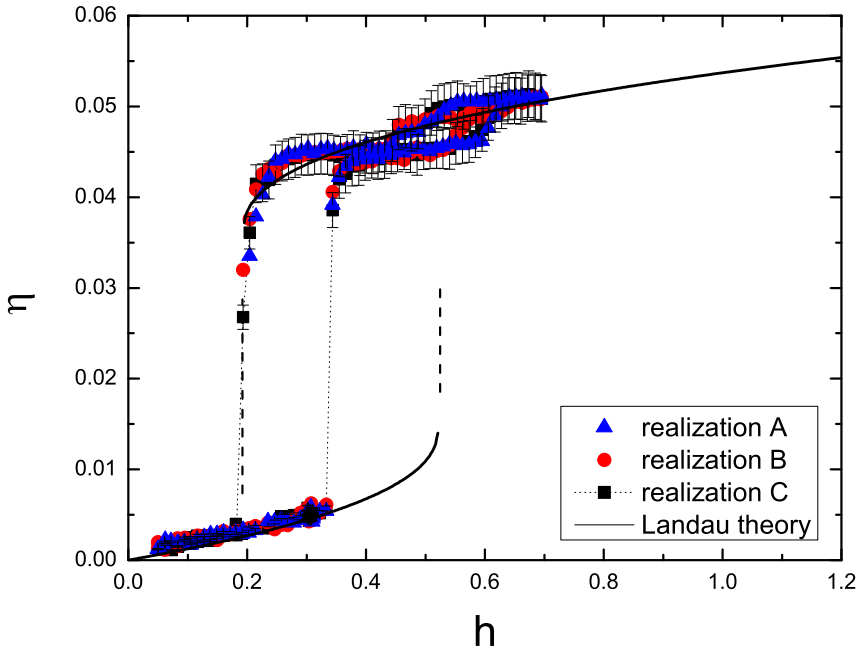


Figure 8.4. Comparison of the numerically observed dependence $\Delta c(h)$ to the solution of the mean–field equation (8.10). The following values of theory parameters are taken $\chi = 85$; $b = 4.55 \times 10^3$ $g = 6.5 \times 10^4$.

first–order phase transition into a state with highly–skewed motif distribution occurs at some critical h. This transition, observed numerically for Erdös–Rényi networks, violates the law of mass action due to correlations between elementary reactions (8.2) in large fields (see Fig. 8.3). This transition is well described by a phenomenological Landau theory with Δc as an order parameter.

To verify that described behavior is a generic phenomenon and not a peculiarity of a particular type of randomization dynamics, we modified the permutation rules allowing an edge connecting two arbitrary vertices (i, j) to be switched at a single step to some other pair (i, k) $(k \neq i, j)$. Under this dynamics the node degrees are not conserved, and the condition (8.5) fails. Accordingly, the dynamics in the motif space becomes effectively two-dimensional with elementary reactions

$$2[2] \rightleftarrows [1] + [3]; \quad 2[1] \rightleftarrows [0] + [2]$$

However, application of an external field $\mathbf{h}$ (which is, in this case, a 2D vector) still leads to a transition into a localized state (the full details of the corresponding simulation will be provided elsewhere). Therefore, one assumes that this phenomenon — localization of the motif distribution under external field into distinct

entropic traps — is apparently universal. We conjecture that stable motif profiles constituting superfamilies [144] may correspond to such stability islands inherent to the complicated underlying entropic landscape of the multidimensional motif space of the underlying network. Study of the concrete details of these landscapes and the basins of attraction of different stability islands in terms of a multidimensional external field **h** would be an interesting and challenging task. The concept of entropy–induced localization discussed above may be instrumental in various other fields. Compare it, for example, with the celebrated Eigen model of biological evolution in the space of heteropolymer sequences [227]. There, the localization–delocalization phase transition, known as the "error catastrophe", separates two states: where the genotype is localized in the vicinity of a preferred pattern, and where it is completely random [228, 229, 230]. The transition occurs due to an interplay between the attraction to a point–like potential well and the entropic repulsion from this well due to the exponential growth of the number of states with increasing Hamming distance from the well. In our case, a different, but complimentary behavior takes place: the nontrivial entropic landscape of the system acts as a source of effective traps, while the uniform external field regulates the transition from one trap to another. It seems that trapping of a complex system into stability islands due to a competition between selection and randomness, provides a generic mechanism of localization in complex biological and social systems.

Conclusions

In this work, we have shown how different mathematical methods can be applied for studying statistical and dynamic objects of complex architecture and, in particular, structures which have no characteristic spatial and/or time scale. Systems with complex architecture are very common in biophysics and their description usually deals with the combinatorial complexity of the problem. In this thesis we have developed some new statistical approaches and models, which allow to analyze the complex biophysical systems. The principal results of our investigations have been published in five research articles; three other manuscripts are submitted.

In Chapter 2 we have developed and implemented a new statistical algorithm for quantitative determination of the binding free energy of two heteropolymer sequences under the assumption that each sequence can form a hierarchical cactus-like secondary structure, typical for RNA molecules. We have proposed in Section 2.3 a constructive way to build a "cost function" characterizing the matching of two RNAs with arbitrary primary sequences. Since base-pairing of two ncRNAs or between ncRNA and DNA plays very important biological role, it is worth estimating theoretically the binding free energy of the ncRNA-target RNA complex by knowing the primary sequences of chains under consideration. Note, that this problem differs from the complete alignment of two RNA sequences: in ncRNA case we align only the sequences of nucleotides which constitute pairs between two RNAs, while the secondary structure of each RNA comes into play only by the combinatorial factors affecting the entropic contribution of chains to the total cost function.

The proposed algorithm is based on two facts: i) the standard alignment problem can be reformulated as a zero–temperature limit of more general statistical problem of binding of two associating heteropolymer chains; ii) the last problem

can be straightforwardly generalized onto the sequences with hierarchical cactus–like structures (i.e. of RNA–type). Taking zero–temperature limit at the very end we arrive at the desired ground state free energy with account for entropy of side cactus–like loops.

Chapter 3 and Chapter 4 are devoted to the study of statistical properties of random biopolymers. We have analyzed important statistical properties of random RNA–RNA complexes, including fluctuations of the binding energy between a pair of macromolecules and loop length distribution in a complex. The results obtained for linear polymer and RNA-like structures show that their energy statistical properties are those of systems in the KPZ universality class. The loop length distribution in linear polymer complexes is precisely similar to that characteristic of independent binding. The loop length distribution in cloverleaf structures is quite well described by a random walk model. Furthermore, using this model, we put forward a hypothesis about critical behavior of random RNA-type heteropolymers.

In Chapter 4 we have demonstrated that alphabets with different number of letters, c, are nonequivalent. This nonequivalence is tightly coupled to the restrictions on the morphology of allowed secondary structures. Indeed, the existence of two regimes (for $c \leq c_{\rm cr}$ and $c > c_{\rm cr}$) is a peculiarity of RNAs and is due to the additional freedom in the formation of the complex cactus-like secondary structures typical for RNAs. For linear matching problem used in DNA comparison, the fraction of nucleotides in the optimal alignment is less than 1 for any alphabet with $c > 1$. In our model the transition between two regimes occurs at $2 < c_{\rm cr} < 4$. The exact value of the critical alphabet size should be sensitive to the microscopic details of the model, and one can list factors which are neglected in our model and which could shift the transition point to the right or to the left from the observed critical value.

We have considered this problem as the planar matching problem, defined by a symmetric random matrix with independent identically distributed entries, taking values 0 and 1. We show that the existence of a perfect planar matching structure is possible only above a certain critical density, p_c, of allowed contacts (i.e. of '1's). Using a formulation of the problem in terms of Dyck paths and a matrix model of planar contact structures, we provide an analytical estimation for the value of the transition point, p_c, in the thermodynamic limit. This estimation is

close to the critical value, $p_c \approx 0.379$, obtained in numerical simulations based on an exact dynamical programming algorithm. We characterize the corresponding critical behavior of the model and discuss the relation of the perfect-imperfect matching transition to the known molten-glass transition in the context of random RNA secondary structure's formation. In particular, we provide strong evidence supporting the conjecture that the molten-glass transition at $T = 0$ occurs at p_c.

In Chapter 5 we have proposed a new model of a heteropolymer chain with RNA-type topology of secondary structure and quenched random distribution of intervals between neighboring monomers. For quantitative analysis of the Random Interval Model (RIM), we have investigated the statistical behavior of height diagrams as a function of the control parameter in the distribution function of intervals. The important result deserving attention, concerns the possibility to pass from the nonlocal recursion relation for the ground state free energy to the local recursion relation if and only if the interaction energy between paired monomers is a concave function of distance. So, for any potential (even random) of concave form, the ground state equation can be essentially simplified resulting in shortening the computational time if these equations are implemented for numerical analysis of secondary structures of polymer chain with RNA-type architecture.

