DOI 10.5281/zenodo.827157

Minireview

Modeling and bioinformatics of bacterial immune systems: understanding regulation of CRISPR/Cas and restriction-modification systems

Jelena Guzina[1], Anđela Rodić[1], Bojana Blagojević[2] and Marko Đorđević[1*]

[1]University of Belgrade – Faculty of Biology, Studentski trg 3, 11000 Belgrade, Serbia
[2]Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia

Summary. Bacterial immune systems protect bacterial cells from foreign DNA, such as viruses and plasmids. They also critically affect bacterial pathogenicity by reducing the flow of genes between bacteria. Two such major systems are restriction-modification and the recently discovered CRISPR/Cas systems. Here we review our work on understanding gene expression regulation in these systems, which takes a systems biology approach, combining modeling, bioinformatics and data analysis from quantitative experiments. Specifically, we address the following: (i) modeling gene expression regulation during restriction-modification system establishment in a naïve bacterial host, (ii) modeling the dynamics of CRISPR/Cas activation, in particular, how the features characterizing system transcription regulation and transcript processing affect the dynamics, (iii) predictions of transcription start sites for alternative σ factors that have been poorly studied up-to-now, but are important as CRISPR/Cas likely responds to bacterial cell envelope stress, (iv) our preliminary results on predictions of different CRISPR/Cas components, in particular, small RNAs associated with the systems, which likely have a key role in their regulation.

Keywords: Bacterial defense systems, bioinformatics, biophysical modeling, CRISPR/Cas, restriction-modification systems.

Introduction

Bacteria are continuously exposed to foreign nucleic acids, such as phage DNA, plasmids or other mobile genetic elements. In order to protect genome integrity, cells are equipped with immune systems that target invasive extrachromosomal elements for degradation (Shabbir et al. 2016), whereby the immune response reduces the rate of horizontal gene transfer (HGT), thus also affecting related aspects of cell functioning (e.g. virulence) (Vasu and Nagaraja 2013; Hatoum-Aslan and Marraffini 2014). Analogous to eukaryotic modes of defense, bacterial immune systems can be recognized as innate or adaptive, where restriction-modification and CRISPR/Cas (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated proteins), respectively, are two major representatives of such systems (Goldberg and Marraffini 2015).

Restriction-modification (RM) systems are considered innate since they target invasive elements without prior immunization with fragments of foreign genetic material. Two major components of RM systems are the enzymes restriction endonuclease (R) and methyltransferase (M) (Fig. 1A), which are frequently encoded on mobile genomic loci (e.g. plasmids), so that these systems easily propagate through bacterial populations (Fig. 1B) (Heitman 1993; Kobayashi et al. 1999). Once an RM system enters the cell, tight regulation of its expression becomes essential for ensuring safe and efficient establishment in the naïve bacterial host. Precisely, R that represents the effector component of a RM system, cuts short specific DNA sequences, irrespective of their location, so that self-targeting can easily arise. To evade autoimmunity, R has to be expressed with a delay with respect to M, as methylation of the genomic sites recognized by R protects them from cleavage (Fig. 1A) (Wilson 1991).

Unlike RM systems, CRISPR/Cas provides adaptive component to bacterial immunity, which arises as a consequence of its dynamical structure (Barrangou et al. 2007; van der Oost et al. 2009). A major system component is the CRISPR array,

*Corresponding author, e-mail: dmarko@bio.bg.ac.rs

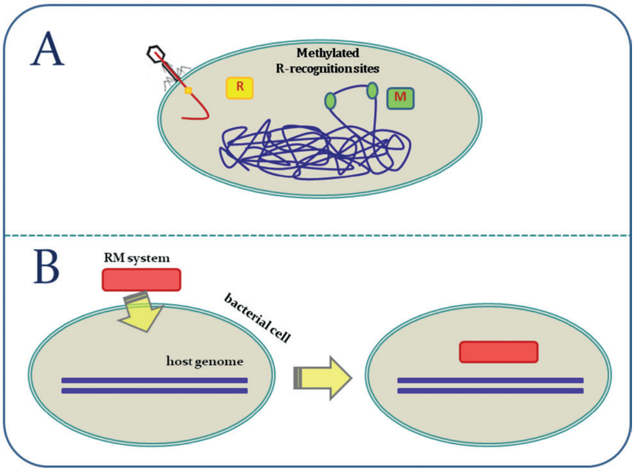

Fig. 1. RM system functioning (A) and establishment (B) in a host bacterial cell. A. Restriction-endonuclease (shown as a yellow rectangle) cuts the DNA at R-specific recognition sites (shown in yellow); Methyltransferase (shown as a green rectangle) methylates R-recognition sites on the host genome, thus protecting these sites (shown in green) from cleavage. **B.** RM systems are usually found on mobile genetic elements (e.g. plasmids), which enables them to efficiently propagate throughout bacterial populations. RM system, entering the bacterial cell (red rectangle), is shown.

which is characterized by a series of tandem repeats separated with unique spacer sequences (Fig. 2) (Al-Attar et al. 2011). The spacers are derived from previously encountered foreign genetic material, so that small interfering RNAs (crRNAs), which are generated upon array expression, target invasive elements based on complementarity; this makes the basic mechanism that confers resistance against foreign DNA/RNA (Bolotin et al. 2005). In addition to CRISPR array, the system also includes Cas proteins with mainly nucleolytic activity, which act as effectors during array immunization with new spacers, crRNA processing/expression and target degradation. CRISPR/Cas components typically remain silent under standard physiological conditions (Pul et al. 2010), whereby sudden activation leads to the production of large crRNA amounts, thus enabling efficient target eradication.

Despite the fact that RM and CRISPR/Cas systems markedly differ mechanistically, they likely embody the same design principles as a consequence of the general characteristics that shape the immune response. Namely, the induction of the CRISPR/Cas system probably faces similar dynamical constraints as the establishment of an RM system in a naïve bacterial host, as both require a rapid transition of the “toxic” (auto-immunogenic) molecule – R or crRNA – from “OFF” to “ON” state (Djordjevic 2013) to enable efficient target eradication. In addition to rapid transition, the expression of “toxic” immune molecules is also characterized by an initial delay, so that crRNAs in CRISPR/Cas are not expressed before the virus genome enters the cell, and M (the antidote) in RM systems has enough time to act.

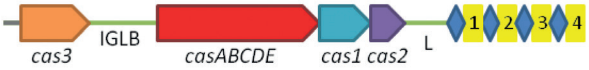

Fig. 2. A typical organization of CRISPR/Cas locus in *E. coli*. CRISPR array is schematically presented with successive blue diamonds (direct repeats) and yellow rectangles (spacers); the upstream *cas* genes, characteristic of Type I CRISPR/Cas systems, are indicated with rightwards-oriented pentagons. Intergenic regions that contain promoters transcribing *cas* genes (IGLB), and CRISPR array (L) are also shown.

