2007

Comparative functional analysis and identification of regulatory control in gene networks using the leucine-responsive regulatory protein and its regulon as a model system

Robert E. Lintner
Medical University of Ohio

Follow this and additional works at: http://utdr.utoledo.edu/theses-dissertations

Recommended Citation
http://utdr.utoledo.edu/theses-dissertations/1304
Health Science Campus

**FINAL APPROVAL OF DISSERTATION**

Doctor of Philosophy in Biomedical Sciences

Comparative Functional Analysis and Identification of Regulatory Control in Gene Networks Using the Leucine-Responsive Regulatory Protein and its Regulon as a Model System

Submitted by:
Robert E. Lintner

In partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biomedical Sciences

**Examination Committee**

**Major Advisor:** Robert Blumenthal, Ph.D.

**Academic Advisory Committee:**
- Eric Lafontaine, Ph.D.
- R. Mark Wooten, Ph.D.
- Alexei Fedorov, Ph.D.
- Kandace Williams, Ph.D.

**Senior Associate Dean**
College of Graduate Studies
Michael S. Bisesi, Ph.D.

**Date of Defense:** April 17, 2007
Comparative Functional Analysis and Identification of Regulatory Control in Gene Networks Using the Leucine–Responsive Regulatory Protein and its Regulon as a Model System

Doctor of Philosophy

Robert E. Lintner

The University of Toledo College of Medicine

2007
DEDICATION

To my wife, Tracy, whose devotion and selfless support made this possible, and to my sons, Ethan and Cole, for filling each day with more happiness than any man deserves.
ACKNOWLEDGMENTS

I thank Dr. Blumenthal for providing the opportunity to work and learn under his guidance. I am grateful for his patience and for his diligence throughout the completion of this work.

I thank Drs. Eric LaFontaine, Alexei Fedorov, Kandace Williams and Mark Wooten for serving as my advisory committee. Thank you for your encouragement, and for the time you spent away from your own work to provide a critical evaluation of mine.

I thank Dr. Pankaj Mishra for his completion of experiments for manuscript I. His important contribution made Figures 2, 4 and 9 possible.

I thank Drs. Arkady Khodursky and Poonam Srivastava for the completion of the microarray analysis presented in manuscript I.

I thank Dr. Elmar Weinhold for providing the SMILing DNA reagents and the protocol used to complete the work presented in manuscript II.
I thank Dr. Venkatesha Basrur for performing the mass spectrometry analysis reported in manuscript II.

I thank the State of Ohio and The University of Toledo for the Pre-doctoral fellowship and tuition waiver provided for me during my five years of study, and the U.S. National Institutes of Health for partial stipend support through a research grant to Dr. Blumenthal.

I also thank the Department of Medical Microbiology and Immunology for providing partial financial assistance.

I thank the department secretaries, Joyce Rodebaugh, Sue Payne and Sharon Ellard for always being available to help with almost anything.

Finally, I thank the members, both past and present, of the Blumenthal, LaFontaine, Wooten and Worth labs for their aid and companionship over the past five years.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedication</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Literature</td>
<td>3</td>
</tr>
<tr>
<td>Manuscript I</td>
<td>20</td>
</tr>
<tr>
<td>Manuscript II</td>
<td>95</td>
</tr>
<tr>
<td>Discussion</td>
<td>150</td>
</tr>
<tr>
<td>Bibliography</td>
<td>154</td>
</tr>
<tr>
<td>Abstract</td>
<td>202</td>
</tr>
</tbody>
</table>
INTRODUCTION

The relative rate at which genome sequences are becoming available offers a unique opportunity for understanding the physiology behind many organisms. In addition to genetic capability a genome sequence may also provide gene regulatory information. In bacteria, genes are controlled primarily at the level of transcription initiation (Lozada-Chavez et al., 2006). Proteins known as transcription factors recognize and bind DNA sequences that are positioned upstream of genes. These regulatory sequences function as the relay between the signal (input) and the response (output), and have been deemed a major contributor of phenotypic variation among genetically similar organisms (Carroll, 2000). Computational efforts aimed at predicting the regulatory networks within a cell are most successful when governed by parameters derived from experimental studies of gene expression and protein binding. The co-development of robust and reliable methods for experimentally and computationally describing gene control is necessary if comparative analyses are to be used in predicting the regulatory patterns of genes based solely on genome sequence.

Regulation of the leucine-responsive regulatory protein (Lrp), and the expression levels of the genes it directly and indirectly controls have been described (Hung et al., 2002; Landgraf et al., 1996; Tani et al., 2002). The combinatorial control for one of the Lrp target genes, gltB, has also been experimentally determined (Paul
et al., 2007). Using Lrp as a model system this dissertation describes the results of tests to determine the validity of assumptions underlying bioinformatic prediction, and reports work on the development of a novel method for identifying unknown regulatory proteins. This work is presented in two manuscripts:

**Manuscript I.** Incomplete Functional Conservation of a Global Regulator Across Related Bacterial Genera: Lrp in Escherichia, Proteus and Vibrio

This manuscript was prepared for and submitted to Nucleic Acids Research. The regulatory extrapolation hypothesis was described and tested, and the regulatory patterns of Lrp in two pathogens, *Proteus mirabilis* and *Vibrio cholerae*, were described for the first time.

**Manuscript II.** A Novel Method for Isolating and Identifying Promoter-Associated Proteins

This manuscript was prepared for Molecular Microbiology. A novel method for the isolation of unknown regulatory proteins crosslinked to promoter DNA *in vivo* is described. The method offers an advantage over conventional methods such as chromatin immunoprecipitation (ChIP) by not requiring the identity of a regulator to be known. Three different isolation approaches were developed and preliminary results for their efficacy are detailed.
LITERATURE

Control of gene expression

RNA polymerase

Discovered in the late 1950s and early 1960s as an essential component in different organisms, RNA polymerase was quickly recognized as being the workhorse of transcription (Huang et al., 1960; Hurwitz et al., 1960; Stevens, 1960; Weiss and Gladstone, 1959). Early biochemical (Burgess et al., 1969) and more recent structural analyses (Busby and Ebright, 1994; Chen et al., 1994; Igarashi and Ishihama, 1991; Kimura and Ishihama, 1995; Ross et al., 1993) of this multi-subunit transcriptase have revealed several key features of the enzyme that are critical to the control of gene expression.

The functions of RNA polymerase are mediated by two distinct multimeric forms: the core enzyme and the holoenzyme (Burgess et al., 1969; Gross et al., 1996). The core enzyme (E) is made up of 5 subunits (α2, β, β′, ω), and is important for transcript elongation and termination. The holoenzyme (Eσ) consists of E in addition to, in E. coli, one of seven possible sigma (σ) factors (Gross et al., 1992). Association of E with a sigma factor is essential for promoter recognition, binding and transcription initiation (Ades et al., 1999). The complex relationship between the relative levels of the different sigma factors and their competition for limiting amounts of RNA polymerase (Bedwell and Nomura, 1986; Nomura et al., 1987; Zhang et al., 2002) is an integral first step in the mechanism behind the activation and repression of gene sets in response to different
environmental conditions (Jenkins et al., 1991; Jishage and Ishihama, 1995; Jishage et al., 1996; Typas et al., 2007). For example, in *Bacillus subtilis* sporulation, a process incited by the alternative $\sigma^H$, is delayed when the primary $\sigma^A$ is overproduced and levels of E are limiting (Hicks and Grossman, 1996).

**Activator-dependent gene expression**

Controlling the frequency of binding and transcription initiation by RNA polymerase (RNAP) is the basic means by which genes are regulated (Record et al., 1996). Activators are proteins that increase transcriptional activity typically by enhancing initiation via the recruitment, isomerization or clearance steps (Shinar et al., 2006).

Activators that function in the recruitment and/or stabilization of RNA polymerase (E$\sigma$) do so by making key co-operative contacts with specific regions of the holoenzyme (Ptashne and Gann, 1997). Two examples of this are the *E. coli* the catabolite repressor protein (CRP) and the bacteriophage $\lambda$ cl activator proteins. CRP recruits RNAP to $P_{lac}$ through direct contacts with the C-terminal domain of the $\alpha$ subunit ($\alpha$CTD), while $\lambda$ cl functions primarily at the isomerization step, and does so by stabilizing DNA-$\sigma$ subunit contacts that are thought to prevent the disassociation of $\sigma$ with the $-35$ hexamer at $P_{RM}$ (Busby and Ebright, 1994; Jeon et al., 1997; Kim and Hu, 1997; Kolb et al., 1993; Li et al., 1994; Li et al., 1997; Zhou et al., 1993). Contacts with other regions of RNAP (e.g. $\beta$, $\beta'$ and $\alpha$NTD) have also been documented (Busby and Ebright, 1997; Miller et al., 1997; Niu et al., 1996; Rhodius and Busby, 1998), but have not been
studied in as much detail. The $\lambda$ cl-$\sigma$ interaction described above is thought to occur through contacts with a flexible linker region associated with the $\alpha$NTD (Dove, et al., 2000).

Binding of the activator to a specific DNA sequence in certain proximity of the –10 and –35 hexamers is crucial if productive contacts are to be made with the subunits of RNAP (Gross et al., 1992). Positive regulators that interact with the $\alpha$CTD typically bind upstream of the –35 at sequences known as UP elements, and these elements can be several bases away due in part to the flexible linker region that connects the $\alpha$CTD to RNAP (Jeon et al., 1997). Activators that contact $\sigma$ subunits usually bind very close to or even overlap the –35, and this is likely due to the fact that there is little flexibility between the $\sigma$ subunit and the $\beta$-$\beta'$ complex (Li et al., 1994; Li et al., 1997).

Not all activators initiate transcription by directly interacting with RNAP. Regulators such as MerR (mercury response regulator) promote initiation by twisting the DNA so as to reposition the –10 and –35 elements relative to one another. By binding between the two hexamers, MerR, facilitates an underwinding of the promoter DNA that allows RNAP to begin transcription (Ansari et al., 1995; Summers, 1992). Other activators play architectural roles, including integration host factor (IHF) and the factor for inversion stimulation (FIS); these function by binding DNA upstream of the –35 hexamer, introducing bends in the DNA that promote contacts between RNAP and other regulators (Danot et al., 1996; Richet et al., 1991), or that aid initiation (Muskhelishvili et al., 1997).
Co-dependence on more than one activator has been observed for a number of genes in *E. coli* (Collado-Vides *et al.*, 1991), and the results from different studies support two basic mechanisms, while leaving room for a third alternative (Hochschild and Joung, 1997). For promoters like that of the *E. coli* asparaginase II (*ansB*), different activators (FNR and CRP) are required because of the independent contacts they each make with the subunits of RNAP (Scott *et al.*, 1995). For other promoters, such as P_{malK}, two regulators (CRP and MalT) are necessary as one repositions the other to facilitate interaction with RNAP (Richet *et al.*, 1991; Richet and Sogaard-Andersen, 1994). The third mechanism, which has yet to be observed in bacteria, presumes a co-operative relationship between two activators that is necessary for one of the regulators to bind a specific target DNA sequence or to interact with a subunit of RNAP (Hochschild and Joung, 1997).

**Repressor-dependent gene expression**

The earliest published models for gene expression involved repressor controlled-systems (Jacob and Monod, 1961; Monod *et al.*, 1963). Probably the most popular and best studied of these is the *lac* operon (Gilbert and Muller-Hill, 1966; Jobe and Bourgeois, 1972; Riggs and Bourgeois, 1968). Used initially by Jacob and Monod as a model to understand the coordinate regulation involved when *E. coli* switches from glucose to lactose for carbon utilization (Jacob and Monod, 1961), the *lac* operon quickly became a “paradigm” for the molecular mechanisms behind control of gene expression by repression. The basic idea
behind repression is that a protein binds and prevents transcription initiation either by inhibiting the binding of RNA polymerase or by interfering with the actual steps of initiation. In the case of the lac operon, binding of the repressor (LacI) to its operator sequences (Gilbert and Maxam, 1973; Riggs et al., 1970) prevents transcription by: 1) inhibiting binding of Eσ, 2) inducing pausing and preventing escape by E, and 3) preventing formation of closed complex 1 (Lee and Goldfarb, 1991; Schlax et al., 1995; Straney and Crothers, 1987).

The action of many repressor proteins relies on their DNA binding capacity such that changes in their affinity for operator site(s) function as an integral part of the repression regulatory mechanism (Wilson et al., 2007). Topological features of promoter DNA, such as looping and supercoiling, are important for the formation and stability of repression complexes (Becker et al., 2005; Brenowitz et al., 1991; Kramer et al., 1988; Mossing and Record, 1986; Reznikoff et al., 1974; Whitson et al., 1987). Loss of affinity for DNA is usually controlled allosterically; binding of a small molecule leads to conformational changes in the repressor that no longer favor its interaction with operator DNA resulting in the relief of repression (Matthews, 1974; Oshima et al., 1972). Two well understood examples of regulatory control by allostery are LacI + 1,6-allolactose (Jobe and Bourgeois, 1972), and TetR + tetracycline (Ramos et al., 2005).

Interaction of activators and repressors

While it is helpful to categorize activator and repressor action for the sake of description, the reality is that initiation from some promoters under certain
conditions is co-dependent on both negative and positive control (Paul et al., 2007; Wu et al., 1998). One example of this control involves the *E. coli nir* promoter. Under conditions where oxygen is limiting and nitrite ions are present the operon coding for a cytoplasmic NADH-dependent nitrite reductase (*nir*) is activated through the integration of a positive signal from the fumarate and nitrate reduction regulatory protein (FNR), a negative (suppressive) signal from the factor for inversion stimulation (FIS) and a positive counter signal from the nitrate response regulator (NarL) (Harborne et al., 1992; Jayaraman et al., 1987; Page et al., 1990; Peakman et al., 1990). In another example, the promoter for glutamate synthase (*gltBD*) from *E. coli* receives signals from three global regulators, the leucine-responsive regulatory protein (positive), the integration host factor (positive) and CRP (negative) in addition to a negative signal from the arginine repressor (Paul et al., 2007).

In yeasts and other higher eukaryotes (Ogata et al., 2003; Pilpel et al., 2001; Robin et al., 2002; van Helden et al., 2000) the cooperative activity of transcriptional activators/repressors bound to their intricately spaced binding sites form key protein-DNA-protein interactions known as enhanceosomes (Carey, 1998; Merika and Thanos, 2001; Remenyi et al., 2004). As indicated by the definition, the formation of these multi-protein networks is dependent on the cooperative binding of activators/repressors such that the assembly is highly sensitive to the critical concentrations of the various regulators (Ogata et al., 2003). In addition, structural proteins that distort DNA by bending, twisting and looping are also necessary as they promote important protein-protein contacts.
(GuhaThakurta and Stormo, 2001). Another essential but poorly understood property of these complexes is the reciprocity between their formation and the recruitment of the transcriptional machinery (Makeev et al., 2003), a relationship that is thought to increase specificity while functioning synergistically to initiate transcription (Bulyk et al., 2004; Carey, 1998; Klingenhoff et al., 1999; Robin et al., 2002).

Examples such as these provide a glimpse into the complex nature of transcription initiation by highlighting the proclivity, even in single celled organisms with relatively small genomes, for integrating multiple inputs from global and local regulatory sources.

**Regulatory networks**

**Basic structure**

The idea of “structure” emerging from the seemingly stochastic processes controlling transcription (see above) hinges on the following principle: that genes are co-regulated such that physiology represents the sum of the integration of expression of pertinent genes (Akutsu et al., 2000; Farr and Kogoma, 1991; Hartemink, 2005; Hughes et al., 2000; Levine and Davidson, 2005; Schumann, 2003; Spector, 1998; Weber and Marahiel, 2003). Viewed in this light, the structural component of regulatory networks stems from how they are connected (Oosawa and Savageau, 2002). Defining connectivity within regulatory networks is a somewhat difficult process that usually involves combining experimental data
(e.g., microarray and chromatin immunoprecipitation) with mathematical models (Balaji et al., 2006; Barrett et al., 2005; Holter et al., 2000; Liao et al., 2003; Yeung et al., 2002). A central feature of experimental and computational efforts to decipher regulatory connections is the DNA binding transcription factor (regulator) and the set of genes (regulon) it controls (Babu et al., 2004; Balaji et al., 2006; Barrett et al., 2005; Hung et al., 2002; Kelley, 2006; Tani et al., 2002). Placing groups of non-contiguous genes under the control of a single regulator (functional definition of a regulon) is an efficient means by which cells can respond to a particular signal, and the successful coordination of these singular co-regulatory responses (stimulon) enhances the organisms ability to adapt to a changing environment (Cases and de Lorenzo, 2005).

The “regulon structure” of an organism is dependent upon its niche (Babu et al., 2004; Madan Babu et al., 2006), and the number of transcription factors in its repertoire can be directly correlated to the size of its genome (Levine and Tjian, 2003; van Nimwegen, 2003) with the quantity of the various structural types (e.g. zinc finger or helix-turn-helix) being restricted by the number of bases in their DNA recognition sequence (Itzkovitz et al., 2006). In addition to structural classification (Aravind et al., 2005; Luscombe et al., 2000; Pabo and Sauer, 1992), regulators can also be grouped based on their regulatory scope (i.e. the number and types of genes they control) to help define their role in a gene network (Cases and de Lorenzo, 2005; Gottesman, 1984).
Defining regulon structure

Studies in *E. coli* have revealed that a small number of global regulators (the current consensus is: Lrp, CRP, IHF, FIS, FNR, ArcA, NarL) are responsible for influencing the expression of more than half of the genome, while nearly a fifth of the other regulators only control one or two genes each (Madan Babu and Teichmann, 2003; Shen-Orr *et al.*, 2002). The structure of gene networks is thus hierarchical and follows a power-law distribution (Oosawa and Savageau, 2002; Yu and Gerstein, 2006). Modeling of regulatory structures is therefore amenable to strategies based on other less ambiguous relationships such as social interactions (Guido *et al.*, 2006; Lee *et al.*, 2002). Using statistically overrepresented local network patterns (motifs) to categorize simple interactions a complex gene regulatory apparatus can be theoretically defined and subsequently applied to experimental data for validation (Brynildsen *et al.*, 2006; Yu and Gerstein, 2006). Much of the information concerning regulatory network structure comes from the decades of experimental analyses performed in *E. coli* and *S. cerevisiae* (Bar-Joseph *et al.*, 2003; Barrett *et al.*, 2005; Milo *et al.*, 2002). Web servers and large databases, such as RegulonDB (Salgado *et al.*, 2006), PRODORIC (Munch *et al.*, 2003), Tractor_DB (Gonzalez *et al.*, 2005), TRANSFAC (Matys *et al.*, 2006) and BioProspector (Liu *et al.*, 2001), contain information about DNA binding sites and offer web based search algorithms for determining network connections (based on model organisms) from genome sequence (discussed in manuscript I).
Predicting regulator binding sites

At the core of regulatory element prediction is the discovery of DNA sequence motifs recognized and bound by transcription factors. Prediction of DNA motifs is basically a statistical process (Li and Tompa, 2006; Tompa et al., 2005). Algorithms are designed to scan large regions of regulatory DNA sequence (~1000 bp) and identify relatively short (Moses et al., 2003) overrepresented motifs (Burge and Karlin, 1997; Pevzner and Sze, 2000). The process is inherently noisy, and additionally complicated by the degeneracy in transcription factor binding sites. For example, in the first six *E. coli* -10 regions cataloged only two out of six positions are conserved (Pribnow, 1975), and in the 12 half sites recognized by the lambda repressor only two of eight positions are conserved (Maniatis et al., 1975). Despite this intrinsic variability regulatory elements do appear to evolve at a much slower rate in comparison to the surrounding sequence in an intergenic region (Moses et al., 2003), which supports the basic assumptions behind the motif approach to identifying binding sites.

