ADVANCES IN APPLIED MICROBIOLOGY

VOLUME 69

Advances in
APPLIED MICROBIOLOGY

VOLUME 69

Edited by

ALLEN I. LASKIN
Somerset, New Jersey, USA

SIMA SARIASLANI
Wilmington, Delaware, USA

GEOFFREY M. GADD
Dundee, Scotland, UK

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
32, Jamestown Road, London NW1 7BY, UK

First edition 2009

ISBN: 978-0-12-374824-9
ISSN: 0065-2164

Printed and bound in the USA
09 10 10 9 8 7 6 5 4 3 2 1

Advances in
APPLIED MICROBIOLOGY

VOLUME 69

CONTRIBUTORS

Ursula Bond
School of Genetics and Microbiology, Trinity College Dublin, Dublin 2, Ireland

Patrick J. Brennan
Department of Microbiology, Immunology and Pathology, Mycobacteria Research Laboratories, Colorado State University, Fort Collins, CO 80523-1682, USA

Marcelo E. Guerin
Department of Microbiology, Immunology and Pathology, Mycobacteria Research Laboratories, Colorado State University, Fort Collins, CO 80523-1682, USA

Paul A. Hoskisson
Strathclyde Institute of Pharmacy and Biological Science, University of Strathclyde, Royal College Building, George Street, Glasgow, United Kingdom.

Mary Jackson
Department of Microbiology, Immunology and Pathology, Mycobacteria Research Laboratories, Colorado State University, Fort Collins, CO 80523-1682, USA

Devinder Kaur
Department of Microbiology, Immunology and Pathology, Mycobacteria Research Laboratories, Colorado State University, Fort Collins, CO 80523-1682, USA

Andrew J. McBain
School of Pharmacy and Pharmaceutical Sciences, The University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom

Sébastien Rigali
Centre d'Ingénierie des Protéines, Université de Liège, Institut de Chimie B6, Sart-Tilman, B-4000 Liège, Belgium

Gilbert Shama
Department of Chemical Engineering, Loughborough University, Loughborough, Leicestershire LE11 3TU, United Kingdom

Henrieta Škovierová
Department of Microbiology, Immunology and Pathology, Mycobacteria Research Laboratories, Colorado State University, Fort Collins, CO 80523-1682, USA

Joy G. Waite
Department of Food Science and Technology, The Ohio State University, Columbus, OH 43210, USA

Ahmed E. Yousef
Department of Food Science and Technology, The Ohio State University, Columbus, OH 43210, USA

CONTENTS

CHAPTER 1

Variation in Form and Function: The Helix-Turn-Helix Regulators of the GntR Superfamily

Paul A. Hoskisson* and **Sébastien Rigali**†

Contents

Abstract

One of the most abundant and widely distributed groups of Helix-turn-helix (HTH) transcription factors is the metabolite-responsive GntR family of regulators (>8500 members in the Pfam database; Jan 2009). These proteins contain a DNA-binding HTH domain at the N terminus of the protein and an effector-binding and/or

* Strathclyde Institute of Pharmacy and Biological Science, University of Strathclyde, Royal College Building, George Street, Glasgow, United Kingdom
† Centre d'Ingénierie des Protéines, Université de Liège, Institut de Chimie B6, Sart-Tilman, B-4000 Liège, Belgium

Advances in Applied Microbiology, Volume 69
ISSN 0065-2164, DOI: 10.1016/S0065-2164(09)69001-8

oligomerisation domain at the C terminus, where upon on binding an effector molecule, a conformational change occurs in the protein which influences the DNA-binding properties of the regulator resulting in repression or activation of transcription. This review summarises what we know about the distribution, structure, function and classification of these regulators and suggests that they may have a future role in biotechnology.

...endless forms most beautiful and most wonderful have been and are being, evolved.

Charles Darwin, 1859

I. INTRODUCTION

As bacteria sense different micro-environments, they modify their gene expression appropriately to enable them to respond to the prevailing conditions. Often the signal sensed within the cell is a metabolic intermediate, and these are sensed by many classes of helix-turn-helix (HTH) transcription factor, through which they modulate gene expression. The effector molecules bound by these proteins are often related catabolic substrates, substrates and/or intermediates of the pathway controlled by the transcription factor.