Common design principles that impose similar dynamical constraints on RM and CRISPR/Cas activity are linked to the equivalent regulatory expression patterns in these systems. To understand the underlying transcription regulation, it is necessary to map transcription start sites (TSS) associated with different components of RM and CRISPR/Cas systems. This, however, is non-trivial since: (i) promoter elements of house-keeping σ factors are highly degenerate, so that a search usually results in a large fraction of false positives (Djordjevic 2014); (ii) information on the specificity of alternative σ factors (related to stress response) is largely missing, which is relevant since CRISPR/Cas is likely induced by cell-envelope stress (Ratner et al. 2015),which, in turn, is connected to Group IV (ECF) σ factors (Raivio and Silhavy 2001; Ratner et al. 2015).

In addition, an important aspect of CRISPR/Cas regulation are small RNAs associated with CRISPR/Cas (tracrRNAs) encoded outside the array, which have an essential role in CRISPR-transcript processing (Deltcheva et al. 2011), and possibly other system functions. Consequently, in this review we briefly present our work on:

1. modeling gene expression regulation during RM system establishment in a naïve bacterial host;
2. modeling dynamics of CRISPR/Cas activation, in particular how key features that characterize systems transcription regulation and transcript processing affect its dynamics;
3. predictions of bacterial TSS, particularly those related to alternative σ factors, which are poorly studied to date, but highly relevant as CRISPR/Cas likely responds to bacterial cell-envelope stress;
4. our preliminary results on predictions of different CRISPR/Cas components, in particular small RNAs associated with the system, which likely have a key role in its regulation.

Modeling *in vivo* expression of restriction-modification systems

Certain dynamical constraints imposed by their immune function have been proposed for RM systems in general. However, RM system dynamics have been observed in live cells in only two cases, as such experimental measurements are complicated by a requirement for synchronous populations of cells transformed with RM system genes (Mruk and Blumenthal 2008; Morozova et al. 2016). In an earlier attempt, Mruk and Blumenthal synchronously introduced the PvuII system genes placed on an M13 phage into naïve cells by phage infection (Mruk and Blumenthal 2008). Our collaborators, on the other hand, conducted the first single-cell measurements of RM system dynamics for the Esp1396I system: they fused sequences encoding fluorescent proteins to the R and M genes and monitored the dynamics of the appearance of fluorescent signals in individual cells, transformed with a plasmid carrying the modified Esp1396I system (Morozova et al. 2016). To check if the regulatory features found in this particular system allow establishing observed dynamics, and if they can provide the proposed dynamical constraints, we constructed a quantitative model of the Esp1396I system regulation, which we will briefly outline below.

Among type II RM systems, whose main characteristic is that R and M are encoded by separate genes, a large group contains a third gene encoding a control (C) protein, which is typically transcribed as a part of the operon with the R gene; the example for such a gene arrangement is the RM system Esp1396I represented in Fig. 3A. C proteins regulate transcription by binding in the form of dimers to their binding sites, partially overlapping with a promoter (Nagornykh et al. 2008). The transcription of Esp1396I system genes was thermodynamically modeled by considering all allowed configurations of the system promoters and determining their statistical weights (Figs. 3B and 3C). The most frequently observed regulation mechanism of the weak C and R operon (CR) promoter (also found in the Esp1396I system, Fig. 3B) involves highly cooperative binding of two C dimers to the left and the right operator sequences (O_L and O_R in Fig. 3A), where a C dimer bound to the high affinity left binding site can recruit either RNA polymerase (RNAP) to the promoter (thus activating transcription; the corresponding configuration has a statistical weight Z^{CR}_{D-RNAP}), or a second C dimer to the low affinity right binding site (establishing a tetramer that represses transcription; configuration Z^{CR}_{T}) (Bogdanova et al. 2008; Nagornykh et al. 2008). In the Esp1396I RM system, transcription of the M gene is also under the control of the C protein (Fig. 3C), whose binding to a single binding site (for a dimer; O_M in Fig. 3A) partially overlapping with the strong M promoter, excludes RNAP binding to the promoter and represses transcription of the M gene (configuration Z^{M}_{D}) (Bogdanova et al. 2009). For both the CR and the M promoter, configurations corresponding to basal transcription (configurations Z^{CR}_{RNAP} and Z^{M}_{RNAP} in Fig. 3, respectively) and empty promoters (statistical weight 1) were also assumed (Bogdanova et al. 2009). According to the classical Shea-Ackers assumption, which states that promoter transcription activity is proportional to the equilibrium probability of RNAP binding (Shea and Ackers 1985), the transcription activities of the CR$(\varphi_{CR})$and the M promoters$(\varphi_M)$are proportional to the probability of establishing their transcriptionally active configurations (for the statistical weights, see Fig. 3 caption):

$$\varphi_{CR} \propto \frac{a+b[C]^2}{1+a+b[C]^2+c[C]^4}\ , \quad \varphi_{M} \propto \frac{f}{1+f+g[C]^2}\ . \qquad (1.1)$$

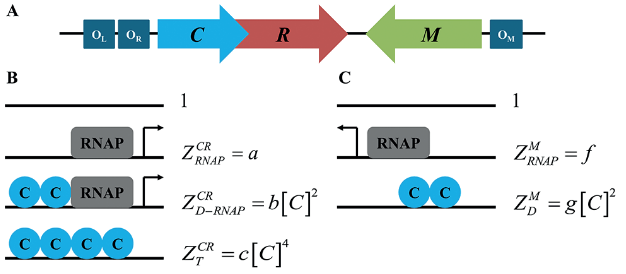

Fig. 3. Modeling transcription regulation in the Esp1396I RM system. A. Esp1396I gene organization scheme. The convergently oriented genes encoding R and M in the Esp1396I system are represented by the red and the green arrows, respectively, while the blue arrow represents the C gene, partially overlapping with the R gene. The dark blue boxes denoted by O_L, O_R and O_M represent operator sequences in the CR and the M promoter, which bind C dimers. **B and C.** The allowed configurations of RNAP (grey rectangle) and C protein (blue circle) molecules on the CR and the M promoter are illustrated, respectively, in B and C, where the transcriptionally active configurations contain an arrow. The corresponding statistical weights (*Z*) of the configurations, indicated on their right, depend on constant RNAP concentration and protein-protein and protein-DNA interaction energies (absorbed into parameters *a, b, c, f* and *g*) and variable C protein concentration.

