European Biophysics Journal (2019) 48:413–424
https://doi.org/10.1007/s00249-019-01366-3

REVIEW

From biophysics to 'omics and systems biology

Marko Djordjevic[1] · Andjela Rodic[1,2] · Stefan Graovac[1,2]

Received: 12 October 2018 / Revised: 12 February 2019 / Accepted: 3 April 2019 / Published online: 9 April 2019

Abstract

Recent decades brought a revolution to biology, driven mainly by exponentially increasing amounts of data coming from "'omics" sciences. To handle these data, bioinformatics often has to combine biologically heterogeneous signals, for which methods from statistics and engineering (e.g. machine learning) are often used. While such an approach is sometimes necessary, it effectively treats the underlying biological processes as a black box. Similarly, systems biology deals with inherently complex systems, characterized by a large number of degrees of freedom, and interactions that are highly non-linear. To deal with this complexity, the underlying physical interactions are often (over)simplified, such as in Boolean modelling of network dynamics. In this review, we argue for the utility of applying a biophysical approach in bioinformatics and systems biology, including discussion of two examples from our research which address sequence analysis and understanding intracellular gene expression dynamics.

Keywords Systems biology · Sequence analysis · Intracellular dynamics · Biophysical modelling · Gene expression regulation

Introduction

The revolution in molecular biology brought a need to analyse previously unprecedented amounts of data, demanding that biology makes a transition from a more qualitative science to a genuine data science. Two significant challenges are: (i) to analyse an exponentially increasing amount of biological sequences stored in databases, and (ii) to predict and interpret in vivo expression dynamics of molecular species inside a cell. To meet these challenges it is clear that advanced quantitative methods are necessary, in terms of both analysing data and formulating theoretical models that can make falsifiable predictions based on available experimental data.

Biophysics traditionally deals with quantitative, i.e. physical representations of complex biological processes, which makes it a natural candidate for providing a framework for both modelling, and developing data analysis methods, in biological problems. This is most evident in systems biology, where one often has to understand how interactions between individual system components (e.g. proteins and their binding sites on DNA in gene circuits) lead to complex system behaviour. Thereby, one typically deals with a large number of components (degrees of freedom) inside a cell, e.g. there are a number of transcription factors (proteins that bind to DNA and regulate gene expression) in a cell, each of them present in many copies. Therefore, statistical thermodynamics becomes a natural framework for dealing with such systems. Second, another characteristic is that input–output relationships for intracellular variables are often highly non-linear. An ubiquitous reason for this non-linearity is saturation of receptors with ligands, i.e. binding occupancy of a ligand to a receptor has a sigmoidal dependence on ligand concentration. In addition, a large cooperativity is often present, where a larger number of weaker interactions contribute to overall binding affinity (Sneppen and Zocchi 2005). Such cooperativity leads to a sharp switch-like response of the system, and consequently to a strong non-linearity in the system, which has to be accounted for by methods from non-linear dynamics. Both statistical thermodynamics and

Special Issue: Regional Biophysics Conference 2018.

✉ Marko Djordjevic
dmarko@bio.bg.ac.rs

[1] Faculty of Biology, Institute of Physiology and Biochemistry, University of Belgrade, Belgrade, Serbia

[2] Interdisciplinary PhD Program in Biophysics, University of Belgrade, Belgrade, Serbia

non-linear dynamics are a standard toolbox used in theoretical biophysics (Phillips et al. 2012).

On the other hand, in different “’omics” sciences (genomics, proteomics) one often has to deal with analysis of exponentially increasing number of stored sequences, or to combine biologically heterogeneous signals to make predictions. To achieve this, methods from computer science and mathematical statistics (e.g. machine learning) are often employed. However, an explicit underlying biological mechanism thereby often remains hidden within a “black box”. On the other hand, when applying a biophysical approach to bioinformatics problems, an explicit (biophysical) understanding of the underlying biological process is used to improve performance of bioinformatics algorithms. Such an approach has proved to be highly beneficial in a number of problems, from modeling structure and folding kinetics of RNA and proteins, to analysing gene expression and predicting regulatory elements (Berg et al. 2004; Berg and von Hippel 1988a; Djordjevic 2013; Djordjevic and Sengupta 2006; Djordjevic et al. 2003; Lee et al. 2017; Locke and Morozov 2015; Mustonen et al. 2008; Sengupta et al. 2002; Vilar 2010; Vilar and Saiz 2013; Zuker et al. 1999). In practice, to deal with the complex nature of the data, when a complete physical model of the underlying process cannot be developed, one combines a biophysical understanding of the relevant process, with appropriate methods from mathematical statistics.

The main goal of this review was to underline the transition of biology from a traditionally more qualitative to a quantitative science, where biophysics has a significant role. Such a transition can arguably be best seen in problems from sequence analysis, and in analysing intracellular expression dynamics, where a massive amount of data, or complex physical interactions, are found. Consequently, in this review we discuss the role of a (bio) physical approach to sequence analysis and intracellular dynamics. We will first discuss two modes of research in biology, the one traditionally present in biology, and the one needed by quantitative measurements, where we will particularly emphasise the complex nature of sequence and dynamics measurements data in biology. We will then follow by an overview of two examples from our research, one in sequence analysis (in particular, regulatory element prediction), and one in analysis of dynamics of gene expression regulation.

Two modes of research in biology

Ways of thinking and researching have evolved with the development of modern techniques that have rapidly become necessary for biological systems analysis. A traditional mode of research in biology is shown in Fig. 1a. One typically starts from a given hypothesis, which is encapsulated by a qualitative model. Hypothesis is then experimentally tested, usually by imposing an appropriate “yes/no” question. Depending on the experimental test, the qualitative model is appropriately revised, and the cycle can be repeated as needed.

Therefore, models are clearly present in traditional biological research, but these models are qualitative rather than quantitative, and there is no evident separation between theory and experiment (Phillips et al. 2012). In fact, a nice example of a (traditional) model in molecular biology is the model of transcription, during which RNA is synthesized based on a DNA template. Physically, transcription is exhibited by a complex molecular machine called RNA polymerase (RNAP). RNAP binds to DNA and synthesizes RNA as it moves along the DNA chain. This mechanism can be even further simplified to a conceptual model shown in Fig. 2a. As can be seen here, the general principle behind this model is complementarity between the synthesized RNA and one of the two DNA strands which is used as a template. Moreover, the model of transcription can be incorporated in a wider model known as the Central Dogma of Molecular Biology, which emphasises that information flows from DNA to protein through RNA (Fig. 2b) (Alberts et al. 2014).