One of the most abundant groups of HTH bacterial metabolite-responsive transcription factors is the GntR family of regulators (>8500 members in the Pfam database; Jan 2009). These multi-domain transcription factors are widely distributed throughout the bacterial world where they play a fundamental role in modulation of gene expression to respond appropriately to the environment context.

This review aims to bring together and summarise our current thinking on GntR regulators, their structure, function, evolution, and how they may be exploited in biotechnology.

II. HELIX-TURN-HELIX DNA-BINDING PROTEINS

The identification of a tri-helical domain and its critical role in DNA binding within the bacteriophage Lambda proteins, cI and cro and the *lac* operon repressor, LacI, were early advances in the pioneering work of Matthews and co-workers (Ohlendorf *et al.*, 1982, 1983) and Sauer *et al.* (1982). The importance of helix two and helix three of the domain led to the identification of what became known as the HTH motif. The third α-helix is often referred to as the 'recognition' helix, which fits within

the major groove of the DNA mediating the protein–DNA interaction (Aravind *et al.*, 2005). Ohlendorf *et al.* (1983) and Sauer *et al.* (1982) suggested, through extensive sequence analysis and secondary structure analysis, that this domain was present in several DNA-binding bacterial activators and repressors, and they hypothesised that these domains descended from a common ancestor.

Throughout the 1980s and 1990s, extensive sequencing, the emergence of whole genome sequencing and experimental work confirmed the ubiquity and central role this domain played in gene regulation in both prokaryotes and eukaryotes and led to the identification of the HTH motif in all domains of life, suggesting that the HTH domain is one of the most ancient protein folds, although it appears to be most prevalent in prokaryotes (Aravind and Koonin, 1999). The development of specific algorithms for recognition of the HTH motif has become indispensible in genome annotation such as that of Dodd and Egan (1990), enabling rapid identification of HTH-containing proteins.

III. GntR REGULATORS

The HTH-containing GntR family is widely distributed throughout the bacteria where they regulate many diverse biological processes. It was named GntR after the first member identified, the *Bacillus subtilis* repressor of the gluconate operon (Haydon and Guest, 1991; Prosite Family PS50949; Pfam family: PF00392). GntR regulators are often located on the chromosome adjacent to the genes that they control, which in many cases allows insight into the metabolites that they may bind. There are however many examples where this is not the case, and identifying their cognate ligands remains a significant barrier to understanding their function.

In general, these proteins contain a DNA-binding HTH domain at the N terminus of the protein and an effector-binding and/or oligomerisation domain at the C terminus (Fig. 1.1). Upon binding an effector molecule at the C-terminal domain, a conformational change occurs in the protein which influences the DNA-binding properties of the regulator resulting in repression or activation of transcription. The DNA-binding domain is conserved throughout the GntR family yet the regions outside the DNA-binding domain are more variable; however, this is not surprising given the diversity of molecules that they bind, and this feature is used to define the GntR-like sub-families (Rigali *et al.*, 2002). Despite the large number of GntR-like regulators identified there are few examples where their effector molecules are known and the complete regulatory circuitry elucidated. Knowledge of this is of particular importance where GntR-like regulators control genes of unknown biochemical function and can

FIGURE 1.1 Schematic representation of a GntR protein. Indicates the N-terminal helix-turn-helix DNA-binding domain and the longer C-terminal effector-binding/oligomerisation domain (E-b/O).

provide information of their cellular function and will enable these processes to be built into modelling frameworks in terms of using systems biology approaches. GntR-like regulators are known to control many fundamental cellular processes such as motility (Jaques and McCarter, 2006), development (Hoskisson *et al.*, 2006), antibiotic production (Hillerich and Westpheling, 2006), antibiotic resistance (Truong-Bolduc and Hooper, 2007), Plasmid transfer (Reuther *et al.*, 2006) and virulence (Casali *et al.*, 2006; Haine *et al.*, 2005). In all these cases the exact ligand regulating gene expression through these proteins is unknown.

There are many cases where GntR-like regulators are not located next to genes that they control (orphan regulators), or without their effectors they are activators of gene expression elsewhere in the genome. One well-studied example is FadR, the fatty acid metabolism regulator in *Escherichia coli*, where it is known to negatively control 12 genes or operons and activate transcription of at least three genes when a fatty acid precursor is bound (DiRusso *et al.*, 1993; See section VIII).