Transcripts (with concentration m_i, where i = *R, M, C* denotes corresponding system components) synthesized from these promoters are degraded with a rate λ_i^m, while proteins (p_i) are generated by transcript translation with a rate $k_i$and are further degraded with a rate λ_i^p, as described by the following dynamical model equations:

$$\frac{dm_i(t)}{dt} = \varphi_i - \lambda_i^m \cdot m_i\ , \quad \frac{dp_i(t)}{dt} = k_i \cdot m_i - \lambda_i^p \cdot p_i\ . \qquad (1.2)$$

It should be noted that the decay terms (λ) in equations include not only degradation of the transcripts and the proteins, but also their dilution due to cell division, which occurred with two very different rates during the first (0-160 min) and second time intervals (after 160 min) of the experiment. Consequently, the cell population dynamics are in part taken into account in the model through the decay terms. However, there are likely significant additional population dynamics effects that should, in principle, be included in the model, e.g. those related to possible changes in the cell metabolism and different plasmid and cell division rates. Namely, our model describing the inherent RM system regulation and assuming constant parameters throughout the experiment (apart from different λ in the two time intervals) can successfully explain the main proposed qualitative features of system dynamics (Fig. 4), i.e. a large accumulation of M early upon plasmid entry into a naïve cell and a delay in the expression of R with respect to M, necessary for complete host genome protection. However, our model cannot completely quantitatively reproduce the system dynamics, i.e. there is a quantitative disagreement between the experimental data and the model predictions for M dynamics in the second time interval (after 160 min), likely arising from the additional population effects that we discussed above.

Design principles behind RM systems

The features of RM systems can be explained in terms of a few simple dynamical constraints that ensure safe and efficient RM system establishment. To this end, we proposed that all RM systems should exhibit the same simple dynamical properties: firstly, in every RM system there should be a significant expression of M prior to R, to avoid autoimmunity (Rodic et al. 2017). Once the host genome is protected (i.e. methylated), R should be rapidly generated, to "immunize" the host cell against virus infection, as fast as possible. Additionally, fluctuations of the toxic molecule R should be minimized, so as to evade that, due to large fluctuations, the toxic molecule amount is not matched by the antidote (M). Consequently, the following three dynamical properties are relevant to characterize RM system dynamics: (i) the time delay of R expression with respect to M; (ii) the transition velocity of the system from "OFF" to "ON" state; (iii) the stability of R steady state levels.

To quantify these dynamical properties, we referred to the predicted system dynamics and the stability of R steady-state levels in the wild type (wt) AhdI system (Fig. 5). Accordingly, we introduced the following dynamical property observables (Rodic et al. 2017): (i) the ratio of the shaded areas in the perturbed and in the wt system for the first 10 min post-system entry as a measure of the time delay (Fig. 5A); (ii) the maximal slope of the sigmoidal R expression curve as a measure of the transition velocity from "OFF" (low R value) to "ON" (high R value) state (Fig. 5A); (iii) a measure of the stability of R steady-state levels (Fig. 5B) as derived in Bogdanova et al. (2008) – note that greater steady-state stability leads to smaller R fluctuations.

We here employed the biophysical model of wt AhdI transcription regulation that we previously developed and which was verified by the *in vitro* experimental measurements of the AhdI transcription activity dependence on C protein concentration (Bogdanova et al. 2008), and also the dynamical model of transcript and protein expression, which was also verified by *in vivo* measurements (see above and Morozova et al. (2016)). The described methodology, which involves a combination of thermodynamic and kinetic modeling, has been successfully applied to various systems in molecular biology (Munro et al. 2016). While there are few studies concerned with modeling some aspects of RM systems expression regulation (Williams et al. 2013), to our knowledge our work is the first to employ this modeling approach to systematically understand the relation between RM system regulation and its dynamics.

In order to explain the (three) AhdI features, we per-

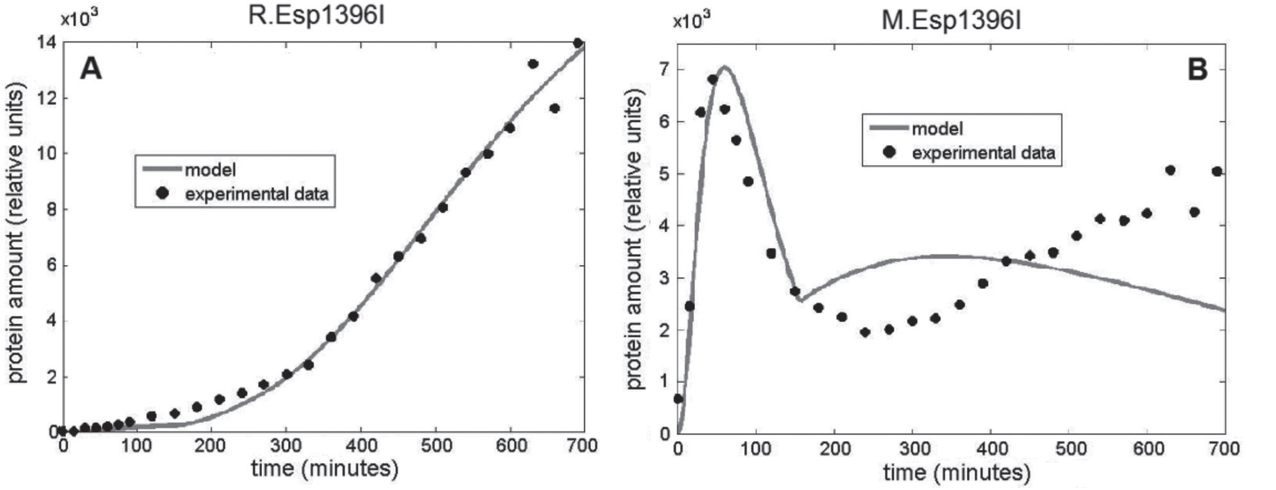

Fig. 4. Predicted Esp1396I RM system expression dynamics vs. experimental data. The change of R and M protein amounts in time is presented, respectively, in **A and B**. Circles correspond to the experimentally measured concentrations of protein fusions, while full lines correspond to the best fit of the model (described by the system of equations 1.1 and 1.2) to the data, obtained by varying parameters in biologically reasonable ranges. Time is set to zero at the point of the first available measurement. Adapted from Morozova et al. (2016).

