OXFORD

Briefings in Bioinformatics, 16(5), 2015, 795–805

doi: 10.1093/bib/bbu045
Advance Access Publication Date: 3 December 2014
Paper

Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks

Patricia Thébault, Romain Bourqui, William Benchimol, Christine Gaspin, Pascal Sirand-Pugnet, Raluca Uricaru and Isabelle Dutour

Corresponding author. Patricia Thébault, University of Bordeaux, Laboratory of Computer Science of Bordeaux, France. Tel: +33 (0)5 40 00 60 95, E-mail: thebault@labri.fr

Abstract

The revolution in high-throughput sequencing technologies has enabled the acquisition of gigabytes of RNA sequences in many different conditions and has highlighted an unexpected number of small RNAs (sRNAs) in bacteria. Ongoing exploitation of these data enables numerous applications for investigating bacterial transacting sRNA-mediated regulation networks. Focusing on sRNAs that regulate mRNA translation in trans, recent works have noted several sRNA-based regulatory pathways that are essential for key cellular processes. Although the number of known bacterial sRNAs is increasing, the experimental validation of their interactions with mRNA targets remains challenging and involves expensive and time-consuming experimental strategies. Hence, bioinformatics is crucial for selecting and prioritizing candidates before designing any experimental work. However, current software for target prediction produces a prohibitive number of candidates because of the lack of biological knowledge regarding the rules governing sRNA–mRNA interactions. Therefore, there is a real need to develop new approaches to help biologists focus on the most promising predicted sRNA–mRNA interactions. In this perspective, this review aims at presenting the advantages of mixing bioinformatics and visualization approaches for analyzing predicted sRNA-mediated regulatory bacterial networks.

Key words: bacterial small RNAs; regulatory network; enrichment methods; visualization software

Patricia Thébault is a full-time Associate Professor of the University of Bordeaux and a permanent member of the Laboratory of Computer Science of Bordeaux. Her research focuses on System Biology.
Romain Bourqui is a full-time Associate Professor of the University of Bordeaux in France and a permanent member of the Laboratory of Computer Science of Bordeaux. His current research interests are information visualization, biological data visualization, graph drawing and graph clustering.
William Benchimol is a Research engineer at University of Bordeaux.
Christine Gaspin is a full-time senior researcher of the French National Institute for Agricultural Research and the director of the Bioinformatics facility of Toulouse. Her main research interest focuses on bioinformatics of noncoding RNAs and related gene regulation networks.
Pascal Sirand-Pugnet is a full-time Associate Professor of the University of Bordeaux in France and the director of the Genome-Transcriptome facility of Bordeaux. His research interest focuses on microbiology, comparative genomics of mollicutes and synthetic biology.
Raluca Uricaru is a full-time Associate Professor of the University of Bordeaux in France and a permanent member of the Laboratory of Computer Science of Bordeaux. Her research focuses on comparative genomics, from the sequence alignment point of view, and on Next-Generation Sequencing problems.
Isabelle Dutour is a full-time Associate Professor of the University of Bordeaux in France and a permanent member of the Laboratory of Computer Science of Bordeaux. Her research focuses on bioinformatics and biological networks.
Submitted: 5 August 2014; **Revised (in revised form):** 5 November 2014

Introduction

The 'regulatory RNA' field is expanding rapidly, and exciting new perspectives have recently emerged (for review see [1, 2]) because of the wide spectrum of regulatory functions that have been identified. Moreover, recent data provided by high-throughput technologies have shown that small RNAs (sRNAs) are much more represented in the bacterial world than previously expected [3]. In contrast to eukaryotic sRNAs (e.g. siRNAs and miRNAs), which have a size of <30 nucleotides, bacterial sRNAs show a wide variety of different structures and sizes, ranging roughly from 50 to 500 nucleotides (513 nt for the *Staphylococcus aureus* RNA III), and are known to be key players in cell regulation (positively or negatively) by interacting with proteins or/and mRNA molecules.

In bacteria, sRNAs are involved in fine-tuning gene expression by many biological processes, such as the modulation of transcription, translation, mRNA stability and DNA maintenance or silencing [4, 5]. Furthermore, crucial regulatory roles of sRNAs have also been noted in the establishment of virulence in several bacterial pathogens, such as *Vibrio cholerae* [6], *S. aureus* [7] and *Listeria monocytogenes* [8]. Moreover, using a genomic comparative analysis, Mandin *et al.* [9] identified a specific pathogenic sRNA subset that includes the pathogenic bacterium *L. monocytogenes* and the nonpathogenic *Listeria innocua*.

The regulation of gene expression by transacting sRNAs generally involves base-pairing with the mRNA. An example of an accepted mechanism of negative regulation involves a transacting sRNA, which will mask the interaction with its mRNA target by base-pairing at the Shine–Dalgarno (SD) site to prevent translation. The change resulting from the interaction between the sRNA and its target modifies the 5′UTR structure of the mRNA and/or blocks ribosomal recruitment. Therefore, the effect of such regulation can either positively or negatively affect the translation or stability of mRNAs, depending on both the folding of the two RNAs and the location of the interaction in the 5′UTR of the mRNA. As is the case for eukaryotic microRNAs, bacterial sRNAs may target several mRNAs [10], and mRNAs may be regulated by one or more sRNAs [11], forming complex interdependent regulation networks. Although high-throughput sequencing technologies have enabled the rapid acquisition of many sRNA sequences (Figure 1), the identification of their targets is still challenging, as the time and cost required for a single experimental validation remains limiting. Therefore, *in silico* methods to help biologists select the most meaningful sRNAs are necessary. Such methods should integrate a maximum amount of knowledge to increase the biological relevance while predicting sRNA–mRNA interactions. However, regardless of the bioinformatics software used for predicting targets, the lack of precise information for the base-paring rules often results in a prohibitive number of predictions [12], even for a small bacterial genome. Therefore, additional approaches are currently being developed to exploit these imperfect predictions.

This review covers two aspects of regulatory network constructions, with a double focus on *in silico* methods dedicated to sRNA–mRNA predictions and data mining tasks.

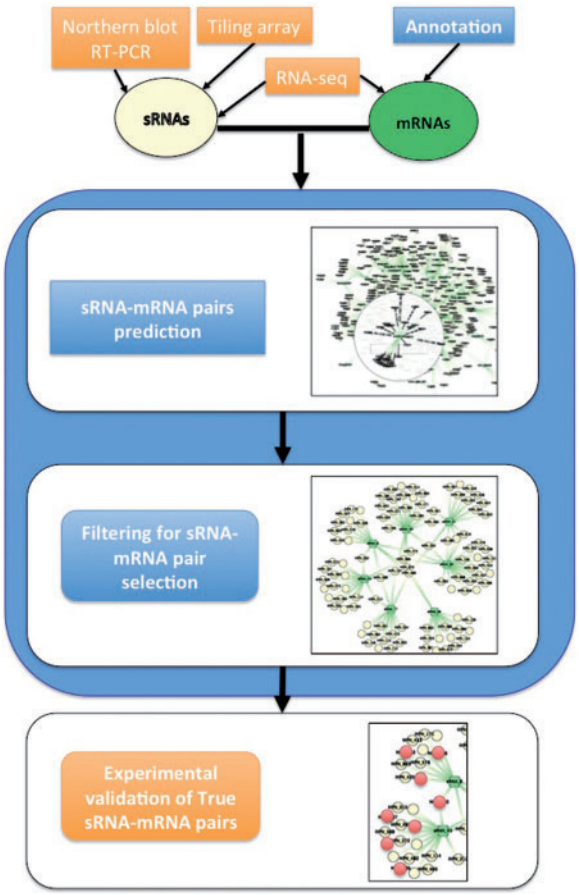

Figure 1. Overview of the exploitation of experimental sRNAs while investigating sRNA-mediated regulatory networks. Experimental and *in silico* stages are given in boxes, respectively, colored in orange and blue. The regulatory network is composed of two node types: sRNA and mRNA. Edges between sRNA–mRNA pairs represent interactions between both RNAs. The focus of this review is presented in the central blue rectangle. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