Chapter 6, 7 and 8 are devoted to statistics in networks. We have shown how the statistical analysis of network connectivity can be used to predict new gene expression regulators. We have proposed that co-expression clusters can be easily identified as highly-connected islands within the integrated network, and that these islands can be used to suggest new regulatory genes for subsequent verification. For this, we have developed a new application of a modified statistical algorithm, based on so-called "shortest path function" (SPF) to rank the nodes that have most effect on gene expression. We have presented an algorithm which uses eQTL data in combination with the majority of published functional interactions in *C. elegans*. We have show that applied to longevity-specific eQTL data for *C. elegans* published in [171] leads to reasonable regulatory gene predictions. Interpretation of the organism-specific integral biological networks and prediction of protein complexes and genetic regulators from a network context may benefit greatly from our study and the new algorithms. It can be applied to other known organism-specific networks.

In the last chapter, random non-directed Erdös–Rényi networks subject to

a dynamics conserving vertex degrees have been considered. We have studied analytically and numerically equilibrium three-vertex motif distributions in the presence of an external field coupled to one of the motifs. For small magnitude of the external field the numerics is well described by chemical kinetics equations based on the law of mass action for the concentrations of motifs. For larger external fields a transition into a state with some trapped motif distribution occurs. We have explained the existence of the transition by employing the notion of the entropy of the motif distribution and describe it in terms of a phenomenological Landau–type theory with a non-zero cubic term. We argue that the localization transition should always occur if the entropy function is non-convex. We conjecture that this phenomenon may be the reason for motifs' pattern formation in real networks.

Refereed Journal Publications

1. S. K. Nechaev, M.V. Tamm and **O.V. Valba**,(2011) "Statistics of non-coding RNAs: alignment and secondary structure prediction", *Journal of Physics A: Mathematical and Theoretical*, Vol. 44. No.19.

2. **O.V. Valba**, S. K. Nechaev, and M.V. Tamm, (2012) "Alignment of RNA Molecules: Binding Energy and Statistical Properties of Random Sequence", *Journal of Experimental and Theoretical Physics*, Vol. 114, No. 2.

3. **O.V. Valba**, S. K. Nechaev, and M.V. Tamm, (2012) "Interaction of RNA Molecules: Binding Energy and the Statistical Properties of Random Sequences", *Russian Journal of Physical C hemistry B*, Vol. 6, No. 3.

4. **O.V. Valba**, M.V. Tamm and S. K. Nechaev, (2012) "New Alphabet-Dependent Morphological Transition in Random RNA Alignment", *Physical Review Letters*, V. 109(1):018102.

5. S.K. Nechaev, A.N. Sobolevskii, **O.V. Valba**,(2013) " Planar diagrams from optimization for concave potentials", *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, Vol. 87, No. 1.

Submitted Contributions

1. **O.V. Valba**, S.K. Nechaev, M. G. Sterken, L.B. Snoek, J. E. Kammenga, O. Vasieva, "On predicting regulatory genes by analysis of functional networks in C.elegans", arXiv:1302.3349.

2. A.Y. Lokhov, S.K. Nechaev, M.V. Tamm, **O.V. Valba**, "New phase transition in random planar diagrams and RNA-type matching", arXiv:1307.2170.

3. V.A. Avetisov, S.K. Nechaev, A.B. Shkarin, M.V. Tamm, **O.V. Valba**, " Islands of stability in motif distributions of random networks", arXiv:1307.0113.

Acknowledgement

I would like to extend my gratitude to my thesis advisor Sergei Nechaev for formulation of interesting problems, fruitful discussions, and the opportunity of taking part in numerous conferences and seminars. Without his guidance and persistent help this dissertation would not have been possible.

I would like to thank Mike Tamm from Moscow State University, who has been involved in most of research activity presented in this work.

I am very grateful to Andrey Lokhov (LPTMS) for collaboration, particularly, for analytical consideration of perfect matching transition from the matrix viewpoint. And also I thank him for the eternal optimism and friendly support.

The optimization transport problem in application to RNA structure model has been done in collaboration with Andrey Sobolevskii (Institute of Transmission Information Problem) who introduced to me the basic grounds of the mathematical area of "optimal transportation".

The research on *C.elegance* was supported by the grant: ERASysbio project GRAPPLE #90201066 - Iterative modeling of gene regulatory interactions underlying stress, disease and ageing in *C. elegans*. I would like to express my gratitude to Andrew Cossins's group in Institute of Integrative Biology (University of Liverpool), M. Madan Babu's Lab (Cambridge), Laboratory of Nematology (Wageningen University) for the hospitality. Especially, I would like to thank Olga Vasieva (Liverpool University) for many discussions and explanations of the biological interpretation of our results.

I am very grateful to V.A. Avetisov and lab of complex system in Institute of Chemical Physics RAS for stimulating discussions, formulation of problems, and warm reception, as well as to the Poncelet Lab at the Independent University (Moscow), where my first results have been discussed at the interdisciplinary seminar of young researchers organized by S. Nechaev, V. Avetisov and A. Sobolevsky.

I highly acknowledge the kind hospitality of LPTMS, which becomes during these years a part of my family. Especially I am very grateful to Emmanuel Trizac for very warm attention and to Claudine le Vaou, who has resolved with extreme efficiency many difficult technical problems during my stay at LPTMS.

I would like to extend my deepest gratitude to my parents Vladimir and Natalia Valba, without whose love, support and understanding I could never have completed this thesis.

Finally, I would like to thank the jury members: Prof. R. Zecchina, Prof. M. Weigt, Prof. O. Martin and Dr. D. Grebenkov for their willing review, valuable and helpful suggestions to improve this thesis.

Appendix A

Derivation of Equation (5.8)

Suppose $X = \{x_i\}_{1 \le i \le 2n}$ with $x_1 < x_2 < \cdots < x_{2n}$ and $X' = \{x'_{i'}\}_{1 \le i' \le 2n'}$ with $x'_1 < x'_2 < \cdots < x'_{2n'}$ are two sets such that $x_{2n} < x'_1$, i.e., X' lies to the right of X.

We will refer to minimum-weight perfect matchings on X and X', i.e., planar (nonintersecting) sets of n (resp. n') arcs connecting the points such that the sum of their weights, which are given by a cost function $w(\cdot, \cdot)$ of concave type, is minimal, as *partial matchings* and to the minimum–weight perfect matching on $X \cup X'$ as *joint matching*.

Call an arc (x_i, x_j) in a nested matching *exposed* if there is no arc $(x_{i'}, x_{j'})$ with x_i, x_j contained between $x_{i'}$ and $x_{j'}$. We call all other arcs in a nested matching non-exposed or *hidden*. Intuitively, exposed arcs are those visible "from above" and hidden arcs are those covered with exposed ones.

We first show, following [122], that whenever an arc (x_i, x_j) is hidden in the partial matching on X, it belongs to the joint optimal matching and is hidden there too. By contradiction, assume that some of hidden arcs in the partial matching on X do not belong to the joint matching. Then there will be at least one exposed arc (x_ℓ, x_r) in the partial matching on X such that some points x_i with $x_\ell < x_i < x_r$ are connected in the joint matching to points outside (x_ℓ, x_r).

Denote all the points in the segment $[x_\ell, x_r]$ that are connected in the joint matching to points on the left of x_ℓ by $z_1 < z_2 < \cdots < z_k$; denote the opposite endpoints of the corresponding arcs by $y_1 > y_2 > \cdots > y_k$, where the inequalities follow from the fact that the joint matching is nested. Likewise denote those points from $[x_\ell, x_r]$ that are connected in the joint matching to points on the

right of x_r by $z'_1 > z'_2 > \cdots > z'_{k'}$ and their counterparts in the joint matching by $y'_1 < y'_2 < \cdots < y'_{k'}$. Observe that although k or k' may be zero, the number $k + k'$ must be positive and even.

Consider now a matching on the segment $[x_\ell, x_r]$ that consists of the following arcs: those arcs of the joint matching whose both ends belong to $[x_\ell, x_r]$; the arcs (z_1, z_2), ..., $(z_{2\kappa-1}, z_{2\kappa})$, where [1] $\kappa = \lfloor k/2 \rfloor$; the arcs (z'_2, z'_1), ..., $(z'_{2\kappa'}, z'_{2\kappa'-1})$, where $\kappa' = \lfloor k'/2 \rfloor$; and $(z_k, z'_{k'})$ if both k and k' are odd. Denote by W' the weight of this matching. It cannot be smaller than the weight W'_0 of the restriction of the optimal partial matching on X to $[x_\ell, x_r]$. For the total weight W of the joint matching on $X \cup X'$ we thus have

$$W \geq W - W' + W'_0. \tag{A.1}$$

We now show that by a suitable sequence of uncrossings the right–hand side here can be further reduced to a matching whose weight is strictly less than W.