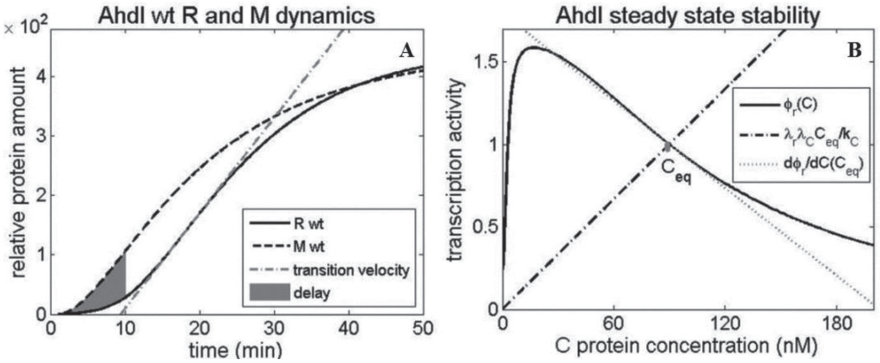

Fig. 5. Quantifying RM system dynamical properties A. R and M expression dynamics for AhdI RM system (Bogdanova et al. 2008). The shaded area presents a measure of a time delay between M (the dashed curve) and R (the solid curve) expression. The maximal slope of the sigmoidal R expression curve (dash-dotted line) is taken as a measure of the transition velocity from OFF to ON state. **B. Stability of the steady-state level**. The steady-state (C_{eq}) is obtained at the intersection of the CR promoter transcription activity (the solid curve) and the dash-dotted line, whose slope depends on the transcript and the protein decay rates and the protein translation rate (see Supplements in (Bogdanova et al. 2008) and (Rodic et al. 2017)). Larger difference in the slopes of the dash-dotted line and the solid curve at their intersection point leads to a more stable steady state. Adapted from Rodic et al. (2017).

turbed them *in silico*, one by one, to observe how this affects the dynamical property observables (Rodic et al. 2017). Firstly, we gradually increased the (initially low) C transcript translation initiation rate k_C towards the value characteristic of R and M transcripts. In Fig. 6A we observe a reduction in the delay between R and M expression, and a decreasing of the R steady-state level as the main effect of this perturbation. This finding can be intuitively explained by the fact that in increasing the translation initiation rate, C is generated faster, which hastens the formation of the activating and repressing complexes on the CR promoter. The effect on the other two observables is negligible. Consequently, this perturbation has a significantly adverse effect on one of the three dynamical properties (the delay between R and M expression), decreasing the ability of the system to protect the host genome from the cleavage.

Next, we gradually lowered the C subunit dissociation constant of dimerization K_1 from the very high value characteristic to the AhdI system, which corresponds to mostly C monomers in the solution, to low values, which correspond to predominantly C dimers in the solution, as shown in Fig. 6B (Bogdanova et al. 2008; Rodic et al. 2017). The three main effects of this perturbation are significant decreases in the time delay, in the transition velocity and in the steady-state levels of R. The stability of R steady-state levels is not significantly affected. Consequently, this perturbation has a significantly adverse effect on two dynamical properties, greatly reducing the ability of the system to protect the host genome from cleavage, and increasing the time window needed for the system to become protected from foreign DNA infection.

Finally, we gradually decreased only the extremely high cooperativity ω in C dimers binding to the CR promoter, which is shown in Fig. 6C (Rodic et al. 2017). We observe that this perturbation affects only the late R dynamics (see the left panel of Fig. 6C), since only efficiency in forming the repressor complex, whose probability is proportional to C^4, is affected, which becomes important only later on, when enough C is generated. Namely, this perturbation significantly decreases the stability of the steady state (see the right panel of Fig. 6C), thus having a significantly adverse effect on one dynamical property but not affecting the others. Also, contrary to the previous two perturbations, it significantly increases the steady-state levels of R, so that exhibiting different perturbations allows a balancing of the amount of the toxic molecule R in the cell.

To summarize, all three AhdI control features, in general, have the same effect on the dynamical properties, i.e. perturbing them makes at least one dynamical property much less optimal, while not notably affecting the other properties. This, together with the fact that decreasing the binding cooperativity ω has the opposite effect on the R steady-state levels from the other two perturbations (which facilitates controlling the toxic molecule R level) can explain the unusually large binding cooperativity in AhdI (Semenova et al. 2005; Bogdanova et al. 2009).

Dynamics of CRISPR/Cas system expression

Despite being intensively used in biotechnology for developing powerful genetic tools, the adaptive prokaryotic immune system CRISPR/Cas still appears to be underexplored when it comes to understanding the mechanism of its natural induction in a cell. In fact, the dynamics of CRISPR/Cas expression upon foreign DNA invasion have not been observed experimentally *in vivo*. What crucially hinders observing these dynamics is that CRISPR/Cas of Type I-E, which is the model system for CRISPR/Cas induction and regulation (most extensively studied in *E. coli*), is silent under