Bacterial sRNAS

Discovering sRNAs by experimental approaches

Two main experimental approaches for deciphering bacterial transcriptomes [3] have been widely described and used for various applications. The first approach consists of the assessment of whole genome tiling arrays by designing probes that cover the complete genome. A limitation of this approach is its dependence on the design of the probes, which is time- and cost-consuming. The second and more frequently used approach is based on a high-throughput sequencing approach following an sRNA enrichment step (see [13] for review). This experimental approach offers the advantage of providing quantitative and qualitative data at reasonable cost [13].

However, obtaining the exhaustive catalog of sRNA encoding genes in one bacterium requires the design of an extended protocol with many variable conditions. For instance, two high-throughput sequencing studies have been performed to analyze the transcriptome of *Escherichia coli* [14, 15]. Raghavan *et al.* [14] designed two conditions according to the absence and presence of Mg^{2+} (involved in the action of the chaperone HFQ for guiding the base-pairing step), whereas Shinhara *et al.* [15] used a single condition during the cell's exponential growth phase in a minimal medium with a unique glucose source. Both studies proposed a list of sRNAs containing most of the 80 known *E. coli* sRNAs and 63 and 113 new sRNA genes, respectively. It is noteworthy that the sequence comparison of the two sets only resulted in a small overlap, specifically, 5 of the 176 new sRNAs. This observation supports the assumption of a strong dependence of sRNA expression on external conditions.

When exploiting RNA-seq or tiling arrays, it must also be mentioned that additional information of particular interest can be added. First, the position of the transcription start site (extremity of the 5′UTR region) can be estimated, which is helpful for specifying sRNA interaction regions onto mRNA transcripts. Second, the co-expression of sRNAs and mRNAs from different conditions can be used to infer putative interactions (a recent work illustrates such analyses with *E. coli K12*) [16]. However, co-expression is not sufficient to prove a direct interaction between sRNAs and mRNAs, and as mentioned by the authors, further bioinformatics and experimental approaches are needed to prove sRNA–mRNA interactions.

What is known about sRNA–mRNA pair features

According to experimental data, the length of the interacting region between sRNAs and mRNAs varies from 5 to 20 bases, with some examples where only few base pairs are necessary [1, 17]. For most of the sRNAs involved in negative regulation, the interaction occurs in the 5′UTR mRNA region in the proximity of the start codon and the SD site. However, the rules governing the interaction have not been fully elucidated, and this lack of information has been noted in several recent publications. In this section, we summarize the features that have been observed or demonstrated, considering both RNA sequences and the experimental data available in the literature.

Richter *et al.* [12] investigated the conservation, from an evolutionary point of view, of sRNA–mRNA interacting regions. By comparing the sequence content, they observed a stronger conservation in sRNAs than in mRNAs. The significance of accessibility from the secondary structural point of view for the interacting regions has also been investigated experimentally by various groups [12, 18], with the general conclusion that the importance of accessible regions remains difficult to evaluate.

Other features of the biochemical functions of the targets are of interest. Gottesman *et al.* [19] assessed some main sRNA functions according to their functional mRNA targets and reported the following: (i) repression of the expression of membrane proteins, (ii) regulation of metabolic activity and (iii) modulation of the synthesis of transcription factors. Regarding the first function, because of the lack of knowledge, it cannot be explained why the cell would promote sRNA-mediated regulation. Focusing on the metabolism of *Mycobacterium tuberculosis*, Gottesman *et al.* hypothesized that the multiple targeting of several enzymes by one sRNA could be a rapid and economic way to adapt the metabolism of the cell to stressful external conditions [20]. Finally, the multiple regulation of one transcription factor by different sRNAs could be beneficial for the cell, enabling rapid coordination of the regulation of different biological pathways.

More formally, Beisel *et al.* [21] suggested classification of regulatory motifs according to their topological and biological features. Hence, the Single-Input Module (SIM) motif is represented by one sRNA interacting with multiple targets. These multiple mRNAs can be related to identical pathways or biological processes, and the requirement for only one regulating molecule allows the cell to rapidly adapt itself. When several stressful conditions appear, the cell can coordinate the regulation of disparate mRNAs by several sRNAs and vice versa by forming a Dense Overlapping Regulon (DOR) motif.

Skippington *et al.* [22] exploited this formalism to describe the predicted orthologous *E. coli* sRNAs in 27 closely related enterobacteria. Based on the type of the regulation motif, they deduced that sRNAs involved in a SIM motif may be more conserved through the phylum of studied bacteria. This observation is consistent with the hypothesis that evolutionary pressure can preserve the base-pairing capacity of sRNAs [11, 18]. As the examples of sRNA-mediated regulations are becoming increasingly complex, the idea that many key cellular processes may be governed by sophisticated regulation networks involving proteins and sRNAs has emerged. Therefore, developing predictive tools to describe such networks at the scale of the cell has become a challenging goal.

Methods and tools to construct sRNA-mediated regulatory networks

Two steps are necessary for the construction of a predicted sRNA-mediated regulatory network. First, it is necessary to acquire the two types of nodes corresponding to mRNAs and sRNAs. Both can be obtained by RNA-seq experimental approaches. Without such experimental data at the genome scale, CDS can be obtained from existing annotated genes, whereas putative sRNAs can be extracted from the RFAM database [23]. Second, it is necessary to predict interactions between sRNAs and their putative mRNA targets, which will be represented as edges between the two types of nodes in the graphs.

Predicting sRNA–mRNA pairs

To detect putative RNA–RNA interactions (for review see [24, 25]), a first class of sequence alignment methods, such as BLAST [26] or ssearch [27], can be used to search for long stretches of complementarity between the mRNA target and the sRNA reversed query sequence. The identity matrix has to be modified to include complementarity, rather than similarity, while considering wobble G–U pairs. To better consider the thermodynamic model in the alignment, the Smith–Waterman algorithm was used as the basis of a variety of implementations, including RiSearch [28] and TargetRNA [29]. Other approaches include specialized versions of RNAfold [30], such as RNAduplex [31] and RNAplex [32]. Such implementations search for the joined minimum free energy (MFE) structure, and they essentially differ by considering bulges and internal loops in different ways. All of these approaches have the common feature of considering intermolecular interactions, while ignoring intramolecular interactions.

A second class of more sophisticated bioinformatics methods takes intramolecular interactions into account while computing intermolecular interactions. The simplest such methods compute the optimal folding of a joined sequence formed by the concatenation of the mRNA, a linker and the sRNA. A modified version of RNAfold is used to compute the interaction, while applying a specific treatment to the linker sequence. Examples of such methods include pairfold [33] and RNAcofold [34]. The main drawback of this approach is that it does not predict pseudoknots. Accordingly, it cannot predict a class of joint structures, such as kissing hairpins.