However, a significant change in modern biology was brought by the so-called “genome revolution”, where advances in technology allowed highly efficient DNA and RNA sequencing. This then led to the generation of

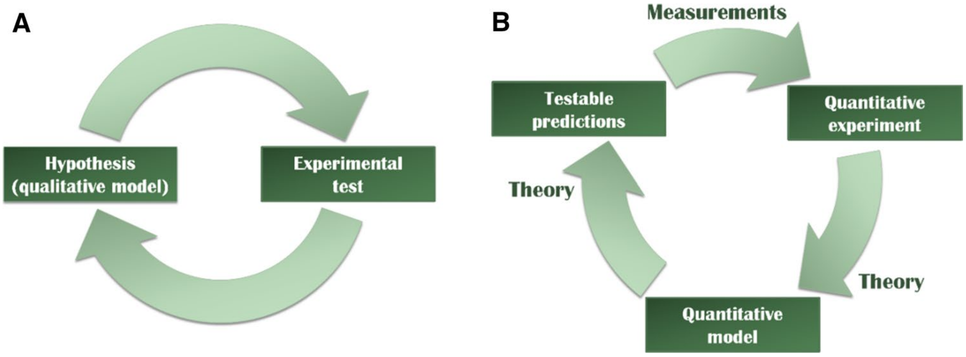


Fig. 1 Role of modelling in biology. **a** Scheme of traditional research in biology. Note no clear separation between theory and experiment. **b** A revised scheme, increasingly characterizing modern biology research, imposed by quantitative measurements. Such scheme characterizes traditionally quantitative sciences (e.g. physics), where there is a clear separation between theory and experiment

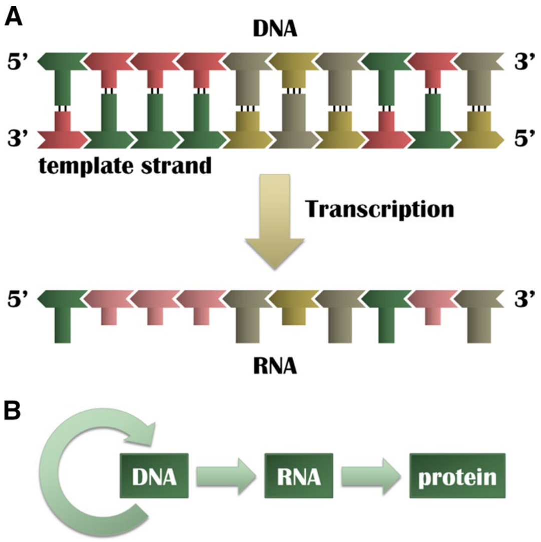


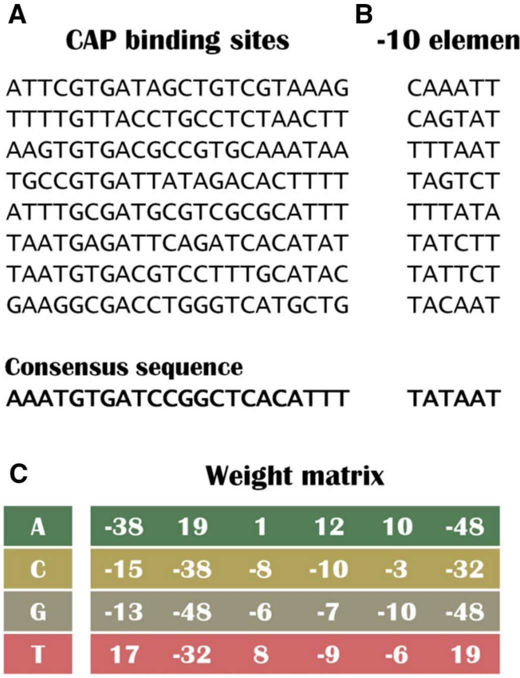
Fig. 2 **a** A conceptual model of transcription, emphasizing that RNA is synthesized based on one of the two DNA strands used as a template. Modified from (Alberts et al. 2014). **b** The Central Dogma of Molecular Biology: DNA is copied during replication, RNA is synthesized from DNA template during transcription, and proteins are synthesized from RNA, defining a flow of information from DNA to proteins

a very large amount of data. For example, currently, there are ~ 80,000 completely or partially sequenced genomes stored in the GeneBank. To get a sense of the quantity of data, even a virus genome is ~ 50,000 bps long, while sequences of bacterial and human genomes are longer by two and five orders of magnitude, respectively. To illustrate this, if the *E. coli* genome was written within a book, it would take 1000 pages to write it down—to write a human genome, 1000 of such books would be needed (Phillips et al. 2012). It is, therefore, clear that informatics resources are necessary for the storage and systematisation of such large quantities of data. Perhaps even more importantly, mathematical methods are necessary to analyse and extract information from these sequences. Such information resources and quantitative methods for analysis of data in molecular biology are known as bioinformatics.

Moreover, the genome revolution brought not only a large amount of sequenced nucleotides, but also an ability to simultaneously measure expression of a large number of genes (e.g. all genes in the genome). Such measurements have been traditionally done on DNA microarrays (chips) where each spot on a microarray corresponds to one gene, and measurements can be done across different time points and experimental conditions; nowadays such studies will more often make use of RNA-seq measurements. In such microarray analysis there is at least one number associated with each of many spots on a microarray. Consequently, to test a biological hypothesis against such measurements, it is clear that a quantitative (rather than qualitative) model has to be used—that is, one needs to compare numbers from a quantitative model with numbers from quantitative experiments. Moreover, recent advances, in particular single-cell techniques, allow also assessing stochastic effects in gene expression, instead of just the population average, which opens up possibilities for testing even more complex hypotheses.

Following the reasoning exposed above, one then comes to a research scheme which is becoming more and more appropriate to modern biology (Fig. 1b). Here, hypothesis together with underlying assumptions is now encapsulated within an appropriate quantitative (theoretical) model. The model leads to falsifiable predictions (often quantitative) that can now be compared with quantitative measurements, and such comparison may lead to a revised theoretical model. The cycle can then be repeated until a theoretical model shows a satisfactory agreement with experimental data. In this research mode, we get a clear division of labour between theory and experiment, which is a situation reminiscent of traditionally more quantitative sciences such as physics.

Beyond the genome revolution, a major problem in biology, which is becoming increasingly important with advancements in experimental biophysical techniques, is to understand in vivo expression dynamics of macromolecules (RNA, proteins) within a cell (De Jong and Geiselmann 2014; Longo and Hasty 2006; Ohno et al. 2014). This is because cell conditions change with time, e.g. due to changing external conditions or due to passing through different stages of the cell cycle or organism development (Klumpp et al. 2009). Understanding in vivo intracellular dynamics is, however, related with difficulties, both from the experimental and the theoretical side. That is, while measuring dynamics of macroscopic objects may be relatively easy, measuring in vivo dynamics of molecules within a cell requires advanced experimental techniques, and is typically related with significant technical difficulties, such as a necessity to synchronise the cell population (Morozova et al. 2016). Similarly, modelling such dynamics is generally also complicated, since, as explained above, the relevant systems employ a large number of degrees of freedom, with interactions that are typically highly non-linear and numbers of molecules of given species that can be very small. Such modelling then requires advanced techniques from theoretical biophysics, such as statistical physics, non-linear dynamics and stochastic modelling. However, despite these difficulties, a joint analysis of intracellular gene expression dynamics, through both experiment and theory, is highly valuable due to two main reasons. First, in those cases where direct measurements of protein dynamics within a cell are already available, theoretical predictions allow interpreting those data (Morozova et al. 2016). Second, when such measurements are not available, one can use modelling to infer relevant parameters, e.g. from equilibrium in vitro measurements, and then use them to make predictions of the relevant in vivo dynamics, therefore gaining valuable understanding of the system function (Bogdanova et al. 2008; Klimuk et al. 2018). Due to this, in this review, we will argue that there is a close connection between biophysical modelling and methods for data analysis in molecular biology (a.k.a. bioinformatics). We will also argue that quantitative (biophysical) understanding of the underlying biological processes can largely improve the analysis and interpretation of the relevant data. In the reminder of the text, we will overview two examples from our research, one from sequence analysis (Djordjevic et al. 2017), and the other from modelling gene expression dynamics (Morozova et al. 2016; Rodic et al. 2017), which illustrate these points.