Motifs are typically identified *de novo*, or by incorporating biological constraints such as gene expression levels and binding properties from ChIP analyses. For both approaches, the binding characteristics (sequence) of a motif can be represented in computation as a consensus sequence or as a position weight matrix. The use of consensus sequences is popular because of its conceptual simplicity (Day and McMorris, 1992). Several algorithms have been designed for consensus sequence building (Cavener, 1987; Choo et al., 1991;
Daniels and Deininger, 1991), and each differs in strength when judged by their
sensitivity and precision, both of which are affected by the allowable mismatches
and the overall ambiguity of the binding specificity (Day and McMorris, 1992).
The advent of whole-genome expression analysis has improved the results
yielded by consensus string matching methods by providing large sets of
coregulated genes whose mRNA levels can be modeled using linear regression
to determine their association with particular motifs (Brazma et al., 1998;
Bussemaker et al., 2001; Keles et al., 2002; van Helden et al., 1998).

Position weight matrices (PWM) [also called position-specific weight
matrices (PSWM) or position specific scoring matrices (PSSM)] were first
employed, biologically, in identifying translation initiation sites (Stormo et al.,
1982b), and are more flexible than consensus methods in that they score all
significant regions within a sequence rather than simply returning only the best
matches. In addition there is a strong correlation between the score of a
sequence identified by a PWM and its functional activity as a promoter (Mulligan
et al., 1984). A fundamental difficulty lies in choosing the correct weights for a
PWM, the success of which is key in distinguishing between true sites and
non–sites. (Stormo, 2000). Several mathematical approaches have been
successfully employed in assigning weights to matrices. For example, to
distinguish the true translation initiation sites in *E. coli*, the Stormo group trained
a machine learning algorithm (perceptron) using experimentally-confirmed sites
and non-sites, and in doing so they were able to develop a “threshold” that
allowed for the distinction between the two groups (Stormo et al., 1982a). Other
statistically based methods have since been developed (Harr et al., 1983; Mulligan et al., 1984; Staden, 1984). One important contribution for de novo prediction based on comparison came from the finding that the information content of a motif is directly related to its binding energy (Berg and von Hippel, 1987; Schneider et al., 1986). This conclusion allowed matrices to be weighted (i.e., true sites distinguished from non-sites) based on the probability of binding by directly discerning the binding energy of a site, and comparing it to the average for all known sites (Heumann et al., 1994; Stormo and Fields, 1998).

PWM are amenable to strategies similar to those used in consensus−based approaches for the incorporation of data from expression and binding analyses, though they seem to be more sensitive and precise. One limitation of PWM stems from the intrinsic assumption that each position within a sequence contributes equally and independently to the overall probability that a regulator will bind, but many regulators are known to make specific contacts with only a few key bases in its recognition sequence (Strzelecka et al., 1995; Wang et al., 1998; Wisedchaisri et al., 2004), and particular positions can be synergistic or antagonistic with one another (Benos et. al., 2002).

The regulatory pattern of most genes is dependent upon the integration of signals from multiple regulatory inputs (see above), so an accurate assessment of promoter regulation requires a consideration of the complete set of proteins responsible for its control (Beer and Tavazoie, 2004; Hvidsten et al., 2005; Pilpel et al., 2001). Identifying the constraints that guide the action of regulatory proteins is an essential part of predicting combinatorial control, although the
binding constraints (e.g. spacing and strand association of regulators) used are not necessarily derived experimentally. Using probabilistic models, such as Bayesian networks, the regulatory affiliations of a particular promoter are described by incorporating the binding constraints with expression analysis (Beer and Tavazoie, 2004; Ryu et al., 2007; Segal et al., 2003). One difficulty with predicting this type of complex control comes from the relatively rare nature of many of the interactions leading to a failure to detect overrepresentation within a large network (Beer and Tavazoie, 2004; Ryu et al., 2007). As more information about the binding properties of different model regulators becomes available for incorporation into algorithms, these studies may eventually offer a significant advantage in that they can systematically decipher the sum of regulatory inputs required for a specific gene expression pattern, which would probably take decades to determine experimentally.

The Lrp regulon

Brief history and description of Lrp activity

First appreciated for its involvement in the transport of branched-chain amino acids (Anderson et al., 1976), the leucine-responsive regulatory protein (Lrp) has since been implicated in affecting the expression of almost 10% of the genes in *E. coli* (Hung et al., 2002; Tani et al., 2002). Lrp’s role in metabolism was first postulated after the discovery that mutations in the *lrp* gene influenced the expression of operons coding for amino acid transport, catabolism and
biosynthesis (Platko et al., 1990; Tuan et al., 1990). The broad regulatory range of Lrp stems from its ability to self assemble into multimeric forms and its interaction (or lack of) with L-leucine and L-alanine (Chen et al., 2001; Chen and Calvo, 2002), the effects of which are discussed in manuscript I of this dissertation. More recent studies have demonstrated Lrp’s role in regulating ribosomal RNA promoters by cooperating with the heat-stable nucleoid-structuring protien (H-NS) to form a nucleoprotein complex that inhibits the interaction of FIS with RNAP leading to repression (Pul et al., 2005; Pul et al., 2007), as well as its control over complex phenotypes associated with virulence and symbiosis (Camacho and Casadesus, 2005; Cowles et al., 2007).

**Structural properties of Lrp**

Lrp and AsnC make up a family of transcriptional regulators that is well conserved among bacterial and archaeal species with orthologs in nearly 50 and 94%, respectively, of the sequenced genomes (Charlier et al., 1997; Napoli et al., 1999). The *E. coli* Lrp monomer has a molecular weight of 18.8 kDa, and contains a helix-turn-helix motif at the N-terminus (de los Rios and Perona, 2007). The C-terminus contains a “regulation of amino acid metabolism” or RAM domain (Ettema et al., 2002) that renders the protein sensitive to allosteric modulation by leucine (de los Rios and Perona, 2007; Platko and Calvo, 1993). There are between 1300 and 3200 Lrp dimers per cell (Ali Azam et al., 1999), and each dimer measures ~55 x 55 x 50 angstroms (de los Rios and Perona, 2007). About 40% of the total Lrp concentration remains unbound by DNA, and
at μM concentrations in vivo (based on E. coli grown in minimal media) Lrp self associates into hexadecamers (Chen et al., 2001; Chen and Calvo, 2002).

The crystal structures of three Lrp/AsnC homologs (Leonard et al., 2001; Thaw et al., 2006) show a symmetrical closed octameric ring (~5.2nm across) with the DNA binding motifs oriented at the periphery; a fourth archaeal ortholog (FL11 from Pyrococcus OT3) differs in that it forms a cylindrical helix containing 12 dimers per turn (Thaw et al., 2006). The E. coli Lrp octamer, which was recently solved bound to 19 bp of DNA, remains open resulting in a linear array of four dimers (de los Rios and Perona, 2007). The structure solvers have postulated, based on free energy estimations and the fact that the other Lrp structures were crystallized in the absence of DNA, that the open form of Lrp exists only when the protein is bound to DNA. Therefore, this model suggests an equilibrium between the closed and open forms of the Lrp octamer that might explain, at least in part, the numerous regulatory effects exerted by Lrp in the presence and absence of leucine (Chen and Calvo, 2002; Hung et al., 2002; Tani et al., 2002). The open form is also amenable to the proposed, but structurally un-supported, role for Lrp in wrapping DNA. In this model, also supported by work with the fim switch inversion mechanism (Kelly et al., 2006), multiple linear octamers induce a continuous supercoil that could produce a chromatin packaging effect (de los Rios and Perona, 2007). Preliminary atomic force microscopy studies are consistent with this possibility (Wiese, 1997).

The crystal structure has also suggested additional roles for the N and C termini of Lrp. In the octameric form the N terminal peptide (10 amino acids) is
positioned such that it may function as a linker to fill the large gaps between adjacent dimers (Xu et al., 1999) an interaction that may provide additional DNA contacts. In addition to directly interacting with DNA (Chen and Calvo, 2002; de los Rios and Perona, 2007), the C terminal peptide appears to bridge the high and low affinity leucine interaction domains (Marasco et al., 1994) and the DNA binding helix-turn-helix (HTH) region, which may be important for dimer assembly on the DNA (Thaw et al., 2006). The involvement of both termini in DNA binding was further substantiated by in vitro binding studies with truncated versions of Lrp (de los Rios and Perona, 2007).

**Regulation of the *E. coli lrp* gene**

Like many other regulators in *E. coli* (Thieffry et al., 1998), Lrp represses its own gene (Wang et al., 1994). The importance of negative autoregulation has been described in studies using synthetic gene circuits (Hlavacek and Savageau, 1996; Savageau, 1976), and is predicted to be essential to the stability and responsiveness of robust networks (Becskei and Serrano, 2000; Hlavacek and Savageau, 1996; Rosenfeld et al., 2002; Savageau, 1976). This oscillatory nature of *lrp* expression combined with the growth rate dependent effects (see manuscript I) may function in meeting the dichotomous coupling requirements of Lrp controlled genes (Wall et al., 2003). Regulators that control their own transcription are coupled to the expression of their target gene(s). The members (regulator and target gene) of a directly coupled gene circuit follow coordinate patterns of expression. In an uncoupled system the expression of the regulator is
constant while that of the target gene changes. Opposite patterns of expression are maintained by the regulator and target gene in an inversely coupled circuit (Hlavacek and Savageau, 1996).
INCOMPLETE FUNCTIONAL CONSERVATION OF A GLOBAL REGULATOR ACROSS RELATED BACTERIAL GENERA: LRP IN ESCHERICHIA, PROTEUS AND VIBRIO

Robert E. Lintner¹, Pankaj K. Mishra¹, Poonam Srivastava², Betsy M. Martinez-Vaz², Arkady B. Khodursky²,³ and Robert M. Blumenthal¹,⁴*

¹Department of Medical Microbiology and Immunology, University of Toledo Health Science Campus, Toledo, OH 43614-5806, USA
²Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, St. Paul, MN, USA
³Biotechnology Institute, University of Minnesota, St. Paul, MN, USA
⁴Program in Bioinformatics and Proteomics/Genomics, University of Toledo Health Science Campus, Toledo, OH 43614-5806, USA

2007
Submitted for publication in Nucleic Acids Research
ABSTRACT

Bacterial genomes are being sequenced rapidly, but few species are physiologically well characterized. Predicting regulation from genome sequences involves extrapolation from better-studied bacteria, using the hypothesis that a conserved regulator, target gene, and regulator-binding site in the target promoter imply conserved regulation. However many species being compared are ecologically and physiologically diverse, and the extrapolation hypothesis has not been well tested. In *E. coli* K-12 the leucine-responsive regulatory protein (Lrp) affects expression of ~400 genes. *Proteus mirabilis* and *Vibrio cholerae* have well-conserved *lrp* orthologs (98% and 92% identity to *E. coli* *lrp*). The functional equivalence of Lrp from these related species was assessed. Heterologous Lrp regulated *gltB*, *livK* and *lrp* transcriptional fusions in an *E. coli* background, though with significant differences. Microarray analysis of these strains revealed that the heterologous Lrp proteins regulate fewer than half of the genes affected by native Lrp. In *P. mirabilis*, heterologous Lrp restored swarming, but with pattern differences. *P. mirabilis* produced substantially more Lrp than *E. coli* or *V. cholerae* under some conditions. Lrp regulation of selected, putative target genes differed among the species. These results suggest that a cautious approach should be taken in extrapolating regulation between bacteria based on genome sequences.
INTRODUCTION

Microbial genomes are being sequenced with increasing frequency and speed. As of 2006, about 300 bacterial genomes have been fully sequenced, and nearly 1000 more such projects are underway (1); this does not include current and planned large-scale metagenomic projects (2,3). A major motivation for this sequencing avalanche is the possibility of learning about a bacterium’s physiology or pathogenesis, without resorting to either labor-intensive classical analyses or the still-expensive and immature tools of systems biology (4-6). Increasingly effective methods are available to generate a “parts list” of genes and pathways from genome sequences of poorly-characterized bacteria (7,8). However understanding the physiology of an organism, in terms of gene regulatory mechanisms and network connections, is currently much more difficult to achieve from sequence analysis alone.

Predicting gene regulatory networks via purely bioinformatic methods is typically accomplished in a three-step process, that involves extrapolating from a well-characterized reference organism such as *E. coli* (9,10). The first two steps involve determining whether valid orthologs for a target (regulated) gene and its regulator (in the reference organism) are present in both organisms. While there is debate over the minimum relatedness orthologs must display for functional and regulatory predictions to be meaningful, these steps are relatively straightforward. The third step, identifying a putative binding site for the regulator upstream of the orthologous target gene, is more complex (11-16). Among other problems, many regulators have degenerate binding motifs and commonly-used
algorithms (consensus string matching and position-specific scoring matrices) often identify many false positives and fail to identify known targets. In addition, the relative strength of a binding site can be as important to the resultant regulatory pattern as the site’s existence (17). Binding strength is difficult to predict from sequence alone (18,19), especially where the structure of the regulator is unknown (20,21).

An even more basic issue involves the overall validity of the underlying hypothesis, which for brevity we refer to as the “regulatory extrapolation hypothesis.” Even if computational methods advance to the point where orthologous genes and binding sites could be identified unambiguously, to what extent would a matching regulator, target gene and binding site correctly predict regulation? Regulatory patterns can be profoundly changed by limited mutation. For example, a single nucleotide change in the soxS promoter results in repression by SoxR, which normally activates, regardless of the redox signal (22), and 1-2 amino acid replacements in a regulatory protein can broaden the range of coactivators or change the effect of an inducer to that of a corepressor (23,24). Some of the regulatory extrapolations involve species that are, ecologically at least, quite different from one another, and (not surprisingly) the environment to which an organism is adapted affects its regulatory architecture (25). In addition to more reliable computational approaches, there is a profound need for better experimentally-based understanding of the extent to which regulatory architecture is conserved among related bacteria adapted to different environments.
The aim of this study is to address this need for assessment by studying a model regulator, and the network of target genes it controls (its “regulon”), in three related bacteria with fully-sequenced genomes. *E. coli* K-12 is the well-studied reference organism (26). *Proteus mirabilis* is, like *E. coli*, among the Enterobacteriaceae, but is a relatively distant member of that family. *Vibrio cholerae* is a member of a different, though related, family – the Vibrionaceae. All three grow on mucosal epithelia. *E. coli* is adapted to growth in the mammalian or avian intestine, *P. mirabilis* is a urinary tract pathogen, and *V. cholerae* is primarily a marine microbe that is an opportunistic pathogen of the human ileum (27). All three organisms are capable of differentiating into elongated, hyperflagellated swarmer cells that spread across solid surfaces (28-32).

The model regulator we are studying is Lrp, the leucine-responsive regulatory protein (33-35). Lrp is highly conserved among Enterobacteriaceae and Vibrionaceae (Fig. 1, sequences above the dotted lines). The Lrp ortholog in *P. mirabilis* differs from that in *E. coli* by only 4/164 amino acids (98% identity), while Lrp from *V. cholerae* shows 92% identity to *E. coli* Lrp. Importantly, none of the changes observed in *P. mirabilis* and *V. cholerae* occur in the helix-turn-helix motif responsible for DNA sequence recognition (Fig. 1, cartoon representation and boxed region), defined via mutation of the *E. coli lrp* gene and x-ray crystallography of an archaeal ortholog (36,37) and recently of the *E. coli* protein itself (38).
The Lrp regulon has been extensively mapped in *E. coli* by microarray analysis, two-dimensional gel electrophoresis, and *lacZ* fusion libraries (39-43). The microarray analysis revealed that Lrp influences the expression of nearly 400 genes, at least 70 of which are directly controlled (A. Khodursky, unpubl. obs.). The RegulonDB database (10,44) recognizes 57 genes as being directly controlled by Lrp, based on literature surveys.

The Lrp regulon is a good model for comparison among species for at least two reasons. First, the regulon is large and includes genes having a range of functions (including biosynthesis, catabolism, transport and virulence). Second, Lrp can generate diverse regulatory patterns that include both activation and repression, and differing sensitivities to the coregulator L-leucine (or L-alanine) (45,46).

To compare the architecture of the Lrp regulons of *E. coli*, *P. mirabilis*, and *V. cholerae*, we began by addressing four general questions. First, are the effects of *lrp* disruption on growth comparable in the three species? Second, are the *lrp* genes functionally interchangeable in complementation assays? Third, do the *lrp* genes themselves have the same expression pattern in the different species? Fourth, are the orthologs of what, in *E. coli K-12*, are Lrp-controlled genes regulated by Lrp in the same manner in the two other species? The answers to these questions suggest that a relatively conservative approach should be taken toward interspecies extrapolations of regulation, even across limited genetic distances.
MATERIALS AND METHODS

Bacterial strains, media, and growth conditions

The bacterial strains used for this study are listed in Table 1. In all cases cells were grown in baffled flasks shaken at 37°C. Morpholinopropane sulfonic acid (MOPS) glucose minimal medium, and MOPS-based defined rich medium (47) were purchased from Teknova (Hollister, CA). In experiments comparing *E. coli* and *V. cholerae* with *P. mirabilis*, media for all strains were supplemented with 0.01 mM nicotinic acid, which was found to be required for the growth of *Proteus mirabilis* (28) and of the *lrp* mutant of *Vibrio cholerae* (unpublished observation). When *lrp* mutants were part of an experiment, minimal media also contained 0.01 mM each of pantothenate and thiamine, which we found to be additional requirements of the *P. mirabilis lrp* mutant, and in some cases 0.1 mM L-cysteine with 0.2 mM L-methionine which were not required but improved growth of this mutant (unpublished observation). For *livK-lacZ* comparison, additional amino acids were used at the following final concentrations: 10 mM L-leucine, 0.4 mM L-isoleucine and 0.4 mM L-valine. Antibiotics were used, where indicated, as follows: 100 µg ampicillin /ml, 15 µg chloramphenicol /ml, 100 µg kanamycin /ml, and 10 µg tetracycline /ml.

The *lrp* alleles are as follows. For *E. coli* and *V. cholerae*, all but the first six and last six codons of the *lrp* ORF were replaced by the gene for chloramphenicol acetyltransferase (*cat*) (our unpublished result; N. Dolganov and G. Schoolnik, unpublished result), with confirmation by PCR amplification and sequencing. Some experiments made use of strains carrying an *E. coli lrp::Tn10*
allele (*lrp-35*, (40)). The *P. mirabilis* allele is a *lrp::miniTn5* disruptant ((48), provided by G. Fraser). The *E. coli* MG1655 *lrp* mutant has the entire *lrp* ORF replaced by the gene for kanamycin, and was constructed using λ*red* recombinase gene replacement system (49) (Betsy Martinez-Vaz and ABK, unpublished results). Other strain information is in Table 1.

**Growth experiments and sample isolation**

Overnight cultures in MOPS glucose or MOPS rich medium were inoculated from (respectively) M9 glucose or LB agar plates containing 0.01 mM nicotinic acid, and grown to early stationary phase. These cultures were then used to inoculate fresh media (1:32). OD\textsubscript{600nm} was measured following sample dilution as needed to maintain OD within the range of 0.08-0.3.

