
This is the **accepted version** of the journal article:

Erill, Ivan; Escribano, Marcos; Campoy Sánchez, Susana; [et al.]. «In silico analysis reveals substantial variability in the gene contents of the Gammaproteobacteria LexA-regulon». *Bioinformatics*, Vol. 19, issue. 17 (Nov. 2003), p. 2225-2236. DOI 10.1093/bioinformatics/btg303

This version is available at <https://ddd.uab.cat/record/287839>

under the terms of the  **CC BY-NC-ND** license

***In silico* analysis reveals substantial variability in the gene contents of the gamma proteobacteria LexA regulon**

Ivan Erill^{1*}[♠], Marcos Escribano^{1♠}, Susana Campoy² and Jordi Barbé²

¹ Biomedical Applications Group, Centro Nacional de Microelectrónica, 08193 Bellaterra, Spain

² Departament de Genètica i Microbiologia / Centre de Recerca en Sanitat Animal (CRESA), Universitat Autònoma de Barcelona 08193 Bellaterra, Spain

ABSTRACT

Motivation: Motif-prediction algorithm capabilities for the analysis of bacterial regulatory networks and the prediction of new regulatory sites can be greatly enhanced by the use of comparative genomics approaches. In this study we make use of a consensus-building algorithm and comparative genomics to conduct an in-depth analysis of the LexA-regulon of gamma proteobacteria, and we use the inferred results to study the evolution of this regulatory network and to examine the usefulness of the control sequences and gene contents of regulons in phylogenetic analysis.

Results: We show, for the first time, the substantial heterogeneity that the LexA regulon of gamma proteobacteria displays in terms of gene content and we analyze possible branching points in its evolution. We also demonstrate the feasibility of using regulon-related information to derive sound phylogenetic inferences.

Availability: Complementary analysis data and both the source code and the Windows-executable files of the consensus-building software are available at <http://www.cnm.es/~ivan/RCGScanner/>.

Contact: ivan.erill@cnm.es; jordi.barbe@uab.es.

INTRODUCTION

The structure and function of bacterial regulons is becoming a widely accepted source of information in the understanding of bacterial physiology and genetics. In essence, a prokaryote regulon can be defined as a network of genes under synchronized transcriptional control by a regulatory protein, or set of proteins, that recognizes a specific binding-motif in the promoter region of the genes it exerts control upon. Protein binding to the operator site may repress or activate transcription of the regulated genes, thus establishing a negative or positive control. This defining property of regulons, the binding of the regulatory protein to a specific recognition sequence in the operator site, has been repeatedly used in *in silico* analyses to predict new regulon members (Lewis et al, 1994; Fernández de Henestrosa *et al.*, 2000; Rodionov *et al.*, 2001) and even to predict previously unreported regulon structures in little-studied species (Gelfand *et al.*, 2000a; McGuire *et al.*, 2000). From the first systematic attempts at defining the informational properties of regulatory regions and the possibility of predicting new regulatory sites by statistically assessing their binding-affinity (Berg and von Hippel, 1987; Berg, 1988), regulatory motif prediction algorithms have evolved fast and have diversified into four main groups, each based on a distinct statistical approach: consensus building algorithms (Stormo and Hartzell, 1989), expectation maximization algorithms (Lawrence *et al.*, 1990), Gibbs sampling-method algorithms (Lawrence *et al.*, 1993) and oligonucleotide frequency analysis (van Helden *et al.*, 1998). Although none of these methods

* To whom correspondence should be addressed.

[♠] The authors wish to express that both authors should be regarded as joint first authors in this work.

strictly requires *a priori* experimental knowledge to work, all of them have been optimized to make use of such heuristics, typically conveyed in the form of experimentally determined regulatory motifs or members of the regulon for a given bacterial genome (Bailey and Elkan, 1995; McCue *et al.*, 2001; Rodionov *et al.*, 2001). More recently, the use of experimental cues to enhance the predicting capabilities of these methods has been assisted by the large-scale introduction of microarray gene-expression experiments (Courcelle *et al.*, 2001; Khil and Camerini-Otero, 2002), which have provided a boon of experimental background to motif-prediction algorithms. Moreover, with the assumption that regulons and regulatory motifs are well-conserved structures among related species (Gelfand *et al.*, 2000b), the wealth of information provided by completely sequenced genomes has also been recently tapped in comparative genomics analyses (Gelfand *et al.*, 2000b; McGuire *et al.*, 2000, McCue *et al.*, 2001; Tan *et al.*, 2001; Rajewsky *et al.*, 2002) that make use of known regulon structures in related genomes to strengthen and focus motif-prediction algorithms.

The assumption that regulon structure is well conserved among related bacterial species is not a bold one. Although regulon members are susceptible to lateral-gene transfer (LGT), the regulon as a whole and its regulatory protein tend to be quite stable from an evolutionary viewpoint (Gelfand *et al.*, 2000b), a fact that is most acute in the case of closely related species, where regulatory motifs are often conserved (McGuire, Hugues, 2000). Regulon conservation has been recently confirmed (Makarova *et al.*, 2001; Rodionov *et al.*, 2001) and positively exploited in the aforementioned comparative genomics assays. Furthermore, the evolutionary stability of a regulon can be correlated with its gene contents (Rajewsky *et al.*, 2002) and the occurrence of self-regulation (Roy *et al.*, 2002). It seems evident that, in the case of a large and self-regulated gene network, regulon structure (s.c. regulatory protein, regulon functional-core genes and regulatory motifs) will tend to be preserved because a mutation either in the gene encoding the regulatory protein or its operator region, will often lead to severe deregulation and, thus, to a substantial disruption in cellular equilibrium. A well-known and documented (Walker, 1994) case of such a large and self-regulated network is the LexA-regulon of the gamma-proteobacteria *Escherichia coli*, the fundamental component of the DNA damage-inducible SOS response (Radman, 1984). The LexA-governed network of *E. coli* has been shown to regulate up to 30 genes (Fernández de Henestrosa *et al.*, 2000), with the LexA protein repressing the system (including the *lexA* gene) by binding to a 16-mer consensus sequence CTG-N₁₀-CAG (the LexA box) in the promoter region of regulated genes (Walker, 1984). Upon DNA damage, ssDNA-activated RecA promotes LexA auto-hydrolysis (Little, 1984), triggering derepression of the system, and activating a set of genes, involving error-prone polymerases (*umuDC*), recombinases (*recA*, *recN*), excision repair nucleases and helicases (*uvrAB*, *uvrD*) and cell-division inhibitors (*sulA*) that contribute to overcome and repair DNA damage (Fernández de Henestrosa *et al.*, 2000). The assumption that the LexA-regulon is a well-conserved structure across substantial evolutionary spans is supported by its described presence in a wide set of bacterial families, ranging from green non-sulfur (Fernández de Henestrosa *et al.*, 2002) and gram-positive bacteria (Winterling *et al.*, 1998) to gamma (Walker, 1984) and alpha proteobacteria (Tapias *et al.*, 1999) that occupy a broad and varied set of ecological niches. In the specific case of gamma proteobacteria, the assumed evolutionary stability of the LexA-regulon is further supported by experimental evidence of regulatory-motif conservation across different species (Garriga *et al.*, 1992) and by the contrasted success of prior studies with motif-prediction algorithms (Lewis *et al.*, 1994; Fernández de Henestrosa *et al.*, 2000; Benítez-Bellón *et al.*, 2002). Besides, the presence of a cell-division inhibitor (*sulA*) in *E. coli* LexA-regulon introduces a bottleneck effect on the evolutionary pathways of this regulon, since it renders LexA⁻ mutants non-viable. Given this hindsight into the structure of the LexA-regulon of *E. coli*, here we test the feasibility of using a consensus-building algorithm as a robust tool to make strong predictions on the regulon structure of different gamma proteobacteria species, and we use the inferred knowledge

to analyze the changes in the gene contents of this regulon that have taken place over small evolutionary distances. Thereafter, we put forward and show that the multifaceted and correlated nature of the information conveyed by a regulon (regulon members, core members, regulatory protein and regulatory motif sequence) can be used as a sound phylogenetic indicator.

