Published in final edited form as:
Nat Biotechnol. 2005 August ; 23(8): 942–944. doi:10.1038/nbt0805-942.

Discovery of DNA regulatory elements with bacteria:

Bacterial one-hybrid selection system offers a low-tech alternative for determining the DNA-binding specificities of transcription factors

Martha L. Bulyk
Martha L. Bulyk is in the Division of Genetics, Department of Medicine, Department of Pathology, and the Harvard-MIT Division of Health Sciences & Technology, Brigham & Women's Hospital and Harvard Medical School, Harvard Medical School New Research Building, Room 466D, 77 Avenue Louis Pasteur, Boston, MA, 02115, USA. mlbulyk@receptor.med.harvard.edu.

With the availability of nearly a decade's worth of genome-scale gene expression profiling and the more recent sequencing of multiple higher eukaryotic genomes, attention is now shifting towards determining the regulatory mechanisms underlying these expression patterns. However, a major challenge in understanding these transcriptional regulatory networks has been the lack of DNA-binding site data for most transcription factors. Without binding site data, it is difficult to identify the target genes directly regulated by a given transcription factor and to identify the *cis* regulatory elements through which this regulation occurs. In this issue, Wolfe and colleagues[1] present their adaptation of a bacterial one-hybrid (B1H) system[2] for determining the DNA-binding specificities of transcription factors. Two important advantages of Wolfe's version of a B1H approach over the previously developed B1H selection system[3] are the incorporation of a negative selectable marker for improved background reduction, and the use of randomized candidate DNA binding sites, both of which have been employed in yeast one-hybrid selection[4]. Using this updated B1H system, the authors identified the DNA-binding site motifs for eight metazoan transcription factors, including one *Drosophila* protein (Odd-skipped (Odd)) whose DNA-binding specificity was previously unknown. Using these newly discovered binding site data, they then predicted and experimentally validated two new target genes of Odd.

DNA-binding proteins are important broadly in both lower organisms and more complex metazoans, in numerous cellular processes such as transcription regulation, DNA repair, and replication. The largest class of these proteins are regulatory transcription factors, which by binding in a sequence-specific fashion to DNA-binding sites in the genome, modulate the expression of their target genes as required for normal progression through the cell cycle and in response to environmental stimuli, and in a cell type and developmental stage specific manner in higher organisms.

Despite the crucial regulatory roles of transcription factors, the DNA-binding specificities of relatively few of them have been characterized in depth. In order to understand how they regulate their target genes, one must be able to identify the DNA-binding sites to which they bind in a given genome. Currently, experimental data on transcription factors' DNA-binding specificities are required to predict such *cis* regulatory elements. However, some methods for high-throughput binding site determination, such as microarray-based readout of chromatin immunoprecipitation ('ChIP-chip')[5-7], require specific antibodies, while other methods, such as *in vitro* selection[8] and protein binding microarrays[9], require purified protein. In contrast, the B1H system not only employs *in vivo* selection, but also offers a low-tech alternative to microarray-based technologies.

In their study, Wolfe and colleagues expressed the DNA-binding domain of a given transcription factor as a fusion to the alpha subunit of RNA polymerase. A library of randomized oligonucleotides cloned into a vector containing the selectable genes *HIS3* and *URA3*. If the given DNA-binding domain (the 'bait') binds a potential DNA target site (the 'prey') in the bacterium, then it will recruit RNA polymerase to the promoter and activate transcription of the reporter genes (Fig. 1A). The two reporter genes *HIS3* and *URA3* are yeast genes that allow for positive and negative selection, respectively, when propagated in a bacterial strain in which the bacterial homologs of these genes have been deleted. Specifically, growth of cells on minimum medium containing 3-amino-triazole (3-AT), which is a competitive inhibitor of HIS3, provides positive selection, while growth on medium containing 5-fluoro-orotic acid (5-FOA), which is converted into a toxic compound by the uracil biosynthesis pathway, provides negative selection. Positive clones are then sequenced, and the sequences of the selected clones are examined with pre-existing motif finding tools (MEME, BioProspector) in order to identify the recognition binding site motif of the query transcription factor (Fig. 1B).