Samples for real-time RT-PCR analysis were isolated at the indicated times by removing an equal number of cells (estimated from culture density) from the flask and immediately adding it to two volumes of RNA stabilization buffer (RNA Protect Bacteria Reagent, Qiagen, Valencia, CA). This prevents the rapid changes in mRNA content that otherwise occur when bacteria are harvested. Samples were mixed, left at room temperature for 10 min, and stored at 4 °C for no more than 5 days.

Samples for microarray analysis were isolated at an OD\textsubscript{600} of ~0.4, at which point 20 ml of culture was mixed with 2.5 ml of ice-cold 5% water-saturated phenol (pH <7.0) in ethanol (50). After 10 min on ice, cells were pelleted, supernatant was removed, and pellets were frozen in liquid nitrogen and stored at
–80°C, if necessary.

**RNA Isolation and cDNA synthesis**

For RT-PCR experiments total RNA was isolated using the RNeasy miniprep kit (Qiagen) using their protocol with an added sonication step. Briefly, cells in the stabilization buffer were harvested by centrifuging at 4°C for 15 min at 5,000 rpm. Supernatants were removed and the pellet was resuspended in 1x TE buffer containing lysozyme (400μg/mL). Lysis buffer was added and the cells were sonicated 3x for 15 s in a cup horn attachment to enhance lysis. Following ethanol precipitation, RNA was bound to the column provided, washed and eluted. To eliminate DNA, the RNA was treated with RQ1 RNAse-free DNase (Promega, Madison, WI) as directed. cDNA was synthesized using total RNA as template, random hexamers (Invitrogen, Carlsbad, CA), and ImPromII reverse transcriptase (Promega). The random primers were annealed at 25°C for 5 min, and the first strand was then extended at 42°C for 1 h. The reverse transcriptase was inactivated by heating to 70°C for 10 min. cDNA samples were stored at -20°C.

For microarray experiments total RNA was extracted by the hot phenol-chloroform method (51), and treated with DNase I in the presence of RNase inhibitor for subsequent labelling by reverse transcription with Cy3-dUTP and Cy5-dUTP fluorescent dyes (Amersham, Little Chalfont, United Kingdom). The RNeasy miniprep kit (see above) was also used in some cases.
Real time RT-PCR analysis

Primer sets (Integrated DNA Technologies, Coralville, IA) were designed for the *adhE*, *gltB*, *lrp* and *recA* genes for each strain. Before each new experiment dilutions of cDNAs were tested to determine the concentration that gave maximally-efficient amplification, and to determine the efficiency for each primer set (23). Cycle threshold (C\_T) values were determined by Roche Lightcycler detection of SYBR green fluorescence. Melting curve (Roche Lightcycler software) and agarose gel analyses were used to confirm the formation of specific products, which ranged in size from 192-202 bp. The standard curve method was used to determine relative amounts of mRNA and levels were normalized to *recA* (52).

Western blot analysis

For each sample equal volumes of cells were centrifuged at 16,000 x g for two min. The supernatants were removed and the cell pellets were stored at -80°C until analysis. Pellets were resuspended in 1X SDS buffer (Novagen, Madison, WI). Cells were lysed by heating to 98°C for ten min, and total protein concentrations were determined using the RC/DC kit and protocol (BioRad, Hercules, CA). Equal amounts of protein were loaded on a 12% acrylamide SDS gel and electrophoresed at 110 V in 1X tris-glycine buffer. Proteins were then electroblotted to polyvinylidene difluoride (PVDF) membranes at 30 V for 1 h using the Xcell blot apparatus (Invitrogen). Proteins were detected by fluorescence using the ECL-plus Western Blotting Detection System (GE Health
Sciences, Piscataway, NJ) as per the manufacturer’s protocol with a 1:125
dilution of rabbit anti-Lrp polyclonal serum (gift of Dr. Joseph Calvo (53)), and a
1:25,000 dilution of horseradish peroxidase-conjugated goat anti-rabbit IgG (gift
of Dr. Darren Sledjeski). Protein bands were visualized on a Storm 840I
phosphorimager (Molecular Dynamics, now GE Healthcare). Densitometric
analysis of Lrp bands was performed using the Molecular Dynamics software,
and the amount of Lrp in each sample was determined by comparison to a
standard curve from purified Lrp dilutions included on each gel.

Microarray experiments
Starting with freshly-streaked single colonies, cultures (supplemented with Ile,
Leu, Val and thiamine as described above) were aerobically grown overnight at
37 °C and then diluted 20-fold into 20 ml of fresh medium. Recombinant cultures
were propagated in the presence of chloramphenicol, and growth was monitored
via OD<sub>600nm</sub>. Cultures were maintained in exponential growth for at least 10
generations by dilution. Relative mRNA abundances between the <i>lrp</i> mutant and
the same strain carrying a <i>lrp</i> gene on plasmid pCC1 (or carrying only the vector)
were determined. This analysis employed <i>E. coli</i> K-12 whole-genome DNA
microarrays including 99% of all annotated open reading frames and the stable
RNA genes. Slide preparation, reverse transcription with the Cy-dyes,
hybridization, and image scanning were performed as previously described (51).
The fluorescent probes were hybridized to an array at 65 °C for 6 h. Intensities in
both channels were smoothed using the Lowess method (54). Dye- and array-
specific noise was removed using the analysis of variance (ANOVA) error model (55). In pair-wise comparisons, differentially expressed genes were identified at an estimated false discovery rate of less than 5% using the SAM package (56). Known Lrp targets were taken from RegulonDB (10,44), and are listed below.

aidB  aroA  b2659  dadA  dadX  fimA  fimC  fimD
fimE  fimF  fimG  fimH  fimI  gabP  gabT  gcvH
gcvP  gcvT  gltB  gltD  gltF  ilvA  ilvD  ilvE
ilvG_1  ilvG_2  ilvH  ilvI  ilvL  ilvM  kbl  livF
livG  livH  livJ  livK  livM  lrp  lysU  malt
micF  ompC  ompF  oppA  oppB  oppC  oppD  oppF
osmC  osmY  sdaA  serA  serC  stpA  tdh  yeiL
ygaF

\textbf{β-galactosidase assays}

Strains were grown to exponential phase in glucose minimal MOPS medium (Teknova). Samples were taken at 20 and 30-min intervals throughout the growth period. Levels of β-galactosidase were determined by o-nitrophenyl-β-D-galactoside (ONPG) hydrolysis as described by Platko \textit{et al.} (57). β-galactosidase levels were plotted against culture absorbance, and points were fitted via linear regression. The resulting slope yields the β-galactosidase activity.
Cloning of *lrp* orthologs

The *lrp* genes (translational start to stop) from *E. coli* O157:H7, *P. mirabilis* HI4320 and *V. cholerae* El tor A1552 were PCR amplified from chromosomal DNA using Pfx DNA polymerase (Invitrogen). The upstream PCR primers contained a consensus *E. coli* ribosome binding site. Fragments were gel purified and cloned into the low-copy pCC1 blunt cloning vector (Epicentre), and transformed into *E. coli* EPI300 per the manufacturer’s protocol. As a vector control, an irrelevant ~1360 bp DNA fragment (kanamycin resistance cassette provided by the manufacturer as a ligation control) was inserted into pCC1. Transformants were selected using chloramphenicol and sequence-confirmed. The recombinants pECLRP, pPMLRP and pVCLRP (Table 1) were isolated using Qiagen miniprep columns. The purified plasmids were then electroporated into *E. coli* BE3780 (Table 1) using a BioRad *E. coli* gene pulser and protocol. For experiments with *P. mirabilis* and *V. cholerae* Δlrp strains, which are already chloramphenicol resistant, these plasmids were digested with BsmI to remove the *cat* gene, and we inserted a kanamycin resistance gene PCR amplified from pACYC177.

Construction of *lacZ* fusions

The promoter regions of the *lrp* and *gltB* genes were PCR amplified from *E. coli* O157:H7, *Proteus mirabilis* HI4320 and *Vibrio cholerae* El tor type N16969 chromosomal DNA using gene specific primers (Table 1) and Pfx DNA polymerase (Invitrogen). The PCR products were digested with BamHI and Sall
and ligated into pBH403, which is a derivative of pKK232-8 and contains a promoterless lacZ gene between two bidirectional transcription terminators. The recombinant plasmids (Table 1) were electroporated into E.coli BE10.2 and PS2209; Proteus mirabilis U6450 and U6450Δlrp; and Vibrio cholerae El tor strain A1552 and A1552 Δlrp.
RESULTS

Differences in growth phenotypes of *lrp* null strains

If Lrp is having similar broad effects on gene expression in *E. coli*, *P. mirabilis*, and *V. cholerae*, then one would expect to see similar effects of a *lrp* null mutation on their growth. Fig. 2A shows the results of growth experiments for wild-type (WT) and *lrp* strains of these three species grown in MOPS glucose and MOPS defined rich media, all supplemented with nicotinate as required by the *P. mirabilis* strains, and with pantothenate and thiamine as required by the *P. mirabilis* ∆*lrp* strain (see Methods). The plot shows WT specific growth rate on the x-axis, and the rate for the *lrp* strain on the y-axis; thus where *lrp* mutation has no effect on growth rate the points would fall on the diagonal line. In the rich medium (closed symbols), *lrp* mutation had little effect on growth of any of the three species. However in the minimal medium, *P. mirabilis* stands out as having a substantial growth rate decrease when *lrp* is mutated (193 min vs. 66 min doubling time for the WT strain). This might represent a *lrp*-dependent partial auxotrophy, in addition to the *lrp*-dependent requirements for pantothenate and thiamine that were satisfied by the medium. However that may be, it is clear that the *lrp* mutation has differential effects in these three species. The effect of the *lrp* allele on *P. mirabilis* growth in minimal medium was fully complemented by supplying the cloned *P. mirabilis* *lrp* gene on a plasmid, and was mostly but incompletely complemented by the *lrp* genes from *E. coli* or *V. cholerae* (Fig. 2B).
Regulation of *E. coli* target genes by Lrp proteins from *P. mirabilis* and *V. cholerae*

If Lrp is functionally conserved among species, as assumed by the regulatory extrapolation hypothesis, then WT *lrp* alleles from the different species should be able to cross complement for gene regulation. We tested the ability of heterologous Lrp proteins to properly regulate (activate or repress) three Lrp-responsive genes in a *lrp* null mutant of *E. coli*. We cloned the *lrp* genes from *P. mirabilis* and *V. cholerae* downstream of PlacUV5 in the low-copy vector pCC1 (Epicentre); a consensus *E. coli* Shine-Dalgarno sequence (58) was introduced as well. As a positive control, we also cloned the *lrp* gene from *E. coli* O157:H7 (which is identical to that of *E. coli* K-12 at the amino acid level). These *lrp*-bearing plasmids, which produce Lrp independently of the normal growth-dependent control that characterizes the native *lrp* gene (59,60), were then used to transform the *E. coli* strains described below. Western blot analysis of a constant amount of total protein, probed with a polyclonal anti-Lrp antiserum (53), revealed comparable accumulation of the various Lrp orthologs (Fig. 3A). The strain carrying the *V. cholerae* *lrp* gene appeared to accumulate ~75% as much Lrp protein as the *E. coli* control (Fig. 3A and data not shown), though this is a minimal estimate because the antiserum was generated against *E. coli* Lrp (92% identical to *V. cholerae* Lrp at the amino acid level).

Lrp directly represses its own promoter in *E. coli*, and this occurs whether or not leucine is present (61). To begin testing the functional equivalency of the different Lrp orthologs we co-transformed strain BE10.2 (ΔlacZ, *lrp-Tn10*) with a
vector containing an *E. coli* Plrp-lacZ fusion and the various *lrp*-bearing plasmids (or vector control). Strains were grown in MOPS glucose, and the specific activity of β-galactosidase was determined and plotted against culture density to more quantitatively assess its level and to assure that the cultures were in balanced growth. Compared to the vector control, *V. cholerae* Lrp repressed Plrp-lacZ to the same extent as did *E. coli* Lrp (~2 fold), while the *P. mirabilis* Lrp (98% identical to *E. coli* Lrp) repressed about twice that much (Fig 3B).

Transcriptional activation is a more demanding process than repression, in the sense that the activator has not only to bind the DNA correctly but also (in most cases) to make productive contacts with RNA polymerase (62,63). Strain BE3780 contains a chromosomal gltB-lacZ operon fusion in the *E. coli* BE10 background (∆lacZ, *lrp-Tn10*) (46). In *E. coli*, glutamate synthase (GltBD, NADPH-dependent glutamine:2-oxoglutarate amidotransferase) is required for glutamate synthesis when the nitrogen source consists of low levels of ammonia (64,65). Lrp directly and strongly activates gltBDF transcription in *E. coli* (46,66), in a process that also requires the global regulator IHF (67). We found that, relative to the vector control, activation of gltB transcription by *P. mirabilis* Lrp was indistinguishable from that of the *E. coli* *lrp* positive control (Fig. 3C). Activation by *V. cholerae* Lrp was ~90% that of the *E. coli* control, as judged by independent linear fits of the *E. coli* and *V. cholerae* data in Fig. 3C (not shown), and this slight difference may in part reflect apparent differences in levels of the respective Lrp orthologs (see above).
Leucine responsiveness of heterologous Lrp proteins

Aside from the ability to activate or repress transcription, Lrp also transduces metabolic signals in the form of amino acid pool levels, in particular the amino acids L-leucine and L-alanine (34). The livKHGMF operon is one of two high-affinity branched chain amino acid transport systems in E. coli (68). The livK gene specifies a periplasmic protein responsible for binding to leucine (69). In the presence of exogenous leucine Lrp represses livK (43,70), but Lrp activates this same gene when leucine is not in the medium (39). Thus a PlivK-lacZ operon fusion should be a particularly sensitive indicator of the responses of Lrp orthologs to leucine. The amino acid residues previously demonstrated to be involved in leucine binding in E. coli Lrp (34,71) are completely conserved in the P. mirabilis and V. cholerae orthologs (boxes in lower half of Fig. 1).

We prepared a PlivK-lacZ operon fusion (pRLIV2), and co-transformed it into E. coli BE10.2 (∆lacZ, lrp-Tn10) together with plasmids bearing the heterologous lrp alleles under the control of PlacUV5 (Table 1). These strains were grown in MOPS glucose media containing isoleucine (I, Ile), leucine (L, Leu) and valine (V, Val). Leu was not used alone as it can lead to starvation for Ile, via feedback inhibition of L-threonine deaminase (72). β-galactosidase activity was determined as described above. In general the Lrp proteins from P. mirabilis and V. cholerae yielded livK regulatory patterns similar to that of the E. coli control, showing activation in the absence of leucine and repression in its presence (Fig 3D and G). However both P. mirabilis and V. cholerae Lrp gave threefold greater repression than E. coli Lrp (Fig 3G). We also looked at the
effects of adding ILV on the regulation of *E. coli* *Plrp* and *PgltB*. Addition of ILV interfered to a moderate extent with the repression of *Plrp* by all three Lrp orthologs (Fig 3E), but the differences observed in the MOPS glucose cultures were maintained, with substantially greater repression by *P. mirabilis* Lrp. Also as expected from previous studies (46,66), activation of *PgltB* by *E. coli* Lrp was moderately reduced in the presence of ILV; the two heterologous Lrp orthologs gave patterns essentially indistinguishable from that of *E. coli* Lrp (Fig 3F).

**Microarray analysis of gene regulation by heterologous Lrp proteins**

The functional conservation of Lrp orthologs was more broadly assessed by using microarrays to analyze gene regulation, when *lrp* alleles from *E. coli*, *P. mirabilis*, or *V. cholerae* (all fused to the same expression sequences) are used to complement *lrp* knockout in *E. coli* K-12. Fewer than half of the genes regulated by *E. coli* Lrp under these conditions were regulated by either of the other two orthologs. Only about 25% of genes differentially regulated by each of the orthologs were regulated in common by all three Lrp proteins (Fig. 4). Similar proportions were seen when activated and repressed genes were considered separately (not shown). [Note: “repression” and “activation” as used here include both direct and indirect effects.]

Next we examined whether the subset of genes that was regulated by both *E. coli* Lrp and one of the other orthologs showed a similar magnitude of regulation (Fig. 5). The effects of *E. coli lrp* (pEcoLrp) on gene expression are shown on the x-axis in every panel. Column A shows the set of genes for which
transcript levels gave statistically-significant decreases when the \( lrp \) mutation was complemented. Column A thus represents genes that are repressed (directly or indirectly) by \( E. coli \) Lrp. Column B includes gene showing direct or indirect activation by \( E. coli \) Lrp. Column C shows the set of 57 genes recognized in RegulonDB (10,44) as being directly controlled by Lrp, whether the control is positive or negative. This set includes genes that are controlled by Lrp, though not under the growth conditions used by us, so the cluster of genes showing no effect of Lrp is not surprising (39,42).

For each column, the top row indicates the effects of “complementing” the \( \Delta lrp \) allele with the vector alone (pCC1). Full complementation relative to that by pEcoLrp would yield a slope of 1.0 for the linear fit. Not surprisingly, the vector gives slopes of <0.2 for all three gene sets. Supplying the \( lrp \) allele from \( P. mirabilis \) (pPmiLrp, bottom row) gives substantial, though not full, complementation of the gene expression pattern, with slopes ranging from 0.81 to 0.93. The \( lrp \) allele from \( V. cholerae \) (pVchLrp), which is more divergent from that of \( E. coli \) than is that of \( P. mirabilis \) (Fig. 1), yields lower overall regulatory complementation – the slopes range from 0.45 to 0.49. Thus small changes in the Lrp sequence, outside of the fully-conserved helix-turn-helix motif, have substantial effects on the magnitude of Lrp-dependent effects on gene expression.

**Effects of orthologous \( lrp \) alleles on swarming behavior in Proteus**

\( P. mirabilis \) undergoes differentiation to form hyperflagellated swarmer cells over
20-fold longer than nonswarmer cells, and yields concentric rings of growth on agar ((28,31); Fig. 6A). It has been shown by others (48) that a *lrp* mutation abolishes swarming in *P. mirabilis* (Fig. 6B). We show here that the *lrp* orthologs from both *E. coli* and *V. cholerae* complement a *P. mirabilis* *lrp* mutant, and regenerate the complex swarming behavior (Fig. 6, panels E and F). The *P. mirabilis* *lrp* gene does not regenerate the exact WT swarming pattern when supplied *in trans* (Fig. 6A, D); this difference may reflect replacement of the native expression sequences with *P{lacUV5} on the plasmid. Nevertheless the three plasmid-borne *lrp* alleles have identical expression sequences, and there is a consistent difference in the swarming patterns resulting from their respective complementation as reflected in growth ring measurements (Fig. 6G).

**Levels of Lrp protein in *E. coli*, *P. mirabilis*, and *V. cholerae* in different media and growth phases**

For orthologous regulatory proteins to generate the same expression patterns of orthologous target genes, it is not enough that the regulatory proteins are *intrinsically* equivalent, meaning that they have the same DNA binding specificity and equivalent interactions with small molecules and with other proteins. They must also share *extrinsic* properties, including accumulation to similar levels under various growth conditions. Potential differences in expression levels were minimized in the complementation experiments described above, by providing each *lrp* ortholog with a common promoter and translation initiation region. However *E. coli* growing exponentially in a minimal glucose medium accumulates
three- to four-fold more Lrp than in rich medium (59,60). To determine whether this pattern of Lrp accumulation is conserved, we used western blot analysis to measure the levels of Lrp throughout a batch growth cycle in *E. coli*, *P. mirabilis* and *V. cholerae* grown in MOPS glucose and MOPS defined rich media (supplemented as described in Methods).