MATERIALS AND METHODS

Experimental data

A thorough description of the gene-set conforming the LexA-regulon in *Escherichia coli* was obtained from published Northern blot (Lewis *et al.*, 1994; Fernández de Henestrosa *et al.*, 2000) and DNA micro-array (Courcelle *et al.*, 2001; Khil and Camerini-Otero, 2002) experimental studies. This data was integrated to make up the basic set of *E. coli* LexA-regulated genes and corresponding binding-motifs shown in Table 1, which was subsequently used as the experimental training set for the consensus-building algorithm.

Genome assemblies and databases

Complete genome assemblies of *Bacillus subtilis* [AL009126], *Escherichia coli* K-12 MG1655 [U00096], *Haemophilus influenzae* Rd [L42023], *Pasteurella multocida* PM70 [AE004439], *Pseudomonas aeruginosa* PA01 [AE004091], *Ralstonia solanacearum* GMI1000 [AL646052], *Sinorhizobium meliloti* 1021 [AL591688], *Salmonella typhimurium* LT2 [AE006468], *Shigella flexneri* 2a str. 301 [AE005674], *Vibrio cholerae* [AE003852] and *Yersinia pestis* strain CO92 [NC_003143] were downloaded from NCBI Genbank database, and a whole genome shotgun of *Klebsiella pneumoniae* MGH78578 [NC_002941] was downloaded from the GSC FTP site at Washington University. Manual orthology searches to assess conservation of LexA-regulon genes and to verify predicted regulon genes were routinely carried out using NCBI TBLASTX server against the *nr* database (Altschul *et al.*, 1990) or by name-querying either NCBI Genbank or TIGR CMR2 databases.

Alignment and phylogeny tools

All automated alignments for orthology searches were carried out using NCBI TBLASTN server with default parameters. Manual protein sequence alignments were performed using INRA Multalin server (Corpet, 1988) and a Blosum62 matrix (Henikoff and Henikoff, 1992). Phylogenetic trees were inferred from aligned DNA (regulatory motif) and protein sequences using Phylip 3.6 DNAML and PROML programs (Felsenstein, 1989), imposing a transition/transversion ratio of 2.0 for DNA sequences and using a PAM Dayhoff matrix (Dayhoff *et al.*, 1978) for protein sequences. Phylogeny trees were plotted using TreeView Windows-based software package (Page, 1996).

Consensus-building software

To analyze regulon structure we developed RCGScanner, a Windows-based standalone software package that integrates a three-step algorithm (see Figure 1) for the prediction of putative regulatory motifs. The first step in the algorithm is a pattern search of user-defined direct or inverted repeats in the form X-n-Y, where X and Y are *a priori* known or estimated sequences and n is a variable nucleotide sequence. The program scans a local DNA sequence file according to the IUB standard (Nomenclature Committee, 1985), looking for matching X-n-Y motifs and allowing up to one mismatch in either the X or Y sequences. To reduce the huge number of false positives that might

arise from a straightforward complete genome scan (Gelfand *et al.*, 2000b), after locating a regulatory motif the program scans the adjacent region and stores only those regulatory sequences that are close (typically 300 bp; all program parameters are user-adjustable) to a coherent open reading frame (ORF). Once the pattern search is completed, the program computes a motif consensus matrix based on experimental knowledge (Berg, 1988), which can be supplied directly or else automatically inferred. If there is enough experimental data for a given species (e.g. *E. coli* LexA-regulon) the program computes the consensus matrix from a collection of user-introduced regulatory motifs (see Table 1). Conversely, when no direct knowledge is available, the program takes a comparative genomics approach, presuming conservation of regulon structure in related bacterial species (Gelfand *et al.*, 2000b). In this case, the program takes as input the protein sequences of regulon genes from a species in which the regulon has been experimentally established, and uses them to query NCBI GenBank database through its TBLASTN server on the unstudied species. Homologies above an identity threshold (typically 80%) are considered conserved orthologs (Rajewsky *et al.*, 2002) and their promoter regions are scanned for putative regulatory motifs. If found, these regulatory motifs will then be used to infer the consensus matrix for the species under consideration. After computation of the consensus matrix, the program uses it to filter putative regulatory motifs by computing their Heterology Index (HI), a statistical measure of the divergence from the consensus sequence (Berg and von Hippel, 1987; Berg, 1998). Two complementary filtering approaches are used here. In direct filtering, sequences are sorted according to their HI value and filtered with a simple threshold method (typically $HI < 8$). In recursive filtering, a more flexible filtering approach is implemented in a similar manner to that already described in the literature (Gelfand *et al.*, 2000a). An initial population of regulatory motifs (i.e. all those found) defines the initial consensus matrix and is filtered with a HI relative-threshold method (typically 1/3 of the mean HI value). Filter-passing motifs are then used to compute a new consensus matrix and the process is iterated until population divergence between consecutive iterations stabilizes below a predefined threshold. The recursive filtering method is more flexible than the direct one, but it is also more sensitive to background noise and local minima (Gelfand *et al.*, 2000a). To overcome noise sensitivity, the program uses the direct filter results as a seed for the initial recursive population, thus focusing the initial search space and improving recursive-filtering results. As a final step, the program automatically queries the NCBI TBLASTN server and the GenBank database to obtain and store functional definitions for each of the genes putatively regulated by a filter-passing motif.

Analysis methods

Software validation methods

To validate software performance, the program was first tested against the most documented case of the LexA-regulon (s.c. *E. coli*; experimental motif consensus: CTGtatatatataCAG, see Table 1). The test against *E. coli* was conducted in a two-step procedure that was later assumed as standard for all analyses. The first step consisted in a sensitive search (CTG-N₁₀-CAG) to assess the efficiency of the pattern search algorithm at detecting experimentally described LexA binding motifs. This search served also to draw an initial estimate of *sensitivity* (i.e. the ability of the filtering algorithm to select described LexA binding motifs against randomly scattered pseudosites) and *specificity* (i.e. the competence of the filtering algorithm at sorting out pseudosites) in broad-spectrum searches, and was used to fine-tune and settle program parameters. The second step consisted in a more restrictive (CTGT-N₈-ACAG) search, to boost specificity and to determine the ability of the program to unambiguously identify regulon structure. Finally, a test against background noise was conducted to estimate the informational relevance of the program results. Restrictive (CTGT-N₈-

ACAG) searches were launched against gram-positive (*B. subtilis*) and alpha proteobacteria (*S. meliloti*) genomes, in which distinct LexA-binding motifs have been experimentally described (Winterling *et al.*, 1998; Tapias *et al.*, 1999), and the results were manually inspected to evaluate their significance.

Regulon analysis methods

A similar two-step procedure was implemented to conduct the full analysis of the LexA-regulon in the selected subgroup of gamma and beta proteobacteria species. For each bacterial species, a first sensitive (CTG-N₁₀-CAG) search was launched and filtered using the automatically inferred consensus matrix derived from conservation of experimentally described *E. coli* LexA-regulon genes (see Table 1). Search results, regardless of selection procedures, were manually inspected to identify putative conservation of LexA binding-motifs controlling homologues of the LexA-regulated genes described in *E. coli*. Next, the subset of these motifs that had been automatically selected by the program was manually picked out and used to recreate a species-specific knowledge table, akin to that experimentally derived for *E. coli* (Table 1). Using this newly inferred knowledge table to compute the consensus matrix and to filter accordingly, a second restrictive (CTGT-N₈-ACAG) search was carried out against each bacterial species, and its selection results were considered putative members of the LexA-regulon for each particular species.