A

B

CAP binding sites	-10 element
ATTCGTGATAGCTGTCGTAAAG	CAAATT
TTTTGTTACCTGCCTCTAACTT	CAGTAT
AAGTGTGACGCCGTGCAAATAA	TTTAAT
TGCCGTGATTATAGACACTTTT	TAGTCT
ATTTGCGATGCGTCGCGCATTT	TTTATA
TAATGAGATTCAGATCACATAT	TATCTT
TAATGTGACGTCCTTTGCATAC	TATTCT
GAAGGCGACCTGGGTCATGCTG	TACAAT
Consensus sequence	
AAATGTGATCCGGCTCACATTT	TATAAT

C

Weight matrix

A	-38	19	1	12	10	-48
C	-15	-38	-8	-10	-3	-32
G	-13	-48	-6	-7	-10	-48
T	17	-32	8	-9	-6	19

Fig. 3 Selected sequences of experimentally determined binding sites from RegulonDB database (Gama-Castro et al. 2011) for **a** CAP transcription factor and **b** σ70. **c** Weight (energy) matrix is inferred from experimentally determined TFBS from RegulonDB (Gama-Castro et al. 2011) database

Sequence analysis: predicting targets of transcription

A classical and still unresolved problem in bioinformatics is predicting a set of genes directly regulated by a given transcription factor (Robison et al. 1998; Stormo 2000). Such genes are also called transcription targets or direct target genes for that transcription factor. To understand how these predictions are made, we will go back to the Central Dogma of Molecular Biology (Fig. 2b). In contrast to the common representation, the process of transcription shown within the Central Dogma is, in fact, typically highly regulated, often via DNA binding proteins called transcription factors. Transcription factors bind to the sequences upstream of the genes and regulate their expression. A classic example of gene expression control by transcription factors is the Lac operon, which encodes the enzymes that degrade sugar lactose within a bacterial cell (Phillips et al. 2012). This operon is under control of two transcription factors, one acting as an activator (CAP, also called CRP), and the other acting as a repressor. So for example, when glucose, which is a primary source of food for bacteria, is absent, but sugar lactose (that has to be degraded) is present, bacteria are "hungry" and the sugar lactose has to be degraded. In such a case, CAP is bound to promoter DNA, but the repressor is not bound, so RNAP can bind to DNA and the downstream gene is transcribed. Transcription factors, such as CAP, bind to short stretches of DNA (typically ~ 20 bps or shorter) which are called transcription factor binding sites (TFBS). Moreover, each transcription factor typically regulates not one, but a number of genes, so it may have many TFBS. Some examples of TFBS for the activator CAP are shown in Fig. 3a. As one can see, these binding sites can be described as a set of similar "words". While there is clearly a common pattern in these binding sites, they are also highly variable, e.g. there are mostly "G"-s at the fifth position in the binding sites, but the bases that appear at most other positions are much less conserved. It is exactly due to this high variability that one typically gets a high number of false positives (i.e. predicted TFBS which, in reality, they are not) in TFBS predictions.

Actually, an even higher variability in binding site sequences is displayed for RNAP. The part of RNAP which binds DNA is the σ factor (Helmann and Chamberlin 1988). Most genes are transcribed by the housekeeping σ factor (σ70), while alternative σ factors transcribe genes under more specific conditions (e.g. that of stress and stringency) (Feklístov et al. 2014; Guzina and Djordjevic 2015; Paget and Helmann 2003). Binding sites for σ70, which are shown in Fig. 3b, consist of several parts that are called promoter elements (Feklístov et al. 2014). For example, if one takes all − 10 elements (elements are denoted according to their typical distance from the transcription start site), they can be represented by one general consensus sequence, but one can notice that actual instances of − 10 elements are quite different from this consensus. Consequently, predicting TFBS is a highly non-trivial problem, and the prediction methods can be divided into two related approaches: (i) Unsupervised search (de-novo motif discovery), which corresponds to identifying binding sites of TF with a priori unknown specificity; and (ii) supervised search, where one starts from a set of experimentally known TBFS and aims to identify new instances of the motif using some recognition rule. We below concentrate on supervised motif recognition, while for unsupervised motif search, one can consult some of the relevant reviews (Bulyk 2003; Das and Dai 2007; Jayaram et al. 2016; Sandve and Drablos 2006).

In supervised search, a standard method to predict TFBS is to form an energy matrix, where elements of this matrix can be interpreted in terms of contributions to protein–DNA binding energy (Berg and von Hippel 1987, 1988b; Djordjevic et al. 2003; Stormo and Fields 1998; Stormo and Zhao 2010). That is, a larger weight, corresponding to a certain base at a certain position in the sequence, is associated with larger contribution to the binding energy due to more frequent presence of this base in binding sites of the given transcription factor. Inferring elements of energy matrix in general leads to a non-linear optimization problem (such as quadratic programming), while in the limit where binding probability can be approximated by a Boltzmann distribution, it reduces to the log ratio of frequencies in the experimentally observed set and in the background model (Djordjevic et al. 2003). Energy matrices have been inferred through different approaches (Djordjevic and Sengupta 2006; Djordjevic et al. 2003; Locke and Morozov 2015; Stormo and Fields 1998; Stormo and Zhao 2010; Vilar 2010; Vilar and Saiz 2013), but searches of binding sites with such energy matrices (and searches of TFBS in general) typically result in a high number of false positives (Stormo 2000; Towsey et al. 2008). For example, in the case of σ70 one typically gets two orders of magnitude more predictions than what would be normally expected (Robison et al. 1998). Some number of those predicted sites may turn out to be true positives, and a significant fraction of false positives may be due to kinetics of promoter recognition—i.e. so-called poised promoters, where RNAP binds strongly, but forms the open complex too slowly to achieve functional transcription (Djordjevic 2013).

One is, however, often interested not so much in predicting individual TFBS, but more in which genes are regulated by a given transcription factor (i.e. transcription targets/direct regulatory targets). A classical approach to the problem of predicting transcription targets is as follows: One first assembles some experimental examples of TFBS from either in vivo experiments (ChIP-seq, ChIP-chip) (Park 2009; Wade et al. 2007), or in vitro experiments (Protein Binding Microarrays, HT-SELEX, SELEX, DNA footprinting, primer extension) (Bulyk 2006; Jagannathan et al. 2006; Newburger and Bulyk 2009; Roulet et al. 2002). From these experimentally assembled TFBS one then infers weight (energy) matrix using some of the many approaches that were previously developed. These methods are based either on statistical methods (Bulyk 2003; Favorov et al. 2005; Levitsky et al. 2014; Stormo 2000) or on a biophysical approach (Djordjevic and Sengupta 2006; Djordjevic et al. 2003; Locke and Morozov 2015; Stormo and Fields 1998; Stormo and Zhao 2010), where using biophysical models generally shows a significantly larger accuracy (Djordjevic and Sengupta 2006; Djordjevic et al. 2003; Homsi et al. 2009). Once the energy matrix is found, one next scans the regions upstream of genes to find the maximal energy matrix score in each region (de Jong et al. 2012; Kim and Ren 2006; Stormo and Zhao 2010; Wade et al. 2007). If this score is above a certain threshold, the downstream gene is classified as a putative target. This procedure, in essence, reduces the problem of finding transcription targets to detecting individual TFBS. However, the problem with such an approach is that, as explained before, such a procedure (based on detecting individual TFBS) leads to high false positive numbers (Robison et al. 1998; Stormo 2000). Therefore, the main challenge, which we addressed in this example, is to do better than this.