The third class of methods starts from the accessible sites of each sequence, which are subsequences of consecutive nucleotides not involved in intramolecular base-pairing. Approaches such as RNAup [35] and IntaRNA [36] rely on the calculation of the partition function by combining both the inter- and intramolecular energies to provide the best energy combination. The complexity of the approach restricts the number of accessible sites. An improvement to the last approach can be found in extensions of the IRIS approach [37]. An interesting study of the joint secondary structure prediction problem can be found in

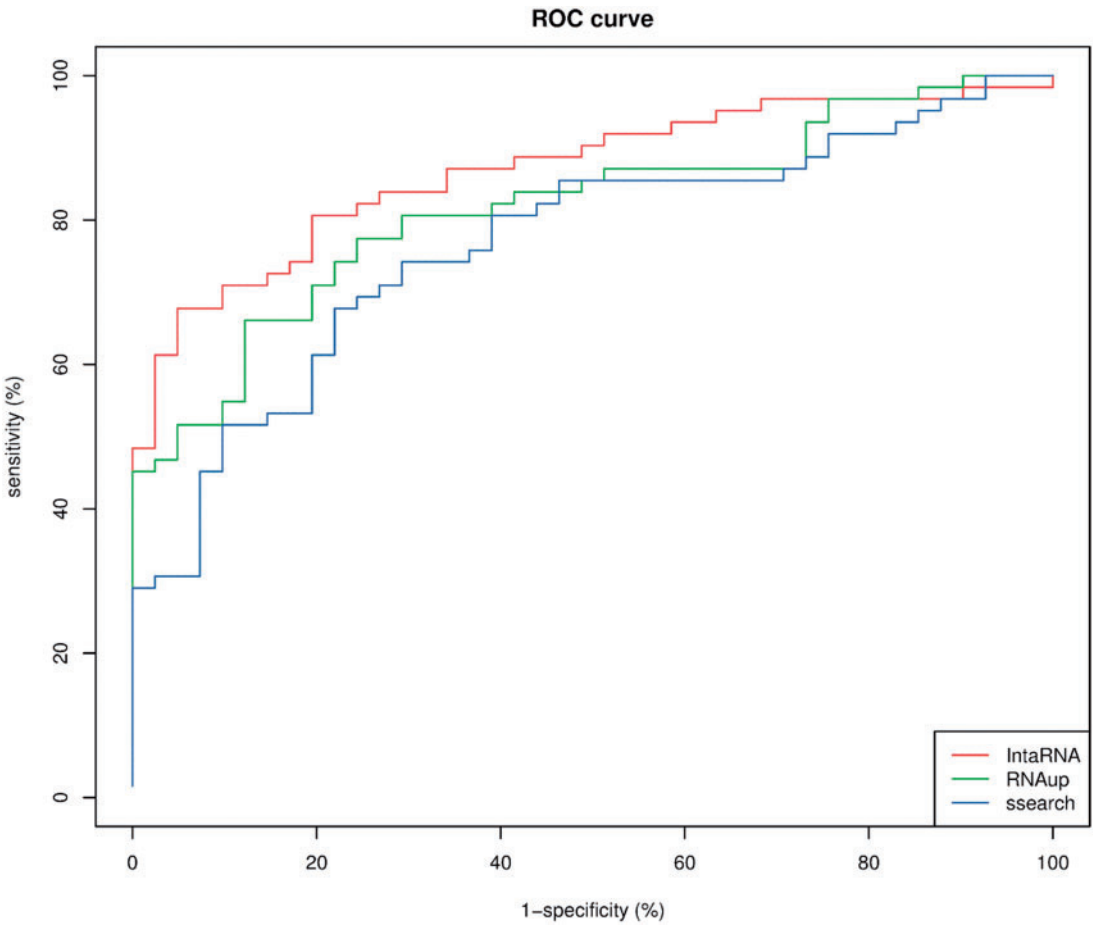

Figure 2. ROC curves. These ROC curves give a sensitivity/specificity report for a software benchmark based on the experimental data provided by sRNATarBase. A colour version of this figure is available online at BIB online: http://bib.oxfordjournals.org.

[38], where the general RNA–RNA interaction problem is shown to be an NP-hard problem under three different models with growing complexity.

In contrast, comparative and probabilistic methods exploit the coevolution of interacting nucleotides and have shown advantages over MFE methods [39]. These methods rely on the identification of conserved structural motifs by using signals of the sequence covariation between species that are phylogenetically related [40]. Their performance is dependent on the quality of the input alignments, and the main drawbacks include a weaker prediction efficiency and the need for a high-capacity CPU.

In this review, we focus on three software programs, RNAup, IntaRNA and ssearch, to illustrate their sensitivity and specificity. The benchmark used is composed of a set of 43 validated sRNAs and 52 mRNAs, with 62 true interactions and 41 noninteractions (see supplementary data) collected from sRNAtarBASE [41] (mainly composed of *E. coli* and *Salmonella enterica* sRNAs). The results of the predictions are represented graphically with receiver operating characteristic (ROC) curves in Figure 2. Based on these results, it is important to mention that sophisticated methods and naïve approaches give comparable results for this data set. The first two software programs exploit sequence information together with structural information, whereas the last one only addresses sequence complementarity. RNAup and IntaRNA, which rely on RNA folding, show a slightly better compromise between sensitivity and specificity. To further our investigation, we used randomly shuffled sequences to evaluate the statistical significance of the RNA composition. We focused on gcvB sRNA, for which (1) more than 10 targets have been experimentally validated and (2) involvement in the pervasive regulation of more that 1% of *Salmonella* genes was previously suggested [10]. As shown in Figure 3, the distribution of the scores given by the three software programs was not higher than that with a real sequence. These observations suggest that use of only the sRNA alphabet with the short length of the interaction regions and the lack of biological information for the interacting rules does not enable discrimination of the results given by the real gcvB sequence.

Facing the general lack of biological information, it is important to obtain the best sensitivity from the first stage of prediction, before the data mining stage, and then to exploit additional biological features to prioritize the biological relevance of the predictions.

Data mining on the sRNA-mediated regulatory network

A widely studied area in bioinformatics relies on biological databases, where a huge amount of biological multipurpose data is available. The integration of these data through data mining approaches can help to formulate relevant arguments when prioritizing predictions for experimental validations. Moreover, for the last decade, a particular interest has been given to enrichment methods to exploit such information by identifying statistically informative annotations.‘