Our results confirm the earlier studies of *E. coli*, in that we saw several-fold higher Lrp levels when cells were grown in MOPS glucose than when they were grown in defined rich medium (Fig. 7, compare panels D and J). When grown in MOPS glucose, *P. mirabilis* (Fig. 7E) and *V. cholerae* (Fig. 7F) produced levels of Lrp similar to that in *E. coli* (Fig. 7D). Furthermore, all three species showed severalfold lower Lrp levels in rich than in minimal medium. There was, however, one substantial difference among the cultures. In defined rich medium, *P. mirabilis* (Fig. 7K) produced up to twice as much Lrp as *E. coli* or *V. cholerae*, with levels highest in stationary phase. This difference is clear even from normalized median values derived by summing all datapoints in each panel (Fig. 7M).

These differences between *E. coli* and *P. mirabilis*, in Lrp protein levels, could well have substantial regulatory significance but needed to be confirmed. Samples were taken from parallel cultures of *E. coli* and *P. mirabilis* during early logarithmic, mid logarithmic, late logarithmic and stationary phases in defined rich medium. Equal amounts of total protein were resolved side-by-side via SDS PAGE. The results of the subsequent western blot (Fig. 7N) confirm that *P. mirabilis* produces substantially more Lrp protein throughout the growth phases,
with the greatest difference (nearly twofold) seen in stationary phase. Thus related bacteria with nearly-identical regulator proteins can have significant differences in the levels of those regulators.

Levels of *lrp* mRNA in *E. coli*, *P. mirabilis*, and *V. cholerae* in different media and growth phases

As regulator levels are often inferred from microarray measurements of mRNA levels, we determined whether the level of *lrp* mRNA also varies with the growth medium in all three organisms (Fig. 8A). At least four QRT-PCR determinations from each of two independent experiments were averaged to generate each plotted value, and the results from minimal medium are plotted against those from defined rich medium. If these two media have no differential effects on *lrp* mRNA levels, then the points would fall on the diagonal line. *E. coli* shows no significant change between log (closed) and stationary phases (open symbols) in either medium, but as expected *lrp* mRNA levels are severalfold lower in rich than in minimal glucose medium. *V. cholerae* is similar to *E. coli* in that *lrp* mRNA levels change little if at all with growth phase, but differs in that there is no significant change with the growth medium, and the log phase levels are about half the *E. coli* value in minimal medium, but 2-3 times the *E. coli* value in defined rich medium. Unlike *E. coli* or *V. cholerae*, when grown in defined rich medium *P. mirabilis* shows about a doubling in stationary versus log phase *lrp* mRNA levels (in minimal glucose medium there is no significant change with growth phase). Like *E. coli*, *P. mirabilis* (in log phase) produces less *lrp* mRNA in defined rich
medium than in minimal glucose medium.

We also measured the levels of \textit{lrp} mRNA for all three organisms during log-phase growth, using a more highly-quantitative method, and the results are consistent with the protein data (Fig. 8B-D). We used a sensitive dilution-response approach that makes use of the fact that our three species-specific pairs of QRT-PCR primers for \textit{lrp} amplify with the same efficiency but are completely specific for their respective template DNAs (data not shown). A baseline amount of \textit{E. coli} cDNA (from a mid-log phase culture in MOPS glucose plus nicotinate) was mixed with varying amounts of test cDNA (from cultures grown in either the glucose or defined rich medium), and the mixes were used as template for simultaneous amplification with two primer pairs. If the test cDNA preparation has the same proportion of \textit{lrp} cDNA as the reference pool, the detected amount of \textit{lrp} cDNA should rise with a slope of 1.0 (actual vs. detected, based on the known amount of total cDNA added); this is shown as a dotted line in each panel. For \textit{E. coli}-derived cDNA, this slope is about 0.75 (glucose minimal culture; Fig. 8B) or 0.4 (defined rich culture), consistent with the medium-dependent effect on Lrp protein levels shown in Fig. 5. \textit{V. cholerae} (Fig. 8D) gave a pattern similar to that of \textit{E. coli}. However – also consistent with the protein data – \textit{P. mirabilis} had substantially more \textit{lrp} cDNA as a proportion of total cDNA, with slopes of about 2 (Fig. 8C).

\textbf{Regulation of orthologous target genes in their native backgrounds}

To determine if differences in Lrp-dependent gene expression occur when the
target gene, Lrp, and host background are all native, we examined promoter regions from orthologs of two genes previously shown to be Lrp responsive in *E. coli* (42): *adhE* and *gltB*. These orthologs were chosen based on percent identity to the *E. coli* protein, and presence of at least one predicted Lrp-binding site. Figs. 9A and B show alignments of the upstream regions for these genes, indicating the Lrp-binding sites as predicted by the PRODORIC virtual footprinter (73).

Wild-type and *lrp* strain pairs were grown in MOPS defined rich medium. Samples were taken in early logarithmic phase (OD$_{600\text{nm}}$ = 0.3), and early stationary phase (1 h after the culture OD vs. time semilogarithmic plot showed divergence from exponential growth). Real-time RT-PCR analysis was used to determine the relative levels of *adhE*, *gltB*, and *recA* mRNAs. The experiment was performed in triplicate and the relative levels of mRNA were determined using the standard curve method (52) and by normalizing to *recA*. There is no effect of Lrp on *recA* expression, at least in *E. coli* and *V. cholerae* under our conditions ((42) and N. Dolganov, pers. commun.).

**Native regulation of *adhE***. AdhE is a fused acetaldehyde-CoA dehydrogenase and iron-dependent alcohol dehydrogenase and pyruvate-formate lyase deactivase (74-76). In *E. coli*, the gene *adhE* is preferentially expressed in stationary phase (77-79), and repressed by Lrp in a leucine-independent manner during exponential growth in minimal glucose medium ((42) and A. Khodursky, pers. commun.). Fig. 10A shows the regulatory pattern of *adhE* in all three organisms. If Lrp had no effect on *adhE* mRNA levels, then the
points would fall on the diagonal line. The fact that all points are above the diagonal line is consistent with Lrp-dependent repression in all cases. However, a more detailed analysis of this data reveals that the regulatory patterns from *P. mirabilis* and *V. cholerae* are different from those in *E. coli* and from one another.

In *E. coli* (circles), there was a very modest decrease in *adhE* mRNA in log phase, but only in the presence of Lrp; in neither case is there a strong Lrp effect. *P. mirabilis*, in contrast (triangles), showed a strong repressive effect of Lrp, though no real growth-phase dependent change in expression. *V. cholerae* (squares) exhibited slight Lrp-associated reduction in log phase expression, but in stationary phase the *adhE* mRNA levels were about 50-fold higher in the *lrp* mutant strain.

**Native regulation of *gltB***. The other target gene, *gltB*, was described earlier. *V. cholerae* has two GltB orthologs, with 73% and 43% identity to *E. coli* GltB. We failed to detect expression of the *gltB* with higher identity to *E. coli* GltB during growth in minimal glucose and defined rich medium (Vc2376, not shown), however the lower-identity ortholog was expressed. In *E. coli* *gltB* is activated 30-40 fold by Lrp when grown in MOPS glucose (40,42,46,66), with the activation dependent on another global regulator, IHF (67,80). We have already shown that the Lrp orthologs from *P. mirabilis* and *V. cholerae* effectively replace *E. coli* Lrp, in an *E. coli* background, for activation of *E. coli PgltB* (Fig. 3, panels C and F). Here we determine whether the *Proteus* and *Vibrio* Lrp orthologs each activate their native *gltB* promoter in the native background. The *P. mirabilis lrp* strain did not grow well in the MOPS minimal medium used in this study, so all
experiments were carried out in MOPS defined rich medium. In another rich medium (LB), activation of gltB by Lrp is greatly reduced relative to minimal glucose, but is still about triple the level in a lrp disruptant (40).

We found that E. coli gltB is activated ~25 fold by Lrp during mid-log and about half as much in early stationary phase (Fig. 10B). P. mirabilis gave several-fold more log-phase gltB expression in the lrp+ than in the lrp strain, with little if any growth-phase-dependent change. V. cholerae gave the most divergent expression pattern: gltB (Vc2373) mRNA levels were halved by Lrp in log phase, but increased about fivefold by Lrp in early stationary phase. In other words, while Lrp activates gltB in E. coli and P. mirabilis, under the same conditions it slightly represses Vc2373 in V. cholerae.

**Lrp regulatory interactions with two promoter regions**

Finally we tested whether orthologs of two promoter regions, that are Lrp-controlled in E. coli, are regulated by Lrp in heterologous hosts. These are the reciprocal experiments to those shown in Fig. 3, where heterologous lrp alleles were moved into an E. coli background to test for control of E. coli target genes.

**Regulation of Plrp.** One promoter set was Plrp from E. coli, P. mirabilis, or V. cholerae (Fig. 9C). The lacZ transcriptional fusions to these promoters were introduced into the lrp mutant or parental strains of these three species in all combinations. Relative LacZ activity was measured using the approach shown in Fig. 3 (determining the slope of a LacZ activity vs. culture density plot), and for each strain pair the lrp result was plotted vs. the lrp+ result – if Lrp did not affect a
given promoter in a given strain pair, the resulting point would fall on the diagonal line.

The *E. coli* host gave highest expression levels for all three *P* _lrp-lacZ_ fusions (Fig. 11A, open symbols), among both the *lrp* and *lrp*⁺ strains, though the *lrp* promoter from *V. cholerae* was the only one to show substantially higher expression in the absence of Lrp. In the *P. mirabilis* background, all three fusions gave the lowest expression levels in the presence of Lrp (Fig. 10A, black symbols), perhaps consistent with the elevated level of Lrp in those cells (Fig. 7K, N). However *P* _lrp_ from *Proteus* itself (triangles) showed only minimal (1.5-fold) reduction of expression in the presence of Lrp, as it did in the two heterologous backgrounds, also perhaps consistent with the elevated Lrp levels in that background.

Overall the data for each *P* _lrp-lacZ_ fusion in its native host is supported by our protein (Fig 7D-F) and mRNA measurements (Fig 8A) with the exception of one. The *P* _lrp-lacZ_ from *V. cholerae* showed low expression in the *V. cholerae* background, whether or not Lrp was present (Fig. 11A, gray square) though this same promoter gave much higher – and Lrp-responsive – expression in the *E. coli* and *P. mirabilis* hosts (Fig 11A, compare y-axis values of filled and open squares). This, and the fact that the *E. coli* and *P. mirabilis* fusions are well expressed in the Vibrio background (Fig 11A, compare gray circle and gray triangle), suggests that *V. cholerae* negatively regulates its *P* _lrp_ via some *Vibrio*-specific factor in addition to Lrp. There are some intriguing and distinctive sequence characteristics of the *V. cholerae* *P* _lrp_ that may explain this behaviour,
and we are investigating these further.

**Regulation of P_{gltB}**. The second promoter set analyzed in this way was P_{gltB} from *E. coli*, *P. mirabilis*, or *V. cholerae* (Fig. 9B). The results shown in Fig. 11B resemble those in Fig. 10B, in that most points are below the diagonal, consistent with activation by Lrp. However the data in Fig. 10B show QRT-PCR measurements of *gltB* mRNA levels in the homologous background only, in log and stationary phases; while the data in Fig. 11B show P_{gltB-lacZ} fusion activity measurements, in log phase only, of all possible homologous and heterologous combinations.

Like the P_{lrp} fusions, the P_{gltB} fusions almost all gave their highest expression levels in the two *E. coli* backgrounds (Fig. 11B, open symbols; the one exception is Vc2373 – see below). Also like the P_{lrp} fusions (again with one exception), the lowest expression levels were in *P. mirabilis* (Fig. 11B, black symbols). For three of the four fusions, this low expression in *P. mirabilis* was true in both the *lrp* and *lrp*\(^+\) strains; the P_{gltB} from *E. coli* was unique in showing substantial Lrp activation of expression in all backgrounds (circles). In *V. cholerae*, the various P_{gltB} promoters give a wide range of responses (gray symbols). P_{gltB} from *Proteus*, which is expressed well in the other two backgrounds, is not detectably expressed in *V. cholerae* (under the conditions used). The two P_{gltB} fusions from *V. cholerae* are well expressed, but neither promoter shows a great deal of activation by Lrp in any of the three backgrounds. The Vc2373 promoter shows particularly diverse responses, from high-level Lrp-insensitive expression in the native host, through mid-level and mildly Lrp-
activated expression in *E. coli*, to lower-level and mildly Lrp-repressed expression in *P. mirabilis*. 
DISCUSSION

Robust methods for predicting gene regulation from DNA sequence data would greatly increase the value of the rapidly-expanding collection of bacterial genome sequences. However current methods rely on a poorly-tested hypothesis – that a well-conserved regulator, and well-conserved target gene downstream of a putative binding site for the regulator, together imply a similar pattern of regulation. For brevity, we refer to this as the “regulatory extrapolation hypothesis,” since it involves inference of a regulatory pattern based on conservation with respect to a well-studied reference organism. Some possibilities regarding this hypothesis are: that it is generally true only among closely-related organisms (genetically, ecologically, or both), that it is generally true for only the most-highly conserved regulators and target genes, or that it is often incorrect even among highly-related genes and organisms.

We have begun to test the regulatory extrapolation hypothesis by examining a well-conserved regulator (Lrp), conserved genes that are Lrp regulatory targets in *E. coli*, and two species of increasing but limited genetic distance from *E. coli*: *Proteus mirabilis* and *Vibrio cholerae*.

Closely-related Lrp proteins have subtle intrinsic differences

The regulatory extrapolation hypothesis relies on a tacit assumption that regulatory proteins with greater than roughly 90% identity in amino sequence are functionally equivalent. We took closely-related Lrp orthologs (all >92% identity) from three species, gave them identical expression sequences for transcription
and translation, put them into the same low-copy vector (pCC1), and introduced
them into the same \textit{E. coli} \textit{K-12} \textit{lrp} and \textit{P. mirabilis} \textit{lrp} backgrounds. The result
was similar overall behavior, supporting the assumption in general, but with
significant functional distinctions that are inconsistent with the assumption. In \textit{E.
coli}, the native \textit{Plrp} (fused to \textit{lacZ}) was repressed equivalently by \textit{E. coli} and \textit{V.
cholerae} Lrp, but substantially more by \textit{P. mirabilis} Lrp (Fig. 3B). These
differences were magnified in the presence of leucine, where \textit{P. mirabilis} Lrp was
unique in showing virtually no effect (Fig. 3E). In contrast, the various Lrp
orthologs gave equivalent activation of \textit{P}gl\textit{ltB} (Figs. 3C, F). \textit{PlivK}, which is
activated by Lrp in the absence of leucine and repressed in its presence, was
regulated equally by all three Lrp orthologs with one exception: in the presence of
leucine, the \textit{E. coli} Lrp represses less than the others (Fig. 3D, G).

Of the 164 aa in all three Lrp orthologs, \textit{P. mirabilis} Lrp differs from \textit{E. coli}
Lrp at four positions (98\% identity), while the \textit{V. cholerae} and \textit{E. coli} orthologs
differ at 12 positions (93\% identity); none of these changes affected the known
DNA-binding helix-turn-helix or the coregulator binding sites (Fig. 1).

Bioinformatic analyses would tend to assume that these orthologs were
functionally equivalent. We used microarray analysis to more globally assess the
ability of orthologous Lrp proteins to properly control the \textit{E. coli} \textit{K-12} Lrp regulon.
Our results confirmed that minor changes in the Lrp amino acid sequence had
substantial effects on the magnitude of Lrp effects.

In the microarray studies we used the slope of a least-squares fit between
two ratios as a measure of overall regulatory concordance (Fig. 5). One ratio is
the expression level of genes in *E. coli* lrp knockout over the level in that strain complemented by a plasmid carrying the *E. coli* lrp gene (on the x-axis in all panels). The second ratio (y-axis) is the expression in *E. coli* lrp- over that in the same strain complemented by a test plasmid (vector control, or *lrp* from *P. mirabilis* or *V. cholerae*). We found that whether we examined genes significantly repressed by Lrp, activated by Lrp (in both cases including both direct and indirect effects), or directly controlled by Lrp in *E. coli*, the results were consistent. Specifically, Lrp from *P. mirabilis* gave differences of 7-19% from *E. coli* Lrp, while Lrp from *V. cholerae* gave differences of 31-51% from *E. coli* Lrp. As expected, the vector control gave differences of 84-98% from *E. coli* Lrp. Additionally, transcriptional effect of *P. mirabilis* Lrp were most similar to *E. coli* Lrp (correlation between 80 and 92%), followed by *V. cholerae* (correlation between 67 and 70%), with vector alone showing no statistically significant similarity (0-20%).

Another assessment of Lrp functionality involved *P. mirabilis* swarming over a solid surface. Swarming is a complex phenomenon; in *Salmonella* about a third of all genes showed swarming-associated changes in expression (81). For the purposes of the present study, swarming thus represents a sensitive indicator of Lrp action. In a *P. mirabilis* background, we found that all three Lrp orthologs restored swarming, but gave repeatable differences in the resulting swarming patterns (Fig. 6).
Orthologous Lrp proteins can have different extrinsic properties

We examined the native regulation of *lrp* in *E. coli*, *P. mirabilis*, and *V. cholerae*. The regulatory extrapolation hypothesis relies on a second tacit assumption: that Lrp levels are similar in the organisms being compared, and change similarly in response to growth conditions. Regulation is, not surprisingly, affected by the level of the regulatory protein; this is specifically true for Lrp (39,66,82). In fact, Lrp protein levels in all three species are reduced in rich medium relative to glucose minimal medium (Fig. 7). For two of the three species, *lrp* mRNA levels are also lower in rich medium (Fig. 8). However we found that, during growth in defined rich medium (especially at higher cell densities), *P. mirabilis* levels of Lrp protein and *lrp* mRNA were about double those in *E. coli* or *V. cholerae* (Figs. 7-8). This higher level may reflect the observation that *P*l*rp* of *P. mirabilis*, unlike that from the other two organisms, shows no evidence of autogenous repression by Lrp (as judged by *P*l*rp*-lacZ fusions; Fig. 11A). This raises interesting questions about how Lrp levels are controlled in *P. mirabilis* (or, for that matter, in *V. cholerae*, where it appears that there is species- and *P*l*rp*-specific negative control; Fig. 11A). But for the purposes of this study, the important point is that Lrp levels differ significantly between the species, so that sequence analysis of the Lrp open reading frame and target gene promoter is not sufficient to predict expression patterns of the target gene.

Orthologous target genes are regulated differently by the same Lrp protein

The third tacit assumption underlying the regulatory extrapolation hypothesis is
the reciprocal of the one described immediately above: that orthologous target genes moved into a common background will be regulated in the same way. We prepared lacZ fusions to both PgltB and Plrp promoters from E. coli, P. mirabilis, and V. cholerae, and introduced them into the lrp and lrp+ strain pairs for all three species in all combinations. Once again the assumption is supported in general (all Plrp combinations are unaffected or repressed by Lrp− Fig. 11A; while all but one PgltB combinations are unaffected or activated by Lrp – Fig. 11B), but not supported by the specifics (e.g., the E. coli PgltB is well-expressed and Lrp-activated in all backgrounds, while the P. mirabilis PgltB ranges from nonexpression to Lrp-activated expression in the different backgrounds; Fig. 11B). [Note: repression and activation by Lrp have been demonstrated for these target genes in E. coli, but have not been proven to occur in the other backgrounds (where the effects might be indirect), and we have used these terms for brevity.]