To demonstrate that their B1H system can work, Wolfe and colleagues first used it to successfully identify the known binding specificities for two mammalian Cys_2His_2 zinc finger proteins, Zif268 (also known as Egr1) and PLAG1, whose DNA binding specificities were previously known. Before proceeding with analysis of additional proteins, the authors grew the original prey library alone in the presence of 5-FOA in an attempt to eliminate self-activating baits and thus reduce the false positive rate. The authors then used this "purified" prey library to determine the DNA binding specificities of four individual transcription factors from *C. elegans* (LAG-1) and *Drosophila* (Dorsal, Paired, Odd), one of which (Odd) had not been characterized previously. Importantly, they also identified the binding specificities of the *Drosophila* proteins Runt and Big-brother (Bgb), which bind DNA with high affinity only as a heterodimer, thus showing that Wolfe's B1H method works not just for monomeric proteins but also for proteins that bind DNA as complexes. In addition, these six proteins represent a number of structural classes of DNA binding domains (Rel homology region, CSL-type DNA binding domain, CBFα/β, paired domain and homeodomain) in addition to the Cys_2His_2 zinc finger domain, thus demonstrating the generality of the B1H approach.

Next, to explore the biological relevance of the DNA-binding site motif that they determined for Odd, Wolfe and colleagues searched the *D. melanogaster* and *D. pseudoobscura* genomes for conserved, syntenic regions that contained at least two Odd binding sites. This type of search is used typically because *cis* regulatory modules frequently contain multiple copies of a given motif, and searches for individual binding sites can result in many false positive target gene predictions. Although it is currently unclear what is the most appropriate way to measure functional conservation of binding sites, phylogenetic conservation within syntenic regions will likely enrich for regulatory regions. A number of the regions that the authors found in their search were adjacent to genes with similar biological functions as that of Odd, including two genes (*gooseberry (gsb)* and *Goosecoid (Gsc)*) that had not been previously identified as direct targets of Odd regulation. *In situ* hybridizations indicated diminished expression of *gsb* and *Gsc* upon induction of ectopically expressed Odd, thus validating that Odd is regulating these genes. Although the original prey library, consisting of 2×10^7 unique clones, contained only a very small fraction of all possible 18 bp sequences ($\sim 7 \times 10^{10}$), this subset of clones still covers enough sequence space to allow a sufficiently large subset of binding site space to be sampled for most proteins. Nevertheless, without a more complex prey library, it may prove difficult to determine the binding specificities of transcription factors with lengthy binding sites (i.e., much longer than 12 bp). Since self-activating sequences are removed in generating the purified prey library, a query transcription factor that has a close homolog in *Escherichia coli* that is active in the selection strain would also likely fail to be characterized by this B1H approach.

Wolfe and colleagues used multiple stringencies (i.e., concentrations of 3-AT) in their positive selection step to identify positive clones. Nevertheless, because the background in their Runt/Bgb selections was unacceptably high at even the highest 3-AT concentration, an additional negative selection step was required (here, the same concentration of 5-FOA was used as in selection of the purified prey library which retained some self-activating sequences). Thus, to apply the B1H approach generally, one would likely need to perform selections at a range of 3-AT concentrations, with presumably higher affinity binding sites being selected at higher concentrations of 3-AT, as well as a range of 5-FOA concentrations, to keep the proportion of false positive colonies to a minimum. It is encouraging that for eight of the nine transcription factors examined by B1H in this study, excluding the one protein that resulted in toxicity, the authors were able to successfully identify their DNA binding site motifs, despite the fact that they used only three different 3-AT concentrations over just a five-fold range.

An important point to keep in mind is that the degeneracy of the discovered binding site motifs will be reflective not only of the number of positive clones that are sequenced, but also of the stringency of the selections. If only a small number of clones from a more stringent selection are sequenced, then the motifs will likely represent only the higher affinity binding sites, even though weaker sites may also be biologically significant. Therefore, this B1H system would be improved by incorporation of a high-throughput sequencing step, such as by concatemerization of positive clones prior to sequencing, as in serial analysis of gene expression (SAGE)[10], which would permit the discovery of more accurate motifs by sequencing a greater number of clones, including those from less stringent selection conditions.