To address this challenge, the basic hypothesis behind our approach was as follows: A number of sites with high energy matrix scores appear randomly in DNA sequences, that is, in a long stretch of text (e.g. DNA sequence) a sufficiently short word (e.g. a binding motif) may occur randomly (Kim and Ren 2006). Such randomly occurring high scoring binding sites are often called non-sites. Our basic assumption [which is also empirically supported by binding energy/score distributions (Djordjevic et al. 2003; Mustonen et al. 2008; Mustonen and Lassig 2005; Sengupta et al. 2002)] is that these non-sites are under weak negative selection in the genome so that they are mostly not deleted from the genome sequence. Consequently, such non-sites are also predicted as true binding sites in energy matrix searches, leading to a high number of false positives as discussed above. Therefore, to improve the search accuracy, one must somehow deal with these non-sites, i.e. find a way to filter them out when making predictions.

Our basic idea is that instead of individual TFBS one should look at the distribution of the energy matrix scores in the upstream regions of genes (Djordjevic et al. 2017). To motivate this idea, in Fig. 4a we show the distribution of predicted binding energies (black line) in all *E. coli* upstream intergenic regions for CAP transcription factor. As CAP is a pleiotropic regulator, which regulates a number of genes in *E. coli*, it binds to at least some of these upstream intergenic regions. On the other hand, in Fig. 4b we show the distribution (black line) of predicted CAP binding energies in convergent intergenic regions. One does not expect any functional TF binding in convergent intergenic regions since they are downstream of both of the adjacent genes. One can see clear overrepresentation of the binding energy distribution in the upstream intergenic regions, i.e. in these regions where the TF is expected to bind. On the other hand, such overrepresentation is absent in the convergent intergenic regions (where the TF is not expected to bind). This then leads us to our basic idea to predict direct targets by assessing the

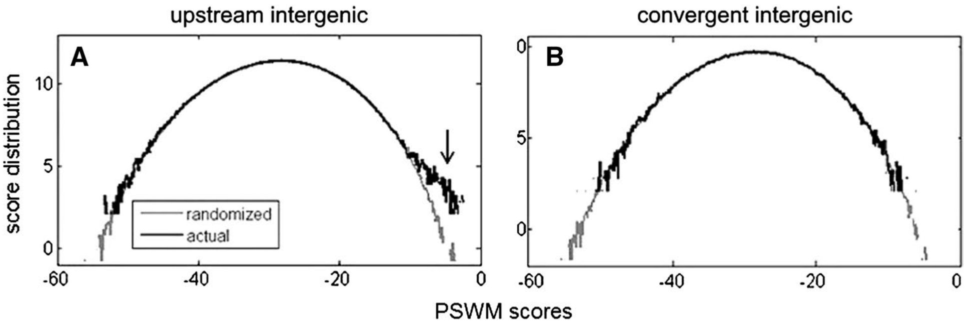


Fig. 4 Predicted binding energy distribution for **a** upstream intergenic regions and **b** convergent intergenic regions. Black and grey lines denote binding energy distributions in actual intergenic regions and in their randomized counterparts, respectively. Convergent intergenic regions are those that are downstream of both of the adjacent genes, while all other intergenic regions (that are upstream of at least one of the adjacent genes) are denoted by upstream intergenic regions. Overrepresentation in the high scoring tail for the upstream intergenic region distribution is indicated by an arrow Figure adopted from (Djordjevic et al. 2017)

difference between the binding energy distribution in an upstream region and the background distribution.

The idea discussed above than boils down to the problem of comparing two distributions. It is convenient that this problem maps to some of the classical work in mathematics. A major advantage of this is that one can robustly assign a measure of statistical significance to any predicted target. The best-known example of a test comparing two distributions is the Kolmogorov–Smirnov (KS) test. This test is non-parametric, and assigns a so-called D score, which quantifies a distance between two cumulative distribution functions (CDFs)—an example of a D score is shown graphically in Fig. 5a. These D scores are calculated very fast, directly from CDFs; consequently, one can scan thousands of distributions/genes on a PC within seconds. Moreover, each of the obtained D scores is associated with a robust significance estimate (P value). A modern competitor of the KS test is the Anderson–Darling test, which is considered more sensitive, but is significantly slower.

In Fig. 5a, an example of how the method works is shown, that is, one can see an example of a positive, i.e. of an upstream region that the method predicts as a target. Here CDF for actual binding energies is above the background, and the associated D score is highly statistically significant. On the other hand, all the cases where actual CDF would be slightly (insignificantly) above the background, or even below the background, would be classified as negatives. Note that the assessment in Fig. 5a is done for individual upstream regions, i.e. the method assesses the upstream regions one by one, while to drive the method, in Fig. 4 we showed global (i.e. in all upstream regions) binding energy distributions.

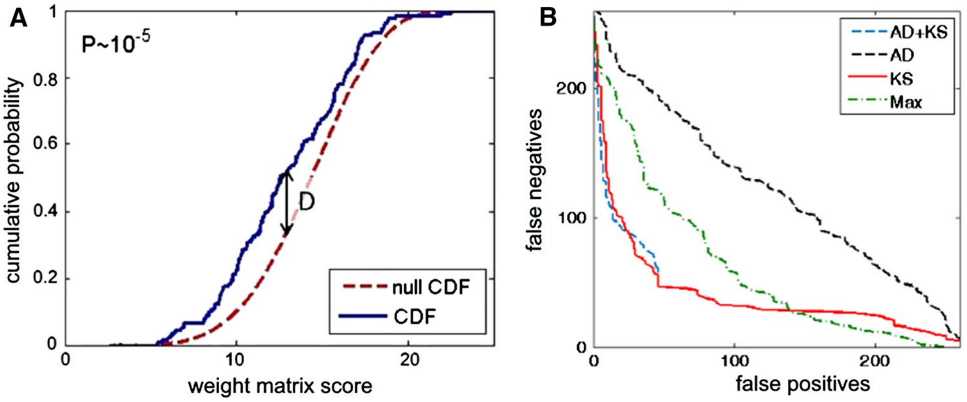


Fig. 5 **a** An example of a predicted positive for the KS-based method. CDFs for predicted binding energies for the actual upstream intergenic region, and for the background (null) are shown by the full blue line, and by the red dashed line, respectively. The associated D score and P value for the difference between the two distributions are shown. **b** ROC curves for the method based on the AD (Anderson–Darling test), the KS (Kolmogorov–Smirnov test) and the AD + KS (hybrid of the Anderson–Darling and the Kolmogorov–Smirnov test) are shown as indicated in the figure legend. Here, “Max” corresponds to the standard method based on assessing TFBS with maximal predicted binding energy, as described in the text Figure adopted from (Djordjevic et al. 2017)

As to the background distribution mentioned above, the question comes of what is an optimal background distribution to use. The usual choice would be to randomize the upstream intergenic regions, i.e. to randomly permute the bases within them so that certain nucleotide statistics are preserved (e.g. a number of trinucleotides). An alternative, natural choice, would be to use the binding energy distribution in the convergent intergenic regions as the background distribution. This choice comes to mind since, as mentioned above, there should be no functional binding to the convergent intergenic regions. In fact, if one compares how well these two choices of background can separate the scores in the positive and putative negative set, one finds that the background based on convergent intergenic regions can much better separate the two D score distributions. The fact that the background based on convergent intergenic regions is more optimal is, in fact, not surprising, because this background also takes into account a likely small (but still existing) negative selection on the non-sites. The randomized intergenic regions are, of course, not sensitive to such negative selection, though they can more accurately reproduce the nucleotide content of the searched region.