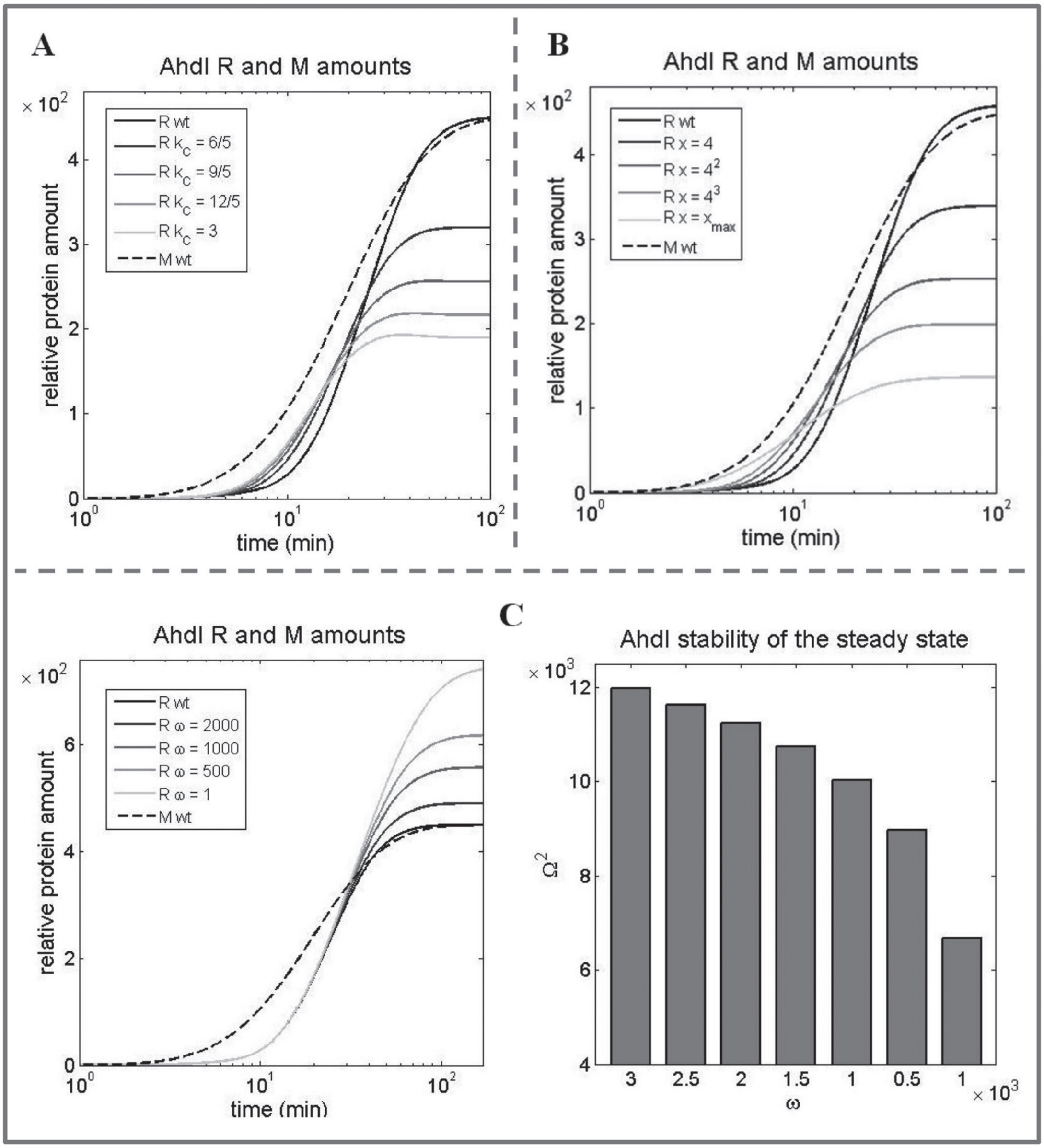

Fig. 6. Perturbing AhdI control features. A. Increasing C transcript translation initiation rate k_C. The effect of gradual k_C increase (from wt putative 3/5 1/min towards 3 1/min, which corresponds to the R and M (Bogdanova et al. 2008) is assessed on the protein expression dynamics, with R (solid) curves fading as k_C increases. The dashed curve corresponds to M expression, which is not affected by any of the three perturbations. **B.** Decreasing dissociation constant of C dimerization K_1. The effect of gradual K_1 decrease from the high value, corresponding to only monomers in the solution, to the low value, corresponding to only dimers in the solution, is assessed on the protein expression dynamics, with R (solid) curves fading as K_1 decreases. The relative protein amounts are derived from *in vitro* wt transcription activity measurements (Bogdanova et al. 2008). x denotes the ratio of K_1 decrease. **C.** Decreasing cooperativity ω of C dimers binding to CR promoter in AhdI. The effect of gradual decrease of extremely high ω, inherent to the wt AhdI system (Bogdanova et al. 2008), to ω corresponding to no binding cooperativity is assessed on the protein expression dynamics (the left figure), with R (solid) curves fading as ω decreases. The stability of R steady-state levels (the right figure). Adapted from Rodic et al. (2017).

normal growth conditions, even in the presence of bacteriophage infection, and the induction mechanism is only partially known (Westra et al. 2010). However, the dynamical properties of CRISPR/Cas induction can be understood by examining how the system regulatory features contribute to the expression dynamics, which can be efficiently performed using quantitative modeling.

Our group previously dynamically modeled pre-crRNA processing into crRNAs upon CasE (processing) protein overexpression (Djordjevic et al. 2012). The proposed model (schematically represented in Fig. 7) takes into account that pre-crRNA is synthesized by transcription of the CRISPR array and then either nonspecifically degraded by an unidentified endonuclease or processed by CasE into crRNAs, which are further relatively slowly degraded. The model predicts that the system feature crucial for enabling the experimentally measured, very large (~2 orders of magnitude) amplification of crRNAs from a small decrease in pre-crRNA concentration upon CasE overexpression, is the rapid, nonspecific degradation of pre-crRNA. Therefore, the unidentified endonuclease is probably an essential component for achieving the fast system transition from “OFF” to “ON” state.

However, CasE proteins, which process pre-crRNA and which determine how the processing rate (*k* in the Fig. 7) depends on time, are gradually synthesized when the induction signal is received. Therefore, to model CRISPR/Cas system induction, in addition to the transcript processing, transcription regulation of the *cas* promoter also has to be incorporated in the model. As the mechanism of transcription induction is not known, to address this problem, we noted clear qualitative similarities in transcription regulation of CRISPR/Cas and RM systems. In particular, while the *cas* promoter is repressed by very cooperative binding of global regulators (such as H-NS proteins), which can be displaced from the promoter by some transcription activators (such as LeuO) (Westra et al. 2010), in the RM systems described above RNAP itself acts as an activator, displacing the recruited C dimer from the repressor position (see Fig. 3B) (Bogdanova et al. 2008, 2009). Therefore, our main idea is to consider a synthetic gene circuit where transcript processing, which is exhibited in the CRISPR/Cas system (Fig. 7), is put under the transcription control of an RM system that was previously studied in detail. Specifically, we assume that *cas* (including *casE*) genes are transcribed together with a gene encoding the C protein from a promoter regulated by the cooperative binding of C dimers, as described above (Fig. 3B). In this way, transcription control of a well-studied RM system serves as a proxy for the transcription control of a much less understood CRISPR/Cas system and can be thermodynamically modeled as described above.

In our future work, we plan to compare the behavior of the model described above with that of a setup in which *cas* genes are constitutively expressed, which we will use to explore: (i) how the cooperative *cas* promoter regulation (see above) is related to the expected sharp switch-like behavior of the system; (ii) how the dynamics of crRNA generation in the cooperative model compares to the limit of infinitely fast (abrupt) system induction (Djordjevic et al. 2012), and (iii) how the fast nonspecific degradation of pre-crRNA (which is the main feature of CRISPR transcript processing) affects the system dynamics.

Predicting CRISPR/Cas system components

As previously mentioned, CRISPR/Cas systems are the focus of current intensive research; however, efforts are predominantly invested into the development of promising biotechnology applications that revolutionize the concepts of programmable genome editing and gene expression regulation (Singh et al. 2017). Consequently, insights into the

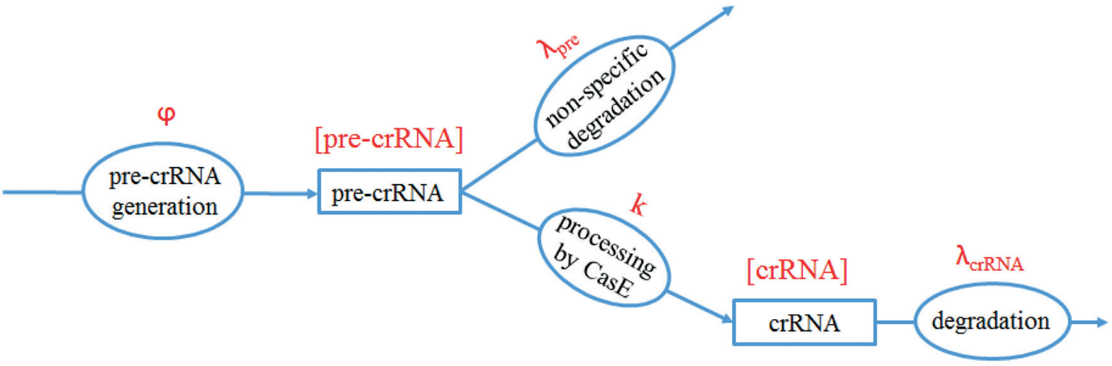

Fig. 7. The model scheme of pre-crRNA processing in CRISPR/Cas system. Notation used: φ – CRISPR promoter transcription activity, λ_{pre} – rate of (nonspecific) pre-crRNA degradation, k – rate of pre-crRNA processing to crRNAs by CasE, λ_{crRNA}– rate of crRNA degradation; square brackets denote concentrations of appropriate RNAs. Adapted from Djordjevic et al. (2012).

mechanisms that control the functioning of native CRISPR/Cas systems remain insufficiently explored. On the other hand, understanding native CRISPR/Cas function is crucial for the advancement of applied CRISPR/Cas research, which depends equally on the diversity of engineered CRISPR-based constructs and the capacity to control these constructs with sufficient precision.