RESULTS AND DISCUSSION

Software validation results

Software validation results were amply satisfactory for the scope of this research. After fine-tuning of parameters, a sensitive (CTG-N₁₀-CAG) search against *E. coli* returned 33,418 putative regulatory motifs, revealing a huge number of false positives (pseudositos) in the genome that is in accordance with previous literature reports (Gelfand *et al.*, 2000b). However, the search did also locate all the 28 documented LexA binding-sites that had been introduced as the training set (Table 1). This was a necessary prerequisite for the study of related genomes, since it guaranteed that, even if not selected, conserved regulatory sites would be found and could be manually tracked by querying results in conserved LexA-regulon homologous regions. Moreover, and taking into account the vast number of pseudositos found with the broad-spectrum search, the program did also fare well in terms of sensitivity (89%, up to 25 of the 28 documented LexA sites were selected), but at the cost of an extremely low specificity (10%, only 42 of the 420 selected motifs were in the promoter region of documented LexA-regulated genes). Therefore, we then examined the reliability of applying a more restrictive search pattern to improve specificity without excessively compromising sensitivity. As expected, the restrictive (CTGT-N₈-ACAG) pattern search returned far less regulatory motifs (1,872), and this had a slight repercussion on sensitivity (71%). However, specificity was boosted by the restrictive search (from 10% to 83%, 20 out of 24 selected motifs corresponded to LexA-regulated genes). This high specificity, combined with the fact that the remnant of selected motifs consisted of previously described damage-inducible genes, such as *minC* and *hlyE* (Courcelle *et al.*, 2001), and previously unreported putative motifs for LexA-regulon genes (see Table 2), led us to conclude that a restrictive search could be a robust indicator of regulon structure for extrapolation into unstudied species. The statistical significance of the results thus obtained and the appropriateness of combining direct and recursive filtering techniques was gauged by examining their accordance with previously published results for *E. coli* LexA binding-site predictions (see Table 1; Benítez-Bellón *et al.*, 2002), by evidencing that all selected motifs

were either experimentally described or new putative sites, and by ascertaining that the results of background noise tests were markedly negative (none of the selected regulatory motifs in *B. subtilis* and *S. meliloti* involved any DNA-repair genes). Therefore, the combined method (i.e. broad-spectrum plus restrictive search) was deemed sound enough to carry out comparative genomics analyses of the LexA-regulon in related bacterial species, since it conveyed the necessary sensitivity to detect most conserved motifs and the required specificity to outline the structure of the regulon in the experimentally unstudied bacteria.

Regulon analysis results

The results of the application of the combined search method on nine different bacterial species, summarized in Table 3, reveal the existence of a conserved set of regulated genes (*lexA*, *recA* and *recN*) among gamma proteobacteria. The existence of such a conserved *regulon core* should be expected in any kind of self-regulated gene network (Gelfand *et al.*, 2000b), and its members ought to define the basic set of essential tasks the regulon was originally set forth to control (e.g. damage-inducible recombination repair). Interestingly, thus, this structure appears to be conserved also in the sole representative of the beta proteobacteria class analyzed in this study (*R. solanacearum*). Even though the *recN* LexA box of *R. solanacearum* appears to be slightly degenerated, the conservation of the regulon core hints for the first time at a more than probable conservation of the gamma LexA-binding motif in the beta proteobacteria class. The results also reinforce the previously proposed idea that *ydbK* and *minC* are damage-inducible genes directly regulated by LexA (Courcelle *et al.*, 2001) and point at some plausible additions to the LexA regulon in different species. Of peculiar interest are the putative LexA regulation of *mdf* and *impA* in the closely related *H. influenzae* and *P. multocida* species, which hints at a probable uptake of the regulation of these genes in a common ancestor, and the putative regulation of pathogenesis-related genes (STM1019, STM2621 and *msgA*, associated with Gifsy-1/2 prophages; STM0925 and STM272, connected to Fels-1/2 prophages) in *S. typhimurium*, a fact that has already been experimentally reported (Benson *et al.*, 2000). Also, the presence of direct LexA regulation for *recG* and *ftsY* in *V. cholerae* suggests a branching point in the evolution of this bacterial species with respect to its closest relatives, which may be connected to the loss of the *sulA* gene in this bacterium. In this respect, it is relevant to pinpoint that the results in Table 3 agree with the hypothesis that *sulA* regulation imposes a sort of bottleneck effect in the evolution of the LexA-regulon, preventing major divergences in conserved regulatory motifs, but not in the gene contents of the regulon. The obvious explanation for this effect is that the regulated presence of *sulA* restricts LexA variability, since any changes that induce a poorer recognition of the *sulA* box will severely handicap the cell's ability to divide. However, the present study indicates that the *sulA* bottleneck effect concerns only a relatively small subgroup of the gamma proteobacteria here checked (*E. coli*, *S. typhimurium* and *Y. pestis*) and it possibly highlights a branching point in the evolution of this bacterial lineage. Even though it could be argued that a *sulA* gene is also present in *P. aeruginosa*, the present study suggests that this *sulA* is not explicitly regulated by a dedicated LexA box (instead, *sulA* seems to be part of the *lexA* operon). Thus, in this species the presence of *sulA* should not induce the same kind of consensus-sequence bottleneck effect, but, rather, a gene-content limitation effect, due to the presumably over-repressed nature of the whole *lexA* operon. On the other hand, Table 3 results reveal an apparent gradual loosening across evolutionary distances of the classical LexA-regulon structure that has been experimentally determined in *E. coli*. This progressive drift seems to place *E. coli* and close relatives at the end of an evolutionary pathway with respect to the LexA regulon, a fact that is in agreement with phylogenetic data otherwise obtained (Fox *et al.*, 1980), with the late appearance of *E. coli* natural habitat (mammals digestive tract) in the fossil record and with the risky but cost-

effective addition of cell division inhibitors (such as *sulA*) to the LexA regulon of *E. coli* close relatives. Most importantly, though, the results shown in Table 3 reveal a clear and smooth evolution of the LexA-regulon in gamma proteobacteria, purporting remarkable plasticity both in terms of the presence/absence of genes and of the nature of their regulatory motifs. It was the logical congruence of these results with previously reported phylogenetic relationships for this class of bacteria (Fox *et al.* 1980; Ochman and Wilson, 1987; Rajewsky *et al.*, 2002; Xie *et al.* 2003) that led us to considerate the feasibility of using regulon data for phylogenetic inference.

Phylogenetic analysis results

To conduct a phylogenetic analysis of the gamma proteobacteria family based on the deduced LexA-regulon structure, we first analyzed which of the multiple informational sources conveyed by a regulon were solid enough to infer phylogenetic relationships. We decided that the regulon core, being strongly preserved in all the analyzed species, could be a sound informational source. Moreover, the regulon core was a very useful structure, because it conveyed two separate, but clearly correlated, kinds of information: protein and regulatory motif sequences for each of the core genes. Additionally, insight into Table 3 results prompted us to esteem that regulon structure, whether as the presence/absence of gene regulation or as divergences in the regulatory motif, could also be a reliable source of information. Lastly, it must be noted that we discarded another plausible source of regulon-related information, the consensus sequence for each bacterial species (Figure 2). Consensus sequence was not employed on the grounds of its low statistical weight (it had been computed from a different number of genes in each species), its low informational content (it is an averaging measure) and the previously outlined possibility (Rajewsky *et al.*, 2002) that the consensus sequence may not be such a robust indicator of binding affinity as predicted (Berg, 1988). This later hypothesis was addressed here in conjunction with microarray gene-expression data (Courcelle *et al.*, 2001; Khil and Camerini-Otero, 2002). Although the idea that regulatory motifs ought to display better binding affinities when closer to the consensus is theoretically sound, we found that the LexA boxes of genes with consistently reported high-induction ratios (*sulA*, *recA* and *recN*) displayed relatively high HI values (data not shown). In a negatively regulated gene network, like the LexA regulon of gamma proteobacteria, high binding affinities should induce strong repression under normal conditions and, consequently, the highest induction ratios upon derepression of the system. Therefore, although these results do not invalidate the theoretical background of using the consensus sequence as an average species indicator or as the basis of consensus-building algorithms, they do cast serious doubts on the validity of using low HI values to accurately predict high binding affinities. Expression profiles also reinforce the hypothesis that the *sulA* box, due to the markedly detrimental effects of *sulA* deregulation, must display a high binding affinity (s.c. high induction levels) and that, as mentioned before, this requirement imposes severe constraints on the variability of the LexA protein and the motifs it recognizes. As a result, we settled on three different sources of information to derive phylogenetic inferences: regulon core protein sequences, regulon core LexA-box sequences and regulon structure. Regulon structure information was introduced in the form of presence/absence/divergence of LexA regulatory motifs for all LexA-regulon genes experimentally described in *E. coli*. When we plotted the phylogenetic trees inferred by the maximum-likelihood method using these three sources of information, we found that the results (Figure 3) were not only in neat accordance to standard phylogenetic approaches (Fox *et al.*, 1980; Ochman and Wilson, 1987; Rajewsky *et al.*, 2002), but were also strikingly similar between them, suggesting that the three sources of information carried by the regulon are strongly correlated by the own regulon nature. The robustness of this correlation becomes more apparent when considering the different nature of the data used for inferring the

trees. Although cladistic analysis by itself cannot be used as a measure of statistical significance, the fact that trees based on protein and short DNA sequences yield such a remarkable resemblance hints at an active selection process behind the regulon structure, counteracting the expected higher noise ratio of short-length sequence analyses.