Our results present a mixed picture. In general terms, we found that Lrp behaves in similar ways in the three tested species, suggesting that regulatory extrapolation over limited genetic distances can provide valid insights. However we also found significant intrinsic and extrinsic differences among the Lrp orthologs, and differences in the behavior of target gene promoters (that have predicted Lrp-binding sites), despite the fairly close genetic relatedness of the species we examined. These results suggest that while the regulatory extrapolation hypothesis can be informative in broad terms, it should be used with considerable caution for application to specific target genes.
ACKNOWLEDGEMENTS

The authors thank Gary Schoolnik and Nadia Dolganov (Stanford Univ.) for the *Vibrio cholerae lrp*<sup>+/−</sup> pair; Gillian Fraser (Cambridge Univ.) for the *Proteus mirabilis lrp*<sup>+/−</sup> pair; Harry Mobley (Univ. of Michigan) for *Proteus mirabilis* HI4320; Victor DiRita (Univ. of Michigan) for *V. cholerae* DNA; Joseph Calvo (Cornell) for anti-Lrp antiserum; Rick Keck for help with plate photography (Univ. of Toledo); Sadik Khuder (Univ. of Toledo) for help and advice regarding statistical analysis, and Mark Wooten and John Lazarus (Univ. of Toledo) for their help with QRT-PCR. We also thank Drs. Schoolnik and Isabel Novella (Univ. of Toledo) for comments on the manuscript. This work was supported by funds from NIH grant R01 AI54716 to RMB, with subcontract to ABK. REL was also supported, in part, by a graduate fellowship from the University of Toledo Health Science Campus.
REFERENCES


15, 312-320.


pleiotropic transcriptional profiles: a case study of DNA gyrase inhibition.

PLoS Genet, 2, e152.


Figure 1. Sequences of selected Lrp proteins. Lrp proteins from various bacterial species were aligned; species used in this study are underlined (Escherichia coli, Proteus mirabilis, and Vibrio cholerae). A more complete list of Lrp orthologs and paralogs can be found at: http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?pid=16128856. The lightly/yellow-shaded regions indicate N- and C-terminal sequences that are conserved among enterobacterial Lrp orthologs, and the darker/blue-shaded regions indicate substitutions relative to E. coli. The boxed regions indicate the DNA-binding helix-turn-helix motif (top portion, under cartoon representation), and the leucine-binding sites (lower portion of sequence); all boxed regions are completely conserved among the species used in this study. For references, see main text. Other Lrp orthologs shown came from (in order shown): Salmonella enterica serovar Typhimurium, Serratia marcescens, Klebsiella pneumoniae, Yersinia pestis, Haemophilus ducreyi, Haemophilus influenzae Rd, Actinobacillus pleuropneumoniae, Pasteurella multocida, Pseudomonas putida, and Pseudomonas aeruginosa.
**Figure 2. Effects of lrp null mutation on growth rates.** Growth rates were determined from a fit to the exponential portion of the growth curve, extending in all but one case (*P. mirabilis*, glucose minimal medium) through at least four mass doublings. Open symbols refer to growth in MOPS glucose plus required supplements (nicotinate, panthothenate and thiamine; see Methods), while closed symbols represent growth in MOPS glucose defined-rich medium. **A.** *lrp* vs. *lrp* + growth rates. The values shown are the specific growth rate constants, \( k \), calculated as \( \ln 2 / \text{(doubling time, in h)} \). For comparison, \( k \) values of 0.5, 1, and 2 correspond respectively to doubling times of 83, 42, and 21 min. The rich medium results are clustered and therefore not labeled; for the minimal medium, the abbreviations used are Eco (*E. coli*), Pmi (*P. mirabilis*), and Vch (*Vibrio cholerae*). The diagonal line shows where points should fall if *lrp* mutation has no effect on the growth rate in these media. **B.** Complementation of the low *P. mirabilis* growth rate in the glucose minimal medium described in (A). The dashed lines indicate growth data for the *P. mirabilis* *lrp* mutant (open circles; 193 min doubling time) and the mutant bearing the vector control (gray circles; 191 min). Remaining lines show the WT *P. mirabilis* (closed circles; 66 min); and the *lrp* mutant bearing plasmids with the *lrp* + genes from *P. mirabilis* (triangles; 69 min), *V. cholerae* (squares; 81 min), or *E. coli* (diamonds; 81 min).
Figure 3. Regulation of selected target genes by heterologous Lrp proteins. E. coli strains, all carrying \textit{lrp-Tn10} and \textit{Δlac}, were transformed with plasmids carrying various \textit{lrp} alleles (or vector control). Transformants were grown in unsupplemented MOPS glucose medium. **A.** Western blot analysis of Lrp accumulation (Eco, \textit{E. coli} Lrp; Pmi, \textit{P. mirabilis} Lrp; Vch, \textit{V. cholerae} Lrp; pCC1, vector control) using polyclonal antiserum raised against \textit{E. coli} Lrp. The arrow indicates the direction of electrophoresis. **B-D.** \textit{Plrp-lacZ} (B), \textit{PgltB-lacZ} (C) and \textit{PlivK-lacZ} (D) activity were measured via ONPG hydrolysis, and plotted vs. culture density to ensure that the cultures were in balanced growth. The Lrp orthologs used are from \textit{P. mirabilis} ▲ and \textit{V. cholerae} ■, as well as the \textit{E. coli} positive control ● and the vector control u. **E-G.** Isoleucine, Leucine and Valine was added to the medium (“+Leu”) for experiments depicted in the lower panels: \textit{Plrp-lacZ} (E), \textit{PgltB-lacZ} (F) and \textit{PlivK-lacZ} (G). The correlation coefficients for all least squares fits to the data were at least 0.97.
Figure 4. Genome-wide comparison of transcriptional effects of three Lrp orthologs. The Venn diagram shows subsets of *E. coli* genes that were differentially regulated in response to Lrp orthologs from the indicated species. Gene expression was assessed by two-color microarray analysis as described in Methods. The pie chart represents the relative distribution of genes responsive to the *E. coli* Lrp that are also responsive to the other Lrp orthologs.
Genes showing CHANGED expression with indicated Lrp

Eco only

Eco+Pmi +Vch

Eco+Vch only

Eco+Pmi only

Pmi

Vch

Eco

321 (214)

200

544 (382)

165

61

273 (75)
Figure 5. Extent of regulatory conservation for significantly increased or decreased targets. In every panel, the x-axis shows the gene expression ratio for *E. coli* K-12 *lrp-Tn10* relative to that in the same strain complemented by *E. coli lrp* (pEcoLrp). The y-axes indicate the equivalent ratio, where the complementation is by vector alone (pCC1) or the *lrp* alleles from *V. cholerae* (pVchLrp) or *P. mirabilis* (pPmiLrp). Full complementation relative to that by pEcoLrp would yield a slope of 1.0 for the linear fit. **A.** This column shows responses of the gene set yielding statistically-significant increases in expression associated with *lrp* mutation, as reflected by an expression ratio significantly above 1.0 on the x-axis. This set includes genes that are repressed (directly or indirectly) by *E. coli* Lrp. **B.** This column includes the set of genes showing significant decreases in expression associated with *lrp* mutation, indicating direct or indirect activation by *E. coli* Lrp. Correlations between transcript abundances of significantly down-regulated genes in two sets (top: ratios of sets as in A, 22 genes); (middle: ratios of sets as in A, 81 gene); (bottom: ratios of sets as in A, 105 genes). **C.** Third is the set of 57 genes recognized in RegulonDB (10,44) as being directly controlled by Lrp (column C), whether the control is positive or negative. This set includes genes that are controlled by Lrp, but not under the growth conditions used by us, so the cluster of genes showing no effect of Lrp is not surprising (39,42). Correlation between transcript abundances of designated (by RegulonDB) Lrp targets, 57 genes (top, middle and bottom panels as in A). The relative transcript abundances were estimated from at least three independent biological replicas using a linear model similar to one introduced
before (83,84). Significantly expressed genes were identified at a fixed false
discovery rate of 5% at the 90\textsuperscript{th} percentile (56). The lists of differentially affected
genes are provided as supplementary materials, and the list of 57 Lrp targets is
included in the Methods section.
Figure 6. Effect of heterologous Lrp proteins on the swarming phenotype of *P. mirabilis*. *Proteus mirabilis* wild-type (A) or *lrp* null strains (B) were grown in LB medium. Overnight cultures were spotted (2 µl) onto triplicate 1.5% agar LB plates. After 12 h at 37 °C, plates were photographed under normal illumination. At the same time, transformants of the *lrp* mutant strain were assayed in parallel using the same methods. These strains contained pCC1 vector (C), or plasmids carrying *lrp*<sup>+</sup> alleles from *P. mirabilis* (D), *E. coli* (E), or *V. cholerae* (F). Panel G shows the results of measurements (average ± standard error) from the center of each colony to the inner (○) and outer (●) edges of the growth rings.
Figure 7. Lrp protein levels as a function of growth. Wild-type strains of *E. coli*, *P. mirabilis* and *V. cholerae* were grown in MOPS glucose plus nicotinate or MOPS defined-rich media. Growth curves (A-C, MOPS glucose medium; G-I, MOPS defined rich medium) and Lrp protein levels (D-F, glucose; J-L, rich) are shown. Equal amounts of total protein were loaded in each lane, and a standard curve of purified *E. coli* Lrp was included on each gel for quantitation. The data from two independent experiments (●, ○) are shown. M shows the amounts of Lrp (ave ± SE) relative to the *E. coli* value for each medium (open, MOPS glucose; closed, MOPS defined rich). These were calculated from the full set of western blot datapoints in each panel (D-F and J-L), normalized to the *E. coli* result for that medium (panel D or J). N shows a comparative western blot. Cell pellets were boiled and equal amounts of total protein from *E. coli* (Ec) and *P. mirabilis* (Pm) were resolved by SDS polyacrylamide gel electrophoresis. The subsequent blot was probed with polyclonal antiserum raised against *E. coli* Lrp.
Figure 8. Variation of \textit{lrp} mRNA levels with growth phase and medium.

\textbf{A.} Values shown are arbitrary units from standard curve-based QRT-PCR (see Methods), with bars indicating standard errors. At least four points from each of two independent experiments were used to generate each plotted value. Arrows point from the logarithmic phase values to the stationary phase values in each case. The diagonal line indicates the position of points expected if growth medium had no effect on mRNA levels. \textbf{B-D.} Direct comparison of \textit{lrp} mRNA levels. Samples from early log, mid log, late log and stationary phases were isolated from cultures of wild-type \textit{E. coli} and \textit{P. mirabilis} growing in MOPS defined rich medium. \textbf{B.} A baseline amount of total \textit{E. coli} cDNA (from a mid-log phase culture in MOPS glucose plus nicotinate) was mixed with varying amounts of test cDNA (all from log-phase cultures) from glucose (○) or rich (●) cultures. The mixes were used as template for simultaneous amplification with one (\textit{E. coli}) or two primer pairs. If the test cDNA preparation has the same proportion of \textit{lrp} cDNA as the reference pool, the detected amount of \textit{lrp} cDNA should rise with a slope of 1.0 (actual \textit{vs.} detected, based on the varied amounts of test cDNA added); this is shown as a dotted line in each panel.
Figure 9. Sequences upstream of $adhE$, $gltB$ and $lrp$ orthologs. In each case, the sequence ends with the initiation codon. Lrp-binding sites and the transcriptional +1 position are known for $E. coli$ K-12 (85). Demonstrated Lrp binding sites are in underlined lowercase italics, and the -35 and -10 sequences inferred from the known +1 position (for $E. coli$) are boxed. Putative binding sites, predicted by the PRODORIC virtual footprinter (73) are shaded, and the match scores for predicted sites are shown to the right. For $E. coli$ $PgltB$, one of the predicted sites overlaps an actual site, and gives a particularly high match score, though an overlapping actual site in $Plrp$ does not. $V. cholerae$ has two nearly-tandem copies of the $gltBD$ operon on chromosome I. The 5’-most $gltB$ ortholog (“Vch1”, locus tag Vc2373) is 43% identical to Eco $gltB$, while the 3’-most ortholog (“Vch2”, Vc2376) is 73% identical to $gltB$ in amino acid sequence.
Figure 10. Regulation of orthologous target genes in native backgrounds.

Samples were isolated at an OD$_{600nm}$ of 0.3 (log), as well as 1 h after linear growth stopped (stationary), from *E. coli*, *P. mirabilis* and *V. cholerae* wild-type and *lrp* cultures growing in MOPS defined rich medium. QRT-PCR was used to determine the relative levels of *adhE*, *gltB* and *recA* messages, with *recA* serving to provide a Lrp-independent baseline. The experiment was performed in triplicate and the level of message was determined using the standard curve method and normalization to *recA*. A – *adhE*. B – *gltB*. For each plot filled symbols represent log phase levels and open symbols represent stationary phase levels. The symbol shapes indicate the species: *P. mirabilis* (△, ▲), *E. coli* (○, ●), or *V. cholerae* (□, ■). The line indicates the position for data if no effect of Lrp is seen (ratio of 1); points above the line are consistent with repression, while those below the line are consistent with activation by Lrp.
Figure 11. Lrp effects on orthologous promoter regions in three backgrounds. The orthologous P_{lrp} or P_{gltB} regions were amplified from E. coli, P. mirabilis, and V. cholerae and inserted upstream of a promoterless lacZ gene. These plasmids were then used to transform lrp / lrp^{+} strain pairs of all three species, and lacZ was measured vs. culture density to obtain the slopes, and for each least squares fit the correlation coefficient was at least 0.97. The standard error (bars) of each slope was calculated from the residuals using the “summary(lm(y~x))” function from the R statistical package. All strains were grown in MOPS glucose medium supplemented with nicotinate, pantothenate, thiamine, methionine and cysteine (see Methods). A – P_{lrp}-lacZ. B – P_{gltB}-lacZ. Symbols used are indicated in the figure, and are a superset of those used in Fig. 8. The line indicates the position for data if no effect of Lrp is seen (ratio of 1); points above the line are consistent with repression, while those below the line are consistent with activation by Lrp.
TABLE 1. Bacterial strains, plasmids, and oligonucleotide primers used.

<table>
<thead>
<tr>
<th>Strains</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli</em> W3110</td>
<td>F− prototroph</td>
<td>F.C. Neidhardt</td>
</tr>
<tr>
<td>E. coli BE10.2</td>
<td><em>W3110 Δlac-169 lrp35::Tn10</em></td>
<td>R.G. Matthews</td>
</tr>
<tr>
<td>E. coli BE3780</td>
<td><em>lrp::Tn10, Δlac-169, gltB(ψQ35)::lacZ</em></td>
<td>R.G. Matthews</td>
</tr>
<tr>
<td>E. coli EPI300</td>
<td>F− *mcrA Δ(mrr-hsdRMS-mcrBC)<em>Φ80dlacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ− rpsL nupG trfA dhfr</em></td>
<td>R.G. Matthews</td>
</tr>
<tr>
<td>PS2209</td>
<td><em>W3110 Δlac-169</em></td>
<td>F. C. Neidhardt</td>
</tr>
<tr>
<td>MG1655</td>
<td>F− <em>lambda- ilvG- rfb-50 rph-1</em></td>
<td>ATCC 700926</td>
</tr>
<tr>
<td>MG1655 Δlrp</td>
<td>F− <em>lambda- ilvG- rfb-50 rph-1lrp::Kan</em></td>
<td>B. Martinez-Vaz and ABK</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(unpublished data)</td>
</tr>
<tr>
<td><em>Proteus mirabilis</em> HI4320</td>
<td>WT</td>
<td>H.L. Mobley</td>
</tr>
<tr>
<td>P. mirabilis U6450</td>
<td>WT</td>
<td>G.M. Fraser</td>
</tr>
<tr>
<td>P. mirabilis</td>
<td><em>lrp::Tn5</em></td>
<td>G.M. Fraser</td>
</tr>
<tr>
<td>U6450Δlrp</td>
<td>VT</td>
<td>G.K. Schoolnik</td>
</tr>
<tr>
<td><em>Vibrio cholerae El tor</em> A1552</td>
<td><em>lrp-cat</em> derivative of strain A1552</td>
<td>G.K. Schoolnik</td>
</tr>
<tr>
<td>V. cholerae El tor A1552Δlrp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmids</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pCC1</td>
<td>Low copy blunt cloning vector</td>
<td>Epicentre</td>
</tr>
<tr>
<td>pECLRP</td>
<td>pCC1 backbone with <em>lrp</em> gene from <em>E. coli O157:H7</em> inserted at BamHI site</td>
<td>This work</td>
</tr>
<tr>
<td>pPMLRP</td>
<td>pCC1 backbone with <em>lrp</em> gene from <em>P. mirabilis HI4320</em> inserted at BamHI site</td>
<td>This work</td>
</tr>
<tr>
<td>pVCLRP</td>
<td>pCC1 backbone with <em>lrp</em> gene from <em>V. cholerae El tor A1552</em> inserted at BamHI site</td>
<td>This work</td>
</tr>
<tr>
<td>pVEC</td>
<td>pCC1 backbone with Kan cassette cloned</td>
<td>This work</td>
</tr>
</tbody>
</table>
pRLIV2 into BamHI site; used as vector control. pACYC backbone with *livK-lacZ* transcriptional fusion. CAT gene was inactivated by digesting with NcoI, in-fill with Klenow fragment and religated to create a frame shift

pECKAN pECLRP with CAT gene deleted and KAN cassette inserted at BsmI

pPMKAN pPMLRP with CAT gene deleted and KAN cassette inserted at BsmI

pVCKAN pVCLRP with CAT gene deleted and KAN cassette inserted at BsmI

pVEC2 pVEC with CAT gene deleted and KAN cassette inserted at BsmI

pPM2005 pBH403 backbone with *E.coli O157* *gltB* promoter cloned at BamHI to SalI

pPM2007 pBH403 backbone with *Proteus mirabilis* *gltB* promoter cloned at BamHI to SalI

pPM2008 pBH403 backbone with *Vibrio cholerae* Vc2376 *gltB*-2 promoter cloned at BamHI to SalI

pPM2009 pBH403 backbone with *Vibrio cholerae* Vc2373 *gltB*-1 promoter cloned at BamHI to SalI

pPM3001 pBH403 backbone with *E.coli W3110 lrp* promoter region cloned at BamHI to SalI

pPM3003 pBH403 backbone with *Proteus mirabilis lrp* promoter cloned at BamHI to SalI

pPM3006 pBH403 with *Vibrio cholerae lrp* promoter cloned at BamHI to SalI

pPM3001Chl pPM3001 CAT gene was inactivated by digesting with NcoI, in-fill with Klenow and religated to create a frame shift