The arcs (z_1, y_1) and (x_ℓ, x_r) are crossing, so that $w(y_1, z_1) + w(x_\ell, x_r) > w(y_1, x_\ell) + w(z_1, x_r)$. Uncrossing these arcs strictly reduces the right-hand side of (A.1):

$$W > W - W' + W'_0 \\ - w(y_1, z_1) - w(x_\ell, x_r) + w(y_1, x_\ell) + w(z_1, x_r).$$

Now the arcs (y_2, z_2) and (z_1, x_r) are crossing, so $w(y_2, z_2) + w(z_1, x_r) - w(z_1, z_2) > w(y_2, x_r)$ and therefore

$$W > W - W' + W'_0 - w(y_1, z_1) - w(y_2, z_2) - w(x_\ell, x_r) + w(y_1, x_\ell) + w(z_1, z_2) + w(y_2, x_r).$$

Repeating this step $\kappa = \lfloor k/2 \rfloor$ times gives the inequality

$$W > W - W' + W'_0 - w(x_\ell, x_r) - \sum_{1 \leq i \leq 2\kappa} w(y_i, z_i) \\ + \sum_{1 \leq i \leq \kappa} w(z_{2i-1}, z_{2i}) + \sum_{1 \leq i \leq \kappa} w(y_{2i-1}, y_{2i-2}) + w(y_{2\kappa}, x_r),$$

where in the rightmost sum y_0 is defined to be x_ℓ. Note that at this stage all arcs

[1] $\lfloor \xi \rfloor$ is the largest integer n such that $n \leq \xi$.

coming to points $z_1, z_2, \dots$ from outside $[x_\ell, x_r]$ are eliminated from the matching, except possibly (y_k, z_k) if k is odd.

It is now clear by symmetry that a similar reduction step can be performed on arcs going from $z'_1, z'_2, \dots$ to the right.

Finally if k and k' are odd, we uncross the pair of arcs (y_k, x_k) and $(y_{k-1}, y'_{k'-1}$ and finally the pair $(z_k, y'_{k'-1})$ and $(z'_{k'}, y'_{k'})$.

The final estimate for W has the form

$$\begin{aligned} W > W - W' + W'_0 - w(x_\ell, x_r) - \sum_{1\le i\le k} w(y_i, z_i) - \sum_{1\le i'\le k'} w(z'_{i'}, y'_{i'}) \\ + \sum_{1\le i\le \kappa} w(z_{2i-1}, z_{2i}) + \sum_{1\le i'\le \kappa'} w(z'_{2i'}, z'_{2i'-1}) + w(z_k, z'_{k'}) \cdot [k,\, k' \text{ are odd}] \\ + \sum_{1\le i\le \kappa} w(y_{2i-1}, y_{2i-2}) + \sum_{1\le i'\le \kappa'} w(y'_{2i'-2}, y'_{2i'-1}) + w(y_k, y'_{k'}) \cdot [k,\, k' \text{ are even}], \end{aligned} \tag{A.2}$$

where notation such as $[k,\, k' \text{ are odd}]$ means 1 if k, k' are odd and 0 otherwise.

The right-hand side of (A.2) contains four groups of terms: first,

$$W - \sum_{1\le i\le k} w(y_i, z_i) - \sum_{1\le i'\le k'} w(z'_{i'}, y'_{i'}),$$

corresponding to the joint matching without the arcs connecting points inside $[x_\ell, x_r]$ to points outside this segment; second,

$$\begin{aligned} W' - \sum_{1\le i\le \kappa} w(z_{2i-1}, z_{2i}) - \sum_{1\le i'\le \kappa'} w(z'_{2i'}, z'_{2i'-1}) \\ - w(z_k, z'_{k'}) \cdot [k,\, k' \text{ are odd}], \end{aligned}$$

which comes with a negative sign and corresponds to the arcs of the joint matching with both ends inside $[x_\ell, x_r]$, and cancels them from the total; third, $W'_0 - w(x_\ell, x_r)$, with positive sign, which corresponds to the hidden arcs of the partial matching on X inside the exposed arc (x_ℓ, x_r), not including the latter; and finally the terms in the last line of (A.2), corresponding to the arcs matching x_ℓ, x_r, and points $y_1, \dots, y_k, y'_1, \dots, y'_{k'}$, i.e., those points outside $[x_\ell, x_r]$ that were connected in the joint matching to points inside this segment.

Gathering together contributions of these four groups of terms, we observe that all negative terms cancel out and what is left corresponds to a perfect matching with a weight strictly smaller than W, in which all arcs hidden by (x_ℓ, x_r) in the partial matching on X are restored. There may still be some crossings caused by terms of the fourth group and *not* involving the hidden arcs in $[x_\ell, x_r]$; uncrossing them if necessary gives a nested perfect matching whose weight is strictly less than that of the joint matching. This contradicts the assumption that the latter is the minimum-weight matching on $X \cup X'$. Therefore all hidden arcs in the partial matching on X (and, by symmetry, those in the partial matching on X') belong to the joint matching.

Now let i, j be indices of opposite parity and such that $i < j$, and define $W_{i,j}$ to be the weight of the minimum-weight perfect matching on the $j - i + 1$ points $x_i < x_{i+1} < \cdots < x_j$. We can now show, following [122], that for all indices i, j of opposite parity with $1 \le i < j \le 2n$, weights $W_{i,j}$ satisfy the recursion

$$W_{i,j} = \min \big[w(x_i, x_j) + W_{i+1,j-1}; \\ W_{i,j-2} + W_{i+2,j} - W_{i+2,j-2} \big] \quad \text{(A.3)}$$

with "initial conditions"

$$W_{i,i-1} = 0, \quad W_{i+2,i-1} = -w(x_i, x_{i+1}). \quad \text{(A.4)}$$

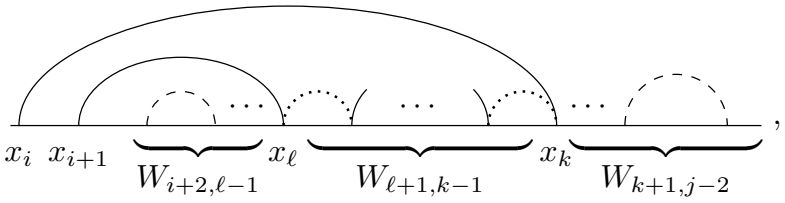
For simplicity we will refer to the minimum-weight perfect matching on points $x_r < x_{r+1} < \cdots < x_s$ as the "matching $W_{r,s}$." Consider first the matching that consists of the arc (x_i, x_j) and all arcs of the matching $W_{i+1,j-1}$, and observe that by optimality the latter its weight $w(x_i, x_j) + W_{i+1,j-1}$ is minimal among all matchings that contain (x_i, x_j).

We now examine the meaning of the expression $W_{i,j-2} + W_{i+2,j} - W_{i+2,j-2}$. Denote the point connected in the matching $W_{i,j-2}$ to x_i by x_k and the point connected to x_{i+1} by x_ℓ. It is easy to see that the pairs of indices i, k and $i+1, \ell$ both have opposite parity. Assume first that

$$x_{i+1} < x_\ell < x_k \le x_{j-2}. \quad \text{(A.5)}$$

Observing that hidden arcs in partial matchings on the sets $X = \{x_i, x_{i+1}\}$

and $X' = \{x_{i+2}, \dots, x_{j-2}\}$ are preserved, and taking into account parity of k and ℓ, we see that x_k and x_ℓ (as well as their neighbors x_{k+1} and $x_{\ell-1}$ if they are contained in $[x_{i+2}, x_{j-2}]$) belong to exposed arcs of the matching $W_{i+2,j-2}$. Thus the matching $W_{i,j-2}$ has the following structure:



where dashed (resp., dotted) arcs correspond to those exposed arcs of the matching $W_{i+2,j-2}$ that belong (resp., do not belong) to $W_{i,j-2}$.

Since points $x_{\ell-1}$ and x_{k+1} belong to exposed arcs in the matching $W_{i+2,j-2}$, the (possibly empty) parts of this matching that correspond to points $x_{i+2} < \dots < x_{\ell-1}$ and $x_{k+1} < \dots < x_{j-2}$ coincide with the (possibly empty) matchings $W_{i+2,\ell-1}$ and $W_{k+1,j-2}$. For the same reason the (possibly empty) part of the matching $W_{i,j-2}$ supported on $x_{\ell+1} < \dots < x_{k-1}$ coincides with $W_{\ell+1,k-1}$. Therefore

$$W_{i,j-2} = w(x_i, x_k) + w(x_{i+1}, x_\ell) + W_{i+2,\ell-1} + W_{\ell+1,k-1} + W_{k+1,j-2}. \quad \text{(A.6)}$$

Taking into account (A.4), observe that in the case $k = i+1$ and $\ell = i$, which was left out in (A.5), this expression still gives the correct formula $W_{i,j-2} = w(x_i, x_{i+1}) + W_{i+2,j-2}$.