Discussion

Conventional phylogenetic analyses (Woese and Fox, 1977; Woese, 1998) have relied mainly in the use of small-subunit ribosomal RNA (16S rRNA). The usefulness of 16S rRNA to infer phylogenetic relationships sprouts from many different wells. On the one hand, ribosomes are essential elements of the translational apparatus and, thus, present in all known life forms, making the 16S rRNA genes universal markers. On the other hand, the very importance of ribosomes for life processes subjects ribosomal genes to a strong selective pressure, meaning that sequence conservation is high in 16S rRNA and that, consequently, its informational content is also elevated. This very same importance makes 16S rRNA genes unlikely candidates for lateral gene transfer (LGT), and this ensures verticality and coherence in the inferred phylogenetic trees. Finally, 16S rRNA genes are relatively large and, thus, they can convey enough informational content to derive long time-span trees, a feat that cannot be accomplished by other universal and highly conserved genes (e.g. tRNA genes) and that has prompted researchers to explore the potential of the even larger 23S rRNA genes in phylogenetic analysis (Pitulle *et al.*, 2002). Nevertheless, there are also some shortcomings associated with the use of 16S rRNA to derive phylogenetic relationships. A major one comes precisely from its strong point, conservation. In fact, 16S rRNA genes are so well conserved that they exhibit little resolving power among closely related bacterial species (Achenbach *et al.*, 2001). Additionally, the natural tendency of cells to duplicate such essential genes leads to varying copy numbers of the gene across different species, causing over and under representation of some of them when conducting phylogenetic analyses. To overcome these difficulties, researchers have used other universal genes with more stable copy numbers (Lloyd and Sharp, 1993; Eisen, 1995) or taxa-specific genes to enhance the resolving power of phylogenetic inferences (Ludwig, 1990; Fukushima, 2002; Ko, 2002), but both these methods still lack an intermediate level of resolution to systematically hold together the results they separately infer. In recent years, and with the advent of sequenced genomes, some new approaches have tried to circumvent this problem by creating multiple protein trees (Feng *et al.*, 1997; Gupta, 2000) or by analyzing gene content and copy number, instead of gene sequence, in complete genomes (Snel *et al.*, 1999; Tekaiia *et al.*, 1999). Still, these methods do not take into account some key aspects that might enhance resolution and understanding, like gene functionality, due to the difficult and subjective handling of such issues. Here we propose the use of computationally deduced regulon structure as a way to exploit functionality associations directly conveyed by nature (instead of subjectively human inferred), and to use this information in association with conventional phylogenetic data sources (i.e. DNA and protein sequence) to derive robust, relatively universal and well-resolving phylogenetic trees.

In general terms, a regulon is a fairly well suited entity to conduct phylogenetic analysis. Although most of them are not universal, regulons are complex structures that are not prone to neither appear out of the blue nor undergo spontaneous deletions. Additionally, there exist regulons, like the CRP-cAMP regulatory network, that are present over vast spans of the life realm. Moreover, many regulons are committed to housekeeping tasks and, thus, they are naturally resilient to mutation and LGT. Even though mutation, deletion and LGT may affect many of the regulated genes, the regulon core ought to be a relatively solid structure (Gelfand *et al.*, 2000b). This applies also to copy number, especially in the case of the regulatory protein. Duplications of

the regulatory protein gene may certainly occur, but the most probable outcome is that the redundant copy will, in time, drift to overtake or complement other regulatory networks (Zuckerandl, 1975; Gelfand *et al.*, 2000b). Thus, the regulon as a complete structure presents double information content: the regulon core, with an evolutionary stable structure, and the global gene set, more prone to variation. This dual nature, glued together by the regulon makeup, offers a simultaneous two-level view on phylogeny that can allow detailed, taxa-specific, and at the same time globally coherent analyses. Furthermore, even when a regulon is not conserved, or undergoes severe changes, in phylogenetically distant species, this fact can be used to derive solid phylogenetic inferences. For instance, the LexA-regulon here studied is not universally conserved, even though it has been shown to be preserved in a wide range of different bacterial lineages and its co-inducer, the RecA protein, has been shown to be a feasible phylogenetic indicator (Lloyd and Sharp, 1993; Eisen, 1995). Nevertheless, and due to the housekeeping functions it carries out (DNA repair), it seems clear that equivalent regulons must exist in those species lacking the LexA-network (Koch and Woodgate, 1998). Therefore, and because of the difficulty of creating working regulons from scratch, regulon loss can be used to pinpoint major evolutionary branching points. Likewise, major divergences between regulons inner structure (e.g. a change in consensus regulatory motif) can also highlight turning points in evolution, as it is the case with the divergent LexA regulatory motifs of alpha (Tapias *et al.*, 1999) and gamma proteobacteria (Walker, 1984) or gram-positive bacteria (Winterling *et al.*, 1998).

CONCLUSION

Our results represent the first published instance of the substantial heterogeneity in gene content displayed by the LexA network in gamma proteobacteria, and point at possible major events (like the acquisition of *sulA*) in the evolution of the LexA regulon in this class of bacteria. We also put forward and test for this particular case the proposition that regulon information, either (or complementarily) obtained by *in silico* or *in vitro* analysis can be used to infer strong phylogenetic relationships in closely related bacteria, and that this method could be extended, with the use of other regulons, to generate a phylogenetic analysis method of both the necessary resolution and adequate consistency to bridge the gap between existing methodologies.

ACKNOWLEDGEMENTS

This work was partly funded by the Consejo de Investigaciones Científicas (CSIC) and by Grants BMC2001-2065 from the Ministerio de Ciencia y Tecnología (MCyT) de España and 2001SGR-206 from the Departament d'Universitats, Recerca i Societat de la Informació (DURSI) de la Generalitat de Catalunya.

REFERENCES

- Achenbach, L. A., Carey, J., Madigan, M. T. (2001) Photosynthetic and phylogenetic primers for detection of anoxygenic phototrophs in natural environments. *Appl. Environ. Microbiol.*, 67, 2922-2926.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403-410.
- Bailey, T. L., Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME, *Proc. IIIrd Int. Conf. Intel. Sys. Mol. Biol.*, 21-29, AAAI Press, Menlo Park, California.
- Benítez-Bellón, E., Moreno-Hagellsieb, G., Collado-Vides, J. (2002) Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA. *Genome Biol.*, 3, research0013.1-0013.16.
- Benson, N. R., Wong, R. M-Y., McClelland, M. (2000) Analysis of the SOS response in *Salmonella enterica* serovar typhimurium using RNA fingerprinting by arbitrarily primed PCR. *J. Bacteriol.*, 182, 3490-3497.