The identification of the small molecules that bind to these regulators has traditionally been difficult and has mainly relied on gene context and bioinformatics to identify possible effector molecules. This area remains a significant challenge to researchers in this field and urgently requires novel methods to aid identification of effector molecules.

IV. DISTRIBUTION OF GntR REGULATORS

Examination and analysis of GntR regulator distribution throughout completely sequenced genomes demonstrate some interesting trends in terms of their abundance and may give clues to how an organism is distributed in a particular ecological niche or the kind of plasticity it experiences within its natural environment.

There are 8561 GntR regulators in the Pfam database (PfAM family GntR: PF00392: Finn *et al.*, 2008). The bulk of these (8561 sequences) are found in 764 bacterial taxa indicating that this protein fold has been widely adopted as a regulatory mechanism. Examination of taxonomic distribution of these regulators throughout the bacteria demonstrates a wide distribution; however, the predominant phyla (from current

sequences available in Pfam) are the Proteobacteria, Firmicutes and the Actinobacteria (Fig. 1.2). Detailed examination of the distribution within well-characterised species (Fig. 1.3) shows an interesting trend, not only with increasing genome size, but also with ecological niche. The trend suggests that organisms that live in complex, highly variable environments such as soil (e.g. *Streptomyces*, *Burkholderia*, *Rhizobium*) have a larger complement of the metabolite-responsive GntR regulators than obligate intracellular parasites and endosymbionts (e.g. *Chlamydia* and *Buchnera*). This trend is reinforced even within genera with *Mycobacterium smegmatis* having a complement of about 60 GntR regulators (Vindal *et al.*, 2007), where all these have been lost in the obligate intracellular pathogen *Mycobacterium leprae* during the extensive gene decay observed in this species (Cole *et al.*, 2001). These data indicate that, whilst there is a trend to increase metabolite-responsive regulators in the genome to enable rapid response to changing conditions in complex environments, this is lost when a stable niche is occupied and this requirement ameliorated.

There are twelve GntR regulators known in the Archaea. Two known from Eukaryotes, one from the sea anemone (*Nematostella vectensis*) and one from *Trichomonas vaginalis*; however, the exact functions of these are unknown. The two known GntRs in viruses are both in bacteriophages,

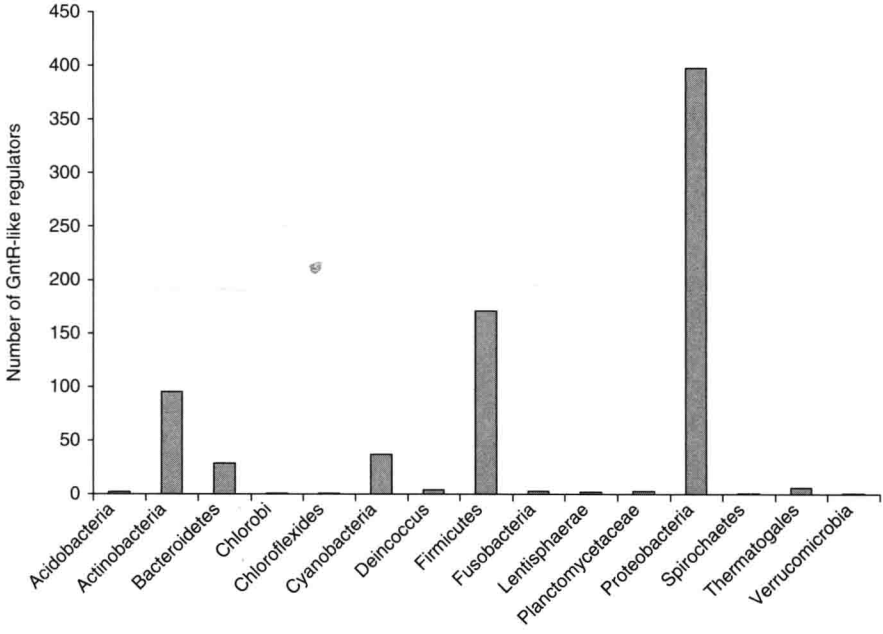

FIGURE 1.2 Distribution of GntR proteins throughout the bacterial Phyla. Please see text in Section IV. Data were taken from the sequences deposited in the Pfam database.