An attractive avenue to improve the knowledge about native CRISPR/Cas systems, which could also lead to more powerful biotech applications, is investigating small CRISPR-associated RNAs. These RNA molecules (tracrRNAs), encoded outside the CRISPR array, are increasingly recognized as carriers of important regulatory and effector roles in the system. Namely, tracrRNAs are indispensable in Type II CRISPR/Cas systems for processing CRISPR array transcripts into mature crRNAs and subsequent targeting of the invasive genetic elements for degradation (in a complex with crRNA and Cas9 nuclease) (Deltcheva et al. 2011). At the same time, the underlying mechanism of action of this effector complex forms the basis for the Cas9:sgRNA paradigm that is extensively exploited for current CRISPR-based biotechnology applications (Hille and Charpentier 2016).

Despite their central role in CRISPR/Cas immunity and immense potential for translational research, small CRISPR-associated RNAs are largely unexplored, since their experimental discovery is complicated by (under standard conditions) a silent CRISPR/Cas system and still limited RNA-seq data in bacteria. An efficient alternative for the systematic identification and analysis of these small RNAs across different bacterial genomes is a bioinformatics-based approach, where the availability of sequenced genomic loci that encode CRISPR/Cas systems is the only prerequisite for computational analysis.

In general, small non-coding RNAs in bacteria are characterized by variable length, a low level of conservation and often indistinguishable secondary structure and nucleotide composition, so that *ab initio* detection, which is based on mining transcription signals (TSS and terminators) associated with small RNA expression units represents the most reliable search procedure (Sridhar and Gunasekaran 2013). However, a major shortfall of such an approach is that TSSs are often predicted with poor accuracy in bacterial genomes (Djordjevic 2014); for example, a standard supervised (information-theory based) search of the housekeeping (RpoD) promoter elements is associated with high rates of false positives.

Namely, due to considerable degeneracy of RpoD promoter elements, accurately aligning the -35 element to the -10 element is highly non-trivial, which was evidenced by our finding that the available -35 element alignments show a significant discrepancy with the biochemical data on σ^{70}-DNA interactions (Djordjevic 2011). In line with this, many implementations of the information-theory method use only the -10 element as the predictor of promoter specificity, which negatively affects the search accuracy. To address this problem, we performed systematic *de novo* MLSA (Multiple Local Sequence Alignment) alignment of RpoD promoter elements in *E. coli*, based on a Gibbs search (for more details on methods see Djordjevic 2011), which provided improved -35 element characterization, along with the identification of the -15 element, a previously unrecognized determinant of RpoD specificity (Djordjevic 2011). As illustrated in Fig. 8, employing this new alignment for a weight matrix-based TSS search resulted in false-positive reduction by 50% (Nikolic et al. 2017), which clearly advocates the implementation of the new alignment within small CRISPR-associated RNA search procedure.

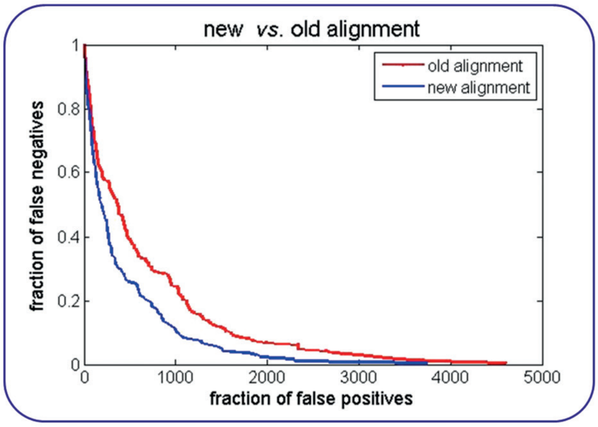

Fig. 8. DET (Detection Error Tradeoff) curve for the old and the new alignment of *E. coli* RpoD promoters. Fraction of false negatives is shown on the *y*-axis, and the fraction of false positives on the *x*-axis. DET-curve for the old alignment is colored red, and for the new alignment in blue. Adapted from Nikolic et al. (2017).

Compared to TSS, a terminator search is characterized by substantially higher accuracy, so that adaptation of the standard algorithm for Rho-independent terminator prediction in bacteria (Ermolaeva et al. 2000) can be used to detect small CRISPR-associated RNAs. Actually, for both TSSs and terminators, the search parameters can be trained against experimentally determined tracrRNAs across Type II CRISPR/Cas systems, where distinguishing true predictions (small RNAs) can be further aided by querying the predicted expression units for complementarity to the array direct repeats. Finally, secondary evidence for these *ab initio* predictions can be obtained through conservation analysis across related bacterial strains and mining available RNA-seq data. This, altogether, will be the core approach in our future research, which will focus on the systematic identification of small associated RNAs across diverse (Type II) CRISPR/Cas systems, with the goal of acquiring deeper insight into the functioning of native CRISPR/Cas systems.

The proposed procedure for small CRISPR-associated RNA detection is based on predicting housekeeping (RpoD) promoter elements; however, CRISPR/Cas induction is also (likely) related to the activity of alternative (ECF) σ factors, that takeover bacterial transcription in response to cell-envelope stress (Ratner et al. 2015). However, ECF promoter prediction is far more challenging, as the binding specificity in this highly versatile group of alternative σ factors (Staron et al. 2009) was largely unknown. Consequently, to address this problem we firstly systematically explored protein and DNA interaction motifs that are involved in transcription initiation by alternative σ factors, as described in the next section.

Transcription by ECF σ factors

Distinct from housekeeping (RpoD) σ factors that globally control bacterial transcription under standard growth conditions, alternative σ factors transcribe more specialized regulons in response to signals related with stress, metabolic changes or development. Among these, ECF σ factors are the most abundant and diverse, yet the underlying mechanisms of ECF transcription initiation are largely unexplored (Helmann 2002). Signaling cascades that activate ECF-specific transcriptional response are mainly triggered at the level of the cell membrane (Brooks and Buchanan 2008), which, on the other hand, is related to the invasion of foreign genetic elements into the bacterial cell. Consequently, equivalent signaling cascades are likely connected with CRISPR/Cas and ECF induction, so the analysis of ECF transcriptional mechanisms might further elucidate the regulatory mechanisms behind CRISPR/Cas activity.