<table>
<thead>
<tr>
<th>Primers – <em>lrp</em> genes</th>
<th>(consensus rbs underlined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>eclrp1</td>
<td>GAAGGAGATATACATATGGTAGATAGC</td>
</tr>
<tr>
<td></td>
<td>AAGAAGCG</td>
</tr>
<tr>
<td>eclrp2</td>
<td>TTAGCGCGTCTTAATAACCAGACGATT</td>
</tr>
<tr>
<td></td>
<td>ACTC</td>
</tr>
<tr>
<td>pmlrp1</td>
<td>GAAGGAGATATACATATGGTAGATAGC</td>
</tr>
<tr>
<td></td>
<td>AAAAAACGTC</td>
</tr>
<tr>
<td>pmlrp2</td>
<td>TTAGCGGTGTATTTAATCACTAGGCC</td>
</tr>
<tr>
<td>vclrp1</td>
<td>GAAGGAGATATACATATGGTAGATAGT</td>
</tr>
<tr>
<td></td>
<td>TATAAAGAAGC</td>
</tr>
<tr>
<td>vclrp2</td>
<td>TTAGCGAGTTTTAATCACTAATTGGGT</td>
</tr>
<tr>
<td>Primers – <em>gltB</em> promoters</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>O157gltb1</td>
<td>CGGGATCCCATACGACATACACTTTTGA</td>
</tr>
<tr>
<td>O157gltb2</td>
<td>ATCG</td>
</tr>
<tr>
<td>pmgltb1</td>
<td>CGGGATCCCGGTATTTTATAAAAAAAAT</td>
</tr>
<tr>
<td>pmgltb2</td>
<td>TTATAGAGGTA</td>
</tr>
<tr>
<td>vc2376-1</td>
<td>CGGGATCCAGCAGCTCTGCTGCC</td>
</tr>
<tr>
<td>vc2376-2</td>
<td>ACGCGTCGAGTTGATGAGGATATCCTCC</td>
</tr>
<tr>
<td>vc2373-1</td>
<td>CGGGATCCATTTCGTAACAAAAATTTTA</td>
</tr>
<tr>
<td>vc2373-2</td>
<td>G</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Primers – <em>lrp</em> promoters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ecolrp+69-1</td>
<td>CGGGATCCCTGACGATTTTGTTAACAAAT</td>
</tr>
<tr>
<td>ecolrp+69-2</td>
<td>TTGTGCAA</td>
</tr>
<tr>
<td>pmlrp+69-1</td>
<td>ACGCGTCGACACTAGCACGGTAAACA</td>
</tr>
<tr>
<td>pmlrp+69-2</td>
<td>TGTTATTTAC</td>
</tr>
<tr>
<td>Vc lrpfull-1</td>
<td>CGGGATCCCTCTCAGCAGGAAAAAAGCCC</td>
</tr>
<tr>
<td>Vc lrpfull-2</td>
<td>G</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Primers – QRT-PCR</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>erart1</td>
<td>ATGGGTATCGACGAAAAAC</td>
</tr>
<tr>
<td>erart2</td>
<td>GGTTTTACGGGAAAGATTTCC</td>
</tr>
<tr>
<td>prart1</td>
<td>ATGGGTATCGACGAAAAAACAACAGG</td>
</tr>
<tr>
<td>prart2</td>
<td>GAAGATTTGACGGCCATAG</td>
</tr>
<tr>
<td>vrart1</td>
<td>ATGGACGAGAATAAACAAGAAG</td>
</tr>
<tr>
<td>vrart2</td>
<td>GTGTCGTTTTACGCGAAGGG</td>
</tr>
<tr>
<td>elrprt1</td>
<td>ATGGGTAGATAGCGAAAAAGGC</td>
</tr>
<tr>
<td>elrprt2</td>
<td>GATGATCCAGAAGATAGGG</td>
</tr>
<tr>
<td>plrprt1</td>
<td>ATGATTTGATAATAAAAACGTCCG</td>
</tr>
<tr>
<td>plrprt2</td>
<td>GGTTTTAACAGCGCAGATTA</td>
</tr>
<tr>
<td>vlrprt1</td>
<td>ATGGGTGAGATATTAAAGAAGGC</td>
</tr>
<tr>
<td>vlrprt2</td>
<td>CTGAGGGTATACAATGCA</td>
</tr>
<tr>
<td>Gene</td>
<td>Sequence</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>eadh1</td>
<td>ATGGCTGTTACTAATGTCGCT</td>
</tr>
<tr>
<td>eadh2</td>
<td>GGTAAAAATCCTTTATCTTCA</td>
</tr>
<tr>
<td>padh1</td>
<td>AACGTTACTGAACCTGATCT</td>
</tr>
<tr>
<td>padh2</td>
<td>ATTTGAGATATATTCTGAAAGCAAG</td>
</tr>
<tr>
<td>vadh1</td>
<td>ATGCCCTGTTAAATATCGCC</td>
</tr>
<tr>
<td>vadh2</td>
<td>ATAAACTCAGAAGCAAGGCATGGT</td>
</tr>
<tr>
<td>eglt1</td>
<td>ATGACACGCAAAACCAGCG</td>
</tr>
<tr>
<td>eglt2</td>
<td>CGCAGTCTACCTTGTGGCTAGG</td>
</tr>
<tr>
<td>pglt1</td>
<td>TGCTCTATCAAAAAACAGCGG</td>
</tr>
<tr>
<td>pglt2</td>
<td>CTGAAAAAACGCGCTTGGG</td>
</tr>
<tr>
<td>vglt1</td>
<td>ACTCAGTGCGTTGCCGTC</td>
</tr>
<tr>
<td>vglt2</td>
<td>TCACAGCCTGACCGCC</td>
</tr>
</tbody>
</table>
A Novel Method for Isolating and Identifying Promoter-Associated Proteins

Robert E. Lintner¹, Elmar Weinhold³ and Robert M. Blumenthal²

¹Department of Medical Microbiology and Immunology, University of Toledo Health Sciences Campus, Toledo, OH 43614-5806, USA
²Program in Bioinformatics and Proteomics/Genomics, University of Toledo Health Sciences Campus, Toledo, OH 43614-5806, USA
³Institute of Organic Chemistry, RWTH Aachen University, Landoltweg 1, D-52056 Aachen, Germany

2007

Prepared for Molecular Microbiology
Summary

Identifying transcription factors and their cognate binding sequences is essential for understanding the regulation of a specific gene, and also for improving the reliability of regulatory network models. Current methods for detecting DNA-protein interactions in vivo generally require an antibody to the regulator of interest, so deduction of even simple regulatory complexes can require the cloning, expression and purification of multiple candidate regulators. In vitro methods, in addition to being time consuming, lack the benefit of being specific for a particular in vivo condition. We have developed a novel in vivo, relatively high-throughput method to isolate regulatory proteins that bind to a promoter of interest under different environmental conditions. This method reduces the discovery time of regulators while facilitating downstream applications for confirming direct binding and determining sequence specificity. Promoter sequences are cloned into a plasmid upstream of affinity isolation cassettes. After in vivo crosslinking and isolation of plasmid DNA by electrophoresis promoter DNA-protein complexes are purified, crosslinks are reversed and proteins are identified using mass spectrometry analysis. This method has been tested using the well-characterized regulatory region from the E. coli glutamate synthase (gltB) gene.
Introduction

Productive interactions between DNA and protein are essential for transcription (Echols, 1986; Von Hippel and McGhee, 1972), and the focus of studies from many different fields has been to elucidate the repertoire of proteins associated with gene regulatory regions (Aiba et al., 2006; Herring et al., 2005; Ren et al., 2000; Whitelaw, 2006). Determining the set of regulators that controls a specific gene is important not only for defining the gene’s position within the regulatory framework of an organism (Tan et al., 2005), but also for highlighting the similarities or distinctions between the regulatory programs of different species or cell types. In addition, the recognition sequences elucidated from these essential experimental analyses contribute to the pool of references needed to build useful computational models of gene networks (Babu et al., 2004; Gelfand, 1999; Gonzalez et al., 2005; Huerta et al., 1998; Perez-Rueda and Collado-Vides, 2000).

Identifying the regulatory proteins associated with a particular promoter can be complicated by the fact that environmental conditions leading to differences in promoter occupation are difficult to model in vitro (Tan et al., 2005). In an attempt to overcome this, many studies now incorporate chromatin immunoprecipitation (ChIP) with expression data from microarray experiments (Hudson and Snyder, 2006; Zhou and Yang, 2006). While this method is effective, it still requires that the identity of at least one regulator be known in advance to immunoprecipitate crosslinked complexes. Efforts to increase the
applicability by epitope tagging different regulators critically reduce the throughput due to the time consuming tasks of cloning, expression and purification.

Other in vitro methods for purifying transcription factors such as DNA affinity chromatography (Gadgil et al., 2001a; Gadgil et al., 2001b), electrophoretic-mobility shift based approaches (Woo et al., 2002) and protein chip methods (Paul et al., 2007) can be extremely labor intensive, and lack the added benefit of physiological information that comes from doing experiments in vivo.

We have developed a novel, relatively high-throughput method for isolating regulatory proteins crosslinked in vivo to their recognition sequences. Our approach eliminates the requirement of a known regulator for immunoprecipitations while allowing for the identification of transcription factors associated with a promoter under different environmental conditions.

Figure 1 provides an overview of the four basic steps in this plasmid-based method. First, the promoter region of interest is cloned onto a species-compatible plasmid containing one of three isolation cassettes (see results). Second, after transforming or transfecting cells with the plasmid they are crosslinked in vivo under the desired conditions, and cells are lysed under non-denaturing conditions. Third, crosslinked plasmid containing the promoter of interest with attached regulators is separated via gel electrophoresis. Finally, promoter regions containing crosslinked proteins are excised from the plasmid by digesting with Ascl and purified via the isolation cassette (see text). Regulatory
proteins are then identified via mass spectrometry, and confirmatory analyses such as electrophoretic mobility shifts (EMSA), DNase I footprinting and analysis of mutants for the regulator or promoter. The goal is to be able to determine rapidly the set of regulators associated with any given promoter, without having to know any of the proteins in advance.
Results

**Gentle lysis of crosslinked cells.** The utility of the method discussed in this study lies in the preferential isolation of a specific promoter crosslinked to the regulatory proteins that control it under a particular condition. The robustness of this method is, therefore, partially dependent upon an efficient non-denaturing lysis protocol (Fig. 1, step 1). There are several good lysis buffers currently on the market, but we found a modified cleared lysate method (Clewell and Helinski, 1969) to be the most efficient (data not shown), and the most versatile as it allowed for downstream applications such as restriction digests without the need for additional yield-reducing buffer exchange or desalting steps (Fig. 2B).

We next determined whether or not plasmid DNA-protein complexes crosslinked *in vivo* by formaldehyde are electrophoretically resolvable. *E. coli* cells containing pRLTG were exposed to formaldehyde for increasing amounts of time, then loaded onto agarose gels. Our results indicate that the amount of resolvable plasmid is inversely proportional to the crosslinking time (Fig 2C). This problem highlights the need for a plasmid DNA-protein complex specific isolation strategy.

**Isolation of plasmid DNA-protein complexes via isopycnic ultracentrifugation.** A study by Clewell *et al.* (Clewell and Helinski, 1969) demonstrated that supercoiled circular DNA-protein complexes can be isolated from whole cell lysates even in the absence of crosslinking. To address the applicability of the gradient approach we sought to simply isolate plasmid DNA
crosslinked to non-specific DNA binding (histone-like) proteins such as HU and Dps (Ishihama, 1999; Swinger and Rice, 2007). We compared the pGFP-associated proteins from *E. coli* and *F. tularensis* (Fig. 3B, and Table 1). For *F. tularensis* all of the enriched protein species came from the mid portion of the gradient, and were specific to the crosslinked culture. For *E. coli* most of the enriched species were also from the mid portion, but were specific to the uncrosslinked culture. Mass spectrometry analysis of selected isolates from *F. tularensis* (Fig. 3B, boxed regions, and Table 1) revealed a potential non-DNA associated protein-protein interaction (SOD and NDK Chen *et al.*, 2003), and two chaperon like proteins (GroEL, ES) that are known to associate with the transcriptional repressor HrcA in *Campylobacter* and *Chlamydial* species for the purpose of regulating their own genes as part of the heat shock response (Susin *et al.*, 2004; Wilson *et al.*, 2005) (Table 1, a,c,e). While this fits in the sense that there may be a DNA association we would not expect to find this complex bound to pGFP. HrcA does not appear to have an ortholog in *F. tularensis*. The other proteins identified do not have any clear or described association direct binding to DNA.

For *E. coli* it is not surprising that we detected protein in the un-crosslinked culture (Clewell, 1969 #2), but the absence of enrichment of even these proteins in the crosslinked sample is cause for concern. Of the proteins identified by mass spectrometry none have any known association with DNA binding (Table 1). One puzzling result, also observed in *F. tularensis*, was the comigration of proteins with large differences in molecular weight (Table 1d, CHMS and 1a, HPI).
Cleavage of these proteins in vivo has not been demonstrated, so it is possible that they may be more sensitive to the heating process used for crosslink reversal.

It is unclear why the crosslinking process seemed to work for *F. tularensis* but was ineffective in *E. coli*. Inefficient formation of crosslinks is one possibility, but we would expect to see more protein bands from the top portion of the gradient (Fig. 3B, lanes 7 and 8) unless, of course, the un-crosslinked proteins are more sensitive to proteolysis and degradation. Incomplete crosslink reversal may also be to blame, although we typically observe only one or two large bands of protein when this occurs (data not shown). Whatever the cause the bottom line is that we did not detect any protein species directly associated with DNA binding for either organism (Table 1). In addition we were unable to resolve distinct plasmid bands via agarose gel electrophoresis (Fig 3, top panels). These results indicate that ultracentrifugation is not, at least under the conditions tested, the ideal first step for enriching plasmid DNA-protein complexes.

**Isolation of plasmid DNA-protein complexes via agarose gel electrophoresis.** The results in Figure 2A show that crosslinked plasmid DNA-protein complexes do resolve reasonably well on agarose gels, albeit with fairly low yields. We next developed a protocol to separate crosslinked plasmid-protein complexes from chromosomal DNA and RNA contamination. Details of the method are in Experimental Procedures. Briefly, *E. coli* cells containing pRLTG (bearing PgltB) were treated with the crosslinker formaldehyde, then lysed and desalted. Concentrated lysates were loaded onto 1% agarose gels (Fig. 4A) and
electrophoresed in sodium borate buffer (Brody and Kern, 2004). Plasmid bands (Fig. 4A, boxed regions) were excised and heated to 98°C in SDS loading buffer containing β-mercaptoethanol to reverse crosslinks and melt agarose. Samples were then quickly (before agarose solidified) loaded onto a 12% SDS-PAGE, and gels were subsequently stained with Coomassie blue (Fig. 4B).

A qualitative analysis of the SDS-PAGE result (Fig. 4B) reveals that identical protein species were isolated regardless of whether cells were exposed to formaldehyde, which is not surprising given the results of Clewell and Helinski nearly 40 years ago (Clewell and Helinski, 1969).

Regulation of the gltB promoter has been well studied in E. coli (Ernsting et al., 1993; Paul et al., 2001; Paul et al., 2007; Wiese et al., 1997), and it is known to bind integration host factor (IHF), the leucine-responsive regulatory protein (Lrp), the catabolite repressor protein (Crp) and the arginine repressor (ArgR). Therefore we would expect to find stained bands that represent the molecular weights of these proteins. Using retardation factor (Rf) analysis to estimate molecular weight, we identified a number of protein candidates (Fig. 4B-C) expected to be, 1) non-specifically associated with plasmid DNA (Dps and Fnr), 2) specific to the gltB promoter (Lrp and Crp), and 3) parts of a holoenzyme complex (α-RNAP and RpoS). These results are promising, though not yet confirmed by mass spectrometry, and suggest that preferential isolation of plasmid DNA-protein complexes via agarose gel electrophoresis may work well as an initial step in isolating regulators associated with a specific promoter.
**Enrichment of promoter DNA-regulatory protein complexes.** We have shown that the isolation of plasmid DNA-protein complexes is possible via agarose gel electrophoresis even in the absence of crosslinking; however, in order to minimize contamination by proteins not specifically associated with the regulatory region of interest, and to increase robustness (i.e. ability to identify low abundance regulators) requires a second purification to enrich for promoter DNA-regulator complexes. We have developed three independent methods for isolating these complexes using a plasmid based approach.

**Isolation via streptavidin binding DNA aptamer.** DNA and RNA aptamers are used for technological applications such as nanomotors (Dittmer *et al.*, 2004; Nutiu and Li, 2005a), and molecular beacons (Nutiu and Li, 2005b; Rupcich *et al.*, 2005), but were initially discovered (Klug and Famulok, 1994) for their ability to bind organic dyes (Ellington and Szostak, 1990), and T4 DNA polymerase (Tuerk and Gold, 1990). The specificity of DNA and RNA aptamers derives from the secondary and tertiary structures they form in solution (Kato *et al.*, 2000; Macaya *et al.*, 1993; Nobile *et al.*, 1998).

We cloned a 40 nucleotide DNA oligonucleotidenucleotide (Fig. 5A) developed to bind streptavidin with nanomolar affinity (Bittker *et al.*, 2002) into a high copy vector (pUC19) downstream of a unique NotI cloning site. The NotI site is used for inserting different promoter regions to be tested. The NotI-aptamer region (Fig. 5B) is flanked by Ascl sites; Ascl has an eight base recognition sequence made up entirely of guanine and cytosine residues, a feature that
makes these sites less abundant in organisms that lack GC-rich genomes (~40
sites in *E. coli* K-12, Dr. J. Posfai, REBASE).

The design of this plasmid (pRLSA) allows cloned promoter regions and
their crosslinked proteins to be isolated via digestion with Ascl, and binding of the
DNA aptamer by streptavidin (Fig. 5B). Purified promoter DNA-protein complexes
can then be subjected to crosslink reversal, and regulators are identified via
mass spectrometry analysis of tryptic digests.

To determine the parameters necessary for a tight and efficient binding
interaction between the DNA aptamer and streptavidin we cloned the *gltB*
promoter into the NotI site of pRLSA. An 800bp *P*_{gltB}-aptamer fragment was
generated and ^32^P labeled (see experimental procedures) via PCR. Purified
fragments were heated to 95°C for 5 min to melt aptamer strands, and rapidly
cooled to promote the formation of the streptavidin binding stem-loop structure
(Bittker *et al.*, 2002).

The results from our gel shift assays (Fig. 6A, B) were inconsistent with
apparent binding occurring at different nanomolar ratios of streptavidin and
MgCl\textsubscript{2}. The aptamer-streptavidin interaction occurs when 45-75 nM streptavidin
is present along with 30 mM MgCl\textsubscript{2}, and adding MgCl\textsubscript{2} to 40 mM was inhibitory
(Fig. 6A). Figure 6B shows that concentrations of streptavidin between 70-85 nM
are effective, and only when used with 30 mM MgCl\textsubscript{2}. Bittker *et al.* used 5 mM
MgCl\textsubscript{2} in their binding assays, but we did not detect binding at this concentration.
One concern is that the melting and snap-cooling process (even if not inhibited
by crosslinking) is not efficient enough. This would reduce the effective
concentration of aptamer molecules available for binding, and might account for some of the variability seen in the nanomaolar concentrations of streptavidin needed for binding. Bittker et al. also found (data not shown) that the aptamer-streptavidin interaction could be blocked by adding excess biotin suggesting that the aptamer and biotin bind in similar places. We added 200µM D-biotin, and found that biotin appeared to block binding only when 75nM streptavidin and 30mM MgCl₂ were present.

Overall the gel shift results indicate that the DNA aptamer approach could work, but that problems with the formation of the stem-loop structure might greatly reduce the efficiency such that only those regulators present in higher concentrations would be detectable.