- Berg, O. G. (1988) Selection of DNA binding sites by regulatory proteins: the LexA protein and the arginine repressor use different strategies for functional specificity. *Nucleic Acids Res.*, 16, 5089-5105.
- Berg, O. G., von Hippel, P. H. (1987) Selection of DNA binding sites by regulatory proteins, statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193, 723-750.
- Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucl. Acids Res.*, 16, 10881-10890.
- Courcelle, J., Khodursky, A., Peter, B., Brown, P. O., Hanawalt, P. C. (2001) Comparative gene expression profiles following UV exposure in wild type and SOS deficient *Escherichia coli*. *Genetics*, 158, 41-64.
- Dayhoff, M., Schwartz, R. M., Orcutt, B. C. (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, 5, 345-352, Dayhoff, M. (Ed.), National Biomedical Research Foundation, Silver Spring, MD.
- Eisen, J. A. (1995) The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J. Mol. Evol.*, 41, 1105-1123.
- Escherichia coli*: SOS repair hypothesis. In *Molecular and Environmental Aspects of Mutagenesis*, Felsenstein, J. (1989) PHYLIP: phylogeny inference package (version 3.2). *Cladistics*, 5, 164-166.
- Feng, D. F., Cho, G., Doolittle, R. F. (1997) Determining divergence times with a protein clock: update and reevaluation. *Proc. Natl. Acad. Sci. USA*, 94, 13028-13033.
- Fernandez de Henestrosa, A. R., Cuñé, J., Erill, I., Magnuson, J. K., Barbé, J. (2002) A green nonsulfur bacterium, *Dehalococcoides ethenogenes*, with the LexA binding sequence found in gram-positive organisms. *J. Bacteriol.*, 184, 6073-6080.
- Fernandez de Henestrosa, A. R., Ogi, T., Aoyagi, S., Chafin, D., Hayes, J. J., Ohmori, H., Woodgate, R. (2000) Identification of additional genes belonging to the LexA regulon in *Escherichia coli*. *Mol. Microbiol.*, 35, 1560-1572.
- Fox, G. E., Stackebrandt, E., Hespel, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsén, K. R., Chen, K. N., Woese, C. R. (1980) The Phylogeny of Prokaryotes. *Science*, 209, 457-463.
- Fukushima, M., Kakinuma, K., Kawaguchi, R. (2002) Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the gyrB gene sequence. *J. Clin. Microbiol.*, 40, 2779-2785.
- Garriga, X., Calero, S., Barbé, J. (1992) Nucleotide sequence analysis and comparison of the *lexA* genes from *Salmonella typhimurium*, *Erwinia carotovora*, *Pseudomonas aeruginosa* and *Pseudomonas putida*. *Mol. Gen. Genet.*, 236, 125-134.
- Gelfand, M. S., Koonin, E. V., Mironov, A. A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative-genomic approach. *Nucleic Acids Res.*, 28, 695-705.
- Gelfand, M. S., Novichkov, P. S., Novichkova, E. S. and Mironov, A. A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. *Briefings in Bioinformatics*, 1, 357-371.
- Gupta, R. S. (2000) The natural evolutionary relationships among prokaryotes. *Crit. Rev. Microbiol.*, 26, 111-131.
- Henikoff, S. Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89, 10915-10519.
- Khil, P.P., Camerini-Otero, R.D. (2002) Over 1000 genes are involved in the DNA damage response of *Escherichia coli*. *Mol. Microbiol.*, 44, 89-105.
- Ko, K. S., Lee, H. K., Park, M. Y., Lee, K. H., Yun, Y. J., Woo, S. Y., Miyamoto H, Kook YH. (2002) Application of RNA polymerase beta-subunit gene (*rpoB*) sequences for the molecular differentiation of *Legionella* species. *J. Clin. Microbiol.*, 40, 2653-2658.
- Koch, W. H., Woodgate, R. (1998) The SOS response In J. A. Nickoloff and M. F. Hoekstra (Eds.), *DNA damage and repair: DNA repair in prokaryotes and lower eukaryotes*. Humana Press, Totowa, New Jersey, 107-134.
- Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., Wootton, J. C. (1993) Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262, 208-214.
- Lawrence, C. E., Reilly, A. A. (1990) An EM Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymers Sequence. *Proteins*, 7, 41-51.
- Lewis, L. K., Harlow, G. R., Gregg-Jolly, L. A., Mount, D. W. (1994) Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in *Escherichia coli*. *J. Mol. Biol.*, 241, 507-523.
- Little, J. W. (1984) Autodigestion of *lexA* and phage repressors. *Proc. Natl. Acad. Sci. USA*, 81, 1375-1379.
- Lloyd, A. T., Sharp, P. M. (1993) Evolution of the *recA* gene and the molecular phylogeny of bacteria. *J. Mol. Evol.*, 37, 399-407.
- Ludwig W, Weizenegger M, Betzl D, Leidel E, Lenz T, Ludvigsen A, Mollenhoff D, Wenzig P, Schleifer KH. (1990) Complete nucleotide sequences of seven eubacterial genes coding for the elongation factor Tu: functional, structural and phylogenetic evaluations. *Arch. Microbiol.*, 153, 241-247.
- Makarova, K. S., Andrey, Mironov, A. A., Gelfand, M. S. (2001) Conservation of the binding site for the arginine repressor in all bacterial lineages. *Genome Biol.*, 2, research0013.1-0013.8.
- McCue, L. A., Thompson, W., Carmack, C. S., Ryan, M. P., Liu, J. S., Derbyshire, V., Lawrence, C. E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, 29, 774-782.
- McGuire, A. M., Hughes, J. D., Church, G. M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.*, 10, 744- 757.
- Nomenclature Committee (1985) IUB Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences Recommendations 1984. *Eur. J. Biochem.*, 150, 1-5.
- Ochman, H., Wilson, A. C. (1987) Evolution in bacteria: evidence for a universal rate in cellular genomes. *J. Mol. Evol.*, 26, 74-86.
- Page, R. D. M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, 12, 357-358.

- Pitulle, C., Strehse, C., Brown, J. W., Breitschwerdt, E. B. (2002) Investigation of the phylogenetic relationships within the genus *Bartonella* based on comparative sequence analysis of the mpB gene, 16S rDNA and 23S rDNA. *Int. J. Syst. Evol. Microbiol.*, 52, 2075-2080.
- Prakash, L., *et al.* (eds). Springfield: Charles C. Thomas, 128-142.
- Radman, M. (1974) Phenomenology of an inducible mutagenic DNA repair pathway in
- Rajewsky, N., Socci, N., Zapotocky, M., Siggia, E. D. (2002) The Evolution of DNA Regulatory Regions for Proteo-gamma Bacteria by Interspecies Comparisons. *Genome Res.*, 12, 298-308.
- Rodionov, D. A., Mironov, A. A., Gelfand, M. S. (2001) Transcriptional regulation of pentose utilisation systems in the *Bacillus/Clostridium* group of bacteria. *FEMS Microbiol. Lett.*, 205, 305-314.
- Roy, S., Sahu, A., Adhya, S. (2002) Evolution of DNA binding motifs and operators. *Gene*, 285, 169-173.
- Snel, B., Bork, P., Huynen, M. A. (1999) Genome phylogeny based on gene content. *Nature Genet.*, 21, 108-110.
- Stormo, G. D., Hartzell, G. W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA*, 86, 1183-1187.
- Tan K., Moreno-Hagelsieb G, and Collado-Vides J. and Stormo G. D. (2001) A Comparative Genomics Approach to Prediction of New Members of Regulons. *Genome Res.*, 11, 566-584.
- Tapias, A., Fernández, S., Alonso, J. C., Barbé, J. (2002) *Rhodobacter sphaeroides* LexA has dual activity: optimising and repressing recA gene transcription. *Nucleic Acids Res.*, 30, 1539-1546.
- Tekaia, F., Lazcano, A., Dujon, B. (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.*, 9, 550-557.
- van Helden, J., André, B., Collado-Vides, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281, 827-842.
- Walker, G. C. (1984) Mutagenesis and inducible responses to deoxyribonucleic acid damage in *Escherichia coli*. *Microbiol Rev.*, 48, 60-93.
- Winterling, K. W., Chafin, D., Hayes, J. J., Sun, J., Levine, A. S., Yasbin, R. E., Woodgate, R. (1998) The *Bacillus subtilis* DinR binding site: redefinition of the consensus sequence. *J. Bacteriol.*, 180, 2201-2211.
- Woese, C. R. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. USA*, 95, 6854-6859.
- Woese, C. R., Fox, G. E. (1977) Phylogenetic Structure of the Prokaryotic Domains: The Primary Kingdoms. *Proc. Natl. Acad. Sci. USA*, 74, 5088-5090.
- Xie, G., Bonner, C. A., Brettin, T., Gottardo, R., Keyhani, N. O. Jensen, R. A. (2003) Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in *Xylella* species and in heterocystous cyanobacteria. *Genome Biol.*, 4, R14.
- Zuckerkandl, E. (1975) The appearance of new structures and functions in proteins during evolution. *J. Mol. Evol.*, 7, 1-57.