Now assume that in the matching $W_{i+1,j}$ the point x_j is connected to $x_{\ell'}$ and the point x_{j-1} to $x_{k'}$. A similar argument gives

$$W_{i+2,j} = W_{i+2,\ell'-1} + W_{\ell'+1,k'-1} + W_{k'+1,j-2} + w(x_{\ell'}, x_j) + w(x_{k'}, x_{j-1}); \quad \text{(A.7)}$$

in particular, if $\ell' = j-1$ and $k' = j$, then $W_{i+2,j} = W_{i+2,j-2} + w(x_{j-1}, x_j)$.

Suppose that $x_k < x_{\ell'}$. Taking into account that x_k, x_{k+1}, $x_{\ell'-1}$, and $x_{\ell'}$ all

belong to exposed arcs in $W_{i+2,j-2}$, we can write

$$\begin{aligned} W_{k+1,j-2} &= W_{k+1,\ell'-1} + W_{\ell',j-2}, \\ W_{i+2,\ell'-1} &= W_{i+2,k} + W_{k+1,\ell'-1} \end{aligned} \tag{A.8}$$

and

$$W_{i+2,j-2} = W_{i+2,k} + W_{k+1,\ell'-1} + W_{\ell',j-2}. \tag{A.9}$$

Substituting (A.8) into (A.6) and (A.7) and taking into account (A.9), we obtain

$$\begin{aligned} W_{i,j-2} + W_{i+2,j} - W_{i+2,j-2} = w(x_i, x_k) + w(x_{i+1}, x_\ell) \\ + W_{i+2,\ell-1} + W_{\ell+1,k-1} + W_{k+1,\ell'-1} \\ + w(x_{\ell'}, x_j) + W_{\ell'+1,k'-1} + w(x_{k'}, x_{j-1}) + W_{k'+1,j-2}. \end{aligned}$$

The right-hand side of this expression corresponds to a matching that coincides with $W_{i,j-2}$ on $[x_i, x_k]$, with $W_{i+2,j-2}$ on $[x_{k+1}, x_{\ell'-1}]$, and with $W_{i+1,j}$ on $[x_{\ell'}, x_j]$. By optimality, this matching cannot be improved on any of these three segments and is therefore optimal among all matchings in which x_i and x_j belong to different exposed arcs.

It follows that under the assumption that $x_k < x_{\ell'}$ the expression in the right-hand side of (A.3) gives the minimum weight of all matchings on $x_i < x_{i+1} < \cdots < x_j$. Moreover, the only possible candidates for the optimal matching are those constructed above: one that corresponds to $w(x_i, x_j) + W_{i+1,j-1}$ and one given by the right-hand side of the latter formula.

It remains to consider the case $x_k \geq x_{\ell'}$. Since $x_k \neq x_{\ell'}$ for parity reasons, it follows that $x_k > x_{\ell'}$; now a construction similar to the above yields a matching which corresponds to $W_{i,j-2} + W_{i+2,j} - W_{i+2,j-2}$ and in which the arcs (x_i, x_k) and $(x_{\ell'}, x_j)$ are crossed. Uncrossing them leads to a matching with strictly smaller weight, which contains the arc (x_i, x_j) and therefore cannot be better than $w(x_i, x_j) + W_{i+1,j-1}$. This means that (A.3) holds in this case too with $W_{i,j} = w(x_i, x_j) + W_{i+1,j-1}$.

Bibliography

[1] A. K. Mohanty, M. Misra, and L. T. Drzal, *Natural fibers, biopolymers, and biocomposites*. CRC Press, 2005.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. Garland Science, fourth edition ed., 2002.

[3] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information," *Nucleic Acids Research*, vol. 9, p. 133, 1981.

[4] T. Akutsu, "Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots," *Discrete Applied Mathematics*, vol. 104, pp. 45–62, 2000.

[5] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3406–3415, 2003.

[6] D. H. Mathews, "Revolutions in RNA secondary structure prediction," *Journal of Molecular Biology*, vol. 359, no. 3, pp. 526–532, 2006.

[7] J. Zhang, M. Lin, R. Chen, W.Wang, and J. Liang, "Discrete state model and accurate estimation of loop entropy of RNA secondary structures," *The Journal of Chemical Physics*, vol. 128, no. 12, 2008.

[8] I. M. Lifshitz, "Some problems of the statistical theory of biopolymers," *Soviet Physics JETP*, vol. 28, p. 1280, 1969.

[9] W. Saenger, *Principles of nucleic acid structure*, vol. 7. Springer-Verlag New York, 1984.

[10] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids," *Nature*, vol. 4356, 1953.

[11] F. A. Vendeix, A. M. Munoz, and P. F. Agris, "Free energy calculation of modified base-pair formation in explicit solvent: A predictive model," *RNA*, vol. 15, no. 12, pp. 2278–2287, 2009.

[12] G. Varani and W. H. McClain, "The G–U wobble base pair," *EMBO Reports*, vol. 1, no. 1, pp. 18–23, 2000.

[13] J.-L. Chen, "Functional analysis of the pseudoknot structure in human telomerase RNA," *Proceedings of the National Academy of Sciences*, vol. 102, no. 23, pp. 8080–8085, 2005.

[14] C. W. A. Pleij, K. Rietveld, and L. Bosch, "A new principle of rna folding based on pseudoknotting," *Nucleic Acids Research*, vol. 13, no. 5, pp. 1717–31, 1985.

[15] M. Zuker, J. A. Jaeger, and D. H. Turner, "A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison," *Nucleic Acids Research*, vol. 19, no. 10, pp. 2707–2714, 1991.

[16] S. R. Eddy, "How do RNA folding algorithms work?," *Nature Biotechnology*, vol. 22, pp. 1457–58, 2004.

[17] I. Hofacker, W. Fontana, P. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatsh. Chem.*, vol. 125, pp. 167–188, 1994.

[18] S. R. Eddy, "What is dynamic programming?," *Nature Biotechnology*, vol. 22, pp. 909–910, 2004.

[19] E. Rivas and S. R. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots," *Journal of Molecular Biology*, vol. 285, pp. 2053–2068, 1999.

[20] J. Ruan, G. D. Stormo, and W. Zhang, "An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots," *Bioinformatics*, vol. 20, no. 1, pp. 58–66, 2004.

[21] A. A. Mironov, "A method for prediction of conserved rna secondary structures," *Molecular Biology (in Russian)*, vol. 41, pp. 711–18, 2007.

[22] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, pp. 223–30, 1973.

[23] C. Laing and T. Schlick, "Computational approaches to 3D modeling of RNA," *Journal of Physics: Condensed Matter*, vol. 22, no. 28, 2010.

[24] K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai, "IPknot: fast and accurate prediction of rna secondary structures with pseudoknots using integer programming," *Bioinformatics*, vol. 27, no. 13, pp. i85–i93, 2011.

[25] M. Bon and H. Orland, "TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots," *Nucleic Acids Research*, vol. 39, no. 14, 2011.

[26] M. Bon, C. Micheletti, and H. Orland, "McGenus: a Monte-Carlo algorithm to predict RNA secondary structures with pseudoknots," *Nucleic Acids Research*, vol. 41, no. 3, pp. 1895–1900, 2013.

[27] M. H. Bailor, X. Sun, and H. M. Al-Hashimi, "Topology links RNA secondary structure with global conformation, dynamics, and adaptation," *Science*, vol. 327, pp. 202–206, Jan 2010.

[28] V. Ambros, "Development: dicing up RNAs," *Science Signaling*, vol. 293, no. 5531, p. 811, 2001.

[29] S. R. Eddy, "Computational genomics of noncoding RNA genes," *Cell*, vol. 109, pp. 137–140, 2002.

[30] S. Buckingham, "The major world of microRNAs," *Nature*, 2003.

[31] M. Guttman and J. L. Rinn, "Modular regulatory principles of large non-coding RNAs," *Nature*, vol. 482, no. 7385, pp. 339–346, 2012.

[32] I. Martianov, A. Ramadass, A. S. Barros, N. Chow, and A. Akoulitchev, "Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript," *Nature*, vol. 445, no. 7128, pp. 666–670, 2007.

[33] K.-M. Schmitz, C. Mayer, A. Postepska, and I. Grummt, "Interaction of noncoding rna with the rdna promoter mediates recruitment of dnmt3b and

silencing of rrna genes," *Genes & Development*, vol. 24, no. 20, pp. 2264–2269, 2010.