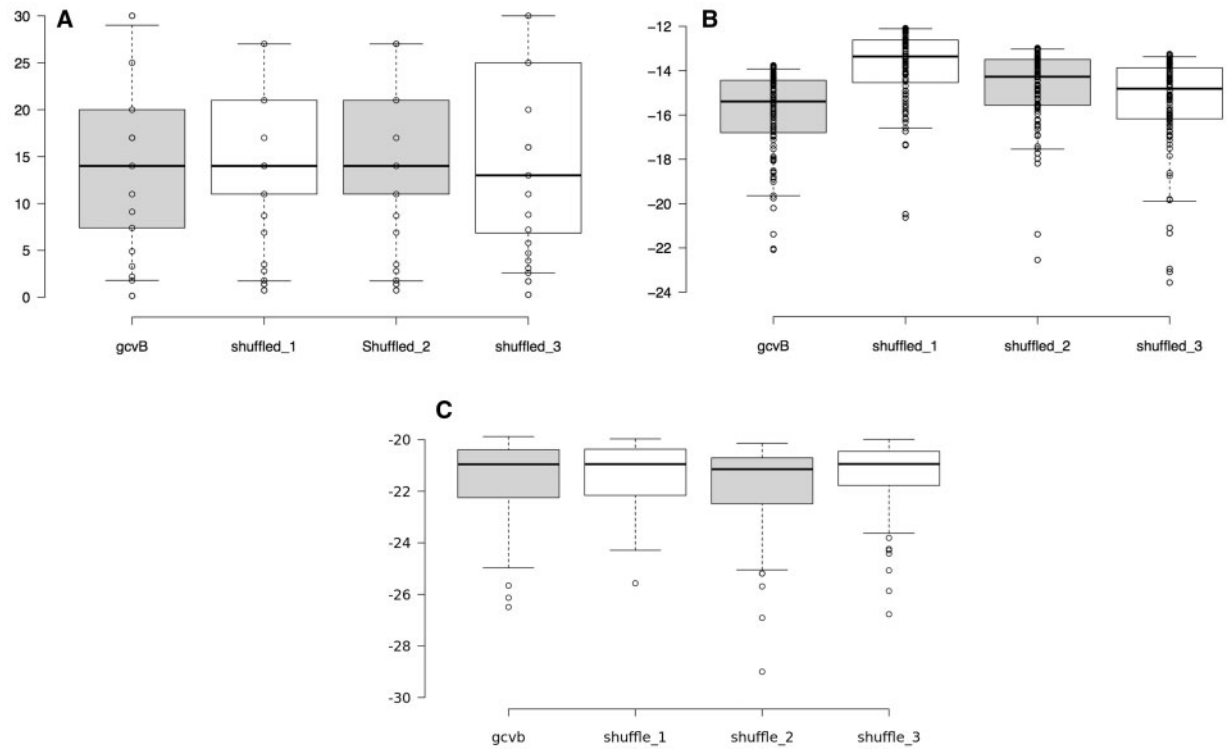

Figure 3. Distribution of the similarity score given by ssearch (**A**) and energy values given by IntaRNA (**B**) and RNAup (**C**). Four sRNA sequences (GcvB and 3 GcvB randomly shuffled sequences, with a global similarity of 50%, with 20% of gaps, and no conserved motifs with a length > 4) were used to calculate the 100 best-scoring interactions with the 5′UTR of the mRNAs in *E. coli*.

Functional enrichment analysis

In enrichment methods, the underlying idea is to compare groups of genes that are put together according to biological criteria (for instance, groups of mRNAs that share a predicted interaction with one unique sRNA) with a second catalog of gene groups derived from a biological database and sharing a biological annotation [42, 43].

The two most widely used biological databases are the Gene Ontology (GO) and KEGG databases, which describe the hierarchical relationships between semantic annotations describing the functional activity of proteins and the metabolic pathway assignation of enzymes, respectively. Other databases, such as MetaCyc [44], can be used to complete KEGG information because they attempt to provide more exact, true biological pathways, for example, by separately recording the different pathway variants observed in different organisms.

The enrichment strategy provides overrepresented mRNA annotations according to a molecular database of interest. A statistical score is computed according to the presence of enriched annotated terms (e.g. GO or enzymatic activity) in a reference group, called the background (e.g. the total mRNAs of the genome of interest). Following and extending this strategy, more than 68 software programs have been proposed and reviewed [42]. The choice of one of these software programs is not an easy task, and some of the following ranking key points have to be prioritized.

First, the use of standard formats for the mRNA identification is crucial. Moreover, typically, only a few gene model organisms are indexed, regardless of the biological database that is used.

A second key point to consider is the list of available databases, as each database is specialized in different biological features. The GO database is the most widely used and partitions genes into three main groups corresponding to their associated biological processes, cellular localization and molecular functions. It is nevertheless meaningful to note that the GO hierarchy is based on DAGs (Directed Acyclic graph, a graph where a child annotation can be linked to several parent annotations, and vice versa). In consequence, one gene can contain several annotations on multiple levels of precision. This redundant information has been described in several works [45, 46] to increase the difficulty of interpreting results (for instance, relying on various genes with different levels of precision in their annotation, according to the GO database). Recently, Dutkowski *et al.* [47] investigated this problem and proposed a new level of compact ontologies to filter out the GO node annotation. Another approach, implemented in GoSemSim [45], combined the similarity between several GO terms and their topological localization in the GO graph to compute a similarity score. Other tools integrate neighborhood information from the local GO graph into each term via an enrichment computation (for review see [42]); for instance, Bauer *et al.* [48] provide a visualization output to interact with the GO graph.

Finally, Väeremo *et al.* [49] investigated the impact of different statistical metrics but showed no significant differences. Conversely, Naeem *et al.* noted that the quality and the type of data may influence the results through the statistical tests used [43].

By considering all of these key features in the case of non-model bacteria, the web server David-WS [50] proved to be an interesting solution. This platform, which also provides a web server and web services for a huge list of bacteria, integrates a large list of databases, ranging from generalized databases, such as Uniprot, to specific databases that exploit protein domain information or sequence conservation. In addition, the web server David-WS enables the simultaneous exploitation of

the enriched semantic terms given by all databases. After the computation of an enrichment score for each database, an additional clustering step is performed to group genes that share similar annotations, regardless of the database. This clustering step is helpful when dealing with the redundancy problem created by the GO annotation because it adds supplementary information to the biological knowledge.

Analyzing regulatory network with visualization

Information visualization has now been established as a fruitful strategy to address the problem raised by the abundance of information [51]. Schneiderman [52] provided some recommendations for the visual exploration of data, which are now known as the Visual Information-Seeking Mantra: 'Overview first, zoom and filter, then details-on-demand'. These recommendations have been widely applied to biological networks, as the size and complexity of newly acquired networks prohibit manual representation (for an overview on biological data visualization, the reader can refer to a special issue of Nature Methods, issue S3, Volume 7). Figure 4 shows how information visualization can fit the analytical process performed by scientists to extract knowledge from large and complex biological data. Information visualization produces visual representation of input data to help the expert to build hypotheses and test them through bioinformatics algorithms. To increase the user's level of confidence, bioinformatics algorithms can either reduce the scope of the study or integrate additional biological knowledge. Such an analytical loop produces new data that can potentially be used to motivate experimental studies.

During the past decade, the community has produced a large number of software pieces for biological network analysis and visualization (for a general review see [53, 54]). Two categories of visualization tools can be distinguished: general network visualization tools and network-dedicated tools. Among the general tools, one of the most famous is Cytoscape [55], and others include Tulip [56], Pajek [57] and Gephi [58]. These tools support many features and are generally suitable options for the visualization, analysis and exploration of large networks. Additionally, some of these tools integrate a plug-in management system or a script interpreter [56, 58, 59], allowing experts to increase the number of supported features (for instance, by adding a dedicated clustering algorithm). Although these tools have proven their efficiency in general cases, because of their generic nature, they may not fulfill the requirements of all users, such as in cases of well-defined and network-specific tasks. This leads to the second category of visualization tools, which are dedicated to particular types of networks. For instance, some of these tools are designed for the analysis of protein–protein interaction networks [60], metabolic networks [61, 62], signaling networks [63, 64], gene regulatory networks [65] and sRNA-mediated regulatory networks [66] (for detailed reviews, the reader can refer to [67]). Usually, these tools integrate only the features needed to solve tasks specific to the type of network they are related to. Restricting the number of features facilitates the manipulation and exploration of the data, as the abundance of features can make the graphic interface tedious to learn.