One advantage that a B1H system offers over a yeast one-hybrid system[4] is that the higher bacterial transformation efficiency allows more complex libraries to be examined more readily. In this present study by Meng *et al*., only a single large plate was required at each selection stringency, with multiple stringencies used for each transcription factor. Although expression in *E. coli* of proteins from higher eukaryotes will be problematic for some proteins, the authors were able to resolve this problem for one protein (Odd) by substituting rare codons with preferred synonymous codons, while expression of another attempted protein was toxic. Still, the effects of any post-translational modifications that are important for DNA-binding specificity would be missed, as would any conformational changes due to the rest of the protein sequence on the DNA-binding domain, as only DNA-binding domains were examined in this study. Nevertheless, given the lack of binding site data for most transcription factors in both model organisms and the human genome, even imperfect binding site data would be extremely valuable. For example, recent analysis suggests that there are ~1960 transcription factors, corresponding to ~8% of genes, in the human genome[11], and the sequence specificities and functions of most of these proteins have not yet been determined.

This B1H system provides another tool in our arsenal for identifying the DNA-binding specificities of transcription factors, and thus predicting their target genes and genomic DNA regulatory elements. Since co-regulation in higher eukaryotes frequently occurs through binding by a combination of transcription factors, analysis of such binding site data for transcription factors from those genomes will require further studies of homotypic and heterotypic binding site clustering, along with more sophisticated algorithms for the consideration of phylogenetic conservation. The B1H method developed by Wolfe and colleagues should also allow for the examination of the effects of protein-protein interactions on DNA binding, which may further guide the prediction of *cis* regulatory modules based on binding site clustering. As suggested by the authors' studies on Odd, results from these analyses also could be used to predict the regulatory roles of uncharacterized transcription factors. The integration of data from such studies will certainly help to achieve our goals of delineating the regulatory networks that govern cellular gene expression.

References

1. Meng X, Brodsky MH, Wolfe SA. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nature Biotechnol 2005;23
2. Dove SL, Joung JK, Hochschild A. Activation of prokaryotic transcription through arbitrary protein-protein contacts. Nature 1997;386:627–630. [PubMed: 9121589]
3. Joung JK, Ramm EI, Pabo CO. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc. Natl. Acad. Sci. U.S.A 2000;97:7382–7387. [PubMed: 10852947]
4. Wilson TE, Fahrner TJ, Johnston M, Milbrandt J. Identification of the DNA binding site for NGFI-B by genetic selection in yeast. Science 1991;252:1296–1300. [PubMed: 1925541]
5. Lieb JD, Liu X, Botstein D, Brown PO. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat. Genet 2001;28:327–334. [PubMed: 11455386]
6. Iyer VR, et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 2001;409:533–538. [PubMed: 11206552]
7. Ren B, et al. Genome-wide location and function of DNA binding proteins. Science 2000;290:2306–2309. [PubMed: 11125145]
8. Oliphant AR, Brandl CJ, Struhl K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol. Cell. Biol 1989;9:2944–2949. [PubMed: 2674675]
9. Mukherjee S, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat. Genet 2004;36:1331–1339. [PubMed: 15543148]
10. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487. [PubMed: 7570003]
11. Messina DN, Glasscock J, Gish W, Lovett M. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. Genome Res 2004;14:2041–2047. [PubMed: 15489324]