**Isolation via TetR-tetO.** The tetracycline repressor (TetR) protein has a high affinity and specificity for its operator sequences, tetO₁ and tetO₂ (Hillen and Berens, 1994), and is frequently used to control gene and transgene expression (Berens and Hillen, 2003; Sprengel and Hasan, 2007; Zhu et al., 2002) in eukaryotes. We have cloned tetR in-frame with a C-terminal hexahistidine tag, with the intent of using the purified protein as a reagent to isolate crosslinked promoter DNA-regulatory protein complexes. Tandem tet operator (Berens and Hillen, 2003) cassettes were cloned in place of the aptamer sequence (see experimental procedures) downstream of the site used for promoter insertion (NotI, Fig. 7B). After isolation by agarose gels (see above) the crosslinked promoter DNA-regulatory protein complexes are purified in a three step process: first, the in vivo crosslinked plasmids are digested with Ascl; second, TetR is
added, and finally the complexes (TetR-promoter-regulators) are
immunoprecipitated using a TetR antibody or isolated on a nickel column.

For this method to be efficient we need significant quantities of pure
hexahistidine- tagged TetR. We are able to produce large amounts of TetR via
IPTG induction in strain BL21 DE3 (Fig. 7C), but almost all of the synthesized
protein was found in the insoluble fraction. The TetR is apparently being shuttled
into inclusion bodies even in the presence of ethanol (2.5%), and at lower culture
temperatures (de Marco and De Marco, 2004; Steczko et al., 1991) (Fig. 7D).

Isolation via SMILing DNA. Sequence specific Methyltransferase-Induced
Labeling of DNA is a novel method developed for covalently attaching
fluorescent reporters, biotin and other functional molecules to DNA using
aziridine co-factors (Dalhoff et al., 2006; Pljevaljcic et al., 2003; Pljevaljcic et al.,
2004a; Pljevaljcic et al., 2004b). The specificity of labeling makes this method
ideal in terms of our objective, which is to isolate unknown regulators bound to a
promoter from the other proteins in crosslinker treated cells. We generated a
plasmid containing a cassette with five equally-spaced M.BseCI recognition
sequences cloned downstream of the NotI, promoter insertion site, (Fig. 8A).
M.BseCI is one of the methyltransferases capable of using the aziridine
substrates. Following isolation of crosslinked plasmid by agarose gel
electrophoresis (see above) promoter DNA-regulatory protein complex is purified
in three steps. First, Ascl digestion is used to liberate promoter DNA-protein
complex from the rest of the plasmid; second, M.BseCI sites on the affinity DNA
fragment are labeled with biotin (see experimental procedures); third, biotin-
promoter-protein complexes are purified using streptavidin sepharose beads (Fig. 8B).

We first tested the SMI Ling DNA protocol in a purified system using pBR322. After linearizing with R.PstI and labeling reactions (see experimental procedures), the DNA was digested with R.ClaI, which shares the same recognition sequence as M.BseCI. Under these conditions, as judged by the protection from R.ClaI, we found the labeling to be efficient (Fig. 8C). In an attempt to eliminate the agarose isolation step we also tried labeling in whole cell lysates (see experimental methods), but were unable to measure the efficiency because R.ClaI enzyme had undetectable activity in the lysates (Fig. 8D).

The application of this method for isolating unknown regulators relies on the interaction between streptavidin and biotin tagged BseCI sites. To determine if this binding occurs we incubated streptavidin (1.6 mM) with biotin-labeled pBSECI (2 h at room temperature), digested with BspHI and ran the mixture on an agarose gel (Fig. 8E). The results show that the 1662 bp fragment, which contains the BseCI sites, is shifted into the well, and that there is some smearing of the other two bands. The shift in the 1662 bp fragment is larger than expected, but may be explained by the following. After M.BseCI transfers the biotin group from the aziridine co-factor it remains covalently attached to the DNA (Dr. E. Weinhold, pers. communication), such that even if only half of the sites were occupied the size of the complex would be very large (1662 bp DNA + 2 molecules of streptavidin [106kDa] + 2 molecules of M.BseCI [122kDa]). In addition to this the overall net charge of the complex is probably going to be
close to neutral, so migration during electrophoresis will also be affected. The smearing of the other two bands may be due to nonspecific binding of DNA by streptavidin or simply to the fact that the large DNA-protein complex described might inhibit/slow movement of the smaller fragments out of the well.
Activation and repression of transcription in prokaryotes and eukaryotes is a complex process that usually involves the action and interaction of a group of co-regulating proteins (Kadonaga, 2004; Metzner et al., 2004; Rojo, 1999; Tudor et al., 2007). Knowledge of the regulators controlling the expression of a particular gene can provide useful information about the global and local regulatory structure of a response to a specific environmental condition or genetic program (Paul et al., 2007; Zaidi et al., 2002). In addition the presence of a well-characterized regulator may provide insight into the mechanism of control for a specific promoter (Gartenberg and Crothers, 1988; Rice et al., 1996; Spana and Corces, 1990).

Current methods for identifying DNA-protein interactions rely on the immunoprecipitation of chromatin using an antibody to a known protein target, and are geared towards determining the DNA sequence recognized by the protein (Hudson and Snyder, 2006; Johnson and Bresnick, 2002; Kang et al., 2002). We have put in place the basis for a method to be used for identifying the regulatory proteins associated with a particular promoter DNA sequence. A recent study (Paul et al., 2007) identified unknown regulators in vitro by attaching biotinylated promoter DNA (160 bp) to streptavidin-coated chips, and incubating with cleared whole-cell extracts. The masses of promoter-associated proteins were then identified using SELDI mass spectrometry. Our method provides the added benefit of in vivo binding conditions and peptide identification by mass
spectrometry. The method is based on the *in vivo* crosslinking principle of ChIP, but it is different in that the identity of isolated regulators is not required.

Preliminary results are encouraging in that we’ve been able to successfully isolate plasmid DNA crosslinked to protein using an agarose gel electrophoresis protocol (Fig. 4). In addition we have put in place three independent approaches for the separation and enrichment of promoter DNA-protein complexes (Figs. 5-8), although each method requires further development (discussed below).

**Agarose gel electrophoresis as a first step.** The results from the density gradient experiments (Fig. 3) suggest that the level of crosslinking may vary from one copy of the plasmid to the next as indicated by the failure to isolate a single plasmid band (smears in Fig. 3A), which might explain why we failed to detect DNA binding proteins along with regions of the gradient containing DNA. It is possible that the crosslinking didn’t work, or that the CsCl reversed the crosslinks. A similar experiment involving only *E. coli* (data not shown) yielded similar results. It is also apparent that downstream isolation steps would be more difficult due to the high concentration of proteins not specifically associated with the plasmid or the promoter DNA. This in mind we developed a method for “purifying” crosslinked plasmid DNA and associated proteins from the majority of the other proteins and DNA (Fig. 4) in addition to functioning as a buffer exchanger.

One important aspect of this step that has not been fully developed is how to extract the crosslinked-plasmid-protein complexes from the agarose. We
boiled the excised bands to melt the agarose and reverse crosslinks at the same time, but cannot use this approach where we wish to isolate the promoter DNA-protein complex from the rest of the plasmid DNA. It might be possible to use a column (such as Amicon Ultra Free-DA), or to add an agarase (such as β-agarase I, New England Biolabs). Both of these methods are fairly rapid, and should have little to no effect on subsequent steps.

**Enrichment of promoter DNA-protein complexes using the streptavidin binding DNA aptamer.** Because the DNA aptamer must form a specific stem loop structure before it can bind streptavidin (Fig. 5) there remains a need to optimize conditions that promote: 1) strand separation of dsDNA; 2) stabilization of the aptameric structure in a way that does not lead to strand reassociation. One critical question that has not been addressed has to do with the effects of crosslinking on the melting of the double-stranded proaptamer. If optimal streptavidin binding conditions are determined, but binding to streptavidin is not detected after crosslinking *in vivo* then it may be necessary to abandon this approach. Modifications to the heating step in an attempt to eliminate proteins would most likely result in a loss of promoter-specific regulators, and an overall decrease in robustness.

**Enrichment of promoter DNA-protein complexes using a hexahistidine tagged tetracycline repressor.** Due to difficulties in purifying TetR this approach would require the most development effort. To limit the laboriousness of this method TetR must be readily purified under non-denaturing conditions, and this has been accomplished by others (Ettner *et al.*, 1996; Hickman *et al.*, 1990;
Hillen et al., 1982; Tovar et al., 1988). Initial attempts to clone the tetR ORF into pET38b(+) that involved removing a large (50 amino acids) signal sequence between the lac operator and the BamHI site were successful, but we were unable to induce protein expression, which was most likely due to alterations in the spacing of the ribosome binding site. An alternative cloning strategy yielded a fusion protein that retained the entire signal sequence (see experimental procedures), and we suspect that this signal sequence may be contributing to the misfolding of TetR. Preliminary attempts to remove the signal sequence using recombineering techniques (Oppenheim et al., 2004; Yu et al., 2003) were unsuccessful.

The His-tagged tetR (from pTETR) could be cloned into an inducible expression vector such as pMAL-p2E (New England Biolabs), which contains a poly-linker for fusing proteins to the mannose binding protein (MBP). Fusing the TetR-His to the MBP may reduce the formation of inclusion bodies while at the same time providing a reliable means of purification (di Guan, 1988 #51; Maina, 1988 #50; P. Rajesh, pers. commun.). Once the purification of TetR-His is completed then the optimal method (add nickel associated beads and do spin/wash or immunoprecipitation with α−TetR or α−His tag antibody) for purifying promoter DNA-protein complexes should be evaluated; the nickel beads would be the least expensive approach.

An alternative to all of the cloning mentioned above could be to do the crosslinking in a TetR overproducing strain (Oehmichen et al., 1984), and then
use immunoprecipitation or affinity chromatography to isolate promoter-protein complexes.

Enrichment of promoter DNA-protein complexes using sequence-specific labeling of DNA (SMILing). Of the three approaches presented SMILing DNA seems to be the most promising. Labeling cleared lysates is probably not going to be the most efficient method as the 6Baz greatly prefers endogenous AdoMet (E. Weinhold, pers. comm). Because we were unable to optimize the lysate in a manner that facilitated digestion by ClaI we could not evaluate the efficiency of labeling. Preliminary purification of the crosslinked plasmid, for example on agarose gels, may be necessary.

Because the M.BseCl remains covalently linked to the recognition sequence, it might be helpful to consider adding a restriction site and some extra sequence just upstream of the recognition sequences. The streptavidin-isolated biotin-promoter DNA-protein complex could then be digested and the contaminating M.BseCl could be removed by using biotin. This added step may help reduce the complexity of the peptide digests by eliminating two proteins, streptavidin and M.BseCl, but the added purification may also reduce yields.

The results from the streptavidin gel shift indicate the formation of crosslinked complexes, and this is most likely due to an insufficient amount of streptavidin (E. Weinhold, pers. comm.).
Conclusions

We have carried out initial development of a novel and potentially very powerful tool for studying regulatory networks. The establishment of an efficient and robust method such as this would have applications in the identification of unknown gene regulatory proteins in many different prokaryotic and eukaryotic systems. Validation of the promoter DNA-protein isolation methods are required, and the value of the method should be tested in a eukaryotic system. For this, the post-lysis protocol should not need substantial modification, but steps related to the *in vivo* aspect (crosslinking conditions, plasmid type, and extraction) would have to be optimized.
Experimental Procedures

Bacterial strains, media and growth conditions.

Details about the strains used in this study are listed in Table 2. All cultures were grown at 37°C by shaking in baffled flasks or by rotating in 15 mL conical tubes. Cultures of *E. coli* were grown in lysogeny broth (Bertani, 1951), and *F. tularensis* was grown in Mueller-Hinton broth (DIFCO) supplemented with 0.025% ferric pyrophosphate and 0.1% isovitalex (Baker et al., 1985; Fortier et al., 1991). Antibiotics were used as indicated in the following concentrations: carbenicillin 50 µg/mL, kanamycin 50 µg/mL and tetracyline 10 µg/mL.

In vivo crosslinking and preparation of cell lysates.

Unless otherwise noted *in vivo* crosslinking was performed as follows: growing cells were exposed to formaldehyde (1% final concentration) for the time periods indicated. Cells (5 ml) were then pelleted by centrifuging for 2 min at 6000 rcf. The supernatant was discarded, and the cells were lysed by adding 600 µl of lysis buffer (10 mM sodium deoxycholate, 1 mg lysozyme, 15 mM EDTA, and 4.5 ml BugBuster, purchased from EMD Biosciences). Lysates were incubated at room temperature with gentle shaking until cleared, and then centrifuged at 4°C and 16,000 rcf for 15 minutes. Supernatants were removed and stored on ice for up to 1 h.
Construction of plasmids

The plasmids used in this study are listed and briefly described in Table 2.

pRLTO – To clone the tandem right tet operator sequence (Berens and Hillen, 2003) equimolar amounts of oligonucleotides tet1 and tet2 (Table 2) were mixed, heated to 98°C for 1 minute and cooled gradually to 28°C. The annealed oligonucleotides were then end-filled using Klenow polymerase (New England Biolabs, Ipswich, MA), and digested with Ascl (New England Biolabs) for 1 hour. Enzymes, salts and small DNA fragments were removed via a CentriSpin 20 column (Princeton Separations Inc, Adelphia, NJ). The 73 bp fragment was ligated into Ascl-digested pRLSA, and used to transform into E. coli Top10. Positive clones were sequence confirmed.

pRLSA – Equimolar amounts of oligonucleotides sap1 and sap2 (Table 2) were treated as above. The fragment was digested with XbaI (New England Biolabs) for 2 h; after heat inactivation of XbaI the 81 bp fragment was ligated into XbaI-digested pUC19 (Yanisch-Perron et al., 1985). Positive clones were sequence confirmed.

pRLTG and pRLSG – The regulatory region of gltBDF (716 bp upstream of translational start) was isolated from pBE10 (Ernsting et al., 1993) by digesting with BamHI and SalI (New England Biolabs). The fragment was end-filled as above, resolved on a 1% agarose gel, excised and purified using the Qiagen (Valencia, CA) QIAquick gel extraction kit and protocol. Eluted fragments were ligated into NotI (New England Biolabs) digested pRLTO and pRLSA (Table 2).
pBSEC1 – Oligonucleotides bsc1 and bsc2 were mixed, annealed and end-filled as previously described. The fragment was digested with Ascl for 2 h, and the 76 bp fragment was then ligated into Ascl-digested pRLTO (Table 2).

**Cloning and expression of tetR**

Primers tetR1 and tetR2 were used with Pfx Taq polymerase (Invitrogen) to PCR amplify tetR from *E. coli* BE10.2, which contains a Tn10 insertion (Ferrario et al., 1995). The PCR reaction was resolved on an agarose gel, and the 637 bp product was excised and purified using the QIAquick method described previously. The purified fragment was digested with BamHI (New England Biolabs), and ligated into BamHI and EcoRV digested pET38b(+) (EMD Biosciences) resulting in pTETR. This strategy put the tetR gene under the control of the T7 promoter, and in-frame with a C-terminal hexahistidine tag as well as a C-terminal cellulose-binding domain. Positive clones were transformed into *E. coli* BL21 (DE3), which contains an IPTG inducible T7 RNA polymerase.

Overnight cultures of BL21 containing pTETR were inoculated at 1:25 into fresh medium. Expression of tetR was induced by adding 0.4mM Isopropyl β-D-1-thiogalactopyranoside (IPTG), and growing for 12 h at 37°C.

**Western blot analysis**

The total concentration of protein was determined for each sample using the Micro BCA protein assay kit (Pierce, Rockford, IL) as directed. Equal amounts of protein were loaded on a 12% acrylamide SDS gel and electrophoresed at 110 V in 1X tris-glycine buffer. Proteins were then electroblotted to polyvinylidene difluoride (PVDF) membranes at 30 V for 1 h using the Xcell blot apparatus.
(Invitrogen). Proteins were detected via chemiluminescence using the Super Signal ECL western blotting substrate (Pierce) as per the manufacturer's protocol, with a 1:125 dilution of rabbit anti-Lrp polyclonal serum (Willins et al., 1991), and a 1:25,000 dilution of horseradish peroxidase-conjugated goat anti-rabbit IgG (gift of Dr. Darren Sledjeski). Protein bands were visualized on x-ray film using various exposure times.

**32P labeling of DNA fragments and electrophoretic mobility shift assays (EMSA)**

Twelve picomoles of primer sapt1 (Table 1) were radiolabeled with 5'-adenosine-[γ-32P]triphosphate (Amersham Biosciences, Pittsburgh, PA) using T4 polynucleotide kinase (New England Biolabs). The labeling reaction was carried out in T4 polynucleotide kinase buffer (New England Biolabs) for 30 min at 37°C, followed by inactivation of PNK by heating to 70°C for 5 min. Primers [γ-32P]sapt1 and sapt2 were then used to PCR amplify an 800 bp product containing the streptavidin binding aptamer and the gltB control region, and unincorporated radionucleotides were removed and the buffer exchanged via CentriSpin 20 (Princeton Separations) column filtration. Eluted fragments were resolved on a 1% agarose gel, excised and extracted using QIAquick as described above.

DNA aptamer-streptavidin binding reactions were carried out under various conditions (see Bittker et al., 2002). Loading buffer was added [1x final], and samples were loaded onto 5% acrylamide gels. Electrophoresis was carried out at room temperature for 2.5 h at 90V. Gels were exposed to phosphorimager screens (Molecular Dynamics, Sunnyvale, CA), and DNA fragments were
detected via the STORM 840i phosphorimager (Molecular Dynamics, now GE Healthcare). Analysis was performed on a Macintosh using the Molecular Dynamics software package and the public domain NIH Image program (developed at the U.S. National Institutes of Health and available at http://rsb.info.nih.gov/nih-image/).

**Isolation of crosslinked plasmid DNA-protein complexes via isopycnic CsCl gradients**

Cells were crosslinked *in vivo* as described above, but with the following modifications: after adding formaldehyde, cells were shaken on ice for 1.5 h before harvesting by centrifugation. Prior to lysis cell pellets were stored at -80°C for 5 min; this added step improved the overall efficiency of lysis for cell pellets from large volumes (≥50ml) of overnight culture.

Cleared lysates were added to 6.62 grams of CsCl (Gibco BRL) to a final volume of 12.5mL (mix = 1.4 g/cc), and samples were heated intermittently as needed to dissolve CsCl. Samples were then stored at 4°C for 24 h. Gradients were transferred to Beckman Quick Seal™ polyallomer ultracentrifuge tubes; ethidium bromide was added to a final concentration of 0.64mg/ml. Tubes were topped off with mineral oil, heat sealed and mixed. Samples were loaded into a 75Ti rotor (Beckman, Fullerton, CA) and centrifuged under vacuum at 23°C and 56 krpm for 20 h in a Beckman L8-80M ultracentrifuge. After centrifugation was completed the centrifuge was allowed to decelerate without braking. The tubes were removed, and fractions were collected by inserting 21 and 18 gauge needles into the top and bottom, respectively, of the tube. Once collected
fractions were pooled as indicated, and added to Sephadex 6-100 spin columns (defined and equilibrated in TEN buffer as directed, Pfizer, New York, NY) to desalt and remove excess ethidium bromide. Samples were then precipitated with saturated n-butanol, concentrated to 50 µl and the buffer exchanged via CentriSpin 20 columns (Princeton Separations).