Table 1. Experimentally determined regulatory motifs for LexA-regulated genes in *E. coli*; *d* indicates the distance (bp) from motif-end to ORF start codon (Sources: Fernández de Henetrosa *et al.*, 2000; Courcelle *et al.*, 2001). Rightmost columns show, where available, the comparative results of Consensus/Patser (C/P), dyad sweeping (D) and RCGScanner (R) methods in binding-site predictions (Source: Benítez-Bellón *et al.*, 2002).

Gene name	Regulatory motif/s	<i>d</i>	C/P	D	R	Gene name	Regulatory motif	<i>d</i>	C/P	D	R
dinB/dinP	CTgTATACTTTACCAg	17	-	+	+	ssb/lexC	CTgAATgAATATACAg	26	+	+	+
dinG/rarB	TTggCTgTTTATACAg	17	+	+	-	uvrA	CTgTATATTCATTCAg	86	+	+	+
dinI	CTgTATAAATAACCAg	22	+	+	+	uvrB	CTgTTTTTTTATCCAg	77	+	+	+
ftsK/dinH	CTgTTAATCCATACAg	79	-	+	+	uvrD/recL	CTgTATATATACCCAg	5	+	+	+
lexA/dinF	CTgTATATACTCACAg	9	+	+	+	yebG	CTgTATAAAATCACAg	20	+	+	+
	CTgTATATACACCCAg	30			+	sulA/sfiA	CTgTACATCCATACAg	19	+	+	+
pcsA/dinD	CTgTATATAAATACAg	34	+	+	+	umuD	CTgTATATAAAAACAg	22	+	+	+
polB/dinA	CTgTATAAAACCACAg	17	-	+	+	yjiW	CTgATgATATATACAg	21	+	+	+
recA/lexB	CTgTATgCTCATACAg	47	+	+	+	molR	CTggATAAAATTACAg	10	+	-	+
recN/radB	CTgTATATAAAACCAg	29	+	+	+	yigN	CTggACgTTTgTACAg	46	-	-	+
	CTgTACACAATAACAg	51			+	ybfE	CTgggTTTTTAATCAg	15	-	-	-
	ATggTTTTTCATACAg	11			+	ydjM	CTgTACgTATCgACAg	5	+	+	-
ruvAB	CTggATATCTATCCAg	52	+	-	+	ydjQ/cho	CTggATAgATAACCAg	24	-	-	-
sbmC/gyrI	CTgTATATAAAAACAg	31	+	+	+	hokE/ybdY	CTgTATAAATAAACAg	180			+

Table 2. Previously unreported putative LexA binding motifs in *E. coli* SOS genes identified in this work; *d* indicates the distance (bp) from motif-end to ORF start codon (+ indicates an intragenic motif).

Gene name	Putative regulatory motif	<i>d</i>
dinB/dinP	CTgAATCTTTACgCAT	52
dinI	CTggTCCgTTAAACAA	77
lexA/dinF	CTggTTTATTgTgCAg	71
pcsA/dinD	ATgTTTTTTTgCCCAg	77
recN/radB	CTgATTCATCCgAAA	145
ruvA	TTgATTCATTACgCAg	10
	CTgTgCCATTTTTCAg	105
sbmC/gyrI	CTACgAgATTAAgCAg	+6
	CTgCTCgCATAATCAA	82
uvrD/recL	CTgATATAATCAgCAA	23
yebG	TTgCTgCCggACgCAg	155
umuD	CTgCTggCAAgAACAg	42
yjiW	CTgAACgCgCAgCTAg	205
	CTggAAAAAATCAA	221
molR	CTggTAGCATCTgCAT	30
ydjM	CTTTCATCgCTgACAg	180

Table 3. Conserved genes and regulatory motifs among gamma proteobacteria species for all 25 LexA-regulated genes (28 regulatory motifs) experimentally described in *E. coli*, and possible additions to the LexA-regulon in different species. The table was constructed as compound of sensitive and restrictive searches (i.e. combined method), plus manual annotation. Bold characters reflect motifs selected in restrictive searches, while italics indicate selection in broad-spectrum searches and plain characters denote non-selected, manually located, motifs. Due to the incomplete nature of the *K. pneumoniae* genome, which does not allow a systematic determination of conservation/loss of genes, this species was not included in the analysis. (*d* indicates distance to ORF. *Name* corresponds to standard sequence annotation numbers and it presence denotes conservation of the gene in the respective genome; * indicates a conserved intergenic region and ** TIGR annotation numbers, respectively).