[34] J. Lingner, T. R. Hughes, A. Shevchenko, M. Mann, V. Lundblad, and T. R. Cech, "Reverse transcriptase motifs in the catalytic subunit of telomerase," *Science*, vol. 276, no. 5312, pp. 561–567, 1997.

[35] W. Gerlach and R. Giegerich, "Guugle: a utility for fast exact matching under rna complementary rules including g-u base pairing," vol. 22, no. 6, pp. 762–764, 2006.

[36] S. H. Bernhart, H. Tafer, U. Muckstein, C. Flamm, P. F. Stadler, and I. L. Hofacker, "Partition function and base pairing probabilities of RNA heterodimers," *Algorithms for Molecular Biology*, vol. 1, no. 1, 2006.

[37] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce, "Thermodynamic analysis of interacting nucleic acid strands," *SIAM Review*, vol. 49, no. 1, pp. 65–88, 2007.

[38] A. Busch, A. S. Richter, and R. Backofen, "IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions," *Bioinformatics*, vol. 24, no. 24, pp. 2849–56, 2008.

[39] H. Chitsaz, R. Salari, S. C. Sahinalp, and R. Backofen, "A partition function algorithm for interacting nucleic acid strands," *Bioinformatics*, vol. 25, no. 12, pp. 365–73, 2009.

[40] P. G. Higgs, "RNA secondary structure: a comparison of real and random sequences," *Journal de Physique I*, vol. 3, no. 1, pp. 43–59, 1993.

[41] D. Ward, *The RNA world*, vol. 23. Elsevier, 1995.

[42] R. Bundschuh and T. Hwa, "RNA secondary structure formation: a solvable model of heteropolymer folding," *Physical Review Letter*, vol. 83, pp. 1479–1483, 1999.

[43] R. Bundschuh and T. Hwa, "Statistical mechanics of secondary structures formed by random RNA sequences," *Physical Review E*, vol. 65, no. 3, 2002.

[44] A. Pagnani, G. Parisi, and F. Ricci-Tersenghi, "Glassy transition in a disordered model for the RNA secondary structure," *Physical Review Letter*, vol. 84, pp. 2026–30, 2000.

[45] B. Maier, D. Bensimon, and V. Croquette, "Replication by a single DNA polymerase of a stretched single-stranded DNA," *Proceedings of the National Academy of Sciences*, vol. 97, pp. 12002–07, 2000.

[46] M. Tamm and S. Nechaev, "Unzipping of two random heteropolymers: Ground-state energy and finite-size effects," *Physical Review E*, vol. 78, no. 1, 2008.

[47] M. Chee, R. Yang, E. Hubbell, A. Berno, X. C. Huang, D. Stern, J. Winkler, D. J. Lockhart, M. S. Morris, and S. P. A. Fodor, "Accessing genetic information with dna high-density DNA arrays," *Science*, vol. 274, no. 5287, pp. 610–614, 1996.

[48] R. A. Gibbs, "DNA amplification by the polymerase chain reaction," *Analytical Chemistry*, vol. 62, no. 13, pp. 1202–1214, 1990.

[49] G. Valenzuela, Jesus, M. B. Francischetti, Ivo, and M. C. Ribeiro, Jos, "Purification, cloning, and synthesis of a novel salivary anti-thrombin from the Mosquito Anopheles albimanus," *Biochemistry*, vol. 38, no. 34, pp. 11209–11215, 1999.

[50] R. F. Service, "DNA chips survey an entire genome," *Science*, vol. 281, no. 5380, pp. 1122a–1122, 1998.

[51] A. Marshall and J. Hodgson, "DNA chips: An array of possibilities," *Nature Biotechnology*, vol. 16, no. 1, pp. 27–31, 1998.

[52] R. Bundschuh and T. Hwa, "An analytic study of the phase transition line in local sequence alignment with gaps," *Discrete Applied Mathematics*, vol. 104, no. 1, pp. 113–142, 2000.

[53] F. Krzakala, M. Mezard, and M. Muller, "Nature of the glassy phase of RNA secondary structure," *Europhysics Letters*, vol. 57, no. 5, pp. 752–758, 2002.

[54] S. Hui and L. H. Tang, "Ground state and glass transition of the RNA secondary structure," *The European Physical Journal B*, vol. 53, no. 1, pp. 77–84, 2006.

[55] C. Monthus and T. Garel, "Directed polymer in a random medium of dimension 1+1 and 1+3: weights statistics in the low temperature phase," *Journal of Statistical Mechanics:Theory and Experiment*, 2007.

[56] M. Lassig and K. J. Wiese, "Freezing of random RNA," *Physical Review Letter*, vol. 96, 2006.

[57] F. David and K. Wiese, "Systematic field theory of the rna glass transition," *Physical Review Letter*, vol. 98, no. 12, 2007.

[58] M. Kardar, G. Parisi, and Y.-C. Zhang, "Dynamic scaling of growing interfaces," *Physical Review Letter*, vol. 56, no. 9, pp. 889–892, 1986.

[59] K. Khanin, S. Nechaev, G. Oshanin, A. Sobolevski, and O. Vasilyev, "Ballistic deposition patterns beneath a growing Kardar-Parisi-Zhang interface," *Physical Review E*, vol. 82, no. 6, 2010.

[60] A. Y. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules*. Izd. Nauka, 1989.

[61] E. Marinari, A. Pagnani, and F. Ricci-Tersenghi, "Zero-temperature properties of RNA secondary structures," *Physical Review E*, vol. 65, no. 4, 2002.

[62] H. Orland and A. Zee, "RNA folding and large N matrix theory," *Nuclear Physics B*, vol. 620, pp. 456–476, 2002.

[63] E. Brezin, C. Itzykson, G. Parisi, and J.-B. Zuber, "Planar diagrams," *Communications in Mathematical Physics*, vol. 59, no. 1, pp. 35–51, 1978.

[64] A. A. Abrikosov and L. P. Gorkov, *Methods of quantum field theory in statistical physics*. Courier Dover Publications, 1975.

[65] R. Saito, "A proof of the completeness of the non crossed diagrams in spin 1/2 Heisenberg model," *Journal of the Physical Society of Japan*, vol. 59, no. 2, pp. 482–491, 1990.

[66] M. L. Mehta, *Random matrices*, vol. 142. Academic press, 2004.

[67] M. Bon, G. Vernizzi, H. Orland, and A. Zee, "Topological slassification of RNA structures," *Journal of Molecular Biology*, vol. 379, no. 4, pp. 900–911, 2008.

[68] I. Dumitriu and E. Rassart, "Path counting and random matrix theory," *Electronic Journal of Combinatorics*, vol. 7, no. 7, 2003.

[69] A. Edelman and N. R. Rao, "Random matrix theory," *Acta Numerica*, vol. 14, no. 1, pp. 233–297, 2005.

[70] G. Vernizzi, H. Orland, and A. Zee, "Enumeration of RNA structures by matrix models," *Physical Review Letter*, vol. 94, no. 16, 2005.

[71] M. L. Mansfield, "Efficient knot group identification as a tool for studying entanglements of polymers," *The Journal of Chemical Physics*, vol. 127, no. 24, 2007.

[72] J. Ito and D. K. Braithwaite, "Compilation and alignment of DNA polymerase sequences.," *Nucleic Acids Research*, vol. 19, no. 15, p. 4045, 1991.

[73] D. K. Braithwaite and J. Ito, "Compilation, alignment, and phylogenetic relationships of DNA polymerases," *Nucleic Acids Research*, vol. 21, no. 4, p. 787, 1993.

[74] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[75] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Molecular Biology*, vol. 147, pp. 195–197, 1981.

[76] M. S. Waterman, L. Gordon, and R. Arratia, "Phase transitions in sequence matches and nucleic acid structure," *Proceedings of the National Academy of Sciences*, vol. 84, no. 5, pp. 1239–1243, 1987.

[77] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[78] D. Sankoff, "Simultaneous solution of the RNA folding, alignment and protosequence problems," *SIAM Journal on Applied Mathematics*, vol. 45, pp. 810–825, 1985.

[79] A. Apostolico and C. Guerra, "The longest common subsequence problem revisited," *Algorithmica*, vol. 2, no. 1-4, pp. 315–336, 1987.

[80] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of the ACM*, vol. 21, no. 1, pp. 168–173, 1974.

[81] D. Gusfield, *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press, 1997.

[82] V. Chvatal and D. Sankoff, "Longest common subsequences of two random sequences," *Journal of Applied Probability*, pp. 306–315, 1975.

[83] J. G. Deken, "Some limit results for longest common subsequences," *Discrete Mathematics*, vol. 26, no. 1, pp. 17–31, 1979.

[84] M. J. Steele, "Long common subsequences and the proximity of two random strings," *SIAM Journal on Applied Mathematics*, vol. 42, no. 4, pp. 731–737, 1982.

