We present a computational approach that identifies regulatory elements conserved across phylogenetically distant organisms. Intergenic regulatory regions were clustered by orthology of the adjacent genes, and an iterative process was applied to search for significant motifs, enabling new elements of the putative regulon to be added in each cycle. With this approach, we identified highly conserved riboswitches and the Gram positive T-box. Interestingly, we identified many other regulatory systems that appear to depend on conserved RNA structures.

Comparative genomic approaches are central to analyzing the increasing number of whole-genome sequences. Although using this kind of analysis to find regulatory elements is not new, the focus has usually been on one genome or group of closely related genomes [1–3] because sequence conservation of functional intergenic regions (promoters, protein binding sites) is usually low, and quickly diverges.

It came as a surprise to many scientists when specific RNA ‘riboswitches’ were shown to be capable of regulating gene expression by directly sensing a metabolite without the intervention of a protein [4]. RNA riboswitches have since been shown to be involved in various metabolic processes including thiamine, riboflavin, cobalamin, adenine, guanine and lysine biosynthesis [5–11]. We assumed that this type of regulatory sequence would be easily identified given their broad phylogenetic distribution and highly conserved nature.

Searching for interesting motifs
The starting point for our work is a set of orthologous regulatory regions. To obtain these we used the Cluster of Orthologous Groups (COG) of proteins database (http://www.ncbi.nlm.nih.gov/COG/) [12] together with operon predictions based on intergenic distances [13]. In this manner, every protein from 164 fully sequenced bacterial genomes that was associated with a COG was assigned to the intergenic minimal upstream region (iMUR) of the first gene of the predicted operon to which it belongs. To avoid over-representation of similar sequences from related genomes, redundant sequences were eliminated. We obtained ~4000 clusters of orthologous regulatory regions, each belonging to a different COG.

We used the public domain motif discovery tool Multiple EM for Motif Elicitation (MEME) [14] to find a set of over-represented ‘seed motifs’ for each COG (Figure 1a). These motifs were used to identify other members of the putative regulon by searching in all upstream regions using the MEME counterpart Motif Alignment and Search Tool (MAST) [15]. As a result of this
search, new members were added to each group (and some original ones were removed), and a new and more specific set of motifs was obtained, again using MEME. This cycle of locating over-represented motifs (with MEME) followed by searching for new genes containing the motifs (with MAST) was repeated until no new members were found (Figure 1b). The resulting 'refined motifs' are our candidate regulatory elements. Because many of these results are redundant, a clustering process was applied based on the number of genes in common for each motif, resulting in 672 different groups of motifs (for more details, see http://www.ibt.unam.mx/biocomputo/conserved_motifs.html). It is worth mentioning that each group can have several motifs (up to four), and that the motifs are larger than usual protein binding sites (the average length of our motif is 43 nucleotides). For example, the 'thiamine ribo-switch' Group 0012 (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html) was obtained from 20 different COGs that converge to form a common set of genes. Following analysis of the four different motifs that define this group, we found that the most significant one contains the thi box and the next two overlap with part of the structure of the THI-element; the fourth motif appears to contain part of the transcriptional terminator with its poly-U tract (Figure 1c).

Evaluating the motifs
To evaluate the likelihood that the 'refined motifs' in our groups represent biologically important regulatory elements, we first eliminated those that matched known proteins (from the non redundant GenBank database) or RNA genes (rRNA, tRNA, scRNA, snRNA and so on) from fully sequenced genomes. This step was necessary given the possible existence of unannotated small genes or erroneously assigned translation start sites resulting from automatic genome annotation. For example, ribosomal protein L36 (~40 amino acids long) is not annotated in O15 genomes but is picked up by our method, when analyzing the upstream region of ribosomal proteins S13 and S11 (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html, Group 0011 marked as a false positive). The groups of 'refined motifs' were then assigned to operons and their statistical significance was evaluated as follows. A P-value (assuming a hypergeometrical distribution) was calculated for each motif to be over-represented in a given Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (http://www.genome.ad.jp/kegg/pathway.html) [16]. The lower the P-value, the more likely it is that the motifs can be considered to be biologically relevant for the regulation of that pathway. A similar evaluation was performed using COG assignations. We were able to