In Fig. 5b one can see the receiver operating characteristic (ROC) curve, providing false positive vs. false negative comparison between our new method (based on the KS procedure) and the standard approach described above. One can see that the new method leads to a much better performance compared to the standard approach (note that the more accurate method corresponds to the curve which is more bowed, i.e. which has better false positive/false negative trade off). This is, in fact, expected as the KS, in distinction to the standard method, is able to take care of non-sites, thereby significantly reducing the number of false positives. What, however, comes as a surprise is that if the method is implemented through the Anderson–Darling (AD), rather than the KS procedure, we obtain a significantly worse performance compared to both the KS-based procedure and the standard method.

To understand this result, i.e. why the procedure implemented through the KS shows a much better performance, we compared sensitivity and specificity for both implementations. We found that the sensitivity of the AD-based procedure is comparable or even better compared to the sensitivity of the KS-based procedure. This is expected, as, normally, the AD test is considered very sensitive. However, when the specificity of the two implementations was compared, we found that the AD-based procedure shows a much lower specificity compared to the KS. In retrospect, it is not hard to understand this result as, in fact, we have only an approximate null (background) distribution. Due to this the large AD sensitivity becomes a problem, as even small differences between the two distributions are highlighted as significant leading to a large number of false positives. This proposition is also consistent with the result shown in Fig. 5b, where a performance (ROC curve) corresponding to the AD–KS hybrid method is shown. In this hybrid, the KS is used to filter out the upstream regions with clearly insignificant P values. If this is done, the accuracy then becomes comparable to the KS implementation, indicating that AD indeed has a problem in correctly classifying regions with small (insignificant) differences compared to the null distribution.

Moreover, it turns out that for some TF our KS-based procedure shows a comparable performance with the standard method, i.e. based just on ROC curves, no clear difference in the performance of the two methods can be observed. However, even in such cases, we argue that the KS procedure has a distinctive advantage, which originates from the robust assignment of statistical significance, allowed by the KS approach. That is, in the case of individual TFBS (the standard method), it is much more complicated to calculate P values, since the binding sites with predicted high binding energies are located in the tail of the energy distribution, which tends to be highly variable (Hertz and Stormo 1999). Consequently, to classify the hits (predict TFBS), one often resorts to what is known as the standard binding threshold, which is based on correctly classifying ~ 98% TFBS in the training set (Robison et al. 1998). When the prediction accuracies of both methods at the standard threshold values (for the KS $P = 0.05$) are compared, the KS leads to significantly higher prediction accuracy compared to the standard method, even in those cases where ROC curves showed no clear advantage of the KS method. Therefore, the KS method leads to a clearly more optimal choice of the classification threshold, which is in itself a distinct advantage of the method.

Overall, this method provides an example of a novel concept (Djordjevic et al. 2017), where overrepresentation of the scoring distribution that corresponds to the entire searched region is assessed, as opposed to predicting individual binding sites. Two implementations of this concept were explored, based on the KS and the AD tests, which both provide straightforward P value estimates for predicted targets. We showed that the KS-based approach is both faster and more accurate, departing from the current paradigm of the AD being slower, but more sensitive. Consequently, the overviewed KS-based method may provide a both fast and accurate approach for predicting target loci of transcription regulators in a wide range of biological systems. In fact, even when one is interested in individual TFBS apart from the direct targets, first predicting the direct targets significantly narrows the search space, allowing to search upstream of only those genes that are predicted as direct targets. That is, once the target upstream regions are identified, one can score individual motifs just in these regions, with search threshold set in some of the standard ways [e.g. by estimating TF chemical potential, see (Djordjevic et al. 2003)].

Regulatory dynamics of restriction-modification systems

To illustrate the contribution of biophysical modelling in understanding intracellular expression dynamics, we will next overview an example from our research that concerns modelling dynamics of gene expression regulation. In particular, the example addresses regulation and dynamics of enzyme expression during establishment of a bacterial restriction-modification (RM) system in a new host. RM systems are rudimentary bacterial immune systems, i.e. they protect a bacterial cell from foreign DNA, such as that coming from viruses and plasmids (Kobayashi 2001). In fact, two main examples of bacterial immune systems are RM and CRISPR/Cas systems (Goldberg and Marraffini 2015). While RM systems are simpler, they present an important model system in molecular biology. The system expresses two enzymes, restriction endonuclease (R) and methyltransferase (M). R recognizes and cuts specific DNA sequences, while M protects the same DNA sequences that are cut by R. Due to this, R cuts unmethylated sequences of the incoming viruses, while host DNA is methylated and consequently protected from cutting (Kobayashi 2001; Mruk and Kobayashi 2013).

However, one should also have in mind that the RM systems are often mobile, i.e. they can spread from one bacterial host to another. Since the host genome is initially not methylated, it is evident that expression of the R and M has to be regulated, to prevent destruction of the host genome during establishment of the RM system in a bacterial host. Note that, as mentioned above, the sequence specificity of R and M in a given R–M system comes in pairs, so even if the host genome is protected by another methyltransferase, this will not prevent it from being cut by R—i.e. for this, the host genome has to be protected by its corresponding M. Also, after the initial synthesis dynamics of the two enzymes in a naïve bacterial host, steady-state levels of methyltransferase and restriction endonuclease are established; these steady-state levels have to be carefully tuned to ensure that, on the one side, the methylation pattern is maintained (which is necessary due to hemimethylation during DNA replication), and that, on the other side, the methylation of the foreign DNA is prevented. Consequently, in experimental biology and biochemistry, R–M systems are often taken as an experimental model for tightly regulated and highly coordinated gene expression (Kobayashi 2001; Mruk and Kobayashi 2013).

The tight regulation of R–M systems discussed above is often achieved by a specialized TF called control (C) protein (McGeehan et al. 2011; Mruk and Kobayashi 2013; Nagornykh et al. 2008). In Fig. 6a, gene organization in a typical RM system is schematically shown. In this case (of the Esp1396I RM system), C protein binds both to the operator sequence upstream of its own gene (O_{CR}) and to the operator sequence upstream of the M gene (O_M) (McGeehan et al. 2011; Mruk and Kobayashi 2013; Nagornykh et al. 2008). In fact, binding of C protein to DNA is often characterised by large binding cooperativity (Bogdanova et al. 2008, 2009; Nagornykh et al. 2008). First, two monomers must form a dimer in order to bind to DNA. In the absence of RNAP, as soon as one dimer is bound to O_{CR}, it recruits another dimer, forming a tetramer complex that represses transcription (Fig. 6b). Consequently, regulation by C protein is an example for the general concept discussed above, i.e. a large cooperativity that appears in molecular interactions inside a cell. When RNAP is present, it can displace one of the two dimers and initiate transcription (Bogdanova et al. 2008,