Structurally, ECF σ factors are the simplest in the entire σ^{70} family, and, at the same time, characterized by the most versatile protein sequences (including DNA-binding domains). Accordingly, promoter specificity in this group is also highly diverse, as evidenced by the very limited capacity for ECF promoter cross-recognition (Rhodius et al. 2013). Clearly, inferring specificity for unexplored group members through comparative analysis against a number of experimentally characterized representatives is not applicable in the ECF σ group. However, it is this approach that underlies the current paradigm on ECF functioning, which assumes interaction with rigid promoters characterized by obligatory and well-conserved -35 and -10 elements (Staron et al. 2009; Feklistov et al. 2014).

The paradigm on ECF functioning is completely opposite to the mix-and-match mechanism of promoter recognition, which was well established in the housekeeping (RpoD) σ^{70} group (Hook-Barnard and Hinton 2007). Namely, the mix-and-match paradigm allows a flexible promoter element structure as long as the threshold transcription activity is accomplished through mutual complementation of promoter element interaction energies with the σ factor. The most extreme, and altogether best known example of this mechanism is -35 element absence in RpoD promoters, which is accommodated through σ factor interactions with a strong -10 element extension (also recognized as dsDNA).

Contrary to current considerations, we identified this ultimate example of promoter element complementation in ECF promoter sequences, recognized by the outlier group members (phage 7-11 and phiEco32 σ factors), during our systematic computational analysis of ECF promoter specificity, where we employed an extensive comparison of protein and DNA sequences through pairwise and multiple, global and local alignments (Fig. 9), for details see Methods in (Guzina and Djordjevic 2016). The presence of the classical mix-and-match trademark in phage ECF promoters is the first example of promoter recognition flexibility in the group, which we further corroborated by identifying a (putatively interacting) conserved protein motif, immediately C-terminal from the domain σ_2 boundary, through multiple global alignment of ECF protein sequences (Guzina and Djordjevic 2016).

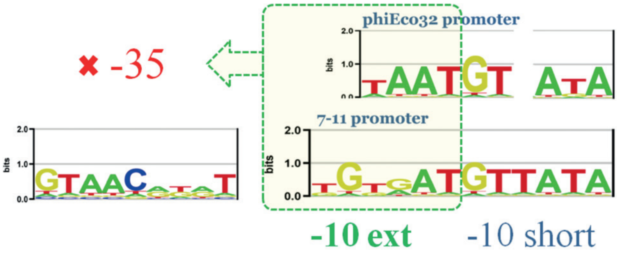

Fig. 9. Alignment of phage 7-11 and phiEco32 ECF promoters. Sequence-logo for 7-11 ECF promoters, with the presence of both -35 elements and long -10 element extensions is shown in the lower part of the figure; the logo for phiEco32 ECF promoters, where the presence of -10 element extension is followed by the absence of the -35 element, is shown in the upper part of the figure. Adapted from Guzina and Djordjevic (2016).

The coexistence of the conserved protein-DNA motifs was inferred in the bacterial ECF02 subgroup (containing experimentally well-characterized σ^E from *E. coli*) through multiple global and local alignments (Guzina and Djordjevic 2016). Interestingly, this novel σ-promoter interaction, whose partial conservation was also found in σ^W of *B. subtilis* (another canonical ECF member belonging to the ECF01 subgroup), appears further away from the domain σ_2/-10 element boundary (Guzina and Djordjevic 2016). At the same time, protein-DNA interactions in the spacer with inversed polarity (i.e. closer to the domain σ_4/-35 element boundary) are present in the ECF32 subgroup, which indicates that ECF σ factors display even greater flexibility during promoter recognition compared to the RpoD group. In fact, the observed flexibility in ECF promoter recognition aligns very well with the common biophysical mechanism of transcription initiation in the σ^{70} family, which is characterized by two major steps – closed and open complex formation (Djordjevic and

Bundschuh 2008). In the first step, σ^{70} factors interact with dsDNA promoter elements, while the second step depends on σ^{70} interactions with ssDNA elements. The interplay between these different energetic contributions determines the transcriptional output on the promoter, whose kinetic profile, in the framework of the mix-and-match mechanism, is indicated by the mutual complementation of the promoter elements, affecting the (most) relevant initiation step(s) for a given σ factor group.

In line with this, a biophysics-based correlation analysis we performed on a larger number of (*E. coli*) σ^E promoters (for more details on the analysis see ref. Guzina and Djordjevic 2017) revealed strong complementation between dsDNA elements, indicating that an efficient bacterial response to stress-related stimuli essentially depends on a high dsDNA-binding affinity of ECF σ factors for their promoters (Guzina and Djordjevic 2017). Correlations found between newly discovered spacer and canonical σ^E (-35 and -10) elements further corroborate the observed kinetic profile of ECF transcription initiation, which could, in turn, provide an alternative regulatory avenue for shaping the dynamics of CRISPR/Cas induction, where rapid expression of effector components (crRNA and Cas) appears as the main underlying signature. In our future research, we will use this detailed analysis of ECF σ factor specificity to develop methods for the accurate detection of TSS associated with these σ factors, which will, in turn, allow more accurate prediction of important CRISPR/Cas components, and consequently a better insight into the native system function.

Conclusion

Here we have reviewed our research on the modeling and bioinformatics of CRISPR/Cas and RM systems. We argue that the results presented to date show that combining experiments with modeling and bioinformatics is an optimal approach to understand the function of these exciting systems. Moreover, such an approach provides a better understanding of the common principles in design of these seemingly mechanistically quite different systems – understanding the principles that unify different biological systems is a major goal of systems biology. We believe that our current results provide a good starting point for understanding the regulation of diverse CRISPR/Cas and RM systems, including newly discovered CRISPR/Cas types. Regarding CRISPR/Cas, this can lead to new and improved biotechnology applications for a system that has already revolutionized the biotechnology field.

Acknowledgments

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, Project No. ON173052.

References

Al-Attar S, Westra ER, van der Oost J, Brouns SJ. 2011. Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biological Chemistry. 392(4):277-289.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 315(5819):1709-1712.

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 Research. 36(5):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 Research. 37(10):3354-3366.

Bolotin A, Quinquis B, Sorokin A, Enrlich SD. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 151(Pt 8):2551-2561.

Brooks BE, Buchanan SK. 2008. Signaling mechanisms for activation of extracytoplasmic function (ECF) sigma factors. Biochimica et Biophysica Acta. 1778(9):1930-1945.