The first aspect that a visualization system has to address is the representation of the network. The main difficulty when building a readable representation is assigning coordinates (in 2D or 3D) to each node of the network. There currently exists a plethora of layout algorithms for particular classes of graphs, such as trees, planar graphs or DAGs, for which effective solutions have been found that give good results, not only in terms of time/space complexity but also in terms of aesthetic criteria.

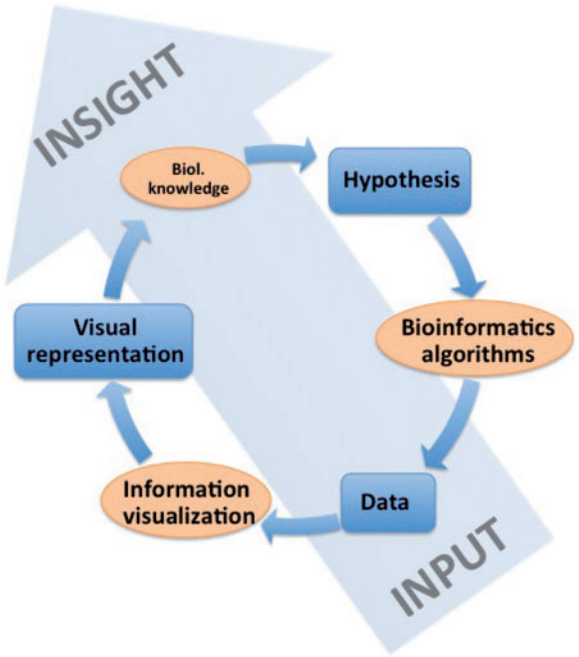

Figure 4. The iterative analysis process that exploits visualization and bioinformatics for getting insight from the input data. The orange ovals depict the involved skills and scientific domains, while the blue squares represent the intermediate results. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

However, real-world graphs, and in particular biological networks, usually do not belong to these classes. Finding an algorithm that gives good results (in terms of aesthetic criteria and information emphasized) for arbitrary graphs is a difficult problem. The most popular approach for drawing such graphs is the force-directed (or organic) approach, as it produces visually pleasant and structurally significant results. Many general visualization tools and network-specific tools [59, 61, 65, 68] support such algorithms. Dedicated algorithms have also been proposed to address specific types of networks and to fulfill user expectations (e.g. [64] for signaling pathways, [69] for metabolic networks). In addition to network layout, some biological network visualization tools allow the user to import and visualize experimental data (such as gene expression or metabolite concentrations). These tools usually integrate visualization dedicated to multidimensional data, such as scatterplots, parallel coordinates and microarray-like representations [55, 61, 68]. Other tools [59, 61, 64] support contextual visualization of time series data. In such cases, the experimental data are superimposed on the nodes of the network by changing their sizes or colors or by using charts or heat-maps.

To facilitate the understanding of the processes occurring in a network and to investigate the putative cellular organization, selection of subparts of the entire data is crucial. Such filtering can be achieved according to three main criteria. First, the subnetwork of interest can be selected by filtering out elements according to the provided data, such as gene expression in a gene regulatory network or metabolite concentrations in metabolic pathways/networks [55, 59, 61, 65]. Another interesting method is based on the computation of graph metrics [56, 57, 61], which has been described in several reviews [70, 71]. In general, when analyzing biological networks, inferring topological metrics helps in generating hypotheses on the robustness and/or dynamics of the network [72, 73].

Finally, use of expert knowledge is of utmost importance to reduce the scope of the analysis by focusing on subparts according to biological information. The integration of biological databases is currently one of the main challenges in the data mining community, and many studies have focused on enrichment analyses for driving the integration of multipurpose 'omics' or annotated data. Although such database integration enables the user to perform more realistic analysis, few tools integrate visualization approaches [74], and if so, they are dedicated to specific biological networks [75, 76].

An illustration with *E. COLI*

The input of the visualization environment is a list of sRNAs and mRNAs extracted from databases and complete genome annotations. When building the network, the two types of genes are automatically assigned to nodes in a bipartite graph where different colors and shapes can distinguish between them, and two nodes are linked by an edge if an interaction between the corresponding RNAs has been predicted. The resulting biological networks can have a prohibitive size due to the large amount of predicted interactions, which hinders their exploration and analysis. Additional information may enrich that first network in an integrated way. For example, the enriched functional annotations of mRNAs allow highlighting a group of genes that belong to the same functional class.

Therefore, the main objective is to extract information from that amount of data, by focusing on subnetworks of interest that represent smaller groups of related genes. With the clustering and layout algorithms, visualization approaches offer a way to integrate and synthetize information and to facilitate the exploration and the manipulation of these large biological networks. First, grouping together nodes that exhibit similar features or behavior can generate subnetworks. Second, newdedicated layout algorithms can be applied to compare the different subnetworks. To contribute to this goal, it is crucial to use an appropriate software to cluster and display nodes according to the information of interest. The filtering algorithms offer generic and specific functionalities to filter out uninteresting elements. In the next paragraph, an example of a strategy is proposed that uses, as the point of departure, the combination of two methods of clustering (or filters) based on, respectively, annotations and topological features. For example, the biologist can extract subnetworks according to particular annotations. He can also select node's neighborhood (e.g. sRNA's putative targets) according to the features of the predicted interactions. The visualization of a specific sRNA region interacting with many other mRNAs may be discovered and considered as an evolutionary constrained region. Multiple views then help to visually mine and facilitate the comparison of two or more subnetworks (e.g. showing validated versus predicted interactions).

We illustrate the advantages of using a combined strategy based on bioinformatics and visualization approaches when dealing with a huge number of predicted targets. In particular, we investigated the sRNA-mediated regulation network related to biofilm formation in the bacterium *E. coli*. Recent publications focusing on the regulation of biofilm formation in enterobacteria are of interest for our case study. Mika *et al.* [11] highlighted the central role of three proteins (FlhD, RpoS, CsgD) whose synthesis is regulated at a transcriptional level by 8 sRNAs (OmrA, OmrB, ArcZ, Mcas, OxyS, DsrA, RprA and GcvB) in a representative DOR motif configuration. Focusing on the same biological process, VanPuyvelde *et al.* [77] reviewed the connections

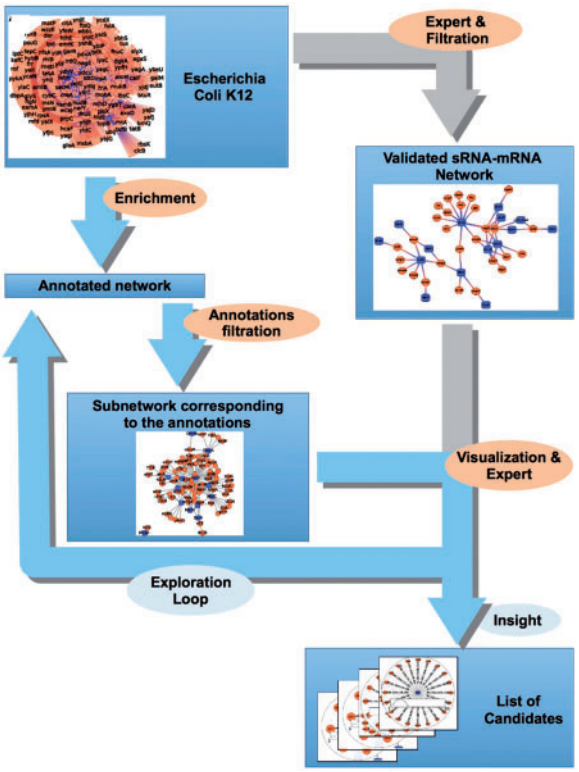

Figure 5. The analysis process mixing visualization and bioinformatics approaches. The blue arrows show the iterative process where the network is successively pruned according to enrichment analysis and visual exploration. The gray arrows show the reconstruction of the validated subnetwork and its exploitation to note conserved features and to identify promising putative interactions in the predicted networks. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