<i>E. coli</i> K-12				<i>S. flexneri</i>				<i>S. typhimurium</i>				<i>Y. pestis</i>				<i>V. cholerae</i>				<i>P. multocida</i>				<i>H. influenzae</i>				<i>P. aeruginosa</i>				<i>R. solanacearum</i>			
Gene	Name	LexA motif	d	Gene	Name	LexA motif	d	Gene	Name	LexA motif	d	Gene	Name	LexA motif	d	Gene	Name	LexA motif	d	Gene	Name	LexA motif	d	Gene	Name	LexA motif	d	Gene	Name	LexA motif	d				
lexA_1	b0403	CTGTATATACACCCAG	9	SP4162	CTGTATATACACCCAG	9	STM4237	30	STM4237	TGTATATACACCCAG	31	YPO0314	CTGTATATACACCCAG	9	VC0092	CTGTATATACACCCAG	9	VC0092	CTGTATATACACCCAG	9	VC0092	CTGTATATACACCCAG	9	VC0092	CTGTATATACACCCAG	9	VC0092	CTGTATATACACCCAG	9	VC0092	CTGTATATACACCCAG	9			
lexA_2	b0403	CTGTATATACACCCAG	30	SP4162	CTGTATATACACCCAG	30	STM4237	30	STM4237	TGTATATACACCCAG	31	YPO0314	CTGTATATACACCCAG	30	VC0092	CTGTATATACACCCAG	30	VC0092	CTGTATATACACCCAG	30	VC0092	CTGTATATACACCCAG	30	VC0092	CTGTATATACACCCAG	30	VC0092	CTGTATATACACCCAG	30	VC0092	CTGTATATACACCCAG	30			
lexA	b2609	CTGTATATACACCCAG	47	SP2722	CTGTATATACACCCAG	47	STM2829	47	STM2829	CTGTATATACACCCAG	64	YPO0307	CTGTATATACACCCAG	64	VC0053	CTGTATATACACCCAG	64	VC0053	CTGTATATACACCCAG	64	VC0053	CTGTATATACACCCAG	64	VC0053	CTGTATATACACCCAG	64	VC0053	CTGTATATACACCCAG	64	VC0053	CTGTATATACACCCAG	64			
recA_1	b2616	CTGTATATATACACAG	29	SP2675	CTGTATATATACACAG	29	STM2684	29	STM2684	CTGTATATATACACAG	29	YPO1105	CTGTATATATACACAG	29	VC0082	CTGTATATATACACAG	29	VC0082	CTGTATATATACACAG	29	VC0082	CTGTATATATACACAG	29	VC0082	CTGTATATATACACAG	29	VC0082	CTGTATATATACACAG	29	VC0082	CTGTATATATACACAG	29			
recA_2	b1861	CTGTATATATACACAG	52	SP1871	CTGTATATATACACAG	52	STM1895	52	STM1895	CTGTATATATACACAG	52	YPO1105	CTGTATATATACACAG	52	VC1846	CTGTATATATACACAG	52	VC1846	CTGTATATATACACAG	52	VC1846	CTGTATATATACACAG	52	VC1846	CTGTATATATACACAG	52	VC1846	CTGTATATATACACAG	52	VC1846	CTGTATATATACACAG	52			
recA_3	b4059	CTGTATATATACACAG	20	SP4145	CTGTATATATACACAG	20	STM2456	90	STM2456	CTGTATATATACACAG	151	YPO0325	CTGTATATATACACAG	151	VC0087	CTGTATATATACACAG	151	VC0087	CTGTATATATACACAG	151	VC0087	CTGTATATATACACAG	151	VC0087	CTGTATATATACACAG	151	VC0087	CTGTATATATACACAG	151	VC0087	CTGTATATATACACAG	151			
recA_4	b0231	CTGTATATATACACAG	17	SP0279	CTGTATATATACACAG	17	STM0313	17	STM0313	CTGTATATATACACAG	3	YPO0323	CTGTATATATACACAG	3	VC2287	CTGTATATATACACAG	3	VC2287	CTGTATATATACACAG	3	VC2287	CTGTATATATACACAG	3	VC2287	CTGTATATATACACAG	3	VC2287	CTGTATATATACACAG	3	VC2287	CTGTATATATACACAG	3			
recA_5	b4058	CTGTATATATACACAG	86	SP4146	CTGTATATATACACAG	86	STM2454	86	STM2454	CTGTATATATACACAG	83	YPO0324	CTGTATATATACACAG	83	VC0394	CTGTATATATACACAG	83	VC0394	CTGTATATATACACAG	83	VC0394	CTGTATATATACACAG	83	VC0394	CTGTATATATACACAG	83	VC0394	CTGTATATATACACAG	83	VC0394	CTGTATATATACACAG	83			
recA_6	b2616	CTGTATATATACACAG	51	SP3891	CTGTATATATACACAG	51	STM2684	51	STM2684	CTGTATATATACACAG	51	YPO1105	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51			
recA_7	b1848	CTGTATATATACACAG	15	SP1858	CTGTATATATACACAG	15	STM1882	150	STM1882	CTGTATATATACACAG	19	YPO1105	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19			
recA_8	b2616	CTGTATATATACACAG	51	SP3891	CTGTATATATACACAG	51	STM2684	51	STM2684	CTGTATATATACACAG	51	YPO1105	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51	VC0082	CTGTATATATACACAG	51			
recA_9	b1848	CTGTATATATACACAG	15	SP1858	CTGTATATATACACAG	15	STM1882	150	STM1882	CTGTATATATACACAG	19	YPO1105	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19	VC0082	CTGTATATATACACAG	19			
recA_10	b3832	CTGTATATATACACAG	46	SP3910	CTGTATATATACACAG	46	STM0881	46	STM0881	CTGTATATATACACAG	47	YPO0325	CTGTATATATACACAG	47	VC0087	CTGTATATATACACAG	47	VC0087	CTGTATATATACACAG	47	VC0087	CTGTATATATACACAG	47	VC0087	CTGTATATATACACAG	47	VC0087	CTGTATATATACACAG	47	VC0087	CTGTATATATACACAG	47			
recA_11	b0799	CTGTATATATACACAG	17	SP0748	CTGTATATATACACAG	17	STM0798	73	STM0798	CTGTATATATACACAG	39	YPO1206	CTGTATATATACACAG	39	VC1855	CTGTATATATACACAG	39	VC1855	CTGTATATATACACAG	39	VC1855	CTGTATATATACACAG	39	VC1855	CTGTATATATACACAG	39	VC1855	CTGTATATATACACAG	39	VC1855	CTGTATATATACACAG	39			
recA_12	b0779	CTGTATATATACACAG	73	SP0729	CTGTATATATACACAG	73	STM0798	73	STM0798	CTGTATATATACACAG	76	YPO1156	CTGTATATATACACAG	76	VC1855	CTGTATATATACACAG	76	VC1855	CTGTATATATACACAG	76	VC1855	CTGTATATATACACAG	76	VC1855	CTGTATATATACACAG	76	VC1855	CTGTATATATACACAG	76	VC1855	CTGTATATATACACAG	76			
recA_13	b0660	CTGTATATATACACAG	17	SP1065	CTGTATATATACACAG	17	STM0097	56	STM0097	CTGTATATATACACAG	57	YPO0618	CTGTATATATACACAG	57	VC0082	CTGTATATATACACAG	57	VC0082	CTGTATATATACACAG	57	VC0082	CTGTATATATACACAG	57	VC0082	CTGTATATATACACAG	57	VC0082	CTGTATATATACACAG	57	VC0082	CTGTATATATACACAG	57			
recA_14	b0690	CTGTATATATACACAG	79	SP1069	CTGTATATATACACAG	79	STM0960	79	STM0960	CTGTATATATACACAG	80	YPO1376	CTGTATATATACACAG	80	VC1903	CTGTATATATACACAG	80	VC1903	CTGTATATATACACAG	80	VC1903	CTGTATATATACACAG	80	VC1903	CTGTATATATACACAG	80	VC1903	CTGTATATATACACAG	80	VC1903	CTGTATATATACACAG	80			
recA_15	b0958	CTGTATATATACACAG	19	SP0958	CTGTATATATACACAG	19	STM1071	19	STM1071	CTGTATATATACACAG	18	YPO1436	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18			
recA_16	b1728	CTGTATATATACACAG	5	SP1192	CTGTATATATACACAG	5	STM1321	5	STM1321	CTGTATATATACACAG	22	YPO1717	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22			
recA_17	b1061	CTGTATATATACACAG	22	SP1067	CTGTATATATACACAG	22	STM1162	36	STM1162	CTGTATATATACACAG	22	YPO1886	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22	VC1903	CTGTATATATACACAG	22			
recA_18	b1183	CTGTATATATACACAG	22	SP1172	CTGTATATATACACAG	22	STM1998	22	STM1998	CTGTATATATACACAG	18	YPO1998	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18	VC1903	CTGTATATATACACAG	18			
recA_19	b4347	CTGTATATATACACAG	21	SP1463	CTGTATATATACACAG	21	STM4523	78	STM4523	CTGTATATATACACAG	27	YPO1998	CTGTATATATACACAG	27	VC1903	CTGTATATATACACAG	27	VC1903	CTGTATATATACACAG	27	VC1903	CTGTATATATACACAG	27	VC1903	CTGTATATATACACAG	27	VC1903	CTGTATATATACACAG	27	VC1903	CTGTATATATACACAG	27			
recA_20	b2115	CTGTATATATACACAG	10	SP2180	CTGTATATATACACAG	10	STM2136	10	STM2136	CTGTATATATACACAG	86	YPO1998	CTGTATATATACACAG	86	VC1903	CTGTATATATACACAG	86	VC1903	CTGTATATATACACAG	86	VC1903	CTGTATATATACACAG	86	VC1903	CTGTATATATACACAG	86	VC1903	CTGTATATATACACAG	86	VC1903	CTGTATATATACACAG	86			
recA_21	b0685	CTGTATATATACACAG	15	SP0685	CTGTATATATACACAG	15	STM0695	15	STM0695	CTGTATATATACACAG	24	YPO1998	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24			
recA_22	b1741	CTGTATATATACACAG	24	SP1485	CTGTATATATACACAG	24	STM1309	24	STM1309	CTGTATATATACACAG	34	YPO1998	CTGTATATATACACAG	34	VC1903	CTGTATATATACACAG	34	VC1903	CTGTATATATACACAG	34	VC1903	CTGTATATATACACAG	34	VC1903	CTGTATATATACACAG	34	VC1903	CTGTATATATACACAG	34	VC1903	CTGTATATATACACAG	34			
recA_23	b3645	CTGTATATATACACAG	31	SP3684	CTGTATATATACACAG	31	STM3064	31	STM3064	CTGTATATATACACAG	24	YPO1998	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24			
recA_24	b2009	CTGTATATATACACAG	31	SP3064	CTGTATATATACACAG	31	STM3064	31	STM3064	CTGTATATATACACAG	24	YPO1998	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24			
recA_25	b1399	CTGTATATATACACAG	180	SP1094	CTGTATATATACACAG	180	STM1094	180	STM1094	CTGTATATATACACAG	24	YPO1998	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24	VC1903	CTGTATATATACACAG	24			