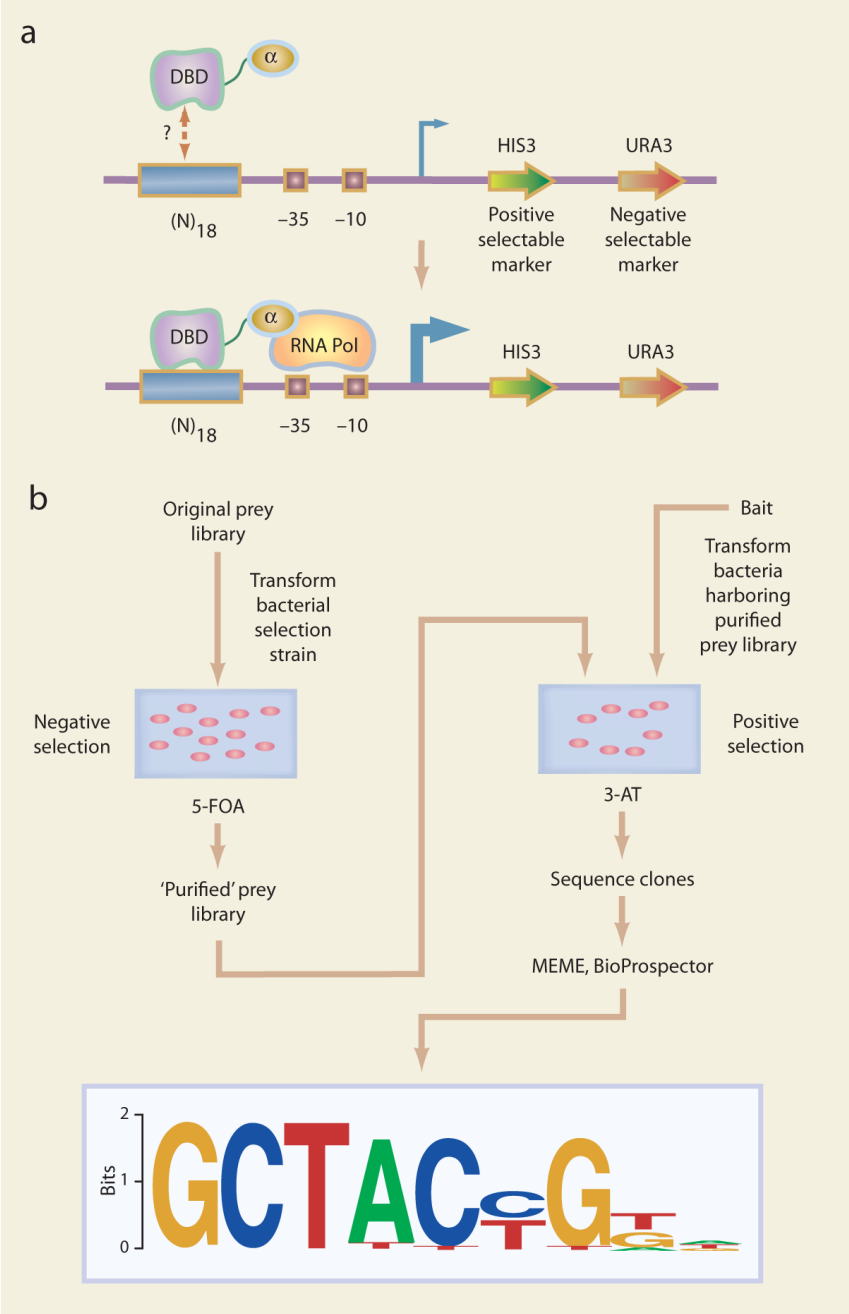

Figure 1.

Schematic diagram of the bacterial one-hybrid selection. (A) A library of randomized 18 bp oligonucleotides are cloned upstream of the *HIS3* (positive) and *URA3* (negative) selectable markers, in a bacterial strain lacking the bacterial *HIS3* and *URA3* homologs (*hisB* and *pyrF*, respectively). A plasmid containing the DNA-binding domain (DBD) of a query transcription factor fused to the alpha subunit of RNA polymerase (the “bait”) is then transformed into bacteria harboring the prey library. If the query DNA-binding domain interacts with a prey DNA sequence, then RNA polymerase is recruited, resulting in the expression of the *HIS3* and *URA3* selectable marker genes. (Adapted from Meng *et al*.) (B) The original prey library of candidate DNA binding sites undergoes a round of negative selection on plates containing 5-

FOA, in order to reduce the proportion of self-activating sequences. The resulting “purified” prey library is then transformed with the bait plasmid, and the bacteria then undergo positive selection at a range of stringencies by growing the cells on a series of plates spanning a range of 3-AT concentrations. Prey from individual colonies are then isolated and sequenced. The prey sequences are then examined with motif finding tools (MEME, BioProspector) in order to identify the DNA binding site motif of the query transcription factor.