between the biofilm regulatory subnetwork and other biological processes, including outer membrane homeostasis, the two-component system, synthesis of extracellular curli and flagellar motility. These processes are interconnected and occur under stressful conditions, such as when the cell has to shift from a growth to a stationary state. Moreover, the amount of published data showing the impact of various biological processes in the formation of biofilm through the action of key regulators is increasing and illustrates the complexity of regulation [77]. Altogether, these processes involve 11 sRNAs (MicM, CyaR, lpeX, Mica, MicC, MicF, Mcas, OmrA, OmrB, RseX and RybB), some of which are also directly involved in biofilm formation. Taken together with an additional report on the pervasive regulation action of GcvB in *Salmonella* [10], we obtained a list of 15 known nonredundant sRNAs, which have been shown to act directly or indirectly on biofilm formation. Investigating the key regulators involved in biofilm formation with regard to various biological networks was the point of departure to illustrate the advantages of integrating multiple computational approaches. To do so, we used rNAV [66], a visualization software program that encapsulates the process of building, enriching and exploring sRNA-mediated networks.

Figure 5 shows the analysis process we followed in this study. One of the key advantages of the visualization strategy is to support an iterative process where the network is successively pruned until a given hypothesis is confirmed. The first step is to build an *E. coli* network using the 15 sRNAs involved in biofilm regulation using INTARNA. We obtained a network composed of 2913 nodes as 5′UTR mRNAs and 6705 edges modeling

the predictive interactions. Such a prohibitive number of putative interactions make the interactive exploration of the data extremely difficult.

The second step of the analysis was to perform a functional enrichment. Using the annotated network, we filtered interesting annotations to reduce the scope of the study to a few RNAs. By visually investigating the regions of the sRNA where the predicted interactions occur, we could then identify promising putative interactions.

In this study, we focused on enriched annotations of interest [77] and extracted three subnetworks with the following keyword annotations (see Figure 6):

1. porin membrane proteins, with 13 sRNAs, 85 mRNAs and 159 potential interactions,
2. two-component system pathway, with 10 sRNAs, 15 mRNAs and 32 potential interactions and
3. flagellum processes, with 10 sRNAs, 13 mRNAs and 23 potential interactions

The addition of these three networks, together with already known interactions (when available), provides a new way of prioritizing new candidates. Hence, the known interactions are of particular interest for identifying further conserved features regarding the experimentally validated pairs of sRNA–mRNA (e.g. the conservation of the position within one sRNA targeting several mRNAs, the RBS interaction position onto the 5′UTR of mRNAs or the presence of an enriched annotation term extracted from the biological databases). We thus extracted the subnetwork corresponding to these validated interactions (see Figure 6). By visually comparing the predicted interactions in these four subnetworks of sRNA of interest (Figure 6 shows the predicted interactions of RybB and GcvB) and taking into

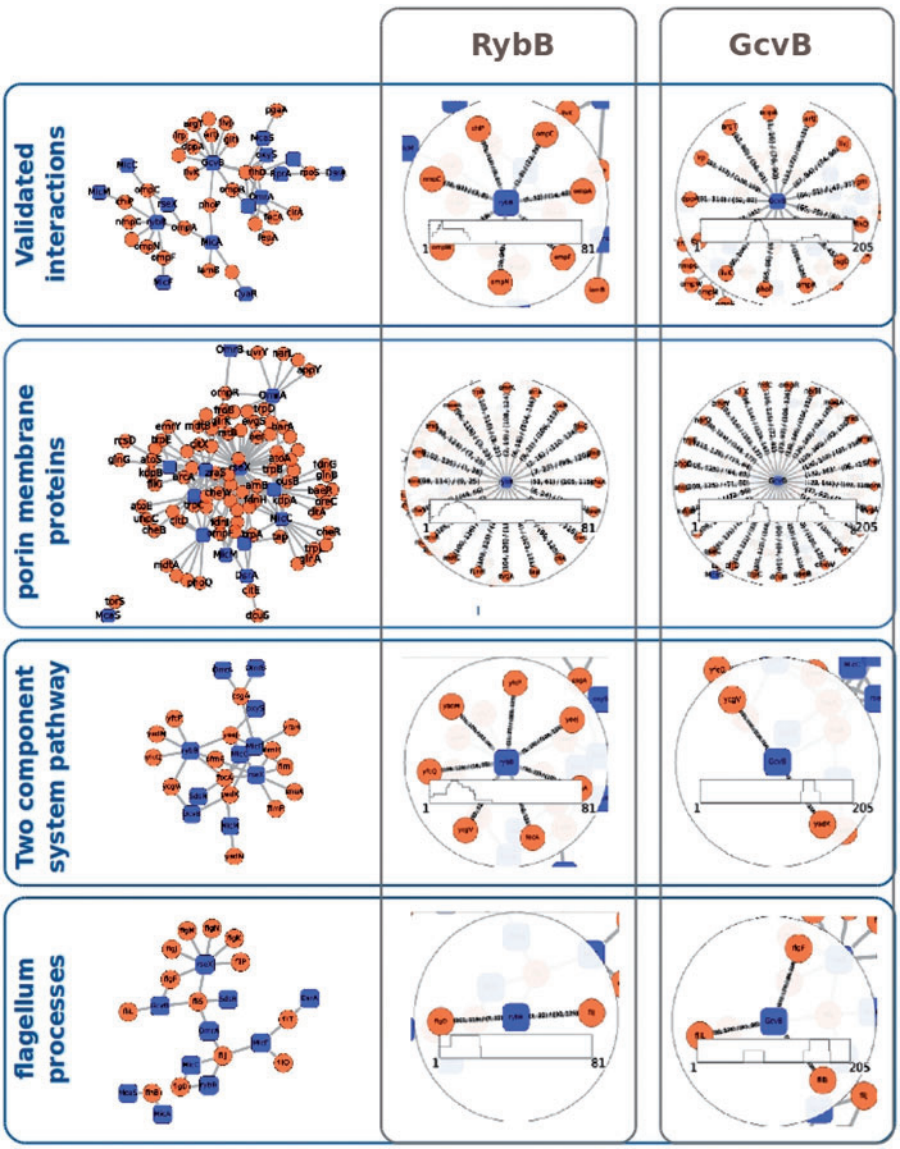

Figure 6. Exploration of the predicted graph using three enriched annotations and interacting regions. RNAs are represented by blue squares and mRNAs by orange circles. The first row shows to the validated interaction network. It depicts the network, together with a focus Rows 2, 3 and 4 show networks pruned according to annotations of interest (respectively, porin membrane proteins, two-component system pathway and flagellum processes). For each resulting subnetwork, according to the enrichment of interest, a focus is depicted on two sRNA and their targets: RybB and GcvB. For each base of one RNA, a curve is displayed. This curve shows the number of interactions each of its bases is involved in. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

account the conservation of interacting sRNA regions, we reduced the number of putative interactions. Following such an analysis process, we highlighted the experimentally validated interactions (7 for RybB and 12 for GcvB) and suggested new interaction candidates (30 for RybB and 28 for GcvB), which could all be potential areas for future investigation. Such types of regulation may support the connections between the biofilm regulatory subnetwork and other biological processes, as previously suggested [77].