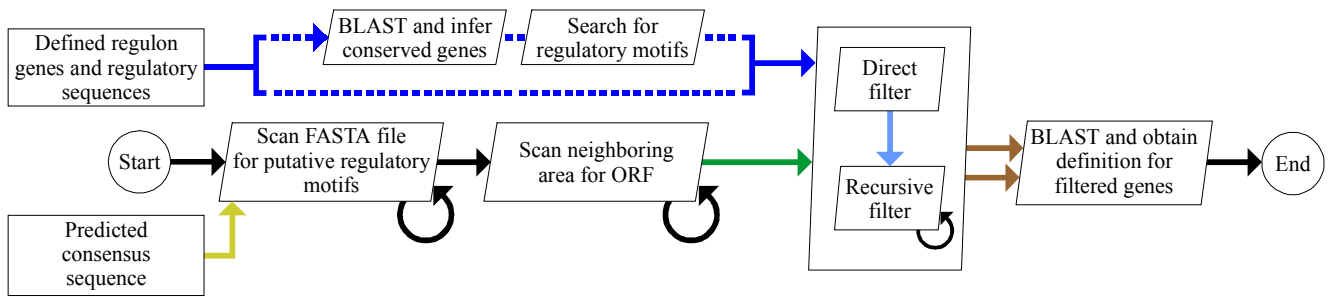


Fig. 1. Schematic diagram of data and program flow in RCGScanner. The search function can backtrack to locate more than one regulatory motif and/or more than one ORF under control of this/these.

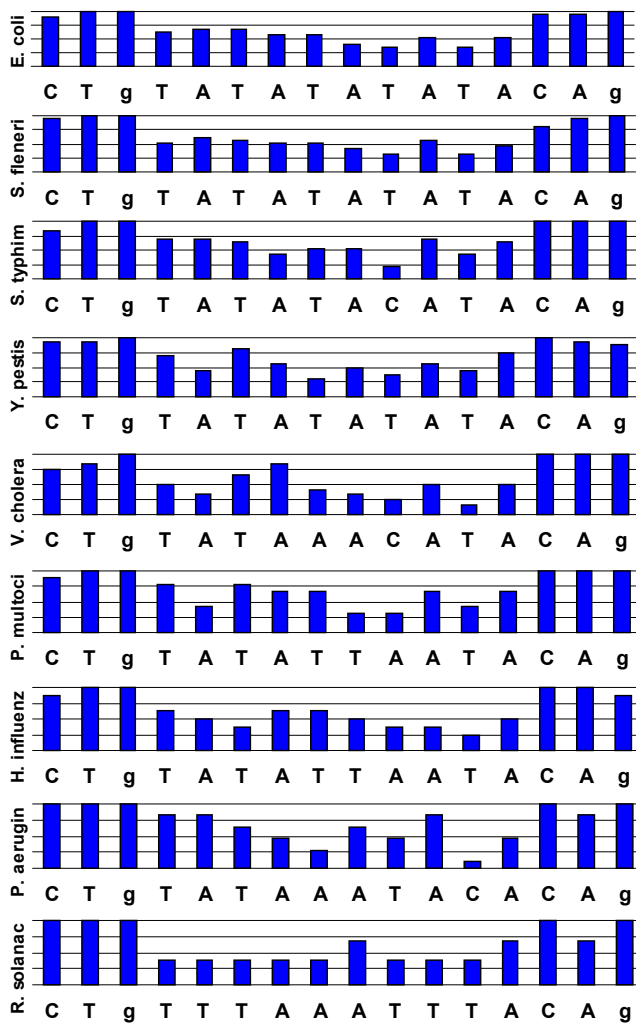


Fig. 2. Consensus LexA regulatory sequences for the studied gamma proteobacteria species. The bars are percentile representations of the statistical occurrence of the consensus base at each position. It is interesting to note the likeness of the consensus sequences for the subset of species with an explicitly regulated *sulA* gene (*E. coli*, *S. flexneri*, *S. typhimurium* and *Y. pestis*; the *sulA* gene of *P. aeruginosa* does not have a dedicated LexA box but seems, instead, to be regulated by the LexA box of the *lexA* operon). Albeit its low statistical significance, such a similitude in consensus sequences endorses the idea that direct *sulA* regulation imposes a bottleneck effect on regulatory protein variability. Consensus sequences were computed from the found regulatory motifs putatively regulating orthologs of described *E. coli* LexA regulon genes (Table 3).

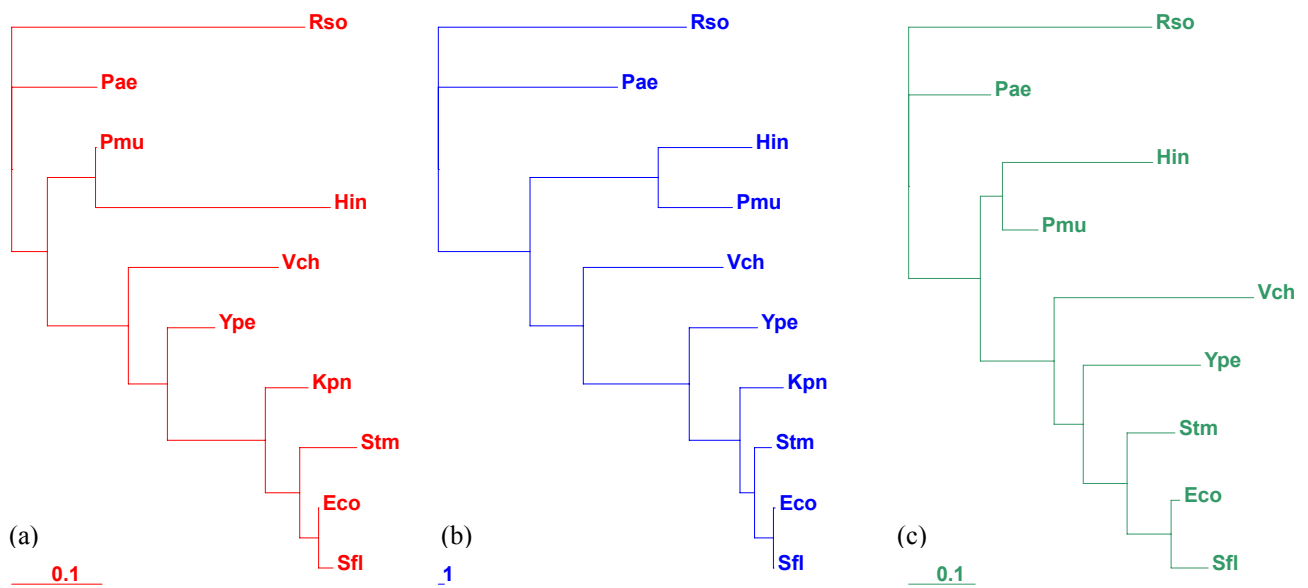


Fig. 3. Phylogeny trees generated using (a) regulon core (*lexA*, *recA* and *recN*) LexA boxes, (b) regulon core protein sequences and (c) all conserved LexA boxes. Note that due to the incomplete nature of the *K. pneumoniae* genome, which prevented foolproof identification of conserved genes, this bacterial species is not included in the analysis leading to tree (c). Also, in (c) analysis, hyphens (-) were placed in DNAML input file to indicate non-preserved regulatory motifs. Rso: *R. solanacearum*; Pae: *P. aeruginosa*; Pmu: *P. multocida*; Hin: *H. influenzae*; Vch: *V. cholerae*; Ype: *Y. pestis*; Kpn: *K. pneumoniae*; Stm: *S. typhimurium*; Eco: *E. coli*; Sfl: *S. flexneri*.