Pure Appl. Chem., Vol. 74, No. 6, pp. 899–905, 2002.

Microbial computational genomics of gene regulation*

Julio Collado-Vides‡, Gabriel Moreno-Hagelsieb, and Arturo Medrano-Soto

Program of Computational Genomics, CIFN-UNAM, Av. Universidad s/n, Cuernavaca, 62100 Morelos, Mexico

Abstract: *Escherichia coli* is a free-living bacterium that condensates a large legacy of knowledge as a result of years of experimental work in molecular biology. It represents a point of departure for analyses and comparisons with the ever-increasing number of finished microbial genomes. For years, we have been gathering knowledge from the literature on transcriptional regulation and operon organization in *E. coli* K-12, and organizing it in a relational database, RegulonDB. RegulonDB contains information of 20–25 % of the expected total sets of regulatory interactions at the level of transcription initiation. We have used this knowledge to generate computational methods to predict the missing sets in the genome of *E. coli*, focusing on prediction of promoters, regulatory sites, regulatory proteins, operons, and transcription units. These predictions constitute separate pieces of a single puzzle. By putting them all together, we shall be able to predict the complete set of regulatory interactions and transcription unit organization of *E. coli*. Orthologous genes in other genomes of known co-regulated sets of genes in *E. coli*, along with their corresponding predicted operons, and their predicted transcriptional regulators, shall permit the extension of the previous goal to many more microbial genomes.

INTRODUCTION

The current accumulated knowledge of gene regulation and gene function in *E. coli* K-12 is unparalleled by about any other model organism. The accumulated knowledge of molecular biology in *E. coli*, summarized in the *E. coli* and *Salmonella* books of Neidhardt and collaborators [1,2], illustrates the legacy that *E. coli* represents.

Our laboratory has been devoted for years to a systematic search in the literature of known mechanisms of regulation of transcription initiation as well as operon organization in *E. coli*. This information is contained in RegulonDB, a relational database with known and computationally predicted elements, that is available at <http://www.cifn.unam.mx/Computational_Biology/regulondb/>.

The knowledge gathered so far gives us 528 known transcription units, 624 mapped promoters, close to 100 terminators [3], 165 DNA-binding transcriptional regulators [4], as well as 2128 genes with a known functional class assigned [5]. This experimentally supported knowledge is complemented with overall comprehensive computational predictions estimating a total of 700 operons [6] and a total of 314 transcriptional regulators [4].

Furthermore, we have also worked in the prediction of operator sites for the binding of transcriptional regulators [7]. Sequence and positional analyses of known sigma 70 promoters have permitted us

*Plenary lecture presented at the International Conference on Bioinformatics 2002: North–South Networking, Bangkok, Thailand, 6–8 February 2002. Other presentations are presented in this issue, pp. 881–914.

‡Corresponding author: E-mail: collado@cifn.unam.mx

to implement a method with an important improvement in specificity for promoter prediction in the complete *E. coli* genome ([8]; A. M. Huerta et al., in preparation).

It should be clear that with the amount of available knowledge, *E. coli* is an excellent model organism to implement and evaluate the performance of computational genomic methods. Once we have used the data to design the methods and tested them in *E. coli*, we will be in a position to expand these questions into other genomes. The importance of the work here outlined shall be appreciated in a similar way as we currently value annotations of gene predictions in sequenced genomes where very little experimental work has been done. As experience accumulates and the accuracy and performance of the computational methods improve, these predictions will be relevant as preliminary knowledge to be confirmed experimentally. Computational methods and their generated annotations offer a first comprehensive interpretation of the biology of the cell whose genome is available.

TOWARD PREDICTING THE COMPLETE SET OF REGULATORY INTERACTIONS AT THE INITIATION OF TRANSCRIPTION LEVEL IN *E. coli*

It is now feasible to consider the goal of a complete annotation or prediction of the regulatory interactions at the level of transcription initiation in *E. coli*. This can be done by integrating, step by step, a set of partial results such as predictions of the repertoire of regulatory proteins, their associated DNA-binding target sites, as well as the promoters and the organization of genes in transcription units and operons. In this paper, we present an overview of such partial results performed in our laboratory, we give a first example of integrated predictions, and, finally, we discuss some questions that remain to be solved to integrate them all with the global aim of predicting the complete set of transcriptional interactions in *E. coli*.