This example illustrates the advantages of coupling bioinformatics and visualization approaches. One advantage involves the graphical representation of the multiple connections between RNAs to bring out hubs as SIM and DOR motifs. Another advantage is that this combined approach highlights interconnected biological processes, helping the biologist to obtain a global view of the regulation networks at the scale of the entire cell.

Discussion and conclusion

Because of the lack of biological information regarding these RNA–RNA interaction rules, prediction tools often yield a prohibitive number of candidates. To circumvent this problem, filtering of the predictions to focus on the most promising ones is crucial. The strategy relies on providing *a posteriori* multipurpose information, according to the heterogeneous data extracted from databases (using enrichment methods) and the visual exploration/analysis of the biological network (using visualization approaches). We reviewed the different methods in both domains in the context of biological networks. To address the large amounts of data generated in network prediction, the integration of additional biological information is often used to filter the results. To do so, enrichment methods allow the exploitation of 'omics' data (derived from multipurpose methods) to identify statistically significant subgroups of sRNA targets sharing an annotation feature.

The exploitation of visualization approaches is also of great interest to enable a general view of the multiple connections between RNAs (such as deciphering hubs) and to explore them while focusing on features for specific sRNAs or mRNAs. A mix of these methods is of great help for focusing on pertinent candidates, with multiscale arguments for designing experimental validation. Moreover, an sRNA-mediated regulatory network can be investigated to infer sRNA functions by indirectly tackling the functions of the targeted mRNAs. Considering that mRNA functions are inferred from annotated genomes, the identification of sRNAs and their relationships with mRNAs is the crucial step. The biological process in which targets are involved could be exploited to propose a functional annotation for their regulating sRNAs. Although more research is required to obtain a complete picture of all regulation networks at the cellular scale, efforts to combine experimental and predicted data in an integrated view will most likely help biologists to understand the complex interactions of an organism with its environment. Moreover, multilevel networks focusing on gene function and regulation will also enhance the ability to predict effects of genome engineering strategies that form the basis of most of the work in the emerging field of synthetic biology.

Even though this review is focused on bacterial sRNA analysis (the case application with *E. coli* can be applied to other bacteria), other applications could be conceived, such as addressing new types of RNAs. However, the biodiversity of interacting molecules that involve RNAs will have to be taken into account for adaptation of this strategy. As an example, when dealing with miRNA in eukaryotes, it will be necessary to take into account the scaling generated by the higher number of genes in eukaryotic cells. Because of the size of the data, both the data representation and the bioinformatics analysis may become more time-consuming. In the interaction prediction step, more appropriate tools should be used to specifically exploit the seed located at the 2–8 nt positions of miRNA as an anchor

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Because of the lack of biological information regarding these RNA–RNA interaction rules, prediction tools often yield a prohibitive number of candidates.
- To address the large amounts of data generated in regulatory network prediction, the integration of additional biological information is often used to filter the results.
- One strategy relies on providing a posteriori multipurpose information, according to the heterogeneous data extracted from databases (using enrichment methods) and the visual exploration/analysis of the biological network (using visualization approaches).

Funding

This work was done under the EVIDEN project (ANR 2010 JCJC 0201 01) and the BACNET project (ANR 2012 IA) supported by the ANR (France); and under the BioBRICK project, AAP Synthetic Biology Bordeaux France (SB2), 2014.

References

1. Storz G, Vogel J, Wassarman K. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* 2011;**43**:880–91.
2. Desnoyers G, Bouchard MP, Massé E. New insights into small RNA-dependent translational regulation in prokaryotes. *Trends Genet* 2013;**29**:92–8.
3. Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* 2010;**11**:9–16.
4. Romby P, Charpentier E. An overview of RNAs with regulatory functions in gram-positive bacteria. *Cell Mol Life Sci* 2010;**67**:217–37.
5. Brantl S, Brückner R. Small regulatory RNAs from low-GC Gram-positive bacteria. *RNA Biol* 2014;**11**:2007.
6. Bradley ES, Bodi K, Ismail AM, *et al*. A genome-wide approach to discovery of small RNAs involved in regulation of virulence in vibrio cholerae. *PLoS Pathogens* 2011;**7**:e1002126.
7. Toledo-Arana A, Repoila F, Cossart P. Small noncoding RNAs controlling pathogenesis. *Curr Opin Microbiol* 2007;**10**:182–8.
8. Toledo-Arana A, Dussurget O, Nikitas G, *et al*. The Listeria transcriptional landscape from saprophytism to virulence. *Nature* 2009;**459**:950–6.