The concentration of total protein was measured for each sample using the Micro BCA protein assay kit (Pierce). 3x SDS loading buffer containing 5% β-mercaptoethanol was added, and samples were heated at 98°C for 40 minutes to reverse crosslinks. After heating samples were centrifuged briefly, and equal amounts of total protein were loaded onto a 12% acrylamide gel. Proteins were resolved by electrophoresis at 90V.

**Mass spectrometry**

Proteins were visualized by staining with colloidal Coomassie Blue (Invitrogen). Bands of interest were excised from gels and digested overnight at 37°C with sequencing-grade modified trypsin (Promega, Madison, WI). Peptides were extracted with 60% acetonitrile : 0.1% TFA. Extracts were concentrated to ~15 µl using a Vacufuge (Eppendorf, Hamburg, Germany). For each sample 2 µl was separated on a 75 µm id X 5 cm X 15 µm Aquasil C18 Picofrit reverse phase column (New Objective, Inc, Woburn, MA). The eluted peptides were directly introduced into a Finnigan™ LCQ-Deca XP Plus ion-trap mass spectrometer equipped with nano-spray source. The mass spectrometer was set to acquire a full MS scan (400-2000 m/z), and a collision induced dissociation (CID) spectrum on the most abundant ion. Analyses of CID spectra were performed manually by
Dr. V. Basrur, or searched against a non-redundant *E. coli* database using the TurboSEQUEST software (Finnigan).

**Sequence specific Methyltransferase Induced Labeling of DNA (SMILing DNA).**

1 µg of plasmid DNA (pBSECI, linearized with R.BspHI) was added to labeling reactions (50 µl final volume) containing: 10 mM Tris/HCl (pH 7.4), 50 mM NaCl, 0.05 mM EDTA, 2 mM β-mercaptoethanol, 80 µM biotin aziridine cofactor and 68.9 nM M.BseCI (gift of Dr. E. Weinhold). DNA was labeled for 3 hours at 55°C. M.BseCl was then heat inactivated by heating to 70°C for 10 minutes.

R.ClaI digests were carried out in Y*-Tango buffer (33 mM Tris/OAc, pH7.9; 10 mM Mg(Oac)₂; 66 mM KOAc, 0.1 mg/ml BSA) for 1 h at 37°C.
Acknowledgements

The authors thank Venkatesha Basrur for performing the mass spectrometry analysis, and Darren Sledjeski for the *Francisella tularensis* strain containing pGFP. REL was supported in part by a graduate fellowship from The University of Toledo Health Science Campus.
References


*Methods* **37**: 16-25.


Figure Legends

Figure 1. Overview of isolation protocol. 1) The regulatory regions of interest are cloned into the NotI site, which is upstream of cassettes that contain the structural or sequence elements needed for purification of the DNA-protein complex (see text). The plasmids are transformed or transfected into cells and grown under the desired conditions. 2) Promoter DNA is crosslinked to transcription factors by adding formaldehyde, and the cells are lysed under non-denaturing conditions. 3) Plasmid DNA with crosslinked proteins is isolated on agarose gels. 4) Promoter DNA-protein complexes are liberated by digesting with Ascl, and purified using the appropriate affinity method (see text). Once the promoter DNA-protein complexes are isolated the crosslinks are reversed and the regulators are identified via mass spectrometry.
1 Clone regulatory region

2 Obtain protein-DNA complex
   In vivo cross-linking
   Non-denaturing lysis

3 Separate plasmid DNA

4 Isolate specific promoter-proteins complex
   Digest with Ascl

Post-Isolation:
   Mass spectroscopy to identify
   Confirmation via EMSA,
   footprinting, mutation
Figure 2. Isolation of plasmid using non-denaturing conditions. A) Cells containing pRLTG were lysed using the gentle lysis method described in the text (lane 1), and using buffer P1 from the Qiagen miniprep kit (lane 3). Plasmid from the P1 lysis was mixed with plasmid from the gentle lysis to demonstrate that the difference in mobility was due to the buffer (lane 2). B) R.Pvull is active in the lysis buffer without added yield-reducing buffer exchange or desalting steps. C) Effects of crosslinking on cell lysis and plasmid mobility. Cells were incubated with formaldehyde for the indicated times.
Figure 3. Isolation of crosslinked plasmid via ultracentrifugation.

*F. tularensis* or *E. coli* cells containing pGFP were crosslinked and lysed. Cleared lysates were added to cesium chloride and ultracentrifuged. A) Fractions were eluted from different regions of the gradient and run on agarose gels to identify those that contained DNA. Samples representing the top (T) and middle (M) positions of the gradient were pooled, loaded onto an SDS-PAGE. B) Isolated proteins were stained with Coomassie blue. Bands (boxed regions) were excised from the gel and proteins were identified by mass spectrometry. Letters by boxes correspond to Table 1.
Table 1. Proteins identified from CsCl gradients.

<table>
<thead>
<tr>
<th>Strain</th>
<th>*Gel Posn</th>
<th>Protein Candidate</th>
<th>MW (kDa)</th>
<th>Accession#</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E. coli</strong></td>
<td>a</td>
<td>Catalase Hydroperoxidase II</td>
<td>84</td>
<td>P21179</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elongation Factor G</td>
<td>77</td>
<td>P0A6M8</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>Chaperonin GroEL</td>
<td>57</td>
<td>AAS75782</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bacteriophage P2-related tail formation protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phosphoenolpyruvate carboxykinase</td>
<td>61</td>
<td>ZP_00923910</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59</td>
<td>P22259</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Superoxide dismutase</td>
<td>21</td>
<td>BAB35788</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alkyl hydroperoxide reductase subunit C (SCRP-23)</td>
<td>21</td>
<td>P0AE08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Periplasmic lipoprotein OsmY</td>
<td>21</td>
<td>BAE78364</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ribosome releasing factor (RRF)</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transcription elongation factor GreA</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td><strong>F. tularensis</strong></td>
<td>a</td>
<td>Chaperone GroEL</td>
<td>57</td>
<td>CAJ80153</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peroxidase/catalase (HPI)</td>
<td></td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>Triosephosphate isomerase</td>
<td>24</td>
<td>CAJ80219</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Purine-nucleoside phosphorylase</td>
<td>26</td>
<td>CAJ79900</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Superoxide dismutase [Fe]</td>
<td>21</td>
<td>CAJ80230</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(SOD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nucleoside diphosphate kinase (NDK)</td>
<td>15</td>
<td>YP_763767</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>Bacterioferritin like protein</td>
<td>19</td>
<td>CAJ79057</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Glycine cleavage protein H</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>e</td>
<td>Thioredoxin</td>
<td>12</td>
<td>CAJ80010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chaperonin like protein GroES</td>
<td>10</td>
<td>CAJ80154</td>
</tr>
</tbody>
</table>

*Refers to the position from Fig. 3B
Figure 4. Isolation of crosslinked plasmid via agarose gel electrophoresis.

*E. coli* cells harboring a plasmid with the *gltB* control region (pRLTG) were crosslinked or left untreated. A) Plasmid DNA was separated by electrophoresis on agarose gels. B) Plasmid DNA-protein isolates were excised, boiled to reverse crosslinks and separated by SDS-PAGE. Purified His-Lrp was included for comparison (see text). C) Retardation factor (Rf) analysis of isolated proteins revealed five possible proteins expected to be associated with the plasmid or the promoter. Filled circles indicate the molecular weight standards.
Figure 5. DNA aptamer based isolation of promoter-protein complexes.

A) Nucleotide sequence of DNA streptavidin aptamer (Bittker, 2002 #28). B) Plasmid arrangement and overview of the strategy for isolating promoter-protein complexes using streptavidin to bind the stem loop structure formed by the sense strand of the DNA aptamer (see text for details).
A 5'-TCTGTGAGACGACGACCGGTCGAGGTTTTGTCTCACAG-3'

B

- XbaI
- AseI
- NotI
- NotI
- AseI
- XbaI

**DNA Aptamer**

1. Cross-link and lysis
2. Isolate plasmid
3. AseI digest
4. Melt strands (heat, urea), snap renature
5. Add streptavidin sepharose beads
Figure 6. Pilot DNA aptamer-streptavidin binding experiments. A) EMSA to detect binding of the DNA aptamer by streptavidin (SA) at different ratios of SA and MgCl$_2$, and in the presence or absence of biotin. B) Duplicate EMSA to determine effects of [MgCl$_2$] on SA-aptamer binding.
Figure 7. TetR based isolation of promoter-protein complexes. A) Nucleotide sequence of tet operators. B) Plasmid arrangement and overview of the promoter-protein complex isolation strategy using the TetR-tetO interaction. C) Induction time course of His-TetR, which is under the control of the T7 promoter, in the presence or absence of 2.5% ethanol (EtOH). Cells were grown with or without 0.4 mM IPTG, and samples were lysed by boiling. Proteins were separated by SDS-PAGE and stained with Coomassie brillant blue. Induction times listed are in hours. D) Cells grown at different temperatures (°C) and in the presence/absence of EtOH were lysed under denaturing and non-denaturing conditions. His-TetR is found in the insoluble fraction.
Figure 8. Isolation of promoter-protein complexes via SMILing DNA. A) Nucleotide sequence of the cassette containing multiple M.BseCI recognition sites. B) Plasmid arrangement and overview of promoter-protein complex isolation using streptavidin to bind biotin labeled cassettes. C) In a purified system labeling of BseCI sites with biotin (6Baz) prevents restriction by ClaI at an efficiency similar to that seen in the adenosyl methionine (AdoMet) control. D) Labeling of BseCI sites in cleared lysates. E) Gel shift to demonstrate binding of the biotin labeled fragment by streptavidin (see text).
DISCUSSION

After years of analysis the transcriptional programs of model organisms such as *E. coli* have yet to be fully characterized (Tompa et al., 2005). This is due in large part to the difficult and time-consuming nature of the experimental process. With some additional validation, predictive methods offer a reprieve to the multi-decade task of defining the complex regulatory networks associated with metabolism and virulence. By combining data from large-scale experimental studies with genome sequence information bioinformatic approaches eventually may be able to efficiently and reliably deduce the genetic and regulatory components of physiology. The work presented here provides a framework for validating the “rules” that govern *in silico* analyses by testing basic assumptions about orthologous transcription factors, and by providing a novel high throughput method for identifying regulator integration at specific promoters.

**Functional equivalency of highly conserved regulators**

A basic assumption used in predicting the regulation of genes in two different organisms is that highly conserved regulators are functionally interchangeable. Comparisons of the amino acid sequences of the Lrp orthologs from three distant but related bacterial species suggest that these proteins will recognize and bind similar DNA sequences in addition to responding equally to co-regulatory signals from leucine and alanine. Results from this work support the basic assumption of
equivalency in that heterologous Lrp proteins do activate or repress target genes, and do respond to leucine in a similar manner. However, subtle but key differences in the activity of these regulators at various promoters suggest unpredicted differences in intrinsic properties such as relative affinity for particular recognition sequence variants. Such differences would obviously affect the expression of a given target gene.

Consider the case of a gene that is transiently activated by a regulator with a low affinity for its promoter, where the affinity is affected by a ten amino acid peptide at the N-terminus of the protein that is far from the known DNA-binding motif. Now consider that the ortholog from a second species, having overall amino acid identity is 95%, contains two substitutions in this N-terminal peptide, enhancing its affinity for the promoter. Data presented here suggests that this scenario applies to \( P_{\text{gltB}} \) from \( E. \ coli \). The Lrp ortholog from \( P. \ mirabilis \) appears to bind the promoter with a much higher affinity as indicated by a significant increase in activity, and a lack of response to antagonism by leucine. This combined with the fact that \( P. \ mirabilis \) Lrp has a very small effect on its own \( PgltB \) indicate that predicting the proper regulation of a gene as well as the place of a regulator within a network requires compensation for the distinct functional features of the transcription factor of interest.
Conservation of regulator gene control

A second property of orthologous regulators that is often taken for granted is that they are themselves regulated in response to similar signals in different organisms, and that the actual concentration of the protein is the same or can be ignored. Details presented in this work concerning the regulatory patterns of orthologous Irp genes in three species suggest that in some cases differences in protein levels correlate with changes observed in the magnitude of target gene expression. Overall, however, a clear relationship between regulator levels and target gene control could not be derived. This may be due in part to the limited number of target genes tested, but can also reflect the complexity predictions that bioinformatic approaches are attempting to make. A limitation to comparative studies involves the convention to only test sets of orthologous targets (i.e. genes present in all of the species being compared).

Defining the interplay between transcription factor concentration and regulatory effects requires an initial characterization of the regulons for the organisms of interest, followed by global analyses of the effects of altering optimal regulator expression levels. Results from these types of studies will provide a detailed picture of regulon structure and overlap between species while highlighting the extent to which individual regulator concentrations are a relevant consideration. A likely possibility is that the effect of protein level is going to vary from target to target depending upon the regulatory function of the transcription
factor such that conclusions about network involvement will be more reliable than conclusions about the regulation of a specific target (see manuscript I).

**Prescribing regulator involvement**

Simply determining whether or not a regulatory protein directly interacts with a promoter can be a difficult process, but this characterization is of utmost importance if regulatory interactions are going to be understood. The development of an efficient method for identifying the regulators associated with the control of a specific gene offers an indispensable means for not only defining the network connections but also for determining the strength of the algorithms that predict them.

**Summary**

In general this study has advanced the knowledge required for developing accurate computational methods of predicting gene regulation, by addressing the validity of underlying assumptions. In addition, the results from this study provide new and detailed information concerning the regulation (protein and mRNA levels) of the previously uncharacterized *lrp* genes from *P. mirabilis* and *V. cholerae* throughout growth in rich and minimal media. Information concerning the interchangeability of the global regulator Lrp from three genetically and physiologically distinct bacterial species was also presented as well as new
details about the regulation of *gltB* and *adhE* in *P. mirabilis* and *V. cholerae* during logarithmic and stationary growth.

The novel *in vivo* cross-linking method described and subjected to preliminary development, promises to be an efficient means for identifying the transcription factors associated with a specific promoter. If fully developed, the method would represent a significant contribution to regulon study while providing a distinct advantage over conventional approaches (e.g. chromatin immunoprecipitation) by allowing unknown regulators to be preferentially isolated. The full utility of this method has yet to be realized, but the results presented here suggest that it is sensitive and amenable to numerous species and experimental conditions, and that it does provide information about the direct involvement of a particular regulator in addition to the interaction of other co-regulators leading to a more conclusive picture about the regulatory mechanism of a gene.


distributed architecture behind scale-free transcriptional regulatory networks. J Mol Biol 360, 204-212.


promotes formation of a bent repression loop in lac DNA. J Mol Biol 196, 101-111.


genomic DNA. J Mol Biol 268, 78-94.


Gonzalez, A. D., Espinosa, V., Vasconcelos, A. T., Perez-Rueda, E., and Collado-


Horak, C. E., Luscombe, N. M., Qian, J., Bertone, P., Piccirillo, S., Gerstein, M., and


polymerase alpha subunit: involvement of the C-terminal region in transcription activation by cAMP-CRP. Cell 65, 1015-1022.


Kalinina, O. V., Makeev, V. J., Sutormin, R. A., Gelfand, M. S., and Rakhmaninova, A.
B. (2003). The channel in transporters is formed by residues that are rare in transmembrane helices. In Silico Biol 3, 197-204.


Landick, R., and Oxender, D. L. (1985). The complete nucleotide sequences of the


Maina, C. V., Riggs, P. D., Grandea, A. G., 3rd, Slatko, B. E., Moran, L. S.,
coli vector to express and purify foreign proteins by fusion to and separation from

preferences in the arrangement of binding motifs and hierarchical levels in
organization of transcription regulatory information. Nucleic Acids Res 31, 6016-
6026.

Maniatis, T., Ptashne, M., Backman, K., Kield, D., Flashman, S., Jeffrey, A., and
Maurer, R. (1975). Recognition sequences of repressor and polymerase in the

genes in the context of biological pathways: An application of integer

Marasco, R., Varcamonti, M., La Cara, F., Ricca, E., De Felice, M., and Sacco, M.
(1994). In vivo footprinting analysis of Lrp binding to the ilvIH promoter region of


Martinez-Antonio, A., Salgado, H., Gama-Castro, S., Gutierrez-Rios, R. M., Jimenez-
Jacinto, V., and Collado-Vides, J. (2003). Environmental conditions and
transcriptional regulation in Escherichia coli: a physiological integrative approach.


Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early
T7 promoter. Proc Natl Acad Sci U S A 72, 784-788.


Quay, S. C., and Oxender, D. L. (1977). Regulation of amino acid transport in


Escherichia coli K-12 transcriptional regulatory network, operon organization, and

integrated framework. Biol Direct 1, 11.

Rev Biophys Biomol Struct 34, 379-398.

Design in Molecular Biology (Reading, MA, Addison-Wesley).

Schlax, P. J., Capp, M. W., and Record, M. T., Jr. (1995). Inhibition of transcription

content of binding sites on nucleotide sequences. J Mol Biol 188, 415-431.

Chaperones 8, 207-217.

Scott, S., Busby, S., and Beacham, I. (1995). Transcriptional co-activation at the ansB
promoters: involvement of the activating regions of CRP and FNR when bound in
tandem. Mol Microbiol 18, 521-531.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N.
(2003). Module networks: identifying regulatory modules and their condition-
specific regulators from gene expression data. Nat Genet 34, 166-176.

Use of the lambda Red recombinase system to produce recombinant prophages


Tani, T. H., Khodursky, A., Blumenthal, R. M., Brown, P. O., and Matthews, R. G. 193


Tudor, C., Feige, J. N., Pingali, H., Lohray, V. B., Wahli, W., Desvergne, B.,


ABSTRACT

Predictions of gene regulation from DNA sequences are generally based on the poorly-tested hypothesis that a conserved regulator, conserved target gene, and putative binding sites for the regulator upstream of the target gene, imply a conserved pattern of regulation. Inherent to this framework is the assumption that well-conserved regulators are intrinsically and extrinsically equivalent, a presumption that has not been fully tested.

The equivalency of the leucine-responsive regulatory protein (Lrp) was assessed for three well-conserved orthologs from the gamma proteobacteria *Escherichia coli, Proteus mirabilis* and *Vibrio cholerae*. Using the well-defined *E. coli* system as a model, it was determined that the Lrp orthologs maintained their ability to properly regulate heterologous target genes in the presence and absence of a coregulator. In addition, the orthologous regulators functioned well as part of the complex response associated with hyperflagellated swarmer cell differentiation in the pathogen *P. mirabilis*. However, there were some differences in the effects of Lrp orthologs on the activity of various target promoters, suggesting differences in affinity. The regulatory control of the *lrp* gene itself, in addition to intrinsic differences between orthologs, was found to be vital to its regulatory scope. This was indicated by the variation in *Proteus* swarming phenotypes, and by the reduced number of target genes complemented by *lrp* under the control of P_{lac} compared to the native promoter.
Experimental methods that identify the transcription factors associated with a particular promoter offer essential details about combinatorial regulatory control. An important limitation on whole-genome analysis of regulation is lack of a relatively high-throughput method for identifying regulatory proteins associated with a given promoter, without knowing in advance what any of those proteins are. In general, current methods either lack the ability to identify unknown regulators or lack the context of true in vivo conditions. A plasmid-based in vivo method for unknown regulator identification was proposed, and developed on a preliminary basis. The method allows enrichment of regulators associated with a given promoter, and may be useful in both prokaryotic and eukaryotic systems.