[85] V. Dancik and M. Paterson, "Upper bounds for the expected length of a longest common subsequence of two binary sequences," *Random Structures & Algorithms*, vol. 6, no. 4, pp. 449–58, 1995.

[86] K. S. Alexander, "The rate of convergence of the mean length of the longest common subsequence," *The Annals of Applied Probability*, pp. 1074–1082, 1994.

[87] M. Kiwi, M. Loebl, and J. Matouvsek, "Expected length of the longest common subsequence for large alphabets," *Advances in Mathematics*, vol. 197, no. 2, pp. 480–498, 2005.

[88] M. Zhang and J. T. Marr, "Alignment of molecular sequences seen as random path analysis," *Journal of Theoretical Biology*, vol. 174, pp. 119–129, 1995.

[89] T. Hwa and M. Lassig, "Similarity detection and localization," *Physical Review Letter*, vol. 76, no. 14, pp. 2591–94, 1996.

[90] J. Boutet de Monvel, "Extensive simulations for longest common subsequences," *The European Physical Journal B*, vol. 7, no. 2, pp. 293–308, 1999.

[91] M. Waterman, R. Arratia, and D. Galas, "Pattern recognition in several sequences: consensus and alignment," *Bulletin Of Mathematical Biology*, vol. 46, no. 4, pp. 515–527, 1984.

[92] M. S. Waterman and M. Vingron, "Sequence comparison significance and poisson approximation," *Statistical Science*, vol. 9, pp. 367–381, 1994.

[93] D. Drasdo, T. Hwa, and M. Lassig, "Scaling laws and similarity detection in sequence alignment with gaps," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 115–141, 2000.

[94] L. Comtet, *Advanced Combinatorics: The art of finite and infinite expansions*. Springer, 1974.

[95] M. Muller, "Statistical physics of rna folding," *Physical Review E*, vol. 67, no. 2, 2003.

[96] A. Grosberg, A. Gutin, and E. Shakhnovich, "Conformational entropy of a branched polymer," *Macromolecules*, vol. 28, no. 10, pp. 3718–3727, 1995.

[97] M. Tamm and S. Nechaev, "Necklace-cloverleaf transition in associating RNA-like diblock copolymers," *Physical Review E*, vol. 75, no. 3, 2007.

[98] P.-G. de Gennes, "Statistics of branching and hairpin helices for the dAT copolymer," *Biopolymers*, vol. 6, no. 5, pp. 715–729, 1968.

[99] A. Khokhlov, "On the swelling of branched macromolecules in good solvent," *Vysokomolek. Soed. (Polymer Science USSR)*, vol. 20B, p. 543, 1978.

[100] S. R. Morgan and P. G. Higgs, "Barrier heights between ground states in a model of RNA secondary structure," *Journal of Physics A: Mathematical and General*, vol. 31, no. 14, p. 3153, 1998.

[101] J. B. De Monvel, "Mean-field approximations to the longest common subsequence problem," *Physical Review E*, vol. 62, no. 1, p. 204, 2000.

[102] S. N. Majumdar and S. Nechaev, "Exact asymptotic results for the Bernoulli matching model of sequence alignment," *Physical Review E*, vol. 72, no. 2, p. 020901, 2005.

[103] T. Kriecherbauer and J. Krug, "A pedestrian's view on interacting particle systems, KPZ universality and random matrices," *Journal of Physics A: Mathematical and Theoretical*, vol. 43, no. 40, 2010.

[104] S. Ma, *The modern theory of critical phenomena*. Izd. Mir (in Russian), 1980.

[105] M. V. Tamm, N. G. Lisachenko, I. Y. Erukhimovich, and V. A. Ivanov, "Finite size effects in the equilibrium system of ideal cyclic polymers: Theory and computer simulation," *Polymer Science*, vol. 47, p. 202212, 2005.

[106] S. K. Lando, *Lectures on generating functions*. MCCME, 2007.

[107] W. Feller, *An introduction to probability theory and its applications*. Wiley, 1968.

[108] A. A. Vladimirov, "Uncrossing matching," *Information Transmission Problems (in Russian)*, vol. 49, 2013.

[109] E. Friedgut, "Necessary and sufficient conditions for sharp thresholds and the k-SAT problem," *J. Amer. Math. Soc.*, vol. 12, no. 20, pp. 1017–54, 1999.

[110] G. Grimmett, *What is Percolation?* Springer, 1999.

[111] R. P. Feynman, "Slow electrons in a polar crystal," *Physical Review*, vol. 97, no. 3, p. 660, 1955.

[112] C. E. Shannon and W. Weaver, *A mathematical theory of communication*. American Telephone and Telegraph Company, 1948.

[113] W. Gilbert, "Origin of life: The RNA world," *Nature*, vol. 319, no. 6055, 1986.

[114] G. F. Joyce, "RNA evolution and the origins of life," *Nature*, vol. 338, no. 6212, pp. 217–224, 1989.

[115] T. M. Fink and R. C. Ball, "How many conformations can a protein remember?," *Physical Review Letter*, vol. 87, no. 19, p. 198103, 2001.

[116] A. Y. Grosberg and A. R. Chochlov, *Giant molecules: here, there, and everywhere*. World Scientific, 2011.

[117] V. Kantorovich, L., "On the mass transfer," in *Reports of Ac. USSR 37, 3 (in Russian)*, 1942.

[118] B. O. Koopman, "Search and its optimization," *The American Mathematical Monthly*, vol. 86, no. 7, pp. 527–540, 1979.

[119] V. Kuznetsov, A., H. N. I., and S. Kostevich, L., *Guide to solving problems in mathematical programming*. High School (in Russian), 1978.

[120] A. Schrijver, "A course in combinatorial optimization." Department of Mathematics, University of Amsterdam., 2013.

[121] J. Delon, J. Salomon, and A. Sobolevski, "Fast transport optimization for Monge costs on the circle," *SIAM Journal on Applied Mathematics*, vol. 70, pp. 2239–2258, Jan 2010.

[122] J. Delon, J. Salomon, and A. Sobolevski, "Local matching indicators for transport problems with concave costs," *SIAM Journal on Discrete Mathematics*, vol. 26, no. 2, pp. 801–827, 2012.

[123] R. J. McCann, "Exact solutions to the transportation problem on the line," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 455, no. 1984, pp. 1341–80, 1999.

[124] A. H. Tchen, "Inequalities for distributions with given marginals," *The Annals of Probability*, pp. 814–827, 1980.

[125] S. T. Rachev, "The Monge-Kantorovich mass transference problem and its stochastic applications," *Theory of Probability & Its Applications*, vol. 29, no. 4, pp. 647–676, 1985.

[126] W. Gangbo and R. J. McCann, "The geometry of optimal transportation," *Acta Mathematica*, vol. 177, no. 2, pp. 113–161, 1996.

[127] P. M. Postal, *Cross-over phenomena*. Holt, Rinehart and Winston New York, 1971.

[128] A. Zee, "Random matrix theory and RNA folding," *Acta Physica Polonica B*, vol. 36, 2005.

[129] E. Gudowska-Nowak, R. A. Janik, J. Jurkiewicz, and M. A. Nowak, "Infinite products of large random matrices and matrix-valued diffusion," *Nuclear Physics B*, vol. 670, no. 3, pp. 479–507, 2003.

[130] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau, T. Hao, and et al., "A map of the interactome network of the metazoan C. elegans," *Science*, vol. 303, no. 5657, pp. 540–543, 2004.

[131] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[132] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[133] S. N. Dorogovtsev and J. F. Mendes, "Evolution of networks," *Advances In Physics*, vol. 51, no. 4, pp. 1079–1187, 2002.

[134] R. Diestel, "Graph theory," 2005.

[135] P. Kaluza, *Evolutionary Engineering of Complex Functional Networks*. PhD thesis, Technische Universitat Berlin, 2007.

[136] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, 2005.

[137] A. Wagner and D. A. Fell, "The small world inside large metabolic networks," *Proceedings of the Royal Society of London, Series B: Biological Sciences*, vol. 268, no. 1478, pp. 1803–1810, 2001.

[138] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási, "Functional and topological characterization of protein interaction networks," *Proteomics*, vol. 4, no. 4, pp. 928–942, 2004.

[139] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–55, 2002.

[140] B. A. Huberman and L. A. Adamic, "Internet: growth dynamics of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 131–131, 1999.

[141] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[142] A. Maayan, S. L. Jenkins, S. Neves, A. Hasseldine, E. Grace, B. Dubin-Thaler, N. J. Eungdamrong, G. Weng, P. T. Ram, J. J. Rice, and et al., "Formation of regulatory patterns during signal propagation in a mammalian cellular network," *Science*, vol. 309, no. 5737, pp. 1078–83, 2005.