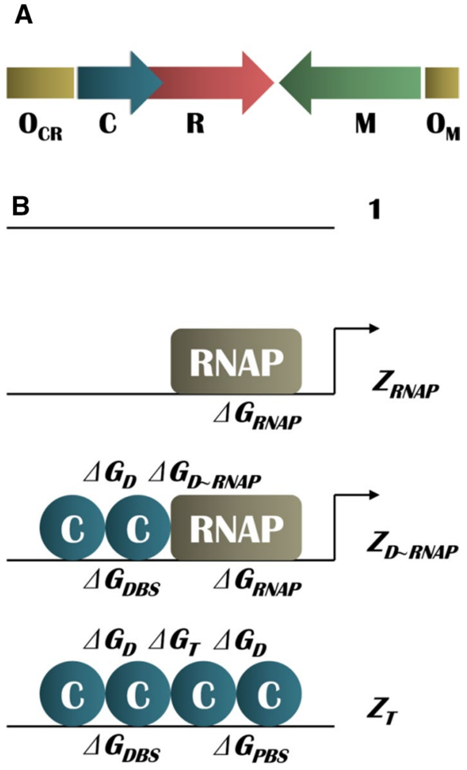

A
O_{CR}
C
R
M
O_M
B
1
RNAP
Z_{RNAP}
ΔG_{RNAP}
ΔG_D $\Delta G_{D\sim RNAP}$
C C RNAP
$Z_{D\sim RNAP}$
ΔG_{DBS} ΔG_{RNAP}
ΔG_D ΔG_T ΔG_D
C C C C
Z_T
ΔG_{DBS} ΔG_{PBS}

Fig. 6 **a** Typical organization of RM loci (genes that encode for RM system). C, R and M denote control protein, restriction enzyme and methyltransferase, respectively. Note that C and R are transcribed together (within the same operon). O_R and O_M denote the upstream operator sequences, to which C protein binds and regulates expression of the downstream genes. **b** Promoter configurations corresponding to O_R. The second and the third configurations correspond to the activation complex (in which the downstream genes are transcribed), while no transcription is exhibited in the first and the fourth configuration. Statistical weights for each configuration are indicated by Z, where indices RNAP, D-RNAP and T denote configurations in which just RNAP is bound, C dimer and RNAP are bound and C tetramer is bound; empty operator sequence corresponds to statistical weight of 1 Modified from (Morozova et al. 2016)

2009; Nagornykh et al. 2008). Control of M gene in this particular system is simpler, i.e. binding of a C dimer to O_M leads to transcription repression (scheme not shown). To model the dynamics of enzyme synthesis it is clear that we must first quantitatively describe how C protein regulates expression of the downstream genes. Briefly, this is done using statistical thermodynamics, where statistical weights (denoted by Z_{RNAP}, $Z_{D\text{-}RNAP}$ and Z_T in Fig. 6b) are assigned to each promoter configuration (Morozova et al. 2016; Sneppen and Zocchi 2005). Note that the promoter configurations, together with the corresponding interaction (free) energies are shown in Fig. 6b for O_{CR}. Each statistical weight contains an entropic contribution, which includes concentrations of relevant molecules (C and RNAP), and an energy term, where appropriate binding free energies (indicated in Fig. 6b) are included. The total promoter transcription activity is then calculated according to the so-called Shea–Ackers approximation, stating that the transcription rate is proportional to the binding occupancy of the promoter by RNAP (Shea and Ackers 1985):

$$\varphi \sim \frac{Z_{RNAP} + Z_{D\sim RNAP}}{1 + Z_{RNAP} + Z_{D\sim RNAP} + Z_T} = \frac{a + b[C]^2}{1 + a + b[C]^2 + c[C]^4} \tag{1}$$

Note that in the equation above, the numerator corresponds to the sum of statistical weights of transcriptionally active configurations, while the denominator corresponds to the sum of all statistical weights (i.e. to the partition function). For simplicity, the statistical weights in the expression above are written in terms of the part that depends on C protein concentration and the part in which all interaction parameters are absorbed [a, b and c in the Eq. (1)]. In fact, it is exactly cooperativity that leads to the quadratic and fourth-degree dependence in the Eq. (1) emphasizing the highly non-linear response due to cooperativity. It turns out that this model shows a very good agreement with in vitro data, i.e. is able to explain both the measurements of the promoter transcription activity done on the wild-type systems and on systems with different mutations in C binding sites (Bogdanova et al. 2008).

However, the next challenge is to check if the model can explain the in vivo data on R and M synthesis in bacterial cells, during RM system establishment. To model that, one can formulate the transcription activity as a function of C protein concentration and use it as an input in a dynamical model that describes R and M synthesis (Morozova et al. 2016). Such synthesis is described by appropriate ordinary non-linear differential equations, where the non-linearity comes from the non-linear dependence of transcription activity on C protein concentration. It was convenient that we could compare our model against the first available in vivo single cell measurements of R and M protein amounts in time in a population of dividing bacterial cells (Morozova et al. 2016). A major problem in measuring the intracellular dynamics is how to synchronize the bacterial cell population (Mruk and Blumenthal 2008). This was here resolved by measuring the protein expression levels in a clonal culture descending from the same (single) cell, which abolishes the need to synchronize the cell population and makes it meaningful to take the population average of the protein expression levels (shown for M in Fig. 7).

We constructed a minimal model, which includes only experimentally, directly observed regulatory interactions. The population dynamics effects were here included through a simple dilution model [see, e.g. (Narang and Pilyugin 2008)], which takes into account that proteins which are typically stable in the cell (with the decay rates of hours or days), are effectively degraded due to dilution by cell division. As shown in Fig. 7 (green line), this minimal model reproduces experimental data reasonably well, i.e. qualitatively it reproduces a massive peak of methyltransferase, as well as a large delay in endonuclease (dynamics not shown) with respect to methyltransferase synthesis, which are proposed as the main qualitative characteristics of RM system expression dynamics. Also, quantitatively, it reproduces the data well in the first ~ 150 min for M (and for R throughout the entire experiment), but at later times, one can clearly observe disagreement with the M data.

Such a comparison of the model predictions with experimental data is an illustration of a more complicated research cycle shown in Fig. 1b. As agreement of the model

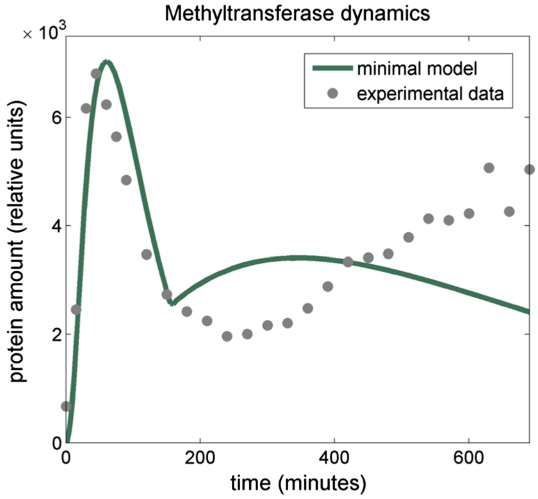


Fig. 7 Methyltransferase expression dynamics: comparison of the model with the experimental data. Grey dots are experimental measurements corresponding to in vivo measurements of enzyme protein amounts. Green curve corresponds to a minimal model, where only an abrupt change in cell division rate is taken into account Figure adapted from (Morozova et al. 2016)

predictions with experimental data is incomplete (i.e. we do not obtain a good agreement of the model with the data for later times), the next step is to revise the model, so that a more complete agreement of the predictions with the model may be achieved. The direction to revise the model is again suggested by the experiment, where it was observed that the cell division rate changes with time, slowing down at ~ 180 min. Moreover, experimental observations show that the plasmid number per cell increases significantly during the course of the experiment. This indicates that the enzyme dynamics in a cell can be notably influenced by interplay of cell division and plasmid division rates, i.e. in such cases there is a necessity to simultaneously consider cell, plasmid and DNA dynamics.