Identification of DNA-binding sites of transcriptional regulators

Previous genome analyses [7,9] have essentially expanded the set of potentially regulated genes subject to the regulation of 55 or 56 regulators with at least 3 or 4 known binding sites. These predictions, based on known sites, have used variations of the standard method of weight matrices, which are known to generate a large number of false positives [10,11]. Given the estimated complete set of 314 or so transcriptional regulators (see below), current predictions cover only around 18 % of the regulatory proteins.

Methods have been implemented to identify potential operator sites in the absence of previous knowledge of any binding site. These methods require another source of information, a family of co-regulated genes that share a common motif (see ref. [12] and references therein; in [13] a list of similar programs and their Web addresses can be found). Such methods use different conceptual frameworks (Gibbs sampling, greedy algorithms such as consensus, or over-representation of nucleotides), giving all as output a candidate motif in the form of a weight matrix. This matrix represents the variability and conservation of sites for a given protein and can then be used to search back in the whole set of upstream regions of a genome to find other potentially co-regulated genes.

Thanks to the recently developed high-throughput DNA micro-array methodologies for post-genomic experimental studies, commonly referred to as the transcriptome, we are now able to directly quantify the output of the gene regulatory network. Despite the fact that there are several alternatives to survey micro-array data [14–19], basically, clustering of genes with similar or complementary expression profiles, through different growth conditions, allows us to infer shared regulatory mechanisms and functional pathways. That is, the sheer abundance of transcriptome data, together with the experimental knowledge gathered in the literature and databases, is providing us with the challenge of developing data-mining strategies that incorporate that knowledge into the underlying models. This will allow detection of yet unknown and/or unexpected relationships among genes that can lead to the definition of the regulatory network structure that best describes the behavior of genes through all assayed expression experiments. Although many strategies to cluster transcriptome data have been reported [20–24], most of them fail to analyze simultaneously supplementary heterogeneous knowledge. We have developed a Bayesian clustering methodology based on multivariate mixtures that classifies heterogeneous types of data (Medrano-Soto et al., in preparation). Such analysis capacity is of great value because it makes possible the classification of expression data by taking into account additional biological knowledge, which can be either directly or indirectly interrelated. We believe that with this analysis some new unknown relationships or interactions between genes may be easier to establish.

Clustering analyses applied on transcriptome data provide groups of co-expressed genes that are good candidates to be potentially co-regulated [25]. By means of the organization of all genes in *E. coli* into transcription units and operons [6], and assuming that the operator sites are usually upstream of the first gene in an operon, it is possible to identify the corresponding upstream regions of a set of co-regulated genes. These sequences can then be searched for operator sites bound by their common transcriptional regulator.

Alternatively, a similar computational strategy can be used for grouping upstream regions of orthologous genes, searching for "phylogenetic footprints" as has been done based on synteny in Eukarya [26].

The first method we implemented to search for a common motif was based on a model of gene regulation in yeast [12]. More recently, we implemented a method specifically dedicated to finding sites with direct or inverted repeats [27]. This method shall prove adequate to finding motifs in bacteria, since bacterial regulators are predominantly dimers of polypeptides with helix-turn-helix (HTH) motifs [4]. Recent results in our laboratory indicate that we can improve the accuracy of the detection by concentrating on statistically significant dyads that tend to overlap with each other within a short sequence region [28]. We have also used overlapping of potential promoter sites, as well as additional information on their relative location to the beginning of the gene, in a method that improves the performance of promoter prediction in regions of 200 bases upstream of the first codon of genes (A. M. Huerta, in preparation).

Prediction of the complete set of transcriptional regulators in *E. coli*

Based on a collection of 150 known transcriptional regulatory proteins, most of them containing the HTH domain, which predominates in Eubacteria, we have implemented predictions that complement this set of known regulators with an almost equal number of predicted regulatory proteins [4]. A total of 314 transcriptional regulators are thus incorporated in RegulonDB, corresponding to around 8 % of genes of the genome. They should then correspond to most of the total repertoire, or vocabulary of regulatory specific motifs to be found among the upstream regions of genes in *E. coli.*

When expanding these analyses to other microbial genomes, we have identified a common origin in the set of transcriptional regulators in all the microbial organisms, based on the conservation of a motif that extends around the HTH DNA-binding motif. Four repressor families are suggested as those present before the divergence of Eubacteria and Archaea. This observation correlates with the fact that repression is in principle easier to establish since all it needs is a protein with a single DNA-binding motif with the target adequately positioned in relation to the promoter, whereas activation requires in addition a motif of interaction with RNA polymerase.