Figure 1. General procedure to identify conserved regulatory motifs. (a) The intergenic Minimal Upstream Regions (iMUR) of the operons that code for each COG (in this example thiC COG0422) from fully sequenced genomes are grouped and over-represented motifs identified using MEME [14]. These motifs constitute the initial, or 'seed motifs' of the group. (b) These motifs are used to identify new members of the putative regulon in the entire set of iMURs using MAST [15]. To obtain 'refined motifs' that better represent the expanded group, we once again used MEME. This cycle is iteratively performed until no new elements are added. (c) At the final step of the process all the genes located by the 'refined motifs' are collected. In the example, the genes are found to belong to the thiamine biosynthesis metabolic pathway, and the conserved motifs correspond to important structural elements of the THI element. Abbreviations: MEME, Multiple EM for Motif Elicitation; MAST, Motif Alignment and Search Tool.
assign a COG or KEGG with a P-value \(<1\times10^{-6}\) in all the cases with the exception of only ten examples that are most probably false positives (motifs found by MEME that do not have any biological meaning). However, because many of the genes in these groups are not functionally annotated, they might still represent biologically relevant elements of unknown function. Finally, the genome context congruence of our putative regulons was verified using gene context analysis (GeConT) [17], a web tool that enables the user to view the neighbours for any set of genes together with their functions and orthologous relationships. The data on our web page (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html) is hyperlinked
to this application so that the co-regulated groups can be visualized easily.

Analyzing the nature of the conserved motifs
When sorting our results by P-values or by conservation, we realized that many of the first groups contain motifs that correspond to previously described riboswitches, known to regulate genes involved in the biosynthesis of different metabolites such as thiamine, riboflavin, cobalamin, adenine and guanine, in addition to the T-box regulator [18] of aminoacyl-tRNA synthetases from Gram positive bacteria (Table 1). Interestingly, we also found important sequence conservation in different families of aminoacyl-tRNA synthetases in Gram negative bacteria. For example, in *Escherichia coli* threeyl-tRNA synthetase, it is known that the mRNA leader region can adopt a tRNA-like structure that is specifically recognized by the corresponding threonyl-tRNA synthetase, establishing an auto-regulatory cycle [19]. The conserved motifs that we identified for this regulatory system correspond to parts of the stem-loop structure that resembles the threomine-tRNA anticodon CGU and to a stable structure that is similar to the acceptor arm of tRNA\(^{\text{Thy}}\). Although it has not been reported previously, the motifs that we found for glutamyl and glutaminyl-tRNA synthetases could participate in a similar mechanism. We also found that many of our groups correspond to ribosomal protein operons – we detected 43 different groups of such operons, most of which correspond to specific phyla (Table 1). Self-regulation has been described for these cases because ribosomal protein L4 is known to bind to its operator, where a complex secondary structure appears to mimic the natural binding site of L4 in the ribosome [20]. We expect that most, if not all, of these operons to be auto-regulated by one of their proteins. Other well-known examples in Table 1 include a pyrimidine biosynthesis group, where our identified motifs correspond to the conserved RNA secondary structure that comprises the binding site for PyrR [21] and the ‘Controlling Inverted Repeat for Chaperon Expression’ (CIRCE), which constitutes the binding site for the HcrA repressor [22]. In all these examples (as occurs with riboswitches), sequence conservation in the regulatory region is a consequence of the constraints imposed by the required RNA structure.

Another case that we found interesting was glycine cleavage. Although regulatory mechanisms have been described for the gcv operon in *E. coli* [23], we found a completely different regulatory system. The organisms that present our glycine cleavage motifs do not include *E. coli* but are mostly actinobacteria, firmicutes and \(\alpha\) and \(\beta\) proteobacteria, and most of them do not encode orthologs of GcvA or GcvR, the two reported specific regulators of the gcv operon [23]. Furthermore, the reported binding sites for GcvA, do not match our motifs. Interestingly, the same glycine cleavage motifs are present in several proteins that are assigned to a Na\(^{+}\) and alanine symporter COG. This could be part of a glycine transport system, and would make this putative regulon more similar to several riboswitches, where metabolic and transporting proteins are regulated by the same element. Many other examples of proposed regulatory systems with their conserved motifs can be found on our web page (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html).