In gene expression regulation, the effects of changing global growth conditions (which in term change the population dynamics) were investigated for system steady states, and for simple synthetic gene circuits (Klumpp and Hwa 2014). The observed R–M system dynamics indicate that analysis of these effects should be extended to the cases of (i) the system early dynamics, i.e. before steady state is established, (ii) natural, rather than just synthetic gene circuits, (iii) growth conditions that change with time. We think that R–M systems can provide a very good model (for both in silico and experimental studies) to quantitatively understand these issues, as they present relatively simple gene circuits, whose dynamics are tightly constrained by gene expression regulation. However, studying the full dynamics in natural gene circuits may also be related to notable technical difficulties. For example, when the full population dynamics is introduced in the minimal model discussed above, it effectively couples R and M dynamics, since the terms describing plasmid division (i.e. changes in the gene copy number) appear in the equations describing both R and M dynamics. Technically, this leads to a significant increase in the dimensionality of the parameter inference problem. This can be dealt with either by some of more general methods for parameter inference in gene regulatory networks [such as genetic algorithms, see, e.g. (Kikuchi et al. 2003)], or by developing methods that are directly suited to the problem, i.e. exploiting the fact that the dynamical equations for all molecular species are coupled through the same terms (corresponding to the population dynamics parameters). Assessing the utility of these approaches, when applied to experimental data, will be further explored in the future.

Furthermore, a common approach to understand different system features in molecular biology is to perturb these features, e.g. by introducing mutations or reengineering the system. However, in terms of understanding the system dynamics, such an approach can be very difficult and labour intensive, as it would require both introducing such perturbations and measuring in vivo system dynamics for each of these perturbations. Alternatively, such perturbations can be made computationally, provided that the model of the wild-type system is available, which is much more straightforward and less labour intensive compared to the experimental approach. In fact, we are up to now unaware of experimental studies where in vivo expression dynamics of molecular species would be measured in parallel with extensive reengineering of the system. Consequently, in the context of studying RM system dynamics, we extensively in silico perturbed these systems (Rodic et al. 2017). The perturbations included characteristic biophysical features of the system, such as its translation rate, dimerization constant, binding cooperativity, etc., and such perturbations were done on RM systems with different architectures (in particular, convergent vs. divergent gene orientation). These perturbations showed that both different system features and system architectures can be explained in terms of a few relatively simple dynamical properties which we call design principles and quantify through appropriate dynamical property observables. In particular, these properties are: (i) a time delay of R with respect to M synthesis, which prevents autoimmunity, i.e. destruction of the host genome by R; (ii) a fast transition of the system from the "OFF" to the "ON" state, which ensures that once the host genome is methylated, it becomes protected from viral infections, (iii) increased stability of R steady state, so that too large fluctuations of the toxic molecule do not kill the host cell. Consequently, diverse features of RM systems can be explained in terms of a few relatively simple dynamical properties of the system (Rodic et al. 2017). In the future, it will be interesting to investigate how the main system dynamical properties would be affected by also perturbing the system population dynamics parameters. Such perturbations would correspond to, e.g. changing global physiological conditions, as changes in these conditions would influence the population dynamics. Such investigation might extend the idea of robustness, (Alon et al. 1999; Barkai and Leibler 1997), initially introduced in the context of intracellular regulation, to changes in the external cell conditions.

Conclusion

The large amount of data in modern biology require development of quantitative methods that enable their analysis and interpretation. While modern biology in general is becoming a genuine data science, accumulation of massive data is arguably most evident in 'omics research, where huge amount of sequence information is being accumulated, and in systems biology, where advanced quantitative measurements are necessary to understand complex interactions between system components. In this review, we argued that biophysics can allow understanding of the biological mechanisms that underlie generation of these complex data. Consequently, bioinformatics (defined as computational molecular biology) and biophysical modelling are complementing each other, and their close integration is necessary for understanding complex biological systems.

To specify the general points made above, we here surveyed two examples where biophysical understanding can contribute to the analysis of massive biological data. The first example concerned sequence analysis and illustrates applications to 'omics data (in this case genome analysis). Here, the new bioinformatics method was developed by considering a physical mechanism of TF binding, where it was proposed that not individual binding energies, but their overrepresentation with respect to background distribution is indicative of function. This physical concept was then translated to an appropriate mathematical framework allowing us to make bioinformatics predictions (in this case of target genes). Moreover, while this concept was tested for bacterial transcription regulators, we expect that it could equally (or even better) be applicable to eukaryotic transcription regulators. This is because in eukaryotes clustering of TFBS is often observed, which would lead to even more pronounced differences between the actual and the background distributions. This illustrates that once the underlying biophysical mechanism is taken into account, it may lead to a data analysis method applicable to a wide range of systems, as this same mechanism may be common in diverse systems.

The second example concerns understanding the dynamics of gene expression regulation and directly illustrates how biophysical modelling can contribute to systems biology. In this example, we have seen that a thermodynamical and dynamical system modelling can explain well both in vitro and in vivo measurements done on RM systems (Bogdanova et al. 2008; Morozova et al. 2016). However, in addition to including the intracellular regulation, a more realistic model of the system also has to include full population dynamics effects. This provides an example of a more complex research cycle in biology, where a theoretical model is being revised based on comparison with quantitative experiment. While key regulatory features in RM systems are different, they can be explained in terms of a few simple design principles (Rodic et al. 2017). Such design principles may provide a common framework for understanding these systems, emphasizing the importance of quantitative modelling of the system dynamics. In fact, understanding similarities in mechanistically otherwise different biological systems (which are often called design principles), is a major goal of systems biology.

Acknowledgements This work was funded by the Swiss National Science foundation under SCOPES Project Number IZ73Z0_152297, and by the Ministry of Education, Science and Technological Development of the Republic of Serbia under Project Number ON173052.

References

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P (2014) Molecular biology of the cell, 6th edn. W. W. Norton & Company, New York

Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397:168

Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387:913

Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol 193:723–750

Berg O, von Hippel P (1988a) Selection of DNA binding sites by regulatory proteins. Trends Biochem Sci 13:207–211

Berg OG, von Hippel PH (1988b) Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. J Mol Biol 200:709–723

Berg J, Willmann S, Lassig M (2004) Adaptive evolution of transcription factor binding sites. BMC Evol Biol 4:42

Bogdanova E, Djordjevic M, Papapanagiotou I, Heyduk T, Kneale G, Severinov K (2008) Transcription regulation of the type II restriction-modification system AhdI. Nucleic Acids Res 36:1429–1442

Bogdanova E, Zakharova M, Streeter S, Taylor J, Heyduk T, Kneale G, Severinov K (2009) Transcription regulation of restriction-modification system Esp1396I. Nucleic Acids Res 37:3354–3366

Bulyk ML (2003) Computational prediction of transcription-factor binding site locations. Genome Biol 5:201

Bulyk ML (2006) DNA microarray technologies for measuring protein–DNA interactions. Curr Opin Biotechnol 17:422–430

Das MK, Dai HK (2007) A survey of DNA motif finding algorithms. BMC Bioinform 8(Suppl 7):S21