Operon organization of genes

Different and sometimes complementary recent genomic analyses have emphasized that functional information can be inferred by simply analyzing conservation of neighborhood of genes across different genomes. Phylogenetic profiles [29], or the "Rosetta stone" [30] method as baptized by the group of Eisenberg, as well as the work of Koonin [31] and Bork [32] all can be labeled as strategies for exploiting the biology of neighborhood.

By means of similar strategies applied to the set of known operons and transcription units contained in RegulonDB, we have shown that operons share a basic common architecture across different bacterial genomes. Briefly, the distribution of intergenic distances are very much similar (Fig. 1), with short intergenic distances—in fact dominated by overlaps of 1 or 4 bases—for genes within operons in all microbial genomes analyzed [33]. This contrasts with a flat distribution of distances at the boundaries of operons. We have extended the method of operon prediction initially implemented in *E. coli* [6] to similar predictions in other microbial genomes. Based on a collection of known operons in *Bacillus subtilis*, we have tested the performance of the method, which shows an estimated accuracy of above 82 %. We have found a way to calibrate the method for each genome [33]. Operons and transcription units are, furthermore, essential in the analysis of transcriptome experiments in bacteria.

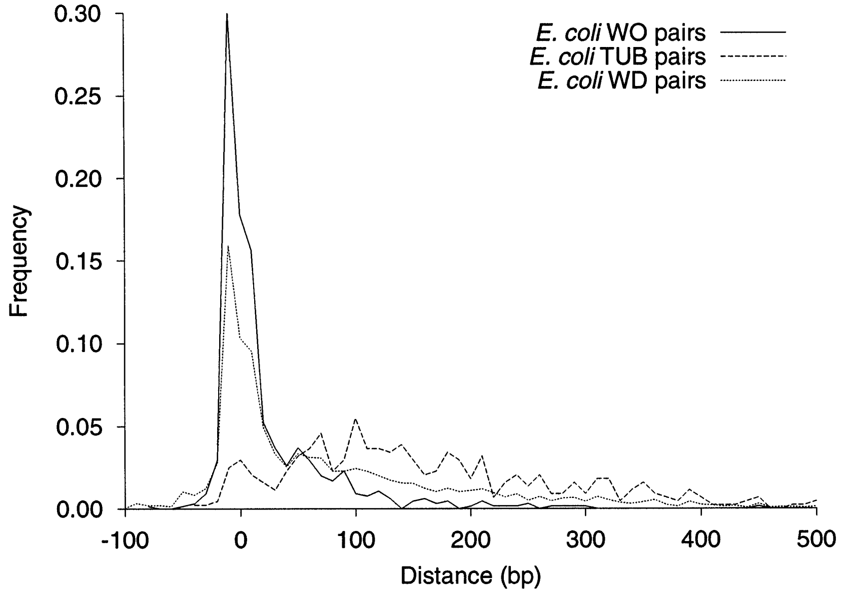

Fig. 1 Intergenic distance distributions of pairs of genes within operons (WO pairs), and of pairs of genes in the same strand, at the borders of transcription units (TUB pairs). We also show the intergenic distance distribution of all adjacent pairs of genes found in the same strand (WD pairs). TUB pairs are the last gene in a transcription unit, and the first one in the next. The distributions are so clearly different, that a distance-based method to discriminate between WO pairs and TUB pairs reaches an accuracy of more than 82 % in *E. coli* [6] and above 84 % in *B. subtilis* [33].