9. Mandin P, Repoila F, Vergassola M, *et al*. Identification of new noncoding RNAs in Listeria monocytogenes and prediction of mRNA targets. *Nucleic Acids Res* 2007;**35**:962–74.
10. Sharma CM, Papenfort K, Pernitzsch SR, *et al*. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. *Mol Microbiol* 2011;**81**:1144–65.
11. Mika F, Hengge R. Small regulatory RNAs in the control of motility and biofilm formation in *E. coli* and Salmonella. *Int J Mol Sci* 2013;**14**:4560–79.
12. Richter AS, Backofen R. Accessibility and conservation: general features of bacterial small RNA-mRNA interactions? *RNA Biol* 2012;**9**:954–65.
13. Lasa I, Toledo-Arana A, Gingeras TR. An effort to make sense of antisense transcription in bacteria. *RNA Biol* 2012;**9**:1039–44.
14. Raghavan R, Groisman EA, Ochman H. Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Res* 2011;**21**:1487–97.
15. Shinhara A, Matsui M, Hiraoka K, *et al*. Deep sequencing reveals as-yet-undiscovered small RNAs in *Escherichia coli*. *BMC Genomics* 2011;**12**:428.
16. Modi SR, Camacho DM, Kohanski MA, *et al*. Functional characterization of bacterial sRNAs using a network biology approach. *Proc Natl Acad Sci USA* 2011;**108**:15522–7.
17. Künne T, Swarts DC, and Brouns JJ. Planting the seed: target recognition of short guide RNAs. *Trends Microbiol* 2014;**22**:74–83.
18. Peer A, Margalit H. Accessibility and evolutionary conservation mark bacterial small-RNA target-binding regions. *J Bacteriol* 2011;**193**:1690–701.
19. Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol* 2011;**3**:pii: a003798.
20. Arnvig KB, Young DB. Identification of small RNAs in Mycobacterium tuberculosis. *Mol Microbiol* 2009;**73**:397–408.
21. Beisel CL, Storz G. Base pairing small RNAs and their roles in global regulatory networks. *FEMS Microbiol Rev* 2010;**34**:866–82.
22. Skippington E, Ragan MA. Evolutionary dynamics of small RNAs in 27 Escherichia coli and Shigella genomes. *Genome Biol Evol* 2012;**4**:330–45.
23. Burge, SW, Daub J, Eberhardt R, *et al*. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 2013;**41**:D226–32.
24. Pichon C, Felden B. Small RNA gene identification and mRNA target predictions in bacteria. *Bioinformatics* 2008;**24**:2807–13.
25. Backofen R, Wolfgang RH. Computational prediction of sRNAs and their targets in bacteria. *RNA Biol* 2010;**7**:33–42.
26. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
27. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 1991;**11**:635–50.
28. Wenzel A, Akbasli E, Gorodkin J. RIsearch: fast RNA-RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics* 2012;**28**:2738–46.
29. Tjaden B. Computational identification of sRNA targets. *Methods Mol Biol* 2012;**905**:227–34.
30. Hofacker IL, Fontana W, Stadler PF, *et al*. Fast folding and comparison of RNA secondary structures. *Monatsh Chem* 1994;**125**: 67–88.
31. Gruber AR, Lorenz R, Bernhart SH, *et al*. The vienna RNA websuite. *Nucleic Acids Res* 2008;**36**:W70–4.
32. Tafer H and Hofacker IL. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* 2008;**24**:2657–63.
33. Andronescu M, Zhang ZC, Condon A. Secondary structure prediction of interacting RNA molecules. *J Mol Biol* 2005;**345**:987–1001.
34. Bernhart SH, Tafer H, Mückstein U, *et al*. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol* 2006;**1**:3.
35. Mückstein U, Tafer H, Hackermüller J, *et al*. Thermodynamics of RNA-RNA binding. *Bioinformatics* 2006;**22**:1177–82.
36. Busch A, Richter AS, Rolf B. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* 2008;**24**:2849–56.
37. Pervouchine DD. IRIS: intermolecular RNA interaction search. *Genome Inform* 2004;**15**:92–101.
38. Alkan C, Karakoç E, Nadeau JH, *et al*. RNA-RNA interaction prediction and antisense RNA target search. *J Comput Biol* 2006;**13**:267–82.
39. Meyer IM. Predicting novel RNA-RNA interactions. *Curr Opin Struct Biol* 2008;**18**:387–93.
40. Babak T, Blencowe BJ, Hughes TR. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics* 2007;**8**:33.
41. Yuan C, Jiayao W, Qian L, *et al*. sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. *RNA* 2010;**16**:2051–7.
42. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009;**37**:1–13.
43. Naeem H, Zimmer R, Tavakkolkhah P, *et al*. Rigorous assessment of gene set enrichment tests. *Bioinformatics* 2012;**28**:1480–6.
44. Caspi R, Altman T, Billington R. *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2014;**42**:D459–71.
45. Yu G, Li F, Qin Y, *et al*. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 2010;**26**:976–8.
46. Wang J, Zhou X, Zhu J, *et al*. GO-function: deriving biologically relevant functions from statistically significant functions. *Brief Bioinform* 2012;**13**:216–27.
47. Dutkowski J, Kramer M, Surma MA, *et al*. A gene ontology inferred from molecular networks. *Nat Biotechnol* 2013;**31**:38–45.
48. Bauer S, Grossmann S, Vingron M, *et al*. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* 2008;**24**:1650–1.
49. Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res* 2013;**41**:4378–91.
50. Huang DW, Sherman BT, Tan Q, *et al*. DAVID Bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* 2007;**35**:W169–75.
51. Thomas JJ, Cook KA. Illuminating the path: the research and development agenda for visual analytics. *IEEE Computer Soc* 2006;**26**:10–3.
52. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of the IEEE Symposium on Visual Languages*. 1996;336–3.
53. Suderman M, Hallett M. Tools for visually exploring biological networks. *Bioinformatics* 2007;**23**:2651–9.
54. Pavlopoulos G, Wegener AL, Schneider R. A survey of visualization tools for biological network analysis. *BioData Min* 2008;**1**:12.

55. Smoot ME, Ono K, Ruscheinski J, *et al*. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011;**27**:431–2.
56. Auber D. Tulip-a huge graph visualization framework. In P. Mutzel and M. Junger, editors. *Graph Drawing. Softwares, Mathematics and Visualization*. Springer-Verlag, 2003; 05–26.
57. Batagelj V, Mrvar A, Mutzel P, *et al*. *Pajek - Analysis and Visualization of Large Networks*. Berlin: Springer-Verlag, 2003, 77–103.
58. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: *International AAAI Conference on Weblogs and Social Media*, Paris, France, 2009.
59. Klukas C, Schreiber FI. Integration of -omics data and networks for biomedical research with VANTED. *J Int Biol* 2010;**7**:112.
60. Iragne F, Nikolski M, Mathieu B, *et al*. ProViz: protein interaction visualization and exploration. *Bioinformatics* 2005;**21**:272–4.
61. Dubois J, Cottret L, Ghozlane A, *et al*. Systrip: a visual environment for the investigation of time-series data in the context of metabolic networks. In: *Proceedings of the 16th International Conference on Information Visualisation (IV'12)*, Montpellier, France, 2012, 204–13.
62. Bourqui R, Cottret L, Lacroix V, *et al*. Metabolic network visualization eliminating node redundance and preserving metabolic pathways. *BMC System Biol* 2007;**1**:29.
63. Berger SI, Iyengar R, Ma'ayan A. AVIS: AJAX viewer of interactive signaling networks. *Bioinformatics* 2007;**23**:2803–5.
64. Barsky A, Munzner T, Gardy J, *et al*. Cerebral: visualizing multiple experimental conditions on a graph with biological context. *IEEE Trans Vis Comput Graph* 2008;**14**:1253–60.
65. Bourqui R, Westenberg MA. Visualizing temporal dynamics at the genomic and metabolic level. In: *Proceedings of the 13th International Conference on Information Visualization*, 2009, 317–22.
66. Dubois J, Ghozlane A, Thébault P, *et al*. *Genome-wide detection of sRNA targets with rNAV*. In: *Proceedings in 3rd IEEE Symposium on Biological Data Visualization*, 13-14 October 2013, Atlanta, USA, 2013, 81–8.
67. Agapito G, Guzzi PH, Cannataro M. Visualization of protein interaction networks: problems and solutions. *BMC Bioinformatics* 2013;**14**(Suppl 1):S1.
68. Funahashi A, Tanimura N, Morohashi M, *et al*. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico* 2003;**1**:159–62.
69. Lambert A, Dubois J, Bourqui, R. Pathway preserving representation of metabolic networks. *Comput Graph Forum* 2011;**30**:1021–30.
70. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**: 101–13.
71. Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. *Brief Bioinform* 2006;**7**: 243–55.
72. Müller-Linow M, Hilgetag CC, Hütt MT. Organization of excitable dynamics in hierarchical biological networks. *PLoS Comput Biol* 2008;**4**:e1000190.
73. Zamal F, Ruths D. On the contributions of topological features to transcriptional regulatory network robustness. *BMC Bioinformatics* 2012;**13**:318.
74. Pesch R, Lysenko A, Hindle M, *et al*. Graph-based sequence annotation using a data integration approach. *J Integr Bioinform* 2008;**5**:2.
75. Baumbach J, Wittkop T, Rademacher K, *et al*. CoryneRegNet 3.0—an interactive systems biology platform for the analysis of gene regulatory networks in corynebacteria and Escherichia coli. *J Biotechnol* 2007;**129**:279–89.
76. Brown KR, Otasek D, Ali M, *et al*. NAViGaTOR: network analysis, visualization and graphing toronto. *Bioinformatics* 2009;**25**:3327–9.
77. Van Puyvelde S, Steenackers HP, Vanderleyden J. Small RNAs regulating biofilm formation and outer membrane homeostasis. *RNA Biol* 2013;**10**:185–91.