Concluding remarks
We have developed a computer method that detects previously reported riboswitches, other already known conserved elements and > 600 statistically significant groups of conserved motifs that would appear to be biologically relevant. We have shown that for a great many regulatory elements their conservation is sufficiently strong to be detected in a single orthologous cluster of genes, without the need to consider additional elements of the regulon or metabolic pathway. In many cases, our motifs coincide with regulatory elements that are reported for specific model organisms such as *E. coli* or *Bacillus subtilis*. We are now able to propose the extent to which these systems have been conserved among fully sequenced bacteria. Several of the signals detected by our methods were related to an RNA secondary structure rather than classic DNA-binding regulators. Although we did not expect to find many conventional protein binding sites (the conserved portion being too small to be detected by our method), we were surprised by the number of structure-dependant regulatory elements that were also highly conserved at the sequence level.

Our study highlights potential new motifs to be further experimentally characterized in terms of their ability to form RNA secondary structures, such as attenuators, and their ability to bind small RNAs, cellular metabolites or regulatory proteins. All this will further help us to understand and define the regulatory mechanisms of these systems.

Acknowledgements
We wish to thank Shirley Ainsworth for bibliographical assistance, as well as Jérôme Verleyen and Roberto Bahena for computer support. We are indebted to Jose Antonio Loza for helping us with the \(P\)-value calculations. C.A.G was supported by fellowships from CONACyT and DGEP-UNAM. This work as partially supported by CONACyT grant 44213-Q.

References
Is avian sex determination unique?: clues from a warbler and from chickens

Shinichi Nakagawa

Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

A recent report on a ZZW female great reed warbler, together with existing data from chickens, might shed new light on an unknown mechanism of avian sex determination.

In contrast to mammals where females carry XX and males XY chromosomes, female birds are heterogametic (ZW), whereas male birds are homogametic (ZZ). Although sex chromosomes are known to have a crucial role in avian sex determination, the underlying mechanism surprisingly remains a mystery. It has been argued that the mechanism depends either on: (i) the female-specific avian W chromosome, which triggers female development (similar to the male-specific Y chromosome, which causes male development in mammals); or (ii) the number of Z chromosomes (one in females and two in males) in a manner that is similar to the dosage effects of X chromosomes found in Drosophila melanogaster and Caenorhabditis elegans [1–3]. To resolve this conundrum, one can study cases of sex chromosome aneuploidy (2A:ZZW and 2A:Z0, where A represents autosome). It has been thought that, depending on which of these karyotypes develops female or male characteristics, one can conclude whether the avian sex is determined by either the presence of a W chromosome or the number of Z chromosomes (in mammals, 2A:XXY develops male characteristics and 2A:X0 develops female characteristics). However, reports on sex chromosome aneuploidy in birds have been virtually nonexistent or unreliable until recently [4].

Clues from a warbler

Arlt et al. [5] have now described an example of an apparent 2A:ZZW great reed warbler (Acrocephalus arundinaceus) female breeding in a natural population. Genetic analysis of blood and skin samples showed that the bird was heterozygous at two Z-linked microsatellite loci (G61 and Aar1; the recombination rate at the two loci is ~0.5, indicating that they are distantly located), suggesting that this female carried two Z chromosomes. Interestingly, only one specific maternal allele was passed on to her male offspring (n = 12), whereas none of the maternal alleles were passed onto her female offspring (n = 5). The possibility of one specific allele being inherited in all twelve males is extremely low (0.5^12 = 0.00024). Therefore, Arlt et al. concluded that this female was trisomic (2A:ZZW) in somatic cells but normally diploid (2A:ZW) in the germline. Alternatively, it is possible that an occurrence of segmental duplications [6] in the Z chromosome could be responsible for this finding, although the observed pattern of inheritance and the distance between the two loci make this explanation highly unlikely. This first report of a possible ZZW female


Corresponding author: Shinichi Nakagawa (s.nakagawa@sheffield.ac.uk). Available online 7 August 2004