Integration of predictions

It is well known that predictions of regulatory motifs—binding sites—and promoters tend to easily generate a high number of false positives. Such predictions have concentrated in finding a single element within a region, and have been improved by using contextual information, such as the distance to the start codon as we mentioned above in the case of promoters (Huerta et al., in preparation). Integration of predictions is another way of using contextual information. One of our first steps towards integrating predictions consisted in the combination of predicting transcription units and DNA-binding sites [34]. We predicted the complete transcription unit (TU) organization of *E. coli* and of *Haemophilus influenzae*. In parallel, we used the sets of known DNA-binding sites for CRP (cAMP receptor protein) and FNR (fumarate and nitrate reduction regulatory protein) in RegulonDB to generate predictions in both genomes using the Consensus/Patser algorithms [35]. We defined orthologous TUs as those containing at least one gene orthologous to those in a TU in the reference genome (Fig. 2). In that way, we were able to classify the predicted sites and probably co-regulated genes according to the score of the predicted binding sites, the correspondence with TUs known to be regulated by the regulatory protein under analysis, or the existence of a predicted site in orthologous TUs, giving each sets of predictions a status in the confidence according to the reinforcement obtained from all this information.

This example illustrates that, as the individual programs of predictions improve, we shall be able to attain a level of higher integration and prediction of the complete set of regulatory interactions.

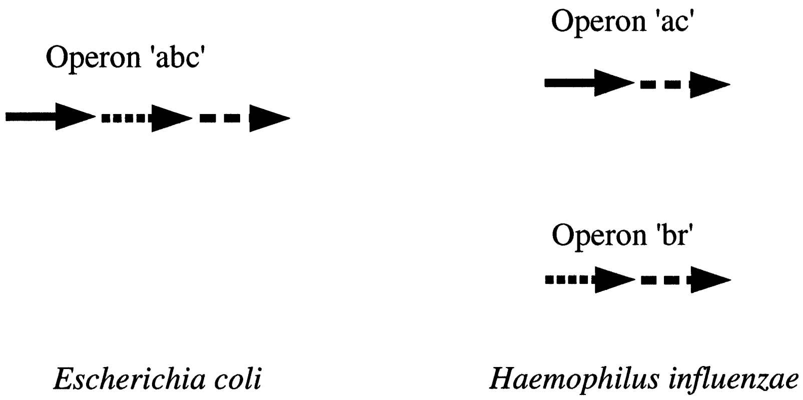

Fig. 2 Orthologous transcription units. The orthology is defined in terms of transcription units (TUs) having an orthologous gene in common. Genes are represented as arrows. In the hypothetical example shown here, the TU in *E. coli* is an operon containing genes ‘a,’ ‘b,’ and ‘c.’ The genes are reorganized in *H. influenzae* into two TUs, one operon containing genes ‘a’ and ‘c,’ another containing genes ‘b’ and ‘r.’ Both TUs in *H. influenzae* would be orthologous TUs to the operon in *E. coli.* Cases with re-organized TUs, might provide information about other genes being part of a given regulon, like gene ‘r’ in this example.

DISCUSSION

We have already at hand the repertoire of regulatory proteins in *E. coli* that should have binding sites upstream of promoters in *E. coli.* Similarly, we do have ways to predict the upstream binding sites, that is to say the target sites for the binding of such proteins. A difficult problem that we need to face in the future is the association of each regulatory protein to its family of operator sites. As far as we know, there is no easy solution to this protein-DNA recognition problem, and current computational methods require crystal structure data to be able to generate some predictions [36].

The three main subgoals necessary to attain the goal of the prediction of the complete network of regulation, as mentioned before, are: (i) to predict the set of transcriptional regulators and a similar set of operator families, (ii) to associate which regulatory protein binds to which operator group of sites, and (iii) to predict the regulatory effect of a regulator in transcription.

These three goals will set the basis for future studies on the evolution of the regulation of gene expression, their variability in different bacteria and microorganisms, and the role of transcriptional regulation and operon organization in microbial evolution. This project of microbial computational genomics will provide a comprehensive description of regulon variability in the prokaryotic world. This, together with the evolution of allosteric interactions, will give a basis to determine the role of gene regulation in the diversity and success of the microbial world.

ACKNOWLEDGMENTS

This work has been supported by grant 0028 from Conacyt, and grant GM62205 from NIGMS at NIH. We acknowledge computer support from Víctor del Moral and César Bonavides.