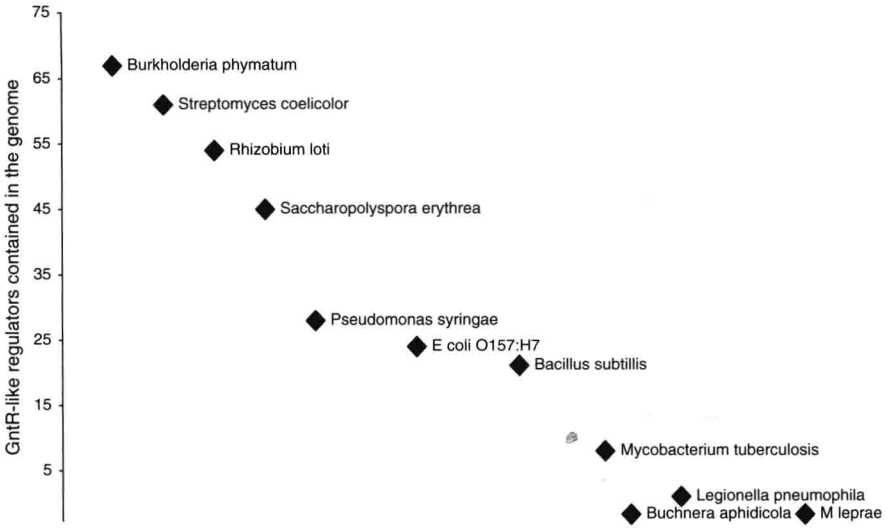

FIGURE 1.3 Distribution of GntR regulators in selected bacterial whole genomes. See text in Section V.

one in the *Streptomyces* phage ϕC31 and one in an enterophage ϕp27. Whilst the function of these within the bacteriophages is unknown, it is likely that they have been acquired from host strains.

V. STRUCTURE AND CLASSIFICATION OF GntR REGULATORS

Haydon and Guest (1991) first described the GntR family based on a common sequence at the N terminus of the proteins. They showed that a highly conserved 69-amino acid N-terminal region, containing a predicted HTH motif was conserved (Fig. 1.4). Further analysis of the domain using Pfam has indicated that the HTH domain can be refined to an average of 62.2 amino acids within the GntR domain (Finn *et al.*, 2008). Whilst overall sequence identity in the N-terminal HTH domain is low, the prediction of secondary structure is highly conserved with the three α-helices, characteristic of the HTH domain being apparent (Fig. 1.4). Despite the abundance of GntR sequences in the databases there are few crystal structures available to fully examine structure/function relationships at a detailed level.

Haydon and Guest (1991) also noted that there was extensive variation in the C-terminal domain suggesting heterogeneity in the effector molecules that they bind. Interestingly analysis of all full-length GntR-like sequences in the Pfam database indicates that on average the N-terminal