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao J, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 471(7340):602-607.

Djordjevic M. 2011. Redefining *Escherichia coli* σ^{70} promoter elements: – 15 motif as a complement of the – 10 motif. Journal of Bacteriology. 193(22):6305-6314.

Djordjevic M. 2013. Modeling bacterial immune systems: strategies for expression of toxic - but useful - molecules. Biosystems. 112(2): 139-144.

Djordjevic M. 2014. Integrating sequence analysis with biophysical modelling for accurate transcription start site prediction. Journal of Integrative Bioinformatics. 11(2):240.

Djordjevic M, Bundschuh R. 2008. Formation of the open complex by bacterial RNA polymerase--a quantitative model. Biophysical Journal. 94(11):4233-4248.

Djordjevic M, Djordjevic M, Severinov K. 2012. CRISPR transcript processing: a mechanism for generating a large number of small interfering RNAs. Biology Direct. 7(1):24.

Ermolaeva MD, Khalak HG, White O, Smith HO, Salzberg SL. 2000. Prediction of transcription terminators in bacterial genomes. Journal of Molecular Biology. 301(1):27-33.

Feklistov A, Sharon BD, Darst SA, Gross SA. 2014. Bacterial sigma factors: a historical, structural, and genomic perspective. Annual Review of Microbiology. 68:357-376.

Goldberg GW, Marraffini LA. 2015. Resistance and tolerance to foreign elements by prokaryotic immune systems - curating the genome. Nature Reviews. Immunology. 15(11):717-724.

Guzina J, Djordjevic M. 2016. Promoter recognition by ECF sigma factors: analyzing DNA and protein interaction motifs. Journal of Bacteriology. 198(14):1927-1938.

Guzina J, Djordjevic M. 2017. Mix-and-matching as a promoter recognition mechanism by ECF sigma factors. BMC Evolutionary Biology. 17(Suppl 1):12.

Hatoum-Aslan A, Marraffini LA. 2014. Impact of CRISPR immunity on the emergence and virulence of bacterial pathogens. Current Opinion in Microbiology. 17:82-90.

Heitman J. 1993. On the origins, structures and functions of restriction-modification enzymes. Genetic Engeneering (N Y). 15:57-108.

Helmann JD. 2002. The extracytoplasmic function (ECF) sigma factors. Advances in Microbial Physiology. 46:47-110.

Hille F, Charpentier E. 2016. CRISPR-Cas: biology, mechanisms and relevance. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 371(1707).

Hook-Barnard IG, Hinton DM. 2007. Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. Gene Regulation and Systems Biology. 1:275-293.

Kobayashi I, Nobusato A, Kobayashi-Takahashi N, Uchiyama I. 1999. Shaping the genome – restriction–modification systems as mobile genetic elements. Current Opinion in Genetics and Development. 9(6):649-656.

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 Research. 44(2):790-800.

Mruk I, Blumenthal RM. 2008. Real-time kinetics of restriction-modification gene expression after entry into a new host cell. Nucleic Acids Research. 36(8):2581-2593.

Munro PD, Ackers GK, Shearwin KE. 2016. Aspects of protein–DNA interactions: a review of quantitative thermodynamic theory for modelling synthetic circuits utilising LacI and CI repressors, IPTG and the reporter gene *lacZ*. Biophysical Reviews. 8(4):331-345.

Nagornykh MO, Bogdanova ES, Protsenko AS, Zakharova MV, Solonin AS, Severinov KV. 2008. [Regulation of gene expression in type II restriction-modification system]. Genetika 44(5):606-615.

Nikolic M, Stankovic T, Djordjevic M. 2017. Contribution of bacterial promoter elements to transcription start site detection accuracy. Journal of Bioinformatics and Computational Biology 15(2):1650038.

Pul U, Wurm R, Arslan Z, Geissen R, Hofmann N, Wagner R. 2010. Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. Molecular Microbiology. 75(6):1495-1512.

Raivio TL, Silhavy TJ. 2001. Periplasmic stress and ECF sigma factors. Annual Reviews of Microbiology. 55:591-624.

Ratner HK, Sampson TR, Weiss DS. 2015. I can see CRISPR now, even when phage are gone: a view on alternative CRISPR-Cas functions from the prokaryotic envelope. Current Opinion in Infectious Diseases. 28(3):267-274.

Rhodius VA, Segall-Shapiro TH, Sharon BD, Ghodasara A, Orlova E, Tabakh H, Brukhardt DH, Clancy K, Peterson TC, Gross CA, et al. 2013. Design of orthogonal genetic switches based on a crosstalk map of sigmas, anti-sigmas, and promoters. Molecular Systems Biology. 9:702.

Rodic A, Blagojevic B, Zdobnov E, Djordjevic M, Djordjevic M. 2017. Understanding key features of bacterial restriction-modification systems through quantitative modeling. BMC Systems Biology 11(Supplement 1):2.

Semenova E, Minakhin L, Bogdanova E, Nagornykh M, Vasilov A, Heyduk T, Solonin A, Zakharova M, Severinov K. 2005. Transcription regulation of the EcoRV restriction-modification system. Nucleic Acids Research. 33(21):6942-6951.

Shabbir MA, Hao H, Shabbir MZ, Wu Q, Sattar A, Yuan Z. 2016. Bacteria vs. Bacteriophages: Parallel Evolution of Immune Arsenals. Frontiers in Microbiology. 7:1292.

Shea MA, Ackers GK. 1985. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. Journal of Molecular Biology. 181(2):211-230.

Singh V, Braddick D, Dhar PK. 2017. Exploring the potential of genome editing CRISPR-Cas9 technology. Gene. 599:1-18.

Sridhar J, Gunasekaran P. 2013. Computational small RNA prediction in bacteria. Bioinformatics and Biology Insights. 7:83-95.

Staron A, Sofia HJ, Dietrich S, Ulrich LE, Liesegang H, Mascher T. 2009. The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) sigma factor protein family. Molecular Microbiology. 74(3):557-581.

van der Oost J, Jore MM,Westra ER, Lundgren M, Brouns SJ. 2009. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends in Biochemical Sciences. 34(8):401-407.

Vasu K, Nagaraja V. 2013. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiology and Molecular Biology Reviews. 77(1):53-72.

Westra E., Pul U, Heidrich N, Jore MM, Lundgren M, Stratmann T, Wurm R, Raine A, Mescher M, Van Heerevald L, et al. 2010. H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Molecular Microbiology. 77(6):1380-1393.

Williams K, Savageau AM, Blumenthal RM. 2013. A bistable hysteretic switch in an activator–repressor regulated restriction–modification system. Nucleic Acids Research. 41(12):6045-6057.

Wilson GG. 1991. Organization of restriction-modification systems. Nucleic Acids Research. 19(10):2539-2566.