Molecular Microbiology (2004) **53**(2), 355–356 doi:10.1111/j.1365-2958.2004.04152.x

MicroCommentary

Building family traditions

Robert F. Schleif
Department of Biology, The John Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA.

In this issue of *Molecular Microbiology*, Yamazaki, Tomono, Ohnishi and Horinouchi (Yamazaki *et al*., 2004) present data showing that the regulator of the A-factor cascade in *Streptomyces griseus*, AdpA, possesses related family members, and they determine both the sites to which AdpA binds and a consensus DNA-binding sequence. Sequence homologues and consensus DNA binding sites of regulatory proteins were first reported more than 20 years ago, so why is this new work noteworthy? The answer is that AdpA is a member of the large and well-known AraC/XylS family. Sequence analysis of this group of interesting and important regulatory proteins has told us disappointingly little about how they function and what they sense, however, and its members have proved frustratingly difficult to study biochemically. This new work on AdpA foretells a welcome change.

AraC protein of *Escherichia coli* was the first member of the family to be studied (Gross and Englesberg, 1959; reviewed by Schleif, 2000). Because AraC itself is poorly soluble and difficult to purify, attention was soon also turned to related family members in the hope that they would be more tractable, and it was guessed that the rhamnose operon in *E. coli* would turn out to be regulated by an AraC homologue. This conjecture turned out to be correct and, indeed, two proteins, RhaR and RhaS, possessing similarity to AraC regulate the operon (Tobin and Schleif, 1987). This seeded the family for further growth that, with the advent of genomics, has become explosive. The NCBI database currently lists more than 2500 homologues in the AraC/XylS family. Many of these, however, are virtually the same protein in closely related bacteria, and one reason why the family is so big is that so many genomes have been sequenced. More telling of the size of the family is that restricting the search to specific strains reveals 39 homologues in *E. coli* K-12, 60 in *Salmonella enterica serovar Typhimurium* LT2, 42 in *Yersinia pestis* biovar *mediaevalis 91001*, 12 in the Ames strain of *Bacillus anthracis* and 50 in *Streptomyces avermitilis* MA-4680.

The DNA-binding domain of AraC comprises the C-terminal third of the protein, and it is through the amino acid sequence of this domain that AraC/XylS family members are identified. Two family members typically possess about 20% sequence identity over this region. Ordinarily, this level of homology is borderline for concluding that two proteins have similar structures and related functions but, because the domain is just over 100 amino acids long, and a number of residues are highly conserved, there is little doubt that most of the proteins that have been identified as AraC/XylS family members are indeed *bona fide* relatives possessing similar structure and function – at least with respect to their DNA-binding domains.

The N-terminal two-thirds of AraC constitute a regulatory domain but, until recently, the putative regulatory domains of AraC/XylS family members were mostly sequence orphans unrelated to other known family members or to other proteins in the greater sequence universe. These regulatory domains certainly did not evolve entirely on their own; they have to be related to other proteins. It should be noted, however, that the isolationist tendencies of the family may be in part a bioinformatics artifact. The majority of the family members have been identified solely from genome sequencing projects. The resulting open reading frames are then automatically compared with all other sequences and, of course, homologues with similar DNA-binding domains are revealed. Perhaps not seen and perhaps not yet analysed are homologies between regulatory modules that reside further down the similarity lists than those between the DNA-binding domains of family members.

The detection of sequence homologues, as reported by Yamazaki *et al*. (2004) for the regulatory module of AdpA, represents a comforting connection between the regulatory modules of some members of the AraC/XylS family and the rest of the protein universe. The relationship of the reported homology family to regulation by AdpA is not at all apparent however. The ‘regulatory’ domain of AdpA shows homology to members of the ThiJ/PfpI/DJ-1 family of proteins. This protein family was identified through genome sequencing and includes a glutamine amidot-

Accepted 29 March, 2004. For correspondence. E-mail rfschleif@jhu.edu; Tel. (+1) 410 516 5206; Fax (+1) 410 516 5213.

ransferase, an intracellular protease and a subunit in an RNA-binding complex. Not only is AdpA related to the outside world, it also possesses family members within the AraC/XylS family. Yamazaki *et al*. (2004) mention that bacteria related to *Streptomyces griseus*, *Streptomyces coelicolor A3* and *S. avermitilis* contain 10 or so genes coding for proteins possessing sequence similarity over the entire length of AdpA! Hence, more of the regulatory domains found in AraC/XylS family are beginning to look normal. They possess related family members and, also, they are related to proteins from outside the family.

Why is the identification of a consensus DNA binding site for AdpA also of note? Ordinarily, the consensus binding site for a particular protein can be deduced readily by examining half a dozen binding sequences. This often fails to identify binding sites of members of the AraC/XylS family however. The reason lies in the biochemistry. A single DNA-binding domain of a protein in the AraC family possesses two helix–turn–helix regions, one at each end of a connecting alpha helix. Thus, one domain is theoretically capable of contacting two adjacent major groove regions in the DNA and, hence, the normally dimeric AraC/XylS family members can potentially contact four major groove regions. Such a large DNA contact region allows a protein to get by without contacting all the bases over the region, and therefore to use a different subset of the bases in binding at different sites. Even a monomeric member of the AraC/XylS family, Rob, seems not to make all the DNA contacts it might (Kwon *et al*., 2000). With such a lackadaisical binding mode, it is not surprising that it is not easy to identify a consensus binding site from a handful of sequences known to be bound by a family member. In the case of AraC (Hendrickson and Schleif, 1985; Brunelle and Schleif, 1989), and now that of AdpA reported here, precise footprinting and protein contact experiments were necessary to locate and align the binding sites, from which a consensus could then be identified and verified with additional experiments. It should be noted that the work reported by Yamazaki *et al*. (2004) required substantial effort to develop a scheme for the partial purification of AdpA, and that detailed footprinting was only possible once this had been done. Thus, as has been the case so many times, genetic studies delineate the problems, then difficult and frequently lengthy biochemical groundwork is needed before elegant experiments combining genetics, physiology, biochemistry, sequencing and often biophysics reveal more answers.

References

Brunelle, A., and Schleif, R. (1989) Determining residue–base interactions between AraC protein and *araI* DNA. *J Mol Biol* **209:** 607–622.

Gross, J., and Englesberg, E. (1959) Determination of the order of mutational sites governing L-arabinose utilization in *Escherichia coli* B/r by transduction with phage P1bt. *Virology* **9:** 314–331.

Hendrickson, W., and Schleif, R. (1985) A dimer of AraC protein contacts three adjacent major groove regions of the *araI* DNA site. *Proc Natl Acad Sci USA* **82:** 3129–3133.

Kwon, H., Bennik, M., Demple, B., and Ellenberger, T. (2000) Crystal structure of the *Escherichia coli* Rob transcription factor in complex with DNA. *Nature Struct Biol* **7:** 424–430.

Schleif, R. (2000) Regulation of the L-arabinose operon of *Escherichia coli*. *Trends Genet* **16:** 559–565.

Tobin, J., and Schleif, R. (1987) Positive regulation of the *Escherichia coli* L-rhamnose operon is mediated by the products of tandemly repeated regulatory genes. *J Mol Biol* **196:** 789–799.

Yamazaki, H., Tomono, A., Ohnishi, Y., and Horinouchi, S. (2004) DNA-binding specificity of AdpA, a transcriptional activator in the A-factor regulatory cascade in *Streptomyces griseus*. *Mol Microbiol* **XX:** 000–000.