REFERENCES

1. F. N. Neidhardt, R. I. Curtiss, E. C. C. Lin, J. L. Ingraham, K. B. Low, B. Magasanik, W. Resnikoff, M. Riley, M. Schaechter, E. Umbarger. *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ASM Press, Washington, DC (1996).
2. F. C. Neidhardt and J. L. Ingraham. *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, American Society for Microbiology, Washington, DC (1987).
3. H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, J. Collado-Vides. *Nucleic Acids Res.* **29**, 72–74 (2001).
4. E. Perez-Rueda and J. Collado-Vides. *Nucleic Acids Res.* **28**, 1838–1847 (2000).
5. M. H. Serres and M. Riley. *Microb. Comp. Genomics* **5**, 205–222 (2000).
6. H. Salgado, G. Moreno-Hagelsieb, T. F. Smith, J. Collado-Vides. *Proc. Natl. Acad. Sci. USA* **97**, 6652–6657 (2000).
7. D. Thieffry, H. Salgado, A. M. Huerta, J. Collado-Vides. *Bioinformatics* **14**, 391–400 (1998).
8. F. R. Blattner, G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, Y. Shao. *Science* **277**, 1453–1474 (1997).
9. K. Robison, A. M. McGuire, G. M. Church. *J. Mol. Biol.* **284**, 241–254 (1998).
10. G. Z. Hertz, G. W. Hartzell, 3rd, G. D. Stormo. *Comput. Appl. Biosci.* **6**, 81–92 (1990).
11. G. D. Stormo. *Methods Enzymol.* **183**, 211–221 (1990).
12. J. van Helden, B. Andre, J. Collado-Vides. *J. Mol. Biol.* **281**, 827–842 (1998).
13. A. Goffeau. *Nat. Biotechnol.* **16**, 907–908 (1998).
14. R. Brent. *Curr. Biol.* **9**, R338-41 (1999).
15. P. O. Brown and D. Botstein. *Nat. Genet.* **21**, 33–37 (1999).
16. R. Somogyi. *Immunology Today* 17–24 (1999).
17. P. A. Clarke, R. te Poele, R. Wooster, P. Workman. *Biochem. Pharmacol.* **62**, 1311–1336 (2001).
18. E. M. Marcotte. *Nat. Biotechnol.* **19**, 626–627 (2001).
19. F. Mehraban and J. E. Tomlinson. *Eur. J. Heart Fail.* **3**, 641–650 (2001).
20. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
21. P. D'Haeseleer, S. Liang, R. Somogyi. *Bioinformatics* **16**, 707–726 (2000).
22. S. Datta. *Gene Expr.* **9**, 249–55 (2001).
23. K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, W. L. Ruzzo. *Bioinformatics* **17**, 977–987 (2001).
24. Y. Xu and V. Olman. *Genome Inform. Ser. Workshop Genome Inform.* **12**, 24–33 (2001).
25. D. E. Bassett, Jr., M. B. Eisen, M. S. Boguski. *Nat. Genet.* **21**, 51–55 (1999).
26. W. W. Wasserman, M. Palumbo, W. Thompson, J. W. Fickett, C. E. Lawrence. *Nat. Genet.* **26**, 225–228 (2000).
27. J. van Helden, A. F. Rios, J. Collado-Vides. *Nucleic Acids Res.* **28**, 1808–1818 (2000).
28. E. Benítez, G. Moreno-Hagelsieb, J. Collado-Vides. GenomeBiology.com **3**:(3): research0013.1-0013.16. (2002).
29. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).

30. E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, D. Eisenberg. *Science* **285**, 751–753 (1999).
31. M. Y. Galperin and E. V. Koonin. *Nat. Biotechnol.* **18**, 609–613 (2000).
32. M. Huynen, B. Snel, W. Lathe, 3rd, P. Bork. *Genome Res.* **10**, 1204–1210 (2000).
33. G. Moreno-Hagelsieb and J. Collado-Vides. *Bioinformatics (ISMB-02)* **18**, Suppl. 1, S–S8 (2002).
34. K. Tan, G. Moreno-Hagelsieb, J. Collado-Vides, G. D. Stormo. *Genome Res.* **11,** 566–584 (2001).
35. G. Z. Hertz and G. D. Stormo. *Bioinformatics* **15**, 563–577 (1999).
36. H. Kono and A. Sarai. *Proteins* **35**, 114–131 (1999).