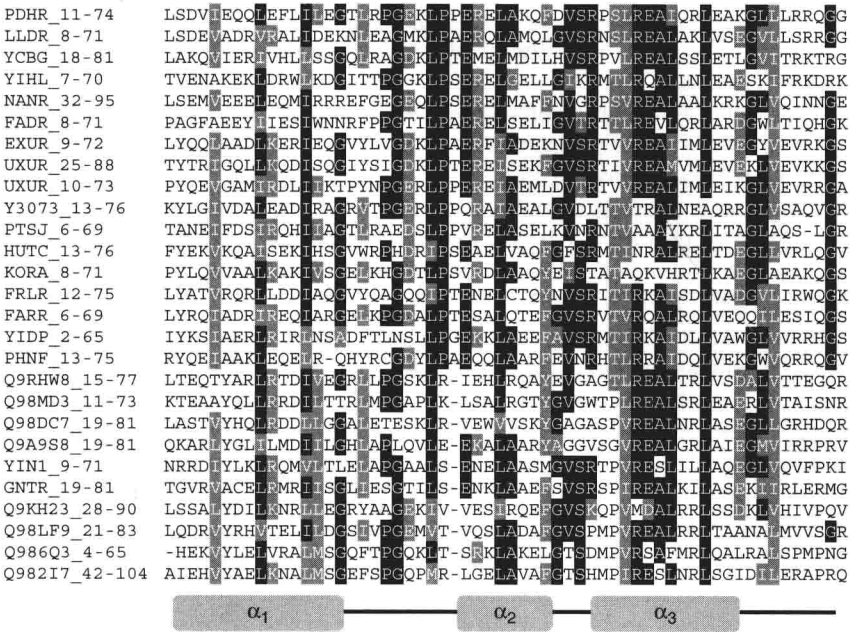

FIGURE 1.4 Alignment of the HTH domain of GntR regulators. This alignment demonstrates conservation of sequence structure within the HTH domain. Alignment was performed using the Seed alignment from Pfam (Finn *et al.*, 2008) used to generate the GntR family hidden Markov Model; the alignment was performed in ClustalW (Larkin *et al.*, 2007) with the residues coloured using Boxshade (http://www.ch.embnet.org/software/BOX_form.html). Secondary structural predictions were performed using the ProteinPredict webserver (Rost *et al.*, 2004) and were checked against the 3D structure of FadR.

domain accounts for 21.8% of the amino acid sequence, yet homology across the whole protein is around 30% reinforcing the concept of extensive C-terminal heterogeneity.

Despite only limited knowledge of two operator sequences at the time, Haydon and Guest (1991) noted that the recognition sequence for the GntR and HutC regulators was identical at seven of the eleven residues. These observations indicated that there are three interacting components of the GntR regulator: the DNA-binding domain, the so-called effector-binding and/or oligomerisation domain (E-b/O) and the *cis*-acting operator sequence. Rigali *et al.* (2002) exploited this idea through extensive analysis of the C-terminal domain of 270 GntR sequences, the N-terminal DNA-binding domain and the operator site. This led to the first extensive work on the family after its initial designation by Haydon and Guest (1991) and the formation of four sub-families within the GntR regulators,

and subsequent work led to the designation of a further three: the AraR, DevA and PlmA sub-families (Franco *et al.*, 2006; Hoskisson *et al.*, 2006; Lee *et al.*, 2003).

Due to the abundance of sequences in the databases, the sub-division of HTH DNA-binding proteins has become an issue in bioinformatics with regard to informing genome annotators. There are relatively few studies where extensive analysis of a protein and its operators has been attempted to inform on protein function (Brown *et al.*, 2003; Busenlehner *et al.*, 2003; Korner *et al.*, 2003; Maddocks and Oyston, 2008; Molina-Henares *et al.*, 2006; Weickert and Adhya, 1992).

The classification of the GntR sub-families was based on alignment of the C terminus and secondary structural predictions to reveal distinct structural topologies within each sub-family (Rigali *et al.*, 2002). Extensive analysis was verified by comparison of the predicted topology with that of the known crystal structure of FadR (van Aalten *et al.*, 2000a) which added confidence to the findings. Comparison of the predicted secondary structures is shown in Fig. 1.5 and illustrates the diversity and length differences in each sub-family described so far.

The most abundant GntR sub-family is FadR, named after the fatty acid biosynthesis and degradation regulator of the same name. This sub-family accounts for approximately 40% of GntR regulators, with the C terminus averaging 160 amino acids and consisting of six or seven α-helices. The crystal structure of the C-terminal of FadR of *Escherichia coli* is known (van Aalten *et al.*, 2000a,b, 2001) and served as a validation of the secondary structural predictions of Rigali *et al.* (2002). FadR is an acyl-CoA-responsive member of the GntR family (van Aalten *et al.*, 2001). The regulator exhibits an unusual protein fold overall. The winged helix-turn-helix (wHTH) is fused to a seven α-helix bundle which has crossover topology, containing a large internal cavity required for binding acyl-CoA (Raman and DiRusso, 1995; van Aalten *et al.*, 2000a,b). One unclear aspect of this study was how an effector-binding domain, located 30 Å from the DNA-binding domain, can affect transcription. van Aalten *et al.* (2001) elucidated this through showing, that upon binding acyl-CoA, the protein backbone undergoes dramatic conformational shifts, which results in a 7.2-Å movement of the DNA recognition helix preventing DNA binding and subsequent transcriptional repression.

Recently a cluster within the FadR sub-family was identified that appears to have evolved in Gram-positive organisms for citrate utilisation, with the authors inferring that this lineage arose through E-b/O domain replacement in an ancestral protein (Blancato *et al.*, 2008).

The second sub-family is the HutC grouping, which represents a highly diverse family in terms of effector molecules and processes regulated and accounts for around 30% of GntR regulators. The HutC sub-family C terminus consists of α-helices and β-sheets and averages 170