[143] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, pp. 910–913, 2002.

[144] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, no. 5663, pp. 1538–42, 2004.

[145] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[146] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[147] N. Barkal and S. Leibler, "Robustness in simple biochemical networks," *Nature*, vol. 387, no. 6636, pp. 913–917, 1997.

[148] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, *et al.*, "Functional profiling of the

saccharomyces cerevisiae genome," *Nature*, vol. 418, no. 6896, pp. 387–391, 2002.

[149] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[150] M. R. Said, T. J. Begley, A. V. Oppenheim, D. A. Lauffenburger, and L. D. Samson, "Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae," *Proceedings of the National Academy of Sciences*, vol. 101, no. 52, pp. 18006–18011, 2004.

[151] B. Vogelstein, D. Lane, and A. J. Levine, "Surfing the p53 network," *Nature*, vol. 408, no. 6810, pp. 307–310, 2000.

[152] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47–C52, 1999.

[153] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 1128–33, 2003.

[154] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. Hao, C. Ooi, B. Godwin, E. Vitols, and et al., "A protein interaction map of Drosophila melanogaster," *Science*, vol. 302, no. 5651, pp. 1727–1736, 2003.

[155] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[156] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 21, pp. 12123–12128, 2003.

[157] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and et al., "Evidence for dynamically organized modularity in the yeast protein–protein interaction network," *Nature*, vol. 430, no. 6995, pp. 88–93, 2004.

[158] G. Balazsi, A. L. Barabasi, and Z. N. Oltvai, "Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli," *Proceedings of the National Academy of Sciences*, vol. 102, no. 22, pp. 7841–46, 2005.

[159] S. Wuchty, Z. N. Oltvai, and A.-L. Barabasi, "Evolutionary conservation of motif constituents in the yeast protein interaction network," *Nature genetics*, vol. 35, no. 2, pp. 176–179, 2003.

[160] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit, "Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction," *Proceedings of the National Academy of Sciences*, vol. 101, no. 16, pp. 5934–39, 2004.

[161] L. V. Zhang, O. D. King, S. L. Wong, D. S. Goldberg, A. H. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F. P. Roth, "Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network," *Journal of biology*, vol. 4, p. 6, 2005.

[162] G. C. Conant and A. Wagner, "Convergent evolution of gene circuits," *Nature Genetics*, vol. 34, no. 3, pp. 264–266, 2003.

[163] M. E. Csete and J. C. Doyle, "Reverse engineering of biological complexity," *Science*, vol. 295, no. 5560, pp. 1664–69, 2002.

[164] S. Mangan and U. Alon, "Structure and function of the feed-forward loop network motif," *Proceedings of the National Academy of Sciences*, vol. 100, no. 21, pp. 11980–85, 2003.

[165] J. A. Papin, T. Hunter, B. O. Palsson, and S. Subramaniam, "Reconstruction of cellular signalling networks and analysis of their properties," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 2, pp. 99–111, 2005.

[166] S. Wasserman, *Social network analysis: Methods and applications*, vol. 8. Cambridge University Press, 1994.

[167] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, and et al., "The genetic landscape of a cell," *Science*, vol. 327, pp. 425–431, Jan 2010.

[168] Y. Li, O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu, J. A. G. Riksen, E. Hazendonk, P. Prins, R. H. A. Plasterk, R. C. Jansen, and et al., "Mapping determinants of gene expression plasticity by genetical genomics in C. elegans," *PLOS Genetics*, vol. 2, no. 12, 2006.

[169] O. A. Alvarez, T. Jagret, S. A. L. M. Kooijman, and J. E. Kammenga, "Responses to stress of Caenorhabditis elegans populations with different reproductive strategies," *Functional Ecology*, vol. 19, no. 4, pp. 656–664, 2005.

[170] S. R. Wicks, R. T. Yeh, W. R. Gish, R. H. Waterston, and R. H. Plasterk, "Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map," *Nature Genetics*, vol. 28, no. 2, pp. 160–164, 2001.

[171] A. Vinuela, L. B. Snoek, J. A. G. Riksen, and J. E. Kammenga, "Genome-wide gene expression regulation as a function of genotype and age in C. elegans," *Genome Research*, vol. 20, no. 7, pp. 929–937, 2010.

[172] A. Vinuela, L. B. Snoek, J. A. G. Riksen, and J. E. Kammenga, "Aging uncouples heritability and expression-QTL in Caenorhabditis elegans," *G3: Genes, Genomes, Genetics*, vol. 2, no. 5, pp. 597–605, 2012.

[173] Y. Li, R. Breitling, L. B. Snoek, K. J. van der Velde, M. A. Swertz, J. Riksen, R. C. Jansen, and J. E. Kammenga, "Global genetic robustness of the alternative splicing machinery in Caenorhabditis elegans," *Genetics*, vol. 186, no. 1, pp. 405–410, 2010.

[174] E. J. Capra, S. M. Skrovanek, and L. Kruglyak, "Comparative developmental expression profiling of two C. elegans isolates," *PLOS ONE*, vol. 3, no. 12, 2008.

[175] J. E. Kammenga, A. Doroszuk, J. A. Riksen, E. Hazendonk, L. Spiridon, A.-J. Petrescu, M. Tijsterman, R. H. Plasterk, and J. Bakker, "A Caenorhabditis elegans wild type defies the temperature–size rule owing to a single nucleotide polymorphism in tra-3," *PLoS genetics*, vol. 3, no. 3, p. e34, 2007.

[176] E. Gutteling, A. Doroszuk, J. Riksen, Z. Prokop, J. Reszka, and J. Kammenga, "Environmental influence on the genetic correlations between life-history traits in Caenorhabditis elegans," *Heredity*, vol. 98, no. 4, pp. 206–213, 2007.

[177] M. de Bono and C. I. Bargmann, "Natural variation in a neuropeptide y receptor homolog modifies social behavior and food response in c. elegans," *Cell*, vol. 94, no. 5, pp. 679–689, 1998.

[178] M. Elvin, L. B. Snoek, M. Frejno, U. Klemstein, J. E. Kammenga, and G. B. Poulin, "A fitness assay for comparing rnai effects across multiple C. elegans genotypes," *BMC Genomics*, vol. 12, no. 1, p. 510, 2011.

[179] E. Gutteling, J. Riksen, J. Bakker, and J. Kammenga, "Mapping phenotypic plasticity and genotype-environment interactions affecting life-history traits in Caenorhabditis elegans," *Heredity*, vol. 98, no. 1, pp. 28–37, 2006.

[180] M. Rodriguez, L. B. Snoek, J. A. G. Riksen, R. P. Bevers, and J. E. Kammenga, "Genetic variation for stress-response hormesis in C. elegans lifespan," *Experimental Gerontology*, vol. 47, no. 8, pp. 581–587, 2012.

[181] J. L. Anderson, L. Albergotti, S. Proulx, C. Peden, R. B. Huey, and P. C. Phillips, "Thermal preference of Caenorhabditis elegans: a null model and empirical tests," *Journal of Experimental Biology*, vol. 210, no. 17, pp. 3107–3116, 2007.

[182] A. Vinuela, L. B. Snoek, J. A. Riksen, and J. E. Kammenga, "Gene expression modifications by temperature-toxicants interactions in Caenorhabditis elegans," *PLOS ONE*, vol. 6, no. 9, p. e24676, 2011.

[183] J. E. Kammenga, P. C. Phillips, M. De Bono, and A. Doroszuk, "Beyond induced mutants: using worms to study natural variation in genetic pathways," *Trends in Genetics*, vol. 24, no. 4, pp. 178–185, 2008.

[184] "Public functional genomics data: Gene Expression Omnibus." http://www.ncbi.nlm.nih.gov/geo.

[185] A. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev,

D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush, "TM4: a free, open-source system for microarray data management and analysis," *Biotechniques*, vol. 34, pp. 374–8, 2003.

[186] L. B. Snoek, K. J. Van der Velde, D. Arends, Y. Li, A. Beyer, M. Elvin, J. Fisher, A. Hajnal, M. O. Hengartner, G. B. Poulin, and et al., "WormQTL–public archive and analysis web portal for natural variation data in Caenorhabditis spp," *Nucleic Acids Research*, vol. 41, no. D1, pp. D738–D743, 2012.

[187] "STRING - known and predicted protein-protein interactions." http://string-db.org/.

[188] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, and et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. Database, pp. D561–D568, 2010.

[189] "WormBase WS220." http://www.wormbase.org.

[190] "Probabilistic Functional Gene Network of Caenorhabditis elegans - wormnet." http://www.functionalnet.org/wormnet/.

[191] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical Review Letter*, vol. 87, no. 19, 2001.

[192] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, and et al., "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–36, 2006.