De Jong H, Geiselmann J (2014) Fluorescent reporter genes and the analysis of bacterial regulatory networks international workshop on hybrid systems biology. Springer, pp 27–50

de Jong A, Pietersma H, Cordes M, Kuipers OP, Kok J (2012) PePPER: a webserver for prediction of prokaryote promoter elements and regulons. BMC Genom 13:299

Djordjevic M (2013) Efficient transcription initiation in bacteria: an interplay of protein–DNA interaction parameters. Integr Biol 5:796–806

Djordjevic M, Sengupta AM (2006) Quantitative modeling and data analysis of SELEX experiments. Phys Biol 3:13–28

Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. Genome Res 13:2381–2390

Djordjevic M, Djordjevic M, Zdobnov E (2017) Scoring targets of transcription in bacteria rather than focusing on individual binding sites. Front Microbiol 8:2314

Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, Makeev VJ (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. Bioinformatics 21:2240–2245

Feklístov A, Sharon BD, Darst SA, Gross CA (2014) Bacterial sigma factors: a historical, structural, and genomic perspective. Annu Rev Microbiol 68:357–376

Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, García-Sotelo JS, López-Fuentes A (2011) RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Res 39:D98

Goldberg GW, Marraffini LA (2015) Resistance and tolerance to foreign elements by prokaryotic immune systems—curating the genome. Nat Rev Immunol 15:717–724

Guzina J, Djordjevic M (2015) Inferring bacteriophage infection strategies from genome sequence: analysis of bacteriophage 7-11 and related phages. BMC Evol Biol 15:1

Helmann JD, Chamberlin MJ (1988) Structure and function of bacterial sigma factors. Annu Rev Biochem 57:839–872

Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15:563–577

Homsi D, Gupta V, Stormo G (2009) Modeling the quantitative specificity of DNA-binding proteins from example binding sites. PLoS ONE 4:e6736

Jagannathan V, Roulet E, Delorenzi M, Bucher P (2006) HTPSELEX–a database of high-throughput SELEX libraries for transcription factor binding sites. Nucleic Acids Res 34:D90

Jayaram N, Usvyat DR, Martin AC (2016) Evaluating tools for transcription factor binding site prediction. BMC Bioinform. https://doi.org/10.1186/s12859-016-1298-9

Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M (2003) Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics 19:643–650

Kim TH, Ren B (2006) Genome-wide analysis of protein-DNA interactions. Annu Rev Genomics Hum Genet 7:81–102

Klimuk E, Bogdanova E, Nagornykh M, Rodic A, Djordjevic M, Medvedeva S, Pavlova O, Severinov K (2018) Controller protein of restriction–modification system Kpn2I affects transcription of its gene by acting as a transcription elongation roadblock. Nucleic Acids Res 46:10810–10826

Klumpp S, Hwa T (2014) Bacterial growth: global effects on gene expression, growth feedback and proteome partition. Curr Opin Biotechnol 28:96–102

Klumpp S, Zhang Z, Hwa T (2009) Growth rate-dependent global effects on gene expression in bacteria. Cell 139:1366–1375

Kobayashi I (2001) Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res 29:3742–3756

Lee J, Freddolino PL, Zhang Y (2017) Ab initio protein structure prediction From protein structure to function with bioinformatics. Springer, pp 3–35

Levitsky VG, Kulakovskiy IV, Ershov NI, Oshchepkov DY, Makeev VJ, Hodgman T, Merkulova TI (2014) Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. BMC Genom 15:80

Locke G, Morozov AV (2015) A biophysical approach to predicting protein–DNA binding energetics. Genetics 200:1349–1361

Longo D, Hasty J (2006) Dynamics of single-cell gene expression. Mol Syst Biol 2:64

McGeehan J, Ball NJ, Streeter S, Thresh S-J, Kneale G (2011) Recognition of dual symmetry by the controller protein C. Esp1396I based on the structure of the transcriptional activation complex. Nucleic Acids Res 40:4158–4167

Morozova N, Sabantsev A, Bogdanova E, Fedorova Y, Maikova A, Vedyaykin A, Rodic A, Djordjevic M, Khodorkovskii M, Severinov K (2016) Temporal dynamics of methyltransferase and restriction endonuclease accumulation in individual cells after introducing a restriction-modification system. Nucleic Acids Res 44:790–800

Mruk I, Blumenthal RM (2008) Real-time kinetics of restriction-modification gene expression after entry into a new host cell. Nucleic Acids Res 36:2581–2593

Mruk I, Kobayashi I (2013) To be or not to be: regulation of restriction–modification systems and other toxin–antitoxin systems. Nucleic Acids Res 42:70–86

Mustonen V, Lassig M (2005) Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc Natl Acad Sci USA 102:15936–15941

Mustonen V, Kinney J, Callan CG Jr, Lassig M (2008) Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proc Natl Acad Sci USA 105:12376–12381

Nagornykh M, Bogdanova E, Protsenko A, Solonin A, Zakharova M, Severinov K (2008) Regulation of gene expression in a type II restriction-modification system. Russ J Genet 44:523–532

Narang A, Pilyugin SS (2008) Bistability of the lac operon during growth of *Escherichia coli* on lactose and lactose + glucose. Bull Math Biol 70:1032–1064

Newburger DE, Bulyk ML (2009) UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. Nucleic Acids Res 37:D77–D82

Ohno M, Karagiannis P, Taniguchi Y (2014) Protein expression analyses at the single cell level. Molecules 19:13932–13947

Paget M, Helmann J (2003) The sigma70 family of sigma factors. Genome Biol 4:203

Park PJ (2009) ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680

Phillips R, Kondev J, Theriot J, Garcia H (2012) Physical biology of the cell. Garland Science, New York

Robison K, McGuire A, Church G (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J Mol Biol 284:241–254

Rodic A, Blagojevic B, Zdobnov E, Djordjevic M (2017) Understanding key features of bacterial restriction-modification systems through quantitative modeling. BMC Syst Biol 11:377–391

Roulet E, Busso S, Camargo A, Simpson A, Mermod N, Bucher P (2002) High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol 20:831–835

Sandve GK, Drablos F (2006) A survey of motif discovery methods in an integrated framework. Biol Direct 1:11

Sengupta AM, Djordjevic M, Shraiman BI (2002) Specificity and robustness in transcription control networks. Proc Natl Acad Sci USA 99:2072–2077

Shea MA, Ackers GK (1985) The OR control system of bacteriophage lambda: a physical-chemical model for gene regulation. J Mol Biol 181:211–230

Sneppen K, Zocchi G (2005) Physics in molecular biology. Cambridge University Press, Cambridge

Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16:16–23

Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci 23:109–113

Stormo GD, Zhao Y (2010) Determining the specificity of protein–DNA interactions. Nat Rev Genet 11:751–760

Towsey M, Hogan J, Mathews S, Timms P (2008) The in silico prediction of promoters in bacterial genomes. Genome Inform 21:178–189

Vilar JM (2010) Accurate prediction of gene expression by integration of DNA sequence statistics with detailed modeling of transcription regulation. Biophys J 99:2408–2413

Vilar JM, Saiz L (2013) Systems biophysics of gene expression. Biophys J 104:2574–2585

Wade JT, Struhl K, Busby SJ, Grainger DC (2007) Genomic analysis of protein–DNA interactions in bacteria: insights into transcription and chromosome organization. Mol Microbiol 65:21–26

Zuker M, Mathews DH, Turner DH (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide RNA biochemistry and biotechnology. Springer, Dordrecht, pp 11–43