[193] P. Syntichaki and N. Tavernarakis, "Signaling pathways regulating protein synthesis during ageing," *Experimental Gerontology*, vol. 41, no. 10, pp. 1020–1025, 2006.

[194] B. Hamilton, "A systematic rnai screen for longevity genes in c. elegans," *Genes & Development*, vol. 19, no. 13, pp. 1544–1555, 2005.

[195] A. Dillin, D. K. Crawford, and C. Kenyon, "Timing requirements for Insulin/IGF-1 signaling in C.elegans," *Science*, vol. 298, no. 5594, pp. 830–834, 2002.

[196] C. Kenyon, "The first long-lived mutants: discovery of the Insulin/IGF-1 pathway for ageing," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, no. 1561, pp. 9–16, 2010.

[197] C. T. Murphy, S. A. McCarroll, C. I. Bargmann, A. Fraser, R. S. Kamath, J. Ahringer, H. Li, and C. Kenyon, "Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans," *Nature*, vol. 424, no. 6946, pp. 277–283, 2003.

[198] M. Hanazawa, I. Kawasaki, H. Kunitomo, K. Gengyo-Ando, K. L. Bennett, S. Mitani, and Y. Iino, "The Caenorhabditis elegans eukaryotic initiation factor 5A homologue, IFF-1, is required for germ cell proliferation, gametogenesis and localization of the P-granule component PGL-1," *Mechanisms of Development*, vol. 121, no. 3, pp. 213–224, 2004.

[199] R. Schrader, C. Young, D. Kozian, R. Hoffmann, and F. Lottspeich, "Temperature-sensitive eif5a mutant accumulates transcripts targeted to the nonsense-mediated decay pathway," *Journal of Biological Chemistry*, vol. 281, 2006.

[200] M. Kaeberlein, "Spermidine surprise for a long life," *Nature Cell Biology*, vol. 11, no. 11, pp. 1277–1278, 2009.

[201] R. Kaur-Sawhney, S. Liu-Mei, E. F. Hector, and W. G. Arthur, "Relation of polyamine synthesis and titer to aging and senescence in oat leaves," *Plant Physiology*, vol. 69, pp. 405–410, 1982.

[202] T. Eisenberg, H. Knauer, A. Schauer, S. Buttner, C. Ruckenstuhl, D. Carmona-Gutierrez, J. Ring, S. Schroeder, C. Magnes, and L. Antonacci, "Induction of autophagy by spermidine promotes longevity," *Nature Cell Biology*, vol. 11, no. 11, pp. 1305–14, 2009.

[203] T. F. Menne, B. Goyenechea, N. Sanchez-Puig, C. C. Wong, L. M. Tonkin, P. J. Ancliff, R. L. Brost, M. Costanzo, C. Boone, and A. J. Warren, "The

Shwachman-Bodian-Diamond syndrome protein mediates translational activation of ribosomes in yeast," *Nature Genetics*, vol. 39, no. 4, pp. 486–495, 2007.

[204] O. Vasieva, "Role of Shwachman-Bodian-Diamond syndrome protein in translation machinery and cell chemotaxis: a comparative genomics approach," *Advances and Applications in Bioinformatics and Chemistry*, p. 43, 2011.

[205] A. V. Samuelson, C. E. Carr, and G. Ruvkun, "Gene activities that mediate increased life span of C. elegans insulin-like signaling mutants," *Genes & Development*, vol. 21, no. 22, pp. 2976–2994, 2007.

[206] R. E. Navarro, E. Y. Shim, Y. Kohara, A. Singson, and T. K. Blackwell, "Cgh-1, a conserved predicted RNA helicase required for gametogenesis and protection from physiological germline apoptosis in C. elegans," *Development*, vol. 128, pp. 322–32, 2001.

[207] F. Simmer, C. Moorman, A. M. van der Linden, E. Kuijk, P. V. van den Berghe, R. S. Kamath, A. G. Fraser, J. Ahringer, and R. H. A. Plasterk, "Genome-wide RNA of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions," *PLoS Biology*, vol. 1, no. 1, p. e2, 2003.

[208] J. Halaschek-Wiener, J. S. Khattra, S. McKay, A. Pouzyrev, J. M. Stott, G. S. Yang, R. A. Holt, S. J. M. Jones, M. A. Marra, A. R. Brooks-Wilson, and D. L. Riddle, "Analysis of long-lived C. elegans daf-2 mutants using serial analysis of gene expression," *Genome Research*, vol. 15, no. 5, pp. 603–615, 2005.

[209] I. Lee, B. Lehner, C. Crombie, W. Wong, A. G. Fraser, and E. M. Marcotte, "A single gene network accurately predicts phenotypic effects of gene perturbation in caenorhabditis elegans," *Nature Genetics*, vol. 40, no. 2, pp. 181–188, 2008.

[210] J. E. Irazoqui, J. M. Urbach, and F. M. Ausubel, "Evolution of host innate defence: insights from Caenorhabditis elegans and primitive invertebrates," *Nature Reviews Immunology*, vol. 10, no. 1, pp. 47–58, 2010.

[211] D. R. Lorenz, C. R. Cantor, and J. J. Collins, "A network biology approach to aging in yeast," *Proceedings of the National Academy of Sciences*, vol. 106, no. 4, pp. 1145–1150, 2009.

[212] C. Ye, S. J. Galbraith, J. C. Liao, and E. Eskin, "Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast," *PLoS Computational Biology*, vol. 5, no. 3, p. e1000311, 2009.

[213] J. J. B. Keurentjes, J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, L. B. Snoek, A. J. M. Peeters, D. Vreugdenhil, M. Koornneef, and R. C. Jansen, "Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci," *Proceedings of the National Academy of Sciences*, vol. 104, no. 5, pp. 1708–1713, 2007.

[214] I. R. Terpstra, L. B. Snoek, J. J. Keurentjes, A. J. Peeters, and G. Van den Ackerveken, "Regulatory network identification by genetical genomics: signaling downstream of the Arabidopsis receptor-like kinase ERECTA," *Plant Physiology*, vol. 154, no. 3, pp. 1067–1078, 2010.

[215] M. A. West, K. Kim, D. J. Kliebenstein, H. van Leeuwen, R. W. Michelmore, R. Doerge, and D. A. Clair, "Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis," *Genetics*, vol. 175, no. 3, pp. 1441–1450, 2007.

[216] M. V. Rockman, S. S. Skrovanek, and L. Kruglyak, "Selection at linked sites shapes heritable phenotypic variation in C. elegans," *Science*, vol. 330, no. 6002, pp. 372–376, 2010.

[217] B. Stigler and H. M. Chamberlin, "A regulatory network modeled from wild-type gene expression data guides functional predictions in Caenorhabditis elegans development," *BMC Systems Biology*, vol. 6, no. 1, p. 77, 2012.

[218] W. Wang and X. Zhang, "Network-based group variable selection for detecting expression quantitative trait loci (eQTL)," *Bioinformatics*, vol. 12, no. 1, p. 269, 2011.

[219] D. Stanley, N. S. Watson-Haigh, C. J. Cowled, and R. J. Moore, "Genetic architecture of gene expression in the chicken," *BMC Genomics*, vol. 14, no. 1, p. 13, 2013.

[220] S. Suthram, A. Beyer, R. M. Karp, Y. Eldar, and T. Ideker, "eQED: an efficient method for interpreting eQTL associations using protein networks," *Molecular systems biology*, vol. 4, no. 1, 2008.

[221] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[222] V. A. Avetisov, S. K. Nechaev, and A. B. Shkarin, "On the motif distribution in random block-hierarchical networks," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 24, pp. 5895–02, 2010.

[223] M. Muller and K. C. Daoulas, "Calculating the free energy of self-assembled structures by thermodynamic integration," *The Journal of chemical physics*, vol. 128, 2008.

[224] Y. Deng and B. Roux, "Computation of binding free energy with molecular dynamics and grand canonical monte carlo simulations," *The Journal Of Chemical Physics*, vol. 128, 2008.

[225] K. A. Connors, *Chemical kinetics: the study of reaction rates in solution*. Wiley. com, 1990.

[226] L. Landau and E. Lifshitz, *Statistical Physics*. Pergamon: Oxford, 1981.

[227] M. Eigen, "Selforganization of matter and the evolution of biological macromolecules," *Naturwissenschaften*, vol. 58, no. 10, pp. 465–523, 1971.

[228] L. Peliti, "Quasispecies evolution in general mean-field landscapes," *Europhysics Letters*, vol. 57, no. 5, p. 745, 2002.

[229] S. Galluccio, R. Graber, and Y.-C. Zhang, "Diffusion on a hypercubic lattice with pinning potential: exact results for the error-catastrophe problem in biological evolution," *arXiv preprint cond-mat/9601088*, 1996.

[230] F. Slanina, "Selective advantage of topological disorder in biological evolution," tech. rep., 2002.